Dynamic Multilingual Self-Reflection: Optimizing LLM Reasoning with Predictive Language Routing

Anonymous ACL submission

Abstract

This work introduces Iterative Multilingual Self-Reflection (IMSR), a framework that enhances Large Language Model (LLM) reasoning by resolving cross-lingual inconsistencies. IMSR utilizes an agentic cycle where models reconcile dual-language responses through structured debate and cross-lingual self-correction. To optimize performance, we employ a classification model to dynamically select the language pair most likely to yield successful reasoning correction for a given query. Extensive evaluations across various LLMs and benchmarks show that dynamic IMSR significantly outperforms standard self-reflection and majority voting. Ultimately, our method bridges the gap between high- and low-resource languages, synergizing their collective strengths to surpass single-language reasoning capabilities.

1 Introduction

The proliferation of Large Language Models (LLMs) has marked a significant milestone in artificial intelligence, demonstrating remarkable capabilities in complex reasoning. However, as these models are deployed globally, a critical challenge remains: the performance disparity across languages, primarily stemming from imbalanced pretraining data.

For instance, models like GPT-4 consistently exhibit stronger capabilities in high-resource languages compared to low-resource ones (OpenAI et al., 2024; Bang et al., 2023). Our empirical evaluation of GPT-4o mini confirmed this asymmetry; English accuracy exceeded Chinese by 6.25% on MMLU and 11.3% on CommonsenseQA (Hendrycks et al., 2021; Talmor et al., 2019). Such linguistic performance gaps limit model reliability in multilingual contexts and potentially disadvantage non-English users (Ramesh et al., 2023).

These results suggest that while an LLM may possess the underlying capability to solve a task, its reasoning can be misguided by incomplete linguistic understanding. This motivates our core research questions: can these performance gaps be mitigated, and can we exploit cross-lingual discrepancies to improve overall reasoning? Rather than merely achieving parity, our goal is to surpass the performance attainable in any single language by harnessing the collective reasoning strengths of multiple languages.

Existing zero-shot solutions, such as standard self-reflection or static ensemble methods, often fall short. Standard self-reflection frequently suffers from confirmation bias (Huang et al., 2024), while majority voting merely aggregates independent outputs without reconciling underlying logical contradictions. This highlights the need for a framework that leverages diverse linguistic perspectives to cross-examine and correct reasoning errors without requiring massive parameter modifications.

To this end, we propose Iterative Multilingual Self-Reflection (IMSR). Suppose a model is asked the same question in two languages, A and B. IMSR specifically targets cases where the model succeeds in one language but fails in another. Through a structured agentic cycle of self-reflection and multi-agent debate, the model is forced to reconcile these divergent reasoning paths and converge toward the more accurate answer.

Furthermore, since a static language pair may not yield optimal cognitive diversity for every query, we introduce a dynamic language pair routing mechanism. Powered by a classification model, this router predicts which language pair—among multiple combinations—has the highest probability of successful correction for a given query. This dynamic selection ensures the most effective linguistic synergy is deployed for each specific problem.

We evaluated IMSR across state-of-the-art models including GPT-4o mini, Gemini 2.5 Flash Lite, DeepSeek, and QWEN3. Extensive evaluations on reasoning benchmarks demonstrate that IMSR

statistically and significantly outperforms standard self-reflection and majority voting strategies.

In summary, our key contributions are:

- **Resolving Cross-Lingual Inconsistencies:** We introduce the IMSR framework, a zero-shot approach that effectively bridges linguistic performance gaps through cross-lingual debate.
- **Maximizing Efficacy via Dynamic Routing:** We develop a classification model to predict and select optimal language pairs, maximizing self-correction likelihood for any given query.
- **Robust and Accessible Performance Gains:** We empirically demonstrate that IMSR provides significant gains over traditional baselines as a lightweight, plug-and-play solution without the need for model fine-tuning.

2 Related Work

Our work builds on three areas: cross-lingual reasoning inconsistencies, multilingual enhancement methods, and multi-agent reflection techniques.

2.1 Cross-Lingual Reasoning Inconsistencies in LLMs

Studies show that LLMs like GPT-4 often produce inconsistent results across languages (Bang et al., 2023; Lai et al., 2023), mainly due to pre-training data imbalances. Rather than viewing this as a limitation, we treat cross-lingual variation as an opportunity to gain complementary insights.

