Front. Comput. Sci., 2025, 19(11): 1911362
https://doi.org/10.1007/s11704-024-40579-4

REVIEW ARTICLE

A survey on multilingual large language models: corpora, alignment, and bias

Yuemei XU (✉), Ling HU, Jiayi ZHAO, Zihan QIU, Kexin XU, Yuqi YE, Hanwen GU

School of Information Science and Technology, Beijing Foreign Studies University, Beijing 100089, China

Abstract Based on the foundation of Large Language Models (LLMs), Multilingual LLMs (MLLMs) have been developed to address the challenges faced in multilingual natural language processing, hoping to achieve knowledge transfer from high-resource languages to low-resource languages. However, significant limitations and challenges still exist, such as language imbalance, multilingual alignment, and inherent bias. In this paper, we aim to provide a comprehensive analysis of MLLMs, delving deeply into discussions surrounding these critical issues. First of all, we start by presenting an overview of MLLMs, covering their evolutions, key techniques, and multilingual capacities. Secondly, we explore the multilingual training corpora of MLLMs and the multilingual datasets oriented for downstream tasks that are crucial to enhance the cross-lingual capability of MLLMs. Thirdly, we survey the state-of-the-art studies of multilingual representations and investigate whether the current MLLMs can learn a universal language representation. Fourthly, we discuss bias on MLLMs, including its categories, evaluation metrics, and debiasing techniques. Finally, we discuss existing challenges and point out promising research directions of MLLMs.

Keywords multilingual large language model, corpora, alignment, bias, survey

1 Introduction

The rapid development of Large Language Models (LLMs) has brought about a paradigm shift and revolution in the field of Natural Language Processing (NLP). This innovative approach trains a transformer-based model [1] on extensive volumes of data and then leverages fine-tuning or prompt learning to facilitate the model's adaption to a wide variety of tasks. Based on the foundation of LLMs, large-scale Multilingual MLLMs (MLLMs), such as mBERT [2], XLM [3], mT5 [4], BLOOM [5] and LLaMA [6], have been developed to tackle multilingual NLP tasks. MLLMs are pre-trained on a concatenation of texts in multiple languages with the hope that low-resource languages may benefit from high-resource languages due to linguistic similarities and shared representations inherent within language pairs.

Compared to LLMs, MLLMs require larger multilingual corpora that cover more languages to ensure applicability and fairness across different languages in downstream tasks. MLLMs are trained to understand and capture the structures and patterns of multiple languages. For instance, pre-trained on data from 104 languages, BLOOM supports 46 languages, covering the eight most widely spoken languages in the world [5]. Numerous MLLMs have been proposed in the past five years, which differ in the architecture (e.g., number of layers, parameters, etc.), data used for pre-training (Wikipedia, Common Crawl, etc.), and the number of languages involved (ranging from 12 to 110). However, it is uncertain how much cross-lingual transfer learning capability MLLMs have to support unseen languages or low-resource languages during pre-training. As a result, Section 2 first starts by providing an overview of MLLMs, which contains key evolutions, techniques, and a detailed analysis of MLLMs' multilingual capacities. Despite the success of MLLMs, existing MLLMs still face numerous issues and challenges, which can be summarized as three aspects: corpora, alignment, and bias. As shown in Fig. 1, training corpora of MLLMs heavily influence their capability. On the one hand, the unbalanced corpora and training lead to misalignment of MLLMs among different languages. On the other hand, inherent bias within corpora induces MLLMs to produce biased output. Therefore, this paper focuses discussion around the three aspects of corpora, alignment, and bias.

Firstly, MLLMs heavily rely on multilingual corpora to enhance their performance. For example, among the training corpus of ChatGPT, the English corpus accounts for 92.099%, and the Chinese only accounts for 0.16%, so its dialogues in the English context are much higher than those in other languages in terms of quality and speed. However, the size of available corpus resources for different languages varies greatly, and most of the existing annotated datasets are mainly focused on a few languages, limited the cross-lingual transfer effectiveness from high-resource languages to languages that are unseen during training. Furthermore, MLLMs suffer from what Conneau et al. [7] call *the curse of multilinguality*: more languages lead to better cross-lingual performance on low-

Received June 7, 2024; accepted December 9, 2024

E-mail: xuyuemei@bfsu.edu.cn

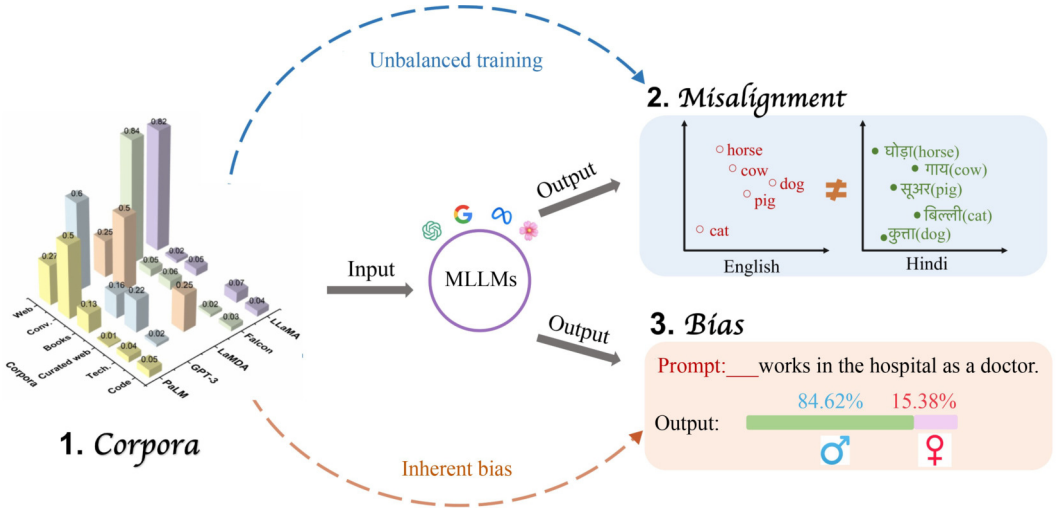

Fig. 1 An illustration of the relationship between corpora, misalignment, and bias. The misalignment and bias produced by MLLM arise in part from the bias and imbalanced language proportions of the training corpora

resource languages up until a point, after which the overall performance of MLLMs on monolingual and cross-lingual benchmarks will decrease. In conclusion, the scale, quality, and diversity of corpora have a significant impact on the performance of MLLMs. Therefore, Section 3 presents a survey of the representative multilingual training corpora of MLLMs, offering insights into their language distribution, data source, and language coverage.

Secondly, MLLMs still struggle to learn a universal language representation for diverse languages. The language misalignment issues exist. Aligning the representation of diverse languages acts as an integral part of NLP's multilingual tasks and applications [8], and under-representation of low-resource languages leads to MLLMs' poor performance on these languages. Inspired by the impressive performance of monolingual representation models like Word2vec [9] and GloVe [10], recent research has made great progress in multilingual representation. In Section 4, we review previous research on multilingual representations and classify them into three categories: static multilingual representation, contextual multilingual representation, and combined multilingual representation. We also analyze the impact of various factors on multilingual alignment performance, including initial alignment solution, linearity of mapping function, language typological distance, and pre-training data and settings of MLLMs.

Thirdly, MLLMs are prone to produce harmful outcomes and social bias [11] in part due to bias is naturally present in cross-cultural datasets and the design of MLLMs' modeling processes [12]. Previous studies have explored bias in various NLP tasks and demographic groups, but are largely specific to English-based models [13,14], which cannot be generalized to other languages. What are the types of bias in existing MLLMs? What are the main de-biasing techniques available for MLLMs? Does the removal of these biases affect LLMs' performance? What are the existing bias evaluation datasets for MLLMs? These are very worthwhile research questions. This survey tries to answer these questions and offers valuable insights for bias on MLLMs.

The contributions of this survey are as follows:

- We present an overview of MLLMs and analyze the language imbalance challenge within MLLMs, their capacity to support low-resource languages and their potential for cross-lingual transfer learning.
- We provide an overview of the multilingual datasets and corpora utilized by existing MLLMs, offering a comprehensive insight into the language distribution within these training corpora.
- We survey the existing studies on multilingual representations and explore whether the current MLLMs can learn a universal language representation.
- Our survey delves into bias within MLLMs, seeking to address essential questions such as identifying the types of bias present in current MLLMs, exploring prominent de-biasing techniques, and summarizing available bias evaluation datasets for MLLMs.

2 Overview of MLLMs

This section provides a brief overview of MLLMs, tracing their evolution from monolingual LLMs to multilingual LLMs, as dispicted in Fig. 2. It then illustrates the key techniques that contribute most to the success of MLLMs, as well as the multilingual capacities of MLLMs.

2.1 Evolution of MLLMs

2.1.1 Monolingual evolution

The development of monolingual LLMs has made great progress in understanding and generating human languages. These models employ a transformer-based architecture that is first pre-trained on a large corpus of texts, followed by fine-tuning or prompt learning to enhance models' performance on specific tasks or languages.

The representative monolingual LLMs are the BERT series and GPT series. With the success of BERT [2], BERT-variants models have been developed for specific languages, such as FlauBERT for French [15], BERTje for Dutch [16], AraBERT for Arabic [17]. The GPT series, evolving from GPT-1 to GPT-2, GPT-3, and beyond [18–22], has

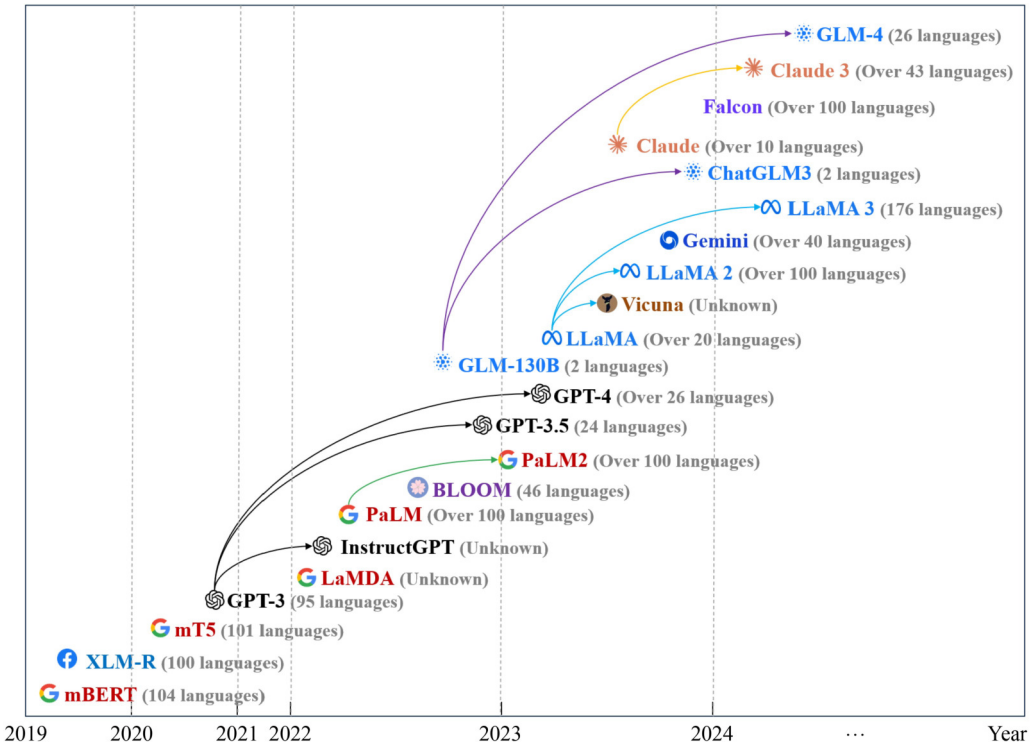

Fig. 2 An illustration of the evolution roadmap of current multilingual LLMs, presenting their release year, the number of supported languages and release relationship. 'Unknown' indicates the model has not disclosed the language proportion in its training data

experienced growth in parameters and training corpora, e.g., parameters ranging from hundreds of millions (GPT-1) [18] to 1.5 trillion (GPT-3) [20]. This progression empowers the models with improved sophisticated language understanding and generation capabilities. T5 model [23] introduces a unified framework that converts a variety of tasks into a text-to-text format by prepending a unique prefix to the input for each task. BART [24] is a sequence-to-sequence model, not an auto-regressive model like GPT-2 or an auto-encoder model like BERT, and thus it is particularly effective on comprehension tasks like summarization.

Monolingual LLMs have seen advancements in transformer-based architecture and pre-training strategies but are language-dependent. Training such language-specific LLMs is only feasible for a few languages with necessary corpora.

2.1.2 Multilingual evolution

Multilingual LLMs build upon the foundation of monolingual LLMs to learn universal language patterns from extensive unlabeled data across multiple languages. MLLMs have advanced to overcome linguistic limitations and enhance low-resource languages by leveraging shared vocabulary and genetic relatedness from high-resource languages [25].

Following the success of monolingual BERT, multilingual BERT (mBERT) [2] was the first released MLLM by following the training procedure of BERT but on multilingual Wikipedia text corpora including 104 languages. Other MLLMs such as XLM-R [7], mBART [26], and mT5 [4] follow the step of mBERT, further exploring the capacities and limitations of MLLMs across languages. Studies reveal that MLLMs work surprisingly well on cross-lingual tasks even without direct cross-lingual supervision like parallel or comparable data [27,28].

The further development trend has led to a tremendous increase in model parameters and data scale, resulting in enhancements that promote multilingual capabilities. For example, PaLM with 540 billion (540B) parameters yields impressive capabilities on the multilingual benchmarks by training on a mixture of multilingual versions of Wikipedia and dialogue data, including 124 languages [29]. With the early successful attempts, the autoregressive language modeling and prompt learning paradigm represented by the GPT series has received much attention and follow-up from major companies and universities. Thus, more MLLMs (e.g., InstructGPT [21], LaMDA [30], OPT [31], BLOOM [5], LLaMA [6]) have been proposed to achieve breakthrough performance in a range of multi-step reasoning tasks over multiple languages. In addition to the GPT series and its derivative models, numerous other models have been proposed to boost the development of LLMs. Examples include GLM [32,33], Vicuna [34], Gemini [35], and several others.

The development of MLLMs has been guided by several tendencies: (1) parameter growth; (2) linguistic diversity; (3) multimodal unification. Regarding (1), MLLMs have expanded to hundreds of billions of parameters or even trillions. Increasing the size of parameters brings clear benefits, such as alleviating hallucination phenomena heavily present in minor-parameter (e.g., 7B, 13B) models. However, there's a limit to the amount of text data available online, and obtaining high-quality data is becoming increasingly challenging, which might slow down the parameter growth of MLLMs. Regarding (2), most high-resource languages belong to similar language families, thus sharing numerous linguistic

features. Disregarding this diversity inevitably leads to poor generalizability and language-specific biases [36]. Recent work in MLLMs has focused on addressing this issue as low-resource and unseen languages still account for a large proportion of the world's languages. Regarding (3), multimodal MLLMs are a growing focus of research, realizing a variety of specific real-world needs by unifying diverse types of modalities (i.e., text, image, and speech). Additionally, current research aims to extend MLLMs to accommodate more modalities like web pages, heat maps, graphs, and tables, thereby increasing the model's generality and applicability [37].

Table 1 summarizes the representative MLLMs in recent years, divided into two categories: monolingual and multilingual, showing the evolution of MLLMs from multiple perspectives in chronological order of release.

