大模型时代的多语言研究综述

高长江　周昊　佘帅杰　钟昊鸣　刘斯哲　赖哲剑　王志军　黄书剑†
计算机软件新技术国家重点实验室, 南京大学
{gaocj, zhouh, shesj, zhonghm, liusz, laizj, wangzj}@smail.nju.edu.cn
huangsj@nju.edu.cn

摘要

进入大语言模型时代以来，传统的多语言研究模式发生了巨大变化。一些传统任务得到了突破性的解决，也出现了多种新任务，以及许多以多语言大模型为基础、面向大模型能力提升的多语言研究工作。本文针对研究领域中的这一新变化，整理归纳了进入大模型时代以来的多语言研究进展，包括多语言大模型、数据集、任务，以及相关的前沿研究方向、研究挑战等，希望能为大模型范式下的多语言研究的未来发展提供参考和帮助。

关键词： 大语言模型； 多语言； 跨语言

A Survey of Multilingual Research in the Large Language Model Era

Changjiang Gao, Hao Zhou, Shuaijie She, Haoming Zhong
Sizhe Liu, Zhejian Lai, Zhijun Wang, Shujian Huang†
National Key Laboratory for Novel Software Technology, Nanjing University
{gaocj, zhouh, shesj, zhonghm, liusz, laizj, wangzj}@smail.nju.edu.cn
huangsj@nju.edu.cn

Abstract

Since we enter the era of large language models, there have been significant changes in the traditional multilingual research paradigm. Some traditional tasks have been solved in a groundbreaking manner, new tasks have emerged, and many multilingual research works based on multilingual large language models and aimed at enhancing the capabilities of these models have been developed. This article focuses on this new change in the research field, summarizing the progress of multilingual research since the era of large models, including multilingual large language models, datasets, tasks, as well as related cutting-edge research directions, research challenges, etc., aiming to provide reference and assistance for the future development of multilingual research under the large model paradigm.

Keywords: Large Language Models , Multilingual , Cross-Lingual

1 引言

近年来，以 Transformer 模型 (Vaswani et al., 2017) 为基础架构，以大规模预训练 (Devlin et al., 2019) 为基本技术的大语言模型对自然语言处理领域的研究范式产生了革命性的影响 (Xu et al., 2024)。这些大模型不仅在多种自然语言处理任务上展现出了远超预期的性能，还显示

出了一些惊人的“涌现”能力，例如上下文学习 (Brown et al., 2020)，思维链能力 (Wei et al., 2023a; Wang and Zhou, 2024) 等。因此，面向大语言模型的研究，已经成为当前自然语言处理领域，乃至更广泛的人工智能领域中的重要趋势。其中，多语言研究作为自然语言处理与语言学学科的重点问题，也受到了强烈的影响。

由于训练语料中包含不同语言的大量数据，现有的大模型往往也是多语言大模型（MLLM）(Briakou et al., 2023; Qin et al., 2024)，并且展现出了强大的多语言能力，包括机器翻译能力 (Garcia et al., 2023; Fu et al., 2023; Li et al., 2024)，多语言推理能力 (Shi et al., 2022)。这些模型既可用于完成传统多语言任务，也为一些新的多语言应用场景提供了可能。同时，随着多语言预训练语料、指令对话数据、评估测试数据集的完善，多语言大模型的数量不断增加，质量也在不断提高。

然而，尽管多语言大模型的性能强大、应用前景广泛，此方面的研究仍面临问题与挑战。首先，多语言大模型的最适宜训练方式仍在完善中，在预训练、有监督微调（SFT）、人类反馈强化学习（RLHF）、下游任务微调等不同训练阶段中如何加入多语言信息 (Qin et al., 2024)，这些阶段的多语言训练分别有何影响 (Gao et al., 2024a)，都还在不断研究探索之中。其次，由于高质量训练语料以英文等高资源语言为主，多语言大模型往往表现出语言偏好与不平衡，在不同语言上展现出性能差距，以及不同语言的知识与语义表示没有对齐等。同时，能否拓展模型的语言能力，将已有的以英语为核心的大模型拓展为覆盖更多语言的模型，并且保持、迁移其英语性能，对实际应用也有重要指导意义。以及，由于深度学习模型的“黑箱”特点，研究者对多语言大模型如何产生多语言能力、内部计算过程与状态变化如何，仍没有明确的认识，需要通过新的观察手段与理论进行可解释性分析。最后，多语言大模型在不同语言上还面临公平性、安全性问题，灾难性遗忘与多语言诅咒，多语言训练代价大等挑战，需要进一步研究解决。

对此，本文将立足大模型时代的多语言研究，分别归纳多语言大模型的领域现状、前沿研究方向、面临的挑战，介绍多语言大模型的训练资源、训练手段、需要解决的问题等，为未来此方面的研究提供参考。

多语言模型类型	模型列表
基座模型	OpenAI ChatGPT, Google Gemini, BLOOM, PALM, LLaMA, Baichuan, DeepSeek, GLM, Qwen, ...
指令模型	Chines-LLaMA-Alpaca, BayLing, Vicuna, x-LLaMA/m-LLaMA, ...
特定任务模型	XGLM-7B, QAlign, MAPO, ...

Table 1: 代表性多语言大模型

多语言数据集类型	数据集列表
预训练语料	Wikipeida, mC4, ROOTs, CulturaX, ...
指令对话训练数据	xp3, OpenAssistant Conversation, Bactrain-X, Ayadataset, ...
评估测试数据集	理解任务：XCOPA, XStoryCloze, BELEBELE, XWinograd, PAWS-X, XNLI, ... 生成任务：XL-Sum, MLQA, MKQA, MGSM, FLORES, ...

Table 2: 代表性多语言数据集

2 研究现状概述

本部分将介绍目前多语言研究领域的基本情况，包括代表性多语言大模型、数据集、任务，并将特别介绍大模型时代新出现的多语言任务。

2.1 代表性多语言大模型

2.1.1 多语言基座模型

多语言基座模型是指那些在多种语言上进行了预训练，并且能够同时支持多种语言的模型。这些模型通常是在大规模的跨语言语料库上进行训练，以学习不同语言之间的共享特征和差异（见表 1）。

商业大模型，如 OpenAI GPT 系列模型，以及 Google Gemini 模型等，为目前全球应用较为广泛的多语言大语言基座模型，具有较强的多语言能力，且适用于不同下游任务。然而，由于模型架构、训练数据与训练方式均不公布，难以确定其多语言能力的基础与来源。

BLOOM (Scao et al., 2023) 是由 Big-Science 团队训练的的开源大模型，具有 176B 参数。BLOOM 在包含在 ROOTS 语料库上训练，能支持 59 种不同的语言，包括 46 种编程语言和 13 种自然语言。

PaLM (Chowdhery et al., 2022) 由 Google 发布，包含 540B 参数的大模型，在训练过程中使用了多语言维基百科和对话数据的混合，包括 124 种语言，其中英文词元在总词元中占比 78%。PaLM 在多语言测试上展现出强大的能力。

LLaMA 系列模型 (Touvron et al., 2023b) 是目前应用较为广泛的的开源大模型，其中 LLaMA-2 与 LLaMA-3 都在训练数据中有意提高了多语言预训练语料的占比，使模型的语言建模能力覆盖了较多不同语种。近期发布的 LLaMA-3 对模型的词表也进行了扩充，进一步提高了多语言能力。

