Enhancing Multilingual Reasoning via Steerable Model Merging

Zhuoran Li[1], **Rui Xu**[2], **Jian Yang**[3], **Junnan Liu**[4], **Zhijun Chen**[3], **Qianren Mao**[5]
Hongcheng Guo[2], **Jiaheng Liu**[6], **Likang Xiao**[3], **Ming Li**[7], **Xiaojie Wang**[1*]
[1]Beijing University of Posts and Telecommunications, [2]Fudan University
[3] Beihang University, [4] Monash University, [5]Zhongguancun Laboratory
[6]Nanjing University, [7]Tsinghua University

Abstract

Model merging is an effective technique for composing the capabilities of a multilingual model and a reasoning model. It has achieved promising generalization in multilingual reasoning tasks by aligning feature spaces of different models. However, the merged single model often fails to address the conflicts between source models, leading to suboptimal performance. In other words, the one-size-fits-all merging strategy may not align with the characteristics of different inputs which may require prioritizing certain models over others. To this end, we propose a Steerable Model Merging (**ST-Merge**) framework to modulate the contribution of each source model. To realize this idea, we introduce a gated cross-attention mechanism to weight or filter the two attended source models in an adaptive manner. Extensive experiments demonstrate that ST-Merge consistently outperforms multiple strong baselines on four multilingual reasoning benchmarks across 21 different languages.

1 Introduction

Multilingual reasoning aims to empower Large Language Models (LLMs) to perform complex reasoning tasks across diverse languages. This capability is valuable in circumstances where limited or no annotations are available for low-resource languages. In recent years, reasoning large language models, such as MetaMath (Yu et al., 2024) and Orca (Mitra et al., 2023), have achieved significant performance improvements through parameter-efficient fine-tuning on source language data and direct application to target language data (as shown in Figure 1(a)).

Furthermore, it has been discovered that additional multilingual representation alignment improves the low-resource language reasoning performance by composing an external multilingual encoder to replace or augment the original LLM query embedding. This strategy, named Model Merging, has demonstrated improvements in many multilingual reasoning tasks (Yoon et al., 2024; Huang et al., 2024) (as shown in Figure 1(b)).

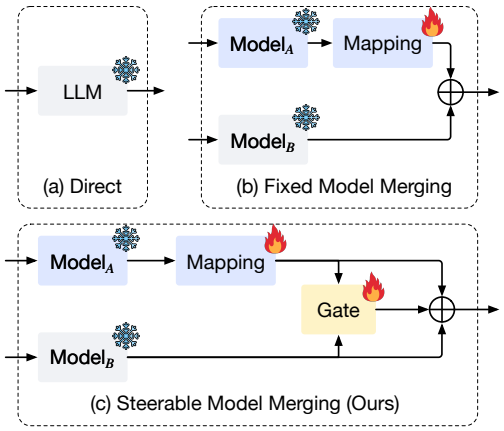


Figure 1: Illustration of our ST-Merge idea. (a) Direct application of LLM to all languages. (b) One-size-fits-all model merging method. (c) The proposed steerable model merging method learns to modulate the contribution of each source model for different inputs. LLM/Model$_B$: the reasoning LLM. Model$_A$: the external multilingual encoder.

However, as a one-size-fits-all strategy, current fixed model merging approaches struggle to strike an optimal balance for inputs across diverse languages. On one hand, over-reliance on the external multilingual encoder can dilute the core reasoning capabilities inherent to the original LLM, potentially leading to *catastrophic forgetting*. On the other hand, insufficient reliance on the external encoder hampers the understanding of low-resource languages, thereby limiting reasoning performance. Existing studies have observed that fixed model merging often causes a degradation in reasoning capabilities for languages in which the LLM is already proficient (Yoon et al., 2024). Therefore, it is imperative to devise an adaptive scheme to mod-

* Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 37266–37277

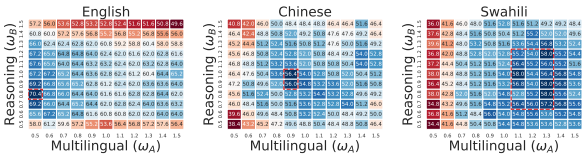


Figure 2: Accuracy on MGSM with different manual weight combinations for the two source models. Darker blue grids indicate higher reasoning accuracy.

ulate the models based on the characteristics of inputs.

In this paper, we first conduct an exploratory analysis by introducing manual scalar weights, ω_A and ω_B, to modulate the representation intensity of the multilingual encoder and the reasoning LLM, respectively. As illustrated in Figure 2, the heatmap of accuracy exhibits divergent collaboration patterns across languages. For English, optimal performance is achieved when the weight for the multilingual encoder (ω_A) is relatively low. This suggests that for languages where the base LLM is already proficient, over-reliance on external multilingual signals may act as noise interfering with the inherent reasoning pathways of LLM. Conversely, for Swahili, performance peaks only when the multilingual encoder contributes significantly (high ω_A). Here, the base LLM lacks the necessary linguistic grounding, and amplifying the external representation is crucial to bridge the semantic gap and activate the reasoning capabilities.

To pursue both effective utilization of low-resource language understanding and the preservation of inherent reasoning abilities, we propose a Steerable Model Merging framework (**ST-Merge**) with gated cross-attention. Instead of relying on a fixed concatenation of representations, our method enables the model to dynamically modulate the contribution of each source model (i.e., the multilingual encoder and the reasoning LLM), allowing for more flexible and adaptive coordination. This design facilitates input-aware modulation modeling, enabling the merged model to shift its inductive bias toward the source most aligned with the current input. As a result, it yields more accurate and targeted reasoning across diverse linguistic contexts. The main contributions of this paper are as follows:

- We propose a steerable model merging (**ST-Merge**) framework that modulates multilingual understanding and reasoning preservation for multilingual reasoning.
- We devise a gated cross-attention mechanism to dynamically modulate the contributions of source models.
- Extensive experiments on four multilingual reasoning benchmarks across 21 languages demonstrate that ST-Merge consistently outperforms strong baselines.