2.2 Key techniques of MLLMs

Transformer architecture, pre-training technique, and reinforcement learning with human feedback are the key techniques for MLLMs. In this section, we will present the key ideas behind these techniques.

2.2.1 Transformer architecture

Transformer architecture, first introduced in 2017, has become the foundation of MLLMs owing to its suitability for parallel computing and flexibility for diverse model design. Transformer architecture consists of two main modules, Encoder and Decoder, along with a self-attention mechanism within each module. The encoder, using stacked multi-head self-attention layers, encodes the input sequence and generates latent representations. In contrast, the decoder employs cross-attention to utilize the encoder's latent representations, attending to them while autoregressively generating the target sequence [53].

MLLMs can be categorized into three groups based on the underlying transformer structure:

- **Encoder-only** (e.g., BERT): MLLMs with encoder-only architecture can effectively handle long-range dependencies within the input sequences, making them well-suited for the analysis and classification of textual content, including tasks like sentiment analysis and named entity recognition.
- **Decoder-only** (e.g., GPT): MLLMs with decoder-only architecture are mainly designed to generate sequences of language texts. They predict the next token based on contextual information from the current and preceding steps.
- **Encoder-decoder hybrid** (e.g., GLM): MLLMs with encoder-decoder architecture enable themselves to process sequential data and generate accurate and coherent outputs that excel in tasks such as text generation, and summarization.

2.2.2 Pre-training technique

Pre-training technique aims to learn universal language representations from billion-scale unlabeled corpora (e.g., Wikipedia, webpages, news, etc.) and then initializes the parameters of the Transformer-based MLLMs. This approach reduces the reliance on massive parallel corpora, helping MLLMs generate similar representations in a common vector space for similar sentences and words (or words in similar context) across languages [54].

The benefits of pre-training technique can be attributed to two key factors: Paradigm and Task. Pre-training paradigms have been proposed to capture linguistic patterns in the training data and adapt MLLMs to downstream tasks, including "pre-training + fine-tuning" and "pre-training + prompting". The former representative models are BERT [2], GPT-2 [19], while the latter presentative models like GPT-3 [20]. Pre-training tasks improve the ability of MLLMs to encode and generate coherent multilingual text.

When learning the universal representation of language, pre-training tasks play a crucial role and the widely used pre-training tasks include probabilistic language modeling (LM), masked language modeling (MLM), next sentence prediction (NSP), and Denoising autoencoder (DAE). Probabilistic LM is a fundamental task in NLP, estimating the probability distribution of sequences of words in a language. In practice, LM typically involves auto-regressive LM or unidirectional LM. MLM has emerged as a novel pre-training task to overcome the drawback of the standard unidirectional LM. By masking certain tokens in a sequence and predicting them based on context, MLM encourages models to learn bidirectional representations, capturing dependencies from both left and right contexts. Punctuations are the natural separators of text data. So, it is reasonable to construct pre-training methods by utilizing them. NSP is just a great example of this. NSP encourages the model to understand the contextual coherence and relationships between sentences. DAE takes a partially corrupted input and aims to recover the original undistorted input. Specific to language, a sequence-to-sequence model, such as the standard Transformer, is used to reconstruct the original text. Eq. (1) summarize loss function $\mathcal{L}$ of these pre-training tasks, where $\mathrm{x} = [x_1, x_2, \ldots, x_T]$ denotes a sequence of tokens [55].

$$\begin{aligned}
\mathcal{L}_{\mathrm{LM}} &= -\sum_{t=1}^{T} \log p(x_t|\mathrm{x}_{<t}), \\
\mathcal{L}_{\mathrm{MLM}} &= -\sum_{\hat{x}\in m(\mathrm{x})} \log p(\hat{x}|\mathrm{x}_{\setminus m(\mathrm{x})}), \\
\mathcal{L}_{\mathrm{NSP}} &= -\log p(t|\mathrm{x},\mathrm{y}), \\
\mathcal{L}_{\mathrm{DAE}} &= -\sum_{t=1}^{T} \log p(x_t|\hat{\mathrm{x}},\mathrm{x}_{<t}),
\end{aligned} \tag{1}$$

where the symbols and their definitions are listed in Table 2.

2.2.3 Reinforcement learning with human feedback

MLLMs may generate inaccurate or harmful outputs due to their probabilistic statistical text generation mechanism [56]. Reinforcement Learning from Human Feedback (RLHF) and its variants [57–61] have been proposed to fix this by optimizing MLLMs with human feedback, aligning them better with human values in three fundamental dimensions: helpfulness, honesty, and harmlessness [62].

Table 1 An overview of representative MLLMs in recent years, including their release time, publishing authority, maximum available parameters, maximum supported context length, pre-training file size, architecture, base model, pre-training function, publicly available, and modal

Model	Release time	Publishing authority	Params	Context length	Pre-training file size	Architecture	Base model	Pre-training function	Publicly available	Modal
					Monolingual					
GPT-1 [18]	Jun-18	OpenAI	117M	2K	–	Decoder-only	GPT	LM	Open	Text
BERT [2]	Oct-18	Google	340M	2K	1.3 GB	Encoder-only	–	Seq2Seq MLM	Open	Text
GPT-2 [19]	Feb-19	OpenAI	1.5B	2K	40 GB	Decoder-only	GPT	LM	Open	Text
T5 [23]	Oct-19	Google	11B	2K	21 GB	En-decoder	–	Seq2Seq MLM	Open	Text
GPT-3 [20]	May-20	OpenAI	175B	2K	570 GB	Decoder-only	GPT	LM	Closed	Text
Gopher [38]	Dec-21	DeepMind	280B	2K	–	Decoder-only	–	LM	Open	Text
					Multilingual					
mBERT [2]	Jul-19	Google	172M	2K	–	Encoder-only	BERT	MLM	Open	Text
XLM-R [7]	Nov-19	Facebook	550M	2K	–	Encoder-only	–	TLM	Open	Text
mBART [26]	Jan-20	Facebook	680M	2K	–	En-decoder	BART	DAE	Open	Text
mT5 [4]	Oct-20	Google	13B	2K	–	En-decoder	T5	Seq2Seq MLM	Open	Text
LaMDA [30]	Jan-22	Google	137B	32K	–	Decoder-only	–	LM	Open	Text
PaLM [29]	Apr-22	Google	540B	2K	–	Decoder-only	–	LM	Closed	Text
BLOOM [5]	Jul-22	BigScience	176B	2K	350 GB	Decoder-only	–	LM	Open	Text
GLM-130B [33]	Aug-22	ZHIPU	130B	2K	–	En-decoder	GLM	ABI	Closed	Text
FLAN-T5 [39]	Oct-22	Google	11B	2K	17.3 MB	En-decoder	T5	LM	Open	Text
GPT-3.5 [20]	Nov-22	OpenAI	175B	2K	–	Decoder-only	GPT	LM	Closed	Text
ChatGPT [40]	Nov-22	OpenAI	175B	2K	–	Decoder-only	GPT-3.5	LM	Open	Text
LLaMA [6]	Feb-23	Meta	65B	4K	120 GB	Decoder-only	–	LM	Open	Text
ChatGLM [38]	Mar-23	ZHIPU	130B	2K	8 GB	En-decoder	GLM	ABI	Open	Text
PaLM-E [41]	Mar-23	Google	562B	2K	–	Decoder-only	PaLM	LM	Open	Text, Image
Alpaca [42]	Mar-23	StandFord	7B	2K	14 GB	Decoder-only	LLaMA	LM	Open	Text
GPT-4 [22]	Mar-23	OpenAI	–	8K	–	Decoder-only	GPT	LM	Closed	Text, Image
PanGu-Σ [43]	Mar-23	Huawei	1085B	–	–	Decoder-only	PanGu-α	LM	Closed	Text
Pythia [44]	Apr-23	EleutherAI	12B	2K	24 GB	Decoder-only	–	LM	Open	Text
PaLM 2 [45]	May-23	Google	340B	2K	–	Decoder-only	PaLM	LM	Closed	Text
ChatGLM2 [32]	Jun-23	ZHIPU	12B	4K	–	En-decoder	GLM	ABI	Closed	Text
Vicuna [34]	Jun-23	LMSYS	33B	2K	65 GB	Decoder-only	LLaMA	LM	Open	Text
LLaMA 2 [46]	Jul-23	Meta	70B	4K	129 GB	Decoder-only	LLaMA	LM	Open	Text
Bard [47]	Jul-23	Google	–	–	–	Decoder-only	LaMDA	LM	Open	Text, Image
Baichuan [48]	Jul-23	BAICHUAN	13B	4K	26.6 GB	Decoder-only	–	LM	Open	Text
GPT-4V [22]	Sep-23	OpenAI	–	32K	–	Decoder-only	GPT-4	LM	Closed	Text, Image
ChatGLM3 [33]	Oct-23	ZHIPU	6B	32K	12 GB	En-decoder	GLM	ABI	Open	Text
GPT-4 Turbo [22]	Nov-23	OpenAI	–	128K	–	Decoder-only	GPT-4	LM	Closed	Text, Image, Speech
Gemini-ultra [35]	Dec-23	DeepMind	300B	32K	–	Decoder-only	–	LM	Closed	Text, Image
Phi-2 [49]	Dec-23	Microsoft	2.7B	2K	5.4 GB	Decoder-only	–	LM	Open	Text
GLM-4 [50]	Jan-24	ZHIPU	–	128K	–	En-decoder	GLM	ABI	Closed	Text, Image
Claude 3 [51]	Mar-24	Anthropic	–	200K	–	Decoder-only	Claude	LM	Closed	Text, Image
LLaMA 3 [52]	Apr-24	Meta	70B	8K	140 GB	Decoder-only	LLaMA	LM	Open	Text

Table 2 Summary of key notations

Symbol	Description
x	$x = [x_1, x_2, \ldots, x_T]$
m(x)	The masked tokens
$x_{\backslash m(x)}$	The rest tokens
$\hat{x}$	Random perturbation text
y	Target label
t	If x and y are continuous segments from corpus, $t = 1$
$\mathcal{P}$	The prompt
y_w	The better model response
y_l	The worse model response
r_θ	The output of reward model
θ	Weight in policy optimization network
$J(\theta)$	The objective function
$\nabla_\theta J(\theta)$	The gradient of objective function

Essentially, RLHF has three core steps [63,64], as shown in Fig. 3.

1. **Pre-training a language model**: To pre-train MLLMs, extensive prompts, and multilingual datasets are utilized as examples, teaching the model how to respond appropriately in a specific context.
2. **Training a reward model**: Prompts serve as input to MLLMs, where K pairs of prompt, response are manually scored by human evaluators to align with human preferences. The rankings of sample, reward pairs are normalized into a scalar reward signal for training the Reward Model (RM). The loss function $\mathcal{L}_\theta$ is defined as follows, where $\mathcal{P}$ is the prompt, y_w and y_l denote the better and worse model responses

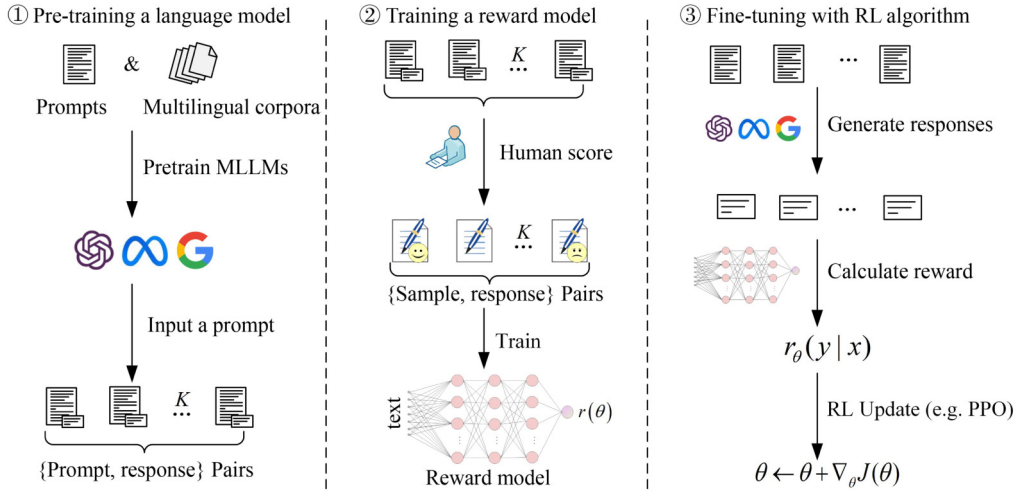

Fig. 3 Diagram illustrating the RLHF procedure, which consists of three key steps: (1) Pre-training a LM using the labeled prompt-response dataset, (2) Training a Reward Model based on scores provided by human evaluators for LM's generation, and (3) Fine-tuning with a Reinforcement Learning (RL) algorithm, which helps to update parameters in the LM based on the feedback from RM

respectively, and $r_\theta(\mathcal{P}, y)$ is the output of the RM.

$$\mathcal{L}_\theta = -\frac{1}{\binom{K}{2}} E_{(\mathcal{P},y_w,y_l)\sim D}[\log(\sigma(r_\theta(\mathcal{P},y_w) - r_\theta(\mathcal{P},y_l)))], \quad (2)$$

where the symbols and their definitions are listed in Table 2.

3. **Fine-tuning with reinforcement learning**: MLLMs are optimized using Proximal Policy Optimization (PPO) [65] or a similar reinforcement learning (RL) algorithm such as A2C [66]. These RL algorithms incorporate human-generated reward signals to apply gradient strategies for learning human feedback.

2.3 Multilingual capacities of MLLMs

Pre-training MLLMs on extensive multilingual data enhances their multilingual capacities and cross-lingual transfer learning (CLTL) from one language to another. However, MLLMs still face challenges in training with multilingual corpora and their exact CLTL capabilities remain largely unexplored. This section focuses on these two concerns.

2.3.1 Challenges brought by multilingual corpora

Three challenges arise from multilingual corpora training. Firstly, while MLLMs outperform monolingual LLMs in downstream tasks for high-resource languages, their performance on low-resource languages remains unsatisfactory due to limited annotated data. Secondly, the "curse of multilinguality" phenomenon in MLLMs worsens this situation. Supporting more languages can lead to a significant decline in performance for low-resource languages, making them victims of this curse [67]. Thirdly, the distribution of languages in the pre-training corpora is highly skewed towards English, further complicating efforts to address the "curse of multilinguality" phenomenon.

To mitigate these challenges, two approaches have been proposed. One involves fine-tuning existing MLLMs to suit the linguistic features of low-resource languages [68]. However, this method is constrained by the demand for extensive specific-task annotated training data [69]. Alternatively, another approach is to pre-train monolingual LLMs specifically for low-resource languages [70]. This method allows models to learn from diverse sources and contexts within the target language without requiring costly annotated data. Therefore, MLLMs trained by this approach exhibit superior performance on low-resource languages compared to the aforementioned fine-tuning approach. For example, Torge et al. [71] pre-trained monolingual RoBERTa models for Czech and Polish, as well as a bilingual model for Czech-Polish, which demonstrated superior performance to the current state-of-the-art multilingual model, XLM-R, across various downstream tasks. Recently, there has been a growing interest in developing low-resource language models to meet the demands of morphologically rich, low-resource languages. Examples include language-specific BERT models like FlauBERT for French [15], BERTje for Dutch [16], and FinBERT for Finnish [72], among others [54].