此外，以 Baichuan (Yang et al., 2023a)，DeepSeek (DeepSeek-AI, 2024)，GLM (Zeng et al., 2022; Du et al., 2022)，Qwen (Bai et al., 2023) 等为代表的一批国产多语言基础模型，使用以中文与英文为主的多语言数据作为预训练语料，语料规模大、质量高且多样化，在中文、英文乃至其他语言的不同任务上都表现出了较强的性能。

2.1.2 多语言指令模型

多语言指令大模型是在以英语为中心的基座大模型基础上，通过多语言指令进行微调而得到的模型。这类模型通过使用多语言指令，将模型在英文上的理解、推理、生成能力扩展到了多语言，提升了模型在多语言上的表现。例如，Chinese-LLaMA-Alpaca 项目 (Cui et al., 2023) 对 LLaMA 系列模型进行了中文的词表扩充、继续预训练与指令微调，提高了中文的相关能力；BayLing (Zhang et al., 2023b) 通过交互式翻译数据和指令数据的混合微调，提高了模型的多语言翻译能力和指令能力；Vicuna (Chiang et al., 2023) 使用了全球各地用户与 ChatGPT 的交互数据微调，使 LLaMA 模型获得了较强的多语言指令能力；m-LLaMA (Zhu et al., 2023) 是在 LLaMA 上使用混合多种不同语言的指令微调数据，其指令微调数据包括了翻译数据和跨语言通用任务数据。

2.1.3 多语言特定任务模型

多语言特定任务大模型更侧重于特定多语言领域或任务的处理，如翻译、多语言问答等。这类模型通常在特定的多语言任务上进行微调，以优化模型在该任务上的性能。Li et al. (2024) 使用翻译数据微调模型，提升了模型的翻译性能。Zhu et al. (2024)，She et al. (2024) 在数学推理任务数据集上，将非英文语言与英文语言进行了对齐，使模型在非英语上表现出与英文相似的推理过程，提升了非英语推理能力。

2.2 代表性多语言数据集

在大型语言模型的训练过程中，数据是其中非常重要的组成部分。目前许多大型语言模型在训练中英语数据占比较大，因此大模型在英语上表现较好，会导致模型更多地倾向于学习英语语言的特征和结构。同时由于缺乏多语言数据集的训练，大多数模型在多语言环境下的性能存在较大差异，在非英语语言上的能力有所欠缺。

为了解决这一问题，有许多工作提出了使用多语言数据集来提升模型多语言能力的方法。通过在预训练阶段和微调阶段引入多语言数据集，模型可以更好地理解和处理各种语言的文本数据，从而提高其在多语言环境下的性能。

2.2.1 预训练语料

维基百科（Wikipeida）是质量较高的多语言数据集，具备较多优点，首先涉及领域众多（科学，历史，艺术，政治，商业），格式工整逻辑清晰。其次包含的语言种类也比较全面（一共支持 294 种语言）。并且维基百科语料会不断更新加入实时的新知识。但是，维基百科中不同语言之间的语料占比相差极大，这会导致仅使用该语料训练的模型（如编码器 mBERT (Kenton and Toutanova, 2019) 在低资源语言上能力较低。

mC4 (Xue et al., 2021) 数据集包含 108 种语言，是 Google 团队从公开的 Common Crawl 网页中爬取，一共包含 9.7T 左右的。在构建数据集时使用了启发式的算法进行自然语言的抽

取以及去重过滤，但是 mC4 数据集仍然存在语言识别度不高，会引入噪音的问题。目前 mT5 (Xue et al., 2021) 模型使用了 mC4 数据集进行端到端训练。

ROOTs（Responsible Open-science Open-collaboration Text Source）(Laurençon et al., 2022) 是 BigScience 团队用于训练 170B 的 BLOOM (Scao et al., 2023) 多语言大模型所用到的数据集。数据集大小约为 1.6 TB，由 59 种语言组成，其中包括 46 种自然语言 (以欧洲和亚洲的语言为主)和13种编程语言。ROOTS 的多语言数据是 BigScience 团队通过使用很多工具爬取（例如 BigScience Catalogue）以及从其他语言的存储库中收集了大量的多语言单语数据。代码数据是从 Github 和 StackExchange 中抽取而来。获得大量数据后，BigScience 团队还使用了许多方法来清洗过滤数据集，例如基于困惑度，字符重复程度来过滤低质量文本，使用 SimHash(Charikar, 2002; Manku et al., 2007) 以及后缀数组 (Manber and Myers, 1993) 的方法来去除重复文本。

CulturaX (Nguyen et al., 2023a) 是由美国俄勒冈大学标注完成的。总括包括 167 种语言，6.3T 词元。在大模型时代，很多多语言大模型，如 BLOOM，PolyLM(Wei et al., 2023b)等，其训练数据集不完全公开，阻碍了开源社区对于 LLM 的分析与理解，例如在无法了解预训练数据集的情况下，开发者们很难对模型产生的幻觉以及有毒的偏见进行归因和分析。CulturaX 创立的初衷是为了提供一个更开明的多语言数据集以供训练，增加了多语言大模型的透明度，让开发者更深入分析和理解多语言大模型。CulturaX 从公开的 OSCAR (Abadji et al., 2022) 和 mC4 数据集中爬取，既保证了内容的时效性（截止至 2023 年 1 月），又保证了低资源语言的占比不至于过低。俄勒冈大学团队使用了一套全新的架构来在数据清洗中去重噪音文本，非自然语言，有毒偏见数据，并有效提升了语言识别的准确率。

2.2.2 指令对话训练数据

在多语言大模型预训练后，模型存储了大量的多语言知识，但此时模型还无法根据人类指令输出正确的回答内容。基于此，许多工作贡献了多语言的指令对话数据，旨在帮助模型获得多语言的指令遵循能力。

xp3 数据集 (Muennighoff et al., 2023) 也是由 BigScience 团队标注的数据集，主要包括 46 种语言，16 种 NLP 任务，用于训练 BLOOMZ 和 mT0。xp3 数据集在 p3 (Victor et al., 2022) 数据集的基础上新增了许多的多语言数据集，并且语言占比分布与 ROOTs 数据集趋近相同。虽然 xp3 数据集的指令数量较多，但是在某些语言还是存在噪音过多的情况。

OpenAssistant Conversation (Köpf et al., 2024) 数据集是由 OpenAssistant 人工进行标注的多语言对话数据集，质量较高，涉及 35 种语言，一共 13 万数据。OpenAssistant Conversation 是以对话树的结构组织而成，具体来说，这一个拥有多轮对话，且对于同一个问题有不同回复的数据集。

Bactrain-X (Li et al., 2023b) 数据集是由阿布扎比大学团队标注的，涉及 51 种语言，每种语言有 6 万 7 千条数据，阿布扎比大学团队首先收集了开源的英语指令数据集 Alpaca (Taori et al., 2023) 和 Dolly (Conover et al., 2023)，其次将指令和输入数据使用谷歌翻译引擎翻译成目标语言,将问题输入到 gpt-3.5-turbo 收集回答。这样组成的（指令，输入，输出）就可作为扩展语言的指令数据集。由于 gpt-3.5-turbo 自身在低资源语言上的局限性，Bactrain-X 在低资源语言上存在较大噪音。

