优化意大利语的 LLM：通过词汇适应减少符号繁殖并提升效率

Luca Moroni[1]*, **Giovanni Puccetti**[2]*, **Pere-Lluis Huguet Cabot**[1], **Andrei Stefan Bejgu**[4]
Edoardo Barba[1], **Alessio Miaschi**[3]
Felice Dell'Orletta[3], **Andrea Esuli**[2], **Roberto Navigli**[1]
[1]Sapienza University of Rome { surname } @diag.uniroma1.it
[2]ISTI-CNR { name.surname } @isti.cnr.it
[3]ILC-CNR { name.surname } @ilc.cnr.it
[4]Babelscape { surname } @babelscape.com

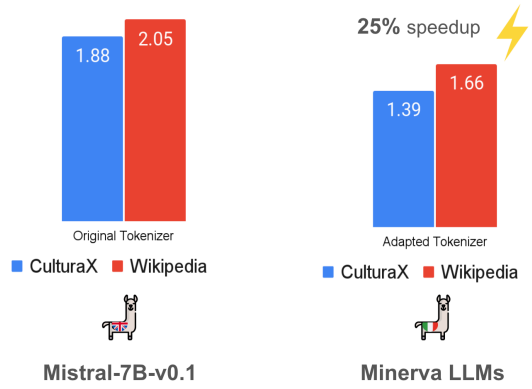

Figure 1: 两个不同的分词器 Mistral-7B-v0.1（左）和 Minerva（右）在 CulturaX（蓝色）和维基百科（红色）意大利语文本中的繁殖率。

Abstract

预训练大型语言模型（LLMs）的数量正在稳定增加，但绝大多数主要为英语语言设计。尽管最先进的 LLMs 可以处理其他语言，由于语言污染或一定程度的多语言预训练数据，它们并未针对非英语语言进行优化，导致编码效率低下（高 token "繁殖性"）和推理速度较慢。在此工作中，我们彻底比较了多种词汇适配技术，以优化英语 LLMs 为意大利语服务，并提出了语义对齐词汇适配（SAVA），一种利用神经映射进行词汇替换的新方法。SAVA 在多个下游任务中表现出竞争力，增强了基于上下文对齐的策略。我们适配了两个 LLMs：*Mistral-7B-v0.1*，减少 token 繁殖性 25 %，以及 *Llama-3.1-8B*，优化词汇并减少了一亿个参数。我们展示了随着词汇的适配，这些模型可以通过在目标语言上进行相对有限阶段的持续训练来恢复其性能。最后，我们在各种多选和生成任务上测试适配模型的能力。[1]

1 介绍

大型语言模型（LLMs）已经获得了巨大的普及，并在广泛的应用中越来越多地被使用 (Radford et al., 2019; Kojima et al., 2022)。尽管它们表现出色，但这些模型主要是以英语为中心的，也就是说，大多数最先进的模型都是在以英语为主要关注点的数据集上设计和预训练的 (Jiang et al., 2023; Dubey et al., 2024; Mesnard et al., 2024)。虽然本地多语言模型——即在多个目标语言中完全预训练——多年来已经发布 (Le Scao et al., 2023)，但它们仍然无法达到与英文预训练模型相媲美的表现水平。主要挑战在于解决那些语言代表性不足的问题，在这些语言中，庞大、干净、开放访问的文本语料库往往稀缺 (Weber et al., 2024; Nguyen et al., 2024)。这种稀缺性是一个问题，因为模型需要大量高质量数据才能达到满意的性能 (Hoffmann et al., 2022)。此外，由于众所周知的多语言性问题，多语言模型通常性能不佳 (Conneau et al., 2020)。

解决这些挑战的一个有希望的方法是将预训练的英语大型语言模型（LLM）调整为支持其他语言 (Chau et al., 2020)。最近的研究指出，将以英语为中心的模型微调以支持其他语言能够带来显著的收益，使得在节省计算资源和训练时间的同时实现高效适应。这种方法减少了训练预算和所需的符号数量，即使在资源有限的情况下也表现出竞争力 (Koto et al., 2021; Minixhofer et al., 2022; Gee et al., 2022; Ostendorff and Rehm, 2023)。

另一个重要方面是语言模型的下游性能，以及分词器在目标语言中的繁殖力。大型语言模型依赖于分词器，它是在一个混合文本（无论是否是大型语言模型的训练数据）上训练的，将原始文本转化为词片段；繁殖力是一个单词被拆分成的平均词块数量 (Brown et al., 1993)。分词器的繁殖力对其训练的语言和文本类型，以及测量繁殖力的文本非常敏感。Figure 1 展示了这个现象的一个例子，比较了 *Minerva-LLMs* 的繁殖力，即一个意大利语优先的大型语言模型家族 (Orlando et al., 2024) 和 *Mistral-7B-v0.1*，一个英语优先的大型语言模型 (Jiang et al., 2023)，在两个意大利语语料库上的表现。

在这项工作中，我们探讨了将两种最先进

* Those authors contributed equally.

[1]我们在 https://github.com/SapienzaNLP/sava 上发布我们的代码和模型

的英文 LLM 适应到意大利语的过程，使用了词汇适应和持续学习。此外，我们引入了一种新的词汇适应技术，称为语义对齐词汇适应（SAVA），并与最近的方法 (Gee et al., 2022; Ostendorff and Rehm, 2023) 进行了全面比较，考察了词汇替换对模型性能在适应过程中的影响。经过词汇适应后，在对意大利语文本进行分词时，我们能将 *Mistral-7B-v0.1* 的冗余减少 25%，将 *Llama-3.1-8B* 减少 16%。关于 *Mistral-7B-v0.1* ，我们没有增加其词汇大小或模型参数，而对于 *Llama-3.1-8B* ，我们有效地将其词汇大小减少 75%，从而缩小最终模型大小 10%。总体而言，我们减少了模型的内存和计算占用。总结来说，主要的贡献是：

- 介绍了一种有效的方法，用于调整生成模型的分词器和词汇，从而在多个下游基准测试中实现优于现有方法的竞争性表现；
- 提供各种分词器适应技术的详细比较分析，重点是低到中等资源场景中的连续训练。
- 分析通过不同的自适应技术学习的嵌入表示，提供对词汇修改如何影响模型性能和泛化能力的更深入理解。

2 相关工作

语言自适应预训练 在目标语言中设计 LLMs 并从头开始训练是从一开始就获得适当的标记词产出并最小化不同语言预训练数据干扰的最佳方法。然而，这种方法通常不可行，尤其是在资源有限和计算预算低的情况下。出于这个原因，最近的几项研究 (de Vries and Nissim, 2021; Gee et al., 2022; Csaki et al., 2024) 侧重于将预训练的 LLMs 适应到新语言。与预训练阶段所需的数据量相比，预训练的 LLMs 可以使用少量数据适应特定语言。一种实现此目标的直接方法是语言自适应预训练（LAPT），Chau et al. (2020) 在多语言环境中利用这一方法，他们在目标语言上测试了多语言 LLMs 的连续训练。有趣的是，LAPT 之前在仅有编码器的架构上由 Gururangan et al. (2020) 提出，他们在生物医学领域成功地适应了 RoBERTa (Zhuang et al., 2021) 模型。在 LAPT 中，模型的架构不 undergo 任何结构变化。这通常会带来性能提升，然而，它并未解决使用不太适合不同语言编码的次优标记器的局限性。关于英语到意大利语模型的 LAPT 研究，曾有几次尝试，最值得注意的是 *LLaMAntino-2-LLMs* ，这是一种在意大利语翻译对话 (Basile et al., 2023) 上对 LLama 2 的微调，以及使用类似方法构建在 Llama-3-8B 之上的最近努力 *LLaMAntino-3-ANITA-8B-Inst-DPO-ITA* (Polignano et al., 2024) 。

