A Survey of Multilingual Reasoning in Language Models

Akash Ghosh[1*] **Debayan Dutta**[1*] **Sriparna Saha**[1] **Chirag Agarwal**[2]

[1]Indian Institute of Technology Patna, India
[2]University of Virginia, USA

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-of-the-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. Rapid growth of evolving developments in this field can be actively tracked on our project page: https://github.com/AkashGhosh/Survey-of-Multilingual-Reasoning-in-Language-Models

1 Introduction

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

Ludwig Wittgenstein

Large Language Models (LLMs) (Vaswani, 2017) have emerged as transformative tools in natural language processing, demonstrating state-of-the-art performance in language generation, translation, and summarization(Jain et al., 2022; Ghosh et al., 2024a,d,b, 2025; Ghosal et al., 2025). 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 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, enabling individuals and systems to solve problems and make complex decisions. Recent advancements have sought to enhance the reasoning capabilities of LLMs using Chain-of-Thought (CoT) (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, where 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.

* Equal contribution. Work done while interning at Aikyam Lab (UVA). Contact author: akash_2321cs19@iitp.ac.in

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8920–8936
November 4-9, 2025

Multilingual reasoning, which combines logical inference with multilingual capabilities, is essential for creating AI systems that effectively operate 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 key contributions of our work are:

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 LLMs have improved their reasoning capabilities. However, extending them 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 neural network architectures designed to model the probability of a sequence of tokens. Formally, LLMs are trained to predict the likelihood of a word (or sub-word token) given the preceding words in a sequence $X = \{x_1, \ldots, x_n\}$, *i.e.,* $P(X) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$, where $P(X)$ is the probability of the entire sequence and $P(x_i|x_1, \ldots, x_{i-1})$ is the conditional probability of the ith token given the preceding tokens.

Reasoning. One of the key reasons behind the success of LLMs in mathematical 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: Derives logically certain conclusions from general premises. If the premises P_i are true, the conclusion C must also be true, *i.e.,* $P_1, P_2, \ldots, P_n \Rightarrow C$.

b) Inductive Reasoning: Infers general rules or patterns from specific observations, leading to conclusions that are likely but not guaranteed, *i.e.,* $P_1, P_2, \ldots, P_n \Rightarrow C_{\text{probabilistic}}$.

c) Abductive Reasoning: Infers the most plausible hypothesis (H_{best}) that explains an observation O, though the inference is not guaranteed to be correct, *i.e.,* $O \Rightarrow H_{\text{best}}$.

d) Analogical Reasoning: Transfers knowledge by identifying relational similarities between domains, *i.e.,* $A : B \approx C : D$.

e) Commonsense Reasoning: Draws on background knowledge of everyday situations to make intuitive, contextually appropriate inferences.

2.2 Desiderata in Multilingual Reasoning

Here, we describe desiderata that lay the foundation for multilingual reasoning in LLMs. 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: $M(P_{l_i}) \rightarrow C_{l_i}, \quad \forall l_i \in L$, 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.,* $C_{l_i} \approx C_{l_j}, \quad \text{if } P_{l_i} \equiv P_{l_j}, \quad \forall l_i, l_j \in L$, 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.,* $\forall l_k \in L_{\text{low-resource}}, \quad M(P_{l_k}) \rightarrow C_{l_k}$.

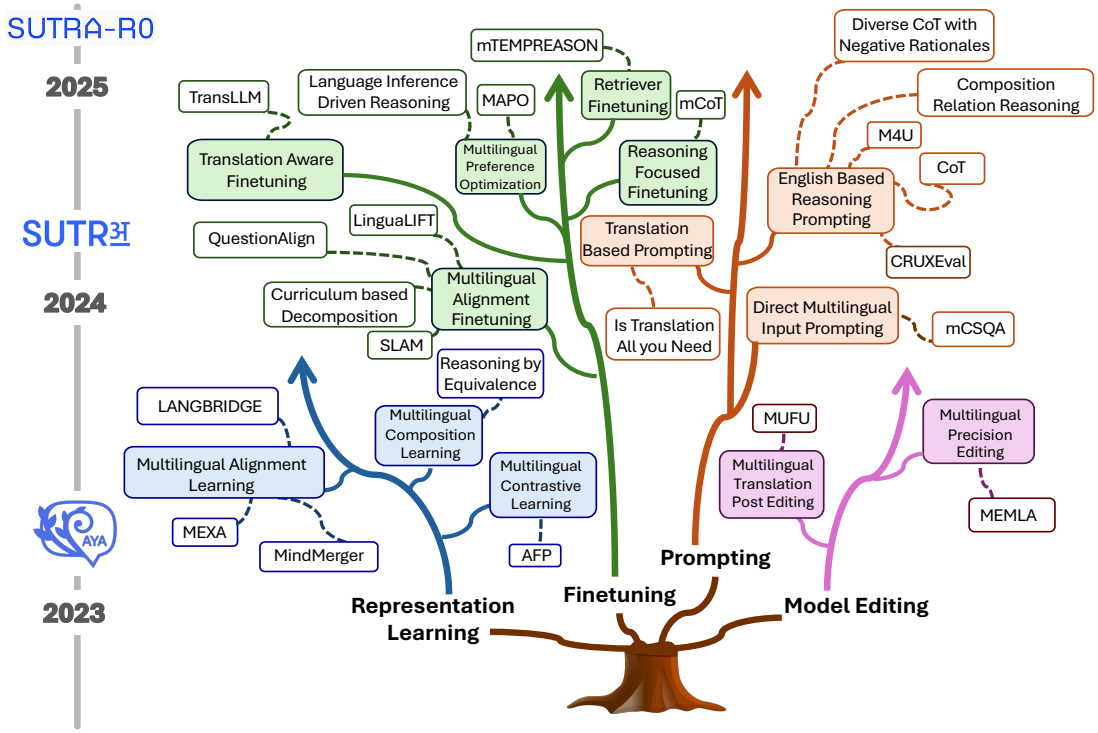

Figure 1: **Taxonomy tree of current Multilingual Reasoning Research.** The thrusts for improving multilingual reasoning mainly include representation learning, fine-tuning, prompting, and model editing. With the emergence of multilingual LLMs, while initial research focused on naive prompting, recent works propose several alignment, editing, and fine-tuning strategies to improve reasoning in multilingual LLMs.

