Patterns

CellPress
OPEN ACCESS

Review

A survey of multilingual large language models

Libo Qin,[1,7,*] Qiguang Chen,[2,7] Yuhang Zhou,[2] Zhi Chen,[3] Yinghui Li,[4] Lizi Liao,[5] Min Li,[1] Wanxiang Che,[2] and Philip S. Yu[6]
[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China
[3]ByteDance, Inc., Shanghai 200082, China
[4]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[5]School of Computing and Information Systems, Singapore Management University, Singapore 188065, Singapore
[6]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60637, USA
[7]These authors contributed equally
*Correspondence: lbqin@csu.edu.cn
https://doi.org/10.1016/j.patter.2024.101118

THE BIGGER PICTURE The rapid advancement of large language models (LLMs) has significantly transformed natural language processing (NLP), enabling machines to understand and generate human-like text. However, most LLMs are predominantly English centric, limiting their applicability in our linguistically diverse world. With over 7,000 languages spoken globally, there is a pressing need for models that can comprehend and generate text across multiple languages. Multilingual large language models (MLLMs) address this gap by processing and producing content in various languages, thereby enhancing global communication and accessibility. This survey provides a comprehensive overview of MLLMs, introducing a systematic taxonomy based on alignment strategies to deepen understanding in this field. By highlighting emerging trends and challenges, this survey aims to guide future research and development, fostering the creation of more inclusive and effective language models that cater to the diverse linguistic landscape of our world.

SUMMARY

Multilingual large language models (MLLMs) leverage advanced large language models to process and respond to queries across multiple languages, achieving significant success in polyglot tasks. Despite these breakthroughs, a comprehensive survey summarizing existing approaches and recent developments remains absent. To this end, this paper presents a unified and thorough review of the field, highlighting recent progress and emerging trends in MLLM research. The contributions of this paper are as follows. (1) Extensive survey: to our knowledge, this is the pioneering thorough review of multilingual alignment in MLLMs. (2) Unified taxonomy: we provide a unified framework to summarize the current progress in MLLMs. (3) Emerging frontiers: key emerging frontiers are identified, alongside a discussion of associated challenges. (4) Abundant resources: we collect abundant open-source resources, including relevant papers, data corpora, and leaderboards. We hope our work can provide the community quick access and spur breakthrough research in MLLMs.

INTRODUCTION

In recent years, remarkable progress has been witnessed in large language models (LLMs),[1–4] which have achieved excellent performance in various natural language processing tasks.[5–7] In addition, LLMs raise surprising emergent capabilities, including in-context learning,[8,9] chain-of-thought reasoning,[10–12] and even planning.[13,14] Nevertheless, the majority of LLMs are English centric, primarily focusing on English tasks,[15,16] which renders them relatively weak in multilingual settings, especially in low-resource scenarios.

Actually, there are over 7,000 languages in the world. With the acceleration of globalization, the success of LLMs should be leveraged to serve diverse countries and languages. To this end, as shown in Figure 1, multilingual large language models (MLLMs) possess the advantage of comprehensively handling multiple languages, gaining increasing attention. Specifically, existing MLLMs can be broadly divided into two groups based on different stages. The first series of works[17–20] leverages multilingual data to tune parameters and boost the overall multilingual performance. The second series of works[11,12,21] also adapts advanced prompting strategies to unlock the deeper multilingual potential of MLLMs during the parameter-frozen inference stage.

While remarkable success has been achieved in MLLMs, there remains a lack of comprehensive review and analysis of recent efforts in the literature, which hinders the development of

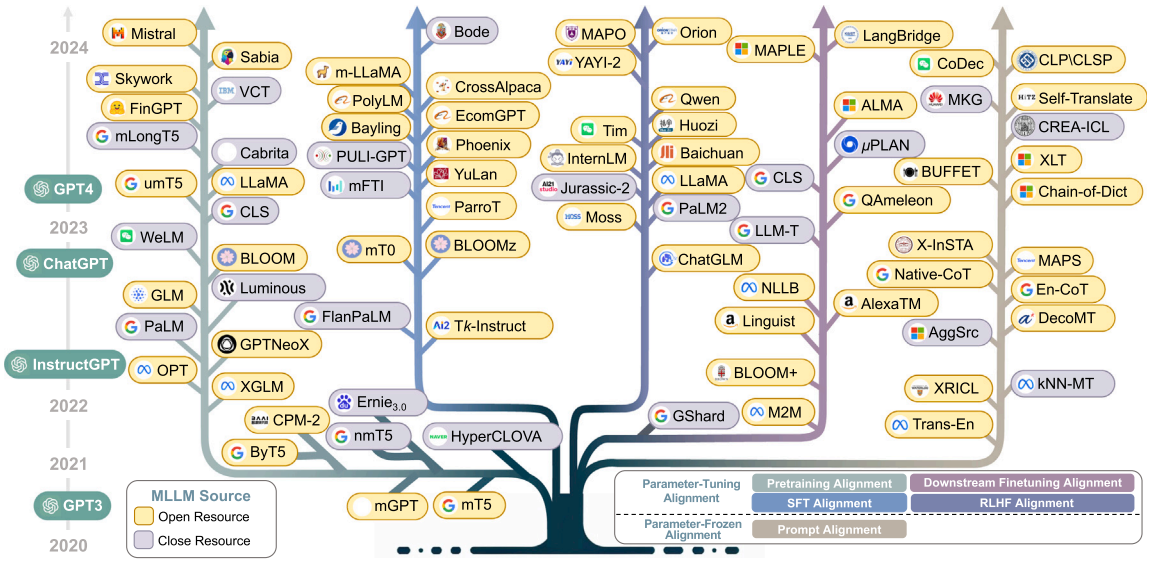

Figure 1. Evolution of selected MLLMs over the past 5 years, where colored branches indicate different alignment stages
For models with multiple alignment stages, the final stage is represented.

MLLMs. To bridge this gap, we attempt to conduct an extensive and detailed analysis of MLLMs. Specifically, we first introduce the widely used data resources and evaluation techniques. Furthermore, due to the key challenge of alignment across languages, we introduce a novel taxonomy according to alignment strategies, aiming to provide a unified perspective in the literature. This taxonomy primarily includes parameter-tuning alignment (PTA) and parameter-frozen alignment (PFA; as shown in Figure 2). In particular, PTA involves fine-tuning model parameters to enhance alignment between English and target languages during pretraining, supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and downstream fine-tuning. On the other hand, PFA refers to alignment achieved by prompting across languages, without requiring adjustments to model parameters. Finally, we highlight some potential frontier areas and the corresponding challenges, especially MLLMs for low-resource languages, hoping to inspire follow-up research.

The main contributions of this work can be summarized as follows. (1) Extensive survey: to the best of our knowledge, we present a comprehensive survey in the MLLM literature based on multilingual alignment. (2) Unified taxonomy: we introduce a unified taxonomy categorizing MLLMs into two alignment types, parameter frozen and parameter tuning, providing a systematic perspective to understand the MLLM literature. (3) Emerging frontiers: we discuss emerging frontiers and highlight their challenges as well as their opportunities, aiming to pave the way for future research developments. (4) Abundant resources: we attempt to organize MLLM resources, including open-source software, diverse corpora, and a curated list of relevant publications.

We hope this work can serve as a valuable resource for researchers and inspire further breakthroughs in future research.

RELATED WORK

Multilingual and cross-lingual natural language processing (NLP) has emerged as a vibrant research area.[22–24] Recent surveys have shed light on multilingual models and cross-lingual transfer, examining language technologies in diverse linguistic and cultural contexts. Previously, Doddapaneni et al.[25] demonstrated that pretrained language models (PLMs) enhance performance in both familiar and unfamiliar languages across various tasks. Similarly, Philippy et al.[26] analyzed the factors affecting zero-shot cross-lingual transfer, offering an in-depth discussion. Additionally, Doğruöz et al.[27] and Winata et al.[28] explored the linguistic and social dynamics of code switching, highlighting its significance in multilingual NLP. The rising demand for global multilingual systems has spurred numerous downstream tasks. For instance, Dabre et al.[24] examined PLMs in machine translation, while Deng et al.[29] focused on their role in information extraction. Panchendrarajan and Zubiaga[30] reviewed methods for identifying fact claims in multilingual and cross-lingual settings. As PLMs are deployed in real-world applications, concerns regarding safety, fairness, and bias have grown. Hershcovich et al.[31] emphasized the cultural sensitivities crucial for effective cross-lingual NLP, while Jiang and Zubiaga[32] discussed offensive language management and dataset challenges. Navigli et al.[33] and Ramesh et al.[34] highlighted the risks of bias, particularly in non-English languages, stressing the need for fairness in multilingual models. Lastly, Yadav and Sitaram[35] expanded their reviews of multilingual PLMs to include multi-modal scenarios. In summary, existing surveys provide comprehensive insights into the technical advancements and challenges of multilingual PLMs, calling for a deeper understanding of these models across diverse cultural and linguistic environments.

With the emergence of LLMs such as GPT-3[1] and GPT-4,[36] various surveys have examined the architecture, capabilities, and limitations of these models, with a particular emphasis on multilingual performance and their alignment with human-like understanding.[4,37–39] Nonetheless, there is a notable lack of comprehensive surveys specifically focused on MLLMs. To address this gap, we undertake a systematic analysis of MLLMs within the contemporary landscape of LLMs.

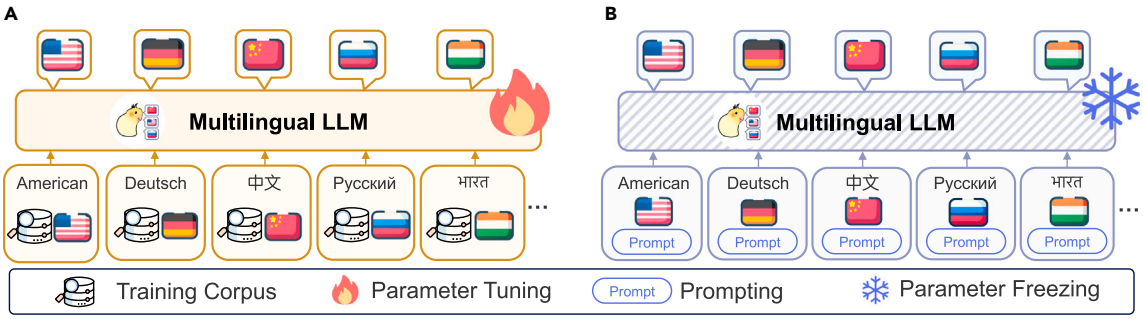

Figure 2. Parameter-tuning alignment vs. parameter-frozen alignment
(A) Parameter-tuning alignment requires the model to fine-tune the MLLM parameters for cross-lingual alignment.
(B) Parameter-frozen alignment directly uses prompts for alignment without parameter tuning.

PRELIMINARY DEFINITIONS

In this section, we formally describe the definitions of monolingual LLMs and MLLMs.

Monolingual LLMs

A monolingual LLM can process only one language at a time. For example, as illustrated in Figure 3A, an English LLM and a Chinese LLM can handle English and Chinese languages separately. Formally, considering a set of languages $\mathcal{L} = \{\mathcal{L}_i\}_{i=0}^{|\mathcal{L}|}$, given an input utterance $\mathcal{X}_i \in \mathcal{L}_i$ in languages $\mathcal{L}_i$, the process of a monolingual LLM ($\mathcal{M}_{mono}$) generating the output $\mathcal{Y}_i$ can be defined as

$$\mathcal{Y}_i = \begin{cases} \mathcal{M}_{mono}(\mathcal{X}_i, \mathcal{L}_i), & mono = \mathcal{L}_i; \\ Unexpect, & mono \neq \mathcal{L}_i, \end{cases} \quad \text{(Equation 1)}$$

where *Unexpect* indicates that the LLM generates an unexpected output in an unintended language; *mono* denotes the single language that the LLM can correctly process.

MLLMs

As shown in Figure 3B, unlike a monolingual LLM, an MLLM is capable of handling and producing content in various languages simultaneously, such as English and Chinese. Formally, for an MLLM $\mathcal{M}_{multi}$, where $multi \subseteq \mathcal{L}$ and $|multi| \geq 2$, the multilingual response of the MLLM is given by

$$\mathcal{Y} = \mathcal{M}_{multi}(\mathcal{X}), \quad \text{(Equation 2)}$$

where $\mathcal{X}$ and $\mathcal{Y}$ belong to multiple languages in *multi*.

RESOURCES FOR TRAINING

In this section, we describe the widely used data resources in pretraining, SFT, and RLHF stages in MLLMs.[4]

Multilingual pretraining data

As shown in Table 1, the widely used multilingual corpora for pretraining in MLLMs can be divided into three categories: (1) manual creation: high-quality pretraining corpora obtained through manual creation and proofreading, including the Bible Corpus[40] and MultiUN.[41] (2) Web crawling: this involves crawling extensive multilingual data from the internet, which include OSCAR,[42] CC-100,[43] mC4,[17] and Redpajama-v.2.[44] Relatively speaking, the data quality obtained through extensive crawling is often poor; however, the sheer volume of data compensates for this by providing substantial world knowledge and long-tail knowledge. Another category of data is extracted from Wikipedia to enhance the knowledge embedded in MLLMs. Common datasets include Wikipedia, WikiMatrix,[45] and WikiExpl.[46] Since Wiki data are authored by humans, they feature high quality and sufficient knowledge density, making them a crucial resource for injecting knowledge into MLLMs. (3) Benchmark adaptation: this refers to re-cleaning or integrating existing benchmarks to enhance data quality, which includes datasets such as OPUS-100,[47] Culturax,[48] OPUS,[49] WMT,[50] and ROOTS.[51] The pretraining data produced by this method are of higher quality than web-crawled data. However, this also results in such data being relatively scarce and lacking diversity.

Multilingual SFT data

Similarly, as shown in Table 2, we categorize the existing multilingual SFT data into four classes: (1) manual creation: this involves acquiring SFT corpora through manual creation and proofreading, which includes Sup-NatInst,[65] OpenAssist,[66] and COIG-PC$_{lite}$. This method ensures the highest quality, but it is costlier and produces a smaller volume of labeled data. (2) Machine translation: this method translates existing monolingual datasets into multilingual instruction datasets, which comprise xP3-MT,[20] MGSM8K$_{Instruct}$,[67] CrossAlpaca,[68,69] MultilingualSIFT,[70] and Bactrain-X.[71] This approach generates extremely large quantities of data with moderate quality. Its advantage lies in rapidly producing a substantial amount of non-English SFT data. However, it may fail to account for the cultural background of specific languages, leading to implicit biases. (3) Benchmark adaptation: this method transforms from existing benchmarks into an instruction format. Widely used datasets include xP3,[20] PolyglotPrompt,[72] and BUFFET.[73] The data quality of this approach is high, but the diversity of tasks and instructions is limited. (4) MLLM-aided generation: such a strategy means that the data are automatically synthesized by MLLMs, containing Vicuna,[74] OverMiss,[75] ShareGPT, BELLE,[76] MultiAlpaca,[77] Guanaco,[78] and Alpaca-4.[79] Data generated from advanced MLLMs may surpass translation quality in high-resource languages. However, it is concerning that data quality may degrade in low-resource languages.

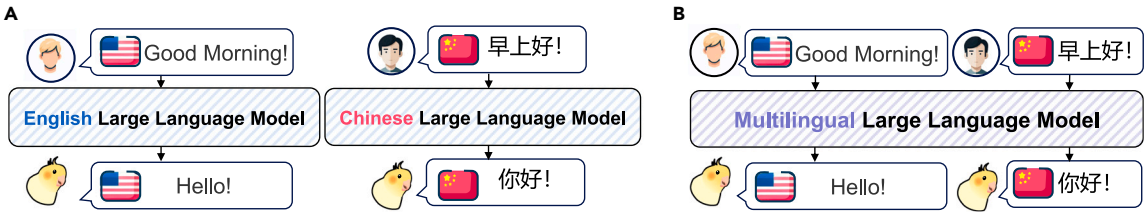

Figure 3. Monolingual large language model vs. multilingual LLM
(A) Monolingual large language model (LLM) can only process one language at a time.
(B) A multilingual LLM is capable of handling and producing content in various languages simultaneously, like English and Chinese.

Multilingual RLHF data

Furthermore, some works leverage multilingual RLHF data to improve alignment. Specifically, Lai et al.[85] leverage multilingual ranking data to train a reward model using RLHF. Similarly, Zeng et al.[86] introduce the TIM dataset to train a more effective reward model in multilingual contexts. This type of data is often labeled with preferences for translation tasks and is of high quality, though there remains room for improvement in task diversity.

MULTILINGUAL PERFORMANCE EVALUATION

To facilitate the comparison of LLMs, extensive efforts have been dedicated to exploring enhanced evaluation methods for multilingual scenarios. This discussion elaborates on MLLM evaluation, covering both (1) evaluation metrics and (2) evaluation benchmarks.

Evaluation metrics

Traditional automatic metric

Using traditional automatic metrics means that we assess the predicted output using decoding probabilities or PLM logits.[87,88] In general, researchers utilize BLEU,[89] BLEURT,[90] chrF++,[91] and COMET[92] for translation evaluation and ROUGE[93] for summary evaluation. Further, Guerreiro et al.[94] propose xCOMET for improved translation evaluation through fine-grained error detection. To assess the general quality of the generated text, the commonly employed approach is to utilize multilingual BERTScore[95] as an evaluation metric. Qin et al.[12] extend Roscoe[96] to multi-language for quality assessment of multilingual CoT. Furthermore, Hlavnova and Ruder[97] develop a comprehensive and robust multilingual checklist system to thoroughly assess the MLLMs' performance.

MLLM-based automatic metric

This approach employs robust MLLMs to score or compare generated outputs for evaluation purposes.[19,71,98] Specifically, Zheng et al.[99] introduce LLM-as-a-judge, where GPT-4 is prompted to assess the performance of other LLMs by comparing its output to the predicted one. However, this method remains unreliable in multilingual settings.[3,26,100,101] Moreover, caution should be exercised particularly in languages where the MLLM is known to perform poorly.[102] Furthermore, Kim et al.[103] and Muller et al.[104] conduct an attribution evaluation to thoroughly assess the robustness of the model.

Human evaluation

Human evaluation involves manually assessing MLLMs through detailed evaluations.[19,71,105,106] Specifically, Lyu et al.[107] initially explore the multilingual challenges of ChatGPT through manually annotated cases. Furthermore, Hu et al.[108] introduce a new multilingual platform to facilitate more convenient manual assessments.

Evaluation benchmarks

Current MLLMs tend to focus more on the alignment performance of non-English languages. Based on different perspectives of alignment, we categorize this into two areas: (1) natural language understanding and (2) natural language generation.

Natural language understanding

Linguistics analysis. For multilingual models, the fundamental requirement is understanding the linguistic differences across languages.[109] The most common multilingual linguistic assessments include part-of-speech (POS) tagging,[110,111] grammar analysis,[112–114] and morphology.[115] Additionally, Zhang et al.[116] and Song et al.[117] provide a comprehensive evaluation of the linguistic acceptability of MLLMs across languages.

Semantic understanding. Compared with linguistics analysis, researchers are more focused on the ability to analyze and understand the specific semantics of multiple languages.[30,100,118–120] The most fundamental aspect of multilingual NLP is performing local semantic understanding, with the most typical task being information extraction,[121] including datasets such as masakhaNER,[122] MASSIVE,[123] MultiCoNER,[124,125] WikiAnn,[126] and SMiLER.[127] The second level involves a semantic understanding of complete sentences, which includes tasks like XNLI,[128] Paws-X,[129] MixATIS++,[130] MTOP,[131] MultiNLU,[132] and PRESTO.[133] Finally, semantic understanding can be further extended to paragraphs, exemplified by question-answering tasks with context, such as MLQA,[134] XQuAD,[135] TyDiQA,[136] X-PARADE,[137] X-CLAIM,[138] Readme++,[139] XKaggle-DBQA,[140] and de Varda and Marelli.[141] Due to the emergence of a large number of multilingual benchmarks in recent years, a series of works has begun to combine various existing semantic understanding tasks into unified evaluations, including XTREME,[142] XTREME-R,[143] XGLUE,[110] MEGA,[144] MEGAVerse,[145] AGIEval,[146] and Superlim.[147] Further, Thapliyal et al.,[148] Changpinyo et al.,[149] Fujinuma et al.,[150] and Kudugunta et al.[54] extend semantic understanding to multi-modal contexts. Given that MLLMs exhibit certain biases[151,152] or vulnerabilities,[153–155] a growing body of work[32,156–158] is dedicated to developing benchmarks specifically designed to rigorously evaluate the performance and reliability of MLLMs in addressing these issues.

