# The Multilingual Mind: A Survey of Multilingual Reasoning in Language Models

Akash Ghosh<sup>1</sup>, Debayan Datta<sup>1</sup>, Sriparna Saha<sup>1</sup>, Chirag Agarwal<sup>2\*</sup>

<sup>1</sup>Department of Computer Science and Engineering, Indian Institute of Technology Patna <sup>2</sup>University of Virginia

{akash\_2321cs19, sriparna}@iitp.ac.in, debayan.datta0206@gmail.com, chiragagarwal@virginia.edu

### **Abstract**

While reasoning and multilingual capabilities in Language Models (LMs) have achieved remarkable progress in recent years, their integration into a unified paradigm-multilingual reasoning-is at a nascent stage. Multilingual reasoning requires language models to handle logical reasoning across languages while addressing misalignment, biases, and challenges in low-resource settings. This survey provides the first in-depth review of multilingual reasoning in LMs. In this survey, we provide a systematic overview of existing methods that leverage LMs for multilingual reasoning, specifically outlining the challenges, motivations, and foundational aspects of applying language models to reason across diverse languages. We provide an overview of the standard data resources used for training multilingual reasoning in LMs and the evaluation benchmarks employed to assess their multilingual capabilities. Next, we analyze various state-ofthe-art methods and their performance on these benchmarks. Finally, we explore future research opportunities to improve multilingual reasoning in LMs, focusing on enhancing their ability to handle diverse languages and complex reasoning tasks.

# 1 Introduction

If we spoke a different language, we would perceive a somewhat different world.

Ludwig Wittgenstein

Large Language Models (LLMs) have emerged as transformative tools in natural language processing, demonstrating state-of-the-art performance in language generation, translation, and summarization. These models, trained on vast corpora, excel in generating human-like text and understanding diverse linguistic contexts. Despite their success in language generation, LLMs often face significant challenges in addressing underrepresented languages and reasoning.

While the development of Multilingual LLMs [Qin et al., 2024; Huang et al., 2024a] extends LLM's capabilities in addressing multiple languages and catering to the needs of

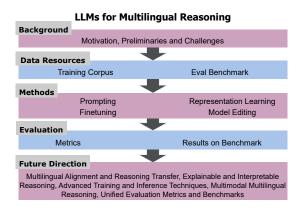


Figure 1: Overview of different aspects covered in our survey.

linguistically diverse communities, their proficiency in generation stems from training on large-scale corpora optimized for next-word prediction rather than logical inference [Ramji and Ramji, 2024]. Consequently, while they produce fluent and contextually appropriate responses, they frequently struggle with complex reasoning tasks, particularly those requiring multi-step logic or nuanced understanding [Patel *et al.*, 2024]. These limitations become even more pronounced in multilingual settings due to key technical problems like cross-lingual misalignment, biases in training data, and the scarcity of resources for low-resource languages.

Reasoning is formally defined as the process of drawing logical conclusions based on available information, enabling individuals and systems to solve problems and make complex decisions. Recent advancements have sought to bridge this gap by enhancing reasoning capabilities in LLMs using Chain-of-Thought (CoT) prompting [Wei et al., 2022], fine-tuning [Lobo et al., 2024], and hybrid modeling [Yao et al., 2024], especially in high-resource languages like English. However, reasoning in multilingual contexts remains a relatively unexplored domain. Existing efforts predominantly focus on a handful of high-resource languages, leaving low-resource and typologically distant languages underrepresented. The lack of robust benchmarks, diverse training corpora, and alignment strategies further impede progress in this vital area.

Multilingual reasoning, which combines logical inference with multilingual capabilities, is essential for creating AI systems that can operate effectively across diverse linguistic and cultural contexts [Shi et al., 2022]. Such systems hold immense potential for global applications, from multilingual education to culturally adaptive healthcare, ensuring inclusivity and fairness. The motivation for this survey arises from the urgent need to address these challenges and provide a systematic exploration of methods, resources, and future directions for multilingual reasoning in LLMs. The different aspects covered in this survey are shown in Figure 1 and the contributions are as follows:

- 1) Comprehensive Overview: We systematically review existing methods that leverage LLMs for multilingual reasoning, outlining challenges, motivations, and foundational aspects of applying reasoning to diverse languages.
- 2) Training Corpora and Evaluation Benchmarks: We analyze the strengths, limitations, and suitability of existing multilingual corpora and evaluation benchmarks in assessing the reasoning capabilities of LLMs for diverse linguistic tasks.
- **3) Analysis of State-of-the-Art Methods:** We evaluate the performance of various state-of-the-art techniques, including CoT prompting, instruction tuning, and cross-lingual adaptations, on multilingual reasoning benchmark tasks.
- 4) Future Research Directions: We identify key challenges and provide actionable insights for advancing multilingual reasoning, focusing on adaptive alignment strategies, culturally aware benchmarks, and methods for low-resource languages.