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.

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 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, *i.e.,* if $P_{l_i} \equiv P_{l_j}, \quad M(P_{l_i}) \approx M(P_{l_j}), \quad \forall l_i, l_j \in L$. Next, we highlight existing works that propose different training corpora and benchmarks for multilingual reasoning in Sec. 3 and then describe previously proposed techniques to improve the multilingual reasoning of LLMs in Sec. 4.

3 Multilingual Reasoning Datasets

Models trained on english corpora exhibit language biases (Lyu et al., 2024), limiting their reasoning capability on non-English languages. Training an LM to solve math problems across languages requires multilingual understanding and mathematical reasoning (Son et al., 2024). Hence, multilingual datasets and benchmarks play a key role in training multilingual LMs and evaluating the effectiveness of various LMs and techniques in handling domain-specific reasoning queries across low- and high-resource languages (Xu et al., 2024; Rasiah et al., 2024; Xue et al., 2024). Below, we detail training datasets (Sec. 3.1) and benchmarks (Sec. 3.2), comprising domains, tasks, and language distribution in current multilingual reasoning datasets.

3.1 Training Corpus

The best strategy to equip an LM 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 language in which the model needs to be adapted. For example, to perform mathematical reasoning (Cobbe et al., 2021; Amini et al., 2019) 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., 2024b) and MultiNLI (Williams et al., 2017)

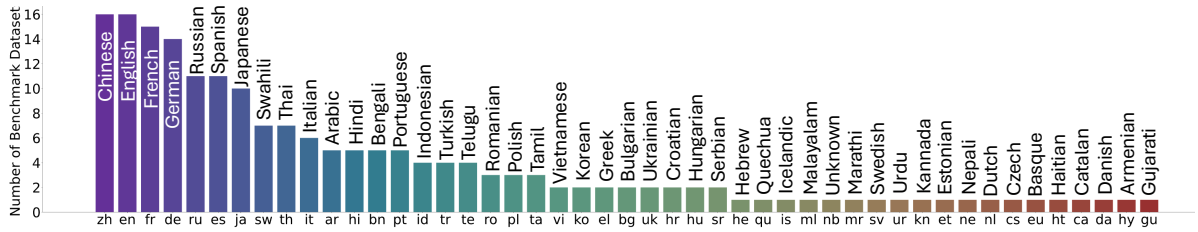

Figure 2: **Language distribution across training corpora and benchmarks for multilingual reasoning.** The y-axis denotes the number of training corpora/benchmark datasets that include a given language (x-axis). We observe a long-tail distribution, denoting that current datasets predominantly cover languages like Chinese, English, French, and German, highlighting the need for benchmarks that represent long-tail languages.

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 LM's representation space. Further, Ponti et al. (2020) introduced XCOPA to show that multilingual pre-training and zero-shot fine-tuning underperform 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.

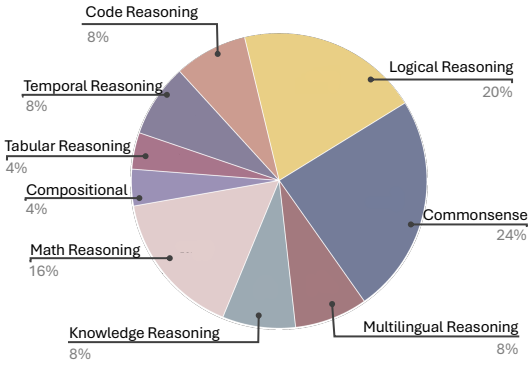

Figure 3: **Distribution of multilingual reasoning datasets.** We find that datasets predominantly comprise logical, commonsense, and math reasoning, and the community needs benchmarks to include compositional and tabular reasoning.

3.2 Evaluation Benchmark

Benchmarks are key to advancing the field of multilingual reasoning as they provide a systematic framework to assess the performance of models across diverse reasoning tasks. Each reasoning task and domain presents unique challenges, making it crucial to have tailored benchmarks that reflect specific requirements and complexities of those tasks. Below, we analyze the evaluation benchmarks on three key aspects, namely languages (Fig. 2), domain (Fig. 3), and task (Fig. 4).

3.2.1 Domains and Tasks Covered

Multilingual reasoning in LMs spans multiple domains, each with its complexities and requirements, and understanding these differences is essential for developing LMs that can effectively adapt to various applications. For instance, Cobbe et al. (2021) highlighted that mathematical reasoning requires structured multi-step logic and datasets. While Ponti et al. (2020) showed that causal reasoning in XCOPA relies on cross-lingual consistency and commonsense inference, Östling and Tiedemann (2016) 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 developing more robust techniques tailored to handle the complexity of a particular domain and task. Figs. 3-4 show the distribution of datasets across various 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 benchmarks, collectively accounting for **54%** of the total (Fig. 4). In contrast, domains like science, ethics, and visual, tabular, and temporal reasoning are underrepresented, covering only **35%**. Notably, **crucial domains such as finance and healthcare still lack dedicated evaluation benchmarks** for multilingual reasoning, highlighting a significant gap in the field.