2 Related Work

2.1 Multilingual Reasoning

Enhancing multilingual reasoning in English-centric LLMs remains a critical challenge. Existing approaches can be broadly categorized into translation-based methods and model merging paradigms. Translation-based strategies, which involve fine-tuning on translated datasets (Zhu et al., 2024) or employing external translators (Shi et al., 2023), have demonstrated significant performance gains. However, these methods incur substantial computational overheads due to the heavy reliance on high-quality parallel corpora and the computational latency of autoregressive decoding. Conversely, model merging has recently emerged as a popular alternative (Yoon et al., 2024; Huang et al., 2024), aiming to combine the strengths of different experts. Despite their promise, current merging techniques typically employ static fusion strategies, often overlooking the inherent feature conflicts and interference between the multilingual encoder and the reasoning LLM. To address this, we propose ST-Merge, which introduces a dynamic gated network. This approach allows for the flexible, context-aware merging of models, effectively resolving inter-model conflicts while maximizing the collaboration between multilingual understanding and logical reasoning.

2.2 Model Merging

Model merging aims to combine the strengths of multiple models into a unified architecture and has been widely used to enhance capabilities such as modality integration (Sung et al., 2023; Chen et al., 2024a) and task generalization (Bandarkar et al., 2025; Du et al., 2025). Existing works can be broadly categorized into two types: homogeneous merging, which combines models with the same architecture, and heterogeneous merging, which merges models across architectural or modality boundaries. Recent studies have explored model merging for cross-lingual transfer learning (Yoon

et al., 2024; Huang et al., 2024), but often suffer from limited controllability and alignment issues in multilingual reasoning settings. In contrast, our work introduces a steerable model merging approach that dynamically modulates the representations, enabling better coordination between multilingual encoder and the reasoning LLM for reasoning across both low-resource and high-resource languages.

2.3 Gated Attention Mechanism

Gated cross-attention mechanisms have been developed to selectively fuse heterogeneous representations by leveraging learnable weights (Chaplot et al., 2018; Lee et al., 2022). These approaches employ multiplicative or residual gating strategies to dynamically weight features, effectively filtering noise and enhancing interpretability in fusion tasks (Kim and Shin, 2021; Ortiz-Perez et al., 2025). Recent studies have extended this paradigm to various architectures, including router-based gating for audio-visual recognition and sparse gating for large language models (Jeong et al., 2025; Qiu et al., 2025). However, while these mechanisms have proven effective in multimodal settings, their potential for steering cross-lingual alignment within a model merging framework remains unexplored.

3 Steerable Model Merging

In this section, we introduce our Steerable Model Merging (**ST-Merge**) method, which is designed to weight and filter attended models conditioned on the specific input question. Figure 3 depicts an overview of the ST-Merge framework. We will provide a detailed description of our approach from the following two stages: feature space alignment and gated cross-attention learning.

Problem Formulation Multilingual reasoning can be formulated as a text generation task. Given an input sequence $\mathbf{x}$ (e.g. a math problem), the model aims to generate the target output sequence $\mathbf{y}$ (e.g. a chain-of-thought and the answer). Formally, the language modeling likelihood of the target output is denoted as:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i}^{L} p(y_i|\mathbf{x}, y_{<i}) \quad (1)$$

Under the paradigm of model merging, we assume there are a multilingual encoder $\mathbf{m}_A$ and a reasoning LLM $\mathbf{m}_B$. Our goal is to learn a merger

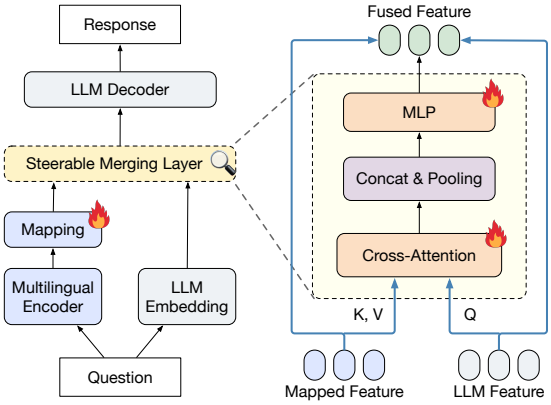


Figure 3: Framework of the proposed steerable model merging **ST-Merge** method for multilingual reasoning.

$\mathbf{m}_{A\oplus B}$ that optimizes the generation probability $p(\mathbf{y}|\mathbf{x})$, ensuring the reasoning accuracy is maintained across different languages.

3.1 Feature Space Alignment

First, we extract distinct features for each input sequence. Specifically, we utilize a multilingual encoder to capture linguistic understanding features and a Large Language Model (LLM) to extract reasoning features.

Multilingual Feature Multilingual feature extraction is performed by the mT5 encoder (Xue et al., 2021). Given an input sequence $\mathbf{x}$, we employ the multilingual model to encode it into a generalized representation $\mathbf{H}_A$, thereby mitigating the complexity of cross-lingual understanding:

$$\mathbf{H}_A = \text{Encoder}(\mathbf{x}) \quad (2)$$

where $\mathbf{H}_A \in \mathbb{R}^{l_A \times d_A}$ is the hidden state output of the last layer in the multilingual encoder, with l_A denoting the sequence length of the original input and d_A the hidden dimension.

Reasoning Feature We utilize LLaMA-based parameter-efficient fine-tuned reasoning LLM (e.g. MetaMath (Yu et al., 2024)) to extract reasoning features. To fully activate the intrinsic reasoning capabilities of the LLM, we append a chain-of-thought prompt $\mathbf{p}$ (e.g., "Let's think step by step") to the original input $\mathbf{x}$, forming a prompted sequence $\mathbf{x}' = [\mathbf{x}; \mathbf{p}]$. We then process this sequence directly through the embedding layer:

$$\mathbf{H}_B = \text{Embedding}(\mathbf{x}') \quad (3)$$

where $\mathbf{H}_B \in \mathbb{R}^{l_B \times d_B}$ represents the representation within the semantic space of the reasoning LLM,

with l_B denoting the sequence length of the original input with the prompt and d_B the LLM embedding dimension.

