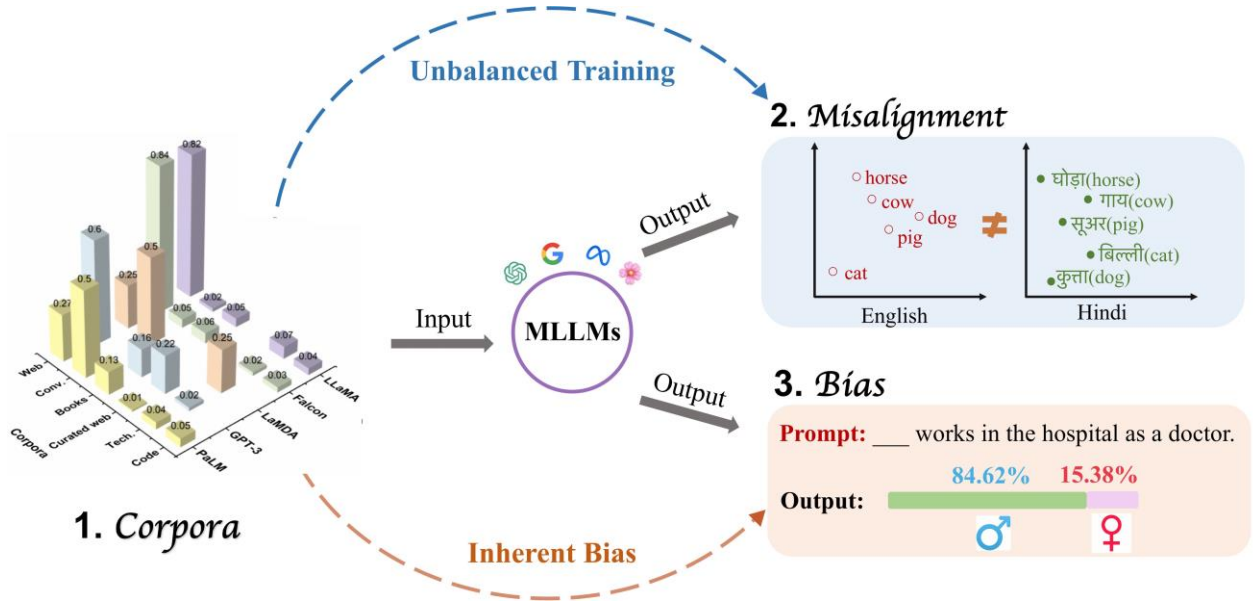
A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias

