Language-Agnostic Latent Injection for Multilingual Hybrid Reasoning

Dang Phuong Nam[1], Nguyen Van Hieu[1], Nguyen Nam Truong[1], Phan Duy Hung[1,*]
[1]FPT University, Hanoi, Vietnam,
Co-author Emails: {namdphe180452, hieunvhe181206, truongnnhs173120}@fpt.edu.vn, hungpd2@fe.edu.vn
***Correspondence:** hungpd2@fe.edu.vn

Abstract

Hybrid latent reasoning expands Large Language Models by replacing discrete text decoding with continuous "soft tokens," yet it paradoxically falls behind text-space baselines such as supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO) on low-resource languages. We trace this degradation to the language-marking problem: soft tokens become entangled with language identity, confining reasoning to suboptimal representational regions. We introduce Language-Agnostic Latent Injection (LALI), a lightweight, parameter-free intervention that applies a rank-1 null-space projection—estimated offline from a strictly disjoint multilingual dataset—to the soft tokens during both reinforcement learning and inference. On MGSM8K and MSVAMP, LALI substantially boosts accuracy for mid-to-low-resource languages on compact models (1.5B, 3B) and transfers zero-shot to unseen languages. Our analysis further uncovers a Performance–Fidelity Paradox: the gains arise from the model pivoting its internal deliberation to high-resource languages; at the 7B scale this bottleneck dissolves on its own, positioning LALI as a targeted tool for parameter-constrained multilingual reasoning.

Keywords: Latent Reasoning, Multilingual Reasoning, Low-Resource Languages, Reinforcement Learning.

1 Introduction

1.1 Problem and Motivation

The advent of Large Language Models (LLMs) has significantly advanced complex problem-solving capabilities. Although much of the initial success relied on discrete text decoding techniques such as Chain-of-Thought (CoT) prompting, a promising new direction reshaping the field is latent reasoning-allowing the model to think and compute entirely within a continuous space. Prominent in this trend is the rise of hybrid latent reasoning methods, typified by the Hybrid Reasoning Policy Optimization (HRPO) algorithm proposed by Yue et al. (2025). Unlike pure latent reasoning, hybrid models directly integrate the latent space with token embeddings through "soft tokens." This integration offers enhanced expressivity, allowing the model to optimize its internal thought process naturally and align more closely with the pre-trained vocabulary.

However, while latent and hybrid reasoning demonstrate remarkable efficacy in high-resource languages (e.g., English), they have yet to effectively address cross-lingual transfer and suffer a severe performance drop when applied to low-resource languages. Lim et al. (2025), Zhao et al. (2025), and Park et al. (2025) have identified the root cause of this degradation as the language-marking problem: due to the connection with the vocabulary space, the injected soft tokens become deeply localized within the input language's representational region, meaning they heavily carry the specific language's identity. Because an LLM's core reasoning capacity predominantly resides in a shared semantic subspace formed from resource-rich pre-training data (Lim et al., 2025; Zhao et al., 2025), repeatedly injecting language-marked tokens forces the reasoning trajectory of low-resource queries to deviate into suboptimal representational regions where the model's logic is weakest.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1152–1162
October 17-21, 2026 ©2026 Association for Computational Linguistics

1.2 Related Work

Latent and Hybrid Reasoning. Latent reasoning has catalyzed a paradigm shift from discrete reasoning chains to continuous computation, extending early efforts to provide models with a "thinking" space via pause tokens (Goyal et al., 2023) or self-taught implicit reasoning mechanisms like Quiet-STaR (Zelikman et al., 2024). Contemporary methods are categorized based on their injection signal axis: direct injection into the hidden state, such as in the purely latent COCONUT method (Hao et al., 2024) and the "soft thinking" technique by Zhang et al. (2025) which allows the model to unlock reasoning potential directly in the continuous concept space, versus the use of soft tokens in the vocabulary space calculated as the probability-weighted average of embeddings based on the token distribution. A prominent representative of the vocab-space branch is the HRPO model (Yue et al., 2025), which incorporates a learnable gating mechanism to mix the hidden state into the token embedding, optimized via reinforcement learning (RL). A major advantage of the vocab-space approach is that the injected signal inherently lies on the embedding manifold, ensuring natural compatibility with pre-trained language models. Nevertheless, it is precisely within this projection structure that the language identity label infiltrates the injected signal. Therefore, instead of continuing to tune historically unstable gating mechanisms, our research intervenes directly at the root of the problem: the injected vector itself. To date, no prior work has attempted to disentangle the language identity component from the injected signal in hybrid latent reasoning mechanisms.

Multilingual Reasoning and Language Subspaces. Prior to latent methods, multilingual reasoning capabilities predominantly relied on eliciting Chain-of-Thought directly in the target language or via English translation (Shi et al., 2023). Currently, the literature on multilingual representation is converging on a clear mechanistic picture: the latent spaces of LLMs factorize into a shared semantic subspace and a language-specific subspace. Pioneering this direction, Xie et al. (2022) demonstrated with the LSAR method that language identity occupies a low-rank subspace and can be completely removed via a parameter-free null-space projection-although this mechanism was previously applied exclusively to static sentence embeddings in retrieval tasks. Furthermore, analyses of LLM latent spaces reveal that the processing pipeline is often dominated by an English-centric latent space, overshadowing the representations of other languages (Wendler et al., 2024). Recently, causal studies have strongly corroborated that ablating the language subspace improves reasoning capabilities: Lim et al. (2025) asserted that cross-lingual performance is hindered by language-specific latent processes, while Zhao et al. (2025) observed across 10 LLMs and 11 languages that language-reasoning disentanglement enhances performance, adhering to the principle that "less language is more." In the context of RL training, multilingual reasoning faces significant instability. Park et al. (2025) described the phenomenon of cross-lingual collapse, where a multilingual model degrades and converges toward the dominant language during RL. Similarly, Liu et al. (2026) pointed out that the multilingual latent reasoning capabilities of current models remain rudimentary.