Feature Alignment Since the representation $\mathbf{H}_A$ resides in the multilingual representation space, which is separate from the reasoning LLM space. The extracted features cannot be used for reasoning directly. Therefore, we project the multilingual features via a mapping layer:

$$\hat{\mathbf{H}}_A = \text{Mapping}(\mathbf{H}_A) \tag{4}$$

where $\hat{\mathbf{H}}_A \in \mathbb{R}^{l_A \times d_B}$ is the projection of $\mathbf{H}_A$ onto the reasoning feature space. Unless otherwise stated, Mapping$(\cdot)$ is implemented as a two-layer Multi-Layer Perceptron (MLP). This transformation aligns the semantic spaces of the multilingual encoder and the reasoning LLM, enabling effective feature merging despite the frozen weights of the base models.

3.2 Gated Cross-Attention Learning

To overcome the limitations of fixed model merging, we introduce a gated cross-attention network to dynamically estimate the optimal weights for the multilingual features $\hat{\mathbf{H}}_A$ and the reasoning features $\mathbf{H}_B$ conditioned on the specific input.

Cross-Attention We first facilitate a comprehensive information interaction between the two types of features to construct a holistic context for gating estimation. Formally, the reasoning feature $\mathbf{H}_B$ acts as the query ($\mathbf{Q}_B$) to attend to the aligned multilingual representation $\hat{\mathbf{H}}_A$, which serves as the key ($\mathbf{K}_A$) and value ($\mathbf{V}_A$). We employ $\mathbf{H}_B$ as the query to ensure the attention mechanism is anchored in the semantic space of the reasoning task.

$$\mathbf{K}_A, \mathbf{V}_A = \hat{\mathbf{H}}_A \mathbf{W}_k^K,\ \hat{\mathbf{H}}_A \mathbf{W}_k^V \tag{5}$$

$$\mathbf{Q}_B = \mathbf{H}_B \mathbf{W}_k^Q \tag{6}$$

$$\text{head}_k = \text{Attn.}(\mathbf{Q}_B, \mathbf{K}_A, \mathbf{V}_A) \tag{7}$$

$$\mathbf{G}_{A\oplus B} = \text{Concat.}_k(\text{head}_k)\mathbf{W}^O \tag{8}$$

where $\mathbf{W}_k^Q, \mathbf{W}_k^K, \mathbf{W}_k^V \in \mathbb{R}^{d_B \times d_k}$ denote the projection matrices for the k-th head, and d_k is the dimension of each attention head. $\mathbf{W}^O \in \mathbb{R}^{d_B \times d_B}$ is the output projection matrix used to aggregate information from all k heads.

Language Embedding We introduce a learnable lightweight language embedding to explicitly inject language identity, which facilitates language differentiation. Given the language ID, we retrieve the corresponding embedding vector $\mathbf{E}_{Lang} \in \mathbb{R}^{d_L}$ and concatenate it with the global context $\mathbf{G}_{A\oplus B} \in \mathbb{R}^{d_B}$, yielding a composite representation $\mathbf{Z} \in \mathbb{R}^{d_B + d_L}$.

Feature Fusion Finally, we employ an MLP layer to project $\mathbf{Z}$ to generate the two weights to modulate the $\hat{\mathbf{H}}_A$ and $\mathbf{H}_B$. We specifically utilize a $1 + \tanh$ activation function to center the weights around 1, as the value of 1 indicates the fixed model merging with original features providing a stable initialization. The final input to the LLM decoder is constructed as follows:

$$[\omega_A, \omega_B] = 1 + \tanh(\text{MLP}(\mathbf{Z})) \tag{9}$$

$$\mathbf{H}_{A\oplus B} = [\langle\text{bos}\rangle; \omega_A \cdot \hat{\mathbf{H}}_A; \langle\text{sep}\rangle; \omega_B \cdot \mathbf{H}_B] \tag{10}$$

where ω_A, ω_B are two input-dependent scalar weights for the multilingual and reasoning representations, respectively; $\langle\text{bos}\rangle$ and $\langle\text{sep}\rangle$ are learnable boundary tokens. The resulting fused embedding $\mathbf{H}_{A\oplus B}$ serves as the steered input to the frozen LLM, guiding the generation of the chain-of-thought reasoning path and the final response.

4 Experiments

4.1 Evaluation Datasets

We evaluate models on four multilingual reasoning datasets across 21 different languages:

Mathematical Reasoning We evaluate on the multilingual math problem datasets MGSM and MSVAMP for this task. **MGSM** (Shi et al., 2023) consists of grade-school level math questions translated by humans into 11 typologically diverse languages. **MSVAMP** (Chen et al., 2024b) extends the SVAMP dataset (Patel et al., 2021) to 10 languages, offering linguistically diverse paraphrases of math problems with varying reasoning structures.

Commonsense Reasoning We evaluate commonsense reasoning using **X-CSQA** (Lin et al., 2021), a multilingual extension of the CommonsenseQA dataset. X-CSQA provides translated versions of CSQA across multiple languages, along with a new data split to support cross-lingual evaluation. The dataset includes 8,888 English training examples, 1,000 development examples per language, and 1,074 test examples per language.