3.2.2 Languages Covered

Comprehensive language coverage is vital for multilingual reasoning, ensuring inclusivity and balanced performance across low- and high-resource linguistic communities. Based on languages, current benchmarks can be primarily classified into

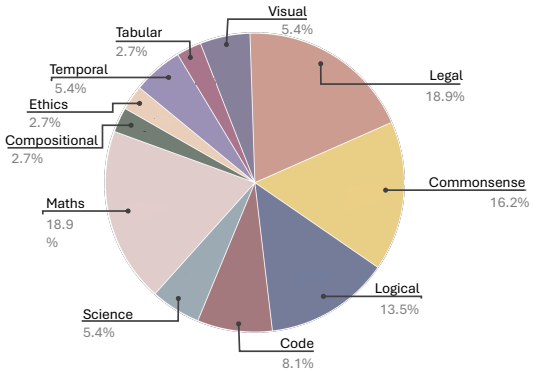

Figure 4: **Distribution of domains in multilingual reasoning datasets.** While legal, commonsense, and math domain dataset cover up to 54% of current multilingual reasoning research, other under-explored domains include ethics, science, visual, and compositional.

human and coding languages. Benchmarks like XNLI (Conneau et al., 2018), mCSQA (Sakai et al., 2024), and m-ARC (Lai et al., 2023) 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 languages remains 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 inequalities.** Datasets like FLORES-200 attempt to balance low- and high-resource languages but fail to achieve comprehensive coverage. To ensure effective LLM performance 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) (see the complete distribution of human languages across benchmarks in Fig. 2). Finally, only four benchmarks (Luo et al., 2024; Xu et al., 2024; Zhang et al., 2024b; Li et al., 2024a) incorporate coding languages across multiple languages.

4 Methods

Multilingual reasoning within LMs 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 Fig. 5), we identify four primary thrusts, *viz.* representation alignment, fine-tuning, prompting, and model editing, collectively contributing to advancing multilingual reasoning in LMs.

a) Representation Alignment. Multilingual reasoning requires consistent representations across languages, but LMs often struggle due to imbalanced training data. Representation alignment ensures that equivalent concepts share similar embeddings, reducing inconsistencies in multilingual inference, vital for reasoning and multilingual generalization. Li et al. (2024b) employs contrastive learning to align multilingual sentence representations by treating translation pairs as positive samples and pulling their embeddings closer, bridging language representation gaps and enhancing model's cross-lingual reasoning and generation capabilities. Multilingual Alignment Learning is another technique that ensures semantic consistency across languages by aligning their representations for improved multilingual performance (Huang et al., 2024b), bridging multilingual encoders with LLMs using minimal parameters to achieve effective alignment without supervision (Yoon et al., 2024; Kargaran et al., 2024). Similarly, Ruan et al. (2025) integrates all encoder layer representations and employs adaptive fusion-enhanced attention to enable layer-wise alignment between the LLM and multilingual encoder, ensuring consistent cross-lingual representations and improving the model's multilingual reasoning capabilities. Finally, an exciting new direction is multilingual compositional learning, which constructs compositional representations by combining equivalent token embeddings across multiple languages (Arora et al., 2024) and formalizing problems in an abstract space and solving them step-by-step using self-training for improved alignment across languages (Ranaldi and Pucci, 2025).

b) Finetuning. It leverages cross-lingual data and tasks to fine-tune models for enhanced reasoning and comprehension, leading to numerous innovative approaches. For instance, LinguaLIFT (Zhang et al., 2024a) uses code-switched fine-tuning along with language alignment layers to effectively bridge the gap between English and low-resource languages, helping maintain the nuance and context across linguistic boundaries. Similarly, QuestionAlign (Zhu et al., 2024b) aligns questions and responses in multiple languages, thereby enhancing cross-lingual understanding and consistency in reasoning and Ko et al. (2025) introduces a strategic fine-tuning approach that anchors reasoning in English and then translates

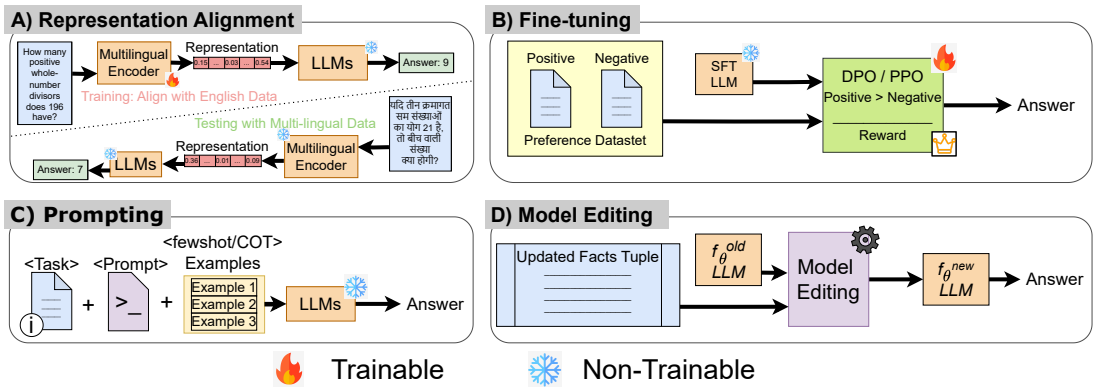

Figure 5: **Taxonomy of Multilingual Reasoning Methods.** A taxonomy of approaches for enhancing multilingual reasoning in models, covering (A) Representation Alignment, (B) Finetuning, (C) Prompting, and (D) Model Editing.

results, significantly reducing cross-lingual performance gaps. Strategic fine-tuning using a small but high-quality bilingual dataset can enhance both the reasoning capabilities and non-English language proficiency of LLMs (Ha, 2025). While these methods have leaned towards extensive fine-tuning, SLAM (Fan et al., 2025) introduces a more parameter-efficient strategy and selectively tunes 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, where TransLLM (Geng et al., 2024) focuses on translation-aware fine-tuning to align different languages, enhancing language understanding but also adapting the model for various cross-lingual tasks. For those aiming at more complex reasoning tasks, **reasoning-focused fine-tuning** has proven beneficial. The Multilingual CoT (mCoT) instruction tuning method (Lai and Nissim, 2024) utilizes a dataset specifically curated for reasoning across languages 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 emphasize the use of language imbalance as a reward signal in models like Direct Preference and Proximal Policy Optimization (She et al., 2024). Recent research has demonstrated that Process Reward Modeling offers fine-grained feedback at each step of the reasoning process, only Wang et al. (2025) has shown its application on non-English language. Finally, an interesting direction moving forward is curriculum-based and retriever-based fine-tuning techniques to enhance multilingual reasoning (Anand et al., 2024; Bajpai and Chakraborty, 2024), where models must not only retrieve relevant information but also compare them to evaluate relationships between them (Agrawal et al., 2024; Ranaldi et al., 2025b; Shao et al., 2024; Yang et al., 2025).