Cultural understanding. Limited by cultural differences, understanding between different languages is not completely

Table 1. Pretraining data resources

Dataset	Storage size (B)	Token size (T)	Language size	Source	Latest update time
Manual					
Bible Corpus[40]	5.2 G	–	833	–	May 2014
MultiUN[41]	–	0.3 B	7	–	December 2014
IIT Bombay[52]	–	0.04 B	2	–	December 2021
Web crawling					
CC-100[43]	–	208 B	116	CommonCrawl	October 2022
mC4[17]	38.5 T	6.3 T	101	CommonCrawl	October 2022
Redpajama-v2[44]	30.4 T	–	5	CommonCrawl	December 2023
OSCAR[42]	6.3 T	800 B	166	CommonCrawl	January 2023
Oromo[53]	0.939 G	0.1 B	11	CommonCrawl	February 2022
Wu Dao 2.0	–	24 B	2	CommonCrawl	October 2023
MADLAD-400[54]	–	3 T	419	CommonCrawl	August 2022
HPLT-Full[55]	230.7 T	5.6 T	75	CommonCrawl	October 2022
HPLT-En-Center[55]	–	1.4 B	18	CommonCrawl	October 2023
Europarl[56]	1.5 G	0.6 B	21	–	May 2012
JW300[57]	–	1.5 B	343	–	July 2019
Glot500[58]	600 G	–	511	–	May 2023
Wikipedia[59]	–	24 B	300	Wikipedia	–
WikiMatrix[45]	65 G	–	85	Wikipedia	April 2021
OPUS-100[47]	2.6 G	–	100	OPUS	July 2020
AfricanNews[60]	12.3 G	–	16	mC4	September 2023
Taxi1500[61]	–	–	1500	Bible Corpus	May 2023
CulturaX[48]	27 T	6.3 T	167	mC4, OSCAR	January 2024
MMedC[62]	–	25.5 B	6	medical website	–
Benchmark adaptation					
ROOTS[51]	1.6 T	–	46	Huggingface	June 2022
OPUS[49]	–	40 B	1,304	–	December 2021
CCMT[63]	–	–	6	–	–
WMT[50]	–	–	32	–	–
IWSLT[64]	4.2 G	–	10	–	–

The term "source" refers to the origin datasets from which the pretraining data are derived. The storage size is measured in bytes (B); where relevant, "G" represents the billion scale and "T" represents the trillion scale. The token size is measured in the number of tokens (T); where relevant, "B" represents the billion scale and "T" represents the trillion scale.

parallel.[159–162] Consequently, researchers have begun exploring ways to evaluate multi-cultural scenarios,[31,163] with the most typical being multi-cultural sentiment analysis.[71,164–168] Furthermore, Zhang et al.[169] expands the multi-cultural scope to the sociopragmatic understanding level. Specifically, Kabra et al.,[170] Wang et al.,[171] Jiang and Joshi,[172] Fung et al.,[173] Li et al.,[174] Son et al.,[175] and Zhou and Zhang[176] propose new benchmarks that require models to fully comprehend diverse cultures. Furthermore, with the emergence of reasoning capabilities, Qin et al.,[12] Liu et al.,[177] and Wang et al.[178] start to evaluate the reasoning abilities of MLLMs across different cultural backgrounds.

Knowledge understanding. A large amount of work has been done to test the degree of knowledge transfer of MLLMs between different languages through examination questions. Specifically, Hardalov et al.,[179] Xuan-Quy et al.,[180] Zhang et al.,[181] Nie et al.,[182] and Ni et al.[183] propose comprehensive knowledge tests in multilingual scenarios. Zhang et al.[16] design a complex translation strategy to translate existing benchmarks for multilingual evaluation. On this basis, M3Exam[184] and EXAMS-V[185] further expand comprehensive knowledge testing to multilingual and multi-modal scenarios. Furthermore, Gekhman et al.[186] test the factual consistency of MLLMs, and Jin et al.,[187] Joseph et al.,[188] Zhao et al.,[189] Wei et al.,[121] Goenaga et al.,[190] Datta et al.,[191] and Thulke et al.[192] propose benchmarks to evaluate the multilingual scientific and professional domain knowledge of current MLLMs.

Natural language generation

Translation. In the process of multilingual alignment, in addition to testing whether multiple languages are aligned in terms of understanding capabilities, researchers often need to consider whether they can also be aligned in terms of output capabilities.[193] The most typical task is machine translation.[24,194] Currently, commonly used datasets include FLORES-101,[195] FLORES-200,[196] WMT,[50] and DiaBLa.[197] Furthermore, Lou et al.[198] propose CCEval for Chinese-centric translation to enable comprehensive evaluation on MLLMs. Due to the significant gap between languages,[15,16,34,187,199–205] Kuparinen

Table 2. Supervised fine-tuning data resource

Dataset	Sample size	Multilingual instruction	Language size	Task size
Manual				
Sup-NatInst[65]	–	–	55	1,616
OpenAssist[66]	–	–	35	–
EcomInstruct[80]	2.5 M	yes	2	12
COIG-PC-lite	650 K	no	2	3,250
Aya Dataset[81]	204 K	no	65	–
Benchmark adaption				
xP3[20]	80 M	no	46	71
BUFFET[73]	–	–	54	15
PolyglotPrompt[72]	–	no	49	6
Translation				
xP3-MT[20]	80 M	yes	46	71
xP3x[82]	168 M	no	101	56
Aya Collection[81]	513 M	no	114	44
MultilingualSIFT[70]	–	yes	11	–
Bactrian-X[71]	–	yes	52	–
MulT[83]	–	yes	6	–
CrossAlpaca[68]	–	–	6	–
MGSM8KInstruct[67]	73.6 K	yes	6	10
XCoT[84]	7.4 K	yes	10	2
MLLM aided				
ShareGPT	–	–	–	–
Vicuna[74]	–	–	–	–
OverMiss[75]	54 K	–	3	1 (translation)
MultiAlpaca[77]	133 K	–	11	–
Guanaco[78]	535 K	–	5	–
Alpaca-4[79]	52 K	–	2	–

The term "multilingual instruction" denotes the presence of instructions in multiple languages to form the specific data input.

et al.,[206] Wassie,[207] Liu et al.,[208] Rakhimova et al.[209] focus more on low-resource-language translation. Additionally, Yang et al.,[210] Gueuwou et al.,[211] Bellagente et al.,[212] Zhong et al.,[213] and Tuo et al.[214] further extend translation and restatement tasks into multi-modal settings for practical scenarios.

Reasoning. Currently, the most commonly used reasoning ability assessments for MLLMs focus on commonsense and mathematical reasoning.[11,12] Specifically, commonsense reasoning includes XCOPA,[215] MARC,[216] XWinograd,[217] GEOMLAMA,[218] X-CSQA,[219] XStoryCloze,[220] ASPEN,[221] and Masakhanews.[222] Mathematical reasoning includes MGSM[21] and WizardMath.[223] Additionally, due to the high cost of annotations for multilingual reasoning, Zhang et al.[16] propose a complex translation and filtering process to construct a multilingual reasoning benchmark.

Coding generation. The generation of code by MLLMs necessitates the capability to produce structured and executable programs based on multilingual natural language instructions. Commonly utilized benchmarks for evaluating this capability include XSPIDER,[140] XSEMPLR,[224] ODEX,[225] Mconala,[226] and HumanEval-XL.[227]

Summarization. To test the summarization ability of MLLMs, summarizing key information from multilingual long texts is required. The simplest example is that from Ryan et al.,[228] who propose a multilingual text reduction benchmark for the evaluation of MLLMs. Secondly, much work focuses on cross-lingual summarization, with typical datasets including XSUM[229] and CrossSum.[230] On this basis, Wang et al.[231] introduce multilingual conversation summarization, and Zhang and Eickhoff[232] propose incorporating code switching into evaluations, making them more practical. Urlana et al.[233] further propose headline summarization for Indian languages. SEAHORSE[234] extends this work to multifaceted multilingual summarization. Additionally, Nguyen et al.[235] and Verma et al.[236] develop summarization benchmarks for multi-modal scenarios.

Dialogue. The communication between models and humans is often interactive; hence, a series of works pay attention to the dialogue ability of MLLMs.[237] Current evaluation sets include xDial-Eval,[238] Multi^{3}WOZ,[239] DIALIGHT,[108] HPD,[240] and X-RiSAWOZ.[241] Since multi-turn dialogues are inherently uncontrollable, traditional metrics are insufficient. To address this, Mendonca et al.[242] utilize PLMs for multi-turn dialogue evaluation. Furthermore, Mendonça et al.[243] propose a new benchmark that enables more robust evaluation by leveraging PLMs. Finally, Ferron et al.[244] introduce the MEEP benchmark to further assess the dialogue participation of MLLMs.

Table 3. The accuracy performance of different MLLMs on XNLI benchmark

Model	en	de	ru	fr	zh	es	vi	tr	sw	ar	el	th	bg	hi	Ur	Avg.
PTA in pretraining stage																
BLOOM-1.1B[20]	33.6	33,3	33.4	33.5	33.4	33.3	33.4	33.6	33.5	33.6	33.5	33.4	33.3	33.5	33.4	33.5
mGPT-1.3B[245]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	40.6
BLOOM-1.7B[20]	33.6	34.1	33.3	34.5	33.3	33.5	33.5	33.4	33.5	34.0	33.4	33.4	33.4	34.5	33.7	33.7
XGLM-1.7B[246]	49.7	–	–	47.9	44.6	37.4	42.8	–	–	45.7	–	–	–	44.4	43.2[a]	–
BLOOM-3B[20]	33.6	35.2	33.6	33.8	35.5	33.9	33.6	33.4	34.0	35.9	33.3	33.3	33.3	33.8	33.4	34.0
BLOOM-7.1B[247]	54.0	39.2	41.3	51.7[a]	48.1[a]	41.5	48.9[a]	38.9	37.7	47.4[a]	36.3	39.3	37.8	47.4	39.9	43.3
XGLM-7.5B[247]	54.1	42.5	45.0	49.9	45.4	39.9	47.2	44.7[a]	44.3[a]	46.4	45.4[a]	45.2[a]	48.9[a]	43.2	42.1	45.6[a]
LLaMA-13B[77]	35.5	35.2	33.6	33.6	34.6	33.4	34.1	34.0	33.2	34.1	34.5	34.6	33.9	35.7	34.1	34.3
mGPT-13B[245]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	42.6
AlexaTM-20B[248]	55.1[a]	47.1[a]	–	50.4	–	47.5[a]	–	–	–	45.1[a]	–	–	–	48.7[a]	–	–
PTA in SFT stage																
BLOOMZ-1.1B[20]	39.0	36.3	35.8	42.7	39.4	41.2	40.5	34.0	35.5	41.3	35.5	33.6	35.5	37.6	35.3	37.5
BLOOMZ-1.7B[20]	49.8	41.1	41.5	48.6	48.1	46.9	46.4	35.5	41.2	47.0	38.4	38.9	38.6	44.5	39.9	43.1
BLOOMZ-3B[20]	52.2	40.8	43.7	50.5	47.7	50.3	46.2	36.1	41.8	48.2	40.4	38.2	39.4	46.2	42.9	44.3
Mistral-7B-Instruct[249]	–	50.4	55.6	59.2	46.0	59.0	33.4	38.8	33.0	34.2	34.2	39.2	46.6	37.0	33.2	–
BLOOMZ-7.1B[247]	51.1	43.7	42.3	48.0	49.7	41.4	48.7	39.8	39.2	48.2	39.7	40.9	40.3	45.6	42.9	44.1
AFP-BLOOM-7.1B[247]	55.0	42.2	42.4	52.7	48.1	43.8	50.0	41.6	42.0	50.0	40.4	42.3	40.3	45.3	45.1	45.4
AFP-XGLM-7.5B[247]	54.8	44.6	48.4	51.3	50.6	41.9	47.7	47.4	45.3	48.8	47.2	48.6	48.8	44.4	42.6	47.5
PolyLM-13B[77]	54.6	50.2	49.0	52.1	35.9	50.0	46.7	44.5	34.4	33.6	33.8	44.6	36.3	34.9	33.5	42.3
mT0x-13B[82]	50.7	47.9	47.7	48.8	45.8	49.6	46.5	44.8	45.1	44.9	48.7	45.8	47.6	45.0	43.3	46.8
Aya-13B[82]	61.5[a]	59.2	58.3[a]	57.4	52.8	59.9	58.3[a]	55.9	55.5	57.0[a]	58.7	55.5	59.5[a]	54.8	52.4	57.1
mT0-13B[20]	61.2	60.1[a]	58.0	59.5[a]	57.1[a]	60.3[a]	58.2	56.8[a]	55.5[a]	56.5	58.8[a]	56.3[a]	59.2	56.1[a]	54.7[a]	57.9[a]
BLOOMZ-176B[20]	60.9	53.1	52.3	57.7	56.2	58.5	55.8	42.7	50.4	54.4	45.9	41.5	47.2	53.5	50.0	52.0
PTA in RLHF stage																
Llama-2-13B-chat[249]	–	41.4	40.2	44.0	38.6	42.8	32.4	34.6	31.6	32.8	34.2	34.0	37.4	31.4	33.6	–
Llama-2-70B-chat[249]	–	44.0	42.0	45.4	42.6	45.6	38.4	38.4	32.6	35.0	37.6	33.0	41.8	34.8	34.8	–
GPT-3-text-davinci-003[11]	63.6	59.4	55.9	60.9	51.6	59.7	49.5	53.9	40.8	51.9	53.2	49.7	54.4	49.8	45.3	53.3
GPT-3.5-turbo[11]	65.4	55.5	50.6	53.2	48.8	59.8	52.1	54.4	49.6	50.9	54.9	44.8	55.7	49.2	44.8	52.6
GPT-4[144]	84.9[a]	78.8[a]	74.3[a]	79.5[a]	74.6[a]	78.8[a]	74.3[a]	76.3[a]	70.9[a]	73.1[a]	79.0[a]	68.8[a]	77.3[a]	72.0[a]	68.1[a]	75.4[a]
PTA in downstream fine-tuning stage																
LLaMA-6.7B[101]	86.9[a]	75.8[a]	73.1[a]	77.0[a]	61.8	77.6	54.7	51.1	40.6	52.7	57.4	46.6	72.8[a]	47.0	45.5	61.4
Pythia-6.9B[101]	83.8	65.9	61.1	70.1	61.8	70.8	55.9	54.4[a]	45.6	56.0	61.5[a]	51.4[a]	61.9	47.3	46.3	59.6
BLOOM-7.1B[101]	81.4	59.7	59.6	75.3	70.7[a]	78.0[a]	70.0[a]	44.2	56.3[a]	69.5[a]	51.2	46.6	50.5	62.8[a]	57.3[a]	62.2[a]

[a]Represents the best performance on different languages at this stage.

TAXONOMY

While the performances of most prevalent MLLMs are exceptional for English, their effectiveness in other languages is notably lower, primarily due to the limited availability of linguistic resources. Consequently, alignment emerges as an effective strategy for improving this performance. As demonstrated in Table 3, efficient alignment can even surpass the model's scaling laws, yielding superior results.

Inspired by this, as shown in Figure 4, we introduce a unified taxonomy focusing on multilingual alignment, which includes PTA and PFA, aiming to provide a systematic framework for researchers to better understand the MLLM literature. Specifically, PTA comprises a series of progressively advanced training and alignment strategies, including pretraining alignment, SFT alignment, RLHF alignment, and downstream fine-tuning alignment. These PTA stages collectively aim to refine model parameters to comprehensively improve multilingual performance. Conversely, PFA focuses on four prompting strategies based on the MLLMs trained with PTA: direct prompting, code-switching prompting, translation alignment prompting, and retrieval-augmented alignment. These PFA methods retain the original parameters to achieve the desired outcomes.

PTA

PTA refers to the process of tuning the parameters of MLLMs to achieve better cross-lingual alignment.[250] As shown in Figure 5, we discuss four categories of PTA, including PTA during the pretraining stage, the SFT stage, the RLHF stage, and the fine-tuning stage.

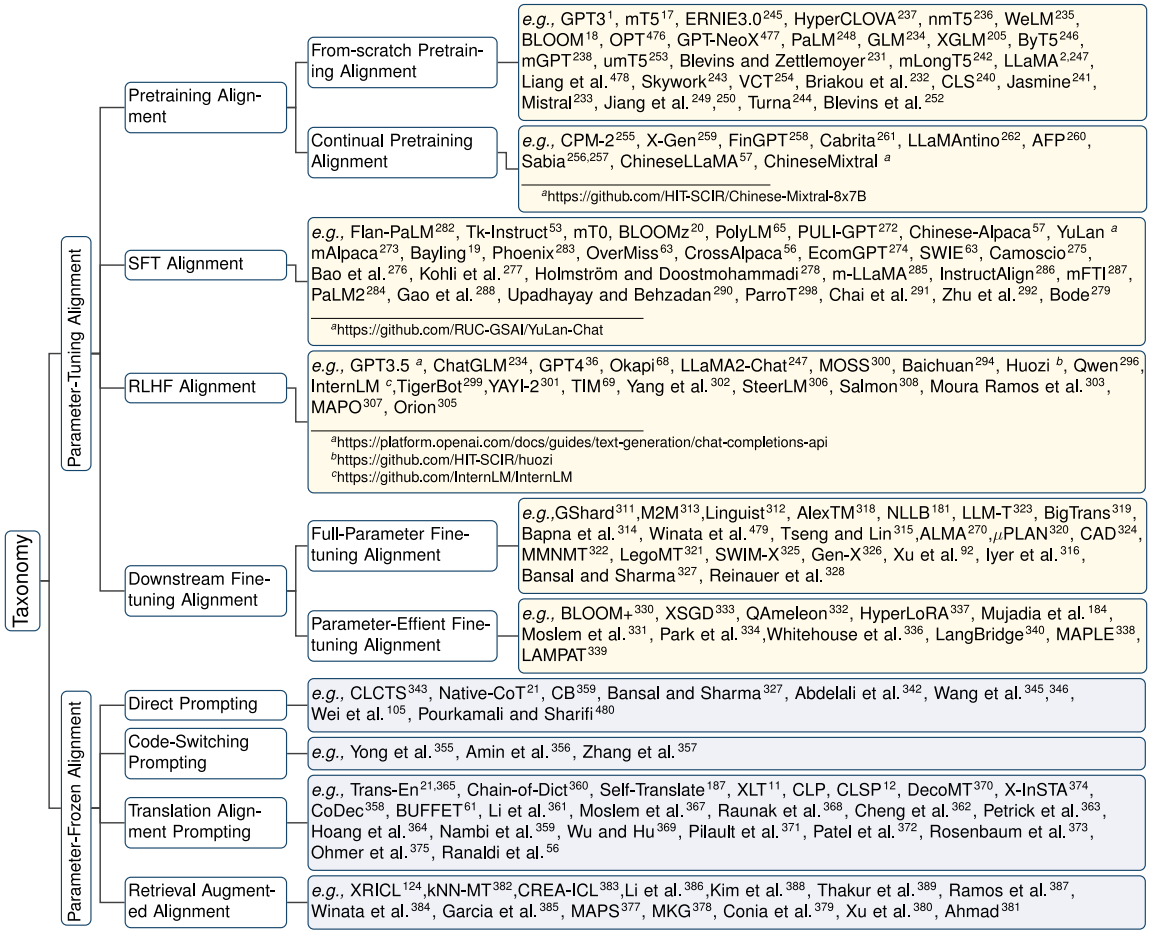

Figure 4. Taxonomy of MLLMs, which includes parameter-tuning alignment methodology and parameter-frozen alignment methodology

PTA in the pretraining stage

From-scratch pretraining alignment. A range of approaches have achieved alignment across languages by tuning the initially random parameters of MLLMs during pretraining (see Figure 5A). Specifically, Blevins and Zettlemoyer,[251] Briakou et al.,[252] and Holmström et al.[253] observe that adding even a small amount of multilingual data during from-scratch pretraining alignment, whether intentional or not, can significantly enhance multilingual performance. Inspired by this, Zeng et al.[254] and Su et al.[255] proactively incorporate bilingual data into their from-scratch pretraining for alignment. Furthermore, a range of studies,[54,245,256–263] such as mT5,[17] Ernie3.0,[264] ByT5,[265] BLOOM,[18] LLaMA,[2,266] PaLM,[267] Mistral,[268] Mixtral,[269] PolyLM,[77] and Nemotron-15B,[270] incorporate multilingual data in the pretraining stage for better alignment. Blevins et al.[271] utilize mixture of experts (MoE) to independently train language models on subsets of multilingual corpora to alleviate the problem of multilingual parameter competition. Furthermore, to enhance the performance of low-resource languages, umT5[272] and XGLM[220] adopt equitable multilingual data sampling methods during from-scratch pretraining for more effective alignment. Muraoka et al.[273] introduce VCT to leverage vision for indirect cross-lingual alignment in from-scratch pretraining.