MGSM	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Lrl.	Hrl.	Avg.
Translate-En [2023]	48.4	37.6	37.6	49.2	46.8	60.4	56.4	47.6	59.6	65.5	41.2	55.1	50.6
MetaMath [2024]	6.8	7.2	6.8	36.4	38.4	55.2	54.4	52.0	57.2	**68.8**	6.9	51.8	38.3
MultiReason [2024]	33.2	40.0	42.0	42.0	42.0	45.2	44.8	45.2	48.0	52.0	38.4	45.6	43.4
QAlign [2024]	32.4	39.6	40.4	44.0	48.4	54.8	56.8	52.4	59.6	68.0	37.5	54.9	49.6
LangBridge [2024]	42.8	50.4	43.2	40.0	45.2	56.4	50.8	52.4	58.0	63.2	45.5	52.3	50.2
MindMerger [2024]	50.4	52.8	57.2	**54.4**	53.6	61.2	57.6	60.8	58.4	66.8	53.5	59.0	57.3
LayAlign [2025]	51.6	**59.2**	58.4	52.0	56.0	62.0	**61.6**	61.6	61.6	66.4	56.4	60.2	59.0
ST-Merge (Ours)	**54.0**	56.8	**58.8**	53.5	**57.2**	**62.4**	61.2	**62.8**	**65.2**	68.0	**56.5**	**61.5**	**60.0**

MSVAMP	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Lrl.	Hrl.	Avg.
Translate-En [2023]	47.9	51.3	43.1	50.4	55.8	43.9	50.9	53.4	51.4	60.6	47.4	52.3	50.9
MetaMath [2024]	14.4	19.5	16.8	53.4	55.0	63.5	64.1	60.3	64.9	66.3	16.9	61.1	47.8
MultiReason [2024]	34.8	38.1	39.8	43.4	42.9	45.6	45.8	45.0	46.1	46.8	37.6	45.1	42.8
QAlign [2024]	41.7	47.7	54.8	58.0	55.7	62.8	63.2	61.1	63.3	65.3	48.1	61.3	57.2
LangBridge [2024]	46.8	46.3	42.1	45.5	50.4	58.1	57.0	55.8	56.9	60.6	45.1	54.9	52.0
MindMerger [2024]	52.0	53.4	54.0	59.0	**61.7**	**64.1**	64.0	**63.3**	65.0	**67.7**	53.1	63.5	60.4
LayAlign [2025]	51.8	55.1	56.9	59.3	58.7	62.5	62.1	58.8	62.0	64.0	54.6	61.1	59.1
ST-Merge (Ours)	**52.8**	**56.3**	**57.6**	**59.7**	**61.7**	63.8	**65.7**	62.2	**66.1**	67.6	**55.6**	**63.9**	**61.4**

Table 1: Accuracy (%) results on MGSM and MSVAMP. We regard Bn, Th, and Sw as low-resource languages, and regard the remaining languages as high-resource languages. Lrl., Hrl., and Avg. represent the average accuracy across low-resource languages, high-resource languages, and all languages, respectively. The best performance is in bold (same for Table 2 and Table 3).

X-CSQA	Sw	Fr	En	Avg.
Translate-En [2023]	36.5	57.2	71.3	52.3
MetaMath [2024]	24.2	63.5	76.3	51.3
MultiReason [2024]	27.6	52.1	67.2	43.8
QAlign [2024]	35.1	60.3	75.7	52.3
LangBridge [2024]	31.8	38.2	44.4	36.1
MindMerger [2024]	45.5	**68.1**	**78.1**	61.0
LayAlign [2025]	53.3	66.5	76.7	62.3
ST-Merge (Ours)	**53.6**	67.2	77.3	**62.5**

Table 2: Accuracy (%) on X-CSQA. Avg. represents the average accuracy across all languages.

XNLI	Sw	Fr	En	Avg.
Translate-En [2023]	65.3	80.4	81.4	75.1
MetaMath [2024]	45.9	82.2	**90.0**	68.7
MultiReason [2024]	56.3	82.9	88.8	71.9
QAlign [2024]	65.2	83.1	89.1	73.5
LangBridge [2024]	71.7	79.9	83.4	76.5
MindMerger [2024]	66.6	83.9	88.7	78.4
LayAlign [2025]	73.0	84.7	88.9	79.7
ST-Merge (Ours)	**73.7**	**84.8**	89.1	**79.9**

Table 3: Accuracy (%) on XNLI. Avg. represents the average accuracy across all languages.

Natural Language Inference We evaluate natural language inference using **XNLI** (Conneau et al., 2018), a widely used multilingual benchmark spanning 15 languages. The task involves determining whether a given *hypothesis* logically follows from a *premise*, categorized as entailment, contradiction, or neutral. The dataset covers languages both typologically close to English (e.g., French, German, Spanish) and more distant (e.g., Arabic, Thai, Swahili), making it well-suited for evaluating cross-lingual generalization.

4.2 Implementation Details

Following prior setup (Huang et al., 2024; Ruan et al., 2025), we train the mapping layer using the Lego-MT corpus (Yuan et al., 2023) via translation tasks. Subsequently, we leverage the MultilingualMath dataset (Yu et al., 2024; Chen et al., 2024b) for the gated cross-attention network learning. We adopt the encoder of mT5-xl (Xue et al., 2021) as the multilingual backbone, and employ MetaMath (Yu et al., 2024) as the large language reasoning model across all experiments, ensuring a fair comparison with prior work (Yoon et al., 2024;

MGSM	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Lrl.	Hrl.	Avg.
ST-Merge (Ours)	**54.0**	**56.8**	**58.8**	**53.5**	**57.2**	**62.4**	**61.2**	**62.8**	**65.2**	**68.0**	**56.5**	**61.5**	**60.0**
w/o Lang. Embed	51.4	54.0	56.4	53.2	56.6	61.4	60.3	62.1	63.9	67.6	53.9	60.7	58.7
w/o Cross-Attention	52.7	54.7	57.8	52.8	55.8	60.8	58.5	62.7	63.4	67.1	55.1	60.2	58.6
w/o Gate Network	50.7	52.4	55.8	51.0	53.6	59.8	56.4	60.7	62.0	66.5	53.0	58.6	56.9
Fix-Merge (Baseline)	50.5	52.9	55.7	50.8	54.8	59.1	56.8	60.7	61.7	66.7	53.0	58.7	57.0

Table 4: Ablation study on MGSM.