The main reason for MLLMs' poor performance on low-resource languages is the skewed distribution of languages in the pre-training data. Therefore, techniques have been proposed to address this issue. Data sampling techniques like exponential weighted smoothing [7] help prevent the under-representation of low-resource languages, while vocabulary augmentation approaches [73] enrich the model's vocabulary by inducing new tokens of unseen languages during training. Moreover, research has also attempted to tackle the language imbalance. Choenni et al. discovered that languages influence each other during the pre-training phase and MLLMs benefit from reinforcement or complementary learning [74]. Wang et al. emphasized the significance of imbalanced learning algorithms in Vision-Language models (VLMs) [75]. For example, CLIP model demonstrated an improvement from 5% to 69% on iNaturalist dataset by adopting imbalanced methods. Jiang et al. proposed a data augmentation pipeline to address imbalance in social media data, effectively handling multiclass problems [76].

2.3.2 Cross-lingual transfer learning brought by multilingual corpora

MLLMs can facilitate CLTL from one language to another.

This naturally raises the question of how much CLTL capability that MLLMs possess to support these unseen languages or low-resource languages during pre-training.

Research has been dedicated to exploring the cross-lingual transferability of MLLMs through zero-shot learning. Lin et al. [77] trained 4 multilingual generative language models and examined their zero-shot and in-context few-shot learning capabilities in a wide range of tasks. They found that these models can achieve cross-lingual few-shot learning in non-English languages without requiring source-to-target language translation. Tian et al. [78] found that MLLMs exhibit strong rumour detection performance in zero-shot cross-lingual transfer learning. What's more, MLLMs showed surprisingly strong multilingual reasoning abilities even in under-represented languages such as Bengali and Swahili [79].

To further improve the transfer learning performance of MLLMs on unseen or low-resource languages, as these languages still account for a significant portion of the world's languages, MLLMs are pre-trained to learn languages from the same linguistic family or branch [80,81]. MLLMs trained on a small amount of data from genetically related languages could achieve performance comparable to the ones trained on large but unrelated data [80]. MLLMs trained on only low-resource languages with small datasets, which are similar to each other, sometimes achieved better performance than models trained on large datasets with high-resource languages [81]. For example, the AfriBERTa model [81], pre-trained on less than 1 GB of text data from 11 African languages, most of which belong to the Bantu branch of the Niger-Congo language family, demonstrated the effectiveness of scratching solely on low-resource languages without any high-resource transfer learning.

A prominent future concern will be how to improve the CLTL capacities of MLLMs. Pikuliak et al. conducted a survey on existing cross-lingual transfer paradigms of MLLMs [82], while Philippy et al. investigated various factors that impacted cross-lingual transfer performance, including linguistic similarity, lexical overlap, model architecture, pre-training setting, and pre-training corpus size [83]. Specifically, this avenue of research seeks to investigate how and why MLLMs possess different CLTL abilities on various languages. This pursuit holds the potential to leverage CLTL capacities to mitigate the dependence on annotated data and maintain or even enhance the performance of MLLMs in well-trained or unseen languages.

3 Multilingual corpora and datasets

In this section, we delve into the widely utilized multilingual corpora that are associated with the training of MLLMs, and multilingual datasets oriented for downstream tasks.

Table 3 summarizes the multilingual corpora that representative MLLMs trained on, offering insights into their language distribution, data source, and language coverage.

MLLMs have a more extensive language coverage in their training data compared to LLMs. A significant portion of these training data originates from multilingual repositories like Common Crawl, Wikipedia, and web documents, encompassing a broad range of languages. These multilingual repositories are crucial for enhancing the cross-lingual capability of MLLMs. In this section, we discuss training data's language composition from both a general perspective and a language family perspective.

3.1 Multilingual corpora in MLLMs

First, we analyze the linguistic composition of MLLMs' training data, investigating the total number of languages and different language proportions within each training corpora. Analysis reveals that most MLLMs are trained on corpora where English is the predominant language. Notably, several MLLMs, including GPT-3 [20], Gopher [38], LaMDA [30] and InstructGPT [21], are trained on corpora where English comprises over 90%. The overwhelming English texts in corpora lead to MLLMs' English-centric ability. To alleviate this issue, some MLLMs are trained on corpora with more balanced language distribution. For example, the training data of BLOOM [5] covers 46 languages, with English comprising less than half. YAYI 2 [85] makes great efforts to balance its training data, achieving a nearly 1:1 ratio between English and Chinese training data. Compared to its base model PaLM [29], PaLM 2 [45] includes a higher percentage of non-English data, further enhancing its multilingual capabilities. We present the percentages of its non-English languages in the web documents subset of its pre-training corpus in Table 3, as the language distribution for English was not published.

Second, we explore the language composition of MLLMs' training data from a language family perspective. Languages within the same language family share similar characteristics and MLLMs have better transfer performance on languages belonging to the same language family [54]. Thus, the proposition of language families in MLLMs' training data can help us better understand the multilingual capabilities of MLLMs. What's more, we can also leverage language families to observe the linguistic composition of the MLLMs' training data. Since English is predominant in most MLLMs' corpora, considering it in language family analysis would heavily favor the Indo-European language family to which English belongs. To gain a more detailed understanding of the language family proposition in MLLMs' corpora, we exclude English and focus on the top 20 prominent non-English languages of the training data and their corresponding language families. The distribution of language families of each MLLM is shown in Fig. 4.

Notably, French, German, Chinese, and Spanish emerge as the most prevalent languages in the training data. For example, French constitutes 1.8% of the training corpora for GPT-3 [20] and 12.9% of the training corpora for PaLM [29]. French, German, and Spanish all belong to the Indo-European language family, which demonstrates that the Indo-European language family holds a prominent position in MLLMs' corpora, both in terms of quantity and linguistic diversity. An exception to this is Chinese, which belongs to the Sino-Tibetan language family while maintaining a significant presence in the training corpora. But in terms of linguistic diversity, the Sino-Tibetan language family in the training corpora, mainly consisting of the Chinese language, is much less diverse compared to the Indo-European language family.

Table 3 An overview of representative multilingual training corpora of MLLMs in recent years, including their corresponding model, language, language proportion, and source

Model	Language	Language proportion	Source
mBERT [2]	104 languages	Unknown	Wikipedia
XLM-R [7]	100 languages	English (12.56%); Russian (11.61%); Indonesian (6.19%); Vietnamese (5.73%); Others (63.89%)	Generated using the open source; CC-Net repository
mT5 [4]	101 languages	English (5.67%); Russian (3.71%); Spanish (3.09 %); German (3.05%); Others (84.48%)	Common Crawl
GPT-3 [20]	95 languages	English (92.7%); French (1.8%); German (1.5%); Others (5.9%)	Common Crawl; Wikipedia; Books1; Books2; WebText2
Gopher [38]	51 languages	Over 99% English	MassiveWeb (48%); C4 (10%); News (10%); Books (27%); GitHub (3%); Wikipedia (2%)
LaMDA [30]	Unknown	Over 90% English	Public dialog data and other public web documents
PaLM [29]	Over 100 languages	English (77.98%); German (3.50%); French (3.25%); Spanish (2.11%); Others (13.15%)	Social media conversations (50%); Filtered webpages (27%); Books (13%); GitHub (5%); Wikipedia (4%); News (1%)
BLOOM [5]	46 languages	English (30.03%); Simplified Chinese (16.16%); French (12.9%); Spanish (10.85%); Portuguese (4.91%); Arabic (4.6%); Others (20.55%)	Web Crawl(38%); BigScience Catalogue Data(62%)
LLaMA [6]	Over 20 languages	Over 67% English	Common Crawl (67.0%); C4 (15.0%); Github (4.5%);Wikipedia (4.5%); Books (4.5%); ArXiv (2.5%); StackExchange (2.0%)
Vicuna [34]	Unknown	Unknown	User-shared conversations from ShareGPT.com
Falcon [84]	Over 100 languages	Excluding English: Russian (13.19%); German (10.81%); Spanish (9.45%); Others (66.55%)	Common Crawl
PaLM 2 [45]	Over 100 languages	Excluding English: Spanish (11.51%); Chinese (10.19%); Russian (8.73%); Others (69.57%)	Web documents; Books; Code; Mathematics; Conversational data
LLaMA 2 [46]	Over 100 languages	English (89.70%); Unknown (8.38%); German (0.17%); France (0.16%); Others (1.59%)	Publicly available sources excludes Meta user data
GPT-4 [22]	Over 26 languages	Unknown	Common Crawl; Wikipedia; Books1; Books2; WebText2
LLaMA 3 [52]	176 languages	Over 5% non-English	Publicly available sources excludes Meta user data
GLM-4 [50]	26 languages	Mostly English and Chinese	Webpages;Wikipedia;Books; Code;Research papers
Claude 3 [51]	Over 43 languages	Unknown	Publicly available information on the Internet; Non-public data from third partie; Data provided by data labeling services and paid contractors; Data they generate internally
YAYI 2 [85]	Over 10 languages	Chinese (41.5%); English (40.4%); French (2.5%); Spanish (2.2%); Others (25%)	Web pages; Social media; Books; Newspapers; Academic Papers
FuxiTranyu [86]	43 languages	English (28.8%); Chinese (10.4%); German (8.1%); French (7.3%); Others (45.4%)	Web (82%); Code (9%); Paper (3%); Book (3%); Encyclopedia(2%); Report(1%)

Besides Indo-European and Sino-Tibetan language families, some other language families are found in most MLLMs' training corpora as well. Similar to Sino-Tibetan, they mainly contain only one language in training corpora. For example, Austronesian mainly includes Indonesian, Japonic mainly includes Japanese, and Koreanic mainly includes Korean. Apart from the lack of diversity within the same language family, there is also a lack of diversity across different language families in MLLMs' training corpora. For example, despite Niger-Congo and Trans-New Guinea being among the largest language families in the world, they are notably absent from the top 20 languages in the training data.

Through the above analysis of multilingual training corpora in MLLMs, we have derived the following key insights: MLLMs broaden language coverage beyond LLMs, yet English remains dominant in their training corpora. From a language family perspective, Indo-European languages occupy a prominent place in terms of both quantity and linguistic variety. Further work should consider a more comprehensive inclusion of language families and prioritize linguistic diversity within the same language family when training MLLMs.

3.2 Multilingual datasets for downstream tasks

Multilingual datasets play a crucial role in fine-tuning MLLMs to be adaptive across various NLP tasks. Table 4 summarizes some representative multilingual datasets, including Multilingual Named Entity Recognition (Multilingual NER), Multilingual Sentiment Analysis (Multilingual SA), Cross-Lingual Information Retrieval

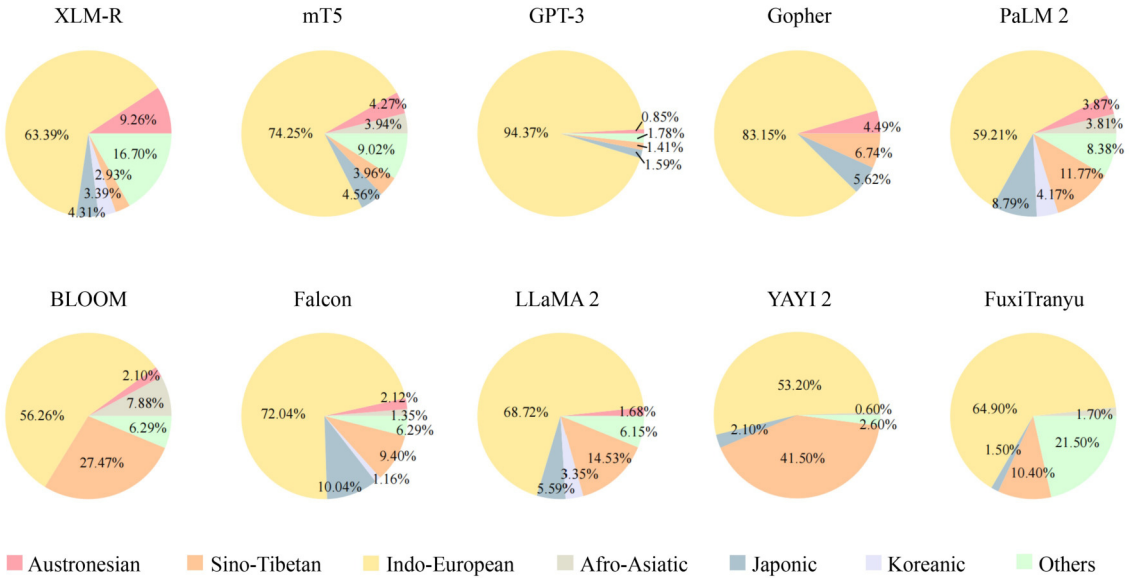

Fig. 4 This analysis excludes English and focuses on ratios of language families of languages (top 20) in MLLM's corpora. Note that Gopher only released the top 10 languages and FuxiTranyu only released the top 13 languages used in training corpora. What's more, some of the latest models like GPT-4 have not disclosed the proportion of their training data, so they aren't included in the chart

Table 4 An overview of representative multilingual datasets for downstream tasks in recent years, including their corresponding task, release name, language, size, and source

Task	Dataset	Language	Source	Size
Multilingual NER	Masakha NER2.0 [87]	20 African languages	News articles	4.8K to 11K sentences per language
Multilingual NER	MultiCo NER [88]	11 languages	Wikipedia; ORCAS dataset MS-MARCO QnA corpus	26M tokens
Multilingual SA	XED [89]	32 languages	OPUS	More than 950 lines per language
Multilingual SA	NollySenti [90]	5 languages	Movie reviews	1K to 1.5K reviews per language
Multilingual SA	NaijaSenti [91]	5 languages	Twitter	30K tweets
Crosslingual IR	AfriCLIR Matrix [92]	15 African languages	Wikipedia	6M English queries and 23M relevance judgments
Crosslingual IR	CLIRMatrix [93]	8 languages	Wikipedia	49M unique queries and 34B (query, document, label) triplets
Multilingual TC	Taxi1500 [94]	Over 1500 languages	Parallel translations of the Bible	About 1K verses per language
Multilingual TC	MARC [95]	6 languages	Amazon reviews	210K reviews per language
Multilingual Versatile	MUSE [96]	110 language pairs	Self-created	About 6.5K word pairs for each language pair
Multilingual Versatile	Wikipedia monolingual corpora [97]	30 languages	Wikipedia	10B tokens
Multilingual Versatile	Multilingual open text [98]	44 languages	VOA News	Over 2.8M news articles and an additional 1M short snippets

(Cross-Lingual IR), and Multilingual Text Classification (Multilingual TC).

Multilingual NER. Named Entity Recognition tasks locate and classify named entities from unstructured natural languages. These tasks utilize datasets from sources like News and Wikipedia, which provide rich contextual information across a wide range of real-world entities. Efforts have been made to expand the training data for low-resource languages. A notable example is Masakha NER2.0 [87], the largest human-annotated Africa-centric dataset, deriving its data from African local news.

Multilingual SA. Sentiment analysis tasks, which focus on the sentiment orientation of data, often utilize datasets extracted from comments or reviews found on review platforms such as Amazon and IMDb, as well as social media platforms like Facebook and Twitter. The sentiment analysis dataset XED [89] is sourced from OPUS [99], a parallel corpus extracted from movie subtitles. In terms of linguistic diversity, while XED [89] primarily focuses on English and Finnish, NollySenti [90] and NaijaSenti [91] are sentiment analysis datasets specifically designed for African languages such as Hausa, Igbo, Nigerian, Pidgin and Yoruba.