Continual pretraining alignment. To address the high computational cost of from-scratch pretraining, continual pretraining alignment is proposed to build the continual training process upon pretrained MLLMs (as shown in Figure 5A). Specifically, CPM-2,[274] Sabia,[275,276] FinGPT,[277] X-Gen,[278] AFP,[247] Cabrita,[279] LLaMAntino,[280] CroissantLLM,[281] MedMT5,[282] and Tang et al.[283] focus on adding more target language data during continual pretraining to enhance general performance. Further, Cui et al.,[69] Yamaguchi et al.,[284] and Lin et al.[285] emphasize extending the MLLMs' vocabularies to adapt to new languages and enable more effective decoding. Singh et al.[286] and Fujii et al.[287] demonstrate that continual pretraining on a specific language significantly enhances model performance across related languages. Blevins et al.[271] extend continual pretraining to the MoE paradigm for improved parameter efficiency. To achieve deeper model alignment, Xu et al.[288] and Guo et al.[289] introduce a novel continuous pretraining paradigm, which is structured in two stages. Initially, the model undergoes pretraining on a substantial corpus of monolingual data. Subsequently, it engages in continual pretraining utilizing multilingual parallel data.

PTA in the SFT stage

As illustrated in Figure 5B, PTA in the SFT stage involves leveraging multiple multilingual task datasets with instruction formats for tuning parameters.[68,75,80,82,290–297] In particular,

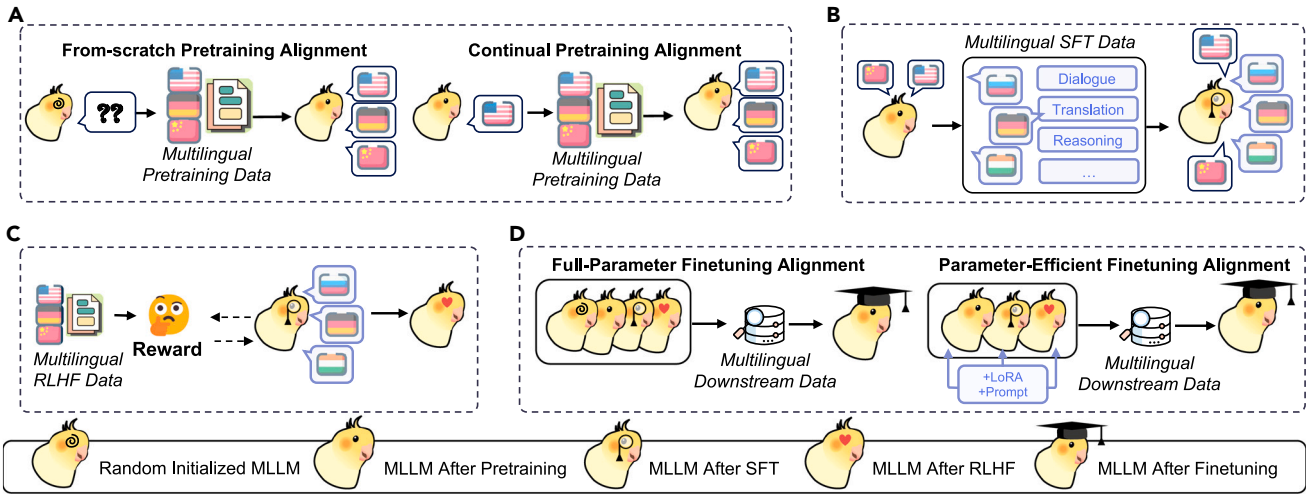

Figure 5. Overview of parameter-tuning alignment methods, which includes PTA in the pretraining stage, PTA in the SFT stage, PTA in the RLHF stage, and PTA in the downstream fine-tuning stage

(A) Parameter-tuning alignment in the pretraining stage involves tuning MLLMs on vast, chaotic datasets. This process comprises both from-scratch pretraining alignment and continual alignment.

(B) Parameter-tuning alignment in the SFT stage refers to utilizing various multilingual task data in an instructional format for parameter tuning.

(C) Parameter-tuning alignment in the RLHF stage pertains to achieving alignment through reinforcement learning from human feedback.

(D) Parameter-tuning alignment in the downstream fine-tuning stage focuses on fine-tuning MLLMs to better adapt to specific downstream tasks, encompassing both full-parameter fine-tuning alignment and parameter-efficient fine-tuning alignment.

MLLMs like Flan-PaLM,[298] mT0, BLOOMz,[20] PolyLM,[77] Tk-Instruct,[65] Chinese-Alpaca,[69] Bayling,[19] Phoenix,[299] and Bode[296] directly incorporate multilingual data in the SFT stage to achieve implicit multilingual alignment across languages. Besides, to address the scarcity of multilingual SFT task data, PaLM2,[300] Zhu et al.,[83] Cahyawijaya et al.,[301] Li et al.,[302] Gao et al.,[303] and Aryabumi et al.[304] introduce translation tasks during the SFT alignment stage to improve alignment. Furthermore, Upadhayay and Behzadan,[305] Chai et al.,[84] and Zhu et al.[306] have begun exploring more effective SFT alignment strategies to optimize the reasoning process.

PTA in the RLHF stage

As shown in Figure 5C, to achieve alignment in the RLHF stage, Okapi,[85] LLaMA2-Chat,[266] ChatGLM,[254,307] Baichuan,[308] Huozi, Chinese-Tiny-LLM,[309] Qwen,[310] InternLM,[311] ParroT,[312] TigerBot,[313] MOSS,[314] YAYI-2,[315] Yang et al.,[316] Moura Ramos et al.,[317] Nemotron-340B,[318] and Orion[319] directly integrate multilingual RLHF data for training multilingual reward models. Additionally, Zeng et al.,[86] Dong et al.,[320] and She et al.[321] introduce a multilingual reward model to compare translation outputs at different levels of granularity. Sun et al.[322] propose the Salmon framework to enhance multilingual RLHF by self-generating rewards for better alignment. Furthermore, Xu et al.[323] introduce contrastive preference optimization (CPO) for translation tasks to address memory or speed inefficiencies in direct preference optimization (DPO).

PTA in the downstream fine-tuning stage

Full-parameter fine-tuning alignment. Full-parameter fine-tuning in MLLMs involves tuning all parameters for downstream tasks (see Figure 5D).[324] Specifically, GShard,[325] Linguist,[326] Fan et al.,[327] Bapna et al.,[328] Tseng and Lin,[329] Iyer et al.,[330] NLLB,[196,331] AlexTM,[248] and BigTrans[332] focus on directly fine-tuning all parameters across various downstream tasks (e.g., information extraction, machine translation). Xu et al.,[109] Huot et al.,[333] Yuan et al.,[334] and Li et al.[335] propose multi-step or fine-grained alignment strategies during full-parameter tuning. To enhance efficiency, Awasthi et al.,[336] De Raedt et al.,[337] Thakur et al.,[338] Whitehouse et al.,[339] Bansal and Sharma,[340] Xu et al.,[288] and Reinauer et al.[341] focus on fine-tuning alignment by knowledge distillation from larger to smaller MLLMs. Furthermore, Zhang et al.[342] identify a scaling law for fine-tuning in translation tasks, significantly advancing the understanding of performance improvements through multilingual fine-tuning alignment.

PEFT. A series of studies employ parameter-efficient fine-tuning (PEFT) alignment approaches to reduce the costs of full-parameter fine-tuning,[199,343,344] which is shown in Figure 5D. Agrawal et al.,[345] Tu et al.,[346] Park et al.,[347] and Dai et al.[348] utilize minimal soft-prompt prefixes for improved fine-tuning alignment. Furthermore, Whitehouse et al.,[349] Xiao et al.,[350] Aggarwal et al.,[351] Le et al.,[352] and Singh et al.[286] introduce methods based on low-rank adaptation (LoRA) to achieve PEFT alignment. Further, Yoon et al.[353] propose a LangBridge model to bridge a multilingual encoder to a single-lingual LLM, effectively achieving promising performance. In addition, Zhao et al.[354] introduce two distinct types of adapters: one tailored for language processing and the other for task-specific applications. These adapters can be effectively combined and integrated, facilitating rapid adaptation to new tasks or datasets.

Performance analysis

To further evaluate the effectiveness of various PTA strategies, as shown in Table 3, we compare different MLLMs trained at various stages using the XNLI[128] benchmark, a widely recognized assessment for multilingual understanding.

Performance in English vs. other languages. All MLLMs exhibit consistent and strong performance in English, attributable to the

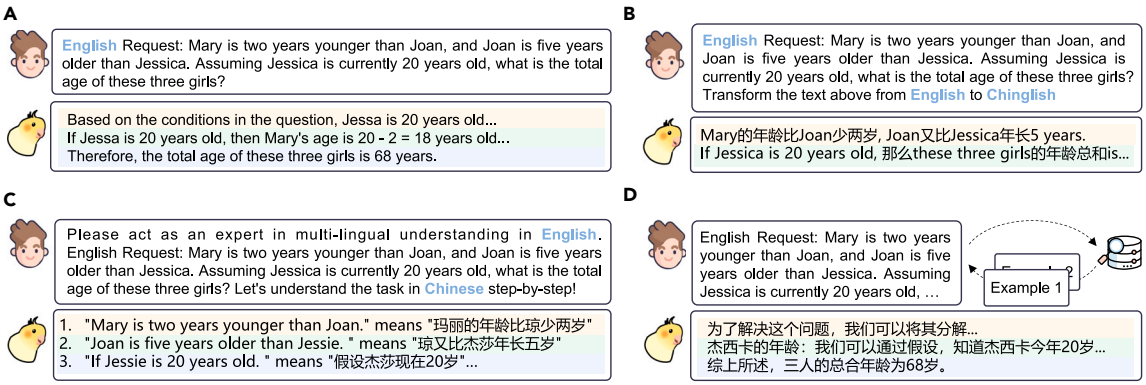

Figure 6. Overview of parameter-frozen alignment methods

The prompts were sourced from Qin et al.[12] and Zhang et al.[355]

(A) Direct prompting involves outputting requests directly without any additional instructions.

(B) Code-switching prompting encourages MLLMs to integrate multilingual words within a single-language utterance.

(C) Translation alignment prompting entails translating the query into other languages to enhance alignment.

(D) Retrieval-augmented alignment incorporates external retrieval mechanisms during prompting to infuse additional knowledge into MLLMs.

extensive availability of English training data and the emphasis on English during model development. For example, GPT-4 achieves an accuracy of 84.9% in English, compared to 68.1% in Urdu. This trend extends to models like GPT-3.5-turbo and Aya-13B, which also perform better in English than non-English languages. These disparities highlight the challenges that MLLMs encounter when addressing a diverse range of languages characterized by varying linguistic structures, data quality, and resource availability.

Impact of training stages on performance. The training stage of MLLMs plays a crucial role in determining their multilingual capabilities. Models in the pretraining stage generally develop fundamental cross-lingual abilities, especially in non-English languages. For instance, BLOOMZ and AFP-BLOOM, which heavily rely on the linguistic performance of the pretrained model BLOOM, exhibit limited average improvements of 0.8% and 2.1%, respectively.

Importance of alignment over model size. Recent findings suggest that the degree of alignment, achieved through pretraining, SFT, and RLHF, is more critical to MLLM performance than simply increasing the model size. For instance, despite having fewer parameters than LLaMA-2-70B-chat, Aya-13B demonstrates superior performance across multiple languages, including those with fewer resources. This observation underscores that effective alignment strategies enable models to generalize more effectively across languages, thereby narrowing the performance gap between high- and low-resource languages.

Takeaways. (1) PTA during the pretraining stage brings essential multilingual capabilities to MLLMs. (2) The effectiveness of alignment in MLLMs is greatly influenced by prior alignment stages (e.g., pretraining significantly impacts SFT).

PFA

In contrast to traditional parameter-tuning approaches,[22] PFA methods aim to perform alignment without any parameter tuning. The most popular approaches employ prompting strategies to elicit the alignment potential of MLLMs. As shown in Figure 6, this section discusses four prompting strategies for alignment without parameter tuning, consisting of (1) direct prompting, (2) code-switching prompting, (3) translation alignment prompting, and (4) retrieval-augmented alignment.

Direct prompting

As shown in Figure 6A, direct prompting refers to directly outputting the request without any additional instructions for implicit alignment through the MLLM itself.[121,220,340,356–360] Even in specific scenarios, MLLMs may actively select the language they excel in or find most suitable for expression, thereby achieving effective language alignment.[361]

Code-switching prompting

As shown in Figure 6B, this approach integrates multilingual words into a single-language utterance, a typical linguistic phenomenon[27,28,362–364] that facilitates effective language alignment.[365,366] Specifically, Yong et al.[367] and Amin et al.[368] demonstrate the effectiveness of MLLMs in cross-lingual alignment through model-generated code-switching texts. Moreover, Zhang et al.[355] highlight the need for fairer and more detailed code-switching optimization in future research.

Translation alignment prompting

Translation alignment prompting approaches involve translating the query into other languages to achieve better alignment[369,370] (see Figure 6C). These approaches can be categorized as follows: (1) key information translation: this approach focuses on extracting key information and executing translation for word-level cross-lingual alignment.[371,372] (2) Direct translation: the model directly translates the whole input, enhancing alignment performance,[106,202,220,373–375] which even exhibits superior results compared to the Google Translation API.[376] (3) Step-by-step translation: instead of direct translation, this method prompts MLLMs to translate the whole input step by step.[377–381] (4) Restatement: beyond preserving the original semantics, some studies focus on prompting MLLMs to restate multilingual inputs to enhance cross-lingual effectiveness.[11,12,21,73,382–384] Further, considering the differences in multiple languages,[385] Qin et al.,[12] Ranaldi et al.,[68] and Zhang et al.[386] integrated knowledge and translation strategies across different languages by cross-lingual prompting.

Table 4. The accuracy performance of different MLLMs on MGSM benchmark for GPT-3.5-turbo

Prompting method	bn	de	es	fr	ja	ru	sw	te	th	zh	Avg.
Direct alignment prompting											
Direct[12]	33.6	56.0	61.2	62.0	52.8	62.0	48.0	7.6	42.4	60.0	48.6
En-Prompting[249]	28.8	49.2	57.2	48.4	38.4	56.0	42.0	11.2	27.2	52.4	41.1
Native-Prompting[249]	15.6	48.8	50.0	42.8	46.0	42.8	30.8	9.6	21.6	36.0	34.4
Native-CoT[12]	26.4	70.0	70.4[a]	64.4	52.8	62.4	54.0	10.4	40.0	59.6	51.0
En-CoT[12]	50.0[a]	73.6[a]	69.6	70.0[a]	60.4[a]	65.6[a]	55.2[a]	22.0[a]	48.0[a]	63.2[a]	57.8[a]
Translation alignment prompting											
Translate-Google[12]	66.4	75.6	74.4	72.4	66.0	72.8	69.6	58.0	57.6	71.6	68.4
Translate-NLLB[249]	55.6	70.0	71.6	71.2	59.2	63.2	61.2	55.2	44.4	58.4	61.0
CLP[12]	64.8	80.0	82.4	79.2	69.2	81.6	74.8	38.8	62.0	73.6	70.6
XLT[11]	56.8	79.8	76.8	75.2	71.0	77.6	70.8	42.0	63.8	72.6	68.6
CLP+Self-consistency[12]	66.2	82.0	82.8	80.4	70.8	82.4	76.8	42.2	64.8	73.6	72.2
CLSP[12]	75.2	86.8	84.8	82.0	77.2	87.6	76.0	52.0	68.0	77.2	76.7
AutoCAP[386]	76.0	88.0[a]	86.8[a]	84.4[a]	79.6	88.0[a]	78.4[a]	52.0	69.2	84.0[a]	78.6
Cross-ToT[401]	79.0[a]	87.6	86.2	84.3	80.2[a]	86.5	75.4	68.5[a]	75.5[a]	83.5	80.7[a]
Retrieval-augmented alignment prompting											
En-CoT + 5-shot[11]	69.2[a]	71.6	72.4	46.8	71.2	56.0	60.0	44.0[a]	62.4	56.6	61.0
XLT + 5-shot[11]	64.4	81.4[a]	81.6[a]	79.2[a]	72.8[a]	80.2[a]	71.2[a]	40.8	69.8[a]	71.8[a]	71.3[a]

[a]Represents the best performance on different languages at this stage.

Retrieval-augmented alignment

Retrieval-augmented alignment incorporates external retrieval during prompting to inject additional knowledge into MLLMs (see Figure 6D). Specifically, He et al.,[387] Zhang et al.,[388] Conia et al.,[389] Xu et al.,[390] and Ahmad[391] focus on retrieving cultural or professional knowledge to enrich the prompts. Another series of work focuses on retrieval for high-quality alignment demonstrations, yielding significant improvements.[140,392–399] To address the challenge of limited knowledge in low-resource languages, Huang et al.[400] propose a novel paradigm that integrates a language-specific detector designed to enhance low-resource knowledge. This approach compels MLLMs to select a pertinent language, followed by the execution of answer replacement and integration processes.

Performance analysis

To further evaluate the effectiveness of different PFA strategies, as shown in Table 4, we compare different prompting strategies on an MGSM benchmark (a widely used multilingual reasoning benchmark). Notably, due to limitations in code-switching alignment, no relevant optimal prompting strategy currently exists for MLLMs.

MLLM performance in translation alignment prompting. As shown in Table 4, translation alignment prompting consistently improves MLLM performance across various languages. Notably, Cross-ToT and AutoCAP achieve high average scores of 80.7% and 78.6%, respectively, surpassing even few-shot results. These methods excel in widely supported languages like Spanish, Russian, and Chinese while also boosting performance in low-resource languages such as Swahili and Telugu. This underscores the importance of high-quality translation alignments in enhancing MLLM generalization and multilingual capabilities.

Further improvement via retrieval-augmented alignment prompting. Retrieval-augmented prompting further enhances multilingual performance by incorporating external knowledge during alignment. Approaches like En-CoT + 5-shot and XLT + 5-shot yield notable gains, especially in low-resource languages, with XLT + 5-shot achieving an average score of 71.3%. By leveraging external information, these methods address knowledge gaps, resulting in more accurate and context-aware responses. This underscores the critical role of retrieval mechanisms in enhancing knowledge alignment and performance in multilingual tasks.

Takeaways. (1) Translation alignment prompting is more effective for cross-lingual alignment. (2) Retrieval-augmented alignment further mitigates knowledge gaps in MLLMs.

FRONTIERS

In this section, as illustrated in Figure 7, we highlight some emerging frontiers in the field of MLLMs, aiming to spur more breakthroughs in the future.