为了应对生育问题，近期的研究集中于通过修改预训练大型语言模型的分词器和词汇来改善语言适应性，以更好地适应目标语言。该领域的若干努力表明了词汇适应技术的有效性。Minixhofer et al. (2022) 和 Liu et al. (2024) 提议替换预训练大型语言模型的分词器及其相应的嵌入层，依赖于双语词典基础或基于图的标记映射。通常，各种词汇适应技术的主要区别在于适应过程中各模型嵌入空间的初始化方式。Ostendorff and Rehm (2023); Dobler and de Melo (2023) 进行了更多的努力，他们使用与所需分词器一起训练的辅助模型的嵌入。他们利用辅助模型中嵌入结构的几何相似性，有效地初始化目标模型的标记表示。同时，Gee et al. (2022) 提出了一种简单的启发式方法，将目标词汇标记初始化为源词汇中对应子标记的平均值。另一个研究由 Koto et al. (2021) 提出了一种适应技术，他们依赖 `FastText` [2] 嵌入空间学习线性映射，以执行基于 BERT 模型的词汇适应。

与以往的研究不同，我们对现有的适应启发式方法进行了彻底的分析，重点研究仅用于解码器的生成模型如何适应意大利语。我们提出了一种新的启发式方法，该方法利用一个专为目标语言优化的辅助嵌入空间来映射和初始化目标词汇的标记。

在本节中，我们将对用于将预训练的 LLM 适配到目标语言的方法进行形式化。接下来的子节概述了用于修改预训练 LLM 词汇的方法，并描述了将其适配到目标语言的过程。最后，我们描述适配的最后一步，也就是持续训练步骤。

2.1 词汇适应

所有的词汇适应方法都具有相似的目标：替换分词器及其词汇表，并用更适合目标语言的词来更换模型嵌入（包括嵌入模块和语言模型头）。

在我们的设置中，我们有一个预训练的源 LLM，M_s ，其嵌入矩阵为 E_s [3] ，分词器为 T_s ，词汇表为 V_s 。为了将我们的模型适应目标语言，我们有一个适合编码目标语言文本的目标分词器 T_t 和词汇表 V_t ，我们希望使其与 M_s 兼容。在某些情况下，我们还可以获得一个辅助模型 M_h ，它是一个通常比 M_s 小的 LLM，其嵌入记为 E_h 。辅助模型是使用 T_t 和 V_t 训

[2] https://fasttext.cc/

[3] 这里，我们假设使用相同的权重，即共享的嵌入模块和语言模型头。当情况不是这样时，该方法是对称的，就好像有两个嵌入矩阵一样。

练的。我们使用上标表示法 E^{t_i} 来表示在矩阵嵌入 E 上的标记 t_i 的表示。

目标是调整源模型嵌入 E_s，这些方法都在它上面运行，从而基于目标分词器 T_t 和目标词汇表 V_t 获得 E_t。

首先，目标嵌入通过在两个词汇表相交的标记中保持与 E_s 相同的表示来初始化，同时对其他剩余的标记应用函数 g：

$$E_t^{t_i} = \begin{cases} g(t_i, \cdot), & t_i \in V_t \setminus V_s \\ E_s^{t_i}, & t_i \in V_s \cap V_t \end{cases}$$

这些方法之间的区别在于使用了 g 来初始化在 V_t 中但不在 V_s 中的标记。该函数可以访问源嵌入 E_s、词汇表 V_s 和分词器 T_s，并有可能访问辅助模型的嵌入、词汇表和分词器，分别为 E_h、V_h 和 T_h。

因此，每种方法都是由其各自的 g 函数定义的，如下所述。

作为基线方法，我们用由正态分布给出的随机表示来初始化交集中外的标记，该正态分布的均值和方差由源嵌入空间定义：

$$g_{random}(t_i, E_s) = \mathcal{N}(\mu(E_s), \sigma^2(E_s))$$

FVT Gee et al. (2022) 引入了用于词汇适应的快速词汇转移（FVT），它由一种有效的方法构成，用于在目标嵌入空间中初始化交集标记。这里，每个目标标记是通过求给定源标记器的嵌入源标记的平均值计算的，即，当我们用 T_s 对目标标记 t_i 进行标记化时得到的结果标记：

$$g_{fvt}(t_i, E_s, T_s) = \frac{1}{|T_s(t_i)|} \cdot \sum_{t_j \in T_s(t_i)} E_s^{t_j}.$$

约束线性规划 Ostendorff and Rehm (2023) 和 Dobler and de Melo (2023) 同时引入了一种启发式方法来初始化库存外代币，该方法依赖于辅助嵌入空间的空间结构。两种方法都在辅助模型 E_h 的嵌入空间上计算 $V_t \setminus V_s$ 中的代币与 $V_t \cap V_s$ 中代币之间的相似性分数。这些相似性用于在目标嵌入矩阵 E_t 中构建库存外代币的表示，这依赖于源嵌入 E_s 表示：

$$g_{clp}(t_i, E_s, V_s, E_h, V_h) = \sum_{t_j \in V_t \cap V_s} E_s^{t_j} \cdot \alpha(E_h^{t_i}, E_h^{t_j})$$

其中，$\alpha(\cdot, \cdot)$ 表示 E_h 中两个符号之间的相似度分数。在这里，我们依赖于 Ostendorff and Rehm (2023) 使用的相似度函数，该函数计算为规范化余弦相似度。

使用线性模型在两个不同模型的嵌入空间之间映射嵌入表示具有理论证明。Moschella et al. (2023) 和 Maiorca et al. (2024) 表明，不同模型的嵌入通过保形变换相关，或更一般地，通过这些空间之间的线性映射相关。受到 Maiorca et al. (2024) 的研究成果和 Koto et al. (2021) 的有趣努力的启发，我们提出了一种为生成模型执行词汇适配的技术，称为语义对齐词汇适配（SAVA）。在我们的方法中，我们依赖于一个来自大型语言模型的辅助模型嵌入 E_h，并学习 $E_h \subseteq \mathbb{R}^m$ 和 $E_s \subseteq \mathbb{R}^n$ 之间的线性映射 ϕ。我们训练了一个单层前馈网络（FFN）以将辅助嵌入空间映射到源嵌入空间上：

$$\phi : x \mapsto y \mid x \in \mathbb{R}^m, y \in \mathbb{R}^n,$$
$$g_{sava}(t_i, E_h) = \phi(E_h^{t_i}) \quad (1)$$

训练 ϕ 的目标是获得助手模型的 token 表示与来源模型的 token 表示之间的映射。为了训练它，我们使用交集 $V_s \cap V_t$ 中的 token，因为它们根据来源和助手模型都有表示，我们可以在 E_s 中的表示和 E_h 中的表示之间训练线性映射。然后，如方程 1 所述，我们使用 ϕ 将不在来源词汇表中的 token（$V_t \setminus V_s$）映射到来源嵌入空间。因此，我们的目标是找到：使

$$\phi(x) = Wx + b,$$

满足，其中 $W \in \mathbb{R}^{n \times m}$ 和 $b \in \mathbb{R}^n$ 是我们线性映射的参数。有关线性映射训练的更多技术细节将在附录 A 中提供。