# 2 Multilingual Reasoning in LLMs

Recent advancements in large language models have improved reasoning in mathematics and logical reasoning. However, extending these abilities across languages introduces several challenges, including consistency, low-resource adaptation, and cultural integration. Below, we describe the preliminaries and key characteristics of multilingual reasoning, focusing on challenges and desiderata for cross-lingual inference.

### 2.1 Preliminaries

**Large Language Models (LLMs).** LLMs are transformer-based [Vaswani, 2017] neural network architectures designed to model the probability of a sequence of tokens. Formally, LLMs are trained to predict the probability of a word (or sub-word token) given the preceding words in a sequence  $X = \{x_1, x_2, \dots, x_n\}$ , formalized as:

$$P(X) = \prod_{i=1}^{n} P(x_i \mid x_1, x_2, \dots, x_{i-1}),$$

where P(X) is the probability of the entire sequence and  $P(x_i|x_1,x_2,\ldots,x_{i-1})$  is the conditional probability of the  $i^{th}$  token given the preceding tokens.

**Reasoning.** One of the key reasons behind the success of LLMs in mathematic and logical tasks is their reasoning capabilities. Formally, reasoning enables LLMs to draw logical conclusions C from premises P using a mapping function: C = f(P). To this end, there are different types of reasoning strategies that an LLM can employ:

a) **Deductive Reasoning:** It derives specific conclusions from general premises. If a given set of premises  $P_i$  are true, the

conclusion C must be true.

$$P_1, P_2, ..., P_n \Rightarrow C,$$

**b) Inductive Reasoning:** Generalizes patterns from specific instances, leading to probabilistic conclusions.

$$P_1, P_2, ..., P_n \Rightarrow C_{\text{probabilistic}}$$

c) Abductive Reasoning: Infers the most plausible explanation  $(H_{best})$  for given observation O.

$$O \Rightarrow H_{\text{best}}$$

**d) Analogical Reasoning:** Identifies relationships between domains and transfers knowledge.

$$A: B \approx C: D$$

**e) Commonsense Reasoning:** Uses real-world knowledge for intuitive decision-making.

## 2.2 Desiderata in Multilingual Reasoning

Here, we describe desiderata that lay the foundation for multilingual reasoning capabilities in LLMs. Multilingual reasoning refers to the capability of models to perform logical inference, problem-solving, and decision-making across multiple languages while maintaining consistency and cultural contextualization. Let  $L = \{l_1, l_2, \ldots, l_m\}$  represent a set of m languages, and let  $P_l$  and  $C_l$  denote the premise and conclusion in a given language  $l_i$ . For a multilingual reasoning model M, the task can be defined as:  $M(P_{l_i}) \to C_{l_i}, \quad \forall l_i \in L$ , where M must satisfy the following key desiderata:

**1. Consistency:** A model should make logically equivalent conclusions across languages for semantically equivalent premises, *i.e.*,

$$C_{l_i} \approx C_{l_j}$$
, if  $P_{l_i} \equiv P_{l_j}$ ,  $\forall l_i, l_j \in L$ ,

where  $\equiv$  indicates semantic equivalence of premises across languages. Consistency ensures that logical conclusions remain invariant of the input language.

**2. Adaptability:** For languages  $l_k \in L_{\text{low-resource}}$ , the model must generalize effectively using cross-lingual transfer from high-resource languages and perform robust reasoning, *i.e.*,

$$\forall l_k \in L_{\text{low-resource}}, \quad M(P_{l_k}) \to C_{l_k},$$

where transfer is performed from  $l_k \in L_{\text{high-resource}}$ .

**3. Cultural Contextualization:** Reasoning should consider cultural and contextual differences inherent to each language, *i.e.*, for a context  $c_{l_i}$  specific to language  $l_i$ , the conclusion  $C_{l_i}$  should adapt accordingly:

$$C_{l_i} = f(P_{l_i}, c_{l_i}), \quad \forall l_i \in L,$$

where f is a mapping function that integrates linguistic reasoning with cultural nuances, critical for tasks like healthcare, policy enforcement, and education, where culturally informed reasoning is required.