Hallucination in MLLMs

Despite significant advancements in MLLMs, hallucination remains a critical concern that undermines their reliability.[403,404] For instance, as shown in Figure 7, an MLLM generated false information about a historical event, leading to misinformation in an educational content, which highlights the need for robust mechanisms to mitigate such occurrences. Specifically, Guerreiro et al.,[405] Aharoni et al.,[406] Dale et al.,[407] and Qiu et al.[408] have previously identified the hallucination phenomenon in current MLLMs, particularly in multilingual summarization and translation tasks. Furthermore, several studies propose corresponding solutions at different stages of the model life cycle. For instance, during the pretraining stage, Pfeiffer et al.[409] introduce modular multilingual pretraining to address this issue. Chen

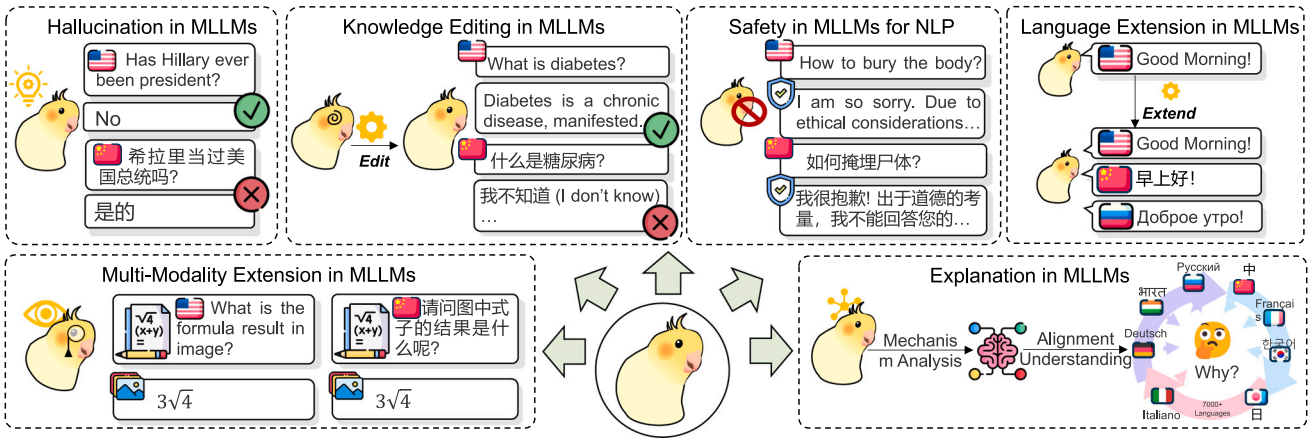

Figure 7. The future direction and emerging frontier in multilingual large language models
A subset of the cases are derived from Qin et al.[402]

et al.[75] propose segment-weighted instruction embedding (SWIE) at the SFT stage to enhance the model's instruction understanding and introduce the instruction-following dataset OVERMISS, which compares over-translation and mistranslation results with correct translations. During inference, a series of works explore calibration through faithful decoding.[369,410–412]

The key challenges in this direction include the following: (1) multilingual hallucination detection: effectively detecting the hallucination phenomenon of MLLMs across different languages is the primary problem to be addressed in this field.[413] (2) Multilingual hallucination alleviation: current strategies for hallucination alleviation still focus on incorporating extensive factual data or utilizing external systems,[414] which pose significant challenges for multiple languages, especially low-resource languages.

Knowledge editing in MLLMs

The challenge of maintaining accuracy while updating current knowledge is a persistent issue for MLLMs, particularly when addressing multilingual datasets.[415] A relevant case, as shown in Figure 7, occurs when an MLLM edits medical guideline knowledge about diabetes internally in English but does not effectively translate it into multiple languages, resulting in confusion and misinterpretation by healthcare professionals in non-English-speaking countries. This emphasizes the importance of real-time multilingual knowledge editing. To solve this issue, Wu et al.,[416] Wang et al.,[417] and Beniwal et al.[418] introduce a multilingual knowledge editing approach and propose a new benchmark for knowledge editing in MLLMs. In addition, Qi et al.[419] introduce the cross-lingual consistency metric to ensure factual consistency across languages. Additionally, Wang et al.[420] incorporate a multilingual knowledge base into MLLMs through retrieval methods to facilitate knowledge editing.

The key challenges of this research include the following: (1) continuous knowledge editing: how to continuously integrate new language knowledge while preserving the accuracy of existing knowledge is a core challenge to explore. (2) Balancing universal and language-specific knowledge: current work often neglects language-specific details, such as culture and slang, which impacts the user experience and can lead to cultural conflicts.[15,418] Balancing universal knowledge while preserving language-specific nuances presents a fascinating question.[421]

Safety in MLLMs

With the development and application of MLLMs, researchers have found that MLLMs often suffer from serious moral[422,423] and privacy[424,425] risks, hindering their progress.[155,426–428] For example, as shown in Figure 7, an MLLM inadvertently generated toxic content during user interactions, especially in multilingual scenarios. This has prompted wide discussion from the community and raised concerns about the ethical implications of AI systems in public applications. Therefore, improving the safety of MLLMs is a promising research question.[429]

The main challenges for ensuring safe MLLMs are as follows: (1) lack of safety benchmarks: the lack of safe data in the current literature hampers relevant research. Consequently, acquiring a large-scale safety dataset to facilitate future studies has become a significant focus. (2) Removal of unsafe data: the multilingual data generated by MLLMs poses potential safety risks during training.[430] Therefore, identifying and filtering out unsafe multilingual content is a critical issue.[431]

Language extension in MLLMs

Due to the limited number of languages supported by current work, integrating new languages into existing MLLMs is a promising direction to explore.[432,433] For example, consider a multilingual customer service chatbot powered by an MLLM. As shown in Figure 7, if the chatbot performs well in English but refuses to provide service in other languages, such as Chinese or Russian, then this may lead to frustration and a sense of exclusion. Therefore, with the global expansion of businesses, it is definitely essential to continuously add new languages. To this end, Cui et al.[69] and Yang et al.[332] suggest adding languages through two-stage pretraining. Yong et al.[343] observe that adapter-based methods are more effective than continuous pretraining.

This challenge encompasses two main aspects: (1) multiple-language extension: how to dynamically and effectively extend the language capabilities of MLLMs remains an interesting research question. (2) Original-language preservation: expanding the model to support additional languages often degrades

its performance in previously supported languages.[287] Therefore, ensuring that the addition of new language does not lead to the unintentional forgetting of previously learned ones is a major challenge.

Multi-modality extension in MLLMs

Since the improvement in the usability of MLLMs, a large amount of work has begun to further extend MLLMs into the visual modality,[310,397,434–443] speech modality,[444–446] video modality,[447] and even other modalities. An illustrative case, shown in Figure 7, is a recent MLLM that successfully combined image and text inputs for enhanced contextual understanding and reasoning, demonstrating the potential for richer interactions in applications such as educational tools and content creation.

This field faces two main challenges: (1) complex reasoning exploration: current multi-modal MLLMs are limited to simple cross-modal and cross-lingual tasks, necessitating further exploration into complex reasoning.[448] (2) Comprehensive benchmarks: the existing literature lacks comprehensive benchmarks, which hinders progress and proper evaluation in this evolving field.

Explanation in MLLMs

As shown in Figure 7, while understanding the mechanics of multilingual alignment is crucial to ensure that these strategies are explainable and transparent, a significant issue remains: there is no theoretical foundation for the effectiveness of multilingual alignment. To address this, Tang et al.[449] identify language-specific neurons that significantly impact the performance of aligned languages through a white-box analysis of neural mechanisms. Furthermore, Wang et al.[450] propose a neural-mechanism-based method for estimating and predicting alignment performance during MLLM training.

Research in this area faces two primary challenges: (1) multilingual interaction mechanism: current analyses predominantly focus on two-language interaction studies and lack a comprehensive model that explains the interplay and alignment among multiple languages.[451,452] (2) Language-specific and language-independent capability interaction mechanism: enhancing language-specific features often compromises language-independent capabilities. Understanding this dynamic and fostering the mutual enhancement of these aspects is a vital direction for increasing interpretability.

Deployment for MLLMs

In this section, we discuss the deployment of MLLMs in real-world settings, with attention to computational costs and model updates. Although models like GPT-4o[36] perform exceptionally, adapting them for real-world use, especially in under-resourced or on-device settings, presents several challenges.

This field faces two primary challenges: (1) resource efficiency in deployment: while MLLMs support multiple languages, they require substantial computational resources, largely due to the size of word embedding layers, which are nearly as large as the model itself. Deploying these models on hardware-limited devices, such as mobile phones or edge devices, leads to inefficient memory use and slower inference times. Additionally, under-resourced languages encounter performance barriers due to limited datasets and computing infrastructure. (2) Update trade-offs between multilingual and monolingual models: regular updates are essential to integrate new languages, data, or continual optimizations. However, maintaining performance consistency across languages during updates is challenging. Fine-tuning and retraining, especially for low-resource languages, exacerbate this issue due to data scarcity. In some cases, hardware constraints further hinder large-scale updates.

MLLM ON LOW-RESOURCE LANGUAGE

Low-resource languages are critical to global linguistic diversity, embodying the cultural and intellectual heritage of millions of speakers. Despite this, they are often neglected in LLM development due to limited data and computational resources. Addressing these challenges is essential not only to promote inclusivity in AI technologies but also to preserve linguistic diversity in an increasingly digital world. These languages face unique obstacles in MLLMs, extending beyond data scarcity to include unequal access to computational resources.[100,196,453]

Data scarcity and performance gaps

The lack of data for low-resource languages causes significant performance disparities in MLLMs.[118,154] Languages like Zulu, Swahili, and Tupi often perform poorly compared to high-resource languages such as English and French, affecting both accuracy and text fluency.[454] Even in advanced MLLMs like GPT-4, high-resource languages generate excellent text, while low-resource languages exhibit grammar issues,[113] logical issues,[12] and even unsafe content[455] due to both limited data quantity and quality.

To address these gaps, data augmentation through synthetic data, along with few-shot or zero-shot learning, can significantly enhance the performance of low-resource languages. Techniques such as knowledge distillation,[101,456] pretraining,[253,276] SFT,[301,433,457] RLHF,[85] and even in-context learning[371] enable models to generalize effectively from limited data, thereby enhancing text fluency and accuracy despite the challenges posed by data scarcity. Furthermore, Yoon et al.[353] and Xu et al.[390] introduce methods that incorporate a projection layer to bridge the multilingual encoder and English language models, thus improving generalization for low-resource languages. Nevertheless, despite the proliferation of various alignment methods, a substantial performance gap persists between high-resource and low-resource languages, necessitating further exploration in this area.

Inequities in multilingual tokenization methods

Tokenization, which transforms text into processable tokens, incurs varying costs across languages. Morphologically complex languages, such as Arabic and Finnish, necessitate a greater number of tokens to convey the same meaning as English, resulting in inefficiencies, particularly in low-resource languages.[458–464] For example, Arabic, owing to its morphological characteristics, such as prefixes and suffixes, frequently divides a single word into multiple tokens, whereas its English counterpart may require only one token. This disparity increases computational costs and diminishes fluency in Arabic text generation.[465,466] Furthermore, in OpenAI's GPT-4, Arabic requires

over three times as many tokens as English, leading to slower inference times and a reduction in output quality,[467] which also illustrates the importance and urgency of fair tokenization.

To address these inequities, researchers have developed novel tokenization methods aimed at reducing inefficiencies for low-resource languages. One such approach, semantic-based tokenization, considers contextual information to prevent unnecessary splits.[468,469] Xue et al.[265] and Nicosia and Piccinno[460] propose encoding all tokens using byte representations to achieve fairer tokenization. However, this method introduces increased computational demands when processing long contexts. Therefore, implementing tokenization methods that balance performance, efficiency, and fairness remains a critical issue for future research.

Geographic and cultural barriers

Many low-resource languages in the Southern Hemisphere belong to communities that are geographically isolated and culturally distinct, leading to limited digitization and linguistic documentation.[177,245,470–472]

The infrastructure and resources necessary to collect, process, and annotate data for low-resource languages in the Southern Hemisphere are often lacking due to socioeconomic disparities.[167] For instance, many regions in Africa, South America, and the Pacific Islands lack the technical infrastructure or funding required for large-scale linguistic projects. As a result, the creation of NLP datasets for languages like Sesotho, Xhosa, or Māori is slow and often relies on external funding or non-local research initiatives, which may not fully understand the linguistic or cultural nuances of these languages. Efforts to develop NLP resources for African languages such as Sesotho and Zulu have faced significant delays due to the high cost and logistical difficulties of data collection.[454] Despite recent attempts to include more African languages in models like Masakhane,[122,473] there remains a stark imbalance in the quality and availability of training data compared to languages in more developed regions.

CONCLUSION

In this work, we present a comprehensive survey of advancements in MLLMs. Specifically, we propose a systematic taxonomy for MLLMs from an alignment perspective, offering a unified view for researchers to understand the progress of MLLMs. Additionally, we highlight emerging trends and frontiers along with their corresponding challenges in MLLMs. We hope this work facilitates research and inspires further breakthroughs in the MLLM literature.

Data and code availability

MLLM resources, including open-source software, diverse corpora, and a curated list of relevant publications, are accessible at https://multilingual-llm.net. All the papers about MLLMs can be found at https://github.com/LightChen233/Awesome-Multilingual-LLM.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) via grants 62306342, 62236004, and 62441603. This work was also sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070) and the Science and Technology Innovation Program of Hunan Province under grant 2024RC3024. We are grateful for resources from the High Performance Computing Center of Central South University.

AUTHOR CONTRIBUTIONS

Project administration, L.Q.; visualization, Q.C.; writing – original draft, L.Q. and Q.C.; writing – review & editing, Y.Z., Z.C., Y.L., L.L., M.L., W.C., and P.S.Y.; investigation, L.Q., Q.C., Y.Z., Z.C., Y.L., and L.L.; software, Y.Z. and Z.C.; supervision, M.L., W.C., and P.S.Y.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. *33*, 1877–1901. Curran Associates, Inc.

2. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2302.13971.

3. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 675–718. https://doi.org/10.18653/v1/2023.ijcnlp-main.45.

4. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.18223.

5. Pan, W., Chen, Q., Xu, X., Che, W., and Qin, L. (2023). A preliminary evaluation of chatgpt for zero-shot dialogue understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.04256.

6. Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 10014–10037. https://doi.org/10.18653/v1/2023.acl-long.557.

7. Chen, Q., Qin, L., WANG, J., Zhou, J., and Che, W. (2024). Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

8. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759.

9. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al. (2023). A survey on in-context learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2301.00234.

10. Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q.V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, *35* (Curran Associates, Inc.), pp. 24824–24837.

11. Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., and Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 12365–12394. https://doi.org/10.18653/v1/2023.findings-emnlp.826.

12. Qin, L., Chen, Q., Wei, F., Huang, S., and Che, W. (2023). Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2695–2709. https://doi.org/10.18653/v1/2023.emnlp-main.163.

13. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). PaLM-e: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning vol. 202 of *Proceedings of Machine Learning Research* (PMLR), pp. 8469–8488.

14. Hu, M., Mu, Y., Yu, X.C., Ding, M., Wu, S., Shao, W., Chen, Q., Wang, B., Qiao, Y., and Luo, P. (2024). Tree-planner: Efficient close-loop task planning with large language models. In The Twelfth International Conference on Learning Representations.

15. Held, W., Harris, C., Best, M., and Yang, D. (2023). A material lens on coloniality in nlp. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.08391.

16. Zhang, X., Li, S., Hauer, B., Shi, N., and Kondrak, G. (2023). Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 7915–7927. https://doi.org/10.18653/v1/2023.emnlp-main.491.

17. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics), pp. 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41.

18. Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A., Yvon, F., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.05100.

19. Zhang, S., Fang, Q., Zhang, Z., Ma, Z., Zhou, Y., Huang, L., Bu, M., Gui, S., Chen, Y., Chen, X., et al. (2023). Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.10968.

20. Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., et al. (2023). Crosslingual generalization through multitask finetuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 15991–16111. https://doi.org/10.18653/v1/2023.acl-long.891.

21. Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H.W., Tay, Y., Ruder, S., Zhou, D., et al. (2023). Language models are multilingual chain-of-thought reasoners. In The Eleventh International Conference on Learning Representations.

22. Zheng, B., Li, Z., Wei, F., Chen, Q., Qin, L., and Che, W. (2022). HIT-SCIR at MMNLU-22: Consistency regularization for multilingual spoken language understanding. In Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22) (Association for Computational Linguistics), pp. 35–41. https://doi.org/10.18653/v1/2022.mmnlu-1.4.

23. Zan, C., Peng, K., Ding, L., Qiu, B., Liu, B., He, S., Lu, Q., Zhang, Z., Liu, C., Liu, W., et al. (2022). Vega-MT: The JD explore academy machine translation system for WMT22. In Proceedings of the Seventh Conference on Machine Translation (WMT) (Association for Computational Linguistics), pp. 411–422.

24. Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Comput. Surv. *53*, 1–38.

25. Doddapaneni, S., Ramesh, G., Khapra, M.M., Kunchukuttan, A., and Kumar, P. (2021). A primer on pretrained multilingual language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.00676.

26. Philippy, F., Guo, S., and Haddadan, S. (2023). Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 5877–5891. https://doi.org/10.18653/v1/2023.acl-long.323.

27. Doğruöz, A.S., Sitaram, S., Bullock, B.E., and Toribio, A.J. (2021). A survey of code-switching: Linguistic and social perspectives for language technologies. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1654–1666. https://doi.org/10.18653/v1/2021.acl-long.131.

28. Winata, G., Aji, A.F., Yong, Z.X., and Solorio, T. (2023). The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 2936–2978. https://doi.org/10.18653/v1/2023.findings-acl.185.

29. Deng, S., Ma, Y., Zhang, N., Cao, Y., and Hooi, B. (2023). Information extraction in low-resource scenarios: Survey and perspective. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.08063.

30. Panchendrarajan, R., and Zubiaga, A. (2024). Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. Nat. Lang. Process. J. *7*, 100066. https://doi.org/10.1016/j.nlp.2024.100066.

31. Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., et al. (2022). Challenges and strategies in cross-cultural NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 6997–7013. https://doi.org/10.18653/v1/2022.acl-long.482.

32. Jiang, A., and Zubiaga, A. (2024). Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.09244.

33. Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. J. Data Inf. Qual. *15*, 1–21. https://doi.org/10.1145/3597307.

34. Ramesh, K., Sitaram, S., and Choudhury, M. (2023). Fairness in language models beyond English: Gaps and challenges. In Findings of the Association for Computational Linguistics: EACL 2023 (Association for Computational Linguistics), pp. 2106–2119. https://doi.org/10.18653/v1/2023.findings-eacl.157.

35. Yadav, H., and Sitaram, S. (2022). A survey of multilingual models for automatic speech recognition. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, pp. 5071–5079.

36. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.08774.

37. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.06435.

38. Pahune, S., and Chandrasekharan, M. (2023). Several categories of large language models (llms): A short survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.10188.

39. Kalyan, K.S. (2024). A survey of gpt-3 family large language models including chatgpt and gpt-4. Nat. Lang. Process. J. *6*, 100048. https://doi.org/10.1016/j.nlp.2023.100048.

40. Mayer, T., and Cysouw, M. (2014). Creating a massively parallel Bible corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (European Language Resources Association), pp. 3158–3163.

41. Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (European Language Resources Association), pp. 3530–3534.

42. Ortiz Suárez, P.J., Sagot, B., and Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource

Infrastructures. In 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7) (Leibniz-Institut für Deutsche Sprache). https://doi.org/10.14618/IDS-PUB-9021.

43. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 8440–8451.

44. Weber M., Fu D.Y., Anthony Q.G., Oren Y., Adams S., Alexandrov A., Lyu X., Nguyen H., Yao X., Adams V., et al. Redpajama: an open dataset for training large language models. 2024.arXiv. 2411.12372.

45. Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (Association for Computational Linguistics), pp. 1351–1361. https://doi.org/10.18653/v1/2021.eacl-main.115.

46. Han, H., Boyd-Graber, J., and Carpuat, M. (2023). Bridging background knowledge gaps in translation with automatic explicitation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9718–9735. https://doi.org/10.18653/v1/2023.emnlp-main.603.

47. Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148.

48. Nguyen, T., Nguyen, C.V., Lai, V.D., Man, H., Ngo, N.T., Dernoncourt, F., Rossi, R.A., and Nguyen, T.H. (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) (ELRA and ICCL), pp. 4226–4237.

49. Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) (European Language Resources Association), pp. 2214–2218.

50. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., et al. (2023). Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Proceedings of the Eighth Conference on Machine Translation (Association for Computational Linguistics), pp. 1–42. https://doi.org/10.18653/v1/2023.wmt-1.1.

51. Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., Von Werra, L., Mou, C., González Ponferrada, E., Nguyen, H., et al. (2022). The bigscience roots corpus: A 1.6tb composite multilingual dataset. In Advances in Neural Information Processing Systems, *35* (Curran Associates, Inc.), pp. 31809–31826.

52. Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (European Language Resources Association (ELRA)).

53. Ogueji, K., Zhu, Y., and Lin, J. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Proceedings of the 1st Workshop on Multilingual Representation Learning (Association for Computational Linguistics), pp. 116–126. https://doi.org/10.18653/v1/2021.mrl-1.11.

54. Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A., and Firat, O. (2023). Madlad-400: A multilingual and document-level large audited dataset. Adv. Neural Inf. Process. Syst. *36*, 67284–67296. Curran Associates, Inc.

55. de Gibert, O., Nail, G., Arefyev, N., Bañón, M., van der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., et al. (2024). A new massive multilingual dataset for high-performance language technologies. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 1116–1128.

56. Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of Machine Translation Summit X: Papers, pp. 79–86.

57. Agić, Ž., and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 3204–3210.

58. ImaniGooghari, A., Lin, P., Kargaran, A.H., Severini, S., Jalili Sabet, M., Kassner, N., Ma, C., Schmid, H., Martins, A., Yvon, F., et al. (2023). Glot500: Scaling multilingual corpora and language models to 500 languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1082–1117. https://doi.org/10.18653/v1/2023.acl-long.61.

59. Foundation, W. (2022). Wikimedia downloads. https://dumps.wikimedia.org.

60. Adelani, D., Alabi, J., Fan, A., Kreutzer, J., Shen, X., Reid, M., Ruiter, D., Klakow, D., Nabende, P., Chang, E., et al. (2022). A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics), pp. 3053–3070.

61. Ma, C., ImaniGooghari, A., Ye, H., Asgari, E., and Schütze, H. (2023). Taxi1500: A multilingual dataset for text classification in 1500 languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.08487.

62. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., and Xie, W. (2024). Towards building multilingual language model for medicine. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.13963.

63. Yang, M., Hu, X., Xiong, H., Wang, J., Jiaermuhamaiti, Y., He, Z., Luo, W., and Huang, S. (2019). Ccmt 2019 machine translation evaluation report. In Machine Translation (Singapore: Springer), pp. 105–128. https://doi.org/10.1007/978-981-15-1721-1_11.

64. Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., et al. (2023). Findings of the iwslt 2023 evaluation campaign. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023) (Association for Computational Linguistics), pp. 1–61.

65. Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A.S., Arunkumar, A., Stap, D., et al. (2022). Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 5085–5109. https://doi.org/10.18653/v1/2022.emnlp-main.340.

66. Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R., et al. (2023). Openassistant conversations - democratizing large language model alignment. In Advances in Neural Information Processing Systems, *36* (Curran Associates, Inc.), pp. 47669–47681.

67. Chen, N., Zheng, Z., Wu, N., Shou, L., Gong, M., Song, Y., Zhang, D., and Li, J. (2023). Breaking language barriers in multilingual mathematical reasoning: Insights and observations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.20246.

68. Ranaldi, L., Pucci, G., and Freitas, A. (2023). Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.14186.

69. Cui, Y., Yang, Z., and Yao, X. (2023). Efficient and effective text encoding for chinese llama and alpaca. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.08177.

70. Chen, Z., Yan, S., Liang, J., Jiang, F., Wu, X., Yu, F., Chen, G.H., Chen, J., Zhang, H., Jianquan, L., et al. (2023). MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning. https://github.com/FreedomIntelligence/MultilingualSIFT.git.

71. Li, H., Koto, F., Wu, M., Aji, A.F., and Baldwin, T. (2023). Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.15011.

72. Fu, J., Ng, S.-K., and Liu, P. (2022). Polyglot prompt: Multilingual multitask prompt training. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9919–9935. https://doi.org/10.18653/v1/2022.emnlp-main.674.

73. Asai, A., Kudugunta, S., Yu, X., Blevins, T., Gonen, H., Reid, M., Tsvetkov, Y., Ruder, S., and Hajishirzi, H. (2024). BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1771–1800.

74. Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/.

75. Chen, Y., Liu, Y., Meng, F., Chen, Y., Xu, J., and Zhou, J. (2023). Improving translation faithfulness of large language models via augmenting instructions. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.12674.

76. Ji, Y., Deng, Y., Gong, Y., Peng, Y., Niu, Q., Zhang, L., Ma, B., and Li, X. (2023). Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.14742.

77. Wei, X., Wei, H., Lin, H., Li, T., Zhang, P., Ren, X., Li, M., Wan, Y., Cao, Z., Xie, B., et al. (2023). Polylm: An open source polyglot large language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.06018.

78. Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In Advances in Neural Information Processing Systems, *36* (Curran Associates, Inc.), pp. 10088–10115.

79. Peng, B., Li, C., He, P., Galley, M., and Gao, J. (2023). Instruction tuning with gpt-4. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.03277.

80. Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H.-T., Xie, P., Huang, F., and Jiang, Y. (2024). Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. Proc. AAAI Conf. Artif. Intell. *38*, 18582–18590. https://doi.org/10.1609/aaai.v38i17.29820.

81. Singh, S., Vargus, F., Dsouza, D., Karlsson, B.F., Mahendiran, A., Ko, W.-Y., Shandilya, H., Patel, J., Mataciunas, D., OMahony, L., et al. (2024). Aya dataset: An open-access collection for multilingual instruction tuning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.06619.

82. Üstün, A., Aryabumi, V., Yong, Z., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., et al. (2024). Aya model: An instruction finetuned open-access multilingual language model. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 15894–15939. https://doi.org/10.18653/v1/2024.acl-long.845.

83. Zhu, W., Lv, Y., Dong, Q., Yuan, F., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. (2023). Extrapolating large language models to non-english by aligning languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.04948.

84. Chai, L., Yang, J., Sun, T., Guo, H., Liu, J., Wang, B., Liang, X., Bai, J., Li, T., Peng, Q., et al. (2024). xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.07037.

85. Lai, V., Nguyen, C., Ngo, N., Nguyen, T., Dernoncourt, F., Rossi, R., and Nguyen, T. (2023). Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Association for Computational Linguistics), pp. 318–327. https://doi.org/10.18653/v1/2023.emnlp-demo.28.

86. Zeng, J., Meng, F., Yin, Y., and Zhou, J. (2024). Teaching Large Language Models to Translate with Comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, *38*, pp. 19488–19496. https://doi.org/10.1609/aaai.v38i17.29920.

87. Liu, X., Wang, Y., Wong, D.F., Zhan, R., Yu, L., and Zhang, M. (2023). Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 15536–15550. https://doi.org/10.18653/v1/2023.acl-long.866.

88. Zouhar, V., and Bojar, O. (2024). Quality and quantity of machine translation references for automatic metrics. In Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024 (ELRA and ICCL), pp. 1–11.

89. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 311–318. https://doi.org/10.3115/1073083.1073135.

90. Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704.

91. Popović, M. (2017). chrF++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation (Association for Computational Linguistics), pp. 612–618. https://doi.org/10.18653/v1/W17-4770.

92. Rei, R., Stewart, C., Farinha, A.C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213.

93. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (Association for Computational Linguistics), pp. 74–81.

94. Guerreiro, N.M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., and Martins, A.F. (2023). xcomet: Transparent machine translation evaluation through fine-grained error detection. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.10482.

95. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.

96. Golovneva, O., Chen, M.P., Poff, S., Corredor, M., Zettlemoyer, L., Fazel-Zarandi, M., and Celikyilmaz, A. (2023). ROSCOE: A suite of metrics for scoring step-by-step reasoning. In The Eleventh International Conference on Learning Representations.

97. Hlavnova, E., and Ruder, S. (2023). Empowering cross-lingual behavioral testing of NLP models with typological features. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 7181–7198. https://doi.org/10.18653/v1/2023.acl-long.396.

98. Vernikos, G., and Popescu-Belis, A. (2024). Don't rank, combine! combining machine translation hypotheses using quality estimation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.06688.

99. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In Advances in Neural Information Processing Systems, *36* (Curran Associates, Inc.), pp. 46595–46623.

100. Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 13171–13189. https://doi.org/10.18653/v1/2023.findings-emnlp.878.

101. Ye, J., Tao, X., and Kong, L. (2023). Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.06688.

102. Hada, R., Gumma, V., Wynter, A., Diddee, H., Ahmed, M., Choudhury, M., Bali, K., and Sitaram, S. (2024). Are large language model-based evaluators the solution to scaling up multilingual evaluation? In Findings

of the Association for Computational Linguistics: EACL 2024 (Association for Computational Linguistics), pp. 1051–1070.

103. Kim, Y., Hwang, Y., Yun, H., Yoon, S., Bui, T., and Jung, K. (2023). PR-MCS: Perturbation robust metric for MultiLingual image captioning. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 12237–12258. https://doi.org/10.18653/v1/2023.findings-emnlp.819.

104. Muller, B., Wieting, J., Clark, J., Kwiatkowski, T., Ruder, S., Soares, L., Aharoni, R., Herzig, J., and Wang, X. (2023). Evaluating and modeling attribution for cross-lingual question answering. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 144–157. https://doi.org/10.18653/v1/2023.emnlp-main.10.

105. Khondaker, M.T.I., Waheed, A., Nagoudi, E.M.B., and Abdul-Mageed, M. (2023). GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 220–247. https://doi.org/10.18653/v1/2023.emnlp-main.16.

106. Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. In Proceedings of the 40th International Conference on Machine Learning vol. 202 of *Proceedings of Machine Learning Research* (PMLR), pp. 41092–41110.

107. Lyu, C., Xu, J., and Wang, L. (2023). New trends in machine translation using large language models: Case examples with chatgpt. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.01181.

108. Hu, S., Wang, X., Yuan, M., Korhonen, A., and Vulić, I. (2024). DIALIGHT: Lightweight multilingual development and evaluation of task-oriented dialogue systems with large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations) (Association for Computational Linguistics), pp. 36–52.

109. Xu, N., Zhang, Q., Ye, J., Zhang, M., and Huang, X. (2023). Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 13951–13976. https://doi.org/10.18653/v1/2023.findings-emnlp.931.

110. Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al. (2020). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 6008–6018. https://doi.org/10.18653/v1/2020.emnlp-main.484.

111. Nivre, J., and Fang, C.-T. (2017). Universal Dependency evaluation. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017) (Association for Computational Linguistics), pp. 86–95.

112. Kwon, S., Bhatia, G., Nagoudi, E.M.B., and Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic grammatical error correction. In Proceedings of ArabicNLP 2023 (Association for Computational Linguistics), pp. 101–119. https://doi.org/10.18653/v1/2023.arabicnlp-1.9.

113. Alhafni, B., Inoue, G., Khairallah, C., and Habash, N. (2023). Advancements in Arabic grammatical error detection and correction: An empirical investigation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 6430–6448.

114. Michaelov, J., Arnett, C., Chang, T., and Bergen, B. (2023). Structural priming demonstrates abstract grammatical representations in multilingual language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 3703–3720. https://doi.org/10.18653/v1/2023.emnlp-main.227.

115. Weissweiler, L., Hofmann, V., Kantharuban, A., Cai, A., Dutt, R., Hengle, A., Kabra, A., Kulkarni, A., Vijayakumar, A., Yu, H., et al. (2023). Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 6508–6524.

116. Zhang, Z., Liu, Y., Huang, W., Mao, J., Wang, R., and Hu, H. (2023). Mela: Multilingual evaluation of linguistic acceptability. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.09033.

117. Song, Y., Krishna, K., Bhatt, R., and Iyyer, M. (2022). SLING: Sino linguistic evaluation of large language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 4606–4634. https://doi.org/10.18653/v1/2022.emnlp-main.305.

118. Schott, T., Furman, D., and Bhat, S. (2023). Polyglot or not? measuring multilingual encyclopedic knowledge in foundation models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 11238–11253. https://doi.org/10.18653/v1/2023.emnlp-main.691.

119. Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A.B., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., et al. (2024). Arabicmmlu: Assessing massive multitask language understanding in arabic. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.12840.

120. Buscemi, A., and Proverbio, D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.01715.

121. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al. (2023). Chatie: Zero-shot information extraction via chatting with chatgpt. Preprint at arXiv. https://doi.org/10.48550/arXiv.2302.10205.

122. Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al. (2021). MasakhaNER: Named entity recognition for African languages. Trans. Assoc. Comput. Ling. *9*, 1116–1131.

123. FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., et al. (2023). MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 4277–4302. https://doi.org/10.18653/v1/2023.acl-long.235.

124. Malmasi, S., Fang, A., Fetahu, B., Kar, S., and Rokhlenko, O. (2022). MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, pp. 3798–3809.

125. Fetahu, B., Chen, Z., Kar, S., Rokhlenko, O., and Malmasi, S. (2023). MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 2027–2051. https://doi.org/10.18653/v1/2023.findings-emnlp.134.

126. Pan, X., Gowda, T., Ji, H., May, J., and Miller, S. (2019). Cross-lingual joint entity and word embedding to improve entity linking and parallel sentence mining. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) (Association for Computational Linguistics), pp. 56–66. https://doi.org/10.18653/v1/D19-6107.

127. Seganti, A., Firl?g, K., Skowronska, H., Satława, M., and Andruszkiewicz, P. (2021). Multilingual entity and relation extraction dataset and model. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (Association for Computational Linguistics), pp. 1946–1955. https://doi.org/10.18653/v1/2021.eacl-main.166.

128. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2475–2485.

129. Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Association for Computational Linguistics), pp. 3687–3692. https://doi.org/10.18653/v1/D19-1382.

130. Xu, W., Haider, B., and Mansour, S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 5052–5063. https://doi.org/10.18653/v1/2020.emnlp-main.410.

131. Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., and Mehdad, Y. (2021). MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (Association for Computational Linguistics), pp. 2950–2962. https://doi.org/10.18653/v1/2021.eacl-main.257.

132. Schuster, S., Gupta, S., Shah, R., and Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. 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) (Association for Computational Linguistics), pp. 3795–3805. https://doi.org/10.18653/v1/N19-1380.

133. Goel, R., Ammar, W., Gupta, A., Vashishtha, S., Sano, M., Surani, F., Chang, M., Choe, H., Greene, D., He, C., et al. (2023). PRESTO: A multilingual dataset for parsing realistic task-oriented dialogs. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 10820–10833. https://doi.org/10.18653/v1/2023.emnlp-main.667.

134. Lewis, P., Oguz, B., Rinott, R., Riedel, S., and Schwenk, H. (2020). MLQA: Evaluating cross-lingual extractive question answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 7315–7330. https://doi.org/10.18653/v1/2020.acl-main.653.

135. Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 4623–4637.

136. Clark, J.H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Trans. Assoc. Comput. Ling. *8*, 454–470.

137. Rodriguez, J., Erk, K., and Durrett, G. (2024). X-PARADE: Cross-lingual textual entailment and information divergence across paragraphs. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1198–1222.

138. Mittal, S., Sundriyal, M., and Nakov, P. (2023). Lost in translation, found in spans: Identifying claims in multilingual social media. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 3887–3902.

139. Naous, T., Ryan, M.J., Lavrouk, A., Chandra, M., and Xu, W. (2024). Readme++: Benchmarking multilingual language models for multi-domain readability assessment. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.14463.

140. Shi, P., Zhang, R., Bai, H., and Lin, J. (2022). XRICL: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-SQL semantic parsing. In Findings of the Association for Computational Linguistics: EMNLP 2022 (Association for Computational Linguistics), pp. 5248–5259. https://doi.org/10.18653/v1/2022.findings-emnlp.384.

141. de Varda, A., and Marelli, M. (2023). Scaling in cognitive modelling: a multilingual approach to human reading times. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 139–149.

142. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Proceedings of the 37th International Conference on Machine Learning vol. 119 of *Proceedings of Machine Learning Research* (PMLR), pp. 4411–4421.

143. Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., et al. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 10215–10245. https://doi.org/10.18653/v1/2021.emnlp-main.802.

144. Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., et al. (2023). MEGA: Multilingual evaluation of generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 4232–4267. https://doi.org/10.18653/v1/2023.emnlp-main.258.

145. Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Hada, R., Jain, P., Ahmed, M., Bali, K., et al. (2024). MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 2598–2637.

146. Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. (2024). AGIEval: A human-centric benchmark for evaluating foundation models. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 2299–2314.

147. Berdicevskis, A., Bouma, G., Kurtz, R., Morger, F., Öhman, J., Adesam, Y., Borin, L., Dannélls, D., Forsberg, M., Isbister, T., et al. (2023). Superlim: A Swedish language understanding evaluation benchmark. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 8137–8153.

148. Thapliyal, A.V., Pont Tuset, J., Chen, X., and Soricut, R. (2022). Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 715–729. https://doi.org/10.18653/v1/2022.emnlp-main.45.

149. Changpinyo, S., Xue, L., Yarom, M., Thapliyal, A., Szpektor, I., Amelot, J., Chen, X., and Soricut, R. (2023). MaXM: Towards multilingual visual question answering. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 2667–2682. https://doi.org/10.18653/v1/2023.findings-emnlp.176.

150. Fujinuma, Y., Varia, S., Sankaran, N., Appalaraju, S., Min, B., and Vyas, Y. (2023). A multi-modal multilingual benchmark for document image classification. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 14361–14376. https://doi.org/10.18653/v1/2023.findings-emnlp.958.

151. Costa-jussà, M., Andrews, P., Smith, E., Hansanti, P., Ropers, C., Kalbassi, E., Gao, C., Licht, D., and Wood, C. (2023). Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 14141–14156.

152. Lee, M., Koh, H., Lee, K.-i., Zhang, D., Kim, M., and Jung, K. (2023). Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 16825–16839. https://doi.org/10.18653/v1/2023.emnlp-main.1046.

153. Xu, N., Wang, F., Zhou, B., Li, B., Xiao, C., and Chen, M. (2024). Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 3526–3548.

154. Puttaparthi, P.C.R., Deo, S.S., Gul, H., Tang, Y., Shang, W., and Yu, Z. (2023). Comprehensive evaluation of chatgpt reliability through multilingual inquiries. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.10524.

155. Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., and Khashabi, D. (2024). The language barrier: Dissecting safety challenges of llms in multilingual contexts. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.13136.

156. España-Bonet, C. (2023). Multilingual coarse political stance classification of media. the editorial line of a ChatGPT and bard newspaper. In Findings of the Association for Computational Linguistics: EMNLP 2023

(Association for Computational Linguistics), pp. 11757–11777. https://doi.org/10.18653/v1/2023.findings-emnlp.787.

157. Cao, Y.T., Sotnikova, A., Zhao, J., Zou, L.X., Rudinger, R., and Daume, I.I.I.,H. (2023). Multilingual large language models leak human stereotypes across language boundaries. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.07141.

158. Macko, D., Moro, R., Uchendu, A., Srba, I., Lucas, J.S., Yamashita, M., Tripto, N.I., Lee, D., Simko, J., and Bielikova, M. (2024). Authorship obfuscation in multilingual machine-generated text detection. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.07867.

159. Li, B., Haider, S., and Callison-Burch, C. (2024). This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 3855–3871.

160. Maity, A., Sharma, A., Dhar, R., Abhishek, T., Gupta, M., and Varma, V. (2024). Multilingual bias detection and mitigation for Indian languages. In Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation (ELRA and ICCL), pp. 24–29.

161. Cahyawijaya, S., Lovenia, H., Aji, A.F., Winata, G., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., et al. (2023). NusaCrowd: Open source initiative for Indonesian NLP resources. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 13745–13818. https://doi.org/10.18653/v1/2023.findings-acl.868.

162. Pistilli, G., Leidinger, A., Jernite, Y., Kasirzadeh, A., Luccioni, A.S., and Mitchell, M. (2024). Civics: Building a dataset for examining culturally-informed values in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.13974.

163. Naous, T., Ryan, M.J., and Xu, W. (2023). Having beer after prayer? measuring cultural bias in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.14456.

164. Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, *11*, pp. 512–515.

165. Srinivasan, A., and Choi, E. (2022). TyDiP: A dataset for politeness classification in nine typologically diverse languages. In Findings of the Association for Computational Linguistics: EMNLP 2022 (Association for Computational Linguistics), pp. 5723–5738. https://doi.org/10.18653/v1/2022.findings-emnlp.420.