2.2 持续训练

虽然通过词汇适应技术重新初始化嵌入能够实现零样本语言建模，但生成的语言模型通常缺乏对新语言的熟练掌握。我们通过对源语言和目标语言的混合进行持续训练来解决这个问题，这使得模型在提高目标语言的同时，能够保持在源语言的性能。

为了实现稳健的比较，我们使用上述所有词汇适配启发式方法将预训练的 LLM 适配到目标语言。我们还展示了通过连续训练基础模型在目标语言上（LAPT）所得的结果。虽然这种方法不太具有破坏性，但它不会改变词汇表或分词器，从而保持其完整性。

3 实验设置

本节描述了我们实验的设置，我们在其中调整了两个流行的 LLM，具体来说是 *Mistral-7B-v0.1* (Jiang et al., 2023) 和 *Llama-3.1-8B* (Dubey et al., 2024)。接下来的几个小节中，我们报告了用于词汇调整、持续训练和评估的设置。

Model	Num. Tokens	Num. Parameters
Mistral-7B-v0.1	32000	7.24B
Mistral-7B-v0.1 a.w. Minerva	32768	7.25B
LLaMa-3-8B	128256	8.03B
LLaMa-3-8B a.w. Minerva	32768	7.25B

Table 1: 模型参数计数和词汇表大小在适应和不适应时的比较（a.w. 代表适应于）。

	Fertility ↓			
Model	CX IT	CX EN	Wp IT	Wp EN
Mistral-7B-v0.1	1.88	1.32	2.05	1.57
Minerva	1.39	1.32	1.66	1.59
LLaMa-3-8B	1.67	1.15	1.80	1.31

Table 2: 不同分词器在 CulturaX（CX）和维基百科（Wp）上的词汇生成能力。

3.1 词汇调整

为将英语模型适应意大利语，我们依赖于 *Minerva-LLMs* 模型系列及其分词器 (Orlando et al., 2024) 。*Minerva-LLMs* 系列的模型是从头在一个意大利语-英语数据集上训练的，即 CulturaX (Nguyen et al., 2024) 。在撰写本文时，已经发布了三个不同的模型，*Minerva-350M* ，*Minerva-1B* ，以及 *Minerva-3B* ，使用相同的分词器。

Minerva-LLMs 分词器与 *Mistral-7B-v0.1* 共享 16,438 个标记，与 *Llama-3.1-8B* 共享 20,358 个标记。对于 CLP 和 SAVA，我们使用 *Minerva-3B* 作为辅助模型。值得注意的是，如 Table 1 所示，使用 *Minerva-LLMs* 分词器来调整像 *Llama-3.1-8B* 这样的大模型显著减少了词汇量（减少了 75 %），因此参数也减少。调整后的 *Llama-3.1-8B* 参数为 72.5 亿个，而原始模型的参数为 80 亿个，结果使模型大小减少了 10 %。

作为进一步的改进，用 *Minerva-LLMs* 替代 *Mistral-7B-v0.1* 和 *Llama-3.1-8B* 标记器对意大利语的生成效果有显著影响。如 Table 2 所示，*Minerva-LLMs* 标记器在两个意大利语文本来源——CulturaX（CX）和 Wikipedia（Wp）上，与 *Mistral-7B-v0.1* 标记器相比，平均提高了 25 个% 的生成效果。在相同设定下，*Llama-3.1-8B* 依靠 *Minerva-LLMs* 标记器在意大利语文本上最多提高了 16 个% 的生成效果。

3.2 持续训练

为了进行持续训练，我们使用 CulturaX，一个大型多语言数据集，该数据集已成功用于对欧盟语言（包括意大利语）的大型持续训练实验中。[4] 我们旨在在固定的计算预算（即标记数量）下比较所有方法。由于计算预算有限，我们决定在训练标记达到 12B 的阈值后停止训练。

我们从 CulturaX 的意大利语和英语部分中抽样训练数据，以创建一个由 75 % 意大利语标记和 25 % 英语标记组成的数据集，正如 Csaki et al. (2024) 所提议的。

我们使用打包方法将所有的标记填充到固定长度的序列中。学习率在所有运行中固定为 10^{-5} 。

对于 *Mistral-7B-v0.1* ，训练在莱昂纳多超级计算机上的 16 个节点上进行（每个节点使用 4 个 64 GB A100），保持全球批量大小为 3072，序列长度为 2048。对于 *Llama-3.1-8B* ，我们将训练数据的序列长度更改为 8192，并将全球批量大小设置为 512。在训练这两个模型时，我们不冻结任何参数，让它们全部更新。我们进行持续训练，使模型能够处理大约 120 亿个标记。具体来说，我们训练 *Mistral-7B-v0.1* 2000 个批次和 *Llama-3.1-8B* 3000 个批次。我们使用 llm-foundry 来训练 [5] ，对于剩余的超参数，我们使用库提供的默认设置。参见 Appendix B 以估算本工作中进行的实验的 CO_2 成本。

3.3 评估

为了评估我们的模型，我们依赖于使用复杂度评估方法的 LM-Evaluation-Harness 库 (Gao et al., 2024) ，用于多项选择（MC）基准测试。作为多项选择基准测试，我们使用 ITA-Bench (Moroni et al., 2024) 的翻译部分，这是一个从英语自动翻译成意大利语的基准测试套件。

在持续训练期间，我们在 0-shot 场景中每 200 批次评估我们的 *Mistral-7B-v0.1* 模型，每 300 批次评估我们的 *Llama-3.1-8B* 模型；通过这种方式，每个后续的检查点都会在相同数量的 tokens 上进行一致评估。为了评估适应后的模型的推理能力，我们使用了多种基准：MMLU (Hendrycks et al., 2021) 、BOOLQ (Clark et al., 2019) 、ARC-easy (Clark et al., 2018) 、PIQA (Bisk et al., 2020) 、SciQ (Welbl et al., 2017) 和 Hellaswag (Zellers et al., 2019) 。

我们还测量了模型在生成任务上的性能，重点关注两个任务：自动翻译，FLoRes 基准 (Costa-jussà et al., 2022) ，以及问答，SQuAD-it (Croce et al., 2018) ，这是一个自动翻译成意大利语的 SQuAD (Rajpurkar et al., 2016) 版本。我们使用了 `vLLM` (Kwon et al., 2023) 作为我们的生成流水线。与生成技术相关的更多详细信息可以在附录 **??** 中找到。