1.3 Contributions

In summary, our main contributions are threefold. First, we construct and release a comprehensive multilingual mathematical reasoning dataset by completely translating GSM8K. Crucially, we enforce a strict disjoint assignment strategy across languages; this non-overlapping design serves as the technical cornerstone that enables the precise isolation and offline estimation of the language subspace, ensuring that language identity is not confounded with problem-specific semantic features. Second, we propose Language-Agnostic Latent Injection (LALI), a lightweight, parameter-free intervention for hybrid latent reasoning. By utilizing the offline-estimated subspace without requiring gradient updates, LALI applies a rank-1 null-space projection to soft tokens during both the reinforcement learning and inference phases. This explicitly strips away

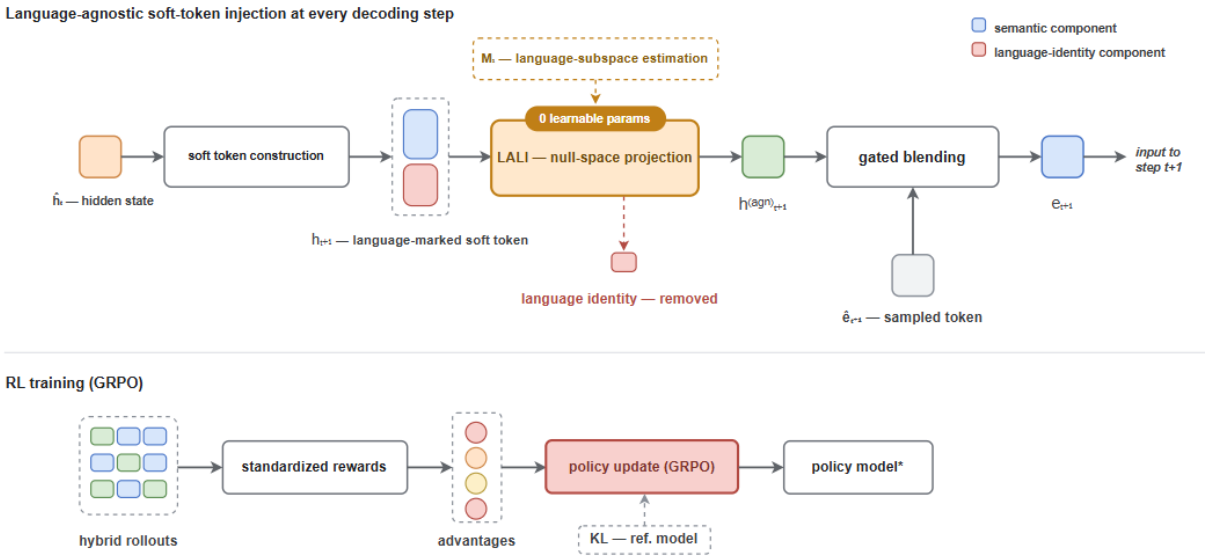


Figure 1: Overview of the LALI pipeline

harmful surface-language representations while preserving universal semantic content, all without introducing any additional learnable parameters to the architecture. Finally, extensive evaluations on MGSM8K and MSVAMP demonstrate that LALI substantially improves reasoning accuracy for low- and mid-resource languages on compact models (1.5B and 3B). Notably, the proposed method exhibits robust zero-shot transfer capabilities, successfully denoising latent representations and boosting performance for entirely unseen languages, such as Spanish and Telugu, which were strictly excluded from the initial subspace estimation.

2 Methodology

Figure 1 provides an overview of the complete pipeline.

2.1 Preliminaries and the Language-Marking Problem

Hybrid Latent Reasoning via Reinforcement Learning (HRPO), proposed by Yue et al. (2025), is a hybrid-latent reasoning method: in the internal reasoning phase, each generation step not only generates a discrete token but also injects a continuous representation-called a soft token-into the input of the next step. At the generation step t, the soft token is created by projecting the final hidden state back into the embedding space (Yue et al., 2025):

$$h_{t+1} = W_e^\top \frac{p_{t+1}}{\|p_{t+1}\|}, \quad p_{t+1} = \text{softmax}\left(\frac{\text{Head}(\hat{h}_t)}{\tau}\right) \tag{1}$$

Where $\hat{h}_t \in \mathbb{R}^d$ is the hidden state at the final layer (d: hidden dimension representation); $\text{Head}(\cdot)$ is the output layer that maps the hidden state to logits over the vocabulary; $\tau > 0$ is the temperature parameter; $p_{t+1} \in \mathbb{R}^{|V|}$ is the next token distribution ($|V|$: vocabulary size); and $W_e \in \mathbb{R}^{|V|\times d}$ is the embedding matrix (tied, frozen). In essence, h_{t+1} is the probability-weighted average of the lexical embedding vectors according to the prediction distribution, and therefore always lies on the embedding manifold and is compatible with the input distribution of the pre-trained model.