Ayadataset (Singh et al., 2024) 是一个旨在减少语言不平等的重要数据集，通过人工筛选和多语种数据收集而成。其数据规模庞大，包含了来自全球 119 个国家的 2997 名合作者的努力共同创建的 204,114 个高质量注释，涵盖了 65 种语言。为了获取这一规模庞大的数据集，Aya 项目与来自世界各地的精通者合作，收集人工筛选的指令和完成实例。通过这种方式，能够获取到更加真实和准确的数据，而不受自动筛选和机器翻译的影响。

2.2.3 评估测试数据集

目前，衡量大型多语言模型的多语言能力是一个备受关注的重要议题。当前的研究工作已经整理并标注了多种语言的文本数据集，这些数据集涵盖了从简单任务到复杂推理任务的广泛范围。评估大型多语言模型的多语言能力通常涉及在这些数据集上执行一系列任务，如语言理解、语言生成等。通过对这些数据集进行全面评估，可以更加深入地了解大型多语言模型在多语言环境下的性能表现，从而为模型的改进和优化提供有力指导（见表 2）。

多语言理解数据集 XCOPA (Ponti et al., 2020) 是一个关于常识推理的数据集，一共覆盖 11 门语言，包括海地克里奥尔语等低资源语言，每个语言含有 600 条数据。与传统的选择题中包含 4 个选项不同，XCOPA 中，每一条数据中的问题仅对应两个选项。

XStoryCloze (Lin et al., 2022)：由 Meta 团队评测 XGLM 模型时将英文的 StoryCloze (Mostafazadeh et al., 2016) 的验证集翻译而来，一共包括 11 门语言。StoryCloze 数据集主要考察模型对于一段故事的理解能力，具体来说，它要求模型从给定的选项中选出一个故事的正确结局。

BELEBELE (Bandarkar et al., 2023):由 Meta 团队标注，是一个涵盖了 122 种语言变体的多项选择机器阅读理解（MRC）数据集，可以更好的测评模型在高资源，中资源，低资源语言上的能力。每个语言包括 900 条数据，每个数据由 1 个段落，1 个问题和四个选项组成，其中段落由 FLOERS-200 (Costa-jussà et al., 2022) 数据集组成。

XWinograd (Muennighoff et al., 2023; Tikhonov and Ryabinin, 2021)：用于评估模型的指代消解能力和常识推理能力的数据集，一共包含 7 种语言，不同语言含有的数据量差异较大。

PAWS-X (Yang et al., 2019) 是一个关于复写的多语言数据集，主要包括 7 种语言，每种语言含有 5 万条左右的数据，其中 2 万 3 千条是人工翻译，另外 2 万 9 千条是机器翻译。

XNLI (Conneau et al., 2018) 是一个关于句子理解的多语言数据集，覆盖 15 种语言。每种语言含有 40 万条数据，每条数据的结构大致如下：给定两个句子，判断这个句子的关系（蕴含，中立，矛盾）。

除上述公开发表的数据集以外，俄勒冈大学团队使用了 DeepL，gpt-3.5-turbo 等模型将很多英文数据集（例如 MMLU (Hendrycks et al., 2020)，ARC 数据集 (Clark et al., 2018)，Truthful_qa 数据集 (Lin et al., 2021)，HellaSwag 数据集 (Zellers et al., 2019)）也翻译成了很多语言，以便更好评测多语言大模型。

多语言生成数据集 XL-Sum (Hasan et al., 2021) 是一个多语言的摘要数据集，涵盖语言范围较广，一共支持 44 种语言。数据来自 BBC 的 135 万篇专业标注的文章摘要对，使用一组经过精心设计的启发式方法提取。XLSum 具有高度抽象、简洁且高质量的特点。

MLQA (Lewis et al., 2019) 是一个多语言问答数据集，覆盖了 7 门语言，MLQA中每个语言包含了5 千个左右的（其中英文有 1 万 2 千个）抽取式问答实例，不同语言之间数据是高度并行的。

MKQA (Longpre et al., 2020) 是一个开放领域的问答评估数据集，涵盖了 26 种语言的数据，每种语言包含 10,000 个问题-答案对，总共 260,000 个问题-答案对。答案基于经过精心策划的、不依赖于特定语言的段落，使得结果可以在各种语言之间进行比较。

MGSM (Cobbe et al., 2021; Shi et al., 2022) 数据集将 GSM8K 的测试集翻译成了其他 10 种语言，每个语言包含 256 个数据，是一个小学难度的数学推理数据集。

FLORES (Costa-jussà et al., 2022) 是一个多语言翻译数据集，数据来源于维基百科，收集到 2000 个句子，由专业人员从英语翻译成其他 200 种语言。

2.3 代表性多语言任务

在大模型时代，多语言处理任务迎来了重大的变革。传统上，多语言任务依赖于针对特定语言对或小范围语言族的专门模型，但随着大模型如 GPT-3、LLaMA (Touvron et al., 2023b) 等的兴起，我们开始看到一种全新的范式。这些大模型通常在多种语言上预训练，能够同时处理多种语言任务，无需特定于某一语言的调整。这种方法不仅简化了模型的部署，还提高了处理低资源语言的能力。接下来，我将分两部分详细介绍代表性的多语言任务：传统任务的大模型方法和大模型时代的新任务。

2.3.1 传统任务的大模型方法

翻译 在当今的大型模型时代，翻译技术已经从以往的 Encoder-Decoder 架构，逐步演变为仅使用 Decoder 的模型。这样的转变主要由于仅 Decoder 模型，比如 GPT 系列 (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023)，在执行语言生成任务时表现出的高效能和灵活性。不同于传统的 Encoder-Decoder 模型，仅 Decoder 模型在进行翻译时无需通过一个编码步骤来理解源语言，而是可以直接在生成步骤中将源语言文本转译为目标语言文本。例如，在执行翻译任务时，仅 Decoder 模型会接收一个用特定源语言写成的句子，并

在前面加上一个指示目标语言的特定前缀或标记，如“Translate to French:”。模型依靠其预先训练的知识来理解源语言，并直接开始生成相应的目标语言文本。

在探索大型语言模型的多语言能力时，设计翻译任务的主要目的是为了验证和增强模型处理和理解多种语言的能力。通过这些翻译任务，研究人员和开发者能够深入探讨模型如何把握并运用不同语言之间的语义和语法结构，以及如何将这些知识应用于准确地将一种语言转换为另一种语言。

多/跨语言信息抽取，摘要，阅读理解 大型语言模型如 ChatGPT 可以处理多语言或跨语言的任务，如信息抽取、摘要和阅读理解，主要是通过有效的提示词设计和模型训练方法来实现。模型的训练数据包含多种语言，使其能够理解和生成多语言内容。 在执行具体任务时，如信息抽取或摘要，设计的提示词会指导模型关注于文本的特定部分或特定类型的信息。 例如，如果需要从一篇中文文章中提取信息并生成英文摘要，可以设计一个提示词让模型首先理解和提取中文文章中的关键信息，然后将这些信息转换成英文摘要。 这一过程通常依赖于模型的多语言理解能力和生成能力，以及其在训练过程中学到的语言转换技能。

此外，通过调整提示词的具体内容和结构，可以优化模型的表现，使其更好地完成特定的多语言或跨语言任务。例如，在要求模型进行阅读理解时，可以通过明确的问题提示模型关注文章的哪一部分内容，或是用何种语言输出答案，以此来提高信息处理的准确性和效率。