[4] https://huggingface.co/occiglot/occiglot-7b-it-en-instruct

[5] https://github.com/mosaicml/llm-foundry

Model	Hellaswag	MMLU	Arc Easy	PIQA	SCIQ	BOOLQ	AVG
Mistral-7B-v0.1	$56.50_{\pm 0.49}$	$47.42_{\pm 0.42}$	$61.67_{\pm 1.01}$	$67.24_{\pm 1.14}$	$84.75_{\pm 1.16}$	$75.01_{\pm 0.75}$	65.43
200 Training Steps							
Random	$55.60_{\pm 0.49}$	$42.48_{\pm 0.42}$	$57.92_{\pm 1.02}$	$68.05_{\pm 1.16}$	$75.46_{\pm 1.39}$	$72.29_{\pm 0.78}$	61.96
FVT	$\underline{56.34}_{\pm 0.49}$	$44.28_{\pm 0.42}$	$\underline{60.42}_{\pm 1.01}$	$69.90_{\pm 1.14}$	$80.48_{\pm 1.28}$	$74.52_{\pm 0.76}$	64.32
CLP	$54.74_{\pm 0.49}$	$42.50_{\pm 0.42}$	$57.62_{\pm 1.02}$	$67.74_{\pm 1.16}$	$76.82_{\pm 1.36}$	$68.07_{\pm 0.81}$	61.24
SAVA	$56.73_{\pm 0.49}$	$\underline{44.23}_{\pm 0.42}$	$60.90_{\pm 1.01}$	$\underline{69.72}_{\pm 1.14}$	$\underline{79.22}_{\pm 1.31}$	$\underline{73.30}_{\pm 0.77}$	$\underline{64.01}$
LAPT	$58.29_{\pm 0.49}$	$49.31_{\pm 0.42}$	$63.00_{\pm 1.00}$	$69.84_{\pm 1.14}$	$84.13_{\pm 1.18}$	$75.07_{\pm 0.75}$	66.60
2000 Training Steps							
Random	$58.43_{\pm 0.49}$	$46.95_{\pm 0.42}$	$62.87_{\pm 1.00}$	$\underline{71.39}_{\pm 1.12}$	$81.62_{\pm 1.25}$	$72.47_{\pm 0.78}$	65.62
FVT	$59.00_{\pm 0.49}$	$\underline{47.35}_{\pm 0.42}$	$63.52_{\pm 0.99}$	$71.51_{\pm 1.12}$	$84.55_{\pm 1.16}$	$75.74_{\pm 0.74}$	66.94
CLP	$\underline{59.21}_{\pm 0.49}$	$47.10_{\pm 0.42}$	$\underline{63.47}_{\pm 0.99}$	$70.77_{\pm 1.13}$	$84.44_{\pm 1.17}$	$76.75_{\pm 0.73}$	$\underline{66.95}$
SAVA	$59.41_{\pm 0.49}$	$47.57_{\pm 0.42}$	$63.39_{\pm 0.99}$	$71.02_{\pm 1.12}$	$84.55_{\pm 1.16}$	$\underline{76.02}_{\pm 0.74}$	66.99
LAPT	$60.51_{\pm 0.48}$	$46.63_{\pm 0.42}$	$64.99_{\pm 0.99}$	$71.21_{\pm 1.12}$	$85.90_{\pm 1.12}$	$76.17_{\pm 0.74}$	67.56

Table 3: *Mistral-7B-v0.1* 适配模型在意大利语翻译基准上的 0-shot 结果。

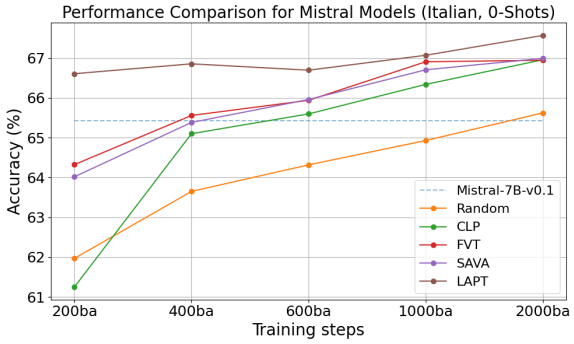

Figure 2: 使用 *Mistral-7B-v0.1* 模型在意大利语翻译的基准数据集上训练期间的平均表现。平均值是基于六个数据集计算的。

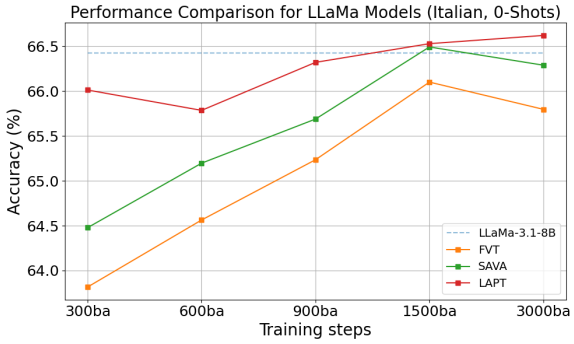

Figure 3: 基于 *Llama-3.1-8B* 模型在意大利语翻译基准上训练期间的平均表现。平均值是在六个数据集上计算得出的。

4 结果

在本节中，我们讨论评估适应模型所获得的结果。我们首先检查选择题基准的得分，然后对生成基准的性能进行单独分析，特别是 FLoRes 和 SQuAD-it。在本节和随后的部分中，我们用 LAPT 缩写表示基本模型在不进行词汇适应的情况下的持续训练。

4.1 多选设置

4.1.1 意大利结果

我们报告了在 Table 3 中，*Mistral-7B-v0.1* 经过 200 和 2000 次批次后的意大利基准测试结果。从表中可以看到，经过适配的模型在训练开始时（200 步设置）就超过了随机机会表现，其中 FVT 和 SAVA 表现出比其他方法（CLP 和随机）更高的性能。与 LAPT 技术相比，所有词汇适应启发式方法表现较差，这是预期中的，因为 LAPT 没有对模型应用任何破坏性结构改变。从 2000 次批次的结果来看，我们可以看到所有经过适配的模型都超过了基础模型的得分，与 LAPT 的表现差距变小。在这种设置下，SAVA 和 FVT 依然表现良好，而随机则落后。

在 Figure 2 中，我们展示了六个意大利任务的平均得分。SAVA 和 FVT 在整个训练过程中始终获得更高的总体得分，尤其是在早期阶段的优势更为明显。这突出了所选启发式方法的影响，尤其是在词汇替换之后。SAVA 和 FVT 在 400 批次时取得的结果与随机方法在训练结束时的结果相当，从而将总训练时间减少了大约 80 %。

在 *Llama-3.1-8B* 的情况下，Table 4 报告了经过 300 和 3000 批适配后模型的分数。我们展示了 FVT 和 SAVA 保持了相当的性能，除了在 BOOLQ 任务中 SAVA 展示了更好的分数，+4% ，即使与 LAPT 设置相比亦如此。与适配后的模型相比，*Llama-3.1-8B* 模型在意大利任务中仍然是一个强有力的基线。在这种设置中，我们通过词汇适应启发式进一步缩小了与 LAPT 模型的性能差距。在 Figure 3 中，我们报告了意大利任务的平均得分，并观察到通过训练步骤持续的改善。