Huang et al., 2024; Ruan et al., 2025). The final model is selected based on the averaged performance of all languages on the dev set. For training, we utilized 4 NVIDIA A100 GPUs with a learning rate of 2e-5, a batch size of 128, a maximum sequence length of 512, and a total of 3 epochs. We conduct experiments with three different random seeds and report the average results.

4.3 Baselines

We compare our method against several state-of-the-art baselines for multilingual reasoning:
Translate-En (Shi et al., 2023) translates non-English inputs to English and uses an English reasoning model.
MetaMath (Yu et al., 2024) is fine-tuned from LLaMA2-7B on an additional mathematical dataset MetaMathQA, which serves as the backbone architectures for baseline methods.
MultiReason (Zhu et al., 2024) enhances reasoning consistency across languages via question alignment and rationale generation.
QAlign (Zhu et al., 2024) aligns questions across languages through fine-tuned translation-based contrastive learning.
LangBridge (Yoon et al., 2024) introduces an alignment layer to bridge non-English inputs to an English-centric reasoning space.
MindMerger (Huang et al., 2024) merges task representations across languages to promote cross-lingual reasoning alignment.
LayAlign (Ruan et al., 2025) integrates representations from all encoder layers to enable layer-wise interaction between the two models.

4.4 Main Results

As shown in Table 1, Table 2, and Table 3, ST-Merge achieves consistent performance gains across all tasks, demonstrating strong generalization. (Please refer to Appendix Table 6, Table 7 for the complete results of X-CSQA and XNLI.) The performance gains on the X-CSQA are relatively limited. We conjecture that the limited gains stem from their discrete choice-format outputs, which provide limited signals and hinder effective optimization of the gate network.

Compared to existing strong multilingual reasoning baselines such as MindMerger (Huang et al., 2024), our method achieves superior performance under identical prompts and training data. Our approach can act as a plug-and-play enhancement over fixed model merging strategies, yielding average gains of +1.7%, +1.3%, +1.5%, and +1.5% over the MindMerger baseline on the four benchmarks, respectively.

Notably, our method consistently improves average performance on high-resource languages. As shown in Table 1, compared to the state-of-the-art baseline, ST-Merge achieves an average high-resource gain +1.3% on MGSM and +0.4% on MSVAMP, respectively. ST-Merge maintains highly competitive accuracy in English (68.0% on MGSM), outperforming other multilingual baselines while effectively narrowing the gap with the state-of-the-art. This demonstrates that ST-Merge can preserve reasoning proficiency in high-resource languages.

4.5 Ablation Study

Table 4 presents the ablation results on the MGSM.

(1) *w/o Lang. Embed*, which removes the language identity embeddings from the gating network. The average accuracy drops from 60.0% to 58.7%, with a notable degradation of 2.6% on low-resource languages. This suggests that without explicit language cues, the gating network lacks the guidance to differentiate between languages. Consequently, the optimization process becomes biased towards dominant high-resource languages, hindering the low-resource language reasoning capacity of the merger model.

Question [Swahili]: John hukimbia maili 60 kwa wiki. Huwa anakimbia siku 3 kwa wiki. Huwa anakimbia saa 3 siku ya kwanza na nusu ya kiasi hicho kwa siku nyingine mbili anazokimbia. Huwa anakimbia kwa kasi gani? (**Answer:** 10)	**Output [Fix-Merge]:** John runs 3 miles on the first day. He runs half that distance on the other two days, so he runs 3/2 = 1.5 miles on each of those days. In total, he runs 3 + 1.5 + 1.5 = 6 miles in a week. Since he runs 60 miles in a week, his speed is 6$_{[\text{✗},0.5523]}$miles / 7 days = 0.857 miles per day. The answer is: 0.857.	**Output [ST-Merge (Ours)]:** John runs 3 miles on the first day. He runs half that distance on the other two days, so he runs 3/2 = 1.5 miles on each of those days. In total, he runs 3 + 1.5 + 1.5 = 6 miles in a week. Since he runs 60 miles in a week, his speed is 6**0**$_{[\text{✓},0.9045]}$/6 = 10 miles per hour. The answer is: 10. (ω_A: 1.1172, ω_B: 1.0547)
Question [French]: Kylar se rend au magasin afin d'acheter des verres pour son nouvel appartement. Un verre coûte 5 $, mais chaque deuxième verre ne coûte que 60% du prix. Kylar veut acheter 16 verres. Combien devra-t-il payer pour les acheter? (**Answer:** 64)	**Output [Fix-Merge]:** The first glass costs $5. The second glass costs 60% of $5, which is 0.6 * $5 = $3. So for every two glasses, Kylar pays $5 + $3 = $8. Kylar wants to buy 16 glasses, so he will pay 16/2 * $8 = $**8**$_{[\text{✗},0.4948]}$0 . The answer is: 80.	**Output [ST-Merge (Ours)]:** The first glass costs $5. The second glass costs 60% of $5, which is 0.6 * $5 = $3. For every two glasses, the price is $5 + $3 = $8. Kylar wants to buy 16 glasses, so he will need to pay 16/2 * $8 = $**6**$_{[\text{✓},0.9614]}$4 . The answer is: 64. ($\omega_A$: 1.0156, ω_B: 1.0503)
Question [Chinese]: 帕姆和弗莱德去参加狂欢节。帕姆坐了2 次云霄飞车，而弗莱德坐了4 次。在此之后，他们各自决定坐雪橇2 次。如果每坐一次需要6 张门票，他们那天用掉了多少张门票? (**Answer:** 60)	**Output [Fix-Merge]:** Pam rides the roller coaster 2 times, so she uses 2 x 6 = 12 tickets. Fred rides the roller coaster 4 times, so he uses 4 x 6 = 24 tickets. After riding the roller coaster, they each decide to ride the ice cream, so they each use 2 x 6 = 12 tickets. In total, they use 12 + **1**$_{[\text{✗},0.7773]}$2 + 24 + 24 = 72 tickets. The answer is: 72.	**Output [ST-Merge (Ours)]:** Pam rides the roller coaster 2 times, so she uses 2 x 6 = 12 tickets. Fred rides the roller coaster 4 times, so he uses 4 x 6 = 24 tickets. After riding the roller coaster, they each decide to ride the snow sled 2 times, so they use 2 x 6 = 12 tickets each. In total, they use 12 + 12 = 24 tickets for the snow sled. Therefore, they use a total of 12 + **2**$_{[\text{✓},0.9993]}$4 + 24 = 60 tickets that day. The answer is: 60. (ω_A: 0.9766, ω_B: 0.9314)

