一个分词器统领全部：跨语言分词器引发的语言可塑性

Diana Abagyan[★1], Alejandro R. Salamanca[1], Andres Felipe Cruz-Salinas[2], Kris Cao[2], Hangyu Lin[2], Acyr Locatelli[2], Marzieh Fadaee[1], Ahmet Üstün[♦1], and Sara Hooker[♦1]

[1]Cohere Labs, [2]Cohere

Corresponding authors: { dianaabagyan, ahmet, sarahooker } @cohere.com

Pretraining massively multilingual Large Language Models (LLMs) for many languages at once is challenging due to limited model capacity, scarce high-quality data, and compute constraints. Moreover, the lack of language coverage of the tokenizer makes it harder to address the gap for new languages purely at the post-training stage. In this work, we study what relatively cheap interventions early on in training improve "language plasticity", or adaptation capabilities of the model post-training to new languages. We focus on tokenizer design and propose using a *universal* tokenizer that is trained for more languages than the primary pretraining languages to enable efficient adaptation in expanding language coverage after pretraining. Our systematic experiments across diverse groups of languages and different training strategies show that a universal tokenizer enables significantly higher language adaptation, with up to 20.2 % increase in win rates compared to tokenizers specific to pretraining languages. Furthermore, a universal tokenizer also leads to better plasticity towards languages that are completely unseen in the tokenizer and pretraining, by up to 5 % win rate gain. We achieve this adaptation to an expanded set of languages with minimal compromise in performance on the majority of languages included in pre-training.

引言

只有少数研究实验室拥有足够的计算资源和专业知识来大规模训练大型 AI 系统 [Maslej et al., 2025; Hooker, 2024] 。大多数研究人员和从业者被迫在可用的预训练模型中选择用于下游任务，即使这些模型未必符合他们的使用案例。在多语言设置中，这种紧张局势表现得尤为明显 [Joshi et al., 2020; Singh et al., 2024; Üstün et al., 2024] ，在预训练中对多语言支持的有限投资往往导致在最先进的大型语言模型中出现显著的语言覆盖不足的问题 [Holtermann et al., 2024] 。

语言覆盖的不平衡导致特定语言用户使用成本越来越高，因为边缘化语言需要更多的 tokens，并且在生成过程中延迟更高 [Ji et al., 2023; Cui et al., 2024; Ahia et al., 2023] ，限制低表现语言的使用者只能使用质量较低的技术 [Held et al., 2023; Durmus et al., 2024; Nicholas & Bhatia, 2023; Ojo et al., 2025] 。进一步加剧这些问题的是，一旦模型被预训练，仅靠后训练很难引导其朝向新

★First author.
♦Principal senior advisors.

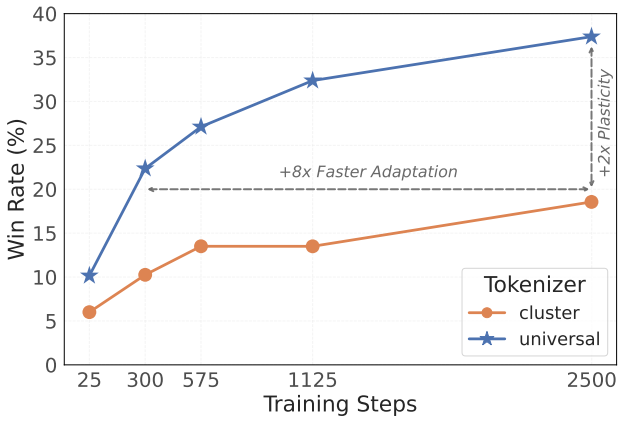


Figure 1: UNIVERSAL 分词器表现出比集群特定的基线分词器高 2 倍的可塑性，并且适应速度提高 8 倍。在继续进行包括主要语言子集和扩展语言子集（§ 2.1 ）的预训练过程中，扩展（新）语言子集的平均胜率。

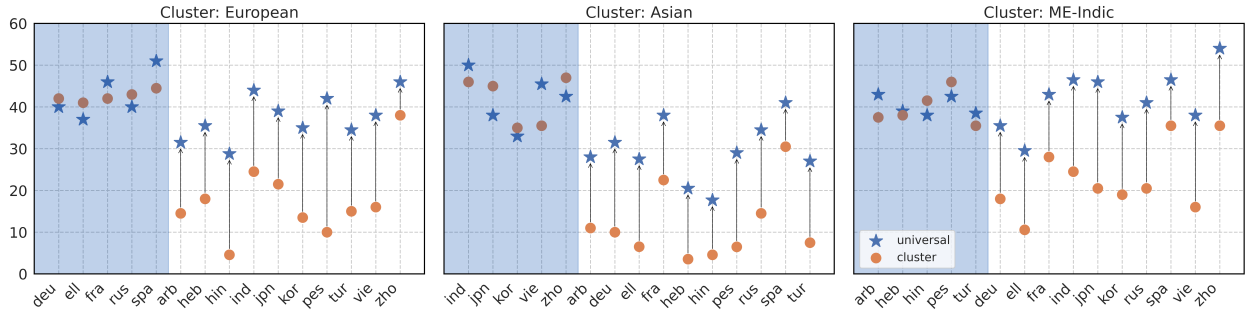


Figure 2: 使用 UNIVERSAL 和 CLUSTER 词汇表训练的模型对抗 Dolly 生成的 [Singh et al., 2024] 获胜率。阴影的蓝色部分代表主要语言的子集，白色部分代表扩展语言。与跨簇的 CLUSTER 词汇表相比，UNIVERSAL 词汇表在扩展子集语言中平均提高了 18.9 % 的获胜率，并在主要语言中提高了 0.3 %。

的行为 [Wang et al., 2025] 。除非在训练中令标记器已针对某种新语言进行校准，否则它通常需要更多的数据和复杂的优化步骤 [Muller et al., 2021] 。

多语言可塑性代表了语言模型能够快速适应语言分布变化以到达下游目标的能力，在我们的案例中，这涉及一组新的焦点语言 [Chen et al., 2023] 。考虑到预训练需要占用大量的计算和成本资源，因此在这个阶段所做的任何能提高下游开发者和研究人员可塑性的干预都是有益的。

在这项工作中，我们研究了最小且高效的预训练干预措施，以降低后续适应成本。特别是，我们将分词识别为具有相对较低干预成本但潜在大下游收益的领域。我们问道：我们能否利用具有广泛语言覆盖率的分词器来提高 LLM 的可塑性，而不影响预训练性能？

我们假设，一个在预训练开始时就引入的，训练于比主要预训练语言更多的语言上的通用分词器，可以快速有效地干预以适应模型到新语言。这显著地不同于以前的工作，该工作专注于预训练后的技术，例如词汇扩展 [Wang et al., 2020] 或重新训练嵌入层 [Artetxe et al., 2020] 。这些技术成本较高，需要像训练预算这样的更多资源，并且在不同语言中取得的成功程度各不相同 [Limisiewicz et al., 2023; Sharthak et al., 2025; Nag et al., 2025] 。

我们系统地通过在预训练规模上进行全面的消融研究，调查多语言分词器的影响，这需要大量的资源投入。我们在 69 种语言中对分词器、语言子集和适应策略进行变更，发现以下结果：

1. UNIVERSAL 标记器显著提高了对新（扩展）语言的适应性，在持续预训练实验中，相较于

专门针对预训练语言的基线标记器，平均胜率提高了 19 %。除了获得更高的适应性之外，UNIVERSAL 标记器在主要语言上的表现几乎相同，在下游评估中与基线标记器相比，差异不超过 2 %。

2. 对于目标适应性调整，其中新语言是唯一的重点，UNIVERSAL 分词器在扩展语言子集上比基线分词器平均提高了 14.6 %。另外，对于完全未见过的语言的适应，这些语言既不包含在分词器中也不包含在预训练中（适应性的最极端情况），UNIVERSAL 分词器在 7 种资源极其匮乏的语言中在胜率上最多可超过基线分词器 5 %。

3. 我们发现，UNIVERSAL tokenizer 能够使适应性能提高 8 倍以上，所需的额外训练大大减少，因此成本也降到最低。我们认为，这极大地有利于那些希望以最小干预来扩展预训练模型语言覆盖范围的从业者。

方法和实验设置

方法论与核心消融

语言覆盖和模型变体。我们的实验包括 62 种类型学和词典多样的语言，这些语言被分成三个具有地理动机的簇：(1) 欧洲语言，(2) 亚洲语言，以及 (3) 中东和印度语言（在全篇中称为 ME-Indic）。对于每个地理簇，我们主要在该簇内的语言上预训练语言模型（称为主要子集），并使用其余语言（称为扩展子集）作为可塑性适应实验的参考点。例如，对于欧洲簇，主要子集包括西班牙语、俄语和葡萄牙语等语言，而扩展子集包括在该簇主流训练数据之外的 10 种语言。此外，我们还考虑了 7 种完全未见过的语言，如僧伽罗语和哈萨克语，这些语言不在分词器或基础模型训练数据中。带有簇的完整语言列表在附录 D 中提供。

适应策略。 我们的目标之一是引入具有高度可塑性和适应性的模型属性。实践者可以选择使用各种不同的训练策略进行语言适应，这取决于他们可用的数据。理想情况下，我们在分词器上做出的选择允许在预训练后的任何方法下改善可塑性。因此，我们在不同的适应策略下评估我们的干预措施，这些策略包括使用主要语言和扩展语言子集中的数据进行继续预训练、针对扩展语言的定向适应以及针对完全未知语言的定向适应。下面我们简要描述这些策略和实验细节：