The HRPO gating mechanism consists of four components-the regulating gate r_t, the information gate i_t, the mixing weight a_t, and the conditional blending operation that creates hybrid embedding vectors e_{t+1}-fully parameterized as follows (Yue et al., 2025):

$$r_t = \sigma(W_a \hat{e}_{t+1} + b_a), \quad i_t = \sigma(W_x \hat{e}_{t+1} + b_x),$$
$$a_t = \exp(-c \cdot \text{softplus}(\Lambda) \odot r_t),$$
$$e_{t+1} = \begin{cases} a_t \odot \hat{e}_{t+1} + \sqrt{1 - a_t^2} \odot (i_t \odot h_{t+1}) & t \in \text{think}, \\ \hat{e}_{t+1} & t \notin \text{think} \end{cases} \tag{2}$$

In there, e_{t+1} is the resulting hybrid input for the next step; $\hat{e}_{t+1}$ is the embedding vector of the discrete token that was just sampled $\hat{x}_{t+1}$, and h_{t+1} is a soft token according to (Eq. 1);

the two gates r_t and i_t use the sigmoid function σ to control the mixing process, while a_t plays the role of mixing the embedding vector and the soft token for the output e_{t+1}, with c being a fixed constant, Λ being the learned parameter vector, and W_a, W_x, b_a, b_x being the learned projections and biases. The hybrid mechanism is only applied in the inference phase ($t \in$ think); the final answer is still generated using standard autoregressive decoding. By initializing $a_t \to 1$, the input in the initial phase is mainly taken from the embedding vector of the sampled token, thereby preserving the generative capacity of the LLM; according to the HRPO design (Yue et al., 2025), as training progresses, a_t is expected to converge to an optimal value range to integrate features from both the hidden tokens and the discrete token.

Gate collapse. Gate collapse is the phenomenon where the gate mechanism learned in (Eq. 2), after training, completely loses its contextual adaptability and degenerates to a fixed operating point. We detected this by attaching a probe to the decoding loop, recording three quantities: gate value a_t; Hidden Rate $H_t = \sqrt{1 - a_t^2}$ - the continuous branch weight; and the entropy of the next token distribution. On this metric, we perform two checks: (i) the analytical distribution of H_t in each language internally; and (ii) a blind test, splitting tokens into subgroups based on entropy to check whether H_t varies-a suitable gate must rely more on the continuous branch when uncertainty is high. All three probes converge on the same conclusion: after training, the gate values only oscillate with negligible amplitude around a fixed operating point at each scale-the mixing weight a_t is pinned at 1 at all scales (equivalently, H_t is locked at a small constant), while the portal i_t stabilizes at a mid-scale value-and the remaining variance does not correlate with the entropy of the next token distribution, with the input language, or with the input magnitude. Functionally, the gate mechanism has lost its adaptability: equation (2) degenerates into a static affine blend where the initialization coefficient is determined instead of context. The addition of an illustrative conditional gate variant with 12.6M parameters also does not change this conclusion: the conditional parameters are inert, and quality improvement is dismissed as an artifact of Seed variance. Full exploratory data is presented in Appendix A.

Central prediction: soft token is language-marked. The language-marking problem is a phenomenon where the soft token provides language identification information in addition to semantic content. This is a direct consequence of structure (1): h_{t+1} is a (normalized) combination of rows of W_e with weight p_{t+1}; for queries on low-resource languages, the high-probability tokens are almost entirely low-resource language tokens, creating soft tokens that pinpoint the specific representational region in the embedding vocabulary space. From this, we argue that soft tokens decompose into two components: a semantic component encoding the next conceptual order of the inference sequence, invariant across languages-and a language-identity component-hostile signals about the writing system and vocabulary of each language. Because the inference capacity of the model resides primarily in the common semantic subspace (shared semantic subspace) formed from resource-rich data (Lim et al., 2025; Zhao et al., 2025), cyclical injection at each generation step, accumulating over hundreds of steps of the inference phase a linguistic identity-heavy expression will anchor the reasoning trajectory of less resource-rich languages in the representation region where the model is weakest. This is the root cause that LALI is addressing.

2.2 LALI: Language-Agnostic Latent Injection

LALI stems from the separation hypothesis stated in Section 2.1: if the soft token consists of a semantic component and a language-identity component, then removing the second component while preserving the first will transform the injection signal into a language-agnostic conceptual representation. Inference on resource-poor languages then takes place in the shared semantic subspace-the representation area where the model has accumulated inference power from resource-rich data. This hypothesis entails a testable

prediction: the improvement is inversely proportional to the resource level of the language, because the language-identity component becomes less harmful as the language becomes more resource-rich, causing the removal to be asymptotically neutral in the high-resource group; we use this prediction as a test axis for the independent mechanism in Section 4.1. Technically, LALI consists of two steps: a single estimation of the language subspace M_s (Eq. 3–4), and a null-space projection applied at every generation step (Eq. 5–6).