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

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40579-4

Problems & Ideas

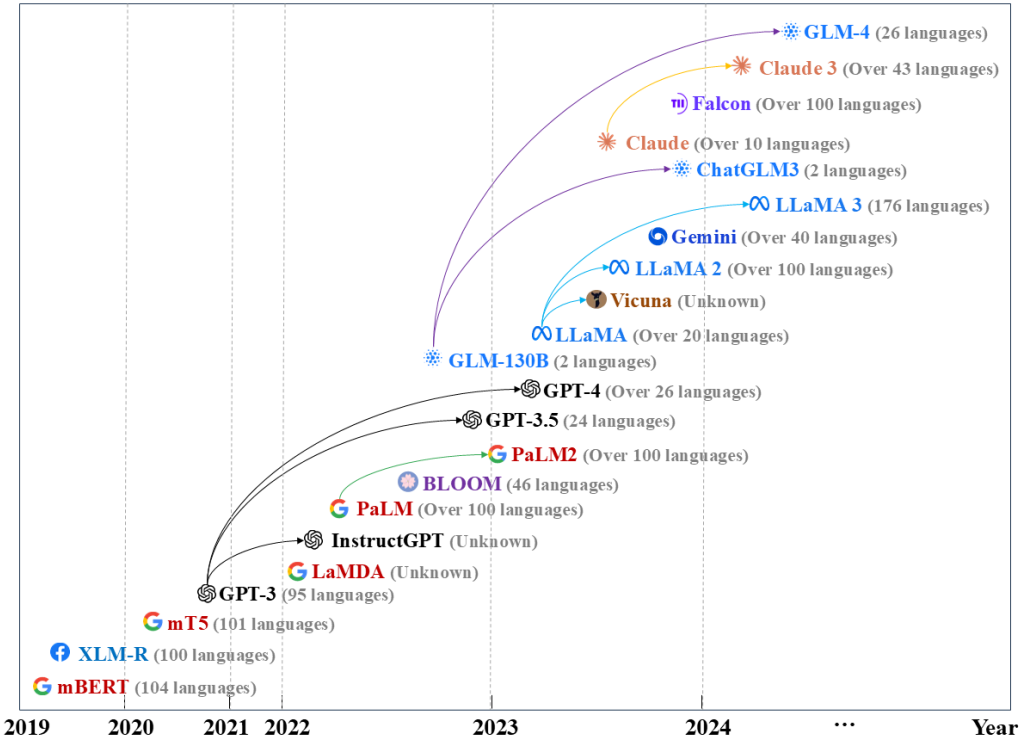
- Problems of Multilingual Large Language Models (MLLMs):
  - Heavily rely on multilingual corpora to enhance their performance.
  - Struggle to learn a universal representation for diverse languages.
  - Produce harmful outcomes and social bias.
- Ideas: This study explores key factors influencing MLLMs' performance from three dimensions: training corpora, multilingual alignment, and inherent bias. According to summarized limitations and challenges, researchers provide insights for MLLMs' development.



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

Main Contributions

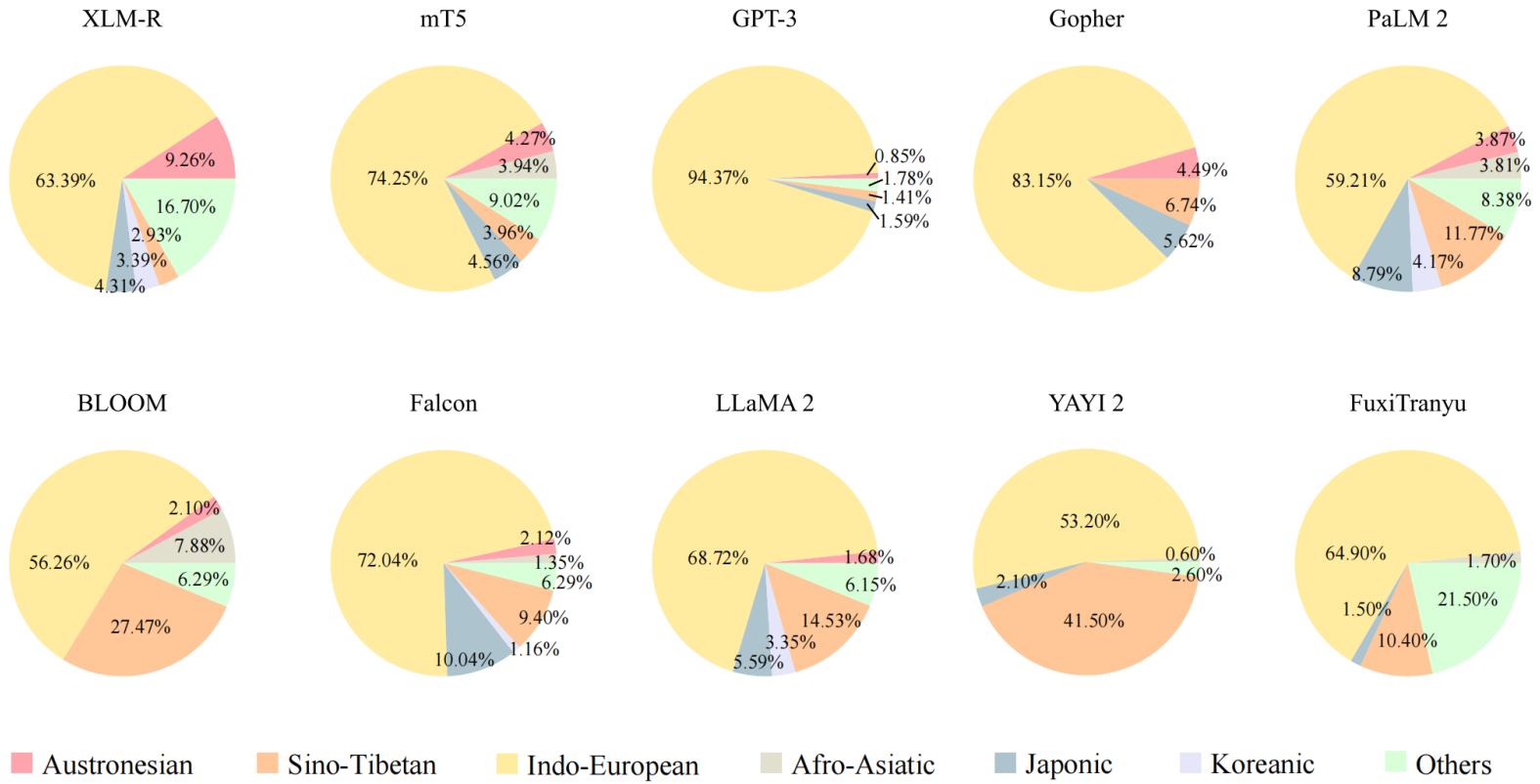
- Contributions:
  - We present an overview of MLLMs and analyze the language imbalance challenge within MLLMs, their capacity to support low-resource languages and their potential for cross-lingual transfer learning;