- 使用主要和扩展语言的数据进行持续预训练：这一策略的目标是增加模型的语言覆盖率，以便支持主要和扩展语言。训练组合的一半由指令微调数据中所有语言的均匀分布构成，另一半是由高质量数据集组成的标准冷却组合（见 § 2.2 ）。这赋予了我们的基础模型遵循指令的能力，并且还允许对主要和扩展语言进行评估。

- 目标适应（扩展语言）：在这些实验中，我们通过监督微调来探索针对特定语言的适应。后续训练数据仅包括每个集群模型的扩展语言子集中的指令风格数据。这使我们能够隔离那些在预训练期间未被关注但在分词器中有所代表的新语言的引入效果。

- 目标适应（完全未知语言）：在最后一组实验中，我们探索了目标适应最极端的设置，即完全未知语言，这些语言在分词器或预训练中都没有出现。在这种情况下，我们考虑每个实验中仅有一种语言的数据可用，因此我们每次针对一种语言微调基础模型。这种切割实验使我们能够在资源极其匮乏的情况下评估我们的方法的适应性。

分词器变体。我们训练了一个大规模的多语言分词器，使用所有 62 种语言的数据，以及仅表示主要语言子集的特定集群分词器。在整篇论文中，我们分别将这些分词器称为 UNIVERSAL 和 CLUSTER 分词器。我们在第 2.3 节中包括了更多关于分词器训练的细节。

实验装置

预训练数据集。模型使用由英语、代码和多语言语料库混合而成的数据进行预训练，其中数据权重分别分配为 55 %、15 % 和 30 %。在多语言训练中提高英语的权重是一种常见做法，因为其任务覆盖率和质量较高，这对于跨语言迁移来说至关重要 [Dash et al., 2025; Singh et al., 2024; Ravisankar et al., 2025] 。我们还包括代码数据，因为即便是自然语言模型，它也已成为训练配方的标准部分，并且发现将其纳入预训练可提升其他任务的表现 [MA et al.; Aryabumi et al., 2024] 。我们使用来自各种公开和专有来源的大量数据语料库。对于使用 UNIVERSAL 分词器进行预训练的模型，我们将训练数据中 5 % 的权重从英文数据重新分配，并均匀分布到所有扩展语言中，以避免词汇表中子词训练不足的情况 [Land & Bartolo, 2024] 。然而，在第 5.4 节中，我们消除了该百分比并显示即使预训练中未包含任何扩展语言子集数据，UNIVERSAL 分词器也显著提升了多语言的灵活性。

降温和指导数据集。 在持续预训练中，我们使用降温数据，这些数据涉及增加高质量数据集的权重，包括文本、数学、代码和指导风格的数据 [Aryabumi et al., 2024] 。最近的研究发现这种方式可以提高下游任务的性能，特别是通过帮助赋予指令跟随能力 [Parmar et al., 2024; Team et al., 2025] 。我们包括了专有和开放数据的高质量混合，其中许多数据是通过多语言数据套利策略创建的 [Odumakinde et al., 2024] ，涵盖了 23 种语言的 100,000 个提示-完成对。最后，在完全看不见的语言实验中，我们模拟一个现实的受限数据场景，仅使用来自 Aya Collection 的 Dolly 训练集翻译的每种语言的 14,800 个指令 [Singh et al., 2024] 。

训练细节。对于我们的实验，如同大多数大型语言模型，我们使用基于 Transformer 的解码器架构 [Vaswani et al., 2017; Radford & Narasimhan, 2018] 。我们的架构包括关键优化措施，例如并行注意力模块 [Chowdhery et al., 2023] 、分组查询注意力 [Ainslie et al., 2023] 、SwiGLU 激活函数 [Shazeer, 2020] 和旋转位置嵌入 [Su et al., 2024] 。

我们的消融实验是广泛的，并且需要大量的预训练运行。鉴于预训练所需的计算量巨大，比如在 128 个 Nvidia H100 GPU 上训练一个 3.3B 参数的模型需要 11 个小时，因此我们只专注于 3.3 亿参数的语言模型进行消融实验。我们在 25,000 步中对每个基本变体训练了 1000 亿个 tokens。考虑到我们运行的实验数量以及评估的多种因素，这个模型大小和训练步数已接近预训练规模下计算可行的极限。总体而言，目标不是模拟一次完整的预训练运行设置，而是获取关于不同方法相对优缺点的充分信号。在持续预训练策略中，我们额外训练了 10.5B 个 tokens，并在各个实验的相应数据集上进行 4 个 epoch 的目标适应。附录 B 中提供了更多的训练和基础设施细节。

基础设施。我们使用 Nvidia H100 GPU 进行训练和评估。我们使用 FAX [Yoo et al., 2022] ，这是一个建立在 JAX [Bradbury et al., 2018] 基础上的高性能训练框架，支持高效的张量和模型并行。

分词器训练

所有的分词器都是使用字节对编码算法 [Sennrich et al., 2016] 训练的。关于分词器训练的更多实现细节在附录 A 中给出。

语言权重。除了改变分词器的覆盖率之外，我们还采取了一种基于数据可用性调整权重的方法，与对某些被训练在 UNIVERSAL 和 CLUSTER 语言覆盖率的处理不同。与传统的在所有数据中进行均匀采样并因最常用语言而占主导地位的方法不同，我们考虑两个因素：(1) 在不同语言中可用数据的自然分布，以及 (2) 由共享同一家族和文字（更可能共享词汇）的语言构成的语言桶。在每个语言桶中，我们在所有语言之间使用均匀的权重。具体而言，对于语言 i ，其中 w_i^d 和 w_i^b 分别表示数据分布和语言桶的权重，我们在分词器数据混合中计算语言权重的方法如下：

$$w_i = \frac{w_i^d . w_i^b}{\sum_n w_n^d . w_n^b} \tag{1}$$

通过这种方式，我们以一种原则性的方法平衡自然数据分布（由于高资源语言而倾斜）与语言分桶，确保对多种字符体系和低资源语言有公平的代表性。我们的预训练实验（第 3 节，附录 **??** ）表明，我们结合语言分桶与按大小比例的数据分布的专门加权，能够实现比统一加权更好的压缩率，并在下游任务中取得更好的性能。在本研究的剩余部分中，除非另有说明，我们在实验中一直使用专门加权。

词汇量。在我们的主要实验中，我们使用了 250k 个标记的词汇。但在 § 5.3 中，我们探索了词汇量如何影响 UNIVERSAL 和 CLUSTER 标记器的性能，并将词汇量在 100k、175k 和 250k 之间变动以理解其影响。

评价

开放式评估。Goldman et al. [2024] 发现生成任务比分类在评估分词器时更有信息性，这可能是由于生成步骤的数量。根据 Üstün et al. [2024] ，生成的质量使用 LLM-as-a-Judge 胜率来评估，其中原始生成用作参考答案。我们使用 Aya 评估数据集 [Singh et al., 2024] 的 `dolly_human_edited` 和 `dolly_machine_translated` 切分作为该任务的测试数据，这些数据通过翻译 Dolly-15k [Conover et al., 2023] 的 200 个保留示例形成。我们使用 15 种适应语言进行开放式评估，这些语言列在附录 D 中。

先前的研究表明，作为评价者的大型语言模型（LLM）是合理的代理，并且在多语言环境中也与人类偏好一致 [Üstün et al., 2024; Singh et al., 2025; Dang et al., 2024; Kreutzer et al., 2025] 。我们使用 Command-A [Cohere et al., 2025] 作为判断模型，考虑到其在多语言环境中作为最佳开源权重判断器的显著表现，得分接近 GPT4o [Gureja et al., 2024; Pombal et al., 2025] 。完整的判断提示在附录 C.3 中提供。

任务特定性能。我们针对多语言评估使用了两个任务特定的评估方法。Belebele [Bandarkar et al., 2024] 是一个代表 122 种语言变体的多项选择题机器阅读理解（MRC）数据集。多语言 MMLU（M-MMLU）[Dac Lai et al., 2023] 是原始 MMLU 数据集 [Hendrycks et al., 2021] 的机器翻译版本，包含从 STEM 到人文学科等主题的问题。

英文专用评估。 此外，我们还在 11 个仅限英语的自然语言推理和常识推理基准上评估模型：ARC-C 和 ARC-E [Chollet, 2019] ， BoolQ [Clark et al., 2019] ， CommonsenseQA [Talmor et al., 2019] ，Hellaswag [Zellers et al., 2019] ，MMLU [Hendrycks et al., 2021] ，OpenBookQA [Mihaylov et al., 2018] ， PIQA [Bisk et al., 2020] ， SIQA [Sap et al., 2019] ， TruthfulQA [Lin et al., 2022] ，和 WinoGrande [Sakaguchi et al., 2019] 。

我们包括特定任务的评估（包括多语言和仅限英语）以理解不同设计选择的相对优点。通常，此时的预训练模型在下游任务中表现不佳，因为模型尚未针对遵循指令 [Wang et al., 2022; Üstün et al., 2024; Aakanksha et al., 2024] 进行优化，或者使用强化学习 [Ahmadian et al., 2024; Dang et al., 2024] 进行对齐。因此，我们并不期望最先进的性能，而是评估不同变体的相对信号。