The language subspace M_s is estimated once. We run the model on a sample set consisting of several responses for each language $l \in \{en, zh, th, sw, bn\}$, collect soft tokens according to (Eq. 1), and calculate the average vector for each language:

$$\mu_l = \frac{1}{N_l} \sum_{n=1}^{N_l} h^{(l,n)} \tag{3}$$

Where $h^{(l,n)} \in \mathbb{R}^d$ is soft token n collected from the query language l; N_l is the total number of soft tokens collected from the query language l; and $\mu_l \in \mathbb{R}^d$ is the corresponding mean vector. The mean vectors are column-sorted into a matrix $M = [\mu_{en}, \ldots, \mu_{bn}] \in \mathbb{R}^{d \times L}$ (L: number of languages participating in the estimation), centered on the global average, and then singular value decomposition (SVD) is performed using the LSAR framework (Xie et al., 2022):

$$\begin{aligned} M_c &= M - \bar{\mu}\mathbf{1}^\top = USV^\top, \\ M_s &= U_{[:,1:r]}, \quad M_s^\top M_s = I_r \end{aligned} \tag{4}$$

Where $\bar{\mu} = \frac{1}{L}\sum_l \mu_l$ is the global mean; $\mathbf{1} \in \mathbb{R}^L$ is the all-ones vector; U, S, V are the matrix of left singular vectors, the diagonal matrix of singular values (in decreasing order), and the matrix of right singular vectors; $M_s = U_{[:,1:r]} \in \mathbb{R}^{d \times r}$ includes the first r columns of U-the directions of maximum variance across languages is orthonormal: $M_s^\top M_s = I_r$, with I_r being the identity matrix of rank r.

The rationale for the decomposition (Eq. 3–4): when averaging over a large number of soft tokens within the same language, with content-discrete problems across languages (data design in Section 3.1), the semantic component-which varies according to the problem content is averaged out, while the language-identification component, defined internally within each language, is preserved. Variance between vectors μ_l primarily reflects the language difference, and the leading singular vectors of M_c approximate the axes of language identification in the embedding space. A rank survey shows that $r = 1$ achieves the best accuracy-consistent with the observation by Xie et al. (2022) in their LSAR study that removing too many dimensions compromises semantic information (full survey including the trade-off: Appendix C). The entire primary experiment uses $r = 1$; in other words, LALI removes exactly one direction from the $d = 2048$ dimensional representation space.

Null-space projection. The only LALI plug-in applied to transform between (Eq. 1–2) is:

$$h_{t+1}^{\text{agn}} = (I - M_s M_s^\top) h_{t+1} \tag{5}$$

Where $I \in \mathbb{R}^{d \times d}$ is the identity matrix and h_{t+1}^{agn} is the language-agnostic soft token. Because $P = I - M_s M_s^\top$ is an orthogonal projection ($P^2 = P, P^\top = P$), the transformation (Eq. 5) removes the projection of the soft token onto the language subspace, while ensuring the preservation of the orthogonal component carrying the semantic content. The blending step then operates on the language-de-labeled signal instead of the raw soft token:

$$e_{t+1} = a_t \odot \hat{e}_{t+1} + \sqrt{1 - a_t^2} \odot (i_t \odot h_{t+1}^{\text{agn}}) \tag{6}$$

The symbols in (Eq. 6) retain the same meaning as in (Eq. 2), and the mixing operation is still only applied in the steps $t \in$ think; the only difference is that LALI does not modify parameter gates or any other HRPO components. Applying the projection before the gate i_t does not affect the results: since the information gate after collapse is almost fixed (Appendix A), $i_t \odot h_{t+1}^{\text{agn}}$ is just a scaling of h_{t+1}^{agn}. Therefore, the projection i_t should yield the same result whether applied before or after.

Tested and applied during training, this ensures the policy is optimized for the latent distribution encountered at evaluation time; Figure 2 illustrates the effect of the LALI-referenced soft token. In terms of implementation, it requires only about five lines of code at the entry point of the latent blending function. Finally, M_s is

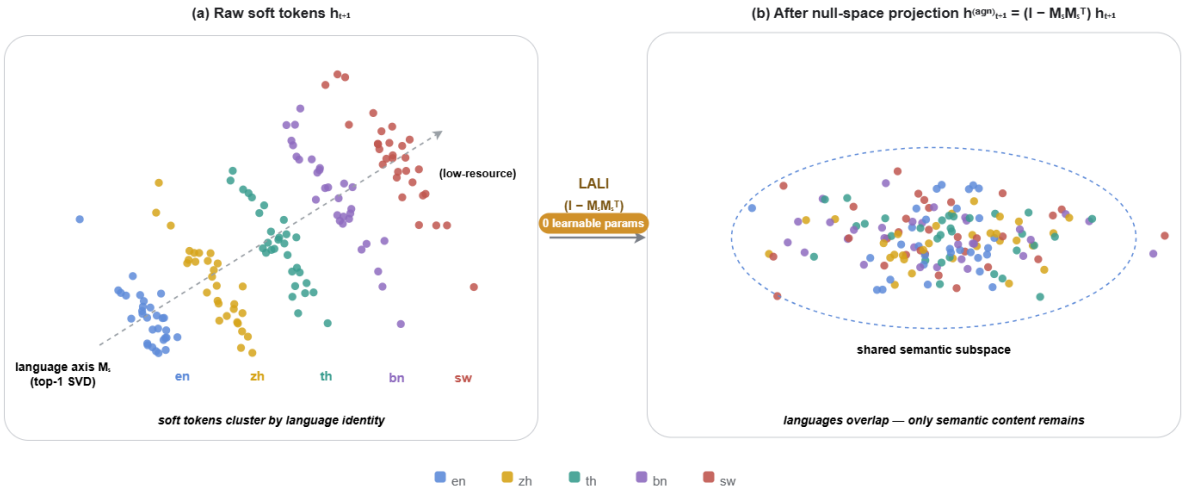


Figure 2: Geometry of the null-space projection

estimated only once and frozen for the model's lifetime: no gradients flow through the projection, and no additional learning parameters are added-in contrast to the parameter set $W_a, W_x, b_a, b_x, \Lambda$ of the gate mechanism-therefore, any improvements in reasoning quality cannot be attributed to an increase in model capacity.

