Language of Thought Shapes Output Diversity in Large Language Models

Anonymous ACL submission

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.

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 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 assumed to be relatively abstract and less language-specific (Pires et al., 2019). These observations indicate the presence of language-correlated geometric structure in the model's thinking space.

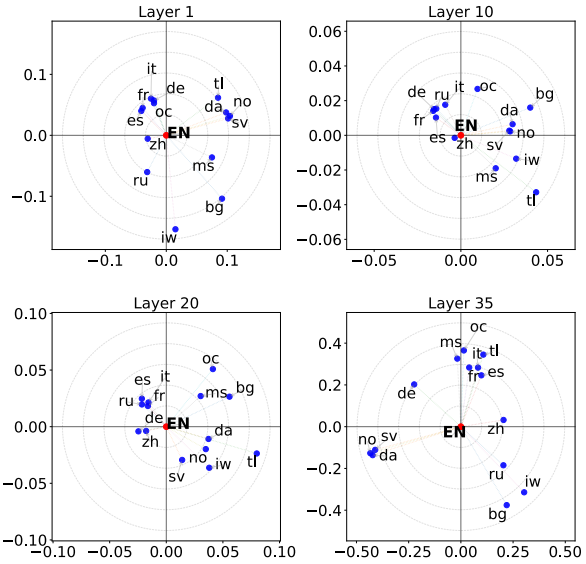

Figure 1: Language geometry of thinking space on Qwen3-8B, with different distance scales across layers for visualization purposes.

Varied Distances to English Thinking We further observe systematic variation in the geometric distance between non-English languages and English. Some languages (e.g., `zh`, `fr`, `es`, `de`) consistently appear closer to English, whereas others (e.g., `iw`, `bg`, `tl`) are embedded farther away. Overall, these results indicate that different languages of thought occupy distinct regions of the model's thinking space, with varied distances to English.

4 Repeated Sampling under Multilingual Thinking

In this and following sections, we further investigate *whether the thinking-space shifts induced by different languages of thought translate into greater output diversity*. In this section, we first introduce a controlled output setting and two repeated sampling strategies. The resulting outputs are used for diversity evaluation in Section 5.

4.1 Output Language Control

Although the model's intermediate thinking T is controlled to be conducted in a specific language and enclosed within `<think>...\think>` (Section 3.1), we further constrain the final output o to English to enable fair output diversity evaluation. This is achieved by inserting an additional English prefix immediately after `</think>`—`Let me provide my answer in English only:`— to guide the model to generate the final response in English. Only the English final outputs are collected for subsequent output diversity evaluation.

Appendix A.1 provides a sanity check indicating that both the thinking and output segments largely follow the intended language control.

4.2 Single-Language Sampling

Section 3.3 shows that different non-English languages occupy distinct thinking regions with varying distances from English. This motivates us to examine *whether switching to a thinking region away from English and performing repeated sampling within that region leads to increased output diversity.* To this end, we introduce the first repeated sampling strategy, *Single-Language Sampling*.

Given an English input, the model's intermediate thinking is constrained to a fixed thinking language l, while the final output is generated in English. We then sample the model M times under this fixed thinking language, and aggregate the resulting English outputs into a set $\mathcal{O}_l$ for diversity evaluation.

4.3 Mixed-Language Sampling

We further examine *whether sampling from distinct thinking regions induced by different languages can yield additional gains in output diversity*. This setting allows us to investigate the compositional effects of multiple thinking languages on output diversity. We thus introduce our second repeated sampling strategy, *Mixed-Language Sampling*.

Specifically, given an English input, we sample the model M times, each time controlling the model to perform intermediate thinking in a different language, while keeping the final output in English. The resulting outputs are aggregated into a set of outputs $\mathcal{O}_{\text{mixed}}$, on which the same diversity evaluation is conducted.