Cluster	Tokenizer	Belebele	M-MMLU	EN Tasks
		Primary Languages		
European	Cluster	41.4	31.1	48.5
	Universal	41.9	30.9	48.4
Asian	Cluster	38.2	29.6	48.2
	Universal	38.1	28.9	48.1
Middle East & Indic	Cluster	38.1	29.2	49.1
	Universal	36.5	28.6	48.2

Table 1: 在三个区域集群的主要语言预训练期间，比较 Cluster 与 Universal 分词器的表现。使用 Universal 分词器的表现与全地理集群模型中的 Cluster 分词器表现相当。

	bul	cat	ces	dan	deu	ell	est	eus	fin	fra	hrv	hun	ita	lit
Uniform	41.0	43.7	42.7	41.5	43.6	41.2	35.7	38.4	34.6	45.0	42.1	36.6	40.6	41.0
Universal	42.1	46.5	45.5	41.1	44.3	42.7	37.9	40.6	32.4	45.3	42.6	37.9	40.1	43.1
	(+1.1)	(+2.8)	(+2.8)	(–0.4)	(+0.7)	(+1.5)	(+2.2)	(+2.2)	(–2.2)	(+0.3)	(+0.5)	(+1.3)	(–0.5)	(+2.1)
	lvs	nld	nob	pol	por	ron	rus	slk	slv	spa	srp	swe	ukr	Average
Uniform	42.3	42.7	42.3	38.4	41.0	40.8	41.0	42.4	39.5	42.5	42.4	42.9	40.9	41.0
Universal	40.5	42.7	42.8	39.8	43.8	41.2	41.4	43.5	41.8	43.8	43.4	43.4	40.0	41.9
	(–1.8)	(+0.0)	(+0.5)	(+1.4)	(+2.8)	(+0.4)	(+0.4)	(+1.1)	(+2.3)	(+1.3)	(+1.0)	(+0.5)	(–0.9)	(+0.9)

Table 2: 在用于欧盟集群模型的预训练时，比较 Universal 与 Uniform 标记器在 Belebele 上的表现。在标记器训练中，我们使用了所有语言（67 种语言；主要和扩展子集），仅改变语言权重。Universal 标记器使用语言存储桶的平衡权重，在 27 种欧洲语言中优于 Uniform 权重，其中 21 种获得了 2.2 相对增益%（平均 41.9 vs 41.0）。

预训练性能结果

- Pretraining with Universal tokenizer produces well-rounded models with competitive results on a variety of tasks for the primary languages, differing from the Cluster tokenizer models by no more than 0.5 % average accuracy across tasks and clusters.
- Our specialized tokenizer weighting that uses language buckets to balance the data availability, leads to better pretraining performance up to 2.8 accuracy increase as measured in the Euro cluster model.

在本节中，我们首先对我们的预训练模型的性能进行基准测试，以确保使用 Universal 标记器不会导致性能下降，因为使用针对比主要集合更广泛的语言优化的标记器时，可能会预期出现性能下降。

Universal tokenizer 在主要语言上不影响性能。正如表中所示 1 ，我们发现我们扩展的 Universal tokenizer 在地理集群模型中与 Cluster 相比具有显著的竞争力。在英文任务中，预训练性能的差异最多不超过 1 % 的平均准确率。对于多语言任务，Universal 和 Cluster tokenizer 之间的最高性能差异在 ME-Indic 集群的 Belebele 上仅为 1.6 % 的平均准确率（38.1 % 对 36.5 %）。总体而言，我们观察到在切换到 Universal tokenizer 时，主要集群语言的性能几乎没有权衡。

事实上，我们观察到 Universal 分词器在 Belebele 的欧元簇上平均表现略有提升（41.9 对比 41.4），并且在亚洲簇上也取得了更接近的性能（38.1 对比 38.2）。作为额外验证，附录 C.1 中的图 6 显示了在预训练期间，欧元簇模型的两个分词器在 Belebele 平均性能上的进展。Universal 分词器在整个预训练过程中表现大致相近，这也暗示了在更长的预训练过程中有相同的趋势。总体而言，这些结果表明，使用 Universal 分词器在主要语言的预训练中并不会导致显著的性能下降。

Cluster	Tokenizer	Dolly Win Rates (%)	
		PRIMARY	EXPANDED
European	CLUSTER	42.8	17.6
	UNIVERSAL	42.8	37.4 (+19.9)
Asian	CLUSTER	41.7	11.7
	UNIVERSAL	41.8	29.5 (+17.8)
Middle East & Indic	CLUSTER	39.7	22.8
	UNIVERSAL	40.2	41.8 (+18.9)

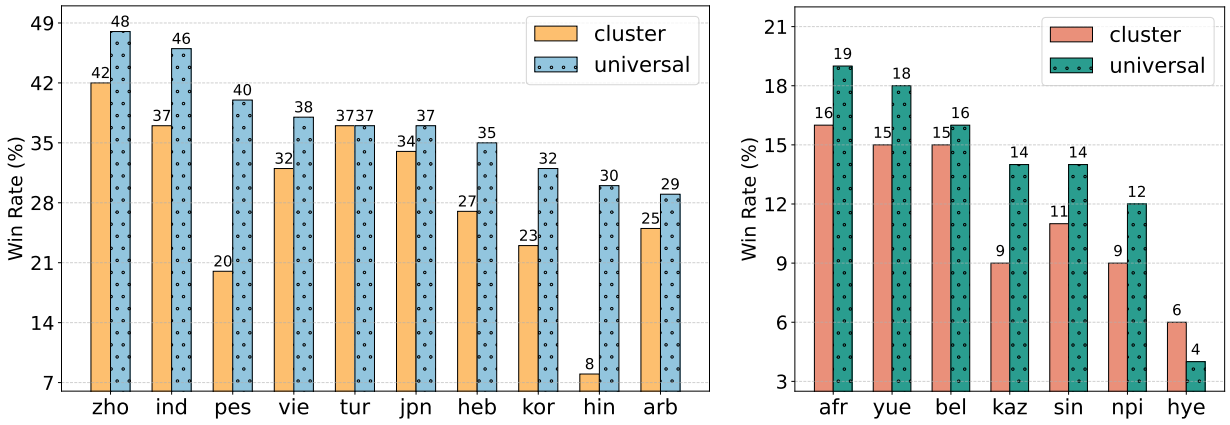
Table 3: 在主要和扩展语言子集上继续预训练后的胜率。UNIVERSAL 分词器在主要语言上的表现与 CLUSTER 相当，并在所有集群的扩展语言子集平均胜率上显示出显著提升（高达 19.9 %）。

	CLUSTER	UNIVERSAL
European	27.2	37.4 (+10.2)
Asian	18.8	34.3 (+15.7)
Middle East & Indic	23.31	41.1 (+17.8)

Table 4: 在目标适应后扩展语言的胜率。UNIVERSAL 标记器在所有簇中相比基线 CLUSTER 标记器显示出更好的性能（最高提升至 17.8%）。

在标记器训练中，使用语言桶对语言进行平衡加权能带来更好的预训练性能。正如 2.3 节中关于标记器训练所述，我们通过脚本和语言家族形成的桶来加权这些语言，以平衡数据的可获得性。为了证明这种加权方案的合理性，我们将 UNIVERSAL 标记器的预训练性能与一个基线标记器（UNIFORM ）进行比较，基线标记器中所有语言的权重是均等的，除了英语外。[1] 我们的消融试验是在欧洲集群中进行的，该集群中的主要语言数量是最多的。在标记器训练中，我们使用所有语言（62 种语言；包括主要和扩展子集），只改变语言的权重。正如表 2 所示，使用语言桶进行平衡加权的 UNIVERSAL 标记器在 27 种欧洲语言中优于 UNIFORM 加权，在 21 种语言上相对提升了 2.2 %（平均为 41.9 对比 41.0）。进一步验证预训练结果，我们在附录 **??** 中提供了这两种标记器在压缩性能上的比较，结果显示 UNIVERSAL 标记器在整体压缩上表现更好。

[1]对于两种分词器的加权，我们使用固定的 30 % 的比例用于英语，因为英语的数据量更大且可用数据的多样性更高。这也保证了分词器加权之间的公平比较。



(a) 扩展语言　　(b) 完全未知的语言。

Figure 3: 通过 SFT 进行目标适应后，Euro 集群模型扩展到多语言 (a) 和完全未知语言 (b) 的语言特定结果。在这两个语言子集中，UNIVERSAL 分词器的表现优于 CLUSTER 分词器，其中相对于集群特定分词器的相对提升在扩展语言（印地语）中达到 22 %，在未知语言（哈萨克语）中达到 5 %。

增强多语言可塑性的研究结果

持续预训练中可塑性的好处

- The UNIVERSAL tokenizer leads to significantly higher plasticity over the CLUSTER specific tokenizers in continued pretraining (on primary & expanded languages) with an average increase in win rate of 18.9 % for the expanded language subsets across clusters.
- The UNIVERSAL tokenizer enables adaptation boost with no drop in primary languages compared to CLUSTER tokenizer, with a near identical performance (with a slight increase of 0.3 % on average) on downstream open-ended generations.

在本节中，我们提出一个问题：在对主要和扩展语言进行持续预训练后，改变分词器的方法是否会带来可塑性优势?