Model	Hellaswag	MMLU	Arc Easy	PIQA	SCIQ	BOOLQ	AVG
LLaMa-3.1-8B	$57.97_{\pm 0.49}$	$54.28_{\pm 0.42}$	$60.46_{\pm 1.01}$	$68.54_{\pm 1.15}$	$82.77_{\pm 1.22}$	$74.52_{\pm 0.76}$	66.42
300 Training Steps							
FVT	$55.61_{\pm 0.49}$	$50.24_{\pm 0.42}$	$59.38_{\pm 1.01}$	$66.99_{\pm 1.17}$	$80.68_{\pm 1.27}$	$70.00_{\pm 0.80}$	63.81
SAVA	$55.48_{\pm 0.49}$	$49.26_{\pm 0.42}$	$59.77_{\pm 1.01}$	$66.62_{\pm 1.17}$	$81.31_{\pm 1.26}$	$74.43_{\pm 0.76}$	64.48
LAPT	$57.92_{\pm 0.49}$	$53.10_{\pm 0.42}$	$61.32_{\pm 1.01}$	$68.97_{\pm 1.15}$	$82.56_{\pm 1.22}$	$72.20_{\pm 0.78}$	66.01
3000 Training Steps							
FVT	$58.44_{\pm 0.49}$	$51.47_{\pm 0.42}$	$62.70_{\pm 1.00}$	$69.53_{\pm 1.14}$	$83.29_{\pm 1.20}$	$69.35_{\pm 0.80}$	65.79
SAVA	$57.82_{\pm 0.49}$	$51.08_{\pm 0.42}$	$63.17_{\pm 1.00}$	$69.78_{\pm 1.14}$	$81.73_{\pm 1.24}$	$74.15_{\pm 0.76}$	66.29
LAPT	$59.35_{\pm 0.49}$	$52.94_{\pm 0.42}$	$62.96_{\pm 1.00}$	$69.72_{\pm 1.14}$	$82.98_{\pm 1.21}$	$71.77_{\pm 0.78}$	66.62

Table 4: 针对 *Llama-3.1-8B* 适应模型的意大利语翻译基准的 0-shot 结果。

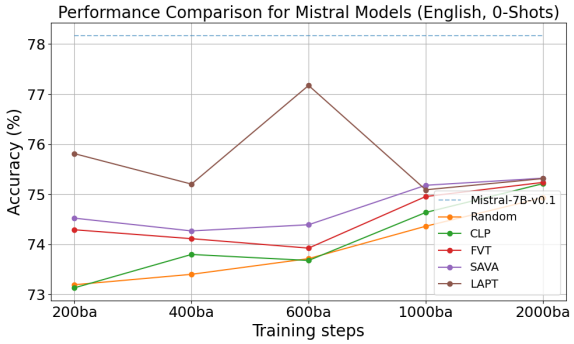

Figure 4: *Mistral-7B-v0.1* 模型在英语基准上的训练期间的平均表现。平均值是在六个数据集上计算的。

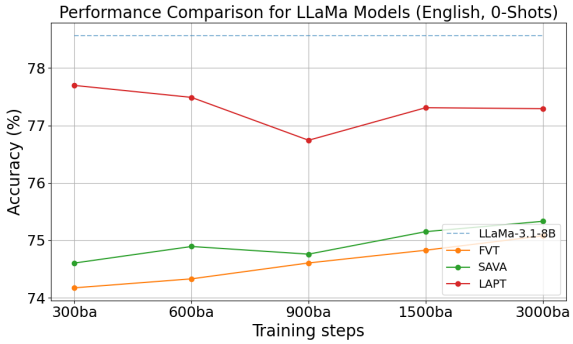

Figure 5: 基于 *Llama-3.1-8B* 模型在训练英文基准时的平均性能。平均值是基于六个数据集计算的。

4.1.2 英文结果

在评估中包含英语可以让我们评估在持续训练过程中，源语言的性能是否得以保留，无论是 *Mistral-7B-v0.1* 还是 *Llama-3.1-8B* 。如同 Section 3 中提到的，我们主要在意大利语数据上进行训练，并辅以一小部分英语数据（占总量的 25 %）。

Figure 4 报告了在英语文本上训练的过程中 *Mistral-7B-v0.1* 的平均分数。我们可以看到所有训练过的模型在适应过程结束时都达到了一个相当的平均分数。所有的适应模型在英语语言中与基本模型相比表现有所下降。

Figure 5 报告了 *Llama-3.1-8B* 模型在训练期间英文基准测试的平均分。在这种设置下，LAPT 平均保持更高的性能；直观上，这可以归因于 *Llama-3.1-8B*（大 75 %）的更大词汇量，这使得在语言适应过程中表现更好，避免了源语言的灾难性遗忘。

对于这两种模型，SAVA 方法使模型在源语言中表现稍微更好。附录 C 中报告了对英语基准测试评估的更详细结果。

基于困惑度评分的多项选择基准测试有其自身的局限性。为了进一步测试我们的模型，我们在两个生成任务上对其进行评估：机器翻译（MT），包括意大利语到英语和英语到意大利语，以及意大利语问答任务。

我们报告了 MT 基准的 COMET-22 (Rei et al., 2022) 和问答任务的 RougeL (Lin, 2004) 。

观察 MT 结果，在 Table 5 中，我们注意到适应后的 *Mistral-7B-v0.1* 模型表现优异，超过了基本模型。这些词汇适应后的模型在英语到意大利语的方向上达到了非常好的结果，其中涉及意大利语文本的生成。我们的研究结果表明，在这种情况下，SAVA 和 FVT 是最有效的词汇适应启发式方法。如在 Table 6 中所示，在 *Llama-3.1-8B* 中也观察到了类似的趋势，其中适应后的模型与基本模型竞争，而 SAVA 和 FVT 取得了与 LAPT 相同的表现。

关于在 SQuAD-it 任务中的结果，Tables 5 and 6 显示 SAVA 获得了非常好的表现，击败了其他启发式方法和 LAPT 方法在这两种模型类型的表现，并且在 *Llama-3.1-8B* 上达到了与基础模型相同的性能。

4.2 训练损失

关于损失轨迹可以做出重要观察。Figure 6 报告了 *Mistral-7B-v0.1* 图，我们可以注意到在训练的早期阶段，各种启发策略之间存在显著差异。SAVA 模型从一开始就表现出更好的适应性，尤其是与 CLP 和随机模型相比时。值

	FLoRes		SQuAD-it
Model	EN-IT	IT-EN	RL
Mistral-7B-v0.1	86.57	87.75	68.92
200 Training Steps			
Random	86.67	87.37	62.1
FVT	87.08	87.55	65.47
CLP	86.58	87.31	64.25
SAVA	87.30	87.59	65.66
LAPT	87.41	87.92	67.35
2000 Training Steps			
Random	88.01	87.92	64.83
FVT	88.29	87.90	66.18
CLP	88.21	87.79	65.99
SAVA	88.31	87.87	67.20
LAPT	88.13	88.02	66.92

Table 5: FLoRes 中 *Mistral-7B-v0.1* 的 5-shot 结果，其中报道了 COMET-22，SQuAD-it 的 2-shot 结果中报道了 RougeL。

	FLoRes		SQuAD-it
Model	EN-IT	IT-EN	RL
Llama-3.1-8B	87.59	88.08	69.21
300 Training Steps			
FVT	87.32	87.65	68.54
SAVA	87.39	87.58	68.70
LAPT	87.82	87.95	67.91
3000 Training Steps			
FVT	88.05	88.02	68.84
SAVA	88.12	88.04	69.05
LAPT	88.11	88.05	66.69

Table 6: 报告了 FLoRes 的 *Llama-3.1-8B* 的 5 次实验结果和 SQuAD-it 的 RougeL 的 2 次实验结果。

得注意的是，CLP 起初似乎落后于随机。从 *Llama-3.1-8B* 损失来看，在 Figure 7 中，我们可以看到两个启发策略表现出类似的轨迹，尽管 SAVA 从一开始就达到了更低的损失。

为了更好地理解不同词汇适应技术的影响，我们分析模型内部和模型间嵌入空间的相似性。具体来说，我们考察不同的适应如何影响嵌入的结构对齐与参考模型的比较（模型内部相似性），以及不同适应模型的嵌入空间如何相互比较（模型间相似性）。