5 How Does Language of Thought Shape Output Diversity?

5.1 Experiment Settings

Datasets and Evaluation Metrics We evaluate output diversity on two benchmarks, NoveltyBench (Zhang et al., 2025) and Infinity-Chat (Jiang et al., 2025), each containing 100 open-ended questions without ground-truth answers. Given an input question, we sample the model M times to obtain a set of outputs $\mathcal{O}$ and evaluate their diversity and quality. Following the evaluation protocols of the original datasets, we consider two output diversity metrics and one output quality metric, as described below.

	en	it	ms	zh	ru	de	iw	bg	da	no	sv	es	tl	oc	fr	avg (non-en)
Distinct Score ↑																
Qwen3-8B	28.55	34.60	33.47	29.00	34.14	35.67	41.33	39.80	36.03	39.69	36.73	32.33	38.35	38.87	33.93	36.00
Qwen3-14B	26.20	30.67	29.23	28.80	31.40	28.93	36.87	32.13	30.13	34.55	32.33	29.73	32.68	33.26	29.53	31.45
Qwen3-32B	35.00	39.33	37.78	37.80	38.67	39.73	43.38	39.93	40.67	40.22	41.80	39.73	41.41	42.96	40.80	40.30
DeepSeek-14B	38.33	43.47	38.07	41.33	44.60	41.14	49.63	47.13	51.85	52.40	50.60	43.60	52.42	45.93	42.27	46.03
Similarity Score ↓																
Qwen3-8B	87.28	85.43	86.53	86.73	85.57	85.14	83.66	84.89	84.79	83.93	85.14	85.76	83.20	80.79	84.57	84.72
Qwen3-14B	87.82	86.68	87.30	86.89	87.20	87.78	85.04	86.94	86.81	86.17	86.46	87.35	87.36	85.72	87.19	86.78
Qwen3-32B	82.10	80.59	81.76	81.61	80.67	78.00	79.64	81.45	79.78	79.54	79.06	79.84	79.71	77.65	80.62	79.99
DeepSeek-14B	81.15	79.98	83.28	82.11	80.17	81.08	76.16	81.34	77.56	77.61	79.27	81.12	76.70	79.81	81.88	79.86
Output Quality ↑																
Qwen3-8B	96.82	95.86	95.72	95.53	96.11	96.69	95.53	96.04	95.09	95.00	96.82	95.72	95.70	95.59	95.40	95.80
Qwen3-14B	96.93	94.94	95.48	95.03	94.70	96.03	96.50	96.00	96.10	96.78	96.16	95.79	95.49	95.87	95.75	95.80
Qwen3-32B	97.36	96.08	95.85	96.22	95.36	94.47	95.57	97.07	95.52	96.87	95.96	94.97	96.04	96.19	94.26	95.70
DeepSeek-14B	95.84	94.75	93.94	94.71	93.69	93.27	89.17	94.52	92.95	92.60	93.66	94.93	90.73	95.45	95.80	93.60

Table 1: Distinct Score (%), Similarity Score (%), and Output Quality across models and thinking languages under *Single-Language Sampling* on NOVELTYBENCH. For each row, the best and worst language results are highlighted.

Metric 1: Distinct Score. We compute *Distinct Score* to measure the functional distinctiveness of $\mathcal{O}$ following Zhang et al. (2025). Specifically, the `deberta-v3-large-generation-similarity` model is used to sequentially judge whether two outputs are functionally equivalent. Each output o_i is compared with all previous outputs $\{o_1, \ldots, o_{i-1}\}$. If o_i is judged equivalent to any o_j $(j < i)$, it is assigned to the same equivalence class; otherwise, it forms a new class. The M outputs are thus clustered into C equivalence classes, and the *Distinct Score* is defined as C/M.

Metric 2: Similarity Score. We also compute the *Similarity Score* following Jiang et al. (2025), which captures semantic similarity among outputs in $\mathcal{O}$. Sentence-level embeddings are first obtained for all generated outputs, and cosine similarity is computed for all output pairs. The final score is obtained by averaging cosine similarities across all pairs. We use `Qwen3-Embedding-8B` for embedding extraction.

Metric 3: Output Quality. To assess whether improvements in output diversity come at the cost of output quality, we evaluate the quality of responses in $\mathcal{O}$ using `gpt-4o-mini`, with scores ranging from 0 to 100. The evaluation considers two dimensions: instruction adherence and overall response quality. Details of the evaluation prompting are provided in Appendix A.2.