使用 UNIVERSAL 标记器训练的模型在 EXPANDED SUBSET 上表现出显著更高的胜率。图 2 和表 3 显示了在欧洲、亚洲和中东-印度语系簇的评估结果，每个 PRIMARY LANGUAGE SUBSET 有 5 种语言，而 EXPANDED SUBSET 的 10 种语言属于其他两个簇。我们看到，与使用 CLUSTER 标记器训练的模型相比，UNIVERSAL 标记器在扩展子集上的所有三个地理簇模型中，胜率平均提高了 18.9 %。各个簇的改进是一致的，其中欧洲、亚洲和中东-印度语系簇模型的胜率分别提高了 +19.9 %、+17.8 % 和 +18.9 %。在所有扩展语言中，波斯语 (+25.8 %)、印地语 (+23.3 %) 和越南语 (+22.0 %) 分别在欧洲、亚洲和中东-印度语系簇中显示出从 UNIVERSAL 标记器中获得的最大益处。

UNIVERSAL 分词器在多个集群上保留了性能。虽然 UNIVERSAL 分词器在扩展的语言上提供了显著的增益，但在图 2 中，我们观察到在所有三个集群中主要语言的性能在两个分词器之间几乎相同。在比较分词器时，所有集群中的主要语言的胜率差异仅为 0.3 %，其中 UNIVERSAL 分词器甚至在欧洲、亚洲和 ME-Indic 模型中分别比 CLUSTER 分词器略高 0.3 %、0.1 % 和 0.5 %。这是有利的，因为这表明在开发模型时，使用 UNIVERSAL 分词器不会对一个提供者感兴趣的主要语言的扩展语言集合的可塑性改进产生权衡。

靶向适应中的可塑性益处

- For the targeted language adaptation through SFT, where only the expanded language data is introduced, the UNIVERSAL tokenizer is much more performant than the CLUSTER tokenizer with 14.6 % average increase across languages and geo-clusters.
- UNIVERSAL tokenizer enables multilingual plasticity not only for languages seen during tokenizer training but also for fully UNSEEN languages with an average improvement of 2 % on 7 under-resourced languages over the CLUSTER tokenizer.

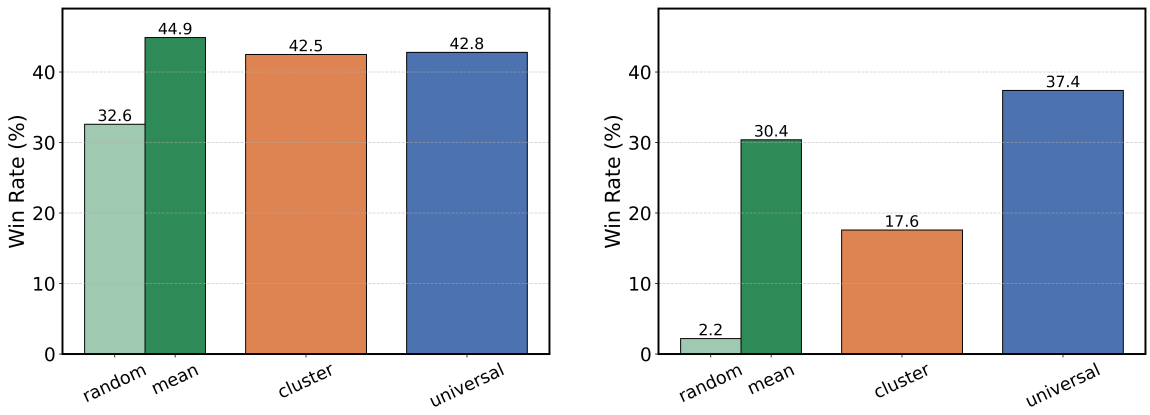
一个非常值得关注的实验设置是更为现实的情境，其中下游开发者仅能访问 EXPANDED 语言的数据。为了模拟这种情境，我们在仅对扩展的语言子集进行有监督的微调是可行的情况下，评估我们干预措施的影响。

在扩展语言集的目标适应中，UNIVERSAL 分词器相比 CLUSTER 分词器具有很高的优势。表 4 显示了每个地理集群中 UNIVERSAL 和 CLUSTER 分词器的平均胜率。UNIVERSAL 分词器在欧洲、亚洲和中东-印度集群中相比 CLUSTER 特定分词器分别实现了 10.2 %、15.7 % 和 17.8 % 的相对胜率提升。在图 3a 中，我们还绘制了欧洲集群中各语言的个体增益。相对于 CLUSTER 分词器，UNIVERSAL 一直能够实现更高的可塑性，其中印地语和波斯语中的相对增益分别高达 22.0 % 和 20.0 %。

UNIVERSAL 分词器在针对完全未见过的语言集进行目标适应时也带来了很大的收益。在最极端的设置中，我们评估了我们分词器干预在适应于分词器和预训练中完全未见过的语言时的好处。图 3b 显示了对 7 种未见过语言的监督微调实验结果。[2] 重要的是，所有这些语言的资源都极其匮乏，并且这种适应是在低数据环境中进行的，因为这代表了开发人员在这些语言中面临的限制。

在最极端的设置中，UNIVERSAL 标记器使得未见语言获得了提升。我们发现 UNIVERSAL 标记器在未见语言上相较于 CLUSTER 标记器实现了改进，平均提升了 2.0 个% 的胜率，其中在尼泊尔语上升至 5.0 个%。鉴于由于这些语言在标记器和预训练中缺乏出现而导致下游性能通常较低，以及数据可用性受限（每种语言仅 15k），我们认为这是未来研究的一个有前途的方向，也是投资于更灵活的标记器设计的另一个理由。

[2] 南非荷兰语、哈萨克语、白俄罗斯语、粤语、尼泊尔语、亚美尼亚语、僧伽罗语



(a) 主要语言　　(b) 扩展语言

Figure 4: 继续预训练后的胜率，比较通过标记替换的跨语言词汇适应 (CVA)。预训练后，Cluster 标记器被替换为 Universal 标记器，保留共享标记的嵌入，并且新标记要么随机初始化（`random`），要么通过共享嵌入的平均值初始化（`mean`）。Universal 标记器在扩展语言上显著超过了两种 CVA 方法。

关键讨论

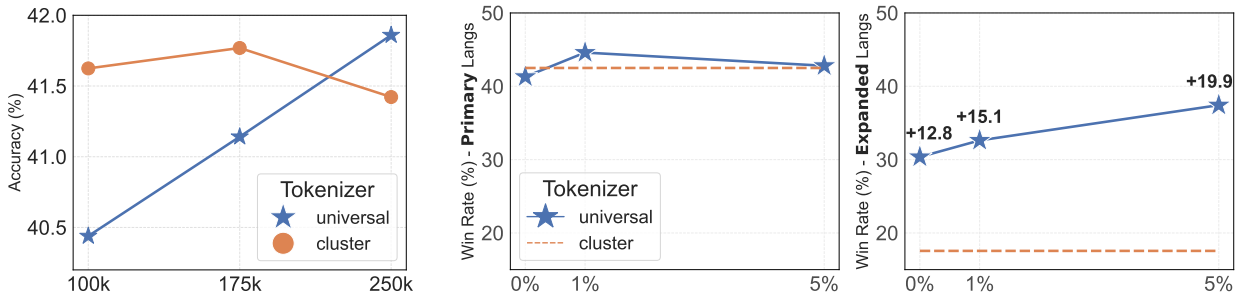
与跨语言词汇适应的比较

- Universal tokenizer outperforms cross-lingual vocabulary adaptation (CVA) by 7 % in the expanded languages where the vocabulary is replaced (before the continued pretraining), and the new tokens are initialized by average embeddings of the shared tokens.
- CVA through tokenizer replacement fails to achieve competitive performance even with the Cluster tokenizer when the new tokens are initialized randomly. CVA (`random`) only reaches 2.2 % win rate in the expanded languages and falls behind the Universal tokenizer by 35.2 % relative drop in win rate.

跨语言词汇适应 (CVA) [Yamaguchi et al., 2024b] 旨在通过扩展或替换现有的分词器以适应新语言，因此在预训练后，令牌嵌入也适应这些新语言，是语言适应的常用方法。CVA 的更详细概述可以在第 6.2 节中找到。在这个消融实验中，我们提出问题：Universal 分词器与 CVA 相比，在适应新语言方面表现如何？

为了与我们的一系列实验有一个完全可比较的设置，我们采用了使用 Cluster 分词器训练的预训练的 Euro 集群模型，并将分词器替换为 Universal 分词器。Cluster 和 Universal 分词器之间共享的标记嵌入被保留，而新标记要么通过从正态分布中采样随机初始化（`random`），要么是共享嵌入的平均值（`mean`）。在词汇替换和标记初始化之后，我们遵循第 2.1 节中描述的相同继续预训练过程。

这次消融实验的结果如图 4 所示。我们发现，当随机初始化新标记时，CVA（标记器替换）即使与 Cluster 标记器相比也难以达到相当的性能，并且显著落后于 Universal 标记器，在主要语言和扩展语言中的胜率分别相差 15.4 % 和 35.2 %。值得注意的是，用共享词汇的均值（图 4 中的 `mean`）初始化新标记时，表现优于随机初始化。虽然标记器替换（`mean`）较未经适配的 Cluster 标记器在扩展语言中的胜率相对提高了 12.8 %，但我们的 Universal 标记器通过平均胜率提升 7 %（37.4 % 对比 30.4 %）实现了更好的适配性能。有趣的是，CVA（`mean`）在主要语言中平均胜率稍微提高了 2.1 %。总体而言，这些结果表明，从一开始就使用 Universal 标记器比在预训练



(a) 词汇量的影响

(b) 在预训练中扩展语言数据的效果

Figure 5: (a) UNIVERSAL 分词器需要更大的词汇量才能在主语言上取得与 CLUSTER 分词器相同（或更好）的预训练性能，正如在 Belebele 中评估的那样。(b) 即使在预训练中没有添加来自扩展语言的数据，UNIVERSAL 分词器相对于 CLUSTER 分词器也表现出显著更高的适应增益。

后再替换更为有效。

通用分词器的适应效率

- UNIVERSAL tokenizer enables +8x faster adaptation in terms of sample efficiency and +2x higher performance for downstream adaptation compared to CLUSTER tokenizer.