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 facilitates dynamic language adaptation and addresses the data imbalance challenge, thereby enhancing cross-lingual consistency, logical alignment, and the robustness of reasoning. For instance, an effective method is Direct Multilingual Input Prompting (Sakai et al., 2024), where the model directly processes inputs in various native languages without translation, preserving the original linguistic nuances. This approach was notably applied in the paper "*Do Moral Judgements*" (Khandelwal et al., 2024), where moral scenarios were directly presented in their native languages to assess the model's reasoning capabilities. Another strategy, Translation-based prompting (Liu et al., 2024) uses translation to convert multilingual inputs into a target 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, 2024b). This is also used to generate diverse CoT with Negative Rationales by incorporating both correct and incorrect reasoning paths to refine multilingual reasoning capabilities (Payoungkhamdee et al., 2024). While in-context learning with natural language can be ambiguous and less effective in low-resource languages, program-based demonstrations offer clearer, structured reasoning that transfers better across languages (Ranaldi et al., 2025a). In addition to the above strategies, Dictionary

Insertion Prompting (DIP) offers a lightweight and practical alternative by inserting English translations of keywords into non-English prompts, bridging linguistic gaps without full translation and enabling clearer reasoning and improved performance in multilingual tasks (Lu et al., 2024).

d) Model Editing. Model 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 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, 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).

An emerging complementary direction investigates inference-time (test-time) compute scaling in enhancing multilingual reasoning. Recent work shows that scaling up compute for English-centric reasoning language models (RLMs) can significantly improve performance across many languages, including low-resource ones, even surpassing larger models (Yong et al., 2025). While most test-time techniques, such as CoT prompting with trial and error, have primarily focused on English, methods like English-Pivoted CoT training (Tran et al., 2025) exploit the model's strong English reasoning capabilities to support multilingual tasks, offering a promising path to bridge alignment gaps for underrepresented languages.

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 detail key metrics for evaluating multilingual reasoning, along with their formal definitions:
1) Accuracy. 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. 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. To ensure multilingual reasoning consistency, models must preserve meaning across languages: i) *Translation Success Rate* 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. They quantify *semantic alignment* and measure whether embeddings across languages remain consistent: i) *Perplexity-Based Alignment (P_{align})*

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

where $P(x_i)$ is the model's probability of predicting token x_i (lower perplexity means better alignment) and ii) *Semantic Alignment* measures the cosine similarity between multilingual sentence embeddings: $S_{\text{align}} = \frac{E_l \cdot E_t}{\|E_l\|\|E_t\|}$, 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 (MGSM (Shi et al., 2022), MSVAMP (Chen et al., 2023)), commonsense (xCSQA (Lin et al., 2021)), logical (xNLI (Conneau et al., 2018)) reasoning benchmarks[1]. Next, we describe the four most popular benchmarks and detail the performance of reasoning techniques, highlighting existing model gaps that limit their reasoning performance.

[1] We only cover benchmarks analyzed by more than four papers.

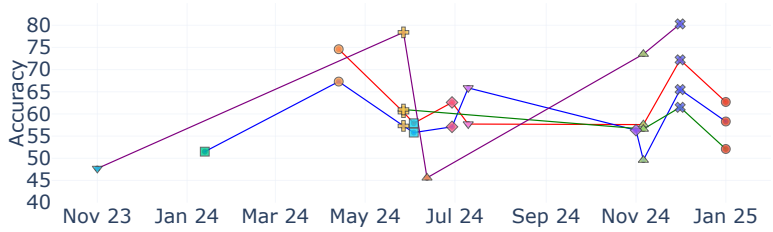

Figure 6: Accuracy trends of various methods on multilingual reasoning benchmarks, including MGSM, MSVAMP, XNLI, and XCSQA. The x-axis represents the arXiv paper submission date, and the y-axis indicates percentage accuracy.

MGSM tests multilingual arithmetic reasoning in LMs with 250 translated math problems in ten diverse languages. While recent trends suggest that advanced post-training techniques like MAPO are key for strong performance, fine-tuning strategies may be more impactful than stronger reasoning architectures or relying on the model's English expertise to improve multilingual performance.

MSVAMP is an out-of-domain multilingual mathematical reasoning dataset comprising 10k problems across ten languages and serves as a comprehensive test bed to evaluate LMs' generalization in multilingual mathematical contexts. We find that advanced preference optimization 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 LMs' cross-lingual commonsense reasoning capabilities. The current trend shows that stronger fine-tuning strategies like two-step fine-tuning or preference optimization show better performance than selectively fine-tuning specific layers as in SLAM.

xNLI evaluates cross-lingual inference across 15 languages. Recent studies suggest that LM integration with external models (Huang et al., 2024b) and multilingual alignment followed by fine-tuning (Zhang et al., 2024a) outperform contrastive learning methods like TCC (Chia et al., 2023), highlighting the need for more structured multilingual adaptation strategies.

