

为低资源语言指导大型语言模型： 巴斯克语的系统性研究

Oscar Sainz Naiara Perez Julen Etxaniz Joseba Fernandez de Landa
Itziar Aldabe Iker García-Ferrero Aimar Zabala Ekhi Azurmendi
German Rigau Eneko Agirre Mikel Artetxe Aitor Soroa
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{ oscar.sainz,a.soroa }@ehu.eus

Abstract

将语言模型与用户意图相结合需要庞大的指令数据集，这类数据集仅适用于有限的语言集。在本文中，我们探讨在资源匮乏的情况下，对传统指令适应流程的替代方案。我们假设一种针对低资源语言的现实情境，其中仅提供以下资源：目标语言的语料库、现有的开放权重多语言基础和指令化基础大型语言模型，以及从指令化基础中抽样生成的合成指令。我们为巴斯克语提供了一套全面的实验，系统性地研究了这些组件的不同组合，并根据 1,680 参与者的基准测试和人类偏好进行评估。我们的结论表明，目标语言的语料库至关重要，合成指令能够生成稳定的模型，更重要的是，使用指令调优的模型作为基础模型可以优于使用未指令化的基础模型，并在扩展时获得更好的结果。使用 Llama 3.1 instruct 70B 作为基础模型，我们的模型在巴斯克语上的效果接近更大尺寸的前沿模型，而除了 1.2B 词语料库外未使用任何巴斯克语数据。我们发布了¹代码、模型、指令数据集和人类偏好，以支持未来低资源语言适应研究的完全可重复性。

1 介绍

大型语言模型 (LLMs)，尤其是开放模型，仍然主要以英语为中心，对世界上绝大多数语言的覆盖非常有限。尽管最近在开放 LLMs 的预训练过程中加入了额外语言的努力，显著的性能差距依然存在。即便是最新的指令微调模型，在处理资源匮乏语言时表现出明显的能力下降 (?)。更为重要的是，与基础模型和指令微调模型相比，后训练过程中英语为中心的性质加大了语言之间的性能差距。

为了克服这些限制，可以通过持续训练并使用有限资源将开放模型适应于新语言。在特别是经过指令调优的模型的情况下，出现了各种努力，通常遵循一个顺序的方法：首先通过持续的预训练来适应基础模型，然后执行指令调优。虽然这种多步骤过程已成为标准做法，但

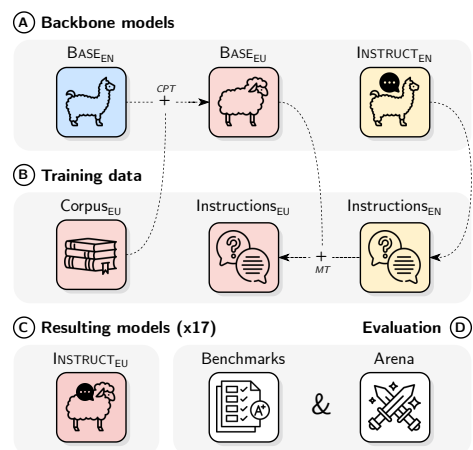


Figure 1: 系统地探索低资源语言的指令微调策略。我们的框架包括：(A) 三个骨干模型（现有基础模型，在目标语言上继续预训练的基础模型，以及现有指令模型）；以及 (B) 种不同的训练数据组合，包括目标语言语料和利用骨干模型采样和/或翻译的合成指令。我们从这些组件的所有可能组合中训练 (C) 个实验模型，并通过静态基准和人类偏好进行 (D) 全面评估，以识别最佳适应路径。

对于其他适应策略的探索很少。我们质疑是否可以在没有专门指令数据的情况下，将指令执行能力直接转移到新语言，或者指令调优模型是否可以通过持续预训练来适应，类似于基础模型。

具体来说，这项工作系统地探讨了不同于传统流程的多种策略，以开发针对低资源语言的指令调优模型，并寻求为巴斯克语确定最佳适应路径作为我们的主要案例研究（见 Fig. 1）。我们有意将探索限制在易于获取或使用开放模型创建的资源上，避免依赖于从商业最先进系统的蒸馏。虽然我们的研究集中于单一语言，但我们的发现可能适用于许多具有相似资源的语言：巴斯克语是一个理想的测试案例，在 Common Crawl 中的排名大约为 50，存在量大约比英语小 1,000 倍²，并且显著缺乏预先

¹存储库：<https://github.com/hitz-zentroa/latxa-instruct>

²<https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html>

存在的指令数据集。这一资源状况反映了全球众多其他低资源语言面临的挑战。

在解决这个研究问题时，我们进一步面临指令跟随大型语言模型（LLM）评估中的一个关键挑战：自动化指标常常忽略用户关心的功能（?）。因此，我们开发了一个结合传统基准测试和众包 LLM 竞技场的评估框架，在该框架中，我们发动了巴斯克语社区进行大规模评估工作，收集了超过 12,000 条来自 1,680 位参与者的偏好注释。该倡议构成了迄今为止对低资源语言进行的人类评价的最大努力。