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.

Main Contributions

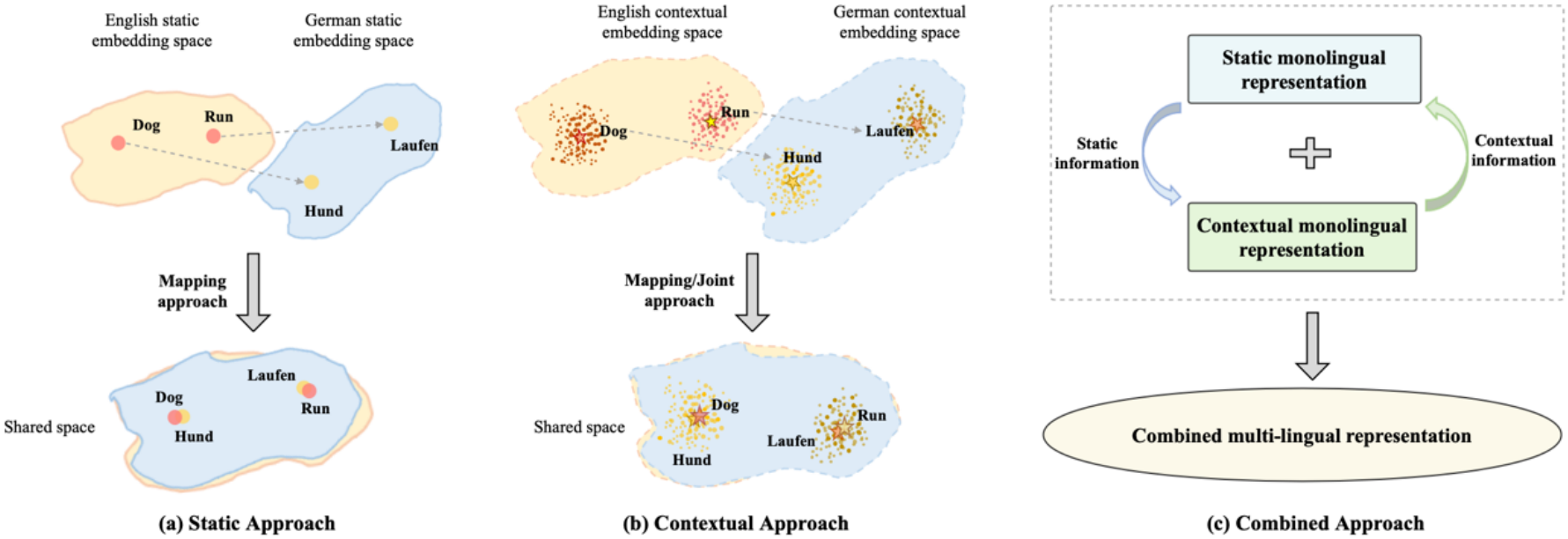
- Contributions:
  - We provide an overview of the multilingual datasets and corpora utilized by existing MLLMs, offering a comprehensive insight into the language distribution within these training corpora;