6 Future Directions

With the rapid development of reasoning models, our community must ensure that models remain unbiased towards low-resource languages. Looking forward, we call on the community to put their collective efforts into the following 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 (Yoon et al., 2024; Huang et al., 2024b). Future research should develop cross-lingual knowledge transfer techniques, enabling models to use high-resource languages as a bridge to enhance reasoning 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 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 cross-lingual alignment, improving interpretability by uncovering cause-and-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. Some exciting techniques in this direction includes post-training RL methods that improve reasoning in low-resource languages (Wu et al., 2024) and efficient inference-time scaling and Agentic frameworks (Khanov et al., 2024; Chakraborty et al., 2024). Preliminary post-training works (Xuan et al., 2025) show that they yield mixed results across languages, with effectiveness depending on the base model and required degree of linguistic diversity, highlighting the need

for language inclusive training approaches.

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. Multimodal Multilingual Reasoning. While there are a few works on visual reasoning in the multilingual context (Das et al., 2024; Gao et al., 2025; Ghosh et al., 2024c), multimodal reasoning (integrating tables, text, image, audio, and video) remains largely unexplored. Advancing this area could enable models to handle complex tasks in low-resource languages and incorporate cross-modal reasoning.

6.New Benchmarks: As multilingual reasoning advances, robust evaluation benchmarks are essential because 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. For instance, Xue et al. (2024) introduces FAMMA which shows significant challenges in the field of Financial Question Answering.

7. Efficient Reasoning Models. An emerging direction in reasoning research is enhancing resource efficiency in reasoning-aware models. Recent works like (Ning et al., 2024) propose strategies for more efficient reasoning, reducing computational costs while maintaining logical consistency. However, this area remains largely unexplored in multilingual settings, offering a key opportunity to develop scalable reasoning models that generalize across languages with minimal resources.

8. Miscellaneous Tasks. LLMs have achieved remarkable performance across a wide range of tasks; however, they continue to struggle with complex compositional reasoning (Zhao and Zhang, 2024a), often performing only marginally better than random guessing. They also face difficulties in reasoning over longer contexts, particularly in low-resource languages (Hengle et al., 2025). Moreover, their reasoning traces frequently exhibit hallucinations (Sahoo et al., 2024), with models failing to reliably integrate information or recognize missing pieces even when the relevant facts are retrievable.

7 Conclusion

Multilingual reasoning in 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 and domain-specific applications. It serves as a call to action for the LLM and reasoning community to focus on advanced alignment techniques, culturally aware reasoning, and scalable architectures. By breaking language barriers and fostering inclusivity, multilingual reasoning can create globally impactful AI systems. Our survey provides a foundation for advancing research in this transformative domain.

8 Limitations

This is the first survey dedicated to the important and emerging topic of multilingual reasoning. We have made every effort to include key studies and recent advancements in this area; however, we acknowledge that some relevant work may have been unintentionally missed. As the field is still in its early stages, this survey does not aim to provide definitive solutions for improving multilingual reasoning. Instead, our goal is to analyze existing approaches and offer a comprehensive evaluation of which techniques demonstrate stronger performance across current benchmarks.

9 Acknowledgement

We would like to thank the anonymous reviewers for their insightful feedback. C.A. is supported, in part, by grants from Capital One, LaCross Institute for Ethical AI in Business, the UVA Environmental Institute, OpenAI Researcher Program, and Cohere. The views expressed are those of the authors and do not reflect the official policy or position of the funding agencies.

References

Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, and Russell Scheinberg. 2024. Evaluating multilingual long-context models for retrieval and reasoning. *arXiv preprint arXiv:2409.18006*.

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. 2024. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. *arXiv*.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL*.

Avinash Anand, Kritarth Prasad, Chhavi Kirtani, Ashwin R Nair, Manvendra Kumar Nema, Raj Jaiswal, and Rajiv Ratn Shah. 2024. Multilingual mathematical reasoning: Advancing open-source llms in hindi and english. *arXiv*.

Avinash Anand, Kritarth Prasad, Chhavi Kirtani, Ashwin R Nair, Manvendra Kumar Nema, Raj Jaiswal, and Rajiv Ratn Shah. 2025. Multilingual mathematical reasoning: Advancing open-source llms in hindi and english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23415–23423.

Gaurav Arora, Srujana Merugu, Shreya Jain, and Vaibhav Saxena. 2024. Towards robust knowledge representations in multilingual llms for equivalence and inheritance based consistent reasoning. *arXiv*.

Ashutosh Bajpai and Tanmoy Chakraborty. 2024. Multilingual llms inherently reward in-language time-sensitive semantic alignment for low-resource languages. *arXiv*.

Ashutosh Bajpai and Tanmoy Chakraborty. 2025. Multilingual llms inherently reward in-language time-sensitive semantic alignment for low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23469–23477.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and 1 others. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning, 2024. *URL https://arxiv.org/abs/2401*, 7037.

Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. 2024. Transfer q star: Principled decoding for llm alignment. *arXiv*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *arXiv*.

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

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

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv*.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. Slam: Towards efficient multilingual reasoning via selective language alignment. *arXiv*.

Junyuan Gao, Jiahe Song, Jiang Wu, Runchuan Zhu, Guanlin Shen, Shasha Wang, Xingjian Wei, Haote Yang, Songyang Zhang, Weijia Li, and 1 others. 2025. Pm4bench: A parallel multilingual multimodal multi-task benchmark for large vision language model. *arXiv preprint arXiv:2503.18484*.

Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, and 1 others. 2024. Why not transform chat large language models to non-english? *arXiv*.

Soumya Suvra Ghosal, Vaibhav Singh, Akash Ghosh, Soumyabrata Pal, Subhadip Baidya, Sriparna Saha, and Dinesh Manocha. 2025. Relic: Enhancing reward model generalization for low-resource indic languages with few-shot examples. *arXiv preprint arXiv:2506.16502*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024b. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024c. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024d.

Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. Infogen: Generating complex statistical infographics from documents. *arXiv preprint arXiv:2507.20046*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-200 evaluation benchmark for low-resource and multilingual machine translation. In *EMNLP*. ACL.