通过这一评估，我们的系统性探索为在低资源语言中开发指令调整模型提供了三个关键见解：(1) 目标语言语料库对性能至关重要——缺乏接触单纯的巴斯克语文本的模型表现出退化，不论其他技术如何；(2) 虽然单语和双语指令数据集都显示出益处，但后者在基准测试和人工评估中产生了一致的结果；(3) 从一个指令调整的英语模型开始优于让一个基础模型学习遵循指令的方法，这对应用于低资源语言的标准流程提出了挑战。

除了这些主要发现之外，我们的工作对该领域还有以下贡献：(4) 首次发布专为巴斯克语调整的指令系列 LLM，包含 8B 和 70B 参数的模型，后者在竞技场中被证明与 GPT-4o 和 Claude Sonnet 具有竞争力；(5) 发布大规模的、合成的英语和巴斯克语的指令调整数据集；以及 (6) 发布首个包含巴斯克语偏好数据集，内含真实用户提示、模型响应和偏好注释，可以支持未来的偏好对齐研究。通过这些贡献，我们旨在推动巴斯克语的语言技术水平，同时建立适用于其他低资源语言的方法论。

2 相关工作

关于为资源匮乏语言开发大型语言模型（LLMs）的研究探讨了多种方法，取得了不同程度的成功。最初尝试从头开始为特定低资源语言开发模型，由于训练数据有限，证明是具有挑战性的。多语言模型开发已成为更加有前景的策略，研究人员利用跨语言迁移学习来改善性能（???）。迄今为止最有效的方法是继续对现有多语言模型进行预训练，这允许进行语言特定的适应，同时受益于大型训练语料库中的丰富语言表示（???）。尽管在开发这些基础模型方面取得了进展，但为资源匮乏语言指导和微调它们的最佳方法仍然基本未被探索（???）。

对资源匮乏语言进行指令微调的研究已经探索了各种方法来克服本地指令数据的不足。不同的研究利用以英语为中心的预训练模型或多语言模型作为跨语言迁移的枢纽架构（?）。关

于数据，研究人员已经探索了结合多语言指令数据集的方法，这些数据集包括对资源匮乏语言的有限覆盖（?）；将现有的英语指令集翻译成目标语言，无论是自动翻译还是人工验证（??）；以及应用数据增强技术，如反向翻译、语言特定的提示以及基于模板的指令生成，以扩展有限的资源（?）。此外，跨语言的上下文学习已显示出有趣的结果（?）。

关于巴斯克语言的适应，已经进行了两项重要研究。? 通过持续预训练来调整 Llama 2 模型开发了 Latxa。同时，? 通过调整 Llama 3.1 创建了 Llama-eus，并随后使用机器翻译数据进行指令调优和偏好对齐，遵循广泛接受的方法。然而，前者仅关注基础模型，不考虑指令调优模型，而后者仅实施了一种单一策略来指导适应后的语言模型。相比之下，这项工作系统地探索了多种策略和组合，以有效地指导（或调整）巴斯克语言模型。

3 资源

指导 LLM 通常依赖于两个组件：基础（或基础性）LLM 和指令数据集。对于非统治性语言来说，获取指令数据集可能非常困难，尤其是在低资源语言场景中。特别是对巴斯克语而言，没有手动生成的，甚至没有质量好的自动生成的大型指令-答案对。因此，如 Fig. 1 所示，我们的可用资源仅限于目标语言的语料库，以及高资源语言的基础和指导模型。从这些有限的资源中，我们通过合成数据生成和模型适应的战略组合，得出创建巴斯克指令调优模型所需的组件。在接下来的部分中，我们描述这些种子资源及其衍生。

对于预训练数据，我们利用了训练 Latxa 的语料库，这是专门为巴斯克语训练的第一个大型语言模型家族。这一语料库包括 4.3 百万篇高质量的巴斯克语文档，大约有 35 亿个 Llama 3.1 标记。在这些来源中，它包含了通过特制的爬虫提取的高质量新闻数据、维基百科以及基于 Common Crawl 的来源，如 CulturaX、Colossal OSCAR 和 HLPT v1.1。该语料库已经过规范化、去重和过滤。数据可在 HuggingFace 中心公开获得。我们将此语料库称为 Corpus EU。

我们使用 Llama 3.1 作为我们的基础 LLM（即未经过微调以遵循对话式指令的模型）。Llama 3.1 是一个公开可用的模型，由于其在英语和其他高资源语言中的强大性能而被广泛采用。我们在整篇论文中将此模型称为 BASE EN。此外，按照 ? 的方法，我们基于 Llama 3.1 训练了一个新的 Latxa 模型，我们将其表示为 BASE EU。对于指令微调的模型，我们采用类

似的策略，使用 Llama 3.1 的指令版，我们称之为 INSTRUCT_{EN}。

3.1 指令抽样和翻译

现有的（英文）指令数据集依赖于高质量、人工制作的指令和响应（例如，没有机器人）、³ 完全自动生成的指令和响应（??），或两者的组合，例如手动编写的提示与自动生成的响应配对（?）。使用这些数据集中的任何一个都会在我们的分析中引入额外的混杂因素（即，从强大的 LLM 中提炼出的知识），这可能导致在此类数据上训练的模型优于我们的 INSTRUCT_{EN}，从而给我们的评估引入噪声。这会引发一个不在本文范围内的独立研究问题：训练模型的最佳指令数据集（组合）是什么？然而，在巴斯克语的情况下，没有公开可用的指令集。以下段落详细说明了为每种语言生成指令的过程。