3 Experimental Setup

3.1 Training Data

A multilingual training dataset is constructed from the GSM8K dataset (Cobbe et al., 2021) utilizing our proposed strictly disjoint assignment strategy. The GSM8K translations, complete with chain-of-thought annotations for English, Chinese, Thai, Swahili, Bengali, and Vietnamese, are machine-translated via the Sonnet 4 model. Each problem is uniquely assigned to a single language, strictly ensuring no cross-lingual overlap. Specifically, the quota for each language is purposefully established to simulate the empirical data distribution trajectory in multilingual NLP systems. The dataset is stratified across three resource tiers: a high-resource group with an abundance of mathematical reasoning data (2,900 English samples, 1,800 Chinese samples); a medium-resource group acting as a transitional anchor (1,100 Vietnamese samples, 700 Thai samples); and a low-resource group suffering from severe digital data scarcity, capped at 500 samples per language (Swahili and Bengali). The total training set comprises 7,454 problems.

This disjoint design is the core factor enabling the identification of the language subspace. Since each language is represented by a completely distinct set of problems, the average value of the soft tokens per language (analyzed in Section 2.2) effectively constitutes the average over distinct textual contents. The mathematical features will converge toward a shared centroid across languages; consequently, the remaining residual variance among the averages will purely reflect language identity characteristics, rather than being confounded by the specific problems that a language happens to contain. Conversely, a parallel corpus-where all languages share the exact same set of problems-would entangle these two factors.

3.2 Evaluation

Accuracy is reported across two standard datasets with diverse linguistic configurations: the MSVAMP dataset (Chen et al., 2024) is evaluated in English, Chinese, Thai, Swahili, Bengali, and Spanish; whereas the entire GSM8K test set is evaluated in English, Chinese, Swahili, Telugu, Spanish, and Vietnamese. Similar to the training set construction methodology, all translations within the test sets are generated using Sonnet 4. Spanish (high-resource) and Telugu (low-resource) are established as held-out languages completely absent from the training set (denoted by †), to verify whether the intervention is capable of zero-shot transfer to language spaces not involved in the M_s

estimation process.

3.3 Baselines and Methods

All evaluated methods utilize a single, fixed prompt and do not impose any constraints on the output language during the reasoning process. This setup is implemented to directly test the natural cross-lingual transfer capability of the model, allowing for the observation of how the model self-structures its thought process rather than being forced into a target format. The comparative analysis includes the following approaches: the base model (un-finetuned), supervised fine-tuning (SFT), text-space GRPO, HRPO (Yue et al., 2025), and LALI. The LALI framework is deployed in two primary phases. In the initial phase, the M_s space is estimated exactly once prior to training, requiring no gradient updates. Specifically, soft tokens are collected by executing forward passes on 64 unlabeled held-out prompts for each of the five core languages (en, zh, th, sw, bn); subsequently, the extraction procedure (detailed in Section 2.2) is applied with a rank setting of $r = 1$. In the second phase, the standard HRPO algorithm is implemented with a frozen M_s. The denoising projection is then applied directly at the injection point throughout both training and inference. This decoupled design ensures that LALI introduces virtually no computational overhead compared to the original HRPO, as the projection matrix is pre-computed offline.

4 Results

4.1 Main Results

The performance of the proposed LALI method is evaluated on two multilingual mathematical reasoning benchmarks: MGSM8K and MSVAMP. Comprehensive empirical results across language models at 1.5B, 3B, and 7B scales are detailed in Table 1. From the collected data, four core observations are derived.

First, LALI establishes superior performance on compact foundational architectures. On the 1.5B and 3B models, the LALI method consistently outperforms reinforcement learning baselines (RL-baselines) such as GRPO and the hybrid latent reasoning method HRPO across the vast majority of evaluated languages. This disparity is particularly pronounced in medium- to low-resource languages. For instance, on the MSVAMP benchmark (1.5B model), LALI increases the accuracy for Bengali (bn) from 29.0 (HRPO) to 43.6, and for Thai (th) from 54.6 to 59.5. This demonstrates that projecting the signal to ablate the language-identity component effectively enables the latent space of smaller models to concentrate more efficiently on core semantic reasoning features. (Note: The SFT method exhibits unusually high accuracy on certain extremely rare languages like Swahili, though RL methods generally dominate the overall performance trends.)

Second, the approach demonstrates breakthrough zero-shot transfer capabilities. One of the most striking findings is LALI's robust transferability to languages entirely absent from the training set (denoted by †). For Spanish (es) on the MSVAMP dataset at the 1.5B scale, LALI achieves 73.7, significantly surpassing HRPO's 61.4 and GRPO's 62.1. Similarly, for an extremely low-resource language such as Telugu (te) on the MGSM dataset at the 3B scale, HRPO's performance stagnates at 7.7, whereas LALI boosts this figure to 16.6 (more than doubling the accuracy). This result confirms that the M_s subspace estimated from the five core languages successfully captures a sufficiently generalized multilingual identity direction, allowing LALI to successfully apply the denoising projection even within entirely novel language spaces.