命名命名实体识别 在大型语言模型中，完成命名实体识别（Named Entity Recognition, NER）主要依赖于模型训练和任务特定的提示词设计。 具体来说，通过精心设计的提示词和适当的少样本学习方法，可以有效地实现 NER 任务。

首先，在提示词设计方面，需要构造一个清晰的任务说明，使模型明白所需执行的具体任务是识别文本中的命名实体。 例如，可以设计一个提示词：“请识别以下文本中的所有人名、地名和组织名称。” 紧接着，提供一段文本，让模型进行实体识别。 其次，使用少样本学习 (Wang et al., 2020c) 不仅提高 NER 任务的效果，同时可以控制模型的输出格式，方便提取出模型输出中的命名实体标签。

情感分析，文本分类 在这种情况下，模型利用其预训练期间学到的知识来直接对文本进行分类，无需任何特定任务的训练。 例如，要进行情感分析，可以给模型提供一个文本并询问：“这段文本的情感是正面的、负面的还是中性的？”模型会根据其预训练中学到的知识来预测答案。 虽然 Decoder-only 的大模型在没有额外训练数据的情况下能够直接进行文本分类和情感分析，但在某些情况下，通过在特定任务的数据上进一步训练（微调）模型可以获得更加可控的输出结果。 对于分类问题而言，也可以通过限制模型只能输出词表中几种特定单词的方式，获得可控输出。

2.3.2 大模型时代的新任务

多语言上下文学习 大模型如 GPT-4 通过上下文学习 (Dong et al., 2022)，在没有显式训练的情况下，仅通过提示指导即可适应新任务。 这种方法特别适用于处理多语言数据，模型可以通过观察少量的例子迅速调整其行为。

多语言指令服从 新一代的大模型如 Codex 和 GPT-4 在多语言环境中不仅可以处理文本任务，还能理解并执行包括代码编写 (Chen et al., 2021)、数学问题解答 (Cobbe et al., 2021) 等复杂任务。 这显示了大模型在处理跨领域、跨语言问题时的强大能力。

总的来说，大模型时代为多语言任务带来了前所未有的机遇和挑战。通过预训练的大型模型，我们能够以前所未有的规模和效率处理多语言信息，但这也对模型的泛化能力和可解释性提出了新的要求。在设计未来的多语言处理系统时，我们需要考虑这些因素，以充分利用大模型的潜力，同时克服其局限性。

3 前沿研究方向

本部分将重点介绍目前多语言大模型研究的若干前沿方向，包括模型训练、跨语言对齐、语言能力扩展，以及多语言大模型内部机制的可解释性分析。图 1 展示了本文所述的多语言研究的不同角度概况。

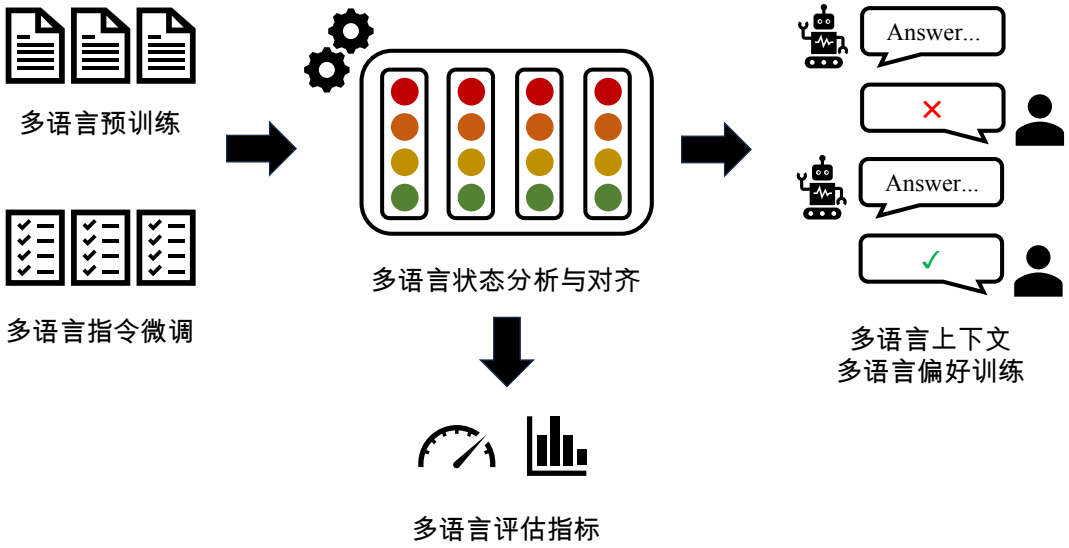

Figure 1: 不同角度的多语言研究示意图

3.1 多语言模型训练

为了得到具有更好多语言能力的大模型，现有的工作从预训练，指令微调以及强化学习偏好优化三个阶段进行。

3.1.1 预训练

在预训练阶段，随机初始化的大模型将在大规模的语料库上做无监督训练，获取知识和语言能力。一个很直接的想法是在预训练语料中加入多语言数据。现有的大模型的预训练语料库中往往包含了一部分多语言数据，如Qwen (Bai et al., 2023)，Mistral (Jiang et al., 2023)，mT5 (Xue et al., 2021)，Erine (Sun et al., 2021)，BLOOM (Scao et al., 2023),LLama1 (Touvron et al., 2023a)，LLama2 (Touvron et al., 2023b)，PolyLM (Wei et al., 2023b)，InternLM (Team, 2023) ，具备一定的多语言能力基础。

尽管在预训练阶段加入多语言的预训练语料有助于获得多语言能力，然而扩充语言数量削弱了每一个语言以及每一个任务的有效信息容量，语种之间竞争有限的模型容量 (Conneau et al., 2020)。因此在预训练阶段，一个需要考虑的问题是各个语言数据的配比问题。多语言的预训练语言模型通常采用基于启发式的温度采样方法来平衡不同语言之间的权重 (Devlin et al., 2019)，对一些低资源语言进行上采样。然而这样的方法容易造成低资源语言的过拟合，也降低了学习效率 (Hernandez et al., 2022; Lee et al., 2021)。一些工作展开了进一步探索，Wang et al. (2020b) 在 DDS (Wang et al., 2020a) 的基础上进行了扩展和改进，提出了 MultiDDS 算法，通过学习一个语言评分器，以优化多语言数据的使用，从而在多种不同语言上实现良好的性能。 Chung et al. (2023) 提出了 UNIMAX，在保证头部语言得到更均匀覆盖的同时，通过明确限制每种语言语料库的重复次数来减轻尾部语言的过拟合问题。

从头开始训练大模型的代价往往相对高昂，也没有办法很好的利用已经训练好的大模型。因此基于已有大模型的继续预训练就成了一个很好的选择。Tang et al. (2020) 展示了一种从已有预训练语言模型如mBART开始继续训练的方法，在拓展语言的同时，很好的保留了已有语言上能力，成功的将原始mBART支持的25种语言拓展到了50种。除了支持语种的大幅度扩充之外，也有一些工作关注于特定一些语言的能力提升，基于已有的英语为中心的大模型，在对应的相关语言的预训练数据上进行了大量的预训练，获得了一定的能力提升，如CPM-2 (Zhang et al., 2021)，Chinese-Llama (Cui et al., 2023)，Chinese-Mixtral (HIT-SCIR, 2024)，SeaLLM (Nguyen et al., 2023b)，Sailor (Dou et al., 2024)等等。