Srishti Gureja, Lester James V Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2024. M-rewardbench: Evaluating reward models in multilingual settings. *arXiv preprint arXiv:2410.15522*.

Huy Hoang Ha. 2025. Pensez: Less data, better reasoning–rethinking french llm. *arXiv preprint arXiv:2503.13661*.

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

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2025. Can llms reason over extended multilingual contexts? towards long-context evaluation beyond retrieval and haystacks. *arXiv preprint arXiv:2504.12845*.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, and 1 others. 2024a. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv*.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024b. Mindmerger: Efficient boosting llm reasoning in non-english languages. *arXiv*.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization. *arXiv preprint arXiv:2212.01669*.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2024. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. *arXiv*.

Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. Do moral judgment and reasoning capability of llms change with language? a study using the multilingual defining issues test. *arXiv*.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. Args: Alignment as reward-guided search. *arXiv*.

Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*.

Huiyuan Lai and Malvina Nissim. 2024. mcot: Multilingual instruction tuning for reasoning consistency in language models. *arXiv*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv*.

Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024a. Eliciting better multilingual structured reasoning from llms through code. *arXiv*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024b. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *NAACL*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024c. Quantifying multilingual performance of large language models across languages. *arXiv e-prints*, pages arXiv–2404.

Zheng Wei Lim, Nitish Gupta, Honglin Yu, and Trevor Cohn. 2024. Mufu: Multilingual fused learning for low-resource translation with llm. *arXiv*.

Yankai Lin, Jiapeng Zhou, Yiming Shen, Wenxuan Zhou, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Xcsqa: A benchmark for cross-lingual conversational question answering. In *EMNLP*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv*.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the impact of fine-tuning on chain-of-thought reasoning. *arXiv*.

Hongyuan Lu, Zixuan Li, and Wai Lam. 2024. Dictionary insertion prompting for multilingual reasoning on multilingual large language models. *arXiv preprint arXiv:2411.01141*.

Xianzhen Luo, Qingfu Zhu, Zhiming Zhang, Libo Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. 2024. Python is not always the best choice: Embracing multilingual program of thoughts. *arXiv preprint arXiv:2402.10691*.

Jiachen Lyu, Katharina Dost, Yun Sing Koh, and Jörg Wicker. 2024. Regional bias in monolingual english language models. *Machine Learning*.

Xuefei Ning, Zifu Wang, Shiyao Li, Zinan Lin, Peiran Yao, Tianyu Fu, Matthew B Blaschko, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Can llms learn by teaching for better reasoning? a preliminary study. *arXiv*.

Robert Östling and Jörg Tiedemann. 2016. Continuous multilinguality with language vectors. *arXiv*.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *arXiv*.

Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. An empirical study of multilingual reasoning distillation for question answering. In *Conference on Empirical Methods in Natural Language Processing*.

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

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv*.

Raghav Ramji and Keshav Ramji. 2024. Inductive linguistic reasoning with large language models. *arXiv*.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025a. When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7369–7396.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582.

Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. 2025b. Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations. *arXiv preprint arXiv:2504.04771*.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. 2024. One law, many languages: Benchmarking multilingual legal reasoning for judicial support.

Zhiwen Ruan, Yixia Li, He Zhu, Longyue Wang, Weihua Luo, Kaifu Zhang, Yun Chen, and Guanhua Chen. 2025. Layalign: Enhancing multilingual reasoning in large language models via layer-wise adaptive fusion and alignment strategy. *arXiv preprint arXiv:2502.11405*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Unveiling hallucination in text, image, video, and audio foundation models: A comprehensive review.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. *arXiv*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv*.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*.

Yueqi Song, Simran Khanuja, and Graham Neubig. 2024. What is missing in multilingual visual reasoning and how to fix it. *arXiv preprint arXiv:2403.01404*.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv*.

Jörg Tiedemann. 2012. Opus: An open source parallel corpus.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. 2025. Scaling test-time compute for low-resource languages: Multilingual reasoning in llms. *arXiv preprint arXiv:2504.02890*.

A Vaswani. 2017. Attention is all you need. *NeurIPS*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv*.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Demystifying multilingual chain-of-thought in process reward modeling. *arXiv preprint arXiv:2502.12663*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Mlake: Multilingual knowledge editing benchmark for large language models. *arXiv preprint arXiv:2404.04990*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv*.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot crosslingual alignment. *arXiv*.

Jiakuan Xie, Pengfei Cao, Yuheng Chen, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Memla: Enhancing multilingual knowledge editing with neuron-masked low-rank adaptation. *arXiv*.

Ruiyang Xu, Jialun Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, Ben He, Shing-Chi Cheung, and Le Sun. 2024. Cruxeval-x: A benchmark for multilingual code reasoning, understanding and execution. *arXiv*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, and 1 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2024. Famma: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint arXiv:2410.04526*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024. Language imbalance driven rewarding for multilingual self-improving. *arXiv preprint arXiv:2410.08964*.

Yahan Yang, Soham Dan, Shuo Li, Dan Roth, and Insup Lee. 2025. Mr. guard: Multilingual reasoning guardrail using curriculum learning. *arXiv preprint arXiv:2504.15241*.

Wenlin Yao, Haitao Mi, and Dong Yu. 2024. Hdflow: Enhancing llm complex problem-solving with hybrid thinking and dynamic workflows. *arXiv*.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

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

Fei Yuan, Yinquan Lu, WenHao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2022. Lego-mt: Learning detachable models for massively multilingual machine translation. *arXiv*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024a. Lingualift: An effective two-stage instruction tuning framework for low-resource language tasks. *arXiv*.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. 2024b. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *arXiv preprint arXiv:2411.09116*.

Jinman Zhao and Xueyan Zhang. 2024a. Exploring the limitations of large language models in compositional relation reasoning. *arXiv preprint arXiv:2403.02615*.