Table 5: Case study on MGSM. The GREEN (RED) highlight indicates a correct (incorrect) reasoning step. The real-valued numbers indicate the next token generation probability. (ω_A, ω_B) represent the learned weights by our method for each language.

(2) *w/o Cross-Attention*, which replaces the fine-grained Cross-Attention mechanism with a simpler concatenation of the representations. Removing this module leads to a significant performance decline across all languages, reducing the average accuracy to 58.6%. This suggests that the token-level interaction provided by cross-attention is essential for deeply considering the context from the two models.

(3) *w/o Gate Network*, which completely eliminates the gating network. In this case, we increase the number of training steps to match the computational budget of the proposed ST-Merge. Despite the extended training, this variant shows no improvement over the static baseline (Fix-Merge, 57.0%) and lags significantly behind the proposed full ST-Merge model (60.0%). This validates the indispensability of the gated cross-attention, demonstrating that the performance boost is driven by the steerable merging strategy rather than simply extending the optimization process.

4.6 Case Study

We present a case study to show that the failed cases of Fix-Merge (fixed model merging strategy) can be rectified by our model. We aim to provide insights into the mechanisms underlying the effectiveness of the proposed steerable model merger.

The proposed ST-Merge framework adaptively modulates feature amplification and suppression to extract the most effective representations for multilingual reasoning. Specifically, ST-Merge leverages the complementary strengths of the backbone LLM, which possesses strong intrinsic reasoning capabilities, and the external multilingual model, which excels in cross-lingual semantic understanding. By utilizing learned weights to selectively enhance these respective strengths while suppressing irrelevant noise, the model is steered toward more accurate reasoning outcomes. As shown in Table 5, in the Swahili case, the static baseline (Fix-Merge) fails to retain the critical entity "60" and hallucinates an incorrect number "6 miles" with a low confidence probability of 0.5523. This indicates that the model was trapped in a state of uncertainty due to the symmetric merging. In contrast, ST-Merge identifies that the input requires prioritizing certain models over others. ST-Merge assigns a higher weight value to the multilingual encoder to boost understanding of the question. Consequently, this asymmetric merging enables the model to generate

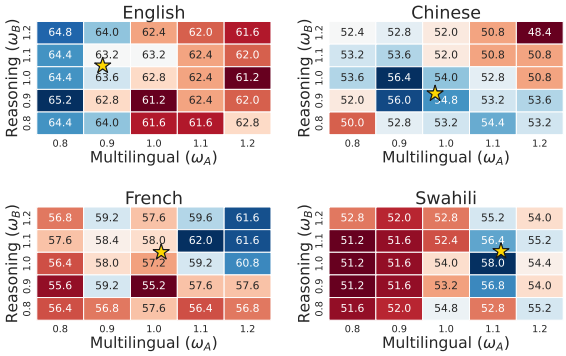
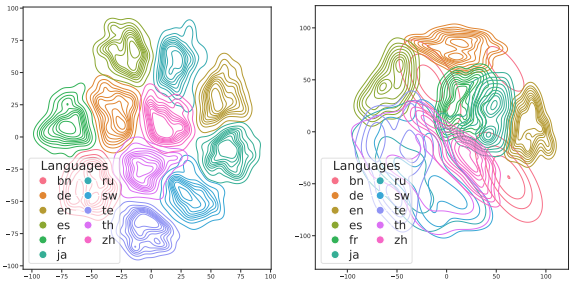


Figure 4: Learned weights analysis of ST-Merge. Darker blue grids indicate higher accuracy. The gold stars represent the learned weights (ω_A, ω_B) by our method for each language.



(a) MetaMath (b) Our ST-Merge

Figure 5: t-SNE visualization of multilingual alignment on MGSM.

the correct answer ("60") with a high confidence of 0.9045, confirming that breaking the symmetry of feature fusion is crucial for robust low-resource reasoning. Similar results can also be observed in the examples of Chinese and French.

4.7 Analysis of Steerable Weights

Figure 4 visualizes the reasoning accuracy on four representative languages from the MGSM dataset (English, Chinese, French, and Swahili) under varying fusion weights ω_A and ω_B. The heatmaps reveal that the optimal weight configuration is highly sensitive to the specific language. For high-resource languages like English and Chinese, the learned weights (marked by gold stars) and optimal regions favor a balanced or reasoning-dominant configuration, leveraging the strong mathematical reasoning capabilities inherent in the base model. In contrast, for a language underrepresented in math reasoning corpora, e.g. Swahili, the model assigns a higher value to ω_A. This suggests that for low-resource languages, the model prioritizes the multilingual module to align representations before performing reasoning. Crucially, our steerable gating mechanism consistently converges to these optimal regions across all languages, demonstrating its ability to adaptively regulate the trade-off between multilingual alignment and mathematical reasoning without manual tuning.