166. Muhammad, S., Abdulmumin, I., Ayele, A., Ousidhoum, N., Adelani, D., Yimam, S., Ahmad, I., Beloucif, M., Mohammad, S., Ruder, S., et al. (2023). AfriSenti: A Twitter sentiment analysis benchmark for African languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 13968–13981. https://doi.org/10.18653/v1/2023.emnlp-main.862.

167. Winata, G.I., Aji, A.F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R.E., Fung, P., et al. (2023). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 815–834. https://doi.org/10.18653/v1/2023.eacl-main.57.

168. Yadav, A., Chandel, S., Chatufale, S., and Bandhakavi, A. (2023). Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.00913.

169. Zhang, C., Doan, K., Liao, Q., and Abdul-Mageed, M. (2023). The skipped beat: A study of sociopragmatic understanding in LLMs for 64 languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2630–2662. https://doi.org/10.18653/v1/2023.emnlp-main.160.

170. Kabra, A., Liu, E., Khanuja, S., Aji, A.F., Winata, G., Cahyawijaya, S., Aremu, A., Ogayo, P., and Neubig, G. (2023). Multi-lingual and multi-cultural figurative language understanding. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 8269–8284. https://doi.org/10.18653/v1/2023.findings-acl.525.

171. Wang, Y., Zhu, Y., Kong, C., Wei, S., Yi, X., Xie, X., and Sang, J. (2023). Cdeval: A benchmark for measuring the cultural dimensions of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.16421.

172. Jiang, M., and Joshi, M. (2024). CPopQA: Ranking cultural concept popularity by LLMs. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 615–630.

173. Fung, Y., Chakrabarty, T., Guo, H., Rambow, O., Muresan, S., and Ji, H. (2023). NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 15217–15230. https://doi.org/10.18653/v1/2023.emnlp-main.941.

174. Li, O., Subramanian, M., Saakyan, A., CH-Wang, S., and Muresan, S. (2023). NormDial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 15732–15744. https://doi.org/10.18653/v1/2023.emnlp-main.974.

175. Son, G., Lee, H., Kim, S., Kim, H., Lee, J.c., Yeom, J.W., Jung, J., Kim, J.w., and Kim, S. (2024). HAE-RAE bench: Evaluation of Korean knowledge in language models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 7993–8007.

176. Zhou, D., and Zhang, Y. (2023). Red ai? inconsistent responses from gpt3. 5 models on political issues in the us and china. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.09917.

177. Liu, C., Koto, F., Baldwin, T., and Gurevych, I. (2024). Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 2016–2039.

178. Wang, B., Liu, Z., Huang, X., Jiao, F., Ding, Y., Aw, A., and Chen, N. (2024). SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 370–390.

179. Hardalov, M., Mihaylov, T., Zlatkova, D., Dinkov, Y., Koychev, I., and Nakov, P. (2020). EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 5427–5444. https://doi.org/10.18653/v1/2020.emnlp-main.438.

180. Xuan-Quy, D., Ngoc-Bich, L., The-Duy, V., Xuan-Dung, P., Bac-Bien, N., Van-Tien, N., Thi-My-Thanh, N., and Hong-Phuoc, N. (2023). Vnhsge: Vietnamese high school graduation examination dataset for large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.12199.

181. Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., and Qiu, X. (2023). Evaluating the performance of large language models on gaokao benchmark. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.12474.

182. Nie, E., Shao, B., Ding, Z., Wang, M., Schmid, H., and Schütze, H. (2024). Bmike-53: Investigating cross-lingual knowledge editing with in-context learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.17764.

183. Ni, S., Tan, M., Bai, Y., Niu, F., Yang, M., Zhang, B., Xu, R., Chen, X., Li, C., and Hu, X. (2024). MoZIP: A multilingual benchmark to evaluate large language models in intellectual property. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 11658–11668.

184. Zhang, W., Aljunied, M., Gao, C., Chia, Y.K., and Bing, L. (2023). M3exam: A multilingual, multimodal, multilevel benchmark for examining

large language models. Adv. Neural Inf. Process. Syst. *36*, 5484–5505. Curran Associates, Inc.

185. Das, R.J., Hristov, S.E., Li, H., Dimitrov, D.I., Koychev, I., and Nakov, P. (2024). Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.10378.

186. Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C., and Szpektor, I. (2023). TrueTeacher: Learning factual consistency evaluation with large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2053–2070. https://doi.org/10.18653/v1/2023.emnlp-main.127.

187. Jin, Y., Chandra, M., Verma, G., Hu, Y., Choudhury, M.D., and Kumar, S. (2024). Ask me in english instead: Cross-lingual evaluation of large language models for healthcare queries. In The Web Conference 2024.

188. Joseph, S., Kazanas, K., Reina, K., Ramanathan, V., Xu, W., Wallace, B., and Li, J.J. (2023). Multilingual simplification of medical texts. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 16662–16692. https://doi.org/10.18653/v1/2023.emnlp-main.1037.

189. Zhao, B., Jin, W., Del Ser, J., and Yang, G. (2023). Chatagri: Exploring potentials of chatgpt on cross-linguistic agricultural text classification. Neurocomputing *557*, 126708. https://doi.org/10.1016/j.neucom.2023.126708.

190. Goenaga, I., Atutxa, A., Gojenola, K., Oronoz, M., and Agerri, R. (2023). Explanatory argument extraction of correct answers in resident medical exams. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.00567.

191. Datta, D., Soni, S., Mukherjee, R., and Ghosh, S. (2023). MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 5291–5302.

192. Thulke, D., Gao, Y., Pelser, P., Brune, R., Jalota, R., Fok, F., Ramos, M., van Wyk, I., Nasir, A., Goldstein, H., et al. (2024). Climategpt: Towards ai synthesizing interdisciplinary research on climate change. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.09646.

193. Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. (2024). Multilingual machine translation with large language models: Empirical results and analysis. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 2765–2781.

194. Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2023). Prompting PaLM for translation: Assessing strategies and performance. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 15406–15427. https://doi.org/10.18653/v1/2023.acl-long.859.

195. Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Trans. Assoc. Comput. Linguist. *10*, 522–538. https://doi.org/10.1162/tacl_a_00474.

196. Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2207.04672.

197. Bawden, R., Bilinski, E., Lavergne, T., and Rosset, S. (2021). Diabla: a corpus of bilingual spontaneous written dialogues for machine translation. Comput. Humanit. *55*, 635–660.

198. Lou, L., Yin, X., Xie, Y., and Xiang, Y. (2023). CCEval: A representative evaluation benchmark for the Chinese-centric multilingual machine translation. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 10176–10184. https://doi.org/10.18653/v1/2023.findings-emnlp.682.

199. Mujadia, V., Urlana, A., Bhaskar, Y., Pavani, P.A., Shravya, K., Krishnamurthy, P., and Sharma, D.M. (2023). Assessing translation capabilities of large language models involving english and indian languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.09216.

200. Fujii, T., Shibata, K., Yamaguchi, A., Morishita, T., and Sogawa, Y. (2023). How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in Japanese. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) (Association for Computational Linguistics), pp. 39–49. https://doi.org/10.18653/v1/2023.acl-srw.5.

201. Khatri, J., Murthy, R., Azad, A.P., and Bhattacharyya, P. (2023). A study of multilingual versus meta-learning for language model pre-training for adaptation to unseen low resource languages. In Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track. Asia-Pacific Association for Machine Translation, pp. 26–34.

202. Etxaniz, J., Azkune, G., Soroa, A., Lacalle, O., and Artetxe, M. (2024). Do multilingual language models think better in English? In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 550–564.

203. Artetxe, M., Goswami, V., Bhosale, S., Fan, A., and Zettlemoyer, L. (2023). Revisiting machine translation for cross-lingual classification. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 6489–6499.

204. Stefanovitch, N., and Piskorski, J. (2023). Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 71–86. https://doi.org/10.18653/v1/2023.emnlp-main.6.

205. Deng, S., Zhang, N., Xiong, F., Pan, J.Z., and Chen, H. (2022). Knowledge extraction in low-resource scenarios: survey and perspective. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.08063.

206. Kuparinen, O., Miletić, A., and Scherrer, Y. (2023). Dialect-to-standard normalization: A large-scale multilingual evaluation. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 13814–13828. https://doi.org/10.18653/v1/2023.findings-emnlp.923.

207. Wassie, A.K. (2024). Machine translation for ge'ez language. In 5th Workshop on African Natural Language Processing.

208. Liu, P., Zhang, L., Farup, T.N., Lauvrak, E.W., Ingvaldsen, J.E., Eide, S., Gulla, J.A., and Yang, Z. (2023). Nlebench+ norglm: A comprehensive empirical analysis and benchmark dataset for generative language models in norwegian. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.01314.

209. Rakhimova, D., Karibayeva, A., and Turarbek, A. (2024). The task of post-editing machine translation for the low-resource language. Appl. Sci. *14*, 486. https://doi.org/10.3390/app14020486.

210. Yang, C.-K., Huang, K.-P., Lu, K.-H., Kuan, C.-Y., Hsiao, C.-Y., and Lee, H.-y. (2023). Investigating zero-shot generalizability on mandarin-english code-switched asr and speech-to-text translation of recent foundation models with self-supervision and weak supervision. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.00273.

211. Gueuwou, S., Siake, S., Leong, C., and Müller, M. (2023). JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 9907–9927. https://doi.org/10.18653/v1/2023.findings-emnlp.664.

212. Bellagente, M., Brack, M., Teufel, H., Friedrich, F., Deiseroth, B., Eichenberg, C., Dai, A.M., Baldock, R., Nanda, S., Oostermeijer, K., et al. (2023). Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation. In Advances in Neural Information Processing Systems, *36* (Curran Associates, Inc.), pp. 59502–59521.

213. Zhong, S., Huang, Z., Gao, S., Wen, W., Lin, L., Zitnik, M., and Zhou, P. (2024). Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13246–13257.

214. Tuo, Y., Xiang, W., He, J.-Y., Geng, Y., and Xie, X. (2024). Anytext: Multilingual visual text generation and editing. In The Twelfth International Conference on Learning Representations.

215. Ponti, E.M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., and Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 2362–2376. https://doi.org/10.18653/v1/2020.emnlp-main.185.

216. Keung, P., Lu, Y., Szarvas, G., and Smith, N.A. (2020). The multilingual Amazon reviews corpus. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 4563–4568. https://doi.org/10.18653/v1/2020.emnlp-main.369.

217. Tikhonov, A., and Ryabinin, M. (2021). It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (Association for Computational Linguistics), pp. 3534–3546. https://doi.org/10.18653/v1/2021.findings-acl.310.

218. Yin, D., Bansal, H., Monajatipoor, M., Li, L.H., and Chang, K.-W. (2022). GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2039–2055. https://doi.org/10.18653/v1/2022.emnlp-main.132.

219. Lin, B.Y., Lee, S., Qiao, X., and Ren, X. (2021). Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1274–1287. https://doi.org/10.18653/v1/2021.acl-long.102.

220. Lin, X.V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. (2022). Few-shot learning with multilingual generative language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9019–9052. https://doi.org/10.18653/v1/2022.emnlp-main.616.

221. Razumovskaia, E., Maynez, J., Louis, A., Lapata, M., and Narayan, S. (2024). Little red riding hood goes around the globe: Crosslingual story planning and generation with large language models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 10616–10631.

222. Adelani, D.I., Masiak, M., Azime, I.A., Alabi, J., Tonja, A.L., Mwase, C., Ogundepo, O., Dossou, B.F.P., Oladipo, A., Nixdorf, D., et al. (2023). MasakhaNEWS: News topic classification for African languages. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 144–159.

223. Luo, H., Sun, Q., Xu, C., Zhao, P., Lou, J., Tao, C., Geng, X., Lin, Q., Chen, S., and Zhang, D. (2023). Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.09583.

224. Zhang, Y., Wang, J., Wang, Z., and Zhang, R. (2023). XSemPLR: Cross-lingual semantic parsing in multiple natural languages and meaning representations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 15918–15947. https://doi.org/10.18653/v1/2023.acl-long.887.

225. Wang, Z., Zhou, S., Fried, D., and Neubig, G. (2023). Execution-based evaluation for open-domain code generation. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 1271–1290. https://doi.org/10.18653/v1/2023.findings-emnlp.89.

226. Wang, Z., Cuenca, G., Zhou, S., Xu, F.F., and Neubig, G. (2023). MCoNaLa: A benchmark for code generation from multiple natural languages. In Findings of the Association for Computational Linguistics: EACL 2023 (Association for Computational Linguistics), pp. 265–273. https://doi.org/10.18653/v1/2023.findings-eacl.20.

227. Peng, Q., Chai, Y., and Li, X. (2024). HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 8383–8394.

228. Ryan, M., Naous, T., and Xu, W. (2023). Revisiting non-English text simplification: A unified multilingual benchmark. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 4898–4927. https://doi.org/10.18653/v1/2023.acl-long.269.

229. Narayan, S., Cohen, S.B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 1797–1807. https://doi.org/10.18653/v1/D18-1206.

230. Bhattacharjee, A., Hasan, T., Ahmad, W.U., Li, Y.-F., Kang, Y.-B., and Shahriyar, R. (2023). CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 2541–2564.

231. Wang, J., Meng, F., Lu, Z., Zheng, D., Li, Z., Qu, J., and Zhou, J. (2022). ClidSum: A benchmark dataset for cross-lingual dialogue summarization. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 7716–7729. https://doi.org/10.18653/v1/2022.emnlp-main.526.

232. Zhang, R., and Eickhoff, C. (2024). CroCoSum: A benchmark dataset for cross-lingual code-switched summarization. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 4113–4126.

233. Urlana, A., Chen, P., Zhao, Z., Cohen, S., Shrivastava, M., and Haddow, B. (2023). PMIndiaSum: Multilingual and cross-lingual headline summarization for languages in India. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 11606–11628. https://doi.org/10.18653/v1/2023.findings-emnlp.777.

234. Clark, E., Rijhwani, S., Gehrmann, S., Maynez, J., Aharoni, R., Nikolaev, V., Sellam, T., Siddhant, A., Das, D., and Parikh, A. (2023). SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9397–9413.

235. Nguyen, L., Scialom, T., Piwowarski, B., and Staiano, J. (2023). LoRaLay: A multilingual and multimodal dataset for long range and layout-aware summarization. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 636–651. https://doi.org/10.18653/v1/2023.eacl-main.46.

236. Verma, Y., Jangra, A., Verma, R., and Saha, S. (2023). Large scale multilingual multi-modal summarization dataset. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 3620–3632. https://doi.org/10.18653/v1/2023.eacl-main.263.

237. Boughorbel, S., and Hawasly, M. (2023). Analyzing multilingual competency of LLMs in multi-turn instruction following: A case study of Arabic. In Proceedings of ArabicNLP 2023 (Association for Computational Linguistics), pp. 128–139.

238. Zhang, C., D'Haro, L., Tang, C., Shi, K., Tang, G., and Li, H. (2023). xDial-eval: A multilingual open-domain dialogue evaluation benchmark. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 5579–5601. https://doi.org/10.18653/v1/2023.findings-emnlp.371.

239. Hu, S., Zhou, H., Hergul, M., Gritta, M., Zhang, G., Iacobacci, I., Vulić, I., and Korhonen, A. (2023). Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. Trans. Assoc. Comput. Ling. *11*, 1396–1415. https://doi.org/10.1162/tacl_a_00609.

240. Chen, N., Wang, Y., Jiang, H., Cai, D., Li, Y., Chen, Z., Wang, L., and Li, J. (2023). Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In Findings of the Association for

Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 8506–8520.

241. Moradshahi, M., Shen, T., Bali, K., Choudhury, M., de Chalendar, G., Goel, A., Kim, S., Kodali, P., Kumaraguru, P., Semmar, N., et al. (2023). X-RiSAWOZ: High-quality end-to-end multilingual dialogue datasets and few-shot agents. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 2773–2794. https://doi.org/10.18653/v1/2023.findings-acl.174.

242. Mendonca, J., Lavie, A., and Trancoso, I. (2023). Towards multilingual automatic open-domain dialogue evaluation. In Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Association for Computational Linguistics), pp. 130–141. https://doi.org/10.18653/v1/2023.sigdial-1.11.

243. Mendonça, J., Pereira, P., Moniz, H., Paulo Carvalho, J., Lavie, A., and Trancoso, I. (2023). Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. In Proceedings of The Eleventh Dialog System Technology Challenge (Association for Computational Linguistics), pp. 133–143.

244. Ferron, A., Shore, A., Mitra, E., and Agrawal, A. (2023). MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 2078–2100. https://doi.org/10.18653/v1/2023.findings-emnlp.137.

245. Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., and Shavrina, T. (2024). mGPT: Few-shot learners go multilingual. Trans. Assoc. Comput. Linguist. *12*, 58–79. https://doi.org/10.1162/tacl_a_00633.

246. Le Scao, T., Wang, T., Hesslow, D., Bekman, S., Bari, M.S., Biderman, S., Elsahar, H., Muennighoff, N., Phang, J., Press, O., et al. (2022). What language model to train if you have one million GPU hours? In Findings of the Association for Computational Linguistics: EMNLP 2022 (Association for Computational Linguistics), pp. 765–782. https://doi.org/10.18653/v1/2022.findings-emnlp.54.

247. Li, C., Wang, S., Zhang, J., and Zong, C. (2024). Improving in-context learning of multilingual generative language models with cross-lingual alignment. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 8058–8076.

248. Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., Peris, C., Rawls, S., Rosenbaum, A., Rumshisky, A., et al. (2022). Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2208.01448.

249. Liu, C., Zhang, W., Zhao, Y., Luu, A.T., and Bing, L. (2024). Is translation all you need? a study on solving multilingual tasks with large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.10258.

250. Wen-Yi, A.W., and Mimno, D. (2023). Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 1124–1131. https://doi.org/10.18653/v1/2023.emnlp-main.71.

251. Blevins, T., and Zettlemoyer, L. (2022). Language contamination helps explains the cross-lingual capabilities of English pretrained models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 3563–3574. https://doi.org/10.18653/v1/2022.emnlp-main.233.

252. Briakou, E., Cherry, C., and Foster, G. (2023). Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 9432–9452.

253. Holmström, O., Kunz, J., and Kuhlmann, M. (2023). Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023) (Association for Computational Linguistics), pp. 92–110.

254. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al. (2023). GLM-130b: An open bilingual pre-trained model. In The Eleventh International Conference on Learning Representations.

255. Su, H., Zhou, X., Yu, H., Shen, X., Chen, Y., Zhu, Z., Yu, Y., and Zhou, J. (2022). Welm: A well-read pre-trained language model for chinese. Preprint at arXiv. https://doi.org/10.48550/arXiv.2209.10372.

256. Kale, M., Siddhant, A., Al-Rfou, R., Xue, L., Constant, N., and Johnson, M. (2021). nmT5 - is parallel data still relevant for pre-training massively multilingual language models? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 683–691. https://doi.org/10.18653/v1/2021.acl-short.87.

257. Kim, B., Kim, H., Lee, S.-W., Lee, G., Kwak, D., Dong Hyeon, J., Park, S., Kim, S., Kim, S., Seo, D., et al. (2021). What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 3405–3424. https://doi.org/10.18653/v1/2021.emnlp-main.274.

258. Chai, Y., Wang, S., Pang, C., Sun, Y., Tian, H., and Wu, H. (2023). ERNIE-code: Beyond English-centric cross-lingual pretraining for programming languages. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 10628–10650.

259. Schioppa, A., Garcia, X., and Firat, O. (2023). Cross-lingual supervision improves large language models pre-training. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.11778.

260. Billah Nagoudi, E.M., Abdul-Mageed, M., Elmadany, A., Inciarte, A., and Islam Khondaker, M.T. (2023). JASMINE: Arabic GPT models for few-shot learning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 16721–16744.

261. Uthus, D., Ontanon, S., Ainslie, J., and Guo, M. (2023). mLongT5: A multilingual and efficient text-to-text transformer for longer sequences. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 9380–9386. https://doi.org/10.18653/v1/2023.findings-emnlp.628.

262. Wei, T., Zhao, L., Zhang, L., Zhu, B., Wang, L., Yang, H., Li, B., Cheng, C., Lü, W., Hu, R., et al. (2023). Skywork: A more open bilingual foundation model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.19341.