在这个消融实验中，我们评估使用 UNIVERSAL 分词器进行适应的速度有多快。更快的适应意味着需要的资源更少，即成本，这对希望适应大语言模型以扩展语言覆盖的从业者来说是非常感兴趣的。

为了评估适应速度，与 CLUSTER 分词器相比，我们在继续对 Euro 集群模型进行预训练期间，对扩展语言的中间检查点进行了评估。图 1 显示了 10 种扩展语言的平均胜率。如图所示，仅需 300 步，UNIVERSAL 分词器就达到了 CLUSTER 在 2500 步时的性能水平，显示出 +8 倍更快的适应速度。鉴于 300 步近似对应于 150K 个样本（相比于 2500 步时的 1.3M），UNIVERSAL 分词器需要更少的数据就能达到与基线最终性能相同的表现，证实了我们建议方案的有效性。

大词汇量的必要性

- The UNIVERSAL tokenizer demonstrates effectiveness without compromising pretraining performance for primary languages, but it necessitates a large vocabulary size, such as 250,000 subwords, to achieve optimal results.

在之前的章节中，我们通过 UNIVERSAL 与 CLUSTER 的分词器比较，建立了在多语言可塑性方面的更高性能。在此消融实验中，我们问：为了避免在主要预训练语言上的性能下降，UNIVERSAL 分词器所需的词汇表大小是多少？

为了确定最佳词汇量大小，我们进行了额外的预训练实验，其中我们将词汇量大小从 100,000 个词标调整到 250,000 个词标，同时调整模型参数，以确保训练参数的总数量保持不变。我们评估了主要预训练语言在 Belebele 上的表现。结果如图 5a 所示。使用 CLUSTER 标记器训练的模型性能没有太大变化，并且在小词汇量大小（100k 和 175k）时超越了 UNIVERSAL 标记器。然而，UNIVERSAL 标记器的性能随着词汇量的增加而提升，并在词汇量大小为 250k 时超越了 CLUSTER 标记器。我们的发现与先前的研究一致，表明大词汇量的益处 [Tao et al., 2024; Huang et al., 2025]

，并建议对于通用标记器的投资需要重新分配权重以确保适当的词汇预算。根据这个消融实验，我们在主要预训练运行中使用 250k 的词汇量大小。

预训练中扩展语言子集的存在

- The UNIVERSAL tokenizer achieves a performance boost of 12.8 % over the CLUSTER tokenizer for the expanded languages even there is no pretraining data is used for these languages. However, including a minimal data up to 5 % increases adaptation performance on the expanded languages from 12.8 % to 19.8 % win rates, without hurting performance in primary pretraining languages.

在用于训练大型语言模型的大型且通常嘈杂的数据集中，语言常常会受到污染 [Blevins & Zettlemoyer, 2022] 。因此，声称一种语言是真正“新”的可能比较困难。在我们的最后一次对比实验中，为了在多语言数据存在的不同假设下测试我们关于可塑性的主张的稳健性，我们对 0 % 、1 % 和 5 % 在预训练中扩大语言的比例进行了评估，针对欧洲集群。

图 5b 显示，即使在新语言（扩展子集）中使用 0 % 多语言比例的最保守情境下，UNIVERSAL 分词器的获胜率比 CLUSTER 分词器提高了 12.8 %。值得注意的是，将该比例增加到 5 % 不会影响主要预训练语言的性能，但会将扩展语言的适应性能从 12.8 % 的获胜率提高到 19.8 %。

相关工作

多语言分词器

分词仍然是多语言模型的一个关键挑战，特别是由于不同脚本之间的低效和差异。Petrov et al. [2023] 显示，许多标准的或以英语为中心的分词器不成比例地分割非拉丁脚本，有时会生成多达相当于英语内容 15 倍的标记。这些不平衡可能导致实际后果，例如增加 API 使用成本 [Ahia et al., 2023; Petrov et al., 2023] ，延长推理时间 [Hofmann et al., 2022; Sun et al., 2023] ，并减少非英语语言的可用上下文窗口 [Velayuthan & Sarveswaran, 2025; Ahia et al., 2023] ，以及降低下游任务性能 [Goldman et al., 2024; Gow-Smith et al., 2022; Fujii et al., 2023] 。此外，非英语语言的低效分词可能会使训练成本膨胀高达 68 % [Ali et al., 2024] 。

为了支持多语言性，mT5 模型引入了一个具有字节回退功能的 250k SentencePiece 词汇，使用温度采样法预训练了 101 种语言，以平衡高资源和低资源语言 [Xue et al., 2021] 。此外，最近的研究探索了旨在更好地反映多样化书写系统结构的分词方法。例如，基于字形的分词，使用 Unicode 字形簇作为原子单位，通过诸如字形对编码（GPE）[Velayuthan & Sarveswaran, 2025] 或 MYTE 之类的方法，一种基于词素的分割策略 [Limisiewicz et al., 2024] ，提供了对复杂书写系统的更好表示。

语言适配后训练

语言模型（PLMs）的语言适应通常有多种方法。在额外训练方面，继续预训练（CPT）和监督微调（SFT）是最常用的方法。继续预训练涉及对适应语言语料库的扩展训练 [Han & Eisenstein, 2019; Muller et al., 2021] ，但这需要大量的数据，对于资源较低的语言来说可能无法获得。监督微调也是一个标准方法 [Kumar et al., 2022; Adelani et al., 2021; Cahyawijaya et al., 2021] ，所需数据比 CPT 少，但可能导致遗忘预训练能力的风险 [Rolnick et al., 2019; Chaudhry et al., 2019] 。特别是，指令微调因其能传授指令跟随能力而受到欢迎 [Gala et al., 2024] 。关于两者优劣的说法是混合的——Ebrahimi & Kann [2021] 发现，在他们的设置中，CPT 比 SFT 更有效，而 Yong et al. [2023] 发现相反的结果。

一个主要的挑战是在分词器中不支持某些脚本和语言。跨语言词汇适应（CVA）通过修改现有的分词器来容纳额外的语言 [Yamaguchi et al., 2024a] ，并需要在目标语言中继续预训练以充分适应 [Fujii et al., 2024] 。CV 有两种常见的方法—词汇扩展，即从目标语言中添加新词并重新使用共享词汇 [Wang et al., 2020; Pfeiffer et al., 2021] ，或词汇替换，即完全替换词汇。新词对应的词嵌入可以随机初始化，使用诸如原始词汇中的一些对应词的平均值的启发式方法 [Minixhofer et al., 2022; Dobler & de Melo, 2023; Downey et al., 2023] ，或基于辅助模型 [Ostendorff & Rehm, 2023] 。更换分词器是麻烦的；一种可能的方法是通过训练一个超网络，将新分词器的词汇映射到现有词嵌入 [Minixhofer et al., 2025] 。这种方法需要继续训练以缩小性能差距，但即便如此也不能超越它。也有建议将语言音译为拉丁字符以规避不支持的脚本 [Muller et al., 2021] ，但这种方法受限于音译的表现。

在未见过的语言中也可以实现具有竞争力的表现，而无需额外的训练或微调步骤。Tanzer et al. [2024] 利用最先进语言模型的长上下文长度，在极低资源、完全未见过的语言中解锁翻译能力，包括在提示中使用该语言的语法书中的上下文。Cahyawijaya et al. [2024] 跨语言使用上下文学习，在提示中引用来自不同高资源语言的任务示例，以实现任务特定能力向低资源语言的迁移。

在这项工作中，我们探讨了在预训练中哪些廉价的干预措施可以在下游优化阶段增加灵活性。我们进行了一项广泛的研究，涉及不同的分词策略、三种语言适应策略，并考虑了跨越 70 种不同语言的数据访问的不同假设。我们发现，在所有情况下，使用具有广泛语言覆盖的 UNIVERSAL 分词器训练的模型能够更好地适应主要预训练集之外的语言，平均胜率提高高达 20.2 %，并且在扩展语

言的定向适应中提高了 17.8 %。即使在完全未见过语言的具有挑战性的低数据环境中，UNIVERSAL 分词器仍显示出高达 5 % 的增益。同时，对主要预训练语言的性能影响可以忽略不计。前期投资于一个庞大的多语言分词器在语言适应方面将是有回报的。

局限性

语言覆盖。我们的实验涉及 69 种语言，涵盖了多样的语言和文字系统，我们系统地研究了多语言数据对分词器训练的影响。虽然我们使用了一份综合的语言列表，但世界上还有更多的语言，这需要研究社区的关注。我们希望我们的工作能鼓励在最先进的语言模型中实现更广泛的语言覆盖。