**4. Cross-Lingual Alignment:** The model must align reasoning processes across typologically diverse languages, where typology refers to linguistic differences in syntax, morphology, and structure (*e.g.*, word order variations between English and

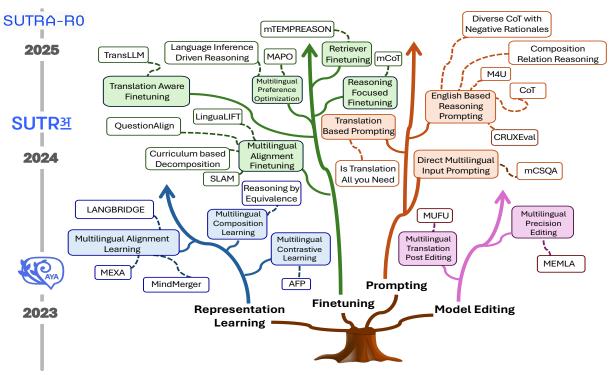


Figure 2: **Taxonomy tree of current Multilingual Research.** The thrusts for improving multilingual reasoning mainly include representation learning, fine-tuning, prompting, and model editing. With the emergence of multilingual LLMs, while initial research focused on naive prompting, recent works propose several alignment, editing, and fine-tuning strategies to improve reasoning in multilingual LLMs.

Japanese). Given the typological variations  $T_{l_i}$  and  $T_{l_j}$  for languages  $l_i$  and  $l_j$ , alignment ensures that reasoning remains consistent and coherent across languages, regardless of their structural differences, *i.e.*,

if 
$$P_{l_i} \equiv P_{l_j}$$
,  $M(P_{l_i}) \approx M(P_{l_j})$ ,  $\forall l_i, l_j \in L$ .

Next, we highlight existing works that propose different training corpus and benchmarks for multilingual reasoning in Sec. 3 and then describe previously proposed techniques to improve multilingual reasoning of LLMs in Sec. 4. The different thrusts of multilingual reasoning research are shown in Figure 2.

# 3 Multilingual Reasoning Datasets

The role of multilingual datasets in reasoning tasks is crucial for developing robust and linguistically diverse LLMs. Models trained on monolingual corpora often exhibit language biases [Lyu et al., 2024], limiting their reasoning capabilities across non-English languages. multilingual datasets play a vital role in ensuring equitable model performance, particularly in high-stakes domains such as science, law, and healthcare [Hendrycks et al., 2020]. Additionally, we need robust benchmarks to evaluate the effectiveness of various LLMs and techniques in handling critical domain-specific reasoning queries across both highresource and low-resource languages [Xu et al., 2024; Rasiah et al., 2024]. Below, we analyze the datasets based on training datasets (Sec. 3.1), and evaluation benchmarks (Sec. 3.2), comprising domains, tasks, and language distribution in current multilingual reasoning datasets.

# **3.1** Training Corpus

The best strategy to equip a language model with a specific type of reasoning is to train the model on it. However, the training objective differs based on the use case, domain, and the language in which the model needs to be adapted. For example, to perform some form of mathematical reasoning (e.g., GSM8K [Cobbe et al., 2021] and OpenMathQA [Amini et al., 2019] are large datasets that are used to improve mathematical reasoning) in a particular language, it needs to be trained with mathematical reasoning datasets, which will differ if we want to adapt the model for legal reasoning. While most training corpora are predominantly based on mathematical reasoning, XCSQA [Zhu et al., 2024] and MultiNLI [Williams et al., 2017] are used for enhancing logical and coding reasoning and sPhinX [Ahuja et al., 2024] is developed to translate instruction-response pairs into 50 languages for fine-tuning. In addition, there are cases where translation datasets like OPUS [Tiedemann, 2012], FLORES-200 [Goyal et al., 2022], and LegoMT [Yuan et al., 2022] are used to map the multilingual representation into the representation space of LLMs. Further, Ponti et al. [2020] introduced XCOPA, showing that multilingual pre-training and zero-shot fine-tuning under-perform compared to translation-based transfer. We argue that, moving forward, selecting the appropriate dataset and training methodology is crucial for optimizing a model's performance in specialized reasoning tasks.

### 3.2 Evaluation Benchmark

Evaluation benchmarks are fundamental to advancing the field of multilingual reasoning as they provide a consistent and sys-

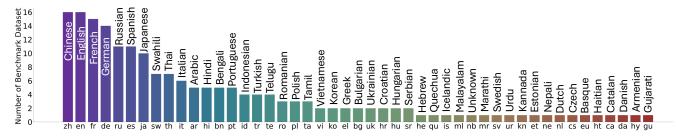


Figure 3: Language distribution across training corpora and benchmarks for multilingual reasoning. The y-axis denotes the number of training corpora/benchmark datasets that include a given language (x-axis). We observe a long-tail distribution denoting that current datasets predominantly cover languages like Chinese, English, French, and German, highlighting the need for benchmarks that represent long-tail languages.