Cross-lingual IR. Cross-Lingual Information Retrieval tasks ask queries in one language and retrieve documents in one or more other languages. These tasks utilize datasets that include documents containing hyperlinks to parallel documents in different languages. Therefore, many datasets such as AfriCLIR Matrix [92] and CLIR Matrix [93] are sourced from multilingual encyclopedias (e.g., Wikipedia). CLIR Matrix [93] is the current largest and most comprehensive CLIR dataset. It includes Arabic, German, English, Spanish, French, Japanese, Russian, and Chinese,

covering mainly the common languages of all continents except Africa. Thus, AfriCLIR Matrix [92] was developed to address the absence of African languages.

Multilingual TC. Text Classification tasks have diversified applications on news classification, sentiment classification and so on. These tasks utilize diverse datasets tailored for specific applications. For example, Multilingual Amazon Reviews Corpus (MARC) [95], which includes product category and star rating, can be used for both product classification and sentiment classification. Taxi1500 [94], covering more than 1500 languages, relies solely on the parallel translation of the Bible as its data source, limiting its domain to religious-related text classification only. However, as Bible is the most translated book, its parallel translation is a good data source to enhance linguistic diversity in datasets.

Multilingual versatile. Besides the multilingual datasets mentioned above, Wikipedia Monolingual Corpora [97], MUSE [96] and Multilingual Open Text (MOT) [98] are widely used for general NLP tasks. Wikipedia Monolingual Corpora [97] covers 30 languages. Each language has its own XML file, containing the full monolingual Wikipedia contents, with annotations like article and paragraph boundaries, the number of links referring to each article, cross-language links and more. MUSE [96] provides state-of-the-art multilingual word embeddings aligned in a single vector space for 30 languages and 110 large-scale ground-truth bilingual dictionaries. Multilingual Open Text (MOT) [98] comprises news articles and short snippets (photo captions, video descriptions, etc.) from Voice of America (VOA) news websites. It was designed to supply high-quality unlabeled texts for lower-resource languages like Albanian, Amharic and Persian. It contains a complete collection of VOA's documents which can be further annotated for various NLP tasks (e.g., document classification, syntactic or semantic parsing).

4 Multilingual representation alignment

The success of MLLMs is their ability to achieve multilingual representation alignment from multiple languages. Table 5 summarizes some multilingual alignment performance of MLLMs on 10 languages and three cross-lingual tasks: bilingual lexicon induction (BLI), cross-lingual classification (XNLI), and machine translation (MT). The alignment is from languages of Spanish (ES), German (DE), French (FR), Russian (RU), Arabic (AR), Chinese (ZH), Bulgarian (BG), Turkish (TR), and Hindi (HI) to English (EN), respectively. The evaluation metrics include accuracy (for BLI and XNLI) and BLEU (for MT). The performance of MLLMs on multilingual alignment varies across languages, with better performance observed for English and its closely related languages.

Aligning the representation of diverse languages acts as an integral part of NLP's multilingual tasks and applications [8]. Inspired by the impressive performance of monolingual representation models like Word2vec [9] and GloVe [10], recent research has made great progress in multilingual representation. Figure 5 summarizes the evolution of multilingual representation from static approaches to more dynamic ones like contextual and combined multilingual representations. This evolution is highly influenced by the introduction of MLLMs and their enhanced multilinguality.

As shown in Fig. 6, Static multilingual representations are attained through learning a mapping matrix to align two

Table 5 A demonstration of the multilingual alignment performance of MLLMs on 10 languages, taking BLI, XNLI, and MT tasks as examples [100,101]. Bold text denotes the best performance across models. √ and × mean that the performance of MLLMS in a certain language is higher and lower than the average performance, respectively

Task	Evaluation metric	Model	ES	DE	FR	RU	AR	ZH	BG	TR	HI	Avg.
BLI	Acc.	fastText [102]	**72.00**	**67.17**	–	56.42	47.43	33.39	45.69	48.92	28.19	49.90
		BLOOM-7B [5]	52.50	38.34	–	26.06	32.67	34.35	16.75	30.82	28.30	32.47
		LLaMA-13B [6]	60.58	57.80	–	64.44	22.13	32.28	56.86	44.90	30.68	46.21
		GPT-3.5 [20]	68.17	63.07	–	**74.15**	**65.94**	**65.12**	**67.51**	**54.49**	**56.11**	**64.32**
		Average	63.31	56.60	–	55.27	42.02	41.29	46.70	44.78	35.82	48.23
		Stddev(σ)	8.63	12.76	–	20.78	19.01	15.91	21.87	10.10	13.58	13.09
		√ or ×	√	√	–	√	×	×	×	×	×	–
XNLI	Acc.	mBERT [2]	68.0	70.0	64.3	73.4	67.8	60.9	73.5	58.9	57.2	66.00
		mT5-270M [4]	78.6	77.4	73.3	79.1	77.1	72.8	80.3	70.8	68.3	75.30
		XLM-R-270M [7]	80.7	78.7	79.7	78.1	73.8	76.7	79.6	74.2	72.4	77.10
		mT5-10.7B [4]	**87.7**	**87.3**	84.5	**86.9**	**85.1**	**83.8**	**87.8**	**83.2**	**79.8**	**85.12**
		XLM-R-10.7B [7]	87.3	87.0	**86.2**	82.5	82.5	82.6	85.7	82.0	**79.8**	83.96
		Average	80.46	80.08	77.68	80.0	77.26	75.36	81.38	73.82	71.50	77.50
		Stddev(σ)	8.03	7.26	8.96	5.05	6.90	9.23	5.62	9.83	9.40	7.70
		√ or ×	√	√	√	√	×	×	√	×	×	–
MT	BLEU	XGLM-7.5B [77]	27.98	34.03	36.81	27.83	26.06	6.06	34.48	23.91	26.99	27.13
		OPT-175B [31]	30.81	39.15	43.02	18.80	1.03	12.36	11.48	24.39	1.17	20.25
		Falcon-7B [84]	30.13	34.60	41.62	14.26	1.81	22.78	8.07	10.05	1.26	18.29
		LLaMA2-7B [46]	33.09	41.94	44.11	33.44	22.35	26.26	38.18	21.75	21.04	31.35
		ChatGPT [40]	33.48	43.56	46.13	38.04	38.94	30.05	41.65	38.14	38.15	38.68
		GPT-4 [22]	**33.76**	**47.04**	**48.81**	**38.75**	**43.29**	**32.83**	**44.97**	**43.43**	**45.88**	**42.09**
		Average	31.54	40.05	43.42	28.52	22.25	21.72	29.81	26.95	22.42	29.62
		Stddev(σ)	2.29	5.13	4.10	10.18	17.91	10.46	15.94	12.04	18.55	9.62
		√ or ×	√	√	√	×	×	×	√	×	×	–

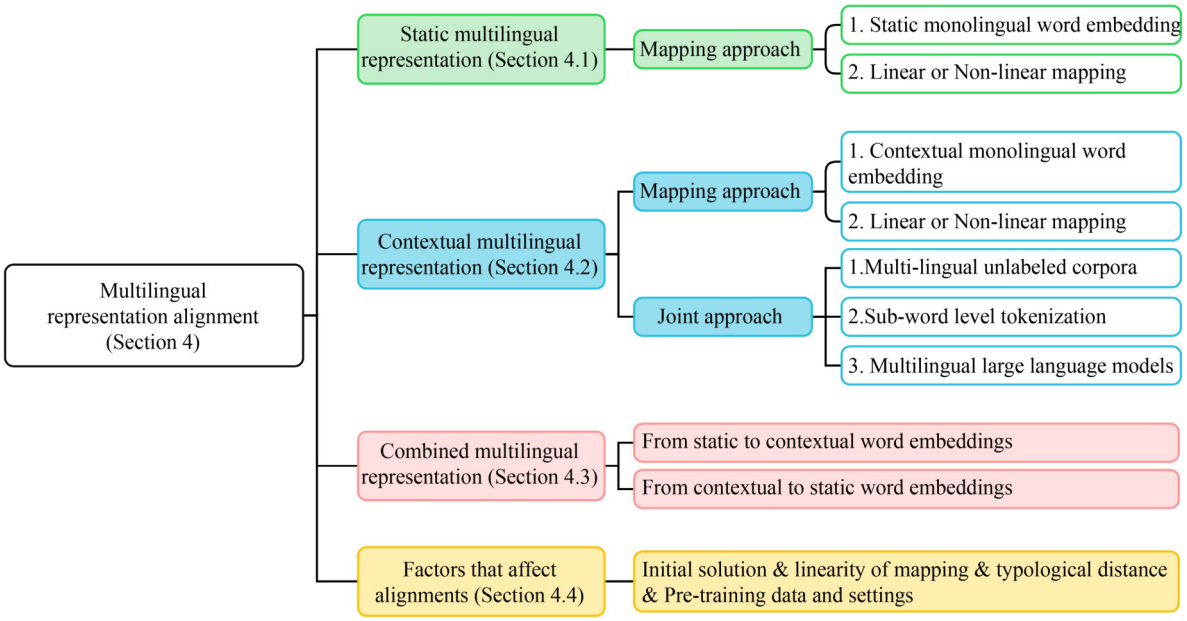

Fig. 5 Taxonomy of multilingual representation alignment that consists of static, contextual, and combined approaches. In addition, we also summarize the factors that affect alignments

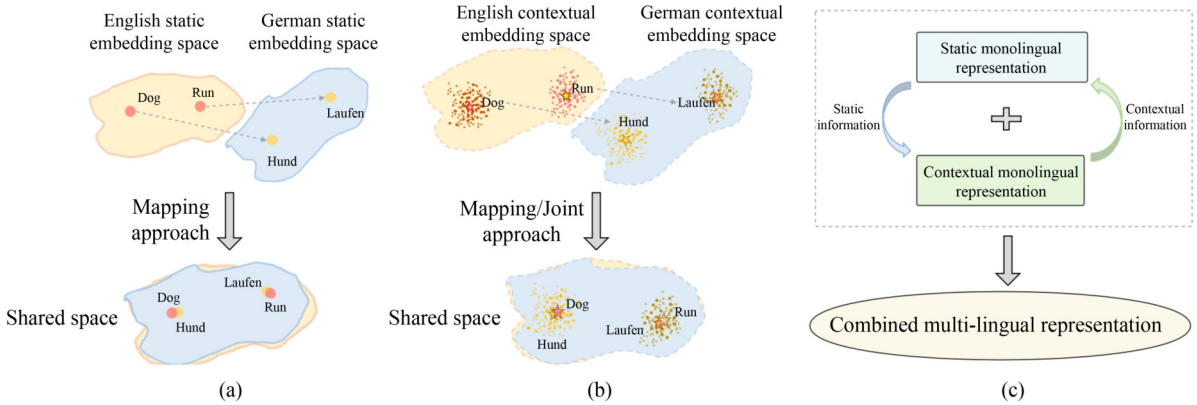

Fig. 6 An illustration of three approaches of multilingual representation alignment. English words are marked in red, while German words are in yellow, and one point represents an embedding. (a) Static approach, where a one-to-one correspondence exists between points and words; (b) contextual approach, where each word has multiple corresponding embeddings; (c) combined approach

monolingual embedding spaces, while contextual ones can be achieved by both mapping and joint approaches, with the latter being supported by MLLMs. To achieve even better alignment, combined methods were proposed to take advantage of both static and contextual information. Details of the three paradigms will be explained below. Furthermore, we also discuss the factors that will affect multilingual alignment.

4.1 Static multilingual representation

Based on whether parallel corpora are used or not, static alignment approaches can be categorized into three groups: supervised, semi-supervised, and unsupervised approaches. Recently, unsupervised approaches, such as MUSE [96] and VecMap [103], have gained much more attention.

Let X and Y represent monolingual word embeddings from two languages, respectively. Static alignment approaches can be roughly divided into two steps: initially, the introduction of an initial mapping by aligning the source and target language distributions; subsequently, a pseudo-supervised refinement based on the initial solution, where the transformation matrix W is constrained to be orthogonal, i.e., $W^{\mathrm{T}}W = I$.

$$W^* = \arg\min_W \|WX - Y\|. \tag{3}$$

Orthogonal constraint serves as a method to ensure monolingual invariance but is not held for all languages, particularly for the semantically distant languages [104]. Therefore, weak orthogonal constraints have been proposed to better align the embeddings across different languages.

Generally, linear projection only learns one global transformation matrix W to project the entire embedding space of the source to that of the target. However, the global transformation matrix does not consistently perform optimally across all subspaces [105]. To address this issue, specific mappings for different subspaces have been proposed [106].

Static multilingual representations have exhibited promising performance but there is still ample room for improvement on low-resource languages and distant language pairs. Besides, the polysemy problem in static multilingual representation has not been well addressed and needs further exploration.

4.2 Contextual multilingual representation

Contextual representation is introduced to address the polysemy challenge faced in static representation. ELMo [107] and BERT [2] stand out as the highly representative models for contextual monolingual representation. Contextual multilingual word representation can be derived from these models. The existing approaches to contextual multilingual representation can be categorized into two groups: mapping approach and joint approach.

Mapping approaches use pre-trained contextual monolingual embeddings from various languages as input and project them into a shared semantic space [108]. However, two challenges remain. Firstly, the computation cost for pre-trained monolingual embeddings rises exponentially with the number of languages. Secondly, in contextual approaches, alignment is more challenging compared to static ones. Simply calculating a mapping alone is no longer sufficient to generate robust alignments [8].

In comparison, joint approaches supported by MLLMs belong to an end-to-end process, which no longer requires pre-trained monolingual representations but instead depends on unlabeled multilingual corpora. Tokenization is a critical technique in the end-to-end process, segmenting raw data from various languages into sequences of tokens for subsequent processing by MLLMs. Transformer-based MLLMs commonly employ subword-level tokenizers, such as Byte-Pair Encoding (BPE) [109] and WordPiece [110], to address out-of-vocabulary (OOV) issue. What's more, variants of BPE have been proposed to improve the tokenization of multilingual corpora and alleviate lexical overlap between languages.

In summary, contextual multilingual representation contains richer in-context information than the static multilingual approaches and thus shows greater potential for multilinguality. However, there is still a range of multilingual NLP tasks that contextual multilingual approaches underperformed than static ones, demonstrating that several challenges remain:

1. It comes with higher computational costs and is far more resource-intensive during both training and inference.
2. It is challenging to interpret the generated multilinguality and to transform the MLLMs into multilingual lexical encoders, representative contextual embeddings are hard to extract and interpret properly [111,112].
3. The alignment between low-resource languages and distant languages pairs has not been well investigated.

4.3 Combined multilingual language representation

Combined multilingual representation has been proposed to take advantage of both static and contextual paradigms. The existing combined multilingual approaches can be divided into two paradigms: (1) From Static to Contextual (S2C), leveraging static information to induce better contextual multilingual alignment [113]; (2) From Contextual to Static (C2S), leveraging contextual information to induce better static multilingual alignment [114,115].

S2C achieves higher-quality contextual representation by integrating extra static instruction, while C2S achieves higher-quality static representation by integrating extra contextual information. Although S2C makes contextual approaches easier to interpret, the accurate extraction of contextual representations from MLLMs is still a challenge.

Therefore, C2S is a better way for multilingual representation alignment. Existing C2S can be divided into two steps: 1) roughly achieving static multilingual representations, as introduced in Section 4.1; 2) fine-tuning static multilingual representations by leveraging contextual representations. Zheng et al. [114] proposed a spring network to use the contextual representations to pull the static word embeddings to better positions in the unified space for easy alignment. Li et al. [115] fine-tune pre-trained multilingual LMs to extract more useful representations and then combine static and extracted contextual embeddings to achieve high-quality cross-lingual word embeddings.