3.1.2 指令微调

在指令微调阶段，模型会在用指令或者模板形式构造好的数据上训练，可以很好的激发模型在预训练阶段习得的知识。因此，通过添加多语言的指令微调能够使模型具备一定程度的多语言指令响应能力。有一些工作 (Muennighoff et al., 2023; Cui et al., 2023; HIT-SCIR, 2024;

Nguyen et al., 2023b; Dou et al., 2024; Wei et al., 2023b; Csaki et al., 2024; Chen et al., 2023a) 收集了来自不同语言的指令微调数据，用于训练模型。这部分的训练数据有综合的指令响应任务、也包括了具体的如数学推理，翻译等任务。还有一些工作 (Muennighoff et al., 2023; Xue et al., 2021) 通过将原有的多语言多任务数据，通过指令模板的形式转化为多语言指令微调数据。相对而言多语言的指令微调数据相对稀缺，因此 BayLing (Zhang et al., 2023b) 通过交互式翻译任务将英语语言生成和指令遵循能力转移到非英语语言的指令遵循型大型语言模型，在多项评估中展现了不错的效果。

3.1.3 人类偏好对齐

RLHF 阶段往往用于调整模型的行为，让模型更好的和人类偏好对齐。这个阶段需要有部分人工标注的偏好数据，通常情况下体现为对于一个指令，利用策略模型采样输出，并人工对这些采样结果进行偏好打分。当前广泛使用的算法有 PPO (Schulman et al., 2017)、DPO (Rafailov et al., 2023)、IPO (Liu et al., 2019b)、CPO (Achiam et al., 2017) 等等。现有工作 (Yang et al., 2023a; Bai et al., 2023; Cui et al., 2023; Du et al., 2022; Wei et al., 2023b; Team, 2023; Chen et al., 2023a) 为了更好的将模型在各个语言上和人类偏好对齐，在这个阶段使用多语言的偏好标注数据来训练。上述的偏好优化算法往往需要基于一个经过多语言指令微调的大模型作为初始化的策略模型，后续提出的 ORPO (Hong et al., 2024) 偏好优化算法可以直接从预训练模型开始训练，节约计算资源消耗的同时也取得了很好的效果。

3.2 跨语言对齐

跨语言对齐，也可以称为跨语言一致性、跨语言迁移等，指的是通过一定技术手段，使得多语言大模型在不同语言上的知识、能力、行为等方面保持一致 (Wang et al., 2023a; Qi et al., 2023; Gao et al., 2024a; Qin et al., 2024)。此类研究的出发点主要有三：1）使大模型可以在不同语言上表现出更平衡的性能，服务不同语言背景的用户 (Wang et al., 2023a)；2）使模型能够复用在不同语言的数据上学习到的知识和偏好，提高训练效率 (Gao et al., 2024a)；3）大多数事实知识、指令服从能力、安全性偏好都是与语言种类无关的，因此模型在这些方面理应具有跨语言对齐的表现 (Ohmer et al., 2023; Wu et al., 2024)。目前，大语言模型的能力日益提高，但它们的跨语言对齐程度却并不令人满意。因此，此方面的研究在将来一段时间内也具有重要意义。 本部分将首先介绍目前对大语言模型跨语言对齐程度的评估工作；随后将根据观察角度与实现方式的不同，分别介绍训练数据、内部状态、模型偏好、指令上下文四个层面的跨语言对齐研究。

3.2.1 对大模型跨语言对齐程度的评估

现有工作对大模型的跨语言对齐程度主要从通用任务能力与事实知识两方面进行了评估。

通用任务能力方面，Lai et al. (2023a) 评估了 ChatGPT 在 37 种语言上执行 7 种代表性 NLP 任务的性能，并与 mT5-XXL (Xue et al., 2021) 等多语言预训练模型对比，发现 ChatGPT 服从非英语指令、完成非英语任务的能力低于英语，且与有监督学习基线的差距较大。Wang et al. (2023a) 提出了 SeaEval 基准测试，用于测量多语言大模型在英语、中文、印地语、越南语等语言上完成包括经典 NLP 任务、复杂推理以及文化理解在内的一系列任务的表现，发现目前的多语言大模型在不同语言的事实、科学和常识性知识等方面表现出了较大的不一致，未达到“平衡的多语言能力”。Zhang et al. (2023a) 考察了部分多语言大模型在语码切换（code-switching）场景下的情感分析、机器翻译、摘要和词级别语种识别任务，发现大模型在这些任务上的性能远不如专门微调过的小模型，提示大模型的多语言能力不能直接泛化到语码切换场景。

事实知识方面，Qi et al. (2023) 针对事实知识的跨语言一致性，提出了 RankC 指标，用于评估 BLOOM, mT5, XLM-RoBERTa (Liu et al., 2019a) 等模型在 17 种语言上的事实知识一致性，发现模型参数增加虽然能使事实准确率更高，但无法提高跨语言一致性；同时，用英语对模型进行知识编辑时，新知识只能迁移到与英语有较高 RankC 分数的语言，说明模型的跨语言事实知识一致性水平较为有限。Wang et al. (2023c) 同样也关注了知识编辑中的跨语言效应，通过构造英语-中文的跨语言知识，评估了英语知识编辑能否对中文对应知识产生影响，发现此种跨语言影响十分有限。Gao et al. (2024a) 则提出了 CLiKA 评测框架，将模型的多语言知识对齐分为性能、一致性、传导性三个层面，评估了现有多语言大模型在 10 种语言上的跨

语言知识对齐程度，并比较了多语言预训练、多语言指令微调对对齐程度的影响，发现当前模型的跨语言对齐程度不够理想，且其跨语言一致性可能主要来源于训练数据的重合，而非语言间的知识传导；同时，尽管多语言混合预训练和指令微调能带来更好的跨语言性能平衡和一致性，它们并不能提高模型的跨语言知识传导程度。

3.2.2 训练数据层面的跨语言对齐

促进模型跨语言对齐能力的常见方法，是在训练数据中加入多语言内容，包括平行语料、非平行语料、多语言指令等，本文将其称为训练数据层面的跨语言对齐。

在大模型时代以前，已经出现了在多语言预训练模型的训练数据中加入多语言内容的尝试。Conneau and Lample (2019) 提出了分别使用多语言非平行语料、平行语料进行预训练的方法，得到了 XLM 系列模型，并在跨语言 NLP 任务、机器翻译等领域上取得了提升。NLLB 团队 (2022) 收集了包含 200 种语言的翻译数据，得到了 NLLB 系列大规模多语言翻译模型，取得了领先的翻译性能。Vu et al. (2022) 则基于 mT5 模型，使用统一的 prompt tuning 框架 (Lester et al., 2021) 在多种语言上进行混合的有监督与无监督微调，在跨语言生成任务上取得了较大提升；还提出了将提示词各部分分离的 factorized prompt 格式，缓解灾难性遗忘。同时，以英语数据为主训练的模型也展现出了一定的多语言能力，Blevins and Zettlemoyer (2022) 研究了这一现象，并提出这可能是由于英语预训练语料中含有比例极低，但绝对数量较多的非英语数据，这些数据的含量与模型的多语言能力密切相关。Briakou et al. (2023) 针对英语预训练模型表现出的零样本或少样本机器翻译能力，调查了 PaLM (Chowdhery et al., 2022) 模型的训练数据，发现其中含有 3000 万以上、涵盖超过 44 种语言的翻译数据对，称为“偶然双语现象”，并认为这些数据与模型的翻译能力相关。这些研究说明了在预训练过程中加入多语言数据，对提升模型的多语言对齐程度较为重要。