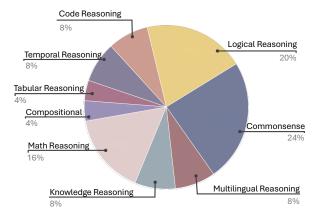


Figure 4: **Distribution of multilingual reasoning datasets.** We observe that datasets predominantly comprise logical, commonsense, and math reasoning datasets, and the community needs more benchmarks to include temporal, compositional, and tabular reasoning.

tematic framework to assess the performance of models across diverse reasoning tasks, such as logical inference, mathematical problem-solving, and cross-lingual understanding. Each reasoning task and domain presents unique challenges, making it crucial to have tailored benchmarks that reflect the specific requirements and complexities of those tasks. Below, we analyze the evaluation benchmarks on three key aspects namely domain (Fig. 4), task (Fig. 5), and languages (Fig. 3) in detail.

#### **Domains and Tasks Covered**

Multilingual reasoning in language models spans multiple domains, each with its own set of complexities and requirements. Understanding these differences is essential for developing language models that can effectively adapt to various applications. For instance, Cobbe et al. [2021] highlighted that mathematical reasoning requires structured multi-step logic and datasets like GSM8K. While Ponti et al. [2020] showed that causal reasoning in XCOPA relies on cross-lingual consistency and commonsense inference, Tiedemann [2012] noted that multilingual reasoning introduces typological challenges. These studies emphasize the need for tailored approaches to address the specific demands of each task and domain. Hence, it is crucial to build reliable and robust benchmarks for all domains and tasks to develop more robust techniques tailored to handle the complexity of a particular domain and task. Figures 4-5 show the distribution of datasets across vari-

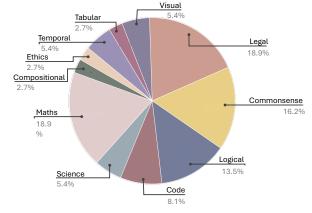


Figure 5: **Distribution of domains in multilingual reasoning datasets.** While legal, commonsense, and math domain dataset cover up to 54% of current multilingual reasoning research, other under-explored domains include ethics, science, visual, and compositional.

ous domains and tasks, highlighting the need to develop more comprehensive benchmarks across multiple domains. Currently, tasks such as math, legal, and commonsense reasoning dominate multilingual evaluation benchmarks, collectively accounting for 54% of the total (see Fig. 5). In contrast, domains like science, visual reasoning, tabular reasoning, temporal reasoning, and ethics are underrepresented, covering only 35%. Furthermore, while commonsense reasoning makes a significant contribution to existing works, other reasoning tasks lag. Notably, crucial domains such as finance and healthcare still lack dedicated evaluation benchmarks for multilingual reasoning, highlighting a significant gap in the field.

# **Languages Covered**

Comprehensive language coverage is vital for multilingual reasoning, ensuring inclusivity and balanced performance across high-resource and low-resource linguistic communities. Based on languages, current benchmarks can be primarily classified into human languages and coding languages. Benchmarks like XNLI [Conneau *et al.*, 2018], mCSQA [Sakai *et al.*, 2024], and MARC [Keung *et al.*, 2020] predominantly focus on high-resource languages like English, Chinese, French, and Spanish. While some efforts include low-resource languages like Swahili (XCOPA [Ponti *et al.*, 2020]), Haitian (M4U [Wang *et al.*, 2024]), and Nepali (mMMLU [Hendrycks *et al.*, 2020]), **their representation remains minimal** and research in these

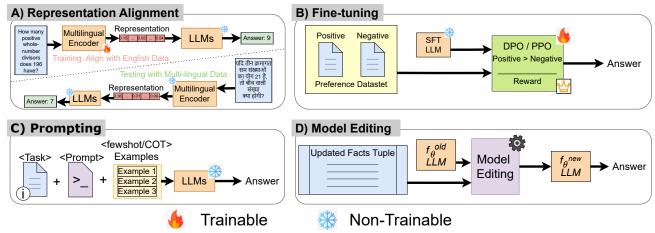


Figure 6: **Taxonomy of Multilingual Reasoning Methods.** A categorization of approaches for enhancing multilingual reasoning in language models, covering (A) Representation Alignment, (B) Finetuning, (C) Prompting, and (D) Model Editing.

languages remain at a nascent stage. Typologically distant and underrepresented languages, such as Kannada, Gujarati (xSTREET [Li et al., 2024a]), and Quechua, are rarely included, **further widening linguistic inequities.** Datasets like FLORES-200 attempt to balance high- and low-resource languages but fail to achieve comprehensive coverage. We note that to ensure LLMs perform effectively across diverse linguistic and cultural contexts, it is critical to include a broader range of low-resource and endangered languages [Goyal et al., 2022; Amini et al., 2019]. The complete distribution of human languages across benchmarks is shown in Figure 3. Finally, only three benchmarks (BBH-Hard, CRUXEval, and TCC) incorporate coding languages across multiple languages.