Third, the degree of improvement follows an inversely proportional relationship with language resources. As predicted by the initial hypothesis established in Section 2.2, the performance gains yielded by LALI tend to be inversely proportional to the resource availability of the language. For high-resource languages such as English (en), the improvement from HRPO to LALI is positive but marginal (e.g., MGSM 3B English increases from 83.2 to 84.5). Conversely, weaker languages (such as Bengali - bn) record substantial leaps (from 55.2 to 63.3 on MSVAMP 3B). This proves that the language identity component acts as a

Table 1: Accuracy (%) on MGSM8K and MSVAMP across model scales (†: held-out languages)

	MGSM8K						MSVAMP					
Method	**en**	**es †**	**zh**	**vi**	**te †**	**sw**	**en**	**es †**	**zh**	**th**	**bn**	**sw**
Qwen2.5 Instruct 1.5B												
SFT	52.5	40.5	44.7	35.3	**9.3**	**11.6**	59.7	50.6	47.7	45.5	30.9	**24.2**
GRPO	68.6	48.2	59.7	48.2	7.1	4.3	78.3	62.1	65.3	59.3	41.9	14.1
HRPO	69.1	54.7	64.5	53.4	5.3	3.8	80.8	61.4	67.1	54.6	29.0	13.0
LALI	**70.4**	**58.9**	**66.6**	**56.9**	6.3	3.9	**81.9**	**73.7**	**67.7**	**59.5**	**43.6**	13.9
Qwen2.5 Instruct 3B												
SFT	64.3	56.9	58.5	49.7	12.6	**24.0**	70.3	61.3	59.5	58.7	47.9	**32.2**
GRPO	80.7	66.5	71.7	67.5	7.7	5.5	86.6	75.1	78.2	66.3	49.3	14.2
HRPO	83.2	73.5	**78.4**	71.2	7.7	7.3	88.5	80.5	**81.5**	72.2	55.2	18.5
LALI	**84.5**	**75.5**	78.0	**71.8**	**16.6**	12.4	**89.9**	**84.2**	80.9	**74.6**	**63.3**	26.6
Qwen2.5 Instruct 7B												
SFT	73.3	65.9	69.6	60.7	24.6	36.8	75.6	73.1	74.0	64.0	56.7	39.6
GRPO	89.9	78.4	83.4	78.0	33.2	22.1	90.1	87.8	87.1	78.8	69.7	47.7
HRPO	**91.1**	**81.8**	**85.4**	**81.6**	**38.4**	**26.0**	**91.9**	**90.3**	**87.8**	**81.1**	**75.8**	49.3
LALI	90.5	81.5	85.2	80.2	29.9	23.4	91.2	88.6	86.4	78.7	69.8	**51.0**

bottleneck hindering the sharing of reasoning capabilities in low-resource languages, and LALI successfully dismantles this barrier.

Finally, performance convergence is observed at larger model scales (scaling behavior). When scaling the foundational architecture to 7B parameters, the gap between LALI and HRPO/GRPO not only narrows, but HRPO overtakes as the leading method across most evaluated languages (e.g., MGSM8K English 7B, HRPO achieves 91.1 compared to LALI's 90.5). This phenomenon is entirely consistent with theoretical expectations: when the capacity of the representation space is sufficiently large (reaching the 7B parameter threshold), the model inherently possesses the ability to simultaneously absorb and process both language identity features and semantic reasoning features without encountering interference. Therefore, LALI's language-identity ablation intervention offers the highest practical value and urgency when solving the optimization of multilingual reasoning on strictly parameter-constrained models (under 7B).

4.2 Analysis: The Trade-off Between Accuracy and Reasoning Fidelity

To decode LALI's performance gains, we analyze Reasoning Language Fidelity-the proportion of CoT responses where the reasoning segment matches the target language. While baselines like SFT maintain near-absolute fidelity ($\sim$ 1.0), LALI intentionally induces a reasoning language pivot. On 1.5B and 3B models, LALI's fidelity on low-resource languages drops drastically, pivoting almost entirely (98% – 100%) to English. This is not a flaw, but the exact mechanism of the null-space projection ($P = I - M_s M_s^\top$). Stripped of its weak surface language anchor, the model is liberated from low-resource representation spaces and automatically falls back to robust English thought streams. This Performance-Fidelity Paradox explains the accuracy surge, though it naturally resolves at the 7B scale, where increased capacity allows the model to recover fidelity (e.g., Telugu reaches 0.86) alongside performance. Crucially, the severity of this trade-off is governed by the projection rank r. Our preliminary ablation reveals that while $r = 1$ maximizes downstream accuracy by completely sacrificing target-language fidelity, increasing the rank to $r = 2$ acts as an intrinsic recovery mechanism, allowing the model to unexpectedly regain native reasoning fidelity while retaining a substantial portion of the accuracy gains. We defer the comprehensive trade-off analysis, along with the full accuracy and fidelity results for the r-ablation, to Appendices B and C.

