Language of Thought Shapes Output Diversity in Large Language Models

Shaoyang Xu, Wenxuan Zhang[*]
Singapore University of Technology and Design
shaoyang_xu@mymail.sutd.edu.sg, wxzhang@sutd.edu.sg

Abstract

Output diversity is crucial for Large Language Models as it underpins pluralism and creativity. In this work, we reveal that controlling the language used during model thinking—the *language of thought*—provides a novel and structural source of output diversity. Our preliminary study shows that different thinking languages occupy distinct regions in a model's thinking space. Based on this observation, we study two repeated sampling strategies under multilingual thinking—*Single-Language Sampling* and *Mixed-Language Sampling*—and conduct diversity evaluation on outputs that are controlled to be in English, regardless of the thinking language used. Across extensive experiments, we demonstrate that switching the thinking language from English to non-English languages consistently increases output diversity, with a clear and consistent positive correlation such that languages farther from English in the thinking space yield larger gains. We further show that aggregating samples across multiple thinking languages yields additional improvements through compositional effects, and that scaling sampling with linguistic heterogeneity expands the model's diversity ceiling. Finally, we show that these findings translate into practical benefits in pluralistic alignment scenarios, leading to broader coverage of cultural knowledge and value orientations in LLM outputs. Our code is publicly available at https://github.com/iNLP-Lab/Multilingual-LoT-Diversity.

1 Introduction

Large Language Models (LLMs) have been globally adopted due to their extensive knowledge and strong reasoning capabilities. Beyond the correctness of individual responses, this widespread use has drawn increasing attention to the *diversity* of LLM-generated outputs. Formally, output diversity quantifies a model's ability to generate multiple distinct responses to open-ended questions without ground-truth answers (Jiang et al., 2025; Zhang et al., 2025). It is recognized as a fundamental objective in pluralistic alignment research (Sorensen et al., 2024; Conitzer et al., 2024), where low diversity can lead to homogenization—often referred to as mode collapse (Jiang et al., 2025; Zhang et al., 2025; Lagzian et al., 2025)—and the over-representation of dominant cultural values (AlKhamissi et al., 2024; Wang et al., 2024). Moreover, diversity is a key indicator of whether AI systems exhibit human-like creativity (Pépin et al., 2024), laying the foundation for innovative problem-solving (Ye et al., 2025; Tian et al., 2024; Chen et al., 2025b; Han et al., 2025), open-ended exploration, and the generation of novel ideas (Guo et al., 2025a; Ruan et al., 2024).

To improve output diversity, temperature scaling is commonly utilized by increasing sampling randomness (Pépin et al., 2024; Tevet and Berant, 2021; Peeperkorn et al., 2024). Other work explored advanced decoding methods (Peeperkorn et al., 2025), aggregating outputs from multiple LLMs (Liang et al., 2024a; Shur-Ofry et al., 2024; Tekin et al., 2024), or increasing prompt variation (Shur-Ofry et al., 2024; Lagzian et al., 2025; Wang et al., 2025a). At training time, several studies proposed diversity-driven RLHF and SFT objectives to encourage more varied generations (Li et al., 2025b; Sun et al., 2025).

Despite their promise, most existing work focuses on English-only or multilingual input settings (Wang et al., 2025a). In contrast, we investigate whether the language used during intermediate thinking—referred to as the *language of thought*—can serve as a controllable and structural source of output diversity. Our investigation is motivated by two observations. First, insights from cognitive science suggest that multilingualism promotes divergent thinking and creativity, as different

[*] Corresponding author

languages encode distinct conceptual and structural biases (Blasi et al., 2022; Kharkhurin et al., 2023). According to the Sapir–Whorf hypothesis (Whorf, 2012), language can shape how concepts are organized and related during thinking. Second, recent studies have demonstrated that modern LLMs are capable of explicit reasoning in multiple languages, with performance differences across languages (Yong et al., 2025; Qi et al., 2025). Together, these insights motivate us to study *language of thought* as a structural property of the model's thinking process, and to examine how varying this property influences output diversity.

To this end, we begin with a preliminary study that explores *whether different thinking languages induce structural differences in the model's thinking space* (§3). Specifically, given the same English input, we control the thinking process to be conducted in different languages and collect the resulting hidden representations. By visualizing these multilingual thinking representations, we observe that different languages correspond to distinct regions in the model's thinking space. Moreover, non-English languages exhibit substantial variation in their distances to English thinking. These observations reveal geometric differences induced by different languages of thought.

Building on these observations, we next examine *whether the thinking-space shifts induced by different languages of thought help output diversity* (§4&5). Although the thinking process is controlled to be conducted in different languages, we further control the model's final outputs to English for fair output diversity evaluation (§4.1). Based on this setup, we perform *repeated sampling* and aggregate the resulting English outputs for diversity evaluation. Specifically, we explore two sampling strategies. The first, *Single-Language Sampling*, performs repeated sampling within a single thinking language (§4.2). The second, *Mixed-Language Sampling*, aggregates English outputs generated through thinking in different languages (§4.3).

We conduct experiments on two benchmarks using two different diversity metrics. Multiple LLMs and 15 thinking languages are evaluated (§5.1). Our main findings are as follows.

First, under *Single-Language Sampling*, we observe that simply switching the language of thought from English to non-English languages consistently leads to higher output diversity. By further computing the correlation between output diversity and the thinking-space distance to English across non-English languages, we identify a clear positive relationship: thinking languages that are geometrically farther from English consistently achieve higher output diversity. These results demonstrate that sampling within thinking regions outside the English-dominant space can systematically mitigate output homogenization. We also evaluate output quality and find that thinking in non-English languages incurs only negligible degradation (§5.2).