Jinman Zhao and Xueyan Zhang. 2024b. Large language model is not a (multilingual) compositional relation reasoner. In *First Conference on Language Modeling*.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, Deyi Xiong, and 1 others. 2024a. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024b. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *arXiv*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024c. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

A Appendix

Related Surveys The earliest surveys (Qin et al., 2024; Xu et al., 2025)—both from April 2024 focus on laying foundational taxonomies of Multilingual LLMs(MLLMs):(Qin et al., 2024) survey resources, taxonomy, and emerging frontiers in MLLMs, while (Xu et al., 2025) delve deeply into multilingual corpora, alignment techniques, and bias issues. Huang et al. (2024a) broadens the scope to multiple perspectives—training/inference, security, cultural domains, and datasets—framing "new frontiers" in multilingual LLM research. Finally, survey by (Zhu et al., 2024a) provides the most comprehensive "systematic" treatment: it covers architectures, pre-training objectives, alignment datasets, a detailed evaluation roadmap (including safety, interpretability, reasoning), and real-world applications across domains. *This survey is the first survey dedicated specifically to multilingual reasoning, drilling deeply into logical inference across languages, its unique challenges (misalignment, bias, low-resource gaps), and the benchmarks and methods tailored to evaluate and improve reasoning capabilities.*

Distribution of languages in Reasoning Datasets.

We show a detailed tabular format of the languages used in different reasoning datasets along with their languages.

af Afrikaans	ar Arabic	be Belarusian	bg Bulgarian
bn Bengali	ca Catalan	cs Czech	da Danish
de German	el Greek	en English	es Spanish
et Estonian	eu Basque	fa Persian	fi Finnish
fr French	ha Hausa	he Hebrew	hi Hindi
hr Croatian	ht Haitian	hu Hungarian	hy Armenian
id Indonesian	id Indonesian	is Icelandic	it Italian
ja Japanese	kn Kannada	ko Korean	lb Luxembourgish
mk Macedonian	ml Malayalam	mr Marathi	nb Norwegian Bokmal
ne Nepali	nl Dutch	pl Polish	pt Portuguese
qu Quechua	ro Romanian	ru Russian	sk Slovak
sl Slovenian	sr Serbian	sv Swedish	tr Turkish
uk Ukrainian	ur Urdu	vi Vietnamese	zh Chinese

Table 1: Language Codes and Their Corresponding Languages

Distribution of papers covering different aspects of Reasoning

Table 2: Multilingual Datasets and their respective papers, domains, and languages.

Dataset	Paper	Domain	Languages
MSVAMP	(She et al., 2024; Yoon et al., 2024; Zhu et al., 2024c,b; Lai and Nissim, 2024; Chai et al., 2024; Huang et al., 2024b; Zhang et al., 2024a; Fan et al., 2025)	Maths	zh, th, ja, en, de, fr, es, bn. sw
MGSM	(She et al., 2024; Yoon et al., 2024; Zhu et al., 2024c,b; Lai and Nissim, 2024; Chai et al., 2024; Huang et al., 2024b; Liu et al., 2024; Zhang et al., 2024a; Fan et al., 2025)	Maths	zh, th, ja, en, de, fr, es, ru, bn. sw, te
MNumGLUESub	(She et al., 2024)	Maths	bn, th, sw, ja, zh, ru, de, es, fr, en
MetaMathQA	(Yoon et al., 2024; Zhu et al., 2024c,b; Lai and Nissim, 2024; Huang et al., 2024b)	Maths	en
Proof-Pile 2	(Yoon et al., 2024)	Maths	en
Exams Dataset	(Payoungkhamdee et al., 2024)	Science and Humanities	ar, de, fr, es, it, pl, vi, pt, sr, hu, tr, bg, hr, mk, sq
M4U Benchmark	(Wang et al., 2024)	Science	zh, en, de
XCSQA	(Zhu et al., 2024b; Zhang et al., 2024a; Fan et al., 2025)	Common Sense	zh, en, de, fr, es, ru, hi
XNLI	(Zhu et al., 2024b; Liu et al., 2024; Zhang et al., 2024a)	Logical	zh, th, ur, en, de, fr, es, ru, el, tr, bg, hi, sw
MultiNLI	(Zhu et al., 2024b), (Huang et al., 2024b)	Logical	en
BBH-Hard	(Luo et al., 2024)	Temporal, Tabular, Spatial	Python, R, C++. Java, Javascript
NLVR2	(Song et al., 2024)	Visual	en
MARVL	(Song et al., 2024)	Visual	id, sw, ta, tr, zh
xSTREET	(Li et al., 2024a)	Logical	ar, zh, ja, en, es, ru
Translated Code Comments (TCC)	(Li et al., 2024a)	Code	Java, JavaScript, Python
mCoT-MATH	(Lai and Nissim, 2024)	Maths	zh, th, ja, en, de, fr, es, ru, bn, hi, te
Reasoning by Equivalence Dataset	(Arora et al., 2024)	Logical	en, fr, es, de, pt, hi
Reasoning by Inheritance Dataset	(Arora et al., 2024)	Logical	en, fr, es, de, pt, hi
XCOT	(Chai et al., 2024)	Maths	de, fr, es, ru, zh, ja, th, te, bn, sw, en
mCSQA	(Sakai et al., 2024)	Common Sense	zh, ja, en, fr, de, pt, ru