Languages and LLMs We conduct experiments on the thinking mode of the Qwen3 family (Yang et al., 2025) with model sizes 8B, 14B, and 32B, as well as DeepSeek-R1-Distill-Qwen-14B (DeepSeek-14B) (DeepSeek-AI et al., 2025). We select 15 thinking languages for evaluation: `en`, `it`, `ms`, `zh`, `ru`, `de`, `iw`, `bg`, `da`, `no`, `sv`, `es`, `tl`, `oc`, and `fr`, from the supported languages of the tested models.

Sampling Parameters Unless otherwise specified, the decoding temperature is set to 0.6. For fair comparison across sampling strategies, the number of samples M is set equal to the number of thinking languages, i.e., $M = 15$.

5.2 Results on Single-Language Sampling

Main Diversity Results Table 1 summarizes the output diversity results on NOVELTYBENCH. On average, switching the thinking language from English to non-English languages yields an improvement of 5.3 to 7.7 points in *Distinct Score* and a reduction of 1.04 to 2.56 points in *Similarity Score*. These results suggest that sampling from thinking regions outside the English-dominant space provides a systematic advantage in output diversity.

We also observe substantial variation in output diversity across thinking languages. Besides `en`, some languages such as `ms` and `zh` consistently exhibit lower diversity, whereas others, including `iw`, `no`, and `oc`, achieve substantially higher diversity across models and metrics. In some cases, individual languages lead to particularly large gains. For example, thinking in `iw` on Qwen3-8B improves the *Distinct Score* by 12.78 points compared to `en`. Taken together with the geometric findings from Section 3.3, these results highlight the strong potential of specific thinking languages— especially those farther from English in the thinking space— for enhancing output diversity.

Correlation with Thinking Distance to English We further examine the relationship between the geometric properties of the thinking space and output diversity. For each language l, we compute its

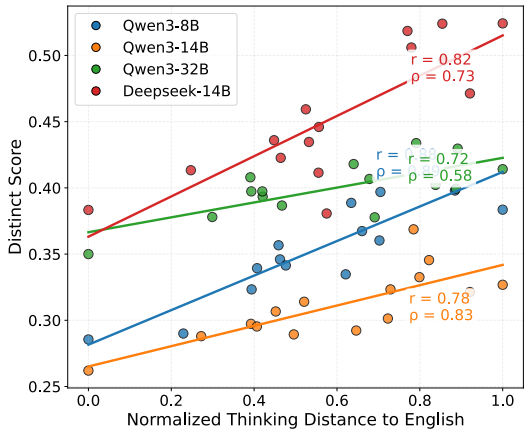

Figure 2: Correlation between the Distinct Score and the thinking distance to English across languages. Pearson's r and Spearman's ρ are reported for each model. Distinct Scores are obtained under *Single-Language Sampling* on NOVELTYBENCH. Thinking distances are normalized to the range $[0, 1]$ for visualization.

Model	S-en	S-non-en avg	S-best	Mixed
NOVELTYBENCH				
Qwen3-8B	28.55	36.00	41.33	**43.73**
Qwen3-14B	26.20	31.45	36.87	**38.00**
Qwen3-32B	35.00	40.30	43.38	**46.53**
DeepSeek-14B	38.33	46.03	**52.42**	52.07
INFINITY-EVAL				
Qwen3-8B	20.67	22.54	24.51	**28.13**
Qwen3-14B	20.40	22.60	**27.07**	26.73
Qwen3-32B	27.00	27.52	28.66	**31.47**
DeepSeek-14B	25.27	31.84	**39.61**	35.33

Table 2: Distinct score (%) comparison of *Mixed-Language Sampling* and *Single-Language Sampling* on NOVELTYBENCH and INFINITY-CHAT. **Bold** indicates the best-performing sampling setting for each model and benchmark.

thinking distance to English, $d(l, \text{en})$, by averaging the layer-wise distances $d_j(l, \text{en})$ across all model layers (Section 3.2), where English has distance zero. We then analyze the correlation between this thinking distance and the output diversity achieved under *Single-Language Sampling* across languages. Figure 2 reports the Pearson and Spearman correlations on NOVELTYBENCH, with output diversity measured by the *Distinct Score*.

We observe a strong positive correlation across different models, with Pearson's r ranging from 0.72 to 0.88 and Spearman's ρ ranging from 0.58 to 0.89. These results corroborate our earlier observations, indicating that the distance to English in the thinking space is informative of the output diversity achievable under *Single-Language Sampling*. More specifically, languages that are geometrically farther from English tend to correspond to more distinct thinking regions, and repeated sampling within such regions is associated with higher output diversity.