为了测量两个嵌入空间之间的相似度，我们依赖于 Moschella et al. (2023) 引入的技术。具体来说，我们随机选择 128 个非前缀标记和 128 个前缀标记从 V_t 中以计算相对嵌入表示，总共得到 256 个锚标记。对每个模型，我们然后调整每个标记相对于这些锚的表示，计算每个维度作为对选定锚的投影。随后，我们基于

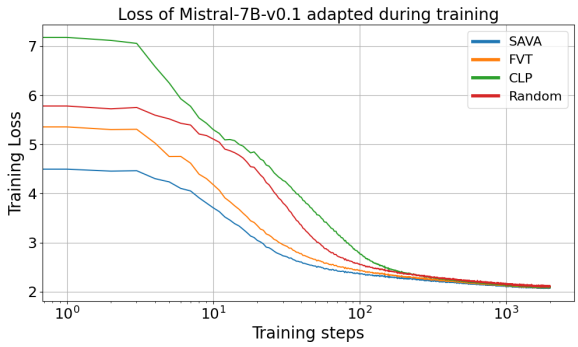

Figure 6: *Mistral-7B-v0.1* 模型在持续训练过程中的损失。

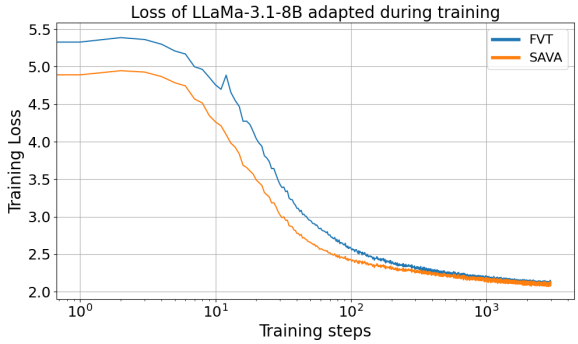

Figure 7: 在 *Llama-3.1-8B* 模型的持续训练中的损失。

此相对表示计算模型之间的余弦相似度，并对结果进行平均，以获得两个不同模型之间的整体相似度得分。

直观地说，一个良好适应的模型应该与 *Minerva-3B* 对齐，因为它作为目标语言的强参考。同样在我们的设置中，*Minerva-3B* 是在平衡的意大利语-英语 CulturaX 数据上预训练的。在 Table 7 中，我们展示了适应模型与 *Minerva-3B* 之间的相似性得分。值得注意的是，CLP 和 SAVA 获得了比其他方法更高的相似性分数。这一结果是预期中的，因为 CLP 和 SAVA 都利用了 *Minerva-3B* 的嵌入空间。有趣的是，SAVA 不仅获得了与 *Minerva-3B* （+3.7）更相似的结构，还表现出更优越的性能，这在前面的部分也是如此。

模型间相似性 为了更深入地了解学习到的嵌入结构之间的差异，Figure 8 提出了使用指定技术调整的 *Mistral-7B-v0.1* 变体之间的相似性评分。我们比较了在持续训练结束时的模型。分析表明模型之间高度相似，但相对表示中的差异高达 10 % 揭示了编码信息的结构变异。该分析表明，即使经过密集训练，适应后模型也不会收敛到相同表示。

在这项工作中，我们广泛探索了各种技术，

	Mistral-7B-v0.1		*Llama-3.1-8B*	
Model	@0ba	@2000ba	@0ba	@3000ba
Random	29.68	31.67	-	-
FVT	33.65	35.30	33.23	33.49
CLP	41.10	42.84	-	-
SAVA	44.81	45.33	41.84	42.02

Table 7: 在训练开始和结束时，*Mistral-7B-v0.1* 适配模型与 *Minerva-3B*（左）以及 *Llama-3.1-8B* 与 *Minerva-3B*（右）之间的相似度得分。

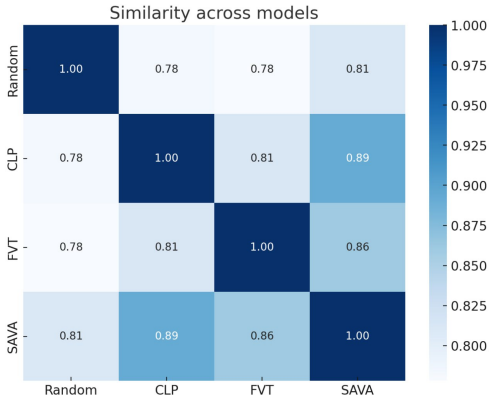

Figure 8: 经过在 12B 个标记上持续训练后的模型相似性。

以将以英语为重点的大型语言模型（如 *Mistral-7B-v0.1* 和 *Llama-3.1-8B*）适应到意大利语。我们引入了一种名为 SAVA 的新启发式方法，该方法利用了一个较小的、原生意大利语语言模型 *Minerva-3B* 的嵌入结构。我们发现，调整英语大型语言模型的词汇表显著提高了语言编码的效果，使生成的标记数量减少了 25 %（对于 *Mistral-7B-v0.1*）和 16 %（对于 *Llama-3.1-8B*）。关于 *Llama-3.1-8B*，我们通过优化其词汇表，修剪了近 10 亿个参数，去除了大约 75 % 的原始标记。我们的评估通过在持续训练阶段的详细分析，揭示了词汇表适应启发式方法之间的性能差异。我们表明，通过相对较少的训练步骤，可以恢复语言能力——在处理了 20 亿个标记后，*Mistral-7B-v0.1* 达到了基础模型的性能。此外，SAVA 启发式在下游任务上显示了出色的性能，SAVA 适应的模型在持续训练期间实现了更快的收敛。此外，与其他分析过的启发式方法相比，SAVA 的嵌入结构与辅助模型的对齐更为紧密。

这项工作开启了多个研究方向。其中一个关键领域是评估 SAVA 方法在不同语言中的扩展性，特别是在中、低资源的环境中。在这种情境下，了解不同启发式方法在较少连续训练步骤中的表现是至关重要的。此外，由于 Minerva-7B 在撰写本文时还不可用，使用它作为辅助模型将是一个合乎逻辑的下一步。

5 局限性

我们研究了以英语为基础的大型语言模型 (Large Language Models) 如何适应意大利语，重点是在调整词汇和分词器的同时达到持续训练模型的表现，并实现较低的冗余率，从而在目标语言中提高效率。

我们将训练数据限制在 CulturaX 数据集上，该数据集由清理过的网络抓取数据组成。包含更高质量的数据集可以提高模型在目标语言中的表现。

我们将分析限制在两个仅解码的大型语言模型：*Mistral-7B-v0.1* 和 *Llama-3.1-8B*。为了进行更全面的研究，可以测试其他以英语为主的模型。然而，上述两个模型是它们参数数量中性能最好的模型之一。此外，我们选择仅关注这两个模型是因为我们必须进行大量的持续训练，而这种训练需要相当多的计算资源。

我们在自动翻译的数据集上评估了适应后的模型，这些数据集用于多项选择任务和开放式问题回答。具体来说，Hellaswag、MMLU、Arc Easy、PIQA、SCIQ 和 BOOLQ 是使用 Tower-Instruct-v0.2 翻译的，这是一种用于自动翻译的开源解决方案，在撰写本文时，它代表了开放式机器翻译模型的最新技术水准。对于生成任务，SQuAD-it 是使用一种半自动的方法进行翻译的。