分词算法。在这项工作中，我们仅关注用于分词器训练的 BPE 算法，这是一种在语言模型中最广泛使用的方法。这一选择是由于每次消融实验的高计算成本，需要大量的计算资源。然而，我们相信我们关于多语言覆盖的发现也适用于其他分词器，如 Unigram 分词器 [Kudo, 2018] 或字节或字符级别的分词 [Xue et al., 2022; Clark et al., 2022] 。我们将这部分探索留待未来的研究。

模型规模。我们所有的预训练实验都在拥有 3.3B 参数的模型上进行，使用的标记预算为 100B，这已经是一个对于资源和计算成本而言非常庞大的任务。鉴于我们的结果在此规模上是有效的，我们预计它们也适用于更大规模的模型和标记预算，正如之前的研究支持所示 [Biderman et al., 2023; Longpre et al., 2024; Aryabumi et al., 2024] 。

我们感谢 Julia Kreutzer、John Dang、Roman Castagné、Aakanksha 以及 Cohere 和 Cohere Labs 的其他同事对我们的支持和深入的反馈。同时，我们也感谢 Shayne Longpre 对预印本的反馈。

Aakanksha, Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. Mix data or merge models? optimizing for diverse multi-task learning, 2024. URL https://arxiv.org/abs/2410.10801.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D' souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. Masakhaner: Named entity recognition for african languages. Transactions of the Association for Computational Linguistics , 9:1116–1131, 10 2021.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pp. 9904–9923, Singapore, Decem-

ber 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. Tokenizer choice for LLM training: Negligible or crucial? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), Findings of the Association for Computational Linguistics: NAACL 2024 , pp. 3907–3924, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.247. URL https://aclanthology.org/2024.findings-naacl.247/.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training, 2024. URL https://arxiv.org/abs/2408.10914.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O' Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning , pp. 2397–2430. PMLR, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In Thirty-Fourth AAAI Conference on Artificial Intelligence , 2020.

Terra Blevins and Luke Zettlemoyer. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pp. 3563–3574, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.233. URL https://aclanthology.org/2022.emnlp-main.233/.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , pp. 8875–8898, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.699. URL https://aclanthology.org/2021.emnlp-main.699/.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. LLMs are few-shot in-context low-resource language learners. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pp. 405–433, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.24. URL https://aclanthology.org/2024.naacl-long.24/.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning, 2019. URL https://arxiv.org/abs/1902.10486.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Adelani, Pontus Lars Erik Saito Stenetorp, Sebastian Riedel, and Mikel Artetxe. Improving language plasticity via pretraining with active forgetting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), Advances in Neural Information Processing Systems , volume 36, pp. 31543–31557. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6450ea28ebbc8437bc38775157818172-Paper-Conference.pdf.

François Chollet. On the measure of intelligence, 2019. URL https://arxiv.org/abs/1911.01547.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research , 24(240): 1–113, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In NAACL , 2019.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. Transactions of the Association for Computational Linguistics , 10:73–91, 2022. doi: 10.1162/tacl_a_00448. URL https://aclanthology.org/2022.tacl-1.5/.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukas Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Chang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025. URL https://arxiv.org/abs/2504.00698.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca, 2024. URL https://arxiv.org/abs/2304.08177.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple

languages with reinforcement learning from human feedback. arXiv e-prints , pp. arXiv–2307, 2023.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya vision: Advancing the frontier of multilingual multimodality, 2025. URL https://arxiv.org/abs/2505.08751.

Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pp. 13440–13454, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.829. URL https://aclanthology.org/2023.emnlp-main.829/.

C.m. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In Duygu Ataman (ed.), Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL) , pp. 268–281, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.mrl-1.20. URL https://aclanthology.org/2023.mrl-1.20/.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. URL https://arxiv.org/abs/2306.16388.

Abteen Ebrahimi and Katharina Kann. How to adapt your pretrained multilingual model to 1600 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), 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) , pp. 4555–4567, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.351. URL https://aclanthology.org/2021.acl-long.351/.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. ArXiv , abs/2404.17790, 2024. URL https://api.semanticscholar.org/CorpusID:269449465.

Takuro Fujii, Koki Shibata, Atsuki Yamaguchi, Terufumi Morishita, and Yasuhiro Sogawa. How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study

in Japanese. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) , pp. 39–49, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.5. URL https://aclanthology.org/2023.acl-srw.5/.

Philip Gage. A new algorithm for data compression. The C Users Journal archive , 12:23–38, 1994. URL https://api.semanticscholar.org/CorpusID:59804030.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. Airavata: Introducing hindi instruction-tuned llm, 2024. URL https://arxiv.org/abs/2401.15006.

Matthias Gallé. Investigating the effectiveness of BPE: The power of shorter sequences. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , pp. 1375–1381, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1141. URL https://aclanthology.org/D19-1141/.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. Unpacking tokenization: Evaluating text compression and its correlation with model performance. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics: ACL 2024 , pp. 2274–2286, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.134. URL https://aclanthology.org/2024.findings-acl.134/.

Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. Improving tokenisation by alternative treatment of spaces. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pp. 11430–11443, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.786. URL https://aclanthology.org/2022.emnlp-main.786/.

Srishti Gureja, Lester James V Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-rewardbench: Evaluating reward models in multilingual settings. arXiv preprint arXiv:2410.15522 , 2024.

Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , pp. 4238–4248, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL https://aclanthology.org/D19-1433/.

William Held, Camille Harris, Michael Best, and Diyi Yang. A material lens on coloniality in nlp, 2023. URL https://arxiv.org/abs/2311.08391.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the

Association for Computational Linguistics (Volume 2: Short Papers) , pp. 385–393, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.43. URL https://aclanthology.org/2022.acl-short.43/.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics: ACL 2024 , pp. 4476–4494, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.265. URL https://aclanthology.org/2024.findings-acl.265/.

Sara Hooker. On the limitations of compute thresholds as a governance strategy. ArXiv , abs/2407.05694, 2024. URL https://api.semanticscholar.org/CorpusID:271051333.

Hongzhi Huang, Defa Zhu, Banggu Wu, Yutao Zeng, Ya Wang, Qiyang Min, and Xun Zhou. Over-tokenized transformer: Vocabulary is generally worth scaling, 2025. URL https://arxiv.org/abs/2501.16975.

Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. Towards better instruction following language models for chinese: Investigating the impact of training data and evaluation, 2023. URL https://arxiv.org/abs/2304.07854.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. D \ 'ej \ a vu: Multilingual llm evaluation through the lens of machine translation evaluation. arXiv preprint arXiv:2504.11829 , 2025.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 66–75, 2018.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pp. 5363–5394, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.360. URL https://aclanthology.org/2022.emnlp-main.360/.

Sander Land and Max Bartolo. Fishing for magikarp: Automatically detecting under-trained tokens in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , pp. 11631–11646, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.649. URL https://aclanthology.org/2024.emnlp-main.649/.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In Anna

Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Findings of the Association for Computational Linguistics: ACL 2023 , pp. 5661–5681, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.350. URL https://aclanthology.org/2023.findings-acl.350/.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. Myte: Morphology-driven byte encoding for better and fairer multilingual language modeling. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 15059–15076. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.804. URL http://dx.doi.org/10.18653/v1/2024.acl-long.804.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pp. 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.179. URL https://aclanthology.org/2024.naacl-long.179/.

YINGWEI MA, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help llms reasoning? In The Twelfth International Conference on Learning Representations .

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. Artificial intelligence index report 2025, 2025. URL https://arxiv.org/abs/2504.07139.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In EMNLP , 2018.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL https://aclanthology.org/2022.naacl-main.293/.

Benjamin Minixhofer, Edoardo M. Ponti, and Ivan Vulić. Zero-shot tokenizer transfer. In Proceedings of the 38th International Conference on Neural Information Processing Systems , NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy,

Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 448–462, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.38. URL https://aclanthology.org/2021.naacl-main.38/.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. Efficient continual pre-training of LLMs for low-resource languages. In Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track) , pp. 304–317, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-194-0. URL https://aclanthology.org/2025.naacl-industry.25/.

Gabriel Nicholas and Aliya Bhatia. Lost in translation: Large language models in non-english content analysis, 2023. URL https://arxiv.org/abs/2306.07377.

Ayomide Odumakinde, Daniel D'souza, Pat Verga, Beyza Ermis, and Sara Hooker. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress, 2024. URL https://arxiv.org/abs/2408.14960.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. Afrobench: How good are large language models on african languages?, 2025. URL https://arxiv.org/abs/2311.07978.