2.2 Approaches to Multilingual Reasoning

Training-based methods such as multilingual fine-tuning can improve cross-lingual performance but are resource-intensive and unsuitable for proprietary models (Indurthi et al., 2024; Ahuja et al., 2025). Prompt-based strategies, e.g., translating problems into English, are lighter (Huang et al., 2023; Qin et al., 2023) but assume a dominant language. Our framework generates solutions in two languages simultaneously and reconciles them via structured dialogue, leveraging strengths from both.

2.3 Multi-Agent Debate, Self-Reflection, and Ensemble Methods

Self-reflection enables models to iteratively critique and refine their own outputs, while ensemble techniques like majority voting aggregate multiple independent responses to improve overall accuracy (Renze and Guven, 2024; Xiong et al., 2025; Wang et al., 2023). However, standard single-language self-reflection often suffers from a lack of inherent cognitive diversity, limiting its ability to break out of flawed reasoning loops. Similarly, majority voting relies on static aggregation of independent outputs rather than active reasoning refinement. We extend these paradigms by treating the LLM as two distinct linguistic agents that iteratively critique each other. This introduces genuine linguistic diversity, allowing our approach to overcome the limitations of standard self-reflection and majority voting by fostering robust, cross-lingual consensus.

2.4 Dynamic Routing and Language Selection

While static prompting strategies are common, a fixed pair of languages may not provide the optimal cognitive diversity for every query. Recent advancements in dynamic model routing demonstrate the efficacy of tailoring selection mechanisms to specific input characteristics. Building on this concept, we introduce a predictive routing mechanism using a classification model. By framing language pair selection as a classification task across multiple distinct language combinations, our router dynamically identifies the specific pair with the highest probability of successfully correcting an initially flawed response. This query-specific routing maximizes the potential of the IMSR framework, ensuring the most effective linguistic pairing is utilized for any given problem.

3 Methods

This section details our framework designed to enhance an LLM's reasoning robustness. The approach consists of two main components: the Iterative Multilingual Self-Reflection (IMSR) pipeline, which orchestrates a structured dialogue between two linguistic agents embodied by the same LLM, and a dynamic routing mechanism powered by a classification model that determines the optimal language pair for the dialogue. The IMSR pipeline itself is zero-shot and requires no modification of the LLM's weights, making it applicable to any pre-trained language model accessible via API.

3.1 Parallel Query Generation

The process begins with a query presented in an initial language. Based on the dynamic routing mechanism (detailed in Section 3.4), the framework selects a specific pair of languages (Language