我们承认，依赖自动翻译的基准测试可能引入了一些噪音，可能掩盖了模型对意大利文本理解的某些能力或问题。由于不存在结构良好的意大利语本地基准，这一限制超出了我们的解决能力。另一限制是仅使用了两个生成基准，我们观察到对于适应模型，结果稍有不同。在生成设置中，SAVA 通常优于其他方法，而 LAPT 模型在下游任务中未能始终提供最佳的平均性能。

未来的工作应该着眼于探索词汇适应模型在生成任务中的能力，并研究模型在目标语言上的生成能力如何影响下游性能。

6 伦理声明

我们主要用意大利语进行实验。这种方法旨在解决使用意大利语工作的实际挑战，意大利语在自然语言处理领域中代表性不足。我们的持续训练是在从开放网络资源中收集的数据上进行的，特别是通过 CulturaX 数据集。由于用于预训练的大规模数据集可能包含个人和敏感信息，因此在将模型部署到实际应用中之前，必须仔细评估这些内容。另一个关键考虑是使用现有的单语或多语种模型作为起点，而不是从

头开始训练新模型。这可能会引入来自原始预训练数据的偏差，可能导致模型反映其他语言的行为和文化影响，而不是目标语言社区的。

7

致谢 Edoardo Barba 和 Alessio Miaschi 全额由 PNRR MUR 项目 PE0000013-公平 资助。Roberto Navigli 和 Felice Dell'Orletta 感谢 PNRR MUR 项目 PE0000013-FAIR 的支持。部分资金由欧洲联盟通过意大利大学和研究部的下一代欧盟计划资助，支持项目为 PNRR - PRIN 2022 (2022EPTPJ9) "WEMB: 从认知语言学到语言工程及其往返的词嵌入"，以及 PNRR 项目 ITSERR（CUP B53C22001770006）。我们感谢 ISCRA 项目 TRAVEL (HP10CY9V7K) 对使用 LEONARDO 超级计算机的支持，该计算机由泛欧高性能计算联合企业拥有，托管于 CINECA（意大利），并感谢 Giuseppe Fiameni 的支持。

References

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. Preprint , arXiv:2312.09993.

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

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics , 19(2):263–311.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In Findings of the Association for Computational Linguistics: EMNLP 2020 , pages 1324–1334, Online. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In 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) , pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv:1803.05457v1 .

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 8440–8451, Online. Association for Computational Linguistics.

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

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In AI*IA 2018 – Advances in Artificial Intelligence , pages 389–402, Cham. Springer International Publishing.

Zoltan Csaki, Bo Li, Jonathan Lingjie Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. SambaLingo: Teaching large language models new languages. In Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024) , pages 1–21, Miami, Florida, USA. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 , pages 836–846, Online. Association for Computational Linguistics.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing , pages 13440–13454, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 .

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. 2022. Fast vocabulary transfer for language model compression. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track , pages 409–416, Abu Dhabi, UAE. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 8342–8360, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR) .

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 .

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825 .

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems , 35:22199–22213.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles .

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. Preprint , arXiv:2211.05100.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In Findings of the Association for Computational Linguistics: NAACL 2024 , pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. 2024. Latent space translation via semantic alignment. Advances in Neural Information Processing Systems , 36.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. volume abs/2403.08295.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Luca Moroni, Simone Conia, Federico Martelli, and Roberto Navigli. 2024. ITA-Bench: Towards a more comprehensive evaluation for Italian LLMs. In Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024) .

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 . OpenReview.net.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) , pages 4226–4237, Torino, Italia. ELRA and ICCL.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Edoardo Barba, Simone Conia, Sergio Orlandini, Giuseppe Fiameni, Roberto Navigli, et al. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. In Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024) .

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning. arXiv preprint arXiv:2301.09626 .

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. Preprint , arXiv:2405.07101.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing , pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT) , pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Findings of the Association for Computational Linguistics ACL 2024 , pages 7407–7416, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. CoRR , abs/2411.12372.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In Proceedings of the 3rd Workshop on Noisy User-generated Text , pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In Proceedings of the 20th Chinese National Conference on Computational Linguistics , pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A SAVA 映射函数的训练

为了实现 SAVA 方法，我们首先需要训练线性映射函数 ϕ。为此，我们使用 latentis 库 [6] 中提供的 `SGDAffineAligner` 方法。

从交集中收集到标记表示对后，我们使用 `ADAM` 优化器和 `MSE Loss` 进行线性映射训练，将学习率设置为 10^{-3} ，并运行 1000 步的优化。

为了增强训练的稳定性，我们在学习 ϕ 之前，首先对标记表示应用标准缩放和 L2 归一化。在训练之后，我们应用逆缩放来恢复原始分布，然后将结果合并到调整后的模型中。

在本节中，我们分析了 SAVA 方法的一些消融实验。我们分析了对助手模型大小的影响，使用了 Minerva 家族的两个较小模型，*Minerva-350M* 和 *Minerva-1B* ，其参数分别为 350M 和 1B。在 Figure 9 中，报告了使用 SAVA 并采用不同助手模型的 *Mistral-7B-v0.1* 的训练损失。从图中可以看到，助手模型的维度对损失轨迹没有巨大影响。为了研究学习映射 ϕ 所用标记数

[6] https://github.com/Flegyas/latentis

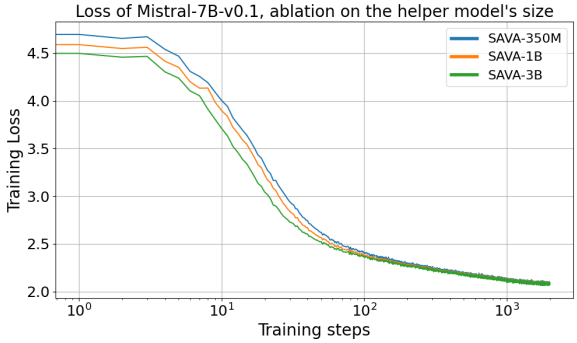

Figure 9: Mistral 模型连续训练中的损失

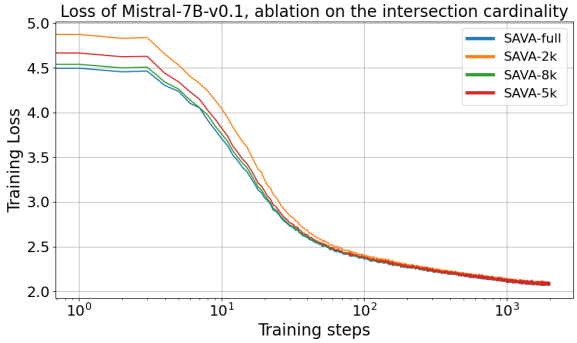

Figure 10: 在持续训练 Mistral 模型中的损失。

量的影响，进行了一个正交实验，在 Figure 10 中，报告了在 *Minerva-3B* 上，使用 SAVA 并依赖于 $V_t \cap V_s$ 中不同标记数量的 *Mistral-7B-v0.1* 的损失。我们观察到，使用更多的标记会导致训练损失更快收敛。从图中可以看到，减少标记数量比减少模型大小对结果的影响更大，特别是对于使用两千个标记的设置。

B 培训资源和环境影响

实验是在 LEONARDO 意大利超级计算机上进行的，该超级计算机的碳效率为 0.432 kgCO $_2$ eq/kWh。总共进行了 50000 小时的计算，使用了类型为 A100 SXM4 80 GB（TDP 为 400W）的硬件。

总排放量估计为 8640 kgCO $_2$ eq，其中 0% 被直接抵消。这些排放量大致分为 95 % 用于持续训练和 5 % 用于评估。

这是一个近似估计，因为计算是在 LEONARDO 定制硬件上进行的，而该硬件在用于估计的工具 中不可用。

我们在生成设置中测试了我们改进后的模型在两个下游任务上的表现，分别是机器翻译和问答。在少样本设置中，我们依赖于被评估模型的上下文能力来进行测试，而无需针对具体任务进行微调步骤。我们依赖于 vLLM 库来支持提示生成，特别是我们用 `temperature=0` 和 `max_tokens=512` 更改了默认参数。

Prompt EN-IT	Prompt IT-EN
Traduci dall'Inglese all'Italiano Text: I love you so much. Translation: Ti amo così tanto.	Translate from Italian to English Text: Ti amo così tanto. Translation: I love you so much.