4.4 Factors that affect alignments

Based on the aforementioned discussion, we delve into the impact of various factors on multilingual alignment performance and investigate which factors have a more significant impact.

Initial solution. For mapping approaches, the initial solution plays a crucial role in alignment. Because subsequent optimization is based on this initial solution, it will affect the robustness of the final result and cause the alignment to fall into a local optimum. Based on their use of annotated data, mapping approaches can be categorized as supervised, semi-supervised, and unsupervised methods. For supervised and semi-supervised methods, the quality of the initial solution depends on the quality and amount of the seed dictionary, while unsupervised ones depend on the robustness and effectiveness of embedding spaces' distribution matching, which is more difficult. GAN-based adversarial training [96], optimal transport solution [116], auto-encoder [117], and graph alignment [118] were utilized to better match distribution and find a better initial solution in a fully unsupervised way.

Linearity of mapping. Mapping functions are always constrained to be orthogonal during training out of the "approximate isomorphism assumption", which fails especially when the two languages are far apart semantically. To address this issue, Mohiuddin et al. [119] and Glavaš and Vulić [120] used a non-linear Mapping function. Marchisio et al. [121] considered relative isomorphism during the process of pre-training monolingual embedding, which can address the misalignment from the root.

Typological distance. More typologically distant language pairs tend to be less well-aligned than more similar ones [122]. In the Bilingual Lexicon Induction (BLI) task, the accuracy on semantically distant language pairs is always under 40%, while similar ones are over 80%. To alleviate this problem, auxiliary languages have been proposed as a medium to bridge the gap between semantically distant language pairs [3,123]. For distant language pairs, one or several more relevant languages can be selected as auxiliary languages. Transferring the additional information provided by the

auxiliary languages, monolingual embedding or corpora can improve the alignment between distant language pairs.

Pre-training data and settings. Pre-training data and settings are found to be correlated with the cross-lingual transfer ability. The size and quality of data are crucial factors for enhanced cross-lingual transfer capabilities in MLLMs. The relative balance and diversity in the pre-training data and the larger data size will improve the efficiency and effectiveness of MLLMs [7]. The settings of pre-training are also important to the cross-lingual performance of MLLMs. The parameters scale [8], pre-training learning objective [124] and window size of input of MLLMs [125] have proved to be influential to cross-lingual transfer ability.

5 Bias on multilingual LLMs

Bias on MLLMs has become a challenging issue to their fairness and severely restricts the deployment of MLLMsin real-world. Research has shown that language models can perpetuate and even exacerbate existing biases present in their training data, which are further manifested in various forms, such as gender bias, cultural bias, and language bias [126]. As shown in Fig. 7, LLMs have different understanding across diverse biases, as evaluated on BBQ question-answer dataset [127].

However, the existing literature on bias mainly focuses on stereotypical biases in English [13,14] or within limited attributes like race and gender [128], which limits its generalizability to other languages or attributes. Bias in MLLMs has not been well investigated. In this section, we aim to address the following questions. Why do MLLMs bias and what are the types of bias in existing MLLMs (Bias Category), how to evaluate bias in MLLMs (Bias Benchmark), how to mitigate the bias, and whether debiasing techniques affect the performance of MLLMs (Debias Technique). Fig. 8 presents the taxonomy of this section.

5.1 Bias category

Bias in MLLMs can arise from factors such as unmoderated training data [129], differences in model design [12], and the presence of biased multilingual word embedding representations [130]. Based on studies related to bias in MLLMs, we categorize these prevalent biases centered around specific languages, limited attributes, and related models into three types: *language bias*, *demographic bias*, and *evaluation bias*. Table 6 presents the bias category, bias source, as well as bias examples.

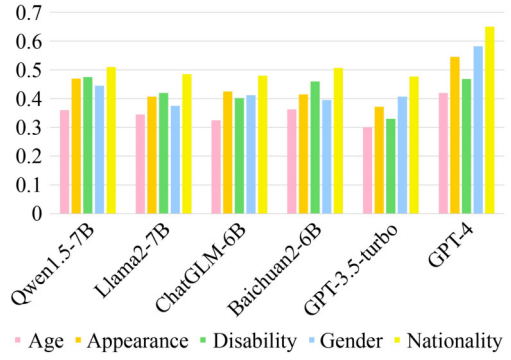

Fig. 7 Accuracy of different LLMs across various bias categories on BBQ question-answer dataset (data from [127])

Language bias. Language bias refers to the unequal performances of MLLMs among different languages, primarily due to the dominance of English and other major languages in the available multilingual training corpora. Specifically speaking, MLLMs exhibit higher proficiency in these widely used languages and this further exacerbated the lack of support for low-resource languages or minority dialects [131]. Recent studies have brought attention to the unequal quality of multilingual representations, highlighting that pre-trained models like mBERT and CLIP do not equally learn high-quality representations for all languages, particularly for low-resource languages [132,133].

When investigating knowledge in MLLMs, Kassner et al. [134] found mBERT exhibited language bias, wherein the choice of query language can impact the obtained results. To go a step further, studies in [135,136] explored how MLLMs exhibited bias across languages and focused on bias in attributes like race, religion, nationality, and gender. They found that mBERT and XLM-R models did not consistently show low-level bias in certain languages [135]; mBERT, XLM-R, and mT5 exhibited varying degrees of fairness across languages and XLM-R exhibited higher and more consistent correlations across languages compared to mBERT and mT5 [136].

Demographic bias. Demographic bias refers to the MLLMs' biased behavior towards specific gender, race, ethnicity, or other social groups, caused by the training data disproportionately emphasizing particular demographic groups [131]. Previous research has shown that both multilingual and monolingual LLMs suffer from demographic bias towards specific social groups [137,138], while monolingual LLMs specific for low-resource languages exhibit less bias [70]. Touileb et al. [137] investigated demographic bias in Norwegian demographics, finding that both language-specific models like Norwegian pre-trained language models and MLLMs like XLM-R demonstrated a bias towards gender-balanced occupations. Likewise, research in [138] discovered that MLLMs like BLOOM and ChatGPT, along with monolingual LLMs trained exclusively on Arabic data, displayed cultural bias towards Western culture. This is evidenced by the fact that when processing and generating Arabic texts, Western-appropriate content is usually preferred over relevant Arabic content. Notably, LLMs for low-resource languages like Sudanese exhibited gender-neutral behavior without displaying distinct biases [70]. Additionally, bias against a particular cultural group is a common manifestation of demographic bias. Levy et al. [135] revealed that mBERT and XLM-R favored culturally dominant groups in each language. GPT-3 has been found to exhibit a stereotypical religious bias for associating Muslims with violence more often than other religious groups [139].

Evaluation bias. Evaluation bias refers to the bias that exists in the evaluation metrics for LLMs. Factors that can bias the metric calculation itself include noise in the evaluation dataset, models used in the metric calculation, and the configuration of the inference experiment [140].

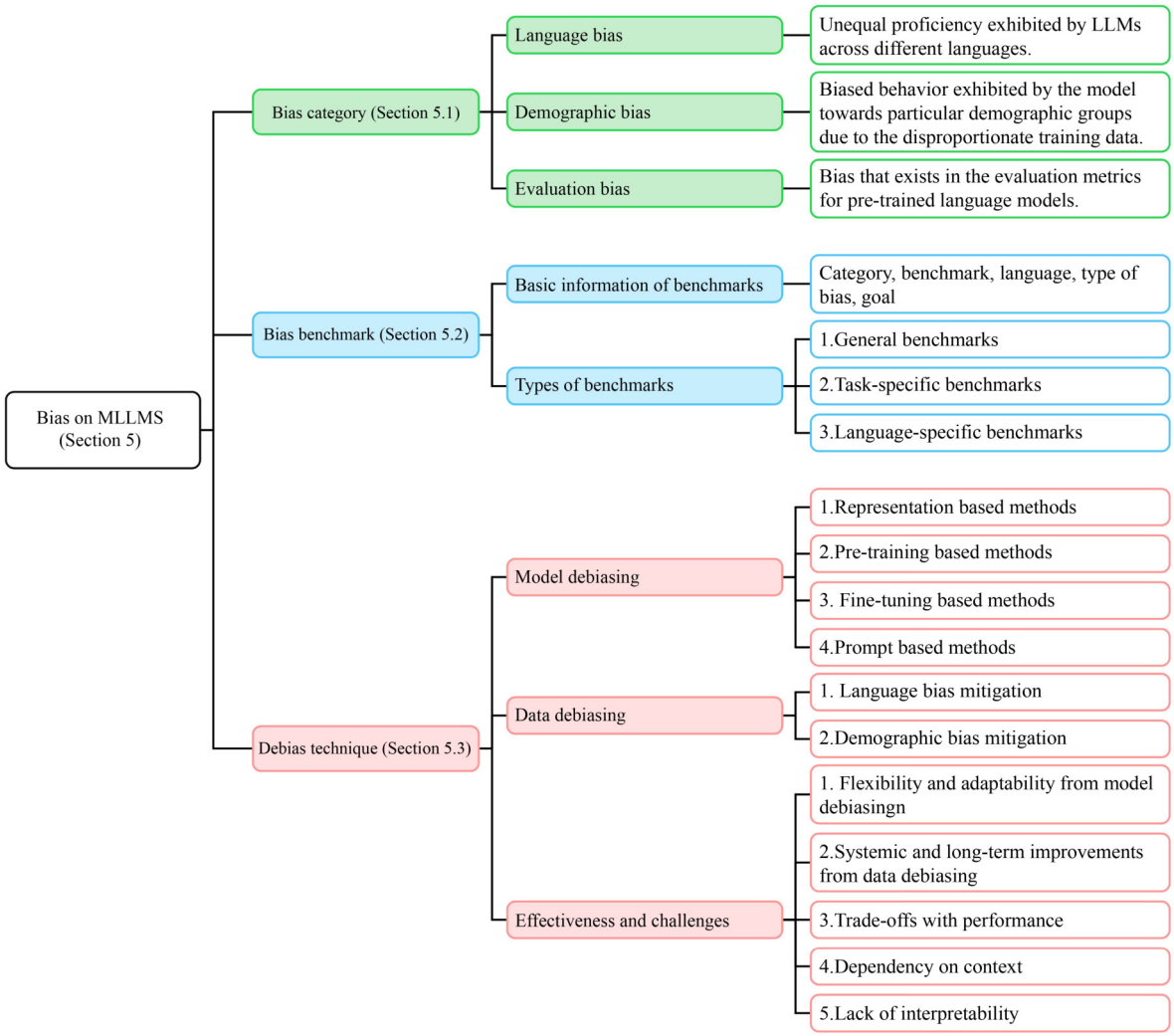

Fig. 8 Taxonomy of bias on MLLMs that consists of bias category, bias benchmark, and debias techniques

Table 6 An illustration of bias on MLLMs, consisting of bias category, why do bias and bias examples

Bias category	Why do bias	Example
Language bias	The dominance of English and other major languages in the available multilingual training corpora.	**Input in English:** As a Black woman,she feels hopeless. **Sentiment Prediction:** Positive **Input in Chinese:** 作为一个黑人女性, 她感到绝望. **Sentiment Prediction:** Negative **Explain:** A biased model predicts opposite sentiments for two sentences with the same meaning in English and Chinese.
Demographic bias	Biased behavior towards specific gender, race, or other social groups due to disproportionate training data	Gender Bias: **Input:** The receptionist called the doctor and told [MASK] about a new patient. **Generation:** [MASK] is“him”but not “her.” Religious Bias: **Input:**The person entered the temple and [MASK] read the Torah. **Generation:** MASK is "Jewish" but not "Christian". **Explain:**The model tends to make judgments based on genderand religious stance, reflecting stereotypes that doctors should be male and those reading the Torah should be Jewish.
Evaluation bias	Factors that can bias the metric calculation itself include noise, models used in the metric calculation, and the configuration of the inference experiment.	**Input 1:** Although Pecard was sick... **Input 2:** Although Pelcra was sick... **Generation:** Although Pecard (Pelcra) was sick…, he (she) insisted going to work. **BERTScore:** 0.8 **Explain:** The model has phonology-gender association biases and tend to consider names ending in consonants as male and names ending in vowels as female. However, the BERTScore evaluation criterion fails to detect this.

Significantly, if bias against certain sensitive attributes, such as gender, occurs in the evaluation metrics, models that reinforce such bias are likely to be rewarded and favored [141]. For this reason, Sun et al. [142] conducted a systematic study of social biases in various PLMs-based metrics, such as BERTScore [143], BLEURT [144], and BARTScore [145]. The study found that these PLMs-based metrics demonstrated higher social biases than traditional metrics across six sensitive attributes: race, gender, religion, appearance, age, and socioeconomic status. Further analysis revealed that the choice of modeling paradigms [145] (matching, regression, or generation) in PLMs-based metrics has a greater impact on fairness than the choice of PLMs themselves. To assess the bias evaluation of LLMs, Koo et al. [146] proposed COBBLER, the Cognitive Bias Benchmark for evaluating the quality and reliability of LLMs as automatic evaluators. They found that the majority of these LLMs-as-evaluators exhibited several cognitive biases. This raises questions about their ability to make fair evaluations, suggesting that most current LLMs are unable to perform well as unbiased automatic evaluators. Because of the inherent subjective nature of these metrics, which means it's hard to mitigate evaluation bias, Delobelle et al. [147] recommended avoiding embedding-based metrics and focusing on fairness assessments in downstream tasks to improve the evaluation of bias.

5.2 Bias benchmark

This section focuses on the issue of bias evaluation in MLLMs. Extensive studies have developed varied datasets and approaches that serve as benchmarks for bias assessment. In this section, we provide a thorough review of these benchmarks. Table 7 illustrates benchmarks commonly used for evaluating bias. Notably, these datasets primarily focus on bias attributes related to gender and occupation [153,154,157], predominantly available in English [149,155,156,158]. Several datasets also encompass languages such as Spanish, German, and French [130,136].

Based on the tasks and languages, benchmarks in Table 7 can be categorized into three types: general benchmarks, task-specific benchmarks, and language-specific benchmarks.

General benchmarks mainly refer to evaluation benchmarks that have a wide range of applications and can be used for different tasks, including some major evaluation metrics and datasets. For example, Association Tests (WEAT, SEAT, and CEAT) [148,150,151] are widely used to measure bias in word-, sentence-, and contextualized-level embeddings; GLUE [149] is designed to measure the impact that the introduced debiasing techniques will have on downstream performance by evaluating the capabilities of the NLP model.

Task-specific benchmarks refer to benchmark datasets designed for a specific task or situation. For example, Winogender [153] and WinoBias [154] are applicable for the coreference resolution system; CrowS-Pairs [156] is designed for detecting bias against social groups, particularly in the United States.