5 Conclusion

This study traced the low-resource failure of hybrid latent reasoning to a concrete mechanistic cause: the injected soft token

carries a language-identity component that anchors reasoning within the model's weakest representational regions, while the hybrid gate itself has degenerated into a constant and offers no leverage. To address this, we proposed LALI—a parameter-free, rank-1 null-space projection leveraging an offline-estimated subspace derived from our strictly disjoint multilingual dataset. LALI removes this harmful component from the injected signal during both reinforcement learning and inference, leaving the core HRPO architecture and reward untouched. Across three model scales and two multilingual benchmarks, LALI delivers its largest gains precisely where theory predicts: low- and mid-resource languages under capacity-constrained models, alongside robust zero-shot transfer to languages absent from both training and subspace estimation. The fidelity analysis reveals the underlying mechanism: freed from its surface language anchor, the model routes low-resource reasoning through its strongest (English-centric) thought streams; at the 7B scale, the model balances the two components on its own and the benefit of the intervention saturates. Since LALI introduces no additional learnable parameters, these improvements are strictly attributable to the explicit latent disentanglement rather than an increase in model capacity.

6 Limitations

While effective for cross-lingual reasoning, LALI has several limitations. First, maximizing performance ($r = 1$) severely degrades native explainability: the model pivots to English reasoning, sacrificing target-language fidelity (Section 4.2) and pedagogical utility. Second, experiments are limited to the Qwen 2.5 family; latent representation geometry may vary across other architectures. Third, validation is restricted to logic-intensive math domains. In culture-dependent tasks, ablating the M_s subspace might inadvertently destroy localized semantic features. Finally, LALI's impact saturates at larger scales (7B). As parameter capacity increases, baseline models inherently learn to balance language identity and reasoning, diminishing the need for explicit null-space projection.

References

Nianyi Chen, Ming Zhu, D. Wu, et al. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. doi:10.18653/v1/2024.findings-emnlp.411.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. Training verifiers to solve math word problems. Computing Research Repository, arXiv:2110.14168.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. Think before you speak: Training language models with pause tokens. Computing Research Repository, arXiv:2310.02226.

Shibo Hao, Sainbayar Sukhbaatar, Dan Su, Xi Li, Zhiyuan Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. Computing Research Repository, arXiv:2412.06769.

Zheng Xin Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. Computing Research Repository, arXiv:2505.13141.

Yuexin Liu, R. Zhao, Hinrich Schütze, and Michael A. Hedderich. 2026. Large reasoning models are (not yet) multilingual latent reasoners. Computing Research Repository, arXiv:2601.02996.

Chanjun Park, J. Kim, J. Lee, S. Bae, Jaegul Choo, and Kang Min Yoo. 2025. Cross-lingual collapse: How language-centric foundation models shape reasoning in large language models. Computing Research Repository, arXiv:2506.05850.

Qwen Team. 2024. Qwen2.5 technical report. Computing Research Repository, arXiv:2412.15115.

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 *International Conference on Learning Representations (ICLR)*. Computing Research Repository, arXiv:2210.03057.

Chris Wendler, Veniamin Veselovsky, G. Monea, and Robert West. 2024. Do llama models speak english in their latent space?. Computing Research Repository, arXiv:2402.10588.

Zhipeng Xie, H. Zhao, T. Yu, and S. Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5617–5633.

Zhenrui Yue, B. Jin, H. Zeng, H. Zhuang, Z. Qin, J. Yoon, L. Shang, Jiawei Han, and Dong Wang. 2025. Hybrid latent reasoning via reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. Computing Research Repository, arXiv:2505.18454.

Eric Zelikman, Georges Harik, Yijia Shao, V. Jayasiri, Noah Haber, and Noah Goodman. 2024. Quiet-STaR: Language models can teach themselves to think before speaking. Computing Research Repository, arXiv:2403.09629.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. 2025. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. Computing Research Repository, arXiv:2505.15778.

W. Zhao, J. Guo, Y. Deng, T. Wu, W. Zhang, Y. Hu, X. Sui, Y. Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. When less language is more: Language-reasoning disentanglement makes LLMs better multilingual reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Appendix A Gate Collapse

This appendix reports the full statistics of the gate-instrumentation probes described in Section 2.1. Table 2 surveys the hybrid gate across all HRPO think-tokens at three scales (a total of $\approx$ 207,000 tokens over five languages: en/es/zh/sw/te). "H_t per language" denotes the average H_t value within each language; "Gap H_t" denotes the largest H_t difference between the five entropy quintiles. If the gating mechanism operated as designed, H_t should vary with content and language; the survey shows that it does not. Across 79,072 think-tokens at the 3B scale, H_t is pinned at 0.0870 with a coefficient of variation of only 0.012%; the per-language means agree to a resolution of 10^{-4}; and when tokens are binned into five entropy quintiles-where the surprisal of the next token ranges from $\sim$0 to 0.12-the H_t difference between quintiles is below 10^{-5} in every case. A per-dimension probe rules out the possibility that oppositely varying dimensions cancel in the mean: the standard deviation across hidden dimensions at 3B is only $\sim$0.0017 (0.17% of the operating value), uniform across all languages. By these criteria, the gate is both content-blind and language-blind: at the convergence checkpoint, "adaptive latent usage" reduces to a fixed operation applied uniformly to every token. Notably, the operating value is determined by the initialization radius $r_{\min}$ (0.99 $\rightarrow\sim$ 0.087 at 1.5B/3B; 0.98 $\rightarrow$ 0.136 at 7B) rather than by model size: gate collapse is a property of the method, scale-invariant, and not an artifact of a single training run.

Appendix B Reasoning-Language Fidelity Matrix

This appendix reports the full reasoning-language fidelity matrix discussed in Section 4.2. For the 3B scale, two additional rows report LALI trained at ranks r=2 and r=3 from the rank sweep of Appendix C. Values in parentheses indicate the dominant pivot language and its share; for example, (en94) means that 94% of the responses in that cell reason in English.