### 4 Methods

Multilingual reasoning within language models has garnered significant attention in recent years, leading to the development of diverse techniques for enhancing their capabilities across diverse languages. Prior works have explored various directions to improve multilingual reasoning. Building upon this body of work (see Figure 6), we identify four primary thrusts, *viz.* representation alignment, fine-tuning, prompting, and model editing, that collectively contribute to advancing multilingual reasoning in language models.

a) Representation Alignment. Multilingual reasoning requires consistent representations across languages, but LLMs often struggle due to imbalanced training data. Representation alignment ensures that equivalent concepts share similar embeddings, reducing inconsistencies in cross-lingual inference, essential for reasoning and multilingual generalization. Li et al. [2024b] employs Multilingual Contrastive Learning to align multilingual sentence representations by treating translation pairs as positive samples and pulling their embeddings closer, bridging representation gaps between languages, enhancing cross-lingual reasoning and generation capabilities. Multilingual Alignment Learning is another form of representation alignment that ensures semantic consistency across languages by aligning their representations for improved multilingual performance, where prior works leverage external multilingual models to enhance reasoning by aligning multilingual capabilities with LLMs [Huang et al., 2024b], bridge multilingual encoders with LLMs using minimal parameters to achieve effective alignment without supervision [Yoon et al., 2024], and evaluate the semantic similarity between English and other languages within LLM embeddings to measure alignment quality [Kargaran et al., 2024]. Furthermore, an exciting new direction in representation alignment is **Multilingual Compositional Learning** which constructs compositional representations by combining equivalent token embeddings across multiple languages [Arora et al., 2024].

b) Finetuning. It leverages cross-lingual data and tasks to fine-tune models for enhanced reasoning and comprehension, leading to numerous innovative approaches. One such approach, exemplified by LinguaLIFT [Zhang et al., 2024], uses code-switched fine-tuning along with language alignment layers to effectively bridge the gap between English and languages with fewer resources, helping maintain the nuance and context across linguistic boundaries. Similarly, QuestionAlign [Zhu et al., 2024] takes a step further by aligning questions and responses in multiple languages, thereby enhancing cross-lingual understanding and consistency in reasoning. While these methods have leaned towards extensive finetuning, SLAM [Fan et al., 2025] introduces a more parameterefficient strategy, which selectively tunes only layers critical for multilingual comprehension, significantly lowering the computational demands while still maintaining or even enhancing the model's reasoning capabilities. Translation has also been harnessed as a powerful tool for knowledge transfer in multilingual settings. TransLLM [Geng et al., 2024], for instance, focuses on translation-aware fine-tuning to align different languages, thereby improving the model's ability to perform reasoning across languages. This method not only enhances language understanding but also adapts the model for various cross-lingual tasks. For those aiming at more complex reasoning tasks, reasoning-focused fine-tuning has proven beneficial. The Multilingual Chain-of-Thought (mCoT) Instruction Tuning method [Lai and Nissim, 2024] utilizes a dataset specifically curated for reasoning across languages (mCoT-MATH) and combines CoT reasoning with instruction tuning to boost consistency and logical problem-solving in

multiple languages. In addition, preference-based techniques to align reasoning outputs across languages emphasizes the use of language imbalance as a reward signal in models like DPO (Direct Preference Optimization) and PPO (Proximal Policy Optimization) [She *et al.*, 2024]. Finally, an interesting direction moving forward is the use of curriculum-based and retriever-based fine-tuning techniques to enhance multilingual reasoning [Anand *et al.*, 2024; Bajpai and Chakraborty, 2024].

- c) Prompting. Prompting has emerged as a key technique for enhancing how LLMs adapt and reason across different languages. By guiding the model through specific strategies, prompting not only facilitates dynamic adaptation but also addresses the challenges posed by data imbalances, thereby enhancing cross-lingual consistency, logical alignment, and the robustness of reasoning. One effective method is Direct Multilingual Input Prompting, as explored in the study by Sakai et al. [2024]. Here, the model directly processes inputs in various native languages without any need for translation, preserving the original linguistic nuances. This approach was notably applied in the paper "Do Moral Judgements" by Khandelwal et al. [2024], where moral scenarios were presented in their native languages to assess the model's reasoning capabilities directly. Another strategy, Translation-Based Prompting [Liu et al., 2024] uses translation to convert multilingual inputs into a pivot language for processing, where tasks are translated into English for reasoning and translated back to the target language for evaluation [Wang et al., 2024; Zhao and Zhang, 2024], where one can also generate diverse CoT with Negative Rationales (d-CoT-nR) by incorporating both correct and incorrect reasoning paths to refine multilingual reasoning capabilities [Payoungkhamdee et al., 2024].
- d) Model Editing. Model (or knowledge) editing is a growing and exciting research area that aims to modify/update the information stored in a model. Formally, model editing strategies update pre-trained models for specific input-output pairs without retraining them and impacting the baseline model performance on other inputs. Multilingual Precision Editing involves making highly specific updates to model knowledge while ensuring minimal impact on unrelated information. Multilingual knowledge Editing with neuron-Masked Low-Rank Adaptation (MEMLA) [Xie et al., 2024] enhances multilingual reasoning by leveraging neuron-masked LoRA-based edits to integrate knowledge across languages and improve multi-hop reasoning capabilities. Further, Multilingual Translation Post-editing refines translations by correcting errors in multilingual outputs for better alignment and fluency, where we can enhance multilingual reasoning by incorporating auxiliary translations into the post-editing process, enabling LLMs to improve semantic alignment and translation quality across languages [Lim et al., 2024].