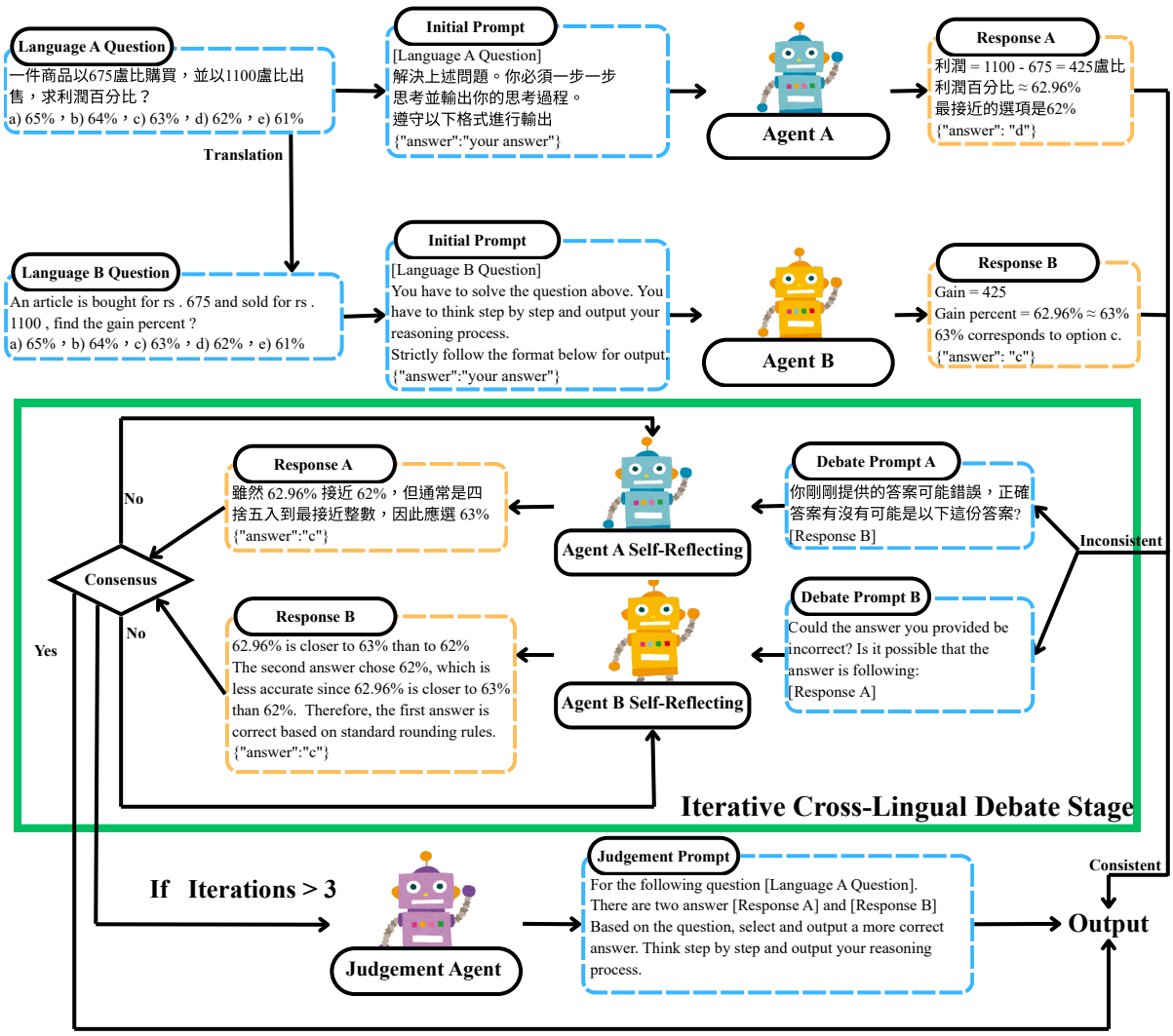

Figure 1: An overview of the Iterative Multilingual Self-Reflection (IMSR) framework, using a Chinese-English example.

A and Language B). To ensure high-quality semantic equivalence and avoid reasoning failures caused by poor translation, we utilize a highly capable LLM as a dedicated translator to translate the initial query into the target languages. The LLM is then prompted to act as two distinct agents (e.g., Agent A and Agent B) operating in parallel, each independently solving the same problem in their respective assigned language using chain-of-thought prompting. Both agents produce detailed reasoning paths and standardized final answers (e.g., in JSON format) in the two different languages. These outputs form the foundation for the following stages.

3.2 Iterative Cross-Lingual Debate

The core of our framework is an iterative debate and self-reflection process, which is triggered only if the initial answers from Agent A and Agent B are inconsistent. The goal of this stage is to force the model to confront and reconcile the divergent reasoning paths that arise from linguistic differences. The debate proceeds in rounds:

1. **Information Exchange:** Agent A is provided with the full reasoning and final answer from Agent B. Simultaneously, Agent B is provided with the output from Agent A.

2. **Challenging:** Each agent is then prompted to reconsider its initial conclusion. The prompt explicitly highlights the discrepancy and challenges the agent to re-evaluate its own logic in light of the alternative reasoning provided by its counterpart. The conversational history is maintained for both agents, allowing them to reference the full context of the debate.

3. **Consensus Checking:** After both agents gen-

erate new, revised responses, their answers are compared. If the answers now align, a consensus has been reached and the process terminates. If the answers remain inconsistent, the framework proceeds to another round of debate, repeating the information exchange and challenge steps.

3.3 Final Judgment Agent

To prevent infinite loops and manage computational cost, the debate is limited to a maximum of three rounds. If no consensus is achieved after the third debate round, an independent "Judgment Agent" is invoked to act as a tie-breaker. This agent is provided with the question alongside the final, conflicting answers and reasoning from both Agent A and Agent B. Using a simple chain-of-thought prompt, the Judgment Agent is instructed to analyze both solutions and select the one it determines to be correct. Its choice becomes the final output.