Table 2: Hybrid gate probing statistics for HRPO during decoding

Model	r_{min}	think token	H_t mean(std)	CV	H_t per language	Entropy Q1-Q5	Gap H_t
1.5B	0.99	47889	0.08727	0.013%	0.0873	0.000-0.359	$< 7e{-}6$
3B	0.99	79072	0.08697	0.012%	0.0870	0.000-0.122	$< e{-}6$
7B	0.98	80111	0.13614	0.008%	0.1361	0.000-0.112	$< 2e{-}6$

Table 3: Reasoning-language fidelity matrix (†: held-out languages)

	MSVAMP						GSM8K					
Method	**en**	**zh**	**th**	**sw**	**bn**	**es †**	**en**	**zh**	**sw**	**te †**	**es**	**vi**
Qwen2.5 Instruct 1.5B												
SFT	1.0	1.0	1.0	0.98	1.0	1.0	1.0	1.0	0.99	1.0	0.99	1.0
GRPO	1.0	1.0	0.03	0.0	0.0	0.97	1.0	1.0	0.01	0.0	0.83	0.1
HRPO	1.0	1.0	0.75	0.12	0.45	0.91	1.0	1.0	0.09	0.1	0.74	0.7
LALI	1.0	1.0	0.02	0.0	0.0	0.01	1.0	1.0	0.01	0.0	0.0	0.02
Qwen2.5 Instruct 3B												
SFT	1.0	0.99	1.0	1.0	1.0	0.98	1.0	1.0	0.99	1.0	1.0	1.0
GRPO	1.0	1.0	0.99	0.99	1.0	1.0	1.0	1.0	0.99	1.0	0.99	1.0
HRPO	1.0	1.0	0.99	0.92	0.97	0.99	1.0	1.0	0.97	0.99	0.98	1.0
LALI	1.0	1.0	0.01	0.01	0.0	0.2	1.0	1.0	0.0	0.15	0.3	0.84
LALI (r=2)	1.0	1.0	0.95	0.06 (en94)	0.66 (en34)	0.88	1.0	1.0	0.17 (en83)	0.96	0.84 (en16)	0.99
LALI (r=3)	1.0	0.92	0.5 (en49)	0.02 (en98)	0.06 (en94)	0.14 (en86)	1.0	0.98	0.05 (en95)	0.18 (en81)	0.13 (en87)	0.45 (en54)
Qwen2.5 Instruct 7B												
SFT	1.0	1.0	1.0	1.0	1.0	0.99	1.0	1.0	0.99	1.0	0.99	1.0
GRPO	1.0	1.0	0.08	0.05	0.18	0.91	1.0	1.0	0.08	0.31	0.94	0.69
HRPO	1.0	0.99	0.02	0.06	0.0	0.0	1.0	0.99	0.03	0.03	0.06	0.02
LALI	1.0	1.0	0.12	0.13	0.41	1.0	1.0	1.0	0.17	0.86	1.0	0.87

Table 4: Rank ablation: accuracy (%) of LALI trained at rank $r \in \{1, 2, 3\}$ in Qwen2.5 3B(†: held-out languages).

	MSVAMP						GSM8K						
Config	**en**	**zh**	**th**	**sw**	**bn**	**es †**	**en**	**zh**	**sw**	**te †**	**es**	**vi**	**low-res**
HRPO	89.2	82.5	72.3	12.9	50.5	80.2	82.5	76.7	4.9	7.7	70.4	69.9	36.4
$r = 1$	**89.9**	80.9	**74.6**	**26.6**	63.3	**84.2**	84.5	**78.0**	**12.4**	16.6	**75.5**	**71.8**	**44.2**
$r = 2$	88.8	**82.6**	73.6	**26.6**	57.8	82.4	**85.1**	77.1	11.8	8.6	71.3	71.2	41.6
$r = 3$	89.3	79.5	71.0	25.7	**63.8**	83.0	82.9	76.0	12.2	**17.4**	74.5	71.1	43.5

Appendix C Rank Ablation: Why $r = 1$

This appendix reports the train-matched rank sweep behind the choice of $r = 1$ in Section 2.2. LALI is retrained with the null-space projection applied at rank $r \in \{1, 2, 3\}$ under an otherwise identical recipe (seed 42); the HRPO reference uses no projection. Table 4 reports accuracy over the 12 evaluation cells, where *low-res* denotes the mean over the low-resource cells th/sw/bn (MSVAMP), sw/te (GSM8K), and vi. The corresponding reasoning-language fidelity of the r=2 and r=3 configurations is reported alongside the main fidelity matrix in Table 3 (Appendix B), annotated with the dominant pivot language; the r=1 configuration coincides with the LALI rows reported there.

The sweep shows that r is not monotone; it governs an accuracy–fidelity trade-off. $r = 1$ attains the highest low-resource accuracy (44.2 mean) but pivots reasoning almost entirely into English (fidelity $\approx$ 0 on th/sw/bn); $r = 2$ retains most of the accuracy gain (41.6) *while restoring target-language reasoning* (th 0.95, bn 0.66, te 0.96, es 0.88, vi 0.99); $r = 3$ degrades both axes. We select $r = 1$ for the main experiments to maximize accuracy; practitioners who require native-language reasoning may prefer $r = 2$.