263. Uludoğan, G., Balal, Z.Y., Akkurt, F., Türker, M., Güngör, O., and Üsküdarlı, S. (2024). Turna: A turkish encoder-decoder language model for enhanced understanding and generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.14373.

264. Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.02137.

265. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. Trans. Assoc. Comput. Ling. *10*, 291–306. https://doi.org/10.1162/tacl_a_00461.

266. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.09288.

267. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. J. Mach. Learn. Res. *24*, 1–113.

268. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.06825.

269. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.D., Hanna, E.B., Bressand, F., et al. (2024).

Mixtral of experts. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.04088.

270. Parmar, J., Prabhumoye, S., Jennings, J., Patwary, M., Subramanian, S., Su, D., Zhu, C., Narayanan, D., Jhunjhunwala, A., Dattagupta, A., et al. (2024). Nemotron-4 15b technical report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.16819.

271. Blevins, T., Limisiewicz, T., Gururangan, S., Li, M., Gonen, H., Smith, N.A., and Zettlemoyer, L. (2024). Breaking the curse of multilinguality with cross-lingual expert language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.10440.

272. Chung, H.W., Garcia, X., Roberts, A., Tay, Y., Firat, O., Narang, S., and Constant, N. (2023). Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In The Eleventh International Conference on Learning Representations.

273. Muraoka, M., Bhattacharjee, B., Merler, M., Blackwood, G., Li, Y., and Zhao, Y. (2023). Cross-lingual transfer of large language model by visually-derived supervision toward low-resource languages. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23 Association for Computing Machinery), pp. 3637–3646. https://doi.org/10.1145/3581783.3611992.

274. Zhang, Z., Gu, Y., Han, X., Chen, S., Xiao, C., Sun, Z., Yao, Y., Qi, F., Guan, J., Ke, P., et al. (2021). Cpm-2: Large-scale cost-effective pre-trained language models. AI Open *2*, 216–224. https://doi.org/10.1016/j.aiopen.2021.12.003.

275. Pires, R., Abonizio, H., Almeida, T.S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In Intelligent Systems (Switzerland: Springer Nature), pp. 226–240. https://doi.org/10.1007/978-3-031-45392-2_15.

276. Almeida, T.S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.09887.

277. Luukkonen, R., Komulainen, V., Luoma, J., Eskelinen, A., Kanerva, J., Kupari, H.-M., Ginter, F., Laippala, V., Muennighoff, N., Piktus, A., et al. (2023). FinGPT: Large generative models for a small language. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2710–2726. https://doi.org/10.18653/v1/2023.emnlp-main.164.

278. Vu, T., Barua, A., Lester, B., Cer, D., Iyyer, M., and Constant, N. (2022). Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9279–9300. https://doi.org/10.18653/v1/2022.emnlp-main.630.

279. Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., and Caridá, V. (2023). Cabrita: closing the gap for foreign languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.11878.

280. Basile, P., Musacchio, E., Polignano, M., Siciliani, L., Fiameni, G., and Semeraro, G. (2023). Llamantino: Llama 2 models for effective text generation in italian language. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.09993.

281. Faysse, M., Fernandes, P., Guerreiro, N., Loison, A., Alves, D., Corro, C., Boizard, N., Alves, J., Rei, R., Martins, P., et al. (2024). Croissantllm: A truly bilingual french-english language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.00786.

282. García-Ferrero, I., Agerri, R., Atutxa Salazar, A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramirez-Romero, J., Rigau, G., et al. (2024). MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 11165–11177.

283. Tang, Y., Han, K., Liu, F., Ni, Y., Tian, Y., Bai, Z., Hu, Y.-Q., Liu, S., JUI, S., and Wang, Y. (2024). Rethinking optimization and architecture for tiny language models. In Forty-first International Conference on Machine Learning.

284. Yamaguchi, A., Villavicencio, A., and Aletras, N. (2024). An empirical study on cross-lingual vocabulary adaptation for efficient generative llm inference. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.10712.

285. Lin, P., Ji, S., Tiedemann, J., Martins, A.F., and Schütze, H. (2024). Mala-500: Massive language adaptation of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.13303.

286. Singh, V., Krishna, A., NJ, K., and Ramakrishnan, G. (2024). A three-pronged approach to cross-lingual adaptation with multilingual llms. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.17377.

287. Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., and Okazaki, N. (2024). Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In First Conference on Language Modeling.

288. Xu, H., Kim, Y.J., Sharaf, A., and Awadalla, H.H. (2024). A paradigm shift in machine translation: Boosting translation performance of large language models. In The Twelfth International Conference on Learning Representations.

289. Guo, J., Yang, H., Li, Z., Wei, D., Shang, H., and Chen, X. (2024). A novel paradigm boosting translation capabilities of large language models. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 639–649.

290. Yang, Z.G., Laki, L.J., Váradi, T., and Prószéky, G. (2023). Mono- and multilingual gpt-3 models for hungarian. In Text, Speech, and Dialogue: 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings (Springer-Verlag), pp. 94–104. https://doi.org/10.1007/978-3-031-40498-6_9.

291. Chen, P., Ji, S., Bogoychev, N., Kutuzov, A., Haddow, B., and Heafield, K. (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), pp. 1347–1356.

292. Santilli, A., and Rodolà, E. (2023). Camoscio: An italian instruction-tuned llama. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.16456.

293. Bao, E., Pérez, A., and Parapar, J. (2023). Conversations in galician: a large language model for an underrepresented language. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.03812.

294. Kohli, G.S., Parida, S., Sekhar, S., Saha, S., Nair, N.B., Agarwal, P., Khosla, S., Patiyal, K., and Dhal, D. (2024). Building a llama2-finetuned llm for odia language utilizing domain knowledge instruction set. In Proceedings of the Third International Conference on AI-ML Systems (AIMLSystems '23 Association for Computing Machinery). https://doi.org/10.1145/3639856.3639890.

295. Holmström, O., and Doostmohammadi, E. (2023). Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa) (University of Tartu Library), pp. 634–642.

296. Garcia, G.L., Paiola, P.H., Morelli, L.H., Candido, G., Júnior, A.C., Jodas, D.S., Afonso, L., Guilherme, I.R., Penteado, B.E., and Papa, J.P. (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based task. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.02909.

297. Chirkova, N., and Nikoulina, V. (2024). Zero-shot cross-lingual transfer in instruction tuning of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.14778.

298. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. J. Mach. Learn. Res. *25*, 1–53.

299. Chen, Z., Jiang, F., Chen, J., Wang, T., Yu, F., Chen, G., Zhang, H., Liang, J., Zhang, C., Zhang, Z., et al. (2023). Phoenix: Democratizing chatgpt across languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.10453.

300. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.10403.

301. Cahyawijaya, S., Lovenia, H., Yu, T., Chung, W., and Fung, P. (2023). InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In Proceedings of the First Workshop in South East Asian Language Processing (Association for Computational Linguistics), pp. 55–78. https://doi.org/10.18653/v1/2023.sealp-1.5.

302. Li, J., Zhou, H., Huang, S., Cheng, S., and Chen, J. (2024). Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Trans. Assoc. Comput. Ling. *12*, 576–592. https://doi.org/10.1162/tacl_a_00655.

303. Gao, P., He, Z., Wu, H., and Wang, H. (2024). Towards boosting many-to-many multilingual machine translation with large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.05861.

304. Aryabumi, V., Dang, J., Talupuru, D., Dash, S., Cairuz, D., Lin, H., Venkitesh, B., Smith, M., Marchisio, K., Ruder, S., et al. (2024). Aya 23: Open weight releases to further multilingual progress. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.15032.

305. Upadhayay, B., and Behzadan, V. (2024). Taco: Enhancing cross-lingual transfer for low-resource languages in LLMs through translation-assisted chain-of-thought processes. In 5th Workshop on practical ML for limited/low resource settings.

306. Zhu, W., Huang, S., Yuan, F., She, S., Chen, J., and Birch, A. (2024). Question translation training for better multilingual reasoning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.07817.

307. GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., et al. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.12793.

308. Yang, A., Xiao, B., Wang, B., Zhang, B., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F., et al. (2023). Baichuan 2: Open large-scale language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.10305.

309. Du, X., Yu, Z., Gao, S., Pan, D., Cheng, Y., Ma, Z., Yuan, R., Qu, X., Liu, J., Zheng, T., et al. (2024). Chinese tiny llm: Pretraining a chinese-centric large language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2404.04167.

310. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023). Qwen-vl: A frontier large vision-language model with versatile abilities. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.12966.

311. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., et al. (2024). Internlm2 technical report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.17297.

312. Jiao, W., Huang, J.-t., Wang, W., He, Z., Liang, T., Wang, X., Shi, S., and Tu, Z. (2023). ParroT: Translating during chat using large language models tuned with human translation and feedback. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 15009–15020. https://doi.org/10.18653/v1/2023.findings-emnlp.1001.

313. Chen, Y., Cai, W., Wu, L., Li, X., Xin, Z., and Fu, C. (2023). Tigerbot: An open multilingual multitask llm. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.08688.

314. Sun, T., Zhang, X., He, Z., Li, P., Cheng, Q., Liu, X., Yan, H., Shao, Y., Tang, Q., Zhang, S., et al. (2024). Moss: An open conversational large language model. Mach. Intell. Res. *21*, 888–905. https://doi.org/10.1007/s11633-024-1502-8.

315. Luo, Y., Kong, Q., Xu, N., Cao, J., Hao, B., Qu, B., Chen, B., Zhu, C., Zhao, C., Zhang, D., et al. (2023). Yayi 2: Multilingual open-source large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.14862.

316. Yang, G., Chen, J., Lin, W., and Byrne, B. (2024). Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 391–398.

317. Moura Ramos, M., Fernandes, P., Farinhas, A., and Martins, A.F. (2023). Aligning neural machine translation models: Human feedback in training and inference. Preprint at arXiv. https://doi.org/10.48550/arXiv.arXiv.e-prints.

318. Adler, B., Agarwal, N., Aithal, A., Anh, D.H., Bhattacharya, P., Brundyn, A., Casper, J., Catanzaro, B., Clay, S., Cohen, J., et al. (2024). Nemotron-4 340b technical report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.11704.

319. Chen, D., Huang, Y., Li, X., Li, Y., Liu, Y., Pan, H., Xu, L., Zhang, D., Zhang, Z., and Han, K. (2024). Orion-14b: Open-source multilingual large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.12246.

320. Dong, Y., Wang, Z., Sreedhar, M., Wu, X., and Kuchaiev, O. (2023). SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 11275–11288. https://doi.org/10.18653/v1/2023.findings-emnlp.754.

321. She, S., Huang, S., Zou, W., Zhu, W., Liu, X., Geng, X., and Chen, J. (2024). Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.06838.

322. Sun, Z., Shen, Y., Zhang, H., Zhou, Q., Chen, Z., Cox, D.D., Yang, Y., and Gan, C. (2024). SALMON: Self-alignment with instructable reward models. In The Twelfth International Conference on Learning Representations.

323. Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Durme, B.V., Murray, K., and Kim, Y.J. (2024). Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In Forty-first International Conference on Machine Learning.

324. Lankford, S., Afli, H., and Way, A. (2023). adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. Information *14*, 638. https://doi.org/10.3390/info14120638.

325. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. (2021). {GS}hard: Scaling giant models with conditional computation and automatic sharding. In International Conference on Learning Representations.

326. Rosenbaum, A., Soltan, S., Hamza, W., Versley, Y., and Boese, M. (2022). LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, pp. 218–241.

327. Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., et al. (2021). Beyond english-centric multilingual machine translation. J. Mach. Learn. Res. *22*, 1–48.

328. Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., et al. (2022). Building machine translation systems for the next thousand languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2205.03983.

329. Tseng, K., and Lin, C.-S. (2022). Enhancing natural language inference of cross-lingual n-shot transfer with multilingual data. In 2022 8th International Conference on Applied System Innovation (ICASI), pp. 68–71. https://doi.org/10.1109/ICASI55125.2022.9774470.

330. Iyer, V., Chen, P., and Birch, A. (2023). Towards effective disambiguation for machine translation with large language models. In Proceedings of the Eighth Conference on Machine Translation (Association for Computational Linguistics), pp. 482–495. https://doi.org/10.18653/v1/2023.wmt-1.44.

331. Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2024). Scaling neural machine translation to 200 languages. Nature *630*, 841–846. https://doi.org/10.1038/s41586-024-07335-x.

332. Yang, W., Li, C., Zhang, J., and Zong, C. (2023). Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.18098.

333. Huot, F., Maynez, J., Alberti, C., Amplayo, R.K., Agrawal, P., Fierro, C., Narayan, S., and Lapata, M. (2024). μPLAN: Summarizing using a content plan as cross-lingual bridge. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 2146–2163.

334. Yuan, F., Lu, Y., Zhu, W., Kong, L., Li, L., Qiao, Y., and Xu, J. (2023). Lego-MT: Learning detachable models for massively multilingual machine translation. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 11518–11533. https://doi.org/10.18653/v1/2023.findings-acl.731.

335. Li, S., Wei, X., Zhu, S., Xie, J., Yang, B., and Xiong, D. (2023). MMNMT: Modularizing multilingual neural machine translation with flexibly assembled MoE and dense blocks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 4978–4990. https://doi.org/10.18653/v1/2023.emnlp-main.303.

336. Awasthi, A., Gupta, N., Samanta, B., Dave, S., Sarawagi, S., and Talukdar, P. (2023). Bootstrapping multilingual semantic parsers using large language models. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 2455–2467. https://doi.org/10.18653/v1/2023.eacl-main.180.

337. De Raedt, M., Bitew, S.K., Godin, F., Demeester, T., and Develder, C. (2023). Zero-shot cross-lingual sentiment classification under distribution shift: an exploratory study. In Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL) (Association for Computational Linguistics), pp. 50–66. https://doi.org/10.18653/v1/2023.mrl-1.5.

338. Thakur, N., Ni, J., Hernandez Abrego, G., Wieting, J., Lin, J., and Cer, D. (2024). Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 7699–7724.

339. Whitehouse, C., Choudhury, M., and Aji, A.F. (2023). LLM-powered data augmentation for enhanced cross-lingual performance. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 671–686. https://doi.org/10.18653/v1/2023.emnlp-main.44.

340. Bansal, P., and Sharma, A. (2023). Large language models as annotators: Enhancing generalization of nlp models at minimal cost. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.15766.

341. Reinauer, R., Simianer, P., Uhlig, K., Mosig, J.E., and Wuebker, J. (2023). Neural machine translation models can learn to be few-shot learners. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.08590.

342. Zhang, B., Liu, Z., Cherry, C., and Firat, O. (2024). When scaling meets LLM finetuning: The effect of data, model and finetuning method. In The Twelfth International Conference on Learning Representations.

343. Yong, Z.X., Schoelkopf, H., Muennighoff, N., Aji, A.F., Adelani, D.I., Almubarak, K., Bari, M.S., Sutawika, L., Kasai, J., Baruwa, A., et al. (2023). BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 11682–11703. https://doi.org/10.18653/v1/2023.acl-long.653.

344. Moslem, Y., Haque, R., and Way, A. (2023). Fine-tuning large language models for adaptive machine translation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.12740.

345. Agrawal, P., Alberti, C., Huot, F., Maynez, J., Ma, J., Ruder, S., Ganchev, K., Das, D., and Lapata, M. (2023). QAmeleon: Multilingual QA with only 5 examples. Trans. Assoc. Comput. Ling. *11*, 1754–1771.

346. Tu, L., Qu, J., Yavuz, S., Joty, S., Liu, W., Xiong, C., and Zhou, Y. (2024). Efficiently aligned cross-lingual transfer learning for conversational tasks using prompt-tuning. In Findings of the Association for Computational Linguistics: EACL 2024 (Association for Computational Linguistics), pp. 1278–1294.

347. Park, N., Park, J., Yoo, K.M., and Yoon, S. (2023). On the analysis of cross-lingual prompt tuning for decoder-based multilingual model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.07820.

348. Dai, S., Liu, X., Luo, P., and Yu, Y. (2024). Act-mnmt auto-constriction turning for multilingual neural machine translation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.06745.

349. Whitehouse, C., Huot, F., Bastings, J., Dehghani, M., Lin, C.-C., and Lapata, M. (2023). Parameter-efficient multilingual summarisation: An empirical study. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.08572.

350. Xiao, Z., Held, W., Liu, Y., and Yang, D. (2023). Task-agnostic low-rank adapters for unseen English dialects. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 7857–7870. https://doi.org/10.18653/v1/2023.emnlp-main.487.

351. Aggarwal, D., Sathe, A., and Sitaram, S. (2024). Maple: Multilingual evaluation of parameter efficient finetuning of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.07598.

352. Le, K.M., Pham, T., Quan, T., and Luu, A.T. (2024). Lampat: Low-rank adaption for multilingual paraphrasing using adversarial training. Proc. AAAI Conf. Artif. Intell. *38*, 18435–18443. https://doi.org/10.1609/aaai.v38i16.29804.

353. Yoon, D., Jang, J., Kim, S., Kim, S., Shafayat, S., and Seo, M. (2024). Langbridge: Multilingual reasoning without multilingual supervision. In ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models.

354. Zhao, Y., Zhang, W., Wang, H., Kawaguchi, K., and Bing, L. (2024). Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.18913.

355. Zhang, R., Cahyawijaya, S., Cruz, J.C.B., Winata, G., and Aji, A.F. (2023). Multilingual large language models are not (yet) code-switchers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 12567–12582. https://doi.org/10.18653/v1/2023.emnlp-main.774.

356. Abdelali, A., Mubarak, H., Chowdhury, S., Hasanain, M., Mousi, B., Boughorbel, S., Abdaljalil, S., El Kheir, Y., Izham, D., Dalvi, F., et al. (2024). LAraBench: Benchmarking Arabic AI with large language models. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 487–520.

357. Zhang, R., Ouni, J., and Eger, S. (2024). Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation. Comput. Ling. *50*, 1001–1047. https://doi.org/10.1162/coli_a_00519.

358. Wang, J., Liang, Y., Meng, F., Zou, B., Li, Z., Qu, J., and Zhou, J. (2023). Zero-shot cross-lingual summarization via large language models. In Proceedings of the 4th New Frontiers in Summarization Workshop (Association for Computational Linguistics), pp. 12–23. https://doi.org/10.18653/v1/2023.newsum-1.2.

359. Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023). Document-level machine translation with large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 16646–16661. https://doi.org/10.18653/v1/2023.emnlp-main.1036.

360. Wei, J., Courbis, A.-L., Lambolais, T., Xu, B., Bernard, P.L., and Dray, G. (2023). Zero-shot bilingual app reviews mining with large language models. In 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 898–904.

361. Holtermann, C., Röttger, P., Dill, T., and Lauscher, A. (2024). Evaluating the elementary multilingual capabilities of large language models with multiq. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.03814.

362. Doğruöz, A.S., Sitaram, S., and Yong, Z.X. (2023). Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 5751–5767. https://doi.org/10.18653/v1/2023.findings-emnlp.382.

363. Khatri, J., Srivastava, V., and Vig, L. (2023). Can you translate for me? code-switched machine translation with large language models. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 83–92. https://doi.org/10.18653/v1/2023.ijcnlp-short.10.

364. Yadav, A., Garg, T., Klemen, M., Ulcar, M., Agarwal, B., and Sikonja, M.R. (2024). Code-mixed sentiment and hate-speech prediction. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.12929.

365. Qin, L., Ni, M., Zhang, Y., and Che, W. (2021). Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20). https://doi.org/10.1145/3502198.

366. Qin, L., Chen, Q., Xie, T., Li, Q., Lou, J.-G., Che, W., and Kan, M.-Y. (2022). GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 2677–2686. https://doi.org/10.18653/v1/2022.acl-long.191.

367. Yong, Z.X., Zhang, R., Forde, J., Wang, S., Subramonian, A., Lovenia, H., Cahyawijaya, S., Winata, G., Sutawika, L., Cruz, J.C.B., et al. (2023). Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching (Association for Computational Linguistics), pp. 43–63.

368. Amin, D., Govilkar, S., Kulkarni, S., Lalit, Y.S., Khwaja, A.A., Xavier, D., and Gupta, S.G. (2023). Marathi-english code-mixed text generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.16202.