Table 8: 用于机器翻译任务的提示

Italian Prompt
Contesto: Il terremoto del Sichuan del 2008 o il terremoto del Gran Sichuan, misurato a 8.0 Ms e 7.9 Mw, e si è verificato alle 02:28:01 PM China ... Domanda: In quale anno si è verificato il terremoto nel Sichuan? Risposta: 2008

Table 9: 用于问答任务的提示

经过大量的试验后，我们注意到提示策略有很大的影响，而模型之间的顺序保持不变。我们在 Tables 8 and 9 中分别报告了用于 FLoRes 和 SQuAD-it 任务的提示。

C 英语多项选择基准测试结果

在本节中，我们对英语基准测试的评估结果进行了详细分析。Table 10 报告了 *Mistral-7B-v0.1* 在六个多项选择基准测试上的表现。从该表中可以看到，SAVA 和 FVT 在适应过程的早期阶段取得了更高的任务分数。对于适应了 *Llama-3.1-8B* 的模型，同样的趋势也很明显，如 Table 11 所示，SAVA 技术在训练的开始和结束时都获得了比 FVT 更高的平均得分。对于这两种模型，单项任务得分均低于基线模型的性能。然而，在适应过程中加入部分英语数据可防止转向意大利语时的灾难性遗忘。

Model	Hellaswag	MMLU	Arc Easy	PIQA	SCIQ	BOOLQ	AVG
Mistral-7B-v0.1	$75.98_{\pm 0.44}$	$57.19_{\pm 0.42}$	$78.55_{\pm 0.94}$	$83.84_{\pm 0.94}$	$95.82_{\pm 0.80}$	$77.64_{\pm 0.78}$	78.17
200 Training Steps							
Random	$72.29_{\pm 0.44}$	$51.59_{\pm 0.42}$	$69.55_{\pm 0.95}$	$81.73_{\pm 0.96}$	$89.97_{\pm 0.97}$	$74.03_{\pm 0.76}$	73.19
FVT	$72.35_{\pm 0.44}$	$53.04_{\pm 0.42}$	$73.08_{\pm 0.92}$	$82.60_{\pm 0.94}$	$92.48_{\pm 0.85}$	$72.20_{\pm 0.78}$	74.29
CLP	$72.59_{\pm 0.44}$	$52.02_{\pm 0.42}$	$70.16_{\pm 0.94}$	$81.55_{\pm 0.96}$	$89.66_{\pm 0.98}$	$72.81_{\pm 0.77}$	73.13
SAVA	$72.81_{\pm 0.44}$	$53.21_{\pm 0.42}$	$74.28_{\pm 0.94}$	$82.47_{\pm 0.96}$	$92.79_{\pm 0.83}$	$71.59_{\pm 0.78}$	74.52
LAPT	$74.13_{\pm 0.43}$	$55.05_{\pm 0.42}$	$75.23_{\pm 0.89}$	$84.02_{\pm 0.91}$	$94.46_{\pm 0.73}$	$71.98_{\pm 0.78}$	75.81
2000 Training Steps							
Random	$72.18_{\pm 0.44}$	$52.11_{\pm 0.42}$	$73.6_{\pm 0.91}$	$82.72_{\pm 0.94}$	$93.21_{\pm 0.81}$	$75.77_{\pm 0.74}$	74.93
FVT	$73.28_{\pm 0.44}$	$52.96_{\pm 0.42}$	$74.76_{\pm 0.90}$	$81.91_{\pm 0.95}$	$94.05_{\pm 0.76}$	$74.46_{\pm 0.76}$	75.23
CLP	$73.37_{\pm 0.44}$	$52.48_{\pm 0.42}$	$74.07_{\pm 0.90}$	$82.47_{\pm 0.94}$	$94.05_{\pm 0.76}$	$74.83_{\pm 0.75}$	75.21
SAVA	$73.02_{\pm 0.44}$	$52.91_{\pm 0.42}$	$74.67_{\pm 0.90}$	$82.29_{\pm 0.94}$	$94.46_{\pm 0.73}$	$74.58_{\pm 0.76}$	75.32
LAPT	$74.26_{\pm 0.43}$	$51.18_{\pm 0.42}$	$73.9_{\pm 0.91}$	$83.65_{\pm 0.92}$	$94.67_{\pm 0.72}$	$74.22_{\pm 0.76}$	75.31

Table 10: 适用于 *Mistral-7B-v0.1* 模型的英语基准的 0 次结果。

Model	Hellaswag	MMLU	Arc Easy	PIQA	SCIQ	BOOLQ	AVG
LLaMa-3.1-8B	$74.21_{\pm 0.43}$	$62.19_{\pm 0.42}$	$77.60_{\pm 0.47}$	$83.03_{\pm 0.93}$	$93.94_{\pm 0.77}$	$80.42_{\pm 0.69}$	78.56
300 Training Steps							
FVT	$72.35_{\pm 0.44}$	$58.22_{\pm 0.42}$	$69.55_{\pm 0.95}$	$81.30_{\pm 0.97}$	$92.27_{\pm 0.86}$	$71.34_{\pm 0.79}$	74.17
SAVA	$72.72_{\pm 0.44}$	$58.19_{\pm 0.42}$	$70.75_{\pm 0.94}$	$81.79_{\pm 0.96}$	$92.90_{\pm 0.83}$	$71.28_{\pm 0.79}$	74.60
LAPT	$74.35_{\pm 0.43}$	$61.74_{\pm 0.41}$	$76.14_{\pm 0.88}$	$83.21_{\pm 0.93}$	$94.05_{\pm 0.76}$	$76.69_{\pm 0.73}$	77.69
3000 Training Steps							
FVT	$73.02_{\pm 0.44}$	$57.85_{\pm 0.42}$	$72.13_{\pm 0.93}$	$82.04_{\pm 1.15}$	$92.90_{\pm 0.83}$	$72.53_{\pm 0.78}$	75.07
SAVA	$72.86_{\pm 0.44}$	$57.94_{\pm 0.42}$	$72.78_{\pm 0.92}$	$81.79_{\pm 0.96}$	$93.31_{\pm 0.80}$	$73.30_{\pm 0.77}$	75.33
LAPT	$74.40_{\pm 0.43}$	$60.50_{\pm 0.42}$	$75.32_{\pm 0.89}$	$82.47_{\pm 0.94}$	$93.63_{\pm 0.78}$	$77.43_{\pm 0.73}$	77.29

Table 11: *Llama-3.1-8B* 适应模型在英语基准测试上的 0-shot 结果。