3.4 Dynamic Language Pair Routing

Relying on a static, pre-defined language pair may not provide the optimal cognitive diversity for every given query. To maximize the efficacy of the IMSR framework, we introduce a predictive routing mechanism using a fine-tuned, lightweight encoder model.

- **Language Combinations:** We select a diverse set of N target languages, resulting in a total of $K = \binom{N}{2}$ unique language pair combinations.

- **Data Preparation and Target Formulation:** To train the router, a single source query is translated into the N target languages, resulting in distinct input queries X for each language. Each linguistic variant is initially treated as an independent training instance. However, because these variants represent the exact same underlying problem, their corresponding target labels y are assigned identical values. To construct this shared label y, we execute the standard IMSR pipeline across all K candidate language pairs for the given problem. We specifically filter the dataset to include only problems where at least one language pair exhibits inconsistent initial answers. Furthermore, we strictly define a successful correction: for a specific language pair, exactly one language yields an initially incorrect answer, but after the IMSR debate process, the final consensus converges to the correct answer. The target y is formulated as a K-dimensional binary array, $y \in \{0, 1\}^K$. For each index representing a specific language pair, the value is set to 1 if it meets this strict correction criterion; otherwise, the value is set to 0. To prevent data leakage during the train/validation split, we group all linguistic variants of the same source query together, treating them as a single problem entity to ensure that identical conceptual questions do not appear in both splits.

- **Inference and Execution:** We fine-tune the lightweight encoder as a multi-label sequence classifier to predict this binary array. During the inference stage, a new query X is fed into the trained router. The model outputs a probability distribution indicating the likelihood of successful correction for each of the K language pairs. Our framework dynamically selects the pair with the highest predicted probability, translates the query into those two languages, and executes the IMSR process (Sections 3.1 - 3.3). This ensures that the most effective linguistic synergy is deployed for each specific problem.

4 Experiments and Results

This section details the experimental setup and presents the results of our evaluation. We designed our experiments to comprehensively assess the effectiveness of our dynamically routed Iterative Multilingual Self-Reflection (IMSR) framework. First, we evaluate the overall performance of IMSR by simultaneously comparing it against purely monolingual zero-shot responses across five diverse languages and established advanced reasoning baselines. Second, we conduct a comprehensive ablation study and routing analysis to validate the necessity of the predictive routing mechanism and to explore the theoretical upper bound of cross-lingual debate.

4.1 Experimental Setup

We evaluated our framework across four state-of-the-art LLMs: DeepSeek v3.2, GPT-4o mini, Gemini 2.5 Flash Lite, and QWEN3-8b. These models were selected to represent a diverse range of architectures. Our benchmark suite includes four complex reasoning datasets with definitive correct answers: MathQA, MMLU, TruthfulQA, and

CommonsenseQA. To ensure high-quality semantic equivalence during the parallel query generation phase, we consistently employed GPT-4.1 mini as the dedicated translator to translate the source queries into our target languages.

To rigorously train and evaluate the XLM-RoBERTa dynamic router, we performed the train-validation dataset split at the *source query* level. Because a single underlying conceptual question is translated into 5 target languages, we strictly grouped all linguistic variants of the same source query into either the training or the validation set. This query-level partitioning effectively prevents data leakage, ensuring that the router is evaluated on entirely unseen problems rather than merely unseen language combinations of familiar questions. The XLM-RoBERTa model was fine-tuned as a multi-label sequence classifier using a learning rate of 1e-5 for 5 epochs.

During inference, the router analyzes the query and selects the optimal language pair from 10 possible combinations (derived from English, Chinese, Japanese, Russian, and Spanish) to execute the IMSR pipeline. We compared our dynamic IMSR approach against several robust baselines:

- **Monolingual Baselines:** The standard zero-shot reasoning performance of the model evaluated independently in English (EN), Chinese (ZH), Japanese (JA), Russian (RU), and Spanish (ES).

- **5-Language Majority Voting (5-Lang MV):** A cross-lingual ensemble approach where the final output is determined by taking the majority vote across the five distinct responses generated by the respective Monolingual Baselines for the same underlying query.