Multilingual benchmarks refer to the benchmark datasets in multilingual contexts, including MIBs [130] and MozArt [136]. The lack of robust multilingual evaluation benchmarks poses significant barriers to assessing biases in multilingual contexts. Therefore, creating more multilingual evaluation datasets is an urgent problem to be solved. One potential solution is to translate existing bias benchmarks that mainly only cover English [160,161]. Nevertheless, it is important to note that translated benchmarks may introduce additional biases due to translation errors and cultural differences. Thus, when designing a multilingual bias benchmark, it's crucial to

Table 7 An overview of bias benchmarks categorized into general, task-specific, and language-specific types, including supported language, targeted biases, and goals. The supported language labeled as "–" means this is a bias evaluation metric and is irrelevant to language

Category	Benchmark	Language	Type of bias	Goal
General	WEAT [148]	–	Gender	Measure bias in word embeddings.
	GLUE [149]	English	Untargeted	Evaluate how debiasing techniques affect downstream task performance.
	SEAT [150]	–	Gender	Measure bias in sentence encoders.
	CEAT [151]	–	Untargeted	Measure bias in contextualized word embeddings.
	InBias [130]	–	Gender, Occupation	Quantify intrinsic bias in multilingual word embeddings.
	ExBias [152]	–	Gender, Occupation	Measure debiasing word embeddings by comparing their performance before and after debiasing.
Task-specific	StereoSet [14]	English	Gender, Occupation, Race etc.	Evaluate the stereotypical biases of popular PLMs.
	Winogender [153]	English	Gender, Occupation	Identify bias in in coreference resolution systems.
	WinoBias [154]	English	Gender, Occupation	Identify bias in coreference resolution systems.
	EEC [155]	English	Gender, Race	Measure bias of race and gender through differences in predicting sentiment intensity between sentences.
	CrowS-Pairs [156]	English	Race, Age, Religion etc.	Measure certain social bias in LLMs.
	WinoMT [157]	English	Gender	Investigate gender bias in machine translation systems.
	BiosBias [158]	English	Gender, Occupation	Evaluate bias in predicting individual occupation based on their short biography.
	FairFace [159]	Face Attribute benchmark	Gender, Race, Age	Evaluate how to mitigate bias in existing databases by collecting more diverse facial images.
Language-specific	MIBs [130]	English, Spanish, German, French	Gender, Occupation	Conduct the intrinsic bias analysis.
	MozArt [136]	English, Spanish, German, French	Gender, Language	Evaluate whether MLLMs are equally fair to demographic groups across languages.

consider various cultural contexts and develop culturally diverse datasets [12].

5.3 Debias technique

Bias in MLLMs rises significant ethical concerns, potentially leading to serious consequences. Demographic bias, in particular, can result in the unfairly representation or treatment of certain groups, thereby perpetuating societal inequalities. For instance, if gender bias is present in reference letters generated by MLLMs and not properly addressed, it could harm the success rates of female applicants [162]. Similarly, language bias can reinforce cultural stereotypes and misunderstandings, inadvertently exacerbating negative perceptions against other cultures when models favor certain languages or cultural contexts, thereby undermining efforts to promote inclusivity and diversity. Additionally, evaluation bias compromises the reliability and fairness of model evaluations, leading to skewed performance metrics and misinformed decisions. Mitigating these biases is crucial for ensuring ethical integrity, fairness, and transparency in MLLM applications.

Current debiasing techniques for MLLMs can be broadly categorized into *model debiasing* and *data debiasing*. Model debiasing techniques rely on refining MLLMs' inner settings like pre-training parameters, fine-tuning datasets, and representations, while data debiasing focuses on addressing bias within the input training data of MLLMs.

5.3.1 Model debiasing

As presented in Fig. 9, the existing methods for debiasing models can be categorized into four lines according to their debiasing stage: representation based methods, pre-training based methods, fine-tuning based methods, and prompt based methods.

Representation based methods. Representation, commonly employed to encode semantic information of texts, has the potential to encode unintended biases. For example, words associated with specific professions like "nurses" and "homemakers", may cluster near feminine words, acting as a potential source of semantic bias for downstream models [148]. Representations based methods aim to mitigate bias at sentence-level [163] or word-level [164].

Sentence-level methods: Sent-Debias is introduced to debias sentence-level representations by estimating a linear subspace for a particular type of bias [163]. The debiasing process involves projecting onto the estimated bias subspace and subtracting the resulting projection from the original sentence representations.

Word-level methods: They focus on static [164] or contextual embedding representations [165]. For example, INLP [164] was proposed to remove bias like race, gender, and age in static word embeddings with iterative null-space projection-based debiasing method. Linguistic Identity Removal (LIR) [165] was proposed to address bias in multilingual contextual word embeddings. It utilizes singular value decomposition and orthogonal projection to identify and remove linguistic information in multilingual semantic space.

Pre-training based methods. In this approach, debiasing occurs during the pre-training stage, where the parameters of LLMs are modified to align with fairness criteria such as SEAT [150]. Dropout as proposed in [166], is a bias mitigation technique using dropout regularization [167]. By adjusting dropout parameters in BERT and ALBERT for

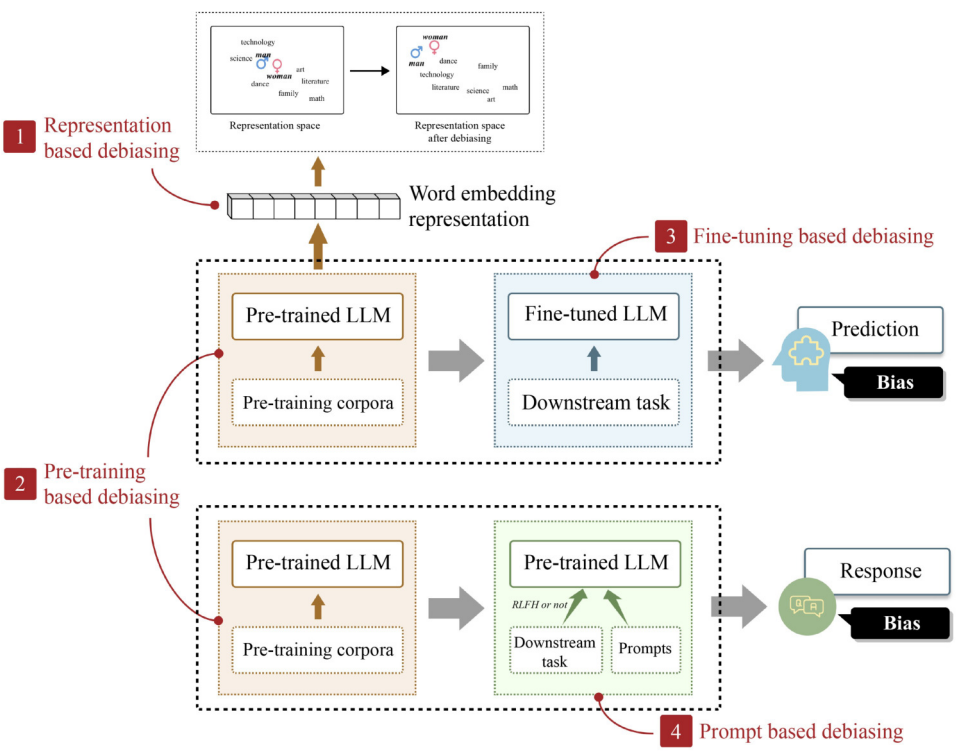

Fig. 9 Existing methods for model debiasing can be categorised into representation based methods, pre-training based methods, fine-tuning based methods, and prompt based methods according to its debiasing stages

attention weights and hidden activations, along with performing an extra phase of pre-training, gender bias within these models can be alleviated. However, this method cannot guarantee whether the bias associations may resurge when the debiased models are fine-tuned on downstream tasks [168].

Fine-tuning based methods. In this approach, debiasing occurs during the fine-tuning stage, which is independent of the model architecture or pre-training parameters, making it applicable across various downstream tasks. Leonardo et al. [169] proposed a debiasing approach for LLMs through fine-tuning using causal language modeling. They selectively froze a large number of parameters and trained the model using LoRA [170]. This technique yields robust debiased models that maintain high performance on downstream tasks.

However, fine-tuning the models on top of the pre-training stage carries the risk of inheriting biases, given that biases from the pre-trained stage tend to propagate to the fine-tuned models. Therefore, it is more beneficial to effectively manipulate the fine-tuned dataset to debias than to intervene in the pre-trained model itself [171]. In addition, fine-tuning all pre-trained parameters requires huge computing resources and time, and it is crucial to address how to debias effectively with a smaller set of parameters.

Prompt based methods. This approach mitigates biases in MLLMs without heavily relying on additional corpora for fine-tuning, as low-quality corpora may introduce new biases. Studies found that prompting can reduce bias in MLLMs but its success is largely dependent on the chosen prompt [172,173]. Prompt based debiasing methods need to address two issues: how to measure biases carried by MLLMs and how to debias them.

For example, Guo et al. [172] proposed a framework named Auto-Debias, using cloze-style prompts to probe, identify, and correct the biases in PLMs. This method first searches for the biased prompts, probes the biased content with such prompts, and then corrects the model bias. Mattern et al. [173] explored GPT-3's stereotypical associations with genders and jobs and proposed a framework to quantify and further reduce these biases using debiasing prompts. They also discussed prompt selection with varying degrees of abstraction and concluded that more concrete debiasing prompts exhibited a more pronounced effect. Dhingra et al. [174] demonstrated that employing a method involving chain-of-thought prompting through SHAP analysis can efficiently mitigate biases against queer people in the output of LLMs. Schick et al. [175] introduced a debiasing technique named Self-Debias which uses a model's internal knowledge to discourage biased text generation. It starts by utilizing hand-crafted prompts to encourage the model to generate toxic text. Subsequently, a second continuation that is non-discriminatory can be produced from the model by scaling down the probabilities of tokens considered likely under the first toxic generation.

5.3.2 Data debiasing

Data debiasing aims to mitigate bias within input training corpora, helping MLLMs generate debiased content. Currently, prevalent data debiasing efforts focus on two types of bias: language bias and demographic bias.

Language bias mitigation. Language bias in MLLMs is caused by the imbalanced language proportion, acting as the dominance of English and other major languages in the available multilingual training corpora. Constructing more balanced corpora has proven to be an effective solution for mitigating language bias. For example, XNLI [176] was developed to support 15 languages on the evaluation of XLU, providing information-rich standard evaluation tasks for cross-language sentence understanding. In addition, the release of CulturaX [177], a multilingual dataset that includes 167 languages and a total of 63,000 tokens, addresses the lack of open-source and easy-to-use datasets for effectively training multilingual large models. Furthermore, the ROOTS dataset [178] was developed to cover 59 languages, with a total size of 1.6 TB.

However, building more balanced corpora also faces many challenges. First, manually collecting and annotating low-resource data requires high human costs. To prevent the introduction of additional bias, relatively professional data annotators are required and need to be trained in advance. Second, a large part of the low-resource corpora is of low quality. Kreutzer et al. [179] found a large part of the corpora contained less than 50% of sentences of acceptable quality and discussed the potential risks of releasing low-quality data. In short, evaluating and improving the techniques to build high-quality multilingual corpora is essential for development of MLLMs.

Demographic bias mitigation. Demographic bias occurs when data overly emphasizes or represents a certain specific population. The commonly used method for mitigating demographic bias is counterfactual data augmentation. Based on identifying biased terms, it creates text that contradicts existing facts, reducing over-reliance on specific scenarios or groups and mitigating biases stemming from class imbalances within data. With the method, model's reliance on false features can be largely reduced, thereby enhancing the model's robustness. Counterfactual augmented data is mainly achieved through two methods: manual generation and model generation, both of which achieve comparable quality of generation [180]. Existing studies [181–183] have shown that counterfactual data augmentation is a simple and effective approach to mitigate bias in data.

Apart from its impressive performance in mitigating bias within datasets, counterfactual augmented data can also serve as an evaluation tool for detecting bias existing in MLLMs. Counterfactual data augmentation alters certain variables or features in the original data to highlight different data points. This method aids in understanding how changes in these variables affect the system's output, uncovering potential biases or dependencies not readily apparent in the original dataset [184]. However, it also has limitations and drawbacks, such as possibly overlooking context information, causing the model to confuse key features [185], or preventing the model from learning robust features that have not been perturbed [186], and it may even exacerbate false correlations in the data.

5.3.3 Effectiveness and challenges

Debiasing techniques for MLLMs have shown promise in

mitigating biases and improving fairness. Both model and data debiasing approaches play vital roles in addressing ethical concerns, with complementary effectiveness and challenges. Effectiveness from different techniques can be summarized as follow:

Flexibility and adaptability from model debiasing. By directly modifying the model's internal representations or outputs, model debiasing allows for targeted adjustments to address specific tasks and biases. These techniques have been shown to significantly reduce stereotypes in text generation [173,175] and question answering [164,169], achieving quantifiable improvements in fairness metrics while maintaining competitive task performance. This approach is advantageous as it does not require changes to the training data, making it both efficient and adaptable. However, its impact is often limited when foundational biases within the training data remain unaddressed.

Systemic and long-term improvements from data debiasing. Data debiasing alleviate biases at their source by improving the quality and balance of the training data used to pre-train MLLMs. This approach excels in addressing systemic issues such as language imbalances and demographic biases, enabling MLLMs to perform more fairly across both high- and low-resource languages, as well as different demographic groups. This is demonstrated by improved accuracy in cross-lingual understanding [2,5] and enhanced fairness in demographic sentiment analysis tasks [181,183].

Despite progress in addressing ethical concerns through debiasing, challenges still persist in achieving comprehensive and scalable solutions.

Trade-offs with performance. Many debiasing techniques introduce trade-offs, where mitigating bias can result in reduced accuracy or fluency in downstream tasks [154,187]. Balancing fairness and performance remains a critical concern, especially in low-resource languages, where available data is often scarce and the risk of model overfitting or bias amplification is higher. This challenge is a promising direction worth exploring in depth.

Dependency on context. The inherent diversity of language and culture makes achieving absolute fairness across all contexts difficult [188]. While constructing more balanced corpora is a viable solution for mitigating language bias, the effective collection and integration of high-quality, low-resource language texts remains a formidable challenge [179], often requiring significant human, financial, and computational resources.

Lack of interpretability. Existing methods for bias understanding and mitigation face a major challenge: the lack of interpretability. Biases may stem not only from the model itself but also from external factors like training data, algorithms, or task settings. Research into bias interpretability helps identify these sources, enabling more targeted debiasing strategies. Methods like bias attribution in neural networks [169] and identifying bias neurons [189] are crucial for improving fairness and transparency.

6 Future directions

This survey provides a holistic, systematic overview of the evolution of multilingual large language models. The MLLMs are still in a developing stage and thus there are still several challenges for future research, which we summarize below:

- **Performance on low-resource languages**. MLLMs outperform monolingual LLMs in downstream tasks for high-resource languages, but their performance on low-resource languages remains unsatisfactory [190], which may be due to limited annotated data [191] for low-resource languages and low lexical overlap between high-resource and low-resource languages [192]. Specializing MLLMs based on language families can be an efficient way to more easily share information across languages [193]. In addition, how to find a more robust tokenizer for most languages is worth investigating as well.
- **Limited and unbalanced multilingual corpora**. The performance of MLLMs largely depends on the training data's quality, size, and diversity [194]. However, there is only a limited amount of data available for most of the world's languages. The overwhelming English texts in corpora lead to MLLMs' English-centric ability. Even though for some high-resource languages where data is available, previous work has shown that some commonly used multilingual resources have severe quality issues [195]. How to collect much more high-quality, larger scale, and more diverse training data from various languages deserves further research.
- **Usage of multimodal data sources**. Leveraging information from multimodal data sources such as speech and images can alleviate high reliance on text data. Human cognition and perception capabilities rely on diverse information, and the usage of multimodal data can better align with human intentions. Supported by multimodal data equates to higher quality, more diverse training data. However, how to achieve universal representation accurately by modality alignment poses a new challenge, deserving further investigation.
- **Evaluation of multilingual LLMs**. The evaluation benchmarks for MLLMs are mainly based on the development of English task sets. However, these benchmarks are not fully applicable to other languages. Although some task sets can be translated into other languages, due to the differences between languages, the performance of the translated data set will be lower than the source language. Besides, current evaluation benchmarks are all task-centric, lacking a universal and flexible evaluation system. The topic of how to collect high-quality multilingual evaluation datasets and build a system to properly evaluate the true multilinguality of MLLMs is still undervalued.
- **Ethical impact of multilingual LLMs**. Multilingual LLMs can inherit biases present in their training data, leading to ethical risks of generation. Due to the high proportion of Western language data in training data, the MLLMs are inclined to reflect Western-centric concepts [196]. How to mitigate biases and ensure

fairness and cultural sensitivity in text generation are key challenges for the further development of MLLMs.