进入大模型时代后，上述研究的思路得到了延续。一些大模型在预训练阶段加入了大量多语言非平行数据或翻译数据，与英语数据混合预训练 (Anil et al., 2023; Bai et al., 2023; Schioppa et al., 2023; Touvron et al., 2023b; Wei et al., 2023b; Scao et al., 2023; Yang et al., 2023a; Yang et al., 2023b)；也有一些模型以英语为主的基座模型为基础，在新语言上继续预训练 (Cui et al., 2023; Larcher et al., 2023)；还有一部分大模型在微调阶段加入多语言任务指令数据，包括对话数据、推理数据、翻译数据等，提高模型的多语言任务执行能力 (Cahyawijaya et al., 2023; Chen et al., 2023b; Kew et al., 2023; Muennighoff et al., 2023; Wei et al., 2023b; Scao et al., 2023; Yang et al., 2023a; Yang et al., 2023b; Zhang et al., 2023b; Jiao et al., 2023; Chen et al., 2024; Chung et al., 2024; Gao et al., 2024b; Li et al., 2024; Zhu et al., 2024; Lai et al., 2023b)。

3.2.3 内部状态层面的跨语言对齐

数据层面的方法是通过输出阶段的监督信号来实现跨语言对齐，但这种方法的粒度较粗，无法在更细的计算过程视角观察跨语言对齐。因此，研究者尝试了构建模型的内部状态对不同语言的相同语义数据的对齐。内部状态层面的跨语言对齐研究起源于无监督翻译领域，后者的核心思想是自动构建不同语言的语义嵌入空间之间的同构映射 (Lample et al., 2018; Søgaard et al., 2018; Cao et al., 2019; Naorem et al., 2024; Shen et al., 2024)，或自动推断双语词典 (Yu et al., 2023b; Ding et al., 2024; Garnier and Guinet, 2024)。

应用到大模型中，有研究使用对比学习等方法，拉近模型在不同语言中的同义上下文表示，从而提高模型的整体多语言能力 (Efimov et al., 2023; Li et al., 2023a; Tao et al., 2023)，或构建更好的多语言嵌入向量 (Zhang et al., 2023c; Zhao and Eger, 2023)。具体到任务中，Wang et al. (2023b) 通过实体表示的对齐，提高模型对知识图谱的跨语言学习效率；Xu et al. (2023b) 利用句法和语义的联合学习，构建适用于文本分类任务的跨语言嵌入模型；Chen et al. (2023b) 引入了一个全局指令表示向量，用于提高模型的翻译忠实度。

在用于提升模型跨语言对齐程度的同时，表示层面的对齐观察也为模型运行机制的可解释性提供了新的视角。Zhao et al. (2023) 通过类比神经科学中的脑区定位方法，在多语言大模型中找到了与多语言能力相关的少量参数，称之为语言核心区。Wendler et al. (2024) 通过精心构造的翻译任务场景，观察了 LLaMA-2 系列模型在处理跨语言数据时的内部状态变化，观察到了其中存在近似正交的“输入空间”、“概念空间”、“输出空间”的证据，且“概念空间”与英语较其他语言更为接近。Zhao et al. (2024) 则提出了平行语言特定神经元探测（PLND）方法，用

于衡量模型的不同神经元在处理多语言输入时的重要性，并认为模型在处理多语言数据时，可能在内部状态中先将输入翻译为英语，进行处理后再翻译回原输入语言。这些研究为进一步提高内部状态层面的跨语言对齐提供了理论支持。

3.2.4 模型偏好与上下文层面的跨语言对齐

除了利用数据层面的任务监督信号，内部状态层面的表示对比信号，还可以通过控制模型输出时的行为偏好与指令上下文，促进其跨语言对齐能力。

行为偏好层面，指的是通过强化学习方法，使模型倾向于给出跨语言一致的回答。这种方法可以追溯到使用强化学习和贝叶斯风险最小化方法优化的神经机器翻译模型 (Ramos et al., 2023; Yang et al., 2024)。应用到大模型的跨语言对齐方面，She et al. (2024) 将跨语言对齐作为一种受鼓励的倾向，提出了 MAPO 框架，针对数学推理场景，以强化学习方法促进模型在英语和非英语场景下给出相同的推理过程的倾向，使模型在相应的评测数据集上展现出了显著性能提升和更高的跨语言推理一致性。Wu et al. (2024) 提出了人类反馈强化学习（RLHF）阶段中的奖励模型的零样本跨语言迁移方法，将用英语数据训练的奖励模型迁移到非英语，从而促进模型行为倾向、安全性等方面的跨语言对齐。

指令上下文层面，指的是不改变模型参数，仅通过设计合适的上下文样例或提示词来提高模型的跨语言对齐程度。Huang et al. (2023) 提出了跨语言思维提示（XLT）方法，让模型先将问题翻译为英语，将其形式化并分布解答，激发模型的跨语言逻辑推理能力，发现可以显著提升模型对各种多语言任务的性能，并缩小英语与非英语的性能差距。Kim et al. (2023) 针对多语言 QA 任务，提出了 In-CLT 方法，在大模型的上下文学习样例中同时提供源语言篇章与目标语言问题，发现能够显著提高模型在目标语言上的 QA 任务性能。Puduppully et al. (2023) 提出了 DecoMT 方法，通过少样本提示，让模型将句子级别翻译任务分解为多个词组级别任务，提高了模型的翻译能力。Yong et al. (2023) 通过让模型模拟双语使用者的提示，鼓励模型生成英语-东南亚语言的高质量语码切换输出。

3.3 语言能力扩展

目前，大型语言模型如 LLaMA 和 Mistral 主要以英语为核心，在处理其他语言时性能明显下降。鉴于此，学术界展开了广泛的研究，旨在扩展和增强这些模型的多语言能力。这些研究可以被概括为两个主要方向：一是单一模型的多语言扩展，二是多模型融合以共同增强多语言处理能力。

在单一模型的多语言扩展方面，研究人员致力于探索如何使单个大模型在处理多种语言时保持高效性能。这包括但不限于多语言继续预训练和跨语言微调等技术。通过这些方法，研究人员希望实现在保持模型在英语上表现优异的同时，提高其在其他语言上的适用性和性能。在涉及到继续预训练的工作中，BigTranslate (Yang et al., 2023b)，Tower (Alves et al., 2024)，ALMA (Xu et al., 2023a) 工作都说明了继续使用单语训练大模型会极大提升模型的多语言能力 （翻译能力），其中ALMA指出，在给定算力的条件下，将更多的算力分配给单语而不是翻译数据会更好的提升模型的翻译能力，在这种实验设置下就可以让大模型的翻译能力与传统的监督模型相接近。在涉及到指令微调的技术中，Bayling (Zhang et al., 2023b)，xllama (Zhu et al., 2023)，SDRRL (Zhang et al., 2024) 都使用了多语言指令微调数据集来提升大模型在下游多语言任务上的性能，其中 xllama，SDRRL 都指出，在指令微调时混入大量的翻译数据会更好帮助模型把存储在英语上的知识迁移到目标语言中。