# 5 Evaluation Metrics and Benchmarks

Evaluating multilingual reasoning in LLMs requires standardized metrics to ensure logical consistency and cross-lingual coherence. Unlike traditional NLP, it must address inference errors, translation drift, and reasoning stability across languages.

### 5.1 Metrics

Here, we categorize the key evaluation metrics for multilingual reasoning, along with their formal definitions:

- 1) Accuracy-Based Metrics. These metrics assess overall correctness in reasoning and multilingual benchmarks: i) *General Accuracy* measures the proportion of correct outputs over total samples and ii) *Zero-Shot Accuracy*, which evaluates model performance on unseen tasks or categories without fine-tuning.
- **2) Reasoning and Consistency Metrics.** These metrics evaluate logical inference and multi-step reasoning ability: i) *Reasoning Accuracy*: Assesses correctness in logical and step-by-step reasoning tasks and ii) *Path Consistency*: Measures coherence between reasoning steps in CoT prompting.
- **3) Translation and Cross-Lingual Metrics.** To ensure multilingual reasoning consistency, models must preserve meaning across languages: i) *Translation Success Rate (TSR)*: Measures correctness and semantic preservation in multilingual translations as the ratio of accurate translations and total translations and ii) *Cross-Lingual Consistency*: Evaluates whether logically equivalent statements yield *consistent reasoning outputs* across different languages.
- **4) Perplexity and Alignment Metrics.** These metrics quantify *semantic alignment* and measure whether embeddings across languages remain consistent: i) *Perplexity-Based Alignment*  $(P_{align})$ :

$$P_{\text{align}} = \exp\left(-\frac{1}{N} \sum_{i=1}^{N} \log P(x_i)\right),\tag{1}$$

where  $P(x_i)$  is the model's probability of predicting token  $x_i$ . Lower perplexity indicates better alignment and ii) *Semantic Alignment Score* ( $S_{align}$ ) that measures the cosine similarity between multilingual sentence embeddings:

$$S_{\text{align}} = \frac{E_l \cdot E_t}{\|E_l\| \|E_t\|},\tag{2}$$

where  $E_l$  and  $E_t$  are sentence embeddings in different languages.

# 5.2 Performance on Benchmarks

Here, we discuss the performance of the aforementioned methods on standard mathematical reasoning benchmarks (MGSM [Shi *et al.*, 2022] and MSVAMP [Chen *et al.*, 2023]), common sense reasoning (xCSQA [Lin *et al.*, 2021]), and logical reasoning (xNLI [Conneau *et al.*, 2018])<sup>1</sup>. In Fig. 7, the x-axis is the timestamp of their arXiv submission, and the y-axis is the predictive performance (accuracy) of the methods on those benchmarks. Next, we describe the four most popular benchmarks and detail the performance of state-of-the-art reasoning techniques on them, highlighting existing gaps in language models that limit their reasoning performance.

MGSM tests multilingual arithmetic reasoning in LLMs with 250 translated math problems in ten diverse languages. While recent trends suggest that advanced post-training techniques like MAPO are crucial for strong performance, fine-tuning strategies may be more impactful than stronger reasoning

<sup>&</sup>lt;sup>1</sup>We only cover benchmarks analyzed by more than four papers.



Figure 7: Accuracy trends of various methodologies on multilingual reasoning benchmarks, including MGSM, MSVAMP, XNLI, and XCSQA. The x-axis represents the date of paper submission on arXiv, while the y-axis indicates percentage accuracy.

architectures or relying on the model's English expertise to improve multilingual performance.

**MSVAMP** is an out-of-domain multilingual mathematical reasoning dataset comprising 10,000 problems across ten languages and serves as a comprehensive test bed to evaluate LLMs' generalization in multilingual mathematical contexts. We find that advanced preference optimization (MAPO) achieves much stronger performance than CoT-based fine-tuning, suggesting advanced fine-tuning techniques are a better direction to beat the current best in this benchmark.

xCSQA is a multilingual extension of the CommonsenseQA dataset, encompassing 12,247 multiple-choice questions translated into 15 languages, designed to assess LLMs' crosslingual commonsense reasoning capabilities. The current trend in this benchmark shows that stronger fine-tuning strategies like two-step fine-tuning (LingualLIFT) or preference optimization (MAPO) show better performance than selectively fine-tuning specific layers as in SLAM.