- **Single-Language Self-Reflection (SR):** A prompt-based verbal reinforcement technique (Shinn et al., 2023) where the model is provided with the original query alongside its initial Monolingual Baseline answer. The model is then instructed to critically evaluate its own logic and generate a corrected final answer within the same language (evaluated in English and Chinese).

Performance is measured by the percentage of correctly answered questions (Accuracy %).

4.2 Main Results and Overall Performance

To assess the effectiveness of our dynamically routed IMSR framework, we evaluated its overall performance by simultaneously comparing it against purely monolingual zero-shot responses, 5-Language Majority Voting (5-Lang MV), and Single-Language Self-Reflection (SR). We define an evaluation scenario as a specific combination of a model and a dataset, yielding 16 distinct scenarios (4 models $\times$ 4 datasets).

As shown in Table 1, the dynamic IMSR framework consistently elevates the models' reasoning capabilities, outperforming both the monolingual baselines and the advanced reasoning strategies in the majority of test cases. Out of the 16 evaluated scenarios, IMSR achieved the highest overall accuracy in 12 scenarios. Statistical significance testing confirms that these performance gains over the strongest baseline in each respective scenario are robust ($p < 0.05$), denoted by an asterisk (*).

The consistent superiority of IMSR highlights a critical structural advantage. While single-language self-reflection often suffers from confirmation bias, IMSR forces the model to confront genuine cognitive diversity through contrasting logic from two different languages. Similarly, while 5-Lang MV relies on the passive, static aggregation of independent outputs, IMSR actively cross-examines, reconciles, and refines the underlying reasoning process.

Notably, the scenarios where IMSR did not surpass the best English baselines (EN-SR) occurred exclusively on CommonsenseQA and MMLU for DeepSeek v3.2 and GPT-4o mini. This behavior stems from the framework's architectural reliance on Chain-of-Thought (CoT) prompting, which is necessary to generate the detailed reasoning paths required for the cross-lingual debate stage. Recent studies have demonstrated that while CoT significantly boosts performance on math and complex symbolic reasoning, it can sometimes degrade performance on knowledge-intensive or short-form question-answering tasks like MMLU and CommonsenseQA (Wang et al., 2024; Sprague et al., 2025). Because these datasets often require direct factual recall rather than complex multi-step logic, forcing a lengthy CoT debate can occasionally introduce hallucinations or derail the reasoning, limiting the framework's ability to refine answers effectively in these specific contexts.

Model	Dataset	Monolingual Baselines					Advanced Methods			IMSR
		EN	ZH	JA	RU	ES	5-Lang MV	ZH-SR	EN-SR	
DeepSeek v3.2	CommonSenseQA	82.25	74.65	73.35	74.50	75.70	80.25	76.40	**83.15**	77.37
	MMLU	90.70	88.80	88.20	88.25	89.30	90.75	89.60	**91.20**	90.32
	MathQA	90.70	91.00	90.80	90.45	90.90	91.45	91.20	91.00	**92.45***
	TruthfulQA	84.58	87.88	87.27	85.56	85.19	88.24	88.01	85.98	**88.24**
GPT-4o mini	CommonSenseQA	79.10	67.80	69.15	70.55	72.65	75.70	70.55	**80.60**	78.14
	MMLU	81.00	74.75	75.10	75.60	77.35	79.45	78.00	**82.35**	80.47
	MathQA	82.35	79.85	76.60	81.15	80.80	84.30	81.20	84.00	**85.29***
	TruthfulQA	72.34	71.36	69.16	68.54	70.75	73.07	72.71	74.42	**76.40***
Gemini 2.5 Flash Lite	CommonSenseQA	73.85	71.10	71.00	71.30	73.55	75.50	71.55	73.95	**75.52***
	MMLU	81.65	82.25	81.35	82.00	82.60	84.90	83.40	86.15	**86.47***
	MathQA	79.80	83.20	83.75	84.20	84.85	85.10	84.89	81.30	**86.21***
	TruthfulQA	78.09	77.60	79.31	77.60	79.44	81.64	79.07	80.78	**83.81***
QWEN3-8b	CommonSenseQA	76.90	66.00	65.40	65.00	66.90	73.50	67.00	77.10	**78.26***
	MMLU	77.75	73.15	71.30	73.35	73.40	77.95	74.65	78.30	**78.81***
	MathQA	81.35	77.20	77.25	80.70	80.90	83.00	78.45	82.40	**83.09***
	TruthfulQA	74.42	75.52	68.30	71.36	73.19	77.85	77.11	75.52	**78.41***

Table 1: Accuracy comparison (%) of the IMSR framework against five monolingual baselines, 5-Language Majority Voting (5-Lang MV), and Self-Reflection (SR). Bold values indicate the highest performance for each dataset across all tested methods. Statistical significance ($p < 0.05$) compared to the strongest baseline in that scenario is denoted by an asterisk (*).