另一方面，多模型融合的研究聚焦于如何结合多个语言模型的知识和能力，以增强整体的多语言处理能力。这涉及到模型融合技术、集成学习方法等方面的研究。通过将不同语言模型的优势互补结合，研究人员期望实现在跨多语言环境下更为有效的文本处理能力。在模型融合方法借助多模态 LLaVA 架构，LangBridge (Yoon et al., 2024)指出，将含有多语言知识的小模型 （例如 mT5）与大模型 MetaMath (Yu et al., 2023a) （推理能力较强，但多语言能力较差）使用一个简单的线性层相融合，在仅使用英语数据下就可以极大提升大模型多语言的推理能力。同时，Bansal et al. (2024) 也尝试将较低资源语言的知识存储到较小的 PaLM 模型中，将较小的，存储较多多资源语言知识的 PaLM 模型与较大的 PaLM 模型使用 cross-attention 的方式去完成模型融合，增强了较大 PaLM 模型在低资源语言上的翻译能力。还有工作 (Blevins et al., 2024) 尝试使用 BTM (Branch Traiin Merge）的方法，对于同一个模型，不同的语言分别独立训练成多个模型后再融合，这种方法在一定程度突破了“多语言诅咒” (Wu and Dredze,

2020)。在集成学习方面，Farinhas et al. (2023) 在大模型完成多语言翻译任务时，采用了不同的生成方法采样了多个结果 （greedy search，beam search）,最终采取集成的方式选中最终候选回答，这样使大模型的翻译能力得到了巨大的长进。

3.4 多语言机制分析

目前的大语言模型普遍具有较强的多语言能力，而大语言模型内部是如何对不同语言进行处理，是一个值得探索的问题。 目前，有许多工作尝试对大语言模型内部处理多语言的机制进行分析。

其中一部分工作从模型生成下一词元时中间隐层状态的语义表示空间出发，将模型的前向计算过程按照不同层分为多个阶段，发现了在不同阶段之间存在语言转换的现象，并在其中部分阶段发现了不同语言间存在对齐的现象。 Wendler et al. (2024) 等人使用 Logit Lens 方法 (nostalgebraist, 2020) 发现，在训练语料主要是英语的大语言模型中（如 LLaMA），即使大语言模型在输出非英语的文本，模型也会在中间层先生成对应于英语词元的隐层状态，再在高层转换成对应输出语言的词元。 具体来说，其认为大语言模型的前向计算可以从低到高分为三个阶段：上下文信息的理解、抽象概念的生成以及从抽象概念到输出词元的转换。特别的，在以英语为中心的大语言模型中，可以观察到在第二阶段中生成的抽象概念的隐层状态正对应于输出词元的英文表示。 类似的，Zhao et al. (2024) 等人提出了 PLND 方法来检测大语言模型中特定于某种语言的神经元。他们发现，在模型的低层和高层中与非英语语言相关的神经元更多，而在中层与英语相关的神经元更多，由此提出了类似的“多语——英语——多语”三阶段模型。他们认为模型在最底层和最高层进行的理解和生成阶段是语言相关的，而在中层进行解决问题时会使用英语的思考能力以及不同语言的事实性知识。 由于大语言模型在处理不同的语言时，都可能会在中间阶段归并到某种语言无关的概念表示，再通过第三阶段进行语言相关的目标词元生成，因此某些低资源语言也能通过与高资源语言（如英语）共享相似的中间表示的方式，共享模型从其他语言学到的通用能力。

另外一部分工作则从模型的计算部件——神经元的角度出发进行分析研究。 Tang et al. (2024) 等人的工作中提出可以将大语言模型中的神经元分为通用神经元与语言特定的神经元，通用神经元在生成各种不同语言时都会被激活，而语言特定神经元在生成特定语言时才会被激活。进一步，他们提出了 LAPE 方法来检测模型中的语言特定神经元。由此，他们发现大语言模型对某种特定语言的处理大部分都是来源于模型顶部和底部的一小部分神经元，并且发现可以通过选择性地干预这些语言特定神经元的激活与抑制，达到控制大语言模型的输出语言的目的。

4 研究面临的挑战

本部分将介绍目前的多语言大模型研究面临的几项重要挑战。

4.1 多语言场景下的公平性与安全性

4.1.1 多语言场景下的公平性

虽然多语言技术的发展使得大模型在非英文上的能力逐渐增强，但模型在不同语言上的差异很难被完全消除，这引发了模型在多语言的公平性问题 (Shliazhko et al., 2022)。由于低资源语言的训练语料稀缺且质量较差 (Yu et al., 2022)，提升大模型在低资源语言上的能力十分困难。除了不同语言性能上的公平性，一些研究 (Levy et al., 2023; Piqueras and Søgaard, 2022) 还揭示了不同语言间的知识、种族、宗教、性别存在着差异。此外，模型在不同语言上的词元占比不同，不同语言词元的切分长度也存在较大差异，这造成了不同语言的计算开销的差异 (Petrov et al., 2024; Ahia et al., 2023)。

4.1.2 多语言场景下的安全性

大型语言模型在理解和生成自然语言方面表现出卓越的能力。然而，这种能力也引发了对潜在安全性问题的担忧，包括生成不安全的内容，例如侮辱或歧视性语言 (Sun et al., 2023) 或泄露私人信息 (Macko et al., 2023)。在多语言场景下，安全性问题更为突出。一方面，多语言数据的清洗和筛选相比于英文数据更加困难，当前大部分多语言数据集的清洗都是不充足的 (Nguyen et al., 2023a)。另一方面，缺少多语言安全性评测基准。Wang et al(Wang et al., 2023d) 构造了覆盖十种不同语言的安全性评测数据集 XSAFETY，是目前唯一的多语言安全性

评测数据集。但是，XSAFETY 覆盖的语言数量有限，构建一个更加全面的多语言安全性评估数据集，仍然是多语言安全性研究的迫切需求。

4.2 多语言诅咒问题

多语言诅咒问题最早由 Conneau et al. (2020) 在多语言预训练模型的研究中提出，具体定义为：当模型的容量固定时，如果不断增加多语言训练数据，模型的多语言能力（尤其是低资源语言能力）会先上升，但在到达一定程度后，几乎所有语言的单语言、跨语言能力都会下降。这种现象提示了模型的多语言能力提升有一定的边界，即多语言能力不能超过模型参数规模、基础语言能力等决定的上限。此后，许多工作都试图突破这一限制，例如使用混合专家模型，为不同语言分配不同的专家 (Blevins et al., 2024; Pfeiffer et al., 2022)。这些工作在缓解多语言诅咒的方面取得了一定效果，但同时也造成了新的问题。例如：当语言种类不断增加时，专家数量需要不断增加，导致模型参数量过大；语言专家中存在重复、冗余参数，并且分专家存储参数的模式对跨语言对齐也产生了挑战。这些问题都需要在未来给予更大的关注。