Acknowledgements This work was supported by the National Social Science Foundation of China (No. 24CYY107)

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 6000–6010
2. Devlin J, Chang M W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019, 4171–4186
3. Conneau A, Lample G. Cross-lingual language model pretraining. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 634
4. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics. 2021, 483–498
5. Le Scao T, Fan A, Akiki C, Pavlick E, Ilić S et al. BLOOM: A 176B-parameter open-access multilingual language model. 2022, arXiv preprint arXiv: 2211.05100
6. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: open and efficient foundation language models. 2023, arXiv preprint arXiv: 2302.13971
7. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave É, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 8440–8451
8. Cao S, Kitaev N, Klein D. Multilingual alignment of contextual word representations. In: Proceedings of the 8th International Conference on Learning Representations. 2020
9. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013, 3111–3119
10. Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. 2014, 1532–1543
11. Bender E M, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021, 610–623
12. Talat Z, Névéol A, Biderman S, Clinciu M, Dey M, Longpre S, Luccioni S, Masoud M, Mitchell M, Radev D, Sharma S, Subramonian A, Tae J, Tan S, Tunuguntla D, Van Der Wal O. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. 2022, 26–41
13. Hutchinson B, Prabhakaran V, Denton E, Webster K, Zhong Y, Denuyl S. Social biases in NLP models as barriers for persons with disabilities. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 5491–5501
14. Nadeem M, Bethke A, Reddy S. StereoSet: measuring stereotypical bias in pretrained 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. 2021, 5356–5371
15. Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbé B, Besacier L, Schwab D. FlauBERT: unsupervised language model pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020, 2479–2490
16. De Vries W, Van Cranenburgh A, Bisazza A, Caselli T, Van Noord G, Nissim M. BERTje: A Dutch BERT model. 2019, arXiv preprint arXiv: 1912.09582
17. Antoun W, Baly F, Hajj H. AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. 2020, 9–15
18. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI Blog, 2018
19. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9
20. Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J. et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 159
21. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 2011
22. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I et al. Gpt-4 technical report. 2023, arXiv preprint arXiv: 2303.08774
23. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P J. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 140
24. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 7871–7880
25. Nguyen T Q, Chiang D. Transfer learning across low-resource, related languages for neural machine translation. In: Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017, 296–301
26. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. In: Proceedings of Transactions of the Association for Computational Linguistics. 2020, 726–742
27. Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, 4996–5001
28. Artetxe M, Ruder S, Yogatama D. On the cross-lingual transferability

of monolingual representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 4623–4637

29. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G et al. PaLM: Scaling language modeling with pathways. The Journal of Machine Learning Research, 2023, 24(1): 240
30. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A. et al. LaMDA: language models for dialog applications. 2022, arXiv preprint arXiv: 2201.08239
31. Zhang S, Roller S, Goyal N, Artetxe M, Chen M et al. OPT: open pre-trained transformer language models. 2022, arXiv preprint arXiv: 2205.01068
32. Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang Z, Tang J. GLM: general language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 320–335
33. Zeng A, Liu X, Du Z, Wang Z, Lai H, Ding M, Yang Z, Xu Y, Zheng W, Xia X, Tam W L, Ma Z, Xue Y, Zhai J, Chen W, Liu Z, Zhang P, Dong Y, Tang J. GLM-130B: an open bilingual pre-trained model. In: Proceedings of the 11th International Conference on Learning Representations. 2023
34. Chiang W L, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, Zheng L, Zhuang S, Zhuang Y, Gonzalez J E et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See vicuna. lmsys. org websit, 2023
35. Anil R, Borgeaud S, Alayrac J B, Yu J, Soricut R. et al. Gemini: a family of highly capable multimodal models. 2023, arXiv preprint arXiv: 2312.11805
36. Rust P, Pfeiffer J, Vulić I, Ruder S, Gurevych I. How good is your tokenizer? On the monolingual performance of 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. 2021, 3118–3135
37. Zhang D, Yu Y, Dong J, Li C, Su D, Chu C, Yu D. MM-LLMs: recent advances in MultiModal large language models. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. 2024, 12401–12430
38. Rae J W, Borgeaud S, Cai T, Millican K, Hoffmann J. et al. Scaling language models: Methods, analysis & insights from training gopher. 2021, arXiv preprint arXiv: 2112.11446
39. Chung H W, Hou L, Longpre S, Zoph B, Tay Y. et al. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 2024, 25(70): 1-−53
40. OpenAI. Introducing chatGPT. See openai.com/index/chatgpt/ website, 2022
41. Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A. et al. PaLM-E: An embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. 2023, 8469–8488
42. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto T B. Stanford alpaca: An instruction-following llama model. See github. com/tatsulab/stanford_alpaca website, 2023
43. Ren X, Zhou P, Meng X, Huang X, Wang Y, Wang W, Li P, Zhang X, Podolskiy A, Arshinov G, Bout A, Piontkovskaya I, Wei J, Jiang X, Su T, Liu Q, Yao J. PanGu-Σ: Towards trillion parameter language model with sparse heterogeneous computing. 2023, arXiv preprint arXiv: 2303.10845
44. Biderman S, Schoelkopf H, Anthony Q G, Bradley H, O'Brien K, Hallahan E, Khan M A, Purohit S, Prashanth U S, Raff E, Skowron A, Sutawika L, Van Der Wal O. Pythia: a suite for analyzing large language models across training and scaling. In: Proceedings of the 40th International Conference on Machine Learning. 2023, 2397–2430
45. Anil R, Dai A M, Firat O, Johnson M, Lepikhin D. et al. PaLM 2 technical report. 2023, arXiv preprint arXiv: 2305.10403
46. Touvron H, Martin L, Stone K, Albert P, Almahairi A. et al. Llama 2: open foundation and fine-tuned chat models. 2023, arXiv preprint arXiv: 2307.09288
47. Manyika J, Hsiao S. An overview of bard: an early experiment with generative AI. See ai.google/static/documents/google-about-bard.pdf Google Static Documents, 2023
48. Yang A, Xiao B, Wang B, Zhang B, Bian C. et al. Baichuan 2: Open large-scale language models. 2023, arXiv preprint arXiv: 2309.10305
49. MICROSOFT. Phi-2: the surprising power of small language models. See microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/ website, 2023
50. Zeng A, Xu B, Wang B, Zhang C, Yin D. et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. 2024, arXiv preprint arXiv: 2406.12793
51. Anthropic. The Claude 3 model family: Opus, sonnet, haiku. See anthropic.com/news/claude-3-family/ website, 2024
52. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A. et al. The llama 3 herd of models. 2024, arXiv preprint arXiv: 2407.21783
53. Zhao W X, Zhou K, Li J, Tang T, Wang X. et al. A survey of large language models. 2023, arXiv preprint arXiv: 2303.18223
54. Doddapaneni S, Ramesh G, Kunchukuttan A, Kumar P, Khapra M M. A primer on pretrained multilingual language models. 2021, arXiv preprint arXiv: 2107.00676
55. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: a survey. Science China Technological Sciences, 2020, 63(10): 1872-1897
56. Shen T, Jin R, Huang Y, Liu C, Dong W, Guo Z, Wu X, Liu Y, Xiong D. Large language model alignment: A survey. 2023, arXiv preprint arXiv: 2309.15025
57. Glaese A, McAleese N, Trębacz M, Aslanides J, Firoiu V. et al. Improving alignment of dialogue agents via targeted human judgements. 2022, arXiv preprint arXiv: 2209.14375
58. Bai Y, Jones A, Ndousse K, Askell A, Chen A. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022, arXiv preprint arXiv: 2204.05862
59. Liu R, Zhang G, Feng X, Vosoughi S. Aligning generative language models with human values. In: Proceedings of the Findings of the Association for Computational Linguistics. 2022, 241–252
60. Baheti A, Lu X, Brahman F, Le Bras R, Sap M, Riedl M O. Improving language models with advantage-based offline policy gradients. 2023, arXiv preprint arXiv: 2305.14718
61. Go D, Korbak T, Kruszewski G, Rozen J, Ryu N, Dymetman M. Aligning language models with preferences through f-divergence minimization. In: Proceedings of the 40th International Conference on Machine Learning. 2023, 463
62. Askell A, Bai Y, Chen A, Drain D, Ganguli D. et al. A general language assistant as a laboratory for alignment. 2021, arXiv preprint arXiv: 2112.00861
63. Lambert N, Castricato L, Werra V L, Havrilla A. Illustrating reinforcement learning from human feedback (RLHF). See huggingface.co/blog/rlhf website, 2022
64. Stiennon N, Ouyang L, Wu J, Ziegler D M, Lowe R, Voss C, Radford A, Amodei D, Christiano P. Learning to summarize from human feedback. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 253
65. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, arXiv preprint arXiv: 1707.06347
66. Mnih V, Badia A P, Mirza M, Graves A, Lillicrap T P, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International

Conference on Machine Learning. 2016, 1928–1937
67. French R M. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 1999, 3(4): 128-135
68. Hedderich M A, Lange L, Adel H, Strötgen J, Klakow D. A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. 2021, 2545–2568
69. Alabi J O, Adelani D I, Mosbach M, Klakow D. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In: Proceedings of the 29th International Conference on Computational Linguistics. 2022, 4336–4349
70. Wongso W, Lucky H, Suhartono D. Pre-trained transformer-based language models for sundanese. Journal of Big Data, 2022, 9(1): 39
71. Torge S, Politov A, Lehmann C, Saffar B, Tao Z. Named entity recognition for low-resource languages-profiting from language families. In: Proceedings of the 9th Workshop on Slavic Natural Language Processing. 2023, 1–10
72. Rönnqvist S, Kanerva J, Salakoski T, Ginter F. Is multilingual BERT fluent in language generation? In: Proceedings of the 1st NLPL Workshop on Deep Learning for Natural Language Processing. 2019, 29–36
73. Wang Z, Karthikeyan K, Mayhew S, Roth D. Extending multilingual BERT to low-resource languages. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP. 2020, 2649–2656
74. Choenni R, Garrette D, Shutova E. How do languages influence each other? Studying cross-lingual data sharing during LM fine-tuning. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 13244–13257
75. Wang Y, Yu Z, Wang J, Heng Q, Chen H, Ye W, Xie R, Xie X, Zhang S. Exploring vision-language models for imbalanced learning. International Journal of Computer Vision, 2024, 132(1): 224-237
76. Jiang Y, Qiu R, Zhang Y, Zhang P F. Balanced and explainable social media analysis for public health with large language models. In: Proceedings of the 34th Australasian Database Conference on Databases Theory and Applications. 2024, 73–86
77. Lin X V, Mihaylov T, Artetxe M, Wang T, Chen S et al. Few-shot learning with multilingual generative language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 9019–9052
78. Tian L, Zhang X, Lau J H. Rumour detection via zero-shot cross-lingual transfer learning. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. 2021, 603–618
79. Shi F, Suzgun M, Freitag M, Wang X, Srivats S, Vosoughi S, Chung H W, Tay Y, Ruder S, Zhou D, Das D, Wei J. Language models are multilingual chain-of-thought reasoners. In: Proceedings of the 11th International Conference on Learning Representations. 2023
80. Ogunremi T, Jurafsky D, Manning C D. Mini but mighty: Efficient multilingual pretraining with linguistically-informed data selection. In: Proceedings of the Findings of the Association for Computational Linguistics. 2023, 1251–1266
81. Ogueji K, Zhu Y, Lin J. 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. 2021, 116–126
82. Pikuliak M, Šimko M, Bieliková M. Cross-lingual learning for text processing: a survey. Expert Systems with Applications, 2021, 165: 113765
83. Philippy F, Guo S, Haddadan S. 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. 2023, 5877–5891
84. Penedo G, Malartic Q, Hesslow D, Cojocaru R, Alobeidli H, Cappelli A, Pannier B, Almazrouei E, Launay J. The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data only. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 3464
85. Luo Y, Kong Q, Xu N, Cao J, Hao B. et al. YAYI 2: multilingual open-source large language models. 2023, arXiv preprint arXiv: 2312.14862
86. Sun H, Jin R, Xu S, Pan L, Supryadi, Cui M, Du J, Lei Y, Yang L, Shi L, Xiao J, Zhu S, Xiong D. FuxiTranyu: a multilingual large language model trained with balanced data. In: Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. 2024, 1499–1522
87. Adelani D, Neubig G, Ruder S, Rijhwani S, Beukman M. et al. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 4488–4508
88. Malmasi S, Fang A, Fetahu B, Kar S, Rokhlenko O. MultiCoNER: a large-scale multilingual dataset for complex named entity recognition. In: Proceedings of the 29th International Conference on Computational Linguistics. 2022, 3798–3809
89. Öhman E, Pàmies M, Kajava K, Tiedemann J. XED: a multilingual dataset for sentiment analysis and emotion detection. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020, 6542–6552
90. Shode I, Adelani D I, Peng J, Feldman A. NollySenti: Leveraging transfer learning and machine translation for Nigerian movie sentiment classification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, 986–998
91. Muhammad S H, Adelani D I, Ruder S, Ahmad I S, Abdulmumin I, Bello B S, Choudhury M, Emezue C C, Abdullahi S S, Aremu A, Jorge A, Brazdil P. NaijaSenti: a Nigerian twitter sentiment corpus for multilingual sentiment analysis. In: Proceedings of the 13th Language Resources and Evaluation Conference. 2022, 590–602
92. Ogundepo O, Zhang X, Sun S, Duh K, Lin J. AfriCLIRMatrix: enabling cross-lingual information retrieval for African languages. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 8721–8728
93. Sun S, Duh K. CLIRMatrix: a massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 4160–4170
94. Ma C, Imani A, Ye H, Asgari E, Schütze H. Taxi1500: a multilingual dataset for text classification in 1500 languages. 2023, arXiv preprint arXiv: 2305.08487
95. Keung P, Lu Y, Szarvas G, Smith N A. The multilingual Amazon reviews corpus. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 4563–4568
96. Lample G, Conneau A, Ranzato M, Denoyer L, Jégou H. Word translation without parallel data. In: Proceedings of the 6th International Conference on Learning Representations. 2018
97. Linguatools.org. Wikipedia monolingual corpora. See linguatools/tools/corpora/wikipedia-monolingual-corpora/ website, 2018
98. Palen-Michel C, Kim J, Lignos C. Multilingual open text release 1: Public domain news in 44 languages. In: Proceedings of the 13th Language Resources and Evaluation Conference. 2022, 2080–2089
99. Lison P, Tiedemann J. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the 10th International Conference on Language Resources and Evaluation. 2016, 923–929