4.3 Ablation Study and Routing Analysis

To rigorously validate the necessity of the XLM-RoBERTa dynamic routing mechanism and assess the theoretical potential of the IMSR framework, we conducted a comprehensive ablation study. We compared our dynamic IMSR against three routing baselines across all 10 language pairs:

- **Random Language Pair:** For each query, a language pair is selected uniformly at random.
- **Best Fixed Pair (Ex-post):** The single static language pair that achieved the highest overall accuracy for a specific model and dataset. Note that this is a theoretical baseline, as identifying the "best" pair requires exhaustively evaluating all pairs ex-post.
- **Upper Bound (Oracle):** The theoretical maximum performance of the IMSR framework. Crucially, a query is considered correct if *any* of the 10 distinct language pairs successfully converges to the correct answer **after completing the full IMSR debate process**. To prevent any misunderstanding, we emphasize that this oracle evaluates the post-debate consensus of the language pairs, not simply whether any of the five independent monolingual baselines happened to generate the correct answer. It strictly measures the absolute potential of cross-lingual reasoning synergy, assuming a perfectly accurate predictive router.

As demonstrated in Table 2, the dynamic IMSR framework consistently outperforms the *Random Language Pair* baseline in 13 out of 16 evaluation scenarios. This confirms that the XLM-RoBERTa router successfully learns meaningful linguistic synergies rather than making arbitrary selections.

More importantly, the results highlight the fundamental flaw of relying on a static language pair. The *Best Fixed Pair* fluctuates drastically depending on the model and dataset. For instance, while English-Chinese (EN-ZH) is the optimal static pair for DeepSeek on MathQA, Japanese-Spanish (JA-ES) is the most effective for Gemini on TruthfulQA. Because the optimal linguistic cognitive diversity is highly context-dependent, dynamically predicting the best pair via IMSR provides a robust, automated solution that consistently yields near-optimal or state-of-the-art results without the need for exhaustive ex-post grid search.

Finally, the *Upper Bound (Oracle)* performance reveals the massive untapped potential of cross-lingual debate. The substantial gap between the current IMSR performance and the upper bound (often exceeding 8-10%) proves that the collective reasoning capability hidden across different linguistic representations is extremely high, paving the way for future research into even more advanced

routing architectures.

5 Discussion

Our results show that the Iterative Multilingual Self-Reflection (IMSR) framework effectively enhances LLM reasoning. By enabling a structured debate between two linguistic agents, it produces answers that are generally more reliable than their monolingual counterparts.

Furthermore, the statistically significant superiority ($p < 0.05$) of IMSR over resource-intensive baselines like Majority Voting and single-language Self-Reflection highlights a critical structural advantage of our approach: mitigating the confirmation bias prevalent in single-language reasoning. While standard self-reflection often traps a model within its initial flawed reasoning loops, and majority voting merely aggregates static outputs without active error correction, IMSR forces the model to confront genuine cognitive diversity.

This improvement comes from actively synthesizing divergent reasoning paths rather than merely selecting the most frequent response. Confronted with a conflicting cross-lingual perspective, the model is compelled to re-evaluate its own logic, identify underlying flaws, and integrate alternative insights.

Crucially, the introduction of the dynamic routing mechanism maximizes the efficacy of this framework. Our experiments demonstrate that relying on a static language pairing is suboptimal. The optimal linguistic diversity required to successfully trigger reasoning correction is highly query-dependent. By accurately predicting the specific language pair that yields the highest probability of successful self-correction for any given problem, the classification model ensures the IMSR framework operates at its full potential.

Ultimately, this dynamic approach not only resolves the widespread issue of reasoning inconsistencies across different languages but actively exploits these variations. By harnessing the collective reasoning strengths and diverse linguistic representations within the LLM, our framework transforms a known pre-training vulnerability into a robust, plug-and-play engine for reasoning enhancement.