**xNLI** evaluates cross-lingual sentence inference across 15 languages, including low-resource ones. Recent studies suggest that LLM integration with external models [Huang *et al.*, 2024b] and multilingual alignment followed by fine-tuning [Zhang *et al.*, 2024] outperform contrastive learning methods like TCC [Chia *et al.*, 2023], highlighting the need for more structured multilingual adaptation strategies.

## 6 Future Directions

As we rapidly develop the new generation of reasoning models, our community must ensure that the models remain unbiased towards languages from underrepresented regions around the world. Looking forward, we call on the community to put their collective efforts into some of the following future directions:

- 1. Multilingual Alignment and Reasoning Transfer. A key challenge in multilingual reasoning is the lack of data in different languages. One promising solution is to leverage existing large datasets and transfer/distill their knowledge in the representation space, as demonstrated by [Yoon et al., 2024; Huang et al., 2024b]. Future research should focus on improving cross-lingual knowledge transfer techniques, enabling models to use high-resource languages as a bridge to enhance reasoning consistency in low-resource languages. Another direction is to generate synthetic datasets using techniques like back-translation and data augmentation, tailored specifically for reasoning tasks.
- 2. Explainable and Interpretable Reasoning. Ensuring

faithful and interpretable reasoning in multilingual LLMs is challenging due to linguistic diversity, translation ambiguities, and reasoning inconsistencies. Studies on English CoT reasoning [Tanneru *et al.*, 2024; Lobo *et al.*, 2024] highlight faithfulness issues, which become more severe when extended to low-resource languages. Causal reasoning can enhance crosslingual alignment, improving interpretability by uncovering cause-effect relationships across languages. Future research should focus on integrating causal reasoning and multilingual CoT frameworks to ensure logical coherence, transparency, and trust in multilingual AI systems.

- **3.** Advanced Training and Inference Techniques. While recent advancements in multilingual reasoning have introduced reasoning-aware fine-tuning and multilingual preference optimization techniques, further efforts are needed to improve training paradigms. Notably, Wu *et al.* [2024] demonstrates that reward signals can be effectively transferred across languages, paving the way for post-training RL methods that improve reasoning in low-resource languages. Additionally, efficient inference-time scaling and agentic frameworks remain under-explored, despite emerging techniques [Khanov *et al.*, 2024; Chakraborty *et al.*, 2024] gaining traction and multi-agent frameworks [Guo *et al.*, 2024] enabling LLMs to simulate agent interactions or refine their reasoning by learning from self-generated reasoning paths.
- **4. Unified Evaluation Metrics.** A comprehensive evaluation framework is a crucial missing component for assessing multilingual reasoning capabilities. Metrics should measure logical consistency, cultural adaptability, and robustness, considering real-world and adversarial multilingual settings.
- **5. Benchmarks:** As multilingual reasoning advances, robust evaluation benchmarks are essential. As reasoning is highly domain-specific in nature, developing targeted benchmarks is crucial, especially in high-stakes fields like healthcare, law, and finance, where accuracy directly affects decision-making.
- **6. Multimodal Multilingual Reasoning.** While there are a few works on visual reasoning in the multilingual context [Das *et al.*, 2024], multimodal reasoning (integrating tables, text, image, audio and video) remains largely unexplored. Advancing this area could enable models handle complex tasks in low-resource languages and incorporate cross-modal reasoning.

# 7 Conclusion

Multilingual reasoning in Large Language Models (LLMs) is a rapidly evolving field, addressing critical challenges like cross-lingual alignment, low-resource language gaps, and cultural adaptation. Our survey highlights advancements in fine-tuning, prompting, and representation learning while identifying gaps in scalability, ethical reasoning, and domain-specific applications. It serves as a call to action for the LLM and reasoning research and development community to focus on advanced alignment techniques, culturally aware reasoning, scalable architectures, and responsible AI outputs. By breaking language barriers and fostering inclusivity, multilingual reasoning has the potential to create globally impactful AI systems. This survey provides a foundation for advancing research in this transformative domain.