4.8 Multilingual Representation Visualization

To examine the alignment of multiple languages, we compare the alignment results of vanilla fine-tuned MetaMath and our ST-Merge in terms of question representation. We select questions spanning eleven different languages from the MGSM datasets to visualize the embedding space. As shown in Figure 5-(a), different languages are distributed in distinct clusters in the embedding space, which indicates that MetaMath can remain highly language-dependent. In contrast, Figure 5-(b) shows that the data distributions of different languages are mixed and overlapping, which demonstrates that ST-Merge achieves more effective alignment of representations across languages compared to MetaMath. This alignment contributes to the superior multilingual reasoning performance of ST-Merge.

5 Conclusion

This work addresses a fundamental challenge of how to effectively coordinate the multilingual encoder and reasoning LLM for multilingual reasoning tasks. Our analysis reveals that fixed "one-size-fits-all" merging strategies potentially introduce a conflict: while improving reasoning performance on low-resource languages with an external multilingual encoder, they often degrade reasoning in high-resource languages where the LLM is already proficient. To address this, we propose a steerable model merging (**ST-Merge**) framework that optimizes the merged model toward balanced multilingual reasoning via dynamic adjustment of weights. Experiments on four multilingual reasoning benchmarks across 21 languages demonstrate consistent gains across both high-resource and low-resource languages. Beyond performance, we further uncover a correlation between reasoning correctness and gating patterns, providing empirical insight into the mechanisms underlying multilingual reasoning generalization. Additionally, our findings suggest that steerable merging strategies represent a promising direction for enhancing the multilingual capabilities of large language models.

Limitations

Our work presents several limitations worth noting. First, to ensure a fair comparison with baseline models, our method primarily conducts experiments using the Llama 2 series models. Future work will involve extending our experiments to additional series models to more comprehensively evaluate the generalizability of our method across diverse backbone architectures. Second, while our method effectively generates weights to improve model collaboration, it lacks fine-grained guidance during the generation process. We hypothesize that a more granular control mechanism during the decoding phase could further enhance performance. In the future, we will explore incorporating token-level or step-aware guidance to address this issue.

Acknowledgments

This work is supported by the National Key R&D Program of China (2024YFF0907003).

References

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. 2018. Gated-attention architectures for task-oriented language grounding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2819–2826. AAAI Press.

Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. 2024a. Model composition for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11246–11262. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024b. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 7001–7016. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, Maosong Sun, and Yang Liu. 2025. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9413–9422. Computer Vision Foundation / IEEE.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting llm reasoning in non-english languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 34161–34187. Curran Associates, Inc.

Boseung Jeong, Jicheol Park, Sungyeon Kim, and Suha Kwak. 2025. Learning audio-guided video representation with gated attention for video-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 26202–26211. Computer Vision Foundation / IEEE.

Yeachan Kim and Bonggun Shin. 2021. An interpretable framework for drug-target interaction with gated cross attention. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2021, 6-7 August 2021, Virtual Event*, volume 149 of *Proceedings of Machine Learning Research*, pages 337–353. PMLR.

Jun-Tae Lee, Sungrack Yun, and Mihir Jain. 2022. Leaky gated cross-attention for weakly supervised multi-modal temporal action localization. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 817–826. IEEE.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andrés Codas, Clarisse Simões, Sahaj Agrawal, Xuxi

Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045.

David Ortiz-Perez, Manuel Benavent-Lledó, Javier Rodríguez-Juan, José García Rodríguez, and David Tomás. 2025. Cognialign: Word-level multimodal speech alignment with gated cross-attention for alzheimer's detection. *Knowl. Based Syst.*, 329:114264.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *CoRR*, abs/2505.06708.

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. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1481–1495, Albuquerque, New Mexico. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. 2023. An empirical study of multimodal model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1563–1575. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

X-CSQA	Sw	Ur	Hi	Ar	Vi	Ja	Pl	Zh	Nl	Ru	It	De	Pt	Fr	Es	En	Avg.
MetaMath [2024]	24.2	25.1	32.9	32.3	50.9	49.1	50.6	56.5	57.5	56.0	56.0	61.2	61.7	63.5	64.0	76.3	51.3
MultiReason [2024]	27.6	29.2	32.0	28.7	38.8	38.7	45.5	43.8	45.9	46.5	50.2	49.1	51.2	52.1	54.3	67.2	43.8
QAlign [2024]	35.1	32.6	37.8	36.3	50.5	49.2	51.3	54.8	56.3	56.3	58.3	58.8	59.8	60.3	63.1	75.7	52.3
LangBridge [2024]	31.8	30.5	30.6	30.6	33.3	33.9	39.8	39.8	38.4	35.1	39.1	37.4	36.3	38.2	38.4	44.4	36.1
Translate-En [2023]	36.5	41.3	48.4	44.6	51.8	47.1	53.3	51.5	55.0	54.4	56.3	57.3	54.7	57.2	55.5	71.3	52.3
MindMerger [2024]	45.5	46.2	48.4	51.4	60.6	53.9	63.3	62.9	63.8	**63.7**	**66.8**	**67.0**	**67.1**	**68.1**	**69.1**	**78.1**	61.0
LayAlign [2025]	53.3	51.7	**53.7**	55.9	**62.0**	56.4	64.8	**64.6**	**66.2**	62.0	66.2	65.2	64.3	66.5	67.3	76.7	62.3
ST-Merge (Ours)	**53.6**	**51.9**	51.6	56.5	61.7	**57.9**	**65.1**	64.1	64.4	63.1	66.6	65.2	66.4	67.2	67.3	77.3	**62.5**

Table 6: Accuracy (%) on X-CSQA. Avg. represents the average accuracy across all languages.