5.1 Computational Cost and Test-Time Compute

A critical consideration when deploying multi-agent or reflection-based frameworks is the computational overhead. In this section, we analyze the training and inference costs of IMSR, framing it within the context of scaling test-time compute.

Training Cost Analysis: The primary upfront cost of our framework lies in data preparation rather than model training. To construct the target labels y for the XLM-RoBERTa router, the IMSR pipeline must be exhaustively executed across all 10 language pairs for every training instance. This data collection process incurs significant API token costs. However, the subsequent fine-tuning of the XLM-RoBERTa model itself is highly efficient. As a lightweight encoder-only model, training converges rapidly (e.g., 5 epochs) with negligible hardware requirements, making it a viable one-time investment.

Inference Cost and Adaptive Compute: During inference, total token consumption (both input context and generated output) directly impacts latency and monetary cost due to the autoregressive nature of LLMs. Advanced ensemble baselines like 5-Language Majority Voting (5-Lang MV) statically scale this cost by forcing the model to process and generate five independent reasoning paths for every single query, regardless of the problem's difficulty. Conversely, single-language Self-Reflection (SR) consumes less compute but often traps the model in confirmation bias, leading to suboptimal accuracy.

In contrast, IMSR operates as a highly efficient *adaptive compute* mechanism. The predictive routing via the classification model adds an $O(1)$ forward pass overhead, which is computationally trivial. More importantly, the total token consumption dynamically adapts to the complexity of the query. To empirically validate this, we tracked the average total tokens (input plus output) per query using the QWEN3-8b model across our benchmarks (see Table 3).

Because IMSR terminates inference early without triggering a multi-round debate when the two linguistic agents immediately reach a consensus, it bypasses unnecessary compute for straightforward queries. Consequently, the expected average token cost of IMSR across the entire dataset remains highly competitive. For example, on MathQA, while IMSR (1206 tokens) requires slightly more compute than a single-language English SR (1053 tokens) to establish the initial cross-lingual diversity, it utilizes nearly 38% fewer tokens than the brute-force 5-Lang MV approach (1950 tokens).

Ultimately, IMSR offers a superior cost-to-

Model	Dataset	Random Pair	Best Fixed Pair (Ex-post)	IMSR (Dynamic)	Upper Bound
DeepSeek v3.2	CommonSenseQA	76.39	**80.80** (EN-JA)	77.37	88.69
	MMLU	89.69	**91.45** (EN-JA)	90.32	94.28
	MathQA	92.03	91.55 (EN-ZH)	**92.45**	93.63
	TruthfulQA	88.00	**89.96** (ZH-JA)	88.24	93.88
GPT-4o mini	CommonSenseQA	74.11	77.80 (EN-ES)	**78.14**	88.28
	MMLU	78.46	**81.40** (EN-ZH)	80.47	89.07
	MathQA	83.90	84.55 (EN-RU)	**85.29**	92.72
	TruthfulQA	74.53	73.56 (EN-ZH)	**76.40**	83.04
Gemini 2.5	CommonSenseQA	75.51	**78.20** (EN-JA)	75.52	88.39
	MMLU	85.61	86.45 (EN-ZH)	**86.47**	92.33
	MathQA	86.35	**87.40** (JA-ES)	86.21	90.27
	TruthfulQA	81.26	83.35 (JA-ES)	**83.81**	90.59
QWEN3-8b	CommonSenseQA	71.75	75.85 (EN-JA)	**78.26**	90.77
	MMLU	76.95	**79.15** (EN-ZH)	78.81	90.35
	MathQA	82.88	**84.40** (RU-ES)	82.69	92.24
	TruthfulQA	76.27	78.34 (EN-ZH)	**78.41**	89.31

Table 2: Ablation study on routing mechanisms. The *Best Fixed Pair* denotes the ex-post optimal static combination for that specific scenario (indicated in parentheses). Bold values indicate the best practical performance (excluding the theoretical upper bound).

performance ratio. It avoids the massive, static token consumption of large ensembles by strategically allocating extra compute *only* when cross-lingual divergence indicates a potential reasoning flaw, making it a highly practical solution for dynamically scaling test-time compute.