369. Zeng, J., Meng, F., Yin, Y., and Zhou, J. (2023). Improving machine translation with large language models: A preliminary study with cooperative decoding. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.02851.

370. Nambi, A., Balloli, V., Ranjit, M., Ganu, T., Ahuja, K., Sitaram, S., and Bali, K. (2023). Breaking language barriers with a leap: Learning strategies for polyglot llms. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.17740.

371. Lu, H., Huang, H., Zhang, D., Yang, H., Lam, W., and Wei, F. (2023). Chain-of-dictionary prompting elicits translation in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.06575.

372. Li, Y., Korhonen, A., and Vulić, I. (2023). On bilingual lexicon induction with large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9577–9599. https://doi.org/10.18653/v1/2023.emnlp-main.595.

373. Cheng, X., Wang, X., Ge, T., Chen, S.-Q., Wei, F., Zhao, D., and Yan, R. (2023). Scale: Synergized collaboration of asymmetric language translation engines. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.17061.

374. Petrick, F., Herold, C., Petrushkov, P., Khadivi, S., and Ney, H. (2023). Document-level language models for machine translation. In Proceedings of the Eighth Conference on Machine Translation (Association for Computational Linguistics), pp. 375–391. https://doi.org/10.18653/v1/2023.wmt-1.39.

375. Hoang, H., Khayrallah, H., and Junczys-Dowmunt, M. (2024). On-the-fly fusion of large language models and machine translation. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 520–532.

376. Intrator, Y., Halfon, M., Goldenberg, R., Tsarfaty, R., Eyal, M., Rivlin, E., Matias, Y., and Aizenberg, N. (2024). Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 829–844.

377. Moslem, Y., Haque, R., Kelleher, J.D., and Way, A. (2023). Adaptive machine translation with large language models. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, pp. 227–237.

378. Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., and Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 12009–12024. https://doi.org/10.18653/v1/2023.findings-emnlp.804.

379. Wu, Y., and Hu, G. (2023). Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In Proceedings of the Eighth Conference on Machine Translation (Association for Computational Linguistics), pp. 166–169. https://doi.org/10.18653/v1/2023.wmt-1.15.

380. Puduppully, R., Kunchukuttan, A., Dabre, R., Aw, A.T., and Chen, N. (2023). DecoMT: Decomposed prompting for machine translation between related languages using large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 4586–4602. https://doi.org/10.18653/v1/2023.emnlp-main.279.

381. Pilault, J., Garcia, X., Bražinskas, A., and Firat, O. (2023). Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 455–483. https://doi.org/10.18653/v1/2023.ijcnlp-main.31.

382. Patel, A., Li, B., Rasooli, M.S., Constant, N., Raffel, C., and Callison-Burch, C. (2023). Bidirectional language models are also few-shot learners. In The Eleventh International Conference on Learning Representations.

383. Rosenbaum, A., Soltan, S., Hamza, W., Damonte, M., Groves, I., and Saffari, A. (2022). CLASP: Few-shot cross-lingual data augmentation for semantic parsing. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Association for Computational Linguistics), pp. 444–462.

384. Tanwar, E., Dutta, S., Borthakur, M., and Chakraborty, T. (2023). Multilingual LLMs are better cross-lingual in-context learners with alignment. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 6292–6307. https://doi.org/10.18653/v1/2023.acl-long.346.

385. Ohmer, X., Bruni, E., and Hupkes, D. (2023). Evaluating task understanding through multilingual consistency: A chatgpt case study. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.11662.

386. Zhang, Y., Chen, Q., Li, M., Che, W., and Qin, L. (2024). AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In Findings of the Association for Computational Linguistics ACL 2024 (Association for Computational Linguistics), pp. 9191–9200. https://doi.org/10.18653/v1/2024.findings-acl.546.

387. He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang, X. (2024). Exploring human-like translation strategy with large language models. Trans. Assoc. Comput. Ling. *12*, 229–246. https://doi.org/10.1162/tacl_a_00642.

388. Zhang, M., Liu, L., Yanqing, Z., Qiao, X., Chang, S., Zhao, X., Zhu, J., Zhu, M., Peng, S., Li, Y., et al. (2023). Leveraging multilingual knowledge graph to boost domain-specific entity translation of ChatGPT. In Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track (Asia-Pacific Association for Machine Translation), pp. 77–87.

389. Conia, S., Li, M., Lee, D., Minhas, U., Ilyas, I., and Li, Y. (2023). Increasing coverage and precision of textual information in multilingual knowledge graphs. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 1612–1634.

390. Xu, S., Li, J., and Xiong, D. (2023). Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 3692–3702. https://doi.org/10.18653/v1/2023.emnlp-main.226.

391. Ahmad, S.R. (2024). Enhancing multilingual information retrieval in mixed human resources environments: A rag model implementation for multicultural enterprise. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.01511.

392. Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2023). In-context examples selection for machine translation. In Findings

of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 8857–8873.

393. Li, X., Nie, E., and Liang, S. (2023). From classification to generation: Insights into crosslingual retrieval augmented ICL. In NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following.

394. Winata, G.I., Huang, L.-K., Vadlamannati, S., and Chandarana, Y. (2023). Multilingual few-shot learning via language model retrieval. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.10964.

395. Garcia, X., Bansal, Y., Cherry, C., Foster, G., Krikun, M., Johnson, M., and Firat, O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. In Proceedings of the 40th International Conference on Machine Learning vol. 202 of *Proceedings of Machine Learning Research* (PMLR), pp. 10867–10878.

396. Li, X., Nie, E., and Liang, S. (2023). Crosslingual retrieval augmented in-context learning for Bangla. In Proceedings of the First Workshop on Bangla Language Processing (BLP-2023) (Association for Computational Linguistics), pp. 136–151. https://doi.org/10.18653/v1/2023.banglalp-1.15.

397. Ramos, R., Martins, B., and Elliott, D. (2023). LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 1635–1651. https://doi.org/10.18653/v1/2023.findings-acl.104.

398. Kim, S., Ki, D., Kim, Y., and Lee, J. (2023). Boosting cross-lingual transferability in multilingual models via in-context learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.15233.

399. Thakur, N., Bonifacio, L., Zhang, X., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Chen, B., Rezagholizadeh, M., et al. (2023). Nomiracl: Knowing when you don't know for robust multilingual retrieval-augmented generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.11361.

400. Huang, Y., Fan, C., Li, Y., Wu, S., Zhou, T., Zhang, X., and Sun, L. (2024). 1+1>2: Can large language models serve as cross-lingual knowledge aggregators?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.14721.

401. Ranaldi, L., Pucci, G., Ranaldi, F., Ruzzetti, E.S., and Zanzotto, F.M. (2024). A tree-of-thoughts to broaden multi-step reasoning across languages. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 1229–1241. https://doi.org/10.18653/v1/2024.findings-naacl.78.

402. Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., Li, M., Che, W., and Yu, P.S. (2024). Large language models meet nlp: A survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.12819.

403. Raunak, V., Menezes, A., and Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics), pp. 1172–1183. https://doi.org/10.18653/v1/2021.naacl-main.92.

404. Xue, B., Wang, H., Wang, R., Wang, S., Wang, Z., Du, Y., and Wong, K.-F. (2024). A comprehensive study of multilingual confidence estimation on large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.13606.

405. Guerreiro, N.M., Alves, D.M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A.F.T. (2023). Hallucinations in large multilingual translation models. Trans. Assoc. Comput. Ling. *11*, 1500–1517. https://doi.org/10.1162/tacl_a_00615.

406. Aharoni, R., Narayan, S., Maynez, J., Herzig, J., Clark, E., and Lapata, M. (2023). Multilingual summarization with factual consistency evaluation. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 3562–3591.

407. Dale, D., Voita, E., Lam, J., Hansanti, P., Ropers, C., Kalbassi, E., Gao, C., Barrault, L., and Costa-jussà, M. (2023). HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 638–653. https://doi.org/10.18653/v1/2023.emnlp-main.42.

408. Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E., and Cohen, S. (2023). Detecting and mitigating hallucinations in multilingual summarisation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 8914–8932. https://doi.org/10.18653/v1/2023.emnlp-main.551.

409. Pfeiffer, J., Piccinno, F., Nicosia, M., Wang, X., Reid, M., and Ruder, S. (2023). mmT5: Modular multilingual pre-training solves source language hallucinations. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 1978–2008. https://doi.org/10.18653/v1/2023.findings-emnlp.132.

410. Ahuja, K., Sitaram, S., Dandapat, S., and Choudhury, M. (2022). On the calibration of massively multilingual language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 4310–4323. https://doi.org/10.18653/v1/2022.emnlp-main.290.

411. Yang, Y., Dan, S., Roth, D., and Lee, I. (2023). Understanding calibration for multilingual question answering models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.08669.

412. Sia, S., DeLucia, A., and Duh, K. (2024). Anti-LM decoding for zero-shot in-context machine translation. In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 3403–3420.

413. Kang, H., Blevins, T., and Zettlemoyer, L. (2024). Comparing hallucination detection metrics for multilingual generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.10496.

414. Ding, H., Pang, L., Wei, Z., Shen, H., and Cheng, X. (2024). Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.10612.

415. Gao, C., Hu, H., Hu, P., Chen, J., Li, J., and Huang, S. (2024). Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 6101–6117.

416. Wu, S., Peng, M., Chen, Y., Su, J., and Sun, M. (2023). Eva-kellm: A new benchmark for evaluating knowledge editing of llms. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.09954.

417. Wang, J., Liang, Y., Sun, Z., Cao, Y., and Xu, J. (2023). Cross-lingual knowledge editing in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.08952.

418. Beniwal, H., D, K., and Singh, M. (2024). Cross-lingual editing in multilingual language models. In Findings of the Association for Computational Linguistics: EACL 2024 (Association for Computational Linguistics), pp. 2078–2128.

419. Qi, J., Fernández, R., and Bisazza, A. (2023). Cross-lingual consistency of factual knowledge in multilingual language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 10650–10666. https://doi.org/10.18653/v1/2023.emnlp-main.658.

420. Wang, W., Haddow, B., and Birch, A. (2023). Retrieval-augmented multilingual knowledge editing. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.13040.

421. Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. (2024). Culturellm: Incorporating cultural differences into large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.10946.

422. Costa-jussà, M., Smith, E., Ropers, C., Licht, D., Maillard, J., Ferrando, J., and Escolano, C. (2023). Toxicity in multilingual machine translation at scale. In Findings of the Association for Computational Linguistics: EMNLP 2023 (Association for Computational Linguistics), pp. 9570–9586.

423. Sánchez, E., Andrews, P., Stenetorp, P., Artetxe, M., and Costa-jussà, M.R. (2023). Gender-specific machine translation with large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.03175.

424. Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., Srba, I., Le, T., Lee, D., Simko, J., et al. (2023). MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9960–9987. https://doi.org/10.18653/v1/2023.emnlp-main.616.

425. He, Z., Zhou, B., Hao, H., Liu, A., Wang, X., Tu, Z., Zhang, Z., and Wang, R. (2024). Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.14007.

426. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-t., Jiao, W., and Lyu, M.R. (2023). All languages matter: On the multilingual safety of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.00905.

427. Ye, M., Sikka, K., Atwell, K., Hassan, S., Divakaran, A., and Alikhani, M. (2023). Multilingual content moderation: A case study on Reddit. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 3828–3844. https://doi.org/10.18653/v1/2023.eacl-main.276.

428. Hämmerl, K., Deiseroth, B., Schramowski, P., Libovickỳ, J., Fraser, A., and Kersting, K. (2022). Do multilingual language models capture differing moral norms?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.09904.

429. Li, J., Liu, Y., Liu, C., Shi, L., Ren, X., Zheng, Y., Liu, Y., and Xue, Y. (2024). A cross-language investigation into jailbreak attacks in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.16765.

430. Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Mohammed Afzal, O., Mahmoud, T., Sasaki, T., et al. (2024). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1369–1407.

431. Bogoychev, N., van der Linde, J., Nail, G., Haddow, B., Zaragoza-Bernabeu, J., Ramírez-Sánchez, G., Weymann, L., Mateiu, T.N., Helcl, J., and Aulamo, M. (2023). Opuscleaner and opustrainer, open source toolkits for training machine translation and large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.14838.

432. Kew, T., Schottmann, F., and Sennrich, R. (2023). Turning english-centric llms into polyglots: How much multilinguality is needed?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.12683.

433. Shaham, U., Herzig, J., Aharoni, R., Szpektor, I., Tsarfaty, R., and Eyal, M. (2024). Multilingual instruction tuning with just a pinch of multilinguality. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.01854.

434. Geigle, G., Jain, A., Timofte, R., and Glavaš, G. (2023). mblip: Efficient bootstrapping of multilingual vision-llms. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.06930.

435. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. (2023). PaLI: A jointly-scaled multilingual language-image model. In The Eleventh International Conference on Learning Representations.

436. Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al. (2024). On scaling up a multilingual vision and language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14432–14444.

437. Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., et al. (2023). Pali-3 vision language models: Smaller, faster, stronger. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.09199.

438. Zhou, C., Liang, Y., Meng, F., Xu, J., Su, J., and Zhou, J. (2023). RC3: Regularized contrastive cross-lingual cross-modal pre-training. In Findings of the Association for Computational Linguistics: ACL 2023 (Association for Computational Linguistics), pp. 11747–11762. https://doi.org/10.18653/v1/2023.findings-acl.746.

439. Hu, J., Yao, Y., Wang, C., WANG, S., Pan, Y., Chen, Q., Yu, T., Wu, H., Zhao, Y., Zhang, H., et al. (2024). Large multilingual models pivot zero-shot multimodal learning across languages. In The Twelfth International Conference on Learning Representations.

440. Zhou, K. (2023). Accessible instruction-following agent. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.06358.

441. He, C., Jin, Z., Xu, C., Qiu, J., Wang, B., Li, W., Yan, H., Wang, J., and Lin, D. (2023). Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.10755.

442. Guo, W., Fang, Q., Yu, D., and Feng, Y. (2023). Bridging the gap between synthetic and authentic images for multimodal machine translation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2863–2874. https://doi.org/10.18653/v1/2023.emnlp-main.173.

443. Yang, J., Guo, H., Yin, Y., Bai, J., Wang, B., Liu, J., Liang, X., Chai, L., Yang, L., and Li, Z. (2024). m3P: Towards multimodal multilingual translation with multimodal prompt. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (ELRA and ICCL), pp. 10858–10871.

444. Huang, Z., Ye, R., Ko, T., Dong, Q., Cheng, S., Wang, M., and Li, H. (2023). Speech translation with large language models: An industrial practice. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.13585.

445. Ronny Huang, W., Allauzen, C., Chen, T., Gupta, K., Hu, K., Qin, J., Zhang, Y., Wang, Y., Chang, S.-Y., and Sainath, T.N. (2024). Multilingual and fully non-autoregressive asr with large language model fusion: A comprehensive study. In ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 13306–13310. https://doi.org/10.1109/ICASSP48485.2024.10448361.

446. Cheng, Y., Zhang, Y., Johnson, M., Macherey, W., and Bapna, A. (2023). Mu 2SLAM: Multitask, multilingual speech and language models. In Proceedings of the 40th International Conference on Machine Learning vol. 202 of *Proceedings of Machine Learning Research* (PMLR), pp. 5504–5520.

447. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.11805.

448. Chen, Q., Qin, L., Zhang, J., Chen, Z., Xu, X., and Che, W. (2024). M 3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.16473.

449. Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., and Wen, J.-R. (2024). Language-specific neurons: The key to multilingual capabilities in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.16438.

450. Wang, H., Minervini, P., and Ponti, E.M. (2024). Probing the emergence of cross-lingual alignment during llm training. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.13229.

451. Yuan, F., Yuan, S., Wu, Z., and Li, L. (2024). How vocabulary sharing facilitates multilingualism in LLaMA? In Findings of the Association for Computational Linguistics: ACL 2024, L.-W. Ku, A. Martins, and V. Srikumar, eds. (Association for Computational Linguistics), pp. 12111–12130.

452. Hu, L., and Xu, Y. (2024). DM-BLI: Dynamic multiple subspaces alignment for unsupervised bilingual lexicon induction. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 2041–2052.

453. Adeyemi, M., Oladipo, A., Pradeep, R., and Lin, J. (2023). Zero-shot cross-lingual reranking with large language models for low-resource languages. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.16159.

454. Ojo, J., Ogueji, K., Stenetorp, P., and Adelani, D.I. (2023). How good are large language models on african languages?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.07978.

455. Yong, Z.X., Menghini, C., and Bach, S. (2023). Low-resource languages jailbreak GPT-4. In Socially Responsible Language Modelling Research.

456. Lin, P., Hu, C., Zhang, Z., Martins, A., and Schuetze, H. (2024). mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In Findings of the Association for Computational Linguistics: EACL 2024 (Association for Computational Linguistics), pp. 276–310.

457. Sengupta, N., Sahu, S.K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., et al. (2023). Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.16149.

458. Koishekenov, Y., Berard, A., and Nikoulina, V. (2023). Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 3567–3585. https://doi.org/10.18653/v1/2023.acl-long.198.

459. Hua, W.-Y., Williams, B., and Shamsi, D. (2023). Lacos-bloom: Low-rank adaptation with contrastive objective on 8 bits siamese-bloom. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.06404.

460. Nicosia, M., and Piccinno, F. (2022). Byte-level massively multilingual semantic parsing. In Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22) (Association for Computational Linguistics), pp. 25–34. https://doi.org/10.18653/v1/2022.mmnlu-1.3.

461. Sun, J., Fernandes, P., Wang, X., and Neubig, G. (2023). A multi-dimensional evaluation of tokenizer-free multilingual pretrained models. In Findings of the Association for Computational Linguistics: EACL 2023 (Association for Computational Linguistics), pp. 1725–1735. https://doi.org/10.18653/v1/2023.findings-eacl.128.

462. Rust, P., Lotz, J.F., Bugliarello, E., Salesky, E., de Lhoneux, M., and Elliott, D. (2023). Language modelling with pixels. In The Eleventh International Conference on Learning Representations.

463. Edman, L., Sarti, G., Toral, A., Noord, G.v., and Bisazza, A. (2024). Are character-level translations worth the wait? comparing ByT5 and mT5 for machine translation. Trans. Assoc. Comput. Ling. *12*, 392–410. https://doi.org/10.1162/tacl_a_00651.

464. Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N., and Tsvetkov, Y. (2023). Do all languages cost the same? tokenization in the era of commercial language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9904–9923.

465. Alyafeai, Z., Al-shaibani, M.S., Ghaleb, M., and Ahmad, I. (2023). Evaluating various tokenizers for arabic text classification. Neural Process. Lett. *55*, 2911–2933.

466. Alkaoud, M., and Syed, M. (2020). On the importance of tokenization in Arabic embedding models. In Proceedings of the Fifth Arabic Natural Language Processing Workshop (Association for Computational Linguistics), pp. 119–129.

467. Petrov, A., La Malfa, E., Torr, P., and Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. Adv. Neural Inf. Process. Syst. *36*, 36963–36990. Curran Associates, Inc.

468. Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J., et al. (2024). Tokenizer choice for LLM training: Negligible or crucial? In Findings of the Association for Computational Linguistics: NAACL 2024 (Association for Computational Linguistics), pp. 3907–3924.

469. Hong, J., Lee, G., and Cho, J. (2024). A simple framework to accelerate multilingual language model for monolingual text generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.10660.

470. Yu, X., Chatterjee, T., Asai, A., Hu, J., and Choi, E. (2022). Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In Findings of the Association for Computational Linguistics: EMNLP 2022 (Association for Computational Linguistics), pp. 3725–3743. https://doi.org/10.18653/v1/2022.findings-emnlp.273.

471. Wendler, C., Veselovsky, V., Monea, G., and West, R. (2024). Do llamas work in english? on the latent language of multilingual transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.10588.

472. AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., and Diab, M. (2024). Investigating cultural alignment of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.13231.

473. Dossou, B.F.P., Tonja, A.L., Yousuf, O., Osei, S., Oppong, A., Shode, I., Awoyomi, O.O., and Emezue, C. (2022). AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP) (Association for Computational Linguistics), pp. 52–64. https://doi.org/10.18653/v1/2022.sustainlp-1.11.