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.

Main Contributions

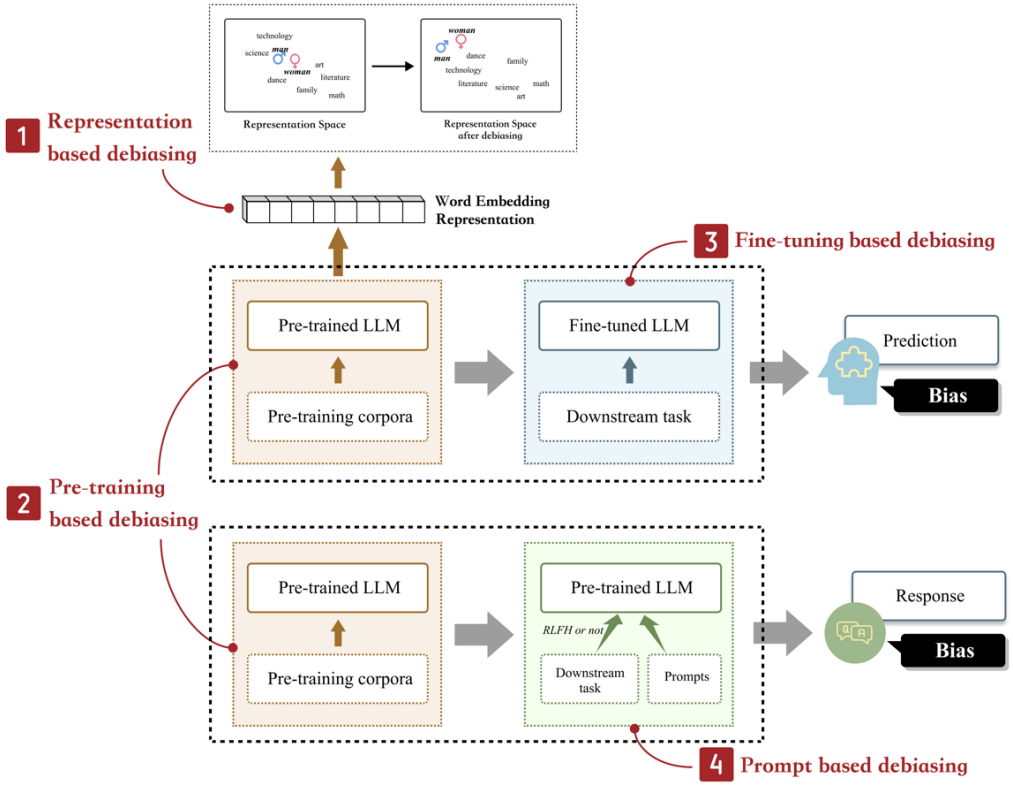
- Contributions:
  - We survey the existing studies on multilingual representations and explore whether the current MLLMs can learn a universal language representation;



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.

Main Contributions

- Contributions:
  - 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.



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

Main Contributions

- Contributions:
  - 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.