4.3 多语言训练的代价

在大模型时代，多语言训练的代价是一个非常重要的议题。多语言模型的训练需要大量的多语言数据 (Kaplan et al., 2020)。这些数据不仅要覆盖多种语言，还需要在各种语言之间保持高质量和平衡性。数据的采集包括从不同地区和文化背景收集文本，这通常涉及到昂贵的数据收集开销。此外为了确保数据的多样性和无害性 (Askell et al., 2021; Perez et al., 2022)，可能需要人工干预来纠正数据问题，那么还要付出额外的人工标注费用。另外，多语言模型通常需要比单一语言模型更大的参数空间和更复杂的网络结构 (Qin et al., 2024)，以便在多种语言之间进行知识共享和转移。这导致了更高的计算成本，包括但不限于 GPU、TPU 等高性能计算资源的使用。此外，这些资源的使用通常需要较长的时间，增加了能源消耗和相应的环境影响 (Scao et al., 2023)。

5 总结

多语言大模型是多语言研究领域在未来一段时间内的重点研究对象。本文从领域情况概述、前沿研究方向、研究面临的挑战三个方面介绍了大模型时代的多语言研究进展，提供了使用、训练、改进和分析多语言大模型所需的参考研究信息。我们希望这项工作能为多语言大模型的将来研究提供可能的帮助。

参考文献

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France, June. European Language Resources Association.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *arXiv preprint arXiv:2305.13707*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick,

Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report, September.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report, September.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. Llm augmented llms: Expanding capabilities through composition. *arXiv preprint arXiv:2401.02412*.

Terra Blevins and Luke Zettlemoyer. 2022. Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv preprint arXiv:2401.10440*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada, July. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-Low Resource Language Alignment via Continual Crosslingual Instruction Tuning, October.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*, September.

Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023a. Tigerbot: An open multilingual multitask llm.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023b. Improving Translation Faithfulness of Large Language Models via Augmenting Instructions, August.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics, March.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways, October.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.

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 preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic 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, August.

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca, June.

DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model, May.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Qiuyu Ding, Hailong Cao, and Tiejun Zhao. 2024. Enhancing Bilingual Lexicon Induction via Bi-directional Translation Pair Retrieving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17898–17906, March.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May. Association for Computational Linguistics.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2023. The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 51–67, Cham. Springer Nature Switzerland.

António Farinhas, José GC de Souza, and André FT Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. *arXiv preprint arXiv:2310.11430*.

Tingchen Fu, Lemao Liu, Deng Cai, Guoping Huang, Shuming Shi, and Rui Yan. 2023. The Reasonableness Behind Unreasonable Translation Capability of Large Language Model. In *The Twelfth International Conference on Learning Representations*, October.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024a. Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly, April.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024b. Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models, February.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation, February.

Paul Garnier and Gauthier Guinet. 2024. Semi-Supervised Learning for Bilingual Lexicon Induction, February.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*.

HIT-SCIR. 2024. Chinese-mixtral-8x7b: An open-source mixture-of-experts llm. https://github.com/HIT-SCIR/Chinese-Mixtral-8x7B.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting, May.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during Chat using Large Language Models tuned with Human Translation and Feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore, December. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed?, December.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2023. Boosting Cross-lingual Transferability in Multilingual Models via In-Context Learning, May.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, April.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback, August.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*, February.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: Closing the gap for foreign languages, August.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. *arXiv preprint arXiv:2305.11242*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023a. Align after Pre-train: Improving Multilingual Generative Models with Cross-lingual Alignment, November.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions, April.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach.

Yongshuai Liu, Jiaxin Ding, and Xin Liu. 2019b. Ipo: Interior-point policy optimization under constraints.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. *arXiv preprint arXiv:2310.13606*.

Udi Manber and Gene Myers. 1993. Suffix arrays: a new method for on-line string searches. *siam Journal on Computing*, 22(5):935–948.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July. Association for Computational Linguistics.

Deepen Naorem, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024. Improving linear orthogonal mapping based cross-lingual representation using ridge regression and graph centrality. *Computer Speech & Language*, 87:101640, August.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023a. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023b. Seallms – large language models for southeast asia.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. *CoRR, abs/2305.11662*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July. Association for Computational Linguistics.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained multilingual models equally fair across languages? *arXiv preprint arXiv:2210.05457*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed Prompting for Machine Translation Between Related Languages using Large Language Models, October.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models, October.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers, April.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André F. T. Martins. 2023. Aligning Neural Machine Translation Models: Human Feedback in Training and Inference, November.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina

Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, June.

Andrea Schioppa, Xavier Garcia, and Orhan Firat. 2023. Cross-Lingual Supervision improves Large Language Models Pre-training, May.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization, April.

Yingli Shen, Wei Bao, Ge Gao, Maoke Zhou, and Xiaobing Zhao. 2024. Unsupervised multilingual machine translation with pretrained cross-lingual encoders. *Knowledge-Based Systems*, 284:111304, January.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models.

Qian Tao, Zhihao Xiong, Bocheng Han, Xiaoyang Fan, and Lusi Li. 2023. A Novel Unsupervised Approach for Cross-Lingual Word Alignment in Low Isomorphic Embedding Spaces. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3027–3041.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-Thought Reasoning Without Prompting, February.

Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, and Graham Neubig. 2020a. Optimizing data usage via differentiable rewards. In *International Conference on Machine Learning*, pages 9983–9995. PMLR.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020b. Balancing training for multilingual neural machine translation. *arXiv preprint arXiv:2004.06748*.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020c. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023a. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning, September.

Chenxu Wang, Zhenhao Huang, Yue Wan, Junyu Wei, Junzhou Zhao, and Pinghui Wang. 2023b. FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs. *Information Fusion*, 89:41–52, January.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023c. Cross-Lingual Knowledge Editing in Large Language Models, September.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023d. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023b. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment, April.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Yuemei Xu, Wanze Du, and Ling Hu. 2023b. A Cross-lingual Sentiment Embedding Model with Semantic and Sentiment Joint Learning. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 82–94, Cham. Springer Nature Switzerland.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias, April.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open Large-scale Language Models, September.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023b. BigTranslate: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages, July.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct Preference Optimization for Neural Machine Translation with Minimum Bayes Risk Decoding, April.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages. In Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali, editors, *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore, December. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. *arXiv preprint arXiv:2401.10695*.

Xinyan Velocity Yu, Akari Asai, Trina Chatterjee, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: a survey of multilingual dataset construction and necessary resources. *arXiv preprint arXiv:2211.15649*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023a. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Shenglong Yu, Wenya Guo, Ying Zhang, and Xiaojie Yuan. 2023b. CD-BLI: Confidence-Based Dual Refinement for Unsupervised Bilingual Lexicon Induction. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 379–391, Cham. Springer Nature Switzerland.

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*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*, September.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021. Cpm-2: Large-scale cost-effective pre-trained language models.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023a. Multilingual Large Language Models Are Not (Yet) Code-Switchers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore, December. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023b. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models, June.

Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. 2023c. VECO 2.0: Cross-lingual Language Model Pre-training with Multi-granularity Contrastive Learning, April.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. *arXiv preprint arXiv:2402.12204*.

Wei Zhao and Steffen Eger. 2023. Constrained Density Matching and Modeling for Cross-lingual Alignment of Contextualized Representations. In *Proceedings of The 14th Asian Conference on Machine Learning*, pages 1245–1260. PMLR, April.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling A Core Linguistic Region in Large Language Models, October.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism?

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question Translation Training for Better Multilingual Reasoning, February.