Second, we further find that *Mixed-Language Sampling* yields additional gains in output diversity. This result indicates that sampling from distinct thinking regions induced by linguistic heterogeneity can further enhance output diversity beyond a single region. Further analysis reveals clear compositional effects among languages: while removing any single language has a relatively small impact on diversity, removing multiple languages leads to a substantially larger degradation (§5.3).

Third, we analyze the effects of the sampling number and temperature, and find that *Mixed-Language Sampling* exhibits a pronounced advantage over *Single-Language Sampling* when further scaling the sampling number, highlighting the role of linguistic heterogeneity in expanding the model's diversity ceiling (§5.4).

Finally, we extend our analysis to pluralistic alignment scenarios (§6). Our results show that *Mixed-Language Sampling* leads to broader coverage of cultural knowledge and values in LLMs, outperforming other sampling strategies, including English sampling, high-temperature decoding, explicit diversity requests, and multilingual prompting. These results highlight the practical utility of our findings in real-world applications.

Overall, our findings establish the *language of thought* as a novel and effective control axis for enhancing output diversity.

2 Related Work

Output Diversity of LLMs Many studies have shown that LLMs often exhibit limited output diversity (Padmakumar and He, 2024; Liang et al., 2024b; Luo et al., 2024; Giorgi et al., 2024). Output diversity evaluation typically considers lexical, syntactic, and semantic dimensions (Guo et al., 2024, 2025b; Lagzian et al., 2025), and employs tools such as Self-BLEU (Zhu et al., 2018) and Sentence-BERT (Reimers and Gurevych, 2019) to compute diversity metrics in NLG tasks (Guo et al., 2024).

Moreover, diversity is often evaluated alongside novelty and creativity in more complex generation settings (Zhang et al., 2025; Lagzian et al., 2025; Pépin et al., 2024; Ye et al., 2025; Tian et al., 2024). Recently, NOVELTYBENCH (Zhang et al., 2025) and INFINITY-CHAT (Jiang et al., 2025) were introduced to assess the ability of LLMs to produce distinct outputs in open-domain dialogue.

Existing approaches to improve output diversity include aggregating outputs from multiple LLMs (Liang et al., 2024a; Shur-Ofry et al., 2024), increasing prompt variation (Liang et al., 2024a; Lagzian et al., 2025; Wang et al., 2025a), and developing diversity-driven RLHF and SFT objectives (Li et al., 2025b; Sun et al., 2025). Unlike these approaches, our work explores the inherent multilingual properties of LLMs as a structural source of output diversity.

Multilingual Reasoning Recent LLMs are trained to perform explicit intermediate reasoning before producing final answers (Muennighoff et al., 2025; Zeng et al., 2025; DeepSeek-AI et al., 2025). Many studies have explored the multilingual generalization of LLM reasoning (Son et al., 2025; Yong et al., 2025; Wang et al., 2025b; Bajpai and Chakraborty, 2025; Qi et al., 2025; Tam et al., 2025; Khairi et al., 2025). Other work has investigated whether multilingualism can improve the performance (Li et al., 2025a; Gao et al., 2025) and efficiency (Ahuja et al., 2025; Chen et al., 2025a) of reasoning. However, none of these studies have examined whether multilingual thinking can enhance the output diversity of LLMs.

3 Language Geometry of Thinking Space

We first conduct a preliminary study to examine *whether different thinking languages induce structural differences in the model's thinking space.*

3.1 Thinking Language Control

All our investigations focus on reasoning-capable LLMs. Given an English input prompt, the model first performs intermediate thinking T, enclosed within `<think>...\think>`, and then generates the final output o, both in English by default.

To control the LLM to perform its intermediate thinking in a target language l, we follow existing multilingual reasoning techniques (Yong et al., 2025; Qi et al., 2025). Specifically, we insert a short prefix, "`Okay, the user is asking`"—translated into l— immediately after the `<think>` token, guiding the subsequent thinking process to be conducted in the target language. The translated prefixes, together with a sanity check of the language control, are provided in Appendix A.1.

3.2 Visualizing Multilingual Thinking Space

Collecting Hidden States Given a set of English input questions, we apply thinking language control to encourage the model to perform thinking in language l for each sample. For a single sample, let the thinking process consist of N tokens $\{t_i^{(l)}\}_{i=1}^N$, and let $h_{i,j}^{(l)}$ denote the hidden state of token $t_i^{(l)}$ at layer j. To obtain a compact representation of the model's thinking behavior, we first average hidden states across all thinking tokens within a sample, and then further average across all samples. This yields a single vector representation $h_j^{(l)}$ that summarizes the model's thinking behavior in language l at layer j. Repeating this process for all thinking languages produces a set of language-specific thinking representations at each layer.

PCA Visualization To visualize the geometry of multilingual thinking space, we first normalize all language representations using ℓ_2 normalization. Viewing English as the anchor, we then compute the cosine distance between each non-English language l and English at layer j as $d_j(l, \text{en}) = 1 - \cos\left(h_j^{(l)}, h_j^{(\text{en})}\right)$. Finally, we apply PCA to the centered representations to obtain a two-dimensional layout for visualization. In the resulting plot, PCA determines only the angular arrangement of languages, while the radial distance of each point is explicitly fixed to its cosine distance to English, i.e., $d_j(l, \text{en})$.

3.3 Observations

We select 14 non-English languages together with English that are officially supported by Qwen3-8B to analyze the multilingual thinking space of the model. Figure 1 shows the resulting geometry at several representative model layers.

Geometric Separation across Thinking Languages We first observe clear geometric separation among thinking representations induced by different thinking languages: representations corresponding to different languages tend to occupy separable regions in the model's thinking space. This separation holds consistently across model layers, including intermediate layers that are often