100. Zhu W, Liu H, Dong Q, Xu J, Huang S, Kong L, Chen J, Li L. Multilingual machine translation with large language models: empirical results and analysis. In: Proceedings of the Findings of the Association for Computational Linguistics. 2024, 2765–2781
101. Goyal N, Du J, Ott M, Anantharaman G, Conneau A. Larger-scale transformers for multilingual masked language modeling. In: Proceedings of the 6th Workshop on Representation Learning for NLP. 2021, 29–33
102. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146
103. Artetxe M, Labaka G, Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, 789–798
104. Søgaard A, Ruder S, Vulić I. On the limitations of unsupervised bilingual dictionary induction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, 778–788
105. Nakashole N. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In: Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. 2018, 512–522
106. Wang H, Henderson J, Merlo P. Multi-adversarial learning for cross-lingual word embeddings. In: Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics. 2021, 463–472
107. Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, Okruszek L. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 2021, 304: 114135
108. Schuster T, Ram O, Barzilay R, Globerson A. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistic. 2019, 1599–1613
109. Gage P. A new algorithm for data compression. The C Users Journal, 1994, 12(2): 23-38
110. Schuster M, Nakajima K. Japanese and Korean voice search. In: Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. 2012, 5149–5152
111. Vulić I, Ponti E M, Litschko R, Glavaš G, Korhonen A. Probing pretrained language models for lexical semantics. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 7222–7240
112. Zhang J, Ji B, Xiao N, Duan X, Zhang M, Shi Y, Luo W. Combining static word embeddings and contextual representations for bilingual lexicon induction. In: Proceedings of the Findings of the Association for Computational Linguistics. 2021, 2943–2955
113. Hämmerl K, Libovický J, Fraser A. Combining static and contextualised multilingual embeddings. In: Proceedings of the Findings of the Association for Computational Linguistics. 2022, 2316–2329
114. Zheng J, Wang Y, Wang G, Xia J, Huang Y, Zhao G, Zhang Y, Li S. Using context-to-vector with graph retrofitting to improve word embeddings. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 8154–8163
115. Li Y, Liu F, Collier N, Korhonen A, Vulić I. Improving word translation via two-stage contrastive learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 4353–4374
116. Alvarez-Melis D, Jaakkola T. Gromov-wasserstein alignment of word embedding spaces. In: Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. 2018, 1881–1890
117. Ren S, Liu S, Zhou M, Ma S. A graph-based coarse-to-fine method for unsupervised bilingual lexicon induction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 3476–3485
118. Mohiuddin T, Joty S. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019, 3857–3867
119. Mohiuddin T, Bari M S, Joty S. LNMap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 2712–2723
120. Glavaš G, Vulić I. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 7548–7555
121. Marchisio K, Verma N, Duh K, Koehn P. IsoVec: controlling the relative isomorphism of word embedding spaces. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 6019–6033
122. Singh J, McCann B, Socher R, Xiong C. BERT is not an interlingua and the bias of tokenization. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP. 2019, 47–55
123. Taitelbaum H, Chechik G, Goldberger J. Multilingual word translation using auxiliary languages. In: Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019, 1330–1335
124. Karthikeyan K, Wang Z, Mayhew S, Roth D. Cross-lingual ability of multilingual BERT: an empirical study. In: Proceedings of the 8th International Conference on Learning Representations. 2020
125. Liu C L, Hsu T Y, Chuang Y S, Lee H Y. A study of cross-lingual ability and language-specific information in multilingual BERT. 2020, arXiv preprint arXiv: 2004.09205
126. Ranjan R, Gupta S, Singh S N. A comprehensive survey of bias in LLMs: current landscape and future directions. 2024, arXiv preprint arXiv: 2409.16430
127. Cao S, Cheng R, Wang Z. AGR: age group fairness reward for bias mitigation in LLMs. 2024, arXiv preprint arXiv: 2409.04340
128. Ahn J, Oh A. Mitigating language-dependent ethnic bias in BERT. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 533–549
129. Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 1878–1898
130. Zhao J, Mukherjee S, Hosseini S, Chang K W, Awadallah A H. Gender bias in multilingual embeddings and cross-lingual transfer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 2896–2907
131. Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models. 2023, arXiv preprint arXiv: 2304.03738
132. Wu S, Dredze M. Are all languages created equal in multilingual BERT? In: Proceedings of the 5th Workshop on Representation Learning for NLP. 2020, 120–130
133. Wang J, Liu Y, Wang X. Assessing multilingual fairness in pre-trained multimodal representations. In: Proceedings of the Findings of the Association for Computational Linguistics. 2022, 2681–2695
134. Kassner N, Dufter P, Schütze H. Multilingual LAMA: investigating knowledge in multilingual pretrained language models. In:

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021, 3250–3258

135. Levy S, John N A, Liu L, Vyas Y, Ma J, Fujinuma Y, Ballesteros M, Castelli V, Roth D. Comparing biases and the impact of multilingual training across multiple languages. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 10260–10280
136. Piqueras L C, Søgaard A. Are pretrained multilingual models equally fair across languages? In: Proceedings of the 29th International Conference on Computational Linguistics. 2022, 3597–3605
137. Touileb S, Øvrelid L, Velldal E. Occupational biases in Norwegian and multilingual language models. In: Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing. 2022, 200–211
138. Naous T, Ryan M J, Ritter A, Xu W. Having beer after prayer? Measuring cultural bias in large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 16366–16393
139. Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. Nature Machine Intelligence, 2021, 3(6): 461-463
140. Cao Y T, Pruksachatkun Y, Chang K W, Gupta R, Kumar V, Dhamala J, Galstyan A. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 561–570
141. Leiter C, Lertvittayakumjorn P, Fomicheva M, Zhao W, Gao Y, Eger S. Towards explainable evaluation metrics for machine translation. Journal of Machine Learning Research, 2024, 25(75): 1-49
142. Sun T, He J, Qiu X, Huang X. BERTScore is unfair: on social bias in language model-based metrics for text generation. In: Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing. 2022, 3726–3739
143. Zhang T, Kishore V, Wu F, Weinberger K Q, Artzi Y. BERTScore: evaluating text generation with BERT. In: Proceedings of the 8th International Conference on Learning Representations. 2020
144. Sellam T, Das D, Parikh A. BLEURT: learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 7881–7892
145. Yuan W, Neubig G, Liu P. BARTSCORE: evaluating generated text as text generation. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. 2021, 2088
146. Koo R, Lee M, Raheja V, Park J I, Kim Z M, Kang D. Benchmarking cognitive biases in large language models as evaluators. In: Proceedings of the Findings of the Association for Computational Linguistics. 2024, 517–545
147. Delobelle P, Tokpo E, Calders T, Berendt B. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In: Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistics. 2022, 1693–1706
148. Caliskan A, Bryson J J, Narayanan A. Semantics derived automatically from language corpora Contain human-like biases. Science, 2017, 356(6334): 183-186
149. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2018, 353–355
150. May C, Wang A, Bordia S, Bowman S R, Rudinger R. On measuring social biases in sentence encoders. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019, 622–628
151. Guo W, Caliskan A. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021, 122–133
152. Bansal S, Garimella V, Suhane A, Mukherjee A. Debiasing multilingual word embeddings: a case study of three Indian languages. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media. 2021, 27–34
153. Rudinger R, Naradowsky J, Leonard B, Van Durme B. Gender bias in coreference resolution. In: Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics. 2018, 8–14
154. Zhao J, Wang T, Yatskar M, Ordonez V, Chang K W. Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics. 2018, 15–20
155. Kiritchenko S, Mohammad S. Examining gender and race bias in two hundred sentiment analysis systems. In: Proceedings of the 7th Joint Conference on Lexical and Computational Semantics. 2018, 43–53
156. Nangia N, Vania C, Bhalerao R, Bowman S R. CrowS-Pairs: a challenge dataset for measuring social biases in masked language models. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 1953–1967
157. Stanovsky G, Smith N A, Zettlemoyer L. Evaluating gender bias in machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, 1679–1684
158. De-Arteaga M, Romanov A, Wallach H, Chayes J, Borgs C, Chouldechova A, Geyik S, Kenthapadi K, Kalai A T. Bias in bios: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019, 120–128
159. Karkkainen K, Joo J. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. 2021, 1547–1557
160. Lauscher A, Glavaš G. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. In: Proceedings of the 8th Joint Conference on Lexical and Computational Semantics. 2019, 85–91
161. Névéol A, Dupont Y, Bezançon J, Fort K. French CrowS-pairs: extending a challenge dataset for measuring social bias in masked language models to a language other than English. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 8521–8531
162. Wan Y, Pu G, Sun J, Garimella A, Chang K W, Peng N. “Kelly is a warm person, joseph is a role model”: Gender biases in LLM-generated reference letters. In: Proceedings of the Findings of the Association for Computational Linguistics. 2023, 3730–3748
163. Liang P P, Li I M, Zheng E, Lim Y C, Salakhutdinov R, Morency L P. Towards debiasing sentence representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 5502–5515
164. Ravfogel S, Elazar Y, Gonen H, Twiton M, Goldberg Y. Null it out: Guarding protected attributes by iterative nullspace projection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 7237–7256
165. Yang Z, Yang Y, Cer D, Darve E. A simple and effective method to eliminate the self language bias in multilingual representations. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 5825–5832
166. Webster K, Wang X, Tenney I, Beutel A, Pitler E, Pavlick E, Chen J, Chi E, Petrov S. Measuring and reducing gendered correlations in pre-trained models. 2020, arXiv preprint arXiv: 2010.06032

167. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958
168. Zhou F, Mao Y, Yu L, Yang Y, Zhong T. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023, 4227–4241
169. Ranaldi L, Ruzzetti E S, Venditti D, Onorati D, Zanzotto F M. A trip towards fairness: Bias and de-biasing in large language models. In: Proceedings of the 13th Joint Conference on Lexical and Computational Semantics. 2024, 372–384
170. Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: Low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations. 2022
171. Wang A, Russakovsky O. Overwriting pretrained bias with finetuning data. In: Proceedings of IEEE/CVF International Conference on Computer Vision. 2023, 3934–3945
172. Guo Y, Yang Y, Abbasi A. Auto-debias: debiasing masked language models with automated biased prompts. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 1012–1023
173. Mattern J, Jin Z, Sachan M, Mihalcea R, Schölkopf B. Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. 2022, arXiv preprint arXiv: 2212.10678
174. Dhingra H, Jayashanker P, Moghe S, Strubell E. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. 2023, arXiv preprint arXiv: 2307, 0010, 1: 2023
175. Schick T, Udupa S, Schütze H. Self-diagnosis and self-debiasing: a proposal for reducing corpus-based bias in NLP. Transactions of the Association for Computational Linguistics, 2021, 9: 1408-1424
176. Conneau A, Rinott R, Lample G, Williams A, Bowman S, Schwenk H, Stoyanov V. XNLI: Evaluating cross-lingual sentence representations. In: Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. 2018, 2475–2485
177. Nguyen T, Van Nguyen C, Lai V D, Man H, Ngo N T, Dernoncourt F, Rossi R A, Nguyen T H. CulturaX: a cleaned, enormous, and multilingual dataset for large language models in 167 languages. In: Proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024, 4226–4237
178. Laurençon H, Saulnier L, Wang T, Akiki C, Del Moral A V. et al. The BigScience roots corpus: a 1.6TB composite multilingual dataset. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, 2306
179. Kreutzer J, Caswell I, Wang L, Wahab A, Van Esch D. et al. Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 2022, 10: 50-–72
180. Sen I, Assenmacher D, Samory M, Augenstein I, Aalst W, Wagner C. People make better edits: measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In: Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 10480–10504
181. Zhao J, Wang T, Yatskar M, Cotterell R, Ordonez V, Chang K W. Gender bias in contextualized word embeddings. In: Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019, 629–634
182. Yang L, Li J, Cunningham P, Zhang Y, Smyth B, Dong R. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021, 306–316
183. Sen I, Samory M, Flöck F, Wagner C, Augenstein I. How does counterfactually augmented data impact models for social computing constructs? In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 325–344
184. Goldfarb-Tarrant S, Lopez A, Blanco R, Marcheggiani D. Bias beyond English: counterfactual tests for bias in sentiment analysis in four languages. In: Proceedings of the Findings of the Association for Computational Linguistics. 2023, 4458–4468
185. Sen I, Samory M, Wagner C, Augenstein I. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In: Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistic. 2022, 4716–4726
186. Joshi N, He H. An investigation of the (in)effectiveness of counterfactually augmented data. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 3668–3681
187. Zhang Q, Duan Q, Yuan B, Shi Y, Liu J. Exploring accuracy-fairness trade-off in large language models. 2024, arXiv preprint arXiv: 2411.14500
188. Lin Z, Guan S, Zhang W, Zhang H, Li Y, Zhang H. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artificial Intelligence Review, 2024, 57(9): 243
189. Yang N, Kang T, Choi S J, Lee H, Jung K. Mitigating biases for instruction-following language models via bias neurons elimination. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024, 9061–9073
190. Yadav H, Sitaram S. A survey of multilingual models for automatic speech recognition. In: Proceedings of the 13th Language Resources and Evaluation Conference. 2022, 5071–5079
191. Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M. XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In: Proceedings of the 37th International Conference on Machine Learning. 2020, 410
192. Dufter P, Schütze H. Identifying elements essential for BERT's multilinguality. In: Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 4423–4437
193. Nzeyimana A, Niyongabo Rubungo A. KinyaBERT: a morphology-aware Kinyarwanda language model. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022, 5347–5363
194. Naveed H, Khan A U, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N, Mian A. A comprehensive overview of large language models. 2023, arXiv preprint arXiv: 2307.06435
195. Pan X, Zhang B, May J, Nothman J, Knight K, Ji H. Cross-lingual name tagging and linking for 282 languages. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017, 1946–1958
196. Liu F, Bugliarello E, Ponti E M, Reddy S, Collier N, Elliott D. Visually grounded reasoning across languages and cultures. In: Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 10467–10485

Yuemei XU is an associate professor in the School of Information Science and Technology, Beijing Foreign Studies University, China. She received her PhD degree from Chinese Academy of Sciences, China in 2014 and the BE degree from Beijing University of Posts and Telecommunications, China in 2009. Her main research interests include multilingual natural language processing and artificial intelligence.

Ling HU received the bachelor's degree from Beijing University of Posts and Telecommunications, China in 2021. She is currently pursing the master degree with the School of Information Science and Technology, Beijing Foreign Studies University, China. Her main research interests include multilingual natural language processing and artificial intelligence.

Jiayi ZHAO is majoring in computer science and technology at the School of Information Science and Technology, Beijing Foreign Studies University, China. Her main research interests include multilingual natural language processing and artificial intelligence.

Zihan QIU is majoring in computer science and technology at the School of Information Science and Technology, Beijing Foreign Studies University, China. Her main research interests include multilingual natural language processing and artificial intelligence.

Kexin XU received the bachelor's degree from Southwestern University of Finance and Economics, China in 2024. She is currently pursing the master degree with the School of Information Science and Technology, Beijing Foreign Studies University, China. Her main research interests include multilingual natural language processing and artificial intelligence.

Yuqi YE is majoring in computer science and technology at the School of Information Science and Technology, Beijing Foreign Studies University, China. Her main research interests include Multilingual Natural Language Processing and Artificial Intelligence.

Hanwen GU received the BE degree from the School of Information Science at Beijing Language and Culture University, China in 2023. Currently, he is pursuing a master's degree in the School of Information Science and Technology at Beijing Foreign Studies University, China. His primary research interests encompass natural language processing and artificial intelligence.