Dataset	5-Lang MV	Self-Reflection		IMSR
		ZH	EN	
MathQA	1950	957	1053	1206
TruthfulQA	1965	834	864	1115
C-SenseQA	1265	594	632	992
MMLU	1740	790	828	1079

Table 3: Empirical inference cost (average total tokens—input plus output—per query) for QWEN3-8b. IMSR operates as an adaptive compute mechanism. Its expected token cost is significantly lower than the brute-force 5-Language Majority Voting ensemble, yet it achieves superior accuracy by strategically investing compute only when cross-lingual inconsistency is detected.

6 Conclusion

We introduced Iterative Multilingual Self-Reflection (IMSR), a zero-shot, prompt-based framework designed to resolve reasoning inconsistencies across different languages by using structured cross-lingual debates to reconcile divergent logic and generate more robust answers. To maximize the efficacy of this framework, we developed a dynamic routing mechanism using an XLM-RoBERTa model. By formulating language selection as a predictive task, our router successfully identifies the optimal language pair—out of 10 combinations derived from 5 diverse languages—that yields the highest probability of self-correction for any given query.

Extensive evaluations across multiple state-of-the-art LLMs demonstrate that our dynamically routed IMSR approach not only consistently surpasses monolingual baselines but also achieves statistically significant improvements ($p < 0.05$) over advanced, resource-intensive strategies such as Majority Voting and single-language Self-Reflection.

Ultimately, this approach provides a practical, scalable method to turn an LLM's inherent cross-lingual variations into a powerful engine for reasoning enhancement without the need to modify model parameters. Future work could explore extending this predictive routing mechanism to an even broader spectrum of linguistic families or integrating it into more complex, multi-step agentic workflows.

Limitations

While our Iterative Multilingual Self-Reflection (IMSR) framework paired with dynamic routing demonstrates significant improvements in reasoning, it has several limitations.

First, the framework's effectiveness is strictly constrained by the model's underlying knowledge. It cannot generate a correct answer if the model is fundamentally incorrect about a topic in both languages. If both Agent A and Agent B produce

incorrect initial answers, it is highly unlikely that a debate between them will spontaneously lead to a correct conclusion.

Second, the method introduces considerable computational overhead and increased latency. Compared to a single-prompt approach, our framework requires a preliminary inference step through the XLM-RoBERTa router, followed by a minimum of two parallel LLM calls. If the agents disagree and a debate is initiated, the number of API calls increases with each round, making it significantly more resource-intensive than standard inference.

Third, the debate process does not guarantee convergence. While the framework limits the debate to a maximum of three rounds to prevent infinite loops, there remains a possibility that the agents fail to reach a consensus, necessitating intervention by a final Judgment Agent, which adds a further layer of heuristic decision-making.

Finally, while the XLM-RoBERTa dynamic router maximizes the efficacy of IMSR, it introduces a dependency on specialized, model-specific training data. Constructing the target labels ($y \in \{0, 1\}^{10}$) requires executing the full IMSR pipeline across all 10 language pairs for every training query to observe which pair succeeds. This data collection process is highly expensive. Furthermore, the router's predictions are inherently tied to the specific performance characteristics of the LLM used during data generation. If the underlying LLM is updated (e.g., from GPT-4o mini to a future iteration) or replaced, the router's accuracy may degrade, necessitating costly data re-collection and model fine-tuning to maintain optimal routing performance.

Acknowledgments

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. 2025. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. *Preprint*, arXiv:2407.09879.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *Preprint*, arXiv:2305.07004.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *Preprint*, arXiv:2310.01798.

Sathish Reddy Indurthi, Wenxuan Zhou, Shamil Chollampatt, Ravi Agrawal, Kaiqiang Song, Lingxiao Zhao, and Chenguang Zhu. 2024. Improving multilingual instruction finetuning via linguistically natural and diverse datasets. *Preprint*, arXiv:2407.01853.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *Preprint*, arXiv:2304.05613.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *Preprint*, arXiv:2310.14799.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond english: Gaps and challenges. *Preprint*, arXiv:2302.12578.

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, page 476–483. IEEE.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg

Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *Preprint*, arXiv:2409.12183.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Dai. 2023. Self-consistency improves chain of thought reasoning in language models. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. 2025. Self-rewarding correction for mathematical reasoning. *Preprint*, arXiv:2502.19613.