Malte Ostendorff and Georg Rehm. Efficient language model training through cross-lingual and progressive transfer learning, 2023. URL https://arxiv.org/abs/2301.09626.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Reuse, don't retrain: A recipe for continued pretraining of language models, 2024. URL https://arxiv.org/abs/2407.07263.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. Fineweb2: A sparkling update with 1000s of languages, December 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-2.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In Proceedings of the 37th International Conference on Neural Information Processing Systems , NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , pp. 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL https://aclanthology.org/2021.emnlp-main.800/.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André FT Martins. M-prometheus: A suite of open multilingual llm judges. arXiv preprint arXiv:2504.04953 , 2025.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Kartik Ravisankar, Hyojung Han, and Marine Carpuat. Can you map it to english? the role of cross-lingual alignment in multilingual performance of llms, 2025. URL https://arxiv.org/abs/2504.09378.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), Advances in Neural Information Processing Systems , volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. Tokenization is more than compression. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , pp. 678–702, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.40. URL https://aclanthology.org/2024.emnlp-main.40/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith (eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Shaurya Sharthak, Vinayak Pahalwan, Adithya Kamath, and Adarsh Shirawalmath. Achieving tokenizer flexibility in language models through heuristic adaptation and supertoken learning, 2025. URL https://arxiv.org/abs/2505.09738.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion, 2025. URL https://arxiv.org/abs/2504.20879.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomput. , 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Jimin Sun, Patrick Fernandes, Xinyi Wang, and Graham Neubig. A multi-dimensional evaluation of tokenizer-free multilingual pretrained models. In Andreas Vlachos and Isabelle Augenstein (eds.), Findings of the Association for Computational Linguistics: EACL 2023 , pp. 1725–1735, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.128. URL https://aclanthology.org/2023.findings-eacl.128/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) , pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book, 2024. URL https://arxiv.org/abs/2309.16575.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems , volume 37, pp. 114147–114179. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cf5a019ae9c11b4be88213ce3f85d85c-Paper-Conference.pdf.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff

Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson,

Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine

Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba

Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude,

Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL https://aclanthology.org/2024.acl-long.845/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems , volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Menan Velayuthan and Kengatharaiyer Sarveswaran. Egalitarian language representation in language models: It all begins with tokenizers, 2024. URL https://arxiv.org/abs/2409.11501.

Menan Velayuthan and Kengatharaiyer Sarveswaran. Egalitarian language representation in language models: It all begins with tokenizers. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), Proceedings of the 31st International Conference on Computational Linguistics , pp. 5987–5996, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.400/.

Shumin Wang, Yuexiang Xie, Bolin Ding, Jinyang Gao, and Yanyong Zhang. Language adaptation of large language models: An empirical study on LLaMA2. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), Proceedings of the 31st International Conference on Computational Linguistics , pp. 7195–7208, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.480/.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective works best for zero-shot generalization? In International Conference on Machine Learning , pp. 22964–22984. PMLR, 2022.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. Extending multilingual BERT to low-resource languages. In Trevor Cohn, Yulan He, and Yang Liu (eds.), Findings of the Association for Computational Linguistics: EMNLP 2020 , pp. 2649–2656, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.240. URL https://aclanthology.org/2020.findings-emnlp.240/.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte

models. Transactions of the Association for Computational Linguistics , 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Findings of the Association for Computational Linguistics: EMNLP 2024 , pp. 6760–6785, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.396. URL https://aclanthology.org/2024.findings-emnlp.396/.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. How can we effectively expand the vocabulary of llms with 0.01gb of target language text?, 2024b. URL https://arxiv.org/abs/2406.11477.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 11682–11703, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.653. URL https://aclanthology.org/2023.acl-long.653/.

Joanna Yoo, Kuba Perlin, Siddhartha Rao Kamalakara, and João GM Araújo. Scalable training of language models using jax pjit and tpuv4. arXiv preprint arXiv:2204.06514 , 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

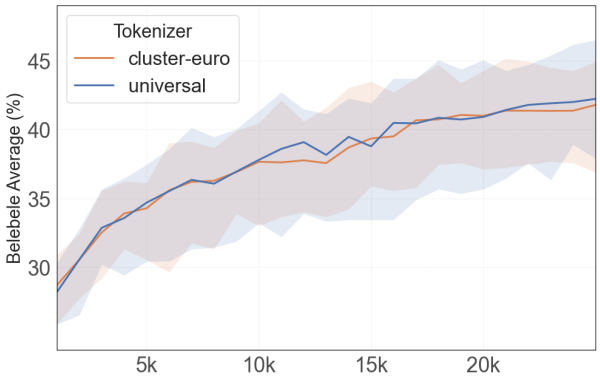


Figure 6: 在预训练期间的主要语言上的平均表现（Euro），在 Belebele 中进行测量。阴影区域表示每种分词器在各语言中的结果。UNIVERSAL 分词器在整个训练期间显示出几乎与 CLUSTER 分词器相同的性能，这也表明在更长的预训练运行中会有类似的性能。

附加分词器细节

我们在所有训练的分词器中使用了 GPT-4o（`gpt4-o200k`）正则表达式进行预标记。[3] 分词器的训练数据是根据第 2.3 节中描述的权重，从预训练数据集合中抽样的。我们使用 [4] 库来训练所有的 BPE 模型。我们将 BPE 训练器的 `min_frequency` 参数设置为 5，以控制合并对的最低频率，并且不使用普通化器。最后，我们抽样了 50GB 的数据来训练所有的分词器。

附加训练细节

超参数我们在学习率 (LR) 上进行了超参数搜索，使用 2×10^{-2} 作为所有预训练实验的峰值 LR。我们使用 512 的批量大小，8192 的序列长度，以及一个预热 2500 步的余弦学习率调度器。对于预训练后的语言适应实验，我们使用常量 LR 1×10^{-4} ，与预训练阶段的结束 LR 对应。

附加结果

额外的预训练结果

图 6 显示了整个预训练过程中，在主语言子集（欧洲集群）上用 Belebele 度量的预训练结果。

扩展语言适应结果

继续预训练

表 5 显示了持续预训练后按语言划分的胜率，分为 PRIMARY 和 EXPANDED 语言

目标适应

表 5 显示了在持续预训练后按语言划分的胜率，分为 PRIMARY 和 EXPANDED 语言

[3] https://github.com/openai/tiktoken/blob/4560a8896f5fb1d35c6f8fd6eee0399f9a1a27ca/tiktoken_ext/openai_public.py#L95

[4] https://github.com/huggingface/tokenizers

		Asian			ME-Indic	
		Cluster	Universal		Cluster	Universal
Primary	ind	56.5	46.5	arb	37.5	43.0
	jpn	52.0	40.5	heb	38.0	39.0
	kor	46.0	40.0	hin	41.5	38.0
	vie	49.5	39.5	pes	46.0	42.5
	zho	52.5	45.0	tur	35.5	38.5
Expanded	arb	16.0	27.0	deu	18.0	35.5
	deu	25.5	27.5	ell	10.6	29.5
	ell	9.0	29.0	fra	28.0	43.0
	fra	31.5	35.5	ind	24.5	46.5
	heb	6.1	24.5	jpn	20.5	46.0
	hin	5.1	27.5	kor	19.0	37.5
	pes	13.5	33.5	rus	20.5	41.0
	rus	17.5	38.0	spa	35.5	46.5
	spa	38.0	43.0	vie	16.0	38.0
	tur	13.0	27.5	zho	35.5	54.0

Table 5: 通过持续的预训练进行语言适应的各语言完整胜率结果。

	Asian			ME-Indic	
	Cluster	Universal		Cluster	Universal
arb	16.5	30.5	deu	20.5	35.5
deu	20.0	30.5	ell	19.1	32.5
ell	13.1	32.0	fra	27.5	39.0
fra	27.5	37.0	ind	20.0	46.5
heb	12.1	32.0	jpn	16.5	41.5
hin	9.1	25.8	kor	16.5	37.0
pes	17.0	41.5	rus	22.5	39.0
rus	20.0	37.5	spa	39.5	46.5
spa	33.0	43.5	vie	16.5	43.0
tur	15.5	30.0	zho	34.5	50.5

Table 6: 目标语言适应胜率

判断胜率提示

<system_prompt>
You are a helpful following assistant whose goal is to select the preferred (least wrong) output for a given instruction.

<user_prompt>
Which of the following answers is the best one for the given instruction? A good answer should follow these rules:
1) It should have correct reasoning,"
2) It should answer the request in the instruction,
3) It should be factually correct and semantically comprehensible,
4) It should be grammatically correct and fluent.

Instruction: instruction
Answer (A): completion_a
Answer (B): completion_b

FIRST provide a concise comparison of the two answers. If one answer is better, explain which you prefer and why. If both answers are identical or equally good or bad, explain why. SECOND, on a new line, state exactly one of 'Answer (A)' or 'Answer (B)' or 'TIE' to indicate your choice of preferred response.
Your response should use the format: Comparison: <concise comparison and explanation> Preferred: <'Answer (A)' or 'Answer (B)' or 'TIE'>.