Dataset	Paper	Domain	Languages
Rulings, Legislation, Court View Generation, Critically Prediction, Law Area Prediction, Judgment Prediction Datasets	(Rasiah et al., 2024)	Legal	de, fr, it, ro, en
mRewardBench	(Gureja et al., 2024)	Logical and CommonSense	ar, cs, de, el, es, fa, fr, he, hi, id, it, ja, ko, nl, pl, pt, ro, ru, tr, uk, vi, zh
Moral Judgement Dataset	(Khandelwal et al., 2024)	Moral	en, zh, hi, ru, es, sw
MCR	(Zhao and Zhang, 2024b)	Compositional	ja, ko, fr
mTEMPREASON	(Bajpai and Chakraborty, 2025)	Temporal	ro, de, fr
XCOPA	(Liu et al., 2024)	Common Sense	zh, it, vi, tr, id, sw, th, et, ta, ht, qu
mARC	(Kargaran et al., 2024)	Common Sense	zh, ja, en, de, fr, es
IndiMathQA	(Anand et al., 2025)	Maths	en, hi
CRUXEval	(Xu et al., 2024)	Code	C#, C++, D, GO, Java, JavaScript, Julia, Luca, Perlm, PHP, R, Racket, Ruby, Rust, Scala, Shell, Swift, TypeScript

Dataset	Paper	Domain	Languages
mMMLU	(Kargaran et al., 2024)	Common Sense	ar, zh, vi, id, en, de, fr, it, nl, eu, es, pt, ca, da, ru, hr, hy, hu, ro, ne, kn, uk, sr, sv, mr, nb, ml, is, bn, hi, ta, te, gu
MMWP Benchmark	(Zhang et al., 2024a)	Maths	af, ar, be, bn, eu, gu, ha, hi, hy, is, kn, lb, mk, ml, mr, ne, sk, sw, ta, te, th, bg, ca, cs, da, fi, hr, hu, id, ko, nb, pl, pt, ro, sl, sr, uk, vi, de, en, es, fr, it, ja, nl, ru, sv, zh

Reasoning Type	Papers
Deductive	Lai and Nissim (2024), Chai et al. (2024), Huang et al. (2024b), Zhang et al. (2024a), Huang et al. (2024b), Fan et al. (2025), Payoungkhamdee et al. (2024), Luo et al. (2024), Song et al. (2024), Li et al. (2024a), Arora et al. (2024), Rasiah et al. (2024), Sakai et al. (2024), Khandelwal et al. (2024), Kargaran et al. (2024), Anand et al. (2025), Xu et al. (2024), She et al. (2024), Zhu et al. (2024b), Li et al. (2024c), Lim et al. (2024), Bajpai and Chakraborty (2025), Li et al. (2024b)
Inductive	Chai et al. (2024), Huang et al. (2024b), Zhang et al. (2024a), Huang et al. (2024b), Fan et al. (2025), Payoungkhamdee et al. (2024), Luo et al. (2024), Song et al. (2024), Li et al. (2024a), Arora et al. (2024), Rasiah et al. (2024), Sakai et al. (2024), Khandelwal et al. (2024), Kargaran et al. (2024), Anand et al. (2025), Xu et al. (2024), She et al. (2024), Zhu et al. (2024b), Li et al. (2024c), Lim et al. (2024), Bajpai and Chakraborty (2025), Li et al. (2024b), Wei et al. (2024), Xie et al. (2024), Yang et al. (2024), Geng et al. (2024), Yang et al. (2025), Ko et al. (2025), Ruan et al. (2025), Lu et al. (2024), Agrawal et al. (2024), Ranaldi et al. (2025b), Ha (2025), Ranaldi et al. (2025a), Ranaldi and Pucci (2025), Xuan et al. (2025), Yoon et al. (2024), Zhu et al. (2024c), Lai and Nissim (2024), Chai et al. (2024), Huang et al. (2024b), Zhang et al. (2024a), Huang et al. (2024b), Fan et al. (2025), Payoungkhamdee et al. (2024), Luo et al. (2024), Song et al. (2024), Li et al. (2024a), Arora et al. (2024), Rasiah et al. (2024), Sakai et al. (2024), Khandelwal et al. (2024), Kargaran et al. (2024), Anand et al. (2025), Xu et al. (2024), She et al. (2024), Zhu et al. (2024b), Li et al. (2024c), Lim et al. (2024), Bajpai and Chakraborty (2025), Li et al. (2024b), Wei et al. (2024), Xie et al. (2024), Yang et al. (2024), Geng et al. (2024), Yang et al. (2025), Ko et al. (2025), Ruan et al. (2025), Lu et al. (2024), Agrawal et al. (2024), Ranaldi et al. (2025b), Ha (2025), Ranaldi et al. (2025a), Ranaldi and Pucci (2025)
Abductive	Huang et al. (2024b), Zhang et al. (2024a)
Analogical	Zhang et al. (2024a), Huang et al. (2024b), Fan et al. (2025), Payoungkhamdee et al. (2024), Luo et al. (2024), Song et al. (2024), Li et al. (2024a), Arora et al. (2024), Rasiah et al. (2024), Sakai et al. (2024), Khandelwal et al. (2024), Kargaran et al. (2024), Anand et al. (2025), Xu et al. (2024), She et al. (2024), Zhu et al. (2024b), Li et al. (2024c), Lim et al. (2024), Bajpai and Chakraborty (2025), Li et al. (2024b), Wei et al. (2024), Xie et al. (2024), Yang et al. (2024), Geng et al. (2024), Yang et al. (2025), Ko et al. (2025), Ruan et al. (2025), Lu et al. (2024), Agrawal et al. (2024), Ranaldi et al. (2025b), Ha (2025), Ranaldi et al. (2025a), Ranaldi and Pucci (2025)
Commonsense	Huang et al. (2024b), Fan et al. (2025), Payoungkhamdee et al. (2024), Luo et al. (2024), Song et al. (2024), Li et al. (2024a), Arora et al. (2024), Rasiah et al. (2024), Sakai et al. (2024), Khandelwal et al. (2024), Kargaran et al. (2024), Anand et al. (2025), Xu et al. (2024), She et al. (2024), Zhu et al. (2024b), Li et al. (2024c), Lim et al. (2024), Bajpai and Chakraborty (2025), Li et al. (2024b), Wei et al. (2024), Xie et al. (2024)

Table 3: Categorization of Papers by Reasoning Type