## References

- Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Awadallah, Monojit Choudhary, Vishrav Chaudhary, and Sunayana Sitaram. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. *arXiv*, 2024.
- Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL*, 2019.
- Avinash Anand, Kritarth Prasad, Chhavi Kirtani, Ashwin R Nair, Manvendra Kumar Nema, Raj Jaiswal, and Rajiv Ratn Shah. Multilingual mathematical reasoning: Advancing open-source llms in hindi and english. *arXiv preprint arXiv:2412.18415*, 2024.
- Gaurav Arora, Srujana Merugu, Shreya Jain, and Vaibhav Saxena. Towards robust knowledge representations in multilingual llms for equivalence and inheritance based consistent reasoning. *arXiv preprint arXiv:2410.14235*, 2024.
- Ashutosh Bajpai and Tanmoy Chakraborty. Multilingual llms inherently reward in-language time-sensitive semantic alignment for low-resource languages. *arXiv preprint arXiv:2412.08090*, 2024.
- Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer q star: Principled decoding for llm alignment. *arXiv preprint arXiv:2405.20495*, 2024.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv* preprint arXiv:2310.20246, 2023.
- Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*, 2023.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv*, 2021.
- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv*, 2018.
- Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv* preprint arXiv:2403.10378, 2024.
- Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. Slam: Towards efficient multilingual reasoning via selective language alignment. *arXiv preprint arXiv:2501.03681*, 2025.

- Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, et al. Why not transform chat large language models to non-english? *arXiv preprint arXiv:2405.13923*, 2024.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc' Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-200 evaluation benchmark for low-resource and multilingual machine translation. In *EMNLP*. ACL, 2022.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv* preprint *arXiv*:2402.01680, 2024.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv*, 2020.
- Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, et al. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv*, 2024.
- Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. Mindmerger: Efficient boosting llm reasoning in non-english languages. arXiv preprint arXiv:2405.17386, 2024.
- Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. Mexa: Multilingual evaluation of english-centric llms via crosslingual alignment. *arXiv preprint arXiv:2410.05873*, 2024.
- Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. The multilingual amazon reviews corpus. *arXiv*, 2020.
- Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. *arXiv preprint arXiv:2402.02135*, 2024.
- Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. *arXiv preprint arXiv:2402.01694*, 2024.
- Huiyuan Lai and Malvina Nissim. mcot: Multilingual instruction tuning for reasoning consistency in language models. *arXiv*, 2024.
- Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. Eliciting better multilingual structured reasoning from llms through code. *arXiv* preprint *arXiv*:2403.02567, 2024.
- Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *NAACL: Human Language Technologies*, 2024.
- Zheng Wei Lim, Nitish Gupta, Honglin Yu, and Trevor Cohn. Mufu: Multilingual fused learning for low-resource translation with llm. *arXiv preprint arXiv:2409.13949*, 2024.

- Yankai Lin, Jiapeng Zhou, Yiming Shen, Wenxuan Zhou, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Xcsqa: A benchmark for cross-lingual conversational question answering. In *EMNLP*, 2021.
- Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*, 2024.
- Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. *arXiv*, 2024.
- Jiachen Lyu, Katharina Dost, Yun Sing Koh, and Jörg Wicker. Regional bias in monolingual english language models. *Machine Learning*, 2024.
- Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *arXiv*, 2024.
- Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. An empirical study of multilingual reasoning distillation for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7739–7751, 2024.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. In *EMNLP*. ACL, 2020.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv*, 2024.
- Raghav Ramji and Keshav Ramji. Inductive linguistic reasoning with large language models. *arXiv*, 2024.
- Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. One law, many languages: Benchmarking multilingual legal reasoning for judicial support, 2024.
- Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. *arXiv*, 2024.
- Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. Mapo: Advancing multilingual reasoning through multilingual alignment-aspreference optimization. *arXiv preprint arXiv:2401.06838*, 2024.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv*, 2022.
- Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv*, 2024.

- Jörg Tiedemann. Opus: An open source parallel corpus, 2012. A Vaswani. Attention is all you need. *NeurIPS*, 2017.
- Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.
- Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv*, 2017.
- Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. *arXiv*, 2024.
- Jiakuan Xie, Pengfei Cao, Yuheng Chen, Yubo Chen, Kang Liu, and Jun Zhao. Memla: Enhancing multilingual knowledge editing with neuron-masked low-rank adaptation. *arXiv preprint arXiv:2406.11566*, 2024.
- Ruiyang Xu, Jialun Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, Ben He, Shing-Chi Cheung, and Le Sun. Cruxeval-x: A benchmark for multilingual code reasoning, understanding and execution. *arXiv*, 2024.
- Wenlin Yao, Haitao Mi, and Dong Yu. Hdflow: Enhancing Ilm complex problem-solving with hybrid thinking and dynamic workflows. *arXiv*, 2024.
- Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. Langbridge: Multilingual reasoning without multilingual supervision. *arXiv* preprint arXiv:2401.10695, 2024.
- Fei Yuan, Yinquan Lu, WenHao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. Lego-mt: Learning detachable models for massively multilingual machine translation. *arXiv*, 2022.
- Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. Lingualift: An effective two-stage instruction tuning framework for low-resource language tasks. *arXiv* preprint arXiv:2412.12499, 2024.
- Jinman Zhao and Xueyan Zhang. Large language model is not a (multilingual) compositional relation reasoner. In *First Conference on Language Modeling*, 2024.
- Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *arXiv*, 2024.