XNLI	Sw	Ur	Hi	Th	Ar	Tr	El	Vi	Zh	Ru	Bg	De	Fr	Es	En	Avg.
MetaMath [2024]	45.9	49.2	55.7	55.4	60.9	61.9	63.7	73.7	74.7	77.6	76.7	80.6	82.2	82.8	**90.0**	68.7
MultiReason [2024]	56.3	57.5	61.7	60.1	61.7	65.6	67.0	73.7	79.1	79.7	78.7	82.3	82.9	83.9	88.8	71.9
QAlign [2024]	65.2	62.2	63.3	65.2	67.0	67.9	66.5	73.7	76.6	79.2	79.4	80.9	83.1	83.8	89.1	73.5
LangBridge [2024]	71.7	66.9	71.1	72.4	75.2	74.8	79.1	78.5	77.4	77.4	79.6	78.8	79.9	80.5	83.4	76.5
Translate-En [2023]	65.3	61.6	68.7	69.5	68.9	74.5	79.3	76.7	74.8	76.0	80.8	80.6	80.4	81.4	87.4	75.1
MindMerger [2024]	66.6	69.4	74.7	71.8	76.2	75.7	78.5	80.3	80.0	80.7	82.4	83.5	83.9	84.4	88.7	78.4
LayAlign [2025]	73.0	71.0	74.7	74.1	**77.6**	76.0	79.6	**80.8**	80.8	81.8	**83.4**	**83.9**	**84.7**	**84.8**	88.9	79.7
ST-Merge (Ours)	**73.7**	**71.8**	**75.1**	**74.2**	**77.6**	**77.2**	**80.0**	80.1	**81.0**	**82.2**	83.3	83.5	**84.7**	**84.8**	89.1	**79.9**

Table 7: Accuracy (%) on XNLI. Avg. represents the average accuracy across all languages.

Method	Params (M)	FLOPs (G)	Train (h)	Infer (m)	Avg. Acc
ST-Merge (Ours)	10282.5	17700.78	5.57	26.3	**60.0**
w/o Gate Network	10265.2	17700.50	4.98	26.3	57.3
Relative Overhead	+0.17%	+0.0016%	+11.85%	0.00%	+4.71%

Table 8: Comparison of computational overhead and performance.

A Example Appendix

A.1 Complete Experimental Results

To facilitate reference, the languages utilized in this work are abbreviated as follows: Bengali (Bn), Thai (Th), Swahili (Sw), Japanese (Ja), Chinese (Zh), German (De), French (Fr), Russian (Ru), Spanish (Es), English (En), Urdu (Ur), Hindi (Hi), Arabic (Ar), Vietnamese (Vi), Polish (Pl), Flemish (Nl), Italian (It), Portuguese (Pt), Turkish (Tr), Greek (El), and Bulgarian (Bg). Due to page limitations, the complete breakdown of results is included here. We report the extensive experimental data on X-CSQA in Table 6 and on XNLI in Table 7. Additionally, comparative results on MGSM are visualized in Figure 6 (manual weights combinations) and Figure 7 (learned weights combinations).

A.2 Computational Overhead Analysis

To verify whether the performance gains of ST-Merge stem from our proposed steerable merging design rather than a simple increase in parameter capacity, we evaluate its computational overhead

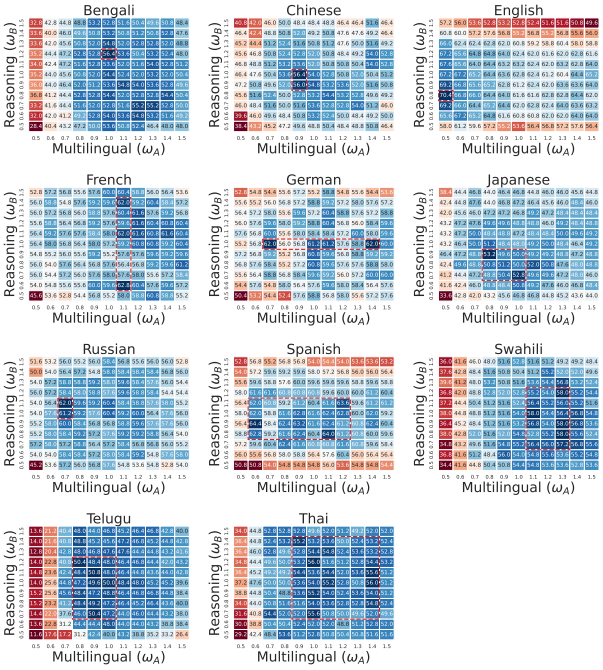


Figure 6: Accuracy on MGSM with different manual weights combinations for the two source models. Darker blue grids indicate higher reasoning accuracy.

against the *w/o Gate Network* variant. As summarized in Table 8, ST-Merge introduces only a marginal parameter increase of 17.3M (+0.17%) and a negligible overhead in FLOPs (+0.0016%). Although the training time increases by 11.85%, the inference latency remains identical to the baseline. Considering the significant absolute improvement of 2.7 points (+4.71%) in average accuracy, these results demonstrate that the effectiveness of ST-Merge is driven by its architectural contribution rather than parameter expansion, ensuring its practicality for resource-constrained applications.

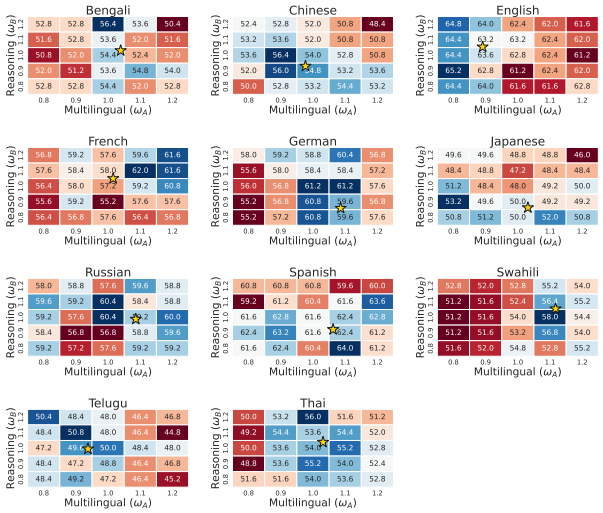


Figure 7: Learned weights analysis of ST-Merge. Darker blue grids indicate higher accuracy. The gold stars represent the learned weights (ω_A, ω_B) by our method for each language.