为了避免外部影响，我们直接从我们的 INSTRUCT_{EN} 模型中抽取指令。我们按照（?）的方法生成英语指令。使用此技术，我们使 INSTRUCT_{EN} 生成不同类型和任务的指令：通用型、代码、数学、算术和翻译。我们总共生成了 400 万条英语指令，然而，在进行超参数搜索之后，我们发现只使用 100 万条指令在整体上取得了更好的结果。我们在 ?? 中分享了该过程的更多细节和示例。

巴斯克语指令。 我们使用 BASE_{EU} 的少量提示翻译了从 INSTRUCT_{EN} 中抽取的指令。现有的英语-巴斯克语机器翻译系统（例如，NLLB（?））主要在句子级的文本数据上进行训练，通常在处理更复杂的输入时表现不佳，包括选择性翻译嵌入在代码片段中的自然语言内容。通过利用像 BASE_{EU} 这样的 LLM，这种模型接触过多样化的数据类型，我们在这种情况下获得了更高质量的翻译。此外，使用在我们自己的实验框架内训练的模型可以避免将外部因素引入我们的流程中。关于翻译指令过程和使用的提示的更多细节在 ?? 中给出。

4 实验设置

我们将实验设置形式化如下。设 $\mathcal{M} = \{\text{BASE}_{\text{EN}}, \text{BASE}_{\text{EU}}, \text{INSTRUCT}_{\text{EN}}\}$ 为骨干模型的集合， $\mathcal{D} = \{\text{Corpus}_{\text{EU}}, \text{Instructions}_{\text{EN}}, \text{Instructions}_{\text{EU}}\}$ 为二进制变量的集合，指示是否使用巴斯克语语料、英语指令和/或巴斯克语指令。因此，可能的配置空间为 $\mathcal{M} \times \mathcal{P}(\mathcal{D})$ ，其中 $\mathcal{P}(\mathcal{D})$ 是 $\mathcal{D}$ 的幂集，产生 $|\mathcal{M}| \times 2^{|\mathcal{D}|} = 3 \times 2^3 = 24$ 种理

论组合。请注意，我们探索同时利用原始文本和指令数据的训练策略。在总共 24 种组合中，我们排除了那些模型在原始训练数据上再次训练的冗余配置。因此，最终得到的不同指令微调模型变体包括 18 种配置：原始的 Llama 3.1 Instruct 8B（即 INSTRUCT_{EN}）和 17 个新的 8B 大小的模型。?? 在 ?? 提供了所有模型变体及其简写名称的完整说明。此外，我们还根据初步基准评估中表现最好的配置训练了一个 70B 的模型。关于基线，我们分析中的主要基线是 INSTRUCT_{EN} 模型，因为它是唯一能够遵循指令的骨干模型。然而，由于我们单独考察每个变量 $\mathcal{D}$ 的效果，具体的比较点在各个案例中有所不同。为提供额外背景，我们还评估了两个以巴斯克语强性能闻名的私有模型：⁴ OpenAI 的 GPT-4o⁵ 和 Anthropic 的 Claude 3.5 Sonnet。⁶

5 评估

我们采用了两种互补的评估方法来评估每个指令调优方案的影响。一方面，我们使用了一些静态基准，通过标准化测试来评估模型的特定能力和知识。另一方面，我们通过 A/B 测试（竞技场风格）进行了人工评估，以捕捉模型表现的定性方面。

5.1 静态基准测试

我们选择了接近实际用例的基准，从多个类别中选取：阅读理解、常识、语言能力、知识、数学推理和偏见。对于每个基准，如果可能，我们评估了巴斯克语、英语和西班牙语版本，以便分析每个微调模型版本的语言特定权衡。我们包括西班牙语是因为它与巴斯克语有密切的社会语言学关系，并且评估我们的指令调优方法如何影响实验中未直接针对的第三种语言。因此，我们总共评估了 29 个基准，详细信息见于 ??。

为了进行这些评估，我们依赖于 Eleuther AI 的 LM 评价工具（?）。大多数数据集被设定为多项选择问题，模型的答案通过选择具有最高对数概率的选项来确定。对于生成任务，答案直接从模型中采样。为了给模型提供上下文示例，我们的评估采用了少量样本设置。所有结果均依据标准、公开的实现测量准确性。有关详细信息，请参阅 ??。

在评估专有模型时，我们无法直接计算对数概率，因为我们无法访问模型权重。这种限制使我们只能评估那些被实现为显式选择题（A，

⁴我们没有在评估中包括被指导的 Llama-eus（?），因为在实验时它尚未公开发布。

⁵gpt-4o-2024-11-20

⁶claude-3-5-sonnet-20241022

³[hf.co/datasets/HuggingFaceH4/no_robots](https://huggingface.co/datasets/HuggingFaceH4/no_robots)