为了内在地评估分词器的质量，我们将压缩率与 Command-A [Cohere et al., 2025] 中使用的公开可用的多语言分词器进行比较。压缩衡量数据在大小（以字节为单位）上的表示效率，而 BPE 则优化这一条件 [Gage, 1994] 。压缩率比较分词器之间的压缩值，由于较低的压缩更为理想，压缩率低于 1 表示一个分词器具有更有利的压缩。之前的工作显示压缩与模型性能，尤其是生成任务的性能相关性较好 [Goldman et al., 2024; Gallé, 2019] ，尽管较低的压缩不是更好分词器的充分条件 [Schmidt et al., 2024] 。然而，长序列长度是开始对语言不公平处理的一种方式，从分词器开始

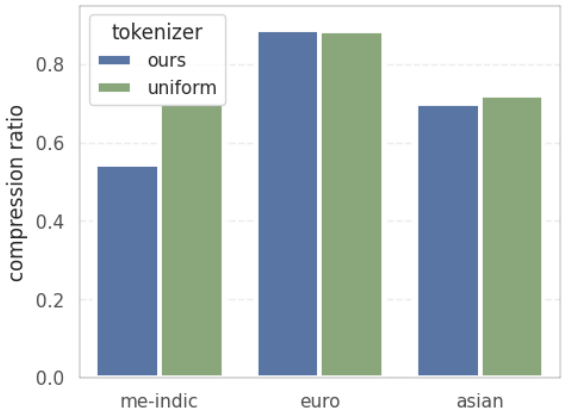


Figure 7: 我们分词器和基线统一分词器的压缩率。我们的分词器使用一种特殊的加权方法，利用训练数据分布和语言分组（§ 2.3），从而实现更好的压缩（越低越好）。

[Velayuthan & Sarveswaran, 2024; Ahia et al., 2023]，因此这是一个需要与下游评估一起考虑的重要指标。

分词器语言加权对压缩比的影响

作为基线，我们评估了统一语言加权，并将其与我们的分词器进行比较，其中我们结合使用了数据分布和语言分桶策略。压缩率是针对 FineWeb-2 [Penedo et al., 2024] 测试分割中开放权重 Command-A [Cohere et al., 2025] 模型的多语言分词器计算的。

图 7 显示了比较结果。可以看到，我们使用特殊权重的分词器比统一基线实现了更好的压缩。请注意，这两个分词器都比 Command-A 实现了更高的总体压缩性能，因为它们在更大语言覆盖范围上进行了训练。

压缩比对下游性能的影响

图表 8 探索了使用 UNIVERSAL 和 CLUSTER 分词器对主要和扩展语言进行训练的欧洲集群模型的压缩比与胜率之间的关系。使用 CLUSTER 的扩展语言表现出较大的压缩比，均超过 1，这表明在该语言中的压缩比差于比较的分词器。同时，这些语言的胜率低于压缩比也较低的主要语言。然而，UNIVERSAL 分词器在主要和扩展语言中表现出相对较高的胜率和较低的压缩比。这个结果佐证了压缩比与下游性能之间的关系，并为 UNIVERSAL 分词器的语言适应性提供了额外的维度。

语言

Table 7: 预训练语言，包括预训练聚类分配。在后训练列中有复选标记但没有聚类分配的语言被用作未见的适应语言。

ISO Code	Language	Script	Family	Subgrouping	Resources	Cluster	In Post-Training
afr	Afrikaans	Latin	Indo-European	Germanic	Mid	-	✓
ara	Arabic	Arabic	Afro-Asiatic	Semitic	High	Me-Indic	✓
amh	Amharic	Ge'ez	Afro-Asiatic	Semitic	Low	-	✗
bel	Belarusian	Cyrillic	Indo-European	Balto-Slavic	Mid	-	✓
ben	Bengali	Bengali	Indo-European	Indo-Aryan	Mid	Me-Indic	✗
bul	Bulgarian	Cyrillic	Indo-European	Balto-Slavic	Mid	Euro	✗
cat	Catalan	Latin	Indo-European	Italic	High	Euro	✗
ces	Czech	Latin	Indo-European	Balto-Slavic	High	Euro	✓
cym	Welsh	Latin	Indo-European	Celtic	Low	Euro	✗
dan	Danish	Latin	Indo-European	Germanic	Mid	Euro	✗
deu	German	Latin	Indo-European	Germanic	High	Euro	✓

ISO Code	Language	Script	Family	Subgrouping	Resources	Cluster	In Post-Training
ell	Greek	Greek	Indo-European	Graeco-Phrygian	Mid	Euro	✓
eng	English	Latin	Indo-European	Germanic	High	Euro	✓
est	Estonian	Latin	Uralic	Finnic	Mid	Euro	✗
eus	Basque	Latin	Basque	-	High	Euro	✗
fil	Filipino	Latin	Austronesian	Malayo-Polynesian	Mid	Asian	✗
fin	Finnish	Latin	Uralic	Finnic	Mid	Euro	✗
fra	French	Latin	Indo-European	Italic	High	Euro	✓
gla	Scottish Gaelic	Latin	Indo-European	Celtic	Low	Euro	✗
gle	Irish	Latin	Indo-European	Celtic	Low	Euro	✗
glg	Galician	Latin	Indo-European	Italic	Mid	Euro	✗
guj	Gujarati	Gujarati	Indo-European	Indo-Aryan	Low	Me-Indic	✗
heb	Hebrew	Hebrew	Afro-Asiatic	Semitic	Mid	Me-Indic	✓
hin	Hindi	Devanagari	Indo-European	Indo-Aryan	High	Me-Indic	✓
hrv	Croatian	Latin	Indo-European	Balto-Slavic	High	Euro	✗
hun	Hungarian	Latin	Uralic	-	High	Euro	✗
hye	Armenian	Armenian	Indo-European	Armenic	Low	-	✓
ibo	Igbo	Latin	Atlantic-Congo	Benue-Congo	Low	-	✗
ind	Indonesian	Latin	Austronesian	Malayo-Polynesian	Mid	Asian	✓
ita	Italian	Latin	Indo-European	Italic	High	Euro	✓
jav	Javanese	Latin	Austronesian	Malayo-Polynesian	Low	Asian	✗
jpn	Japanese	Japanese	Japonic	Japanesic	High	Asian	✓
kaz	Kazakh	Cyrillic	Turkic	Common Turkic	Mid	-	✓
khm	Khmer	Khmer	Austroasiatic	Khmeric	Low	Asian	✗
kor	Korean	Hangul	Koreanic	Korean	Mid	Asian	✓
lao	Lao	Lao	Tai-Kadai	Kam-Tai	Low	Asian	✗
lav	Latvian	Latin	Indo-European	Balto-Slavic	Mid	Euro	✗
lit	Lithuanian	Latin	Indo-European	Balto-Slavic	Mid	Euro	✗
mlt	Maltese	Latin	Afro-Asiatic	Semitic	Low	Me-Indic	✗
msa	Malay	Latin	Austronesian	Malayo-Polynesian	Mid	Asian	✗
mya	Burmese	Myanmar	Sino-Tibetan	Burmo-Qiangic	Low	Asian	✗
nep	Nepali	Devanagari	Indo-European	Indo-Aryan	Low	-	✓
nld	Dutch	Latin	Indo-European	Germanic	High	Euro	✓
nor	Norwegian	Latin	Indo-European	Germanic	Low	Euro	✗
pan	Punjabi	Gurmukhi	Indo-European	Indo-Aryan	Low	Me-Indic	✗
pes	Persian	Arabic	Indo-European	Iranian	High	Me-Indic	✓
pol	Polish	Latin	Indo-European	Balto-Slavic	High	Euro	✓
por	Portuguese	Latin	Indo-European	Italic	High	Euro	✓
ron	Romanian	Latin	Indo-European	Italic	Mid	Euro	✓
rus	Russian	Cyrillic	Indo-European	Balto-Slavic	High	Euro	✓
sin	Sinhala	Sinhala	Indo-European	Indo-Aryan	Low	-	✓
slk	Slovak	Latin	Indo-European	Balto-Slavic	Mid	Euro	✗
slv	Slovenian	Latin	Indo-European	Balto-Slavic	Mid	Euro	✗
spa	Spanish	Latin	Indo-European	Italic	High	Euro	✓
srp	Serbian	Cyrillic	Indo-European	Balto-Slavic	High	Euro	✗
swa	Swahili	Latin	Atlantic-Congo	Benue-Congo	Low	-	✗
swe	Swedish	Latin	Indo-European	Germanic	High	Euro	✗
tam	Tamil	Tamil	Dravidian	South Dravidian	Mid	Me-Indic	✗
tel	Telugu	Telugu	Dravidian	South Dravidian	Low	Me-Indic	✗
tha	Thai	Thai	Tai-Kadai	Kam-Tai	Mid	Asian	✗
tur	Turkish	Latin	Turkic	Common Turkic	High	Me-Indic	✓
ukr	Ukrainian	Cyrillic	Indo-European	Balto-Slavic	Mid	Euro	✓
urd	Urdu	Arabic	Indo-European	Indo-Aryan	Mid	Me-Indic	✗
vie	Vietnamese	Latin	Austroasiatic	Vietic	High	Asian	✓
xho	Xhosa	Latin	Atlantic-Congo	Benue-Congo	Low	-	✗
yor	Yorùbá	Latin	Atlantic-Congo	Benue-Congo	Low	-	✗
yue	Cantonese	Han	Sino-Tibetan	Sinitic	Low	-	✓
zho	Mandarin Chinese	Han	Sino-Tibetan	Sinitic	High	Asian	✓
zul	Zulu	Latin	Atlantic-Congo	Benue-Congo	Low	-	✗

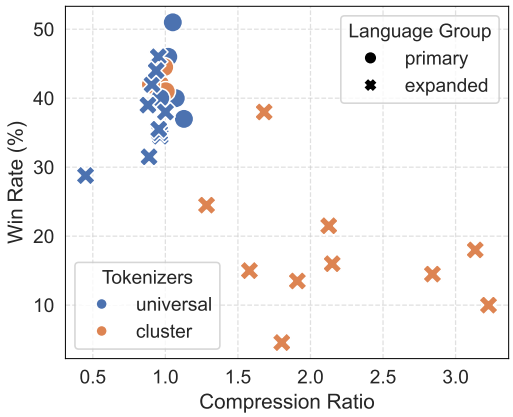


Figure 8: 在欧洲集群中，每种语言的适应结果以及分词器的压缩比。虽然 UNIVERSAL 分词器能够实现更好的压缩，尤其是对于扩展的语言子集，从而实现更好的下游性能，但 CLUSTER 分词器未能代表这些语言，导致较低的适应结果。