Output Diversity vs. Quality Table 1 also reports the output quality results. We observe a mild trade-off between output diversity and quality. While English generally achieves higher output quality, there is no clear pattern in which languages with the highest output diversity consistently suffer the lowest output quality. In some cases, specific languages such as sv and oc achieve strong performance on both dimensions. Overall, thinking in non-English languages results in only a modest decrease of 1.02 to 2.24 points in output quality.

Appendix A.3 provides results on INFINITY-CHAT, which also exhibits similar patterns.

5.3 Results on Mixed-Language Sampling

Comparison with Single-Language Sampling Table 2 compares *Mixed-Language Sampling* with three *Single-Language Sampling* settings: English sampling (S-en), the average performance over non-English sampling (S-non-en avg), and the best-performing single-language sampling (S-best). Across both benchmarks, *Mixed-Language Sampling* consistently improves output diversity over S-en and S-non-en avg.

Moreover, *Mixed-Language Sampling* often matches or even exceeds the performance of the S-best setting. These results indicate that *Mixed-Language Sampling* provides a robust strategy for improving output diversity without requiring prior knowledge of which single language performs best. This advantage arises from the structural differences among languages in the thinking space (Section 3.3): sampling from multiple distinct thinking regions and aggregating the resulting outputs exploits the compositional effects of different languages.

Results based on the *Similarity Score* are reported in Appendix A.4 and show the same trend.

Compositional Effects of Different Languages To further explore the compositional effects of different languages in *Mixed-Language Sampling*, we conduct an ablation study on Qwen3-8B by progressively removing k languages from *Mixed-Language Sampling* ($k = 1, \ldots, 5$). For each value of k, we enumerate all possible combinations of

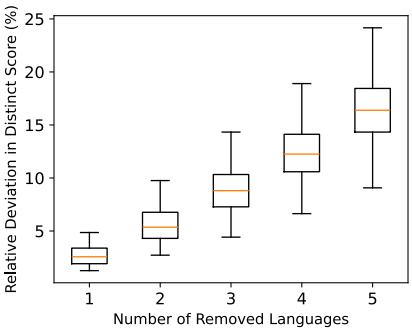

Figure 3: Relative deviation in Distinct Score under the removal of k languages in *Mixed-Language Sampling*.

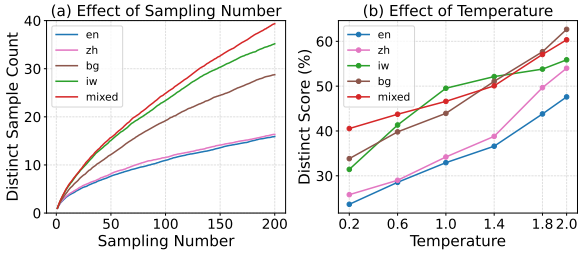

Figure 4: Effects of sampling parameters on output diversity. (a) Distinct sample count as a function of the sampling number M at a fixed temperature (0.6). (b) Distinct Score (%) under different temperatures with a fixed sampling number ($M = 15$).

language removal and measure the relative deviation of the *Distinct Score* from the original result, to quantify the effect of language removal.

Figure 3 shows the relative deviation in *Distinct Score*. We first observe that removing a single language leads to only a small change (2.7% on average), indicating that *Mixed-Language Sampling* does not rely on any individual language to achieve its diversity gains. However, as k increases, the diversity degradation grows rapidly and in a super-linear manner. This suggests that the contributions of different languages are not merely additive; instead, languages provide complementary diversity benefits through their joint participation. Together, these results demonstrate that output diversity under *Mixed-Language Sampling* emerges from the compositional interaction of multiple languages, rather than from any single dominant language.

5.4 Other Analysis

Two parameters are important in repeated sampling: the sampling number M and the temperature. By default, we set $M = 15$ and the temperature to 0.6. In this section, we vary these parameters using Qwen3-8B to examine their effects on two sampling strategies. For *Single-Language Sampling*, we select four representative languages for analysis: `en` and `zh` (lower-performing), and `bg` and `iw` (higher-performing).

5.4.1 Scaling Sampling Number

We first vary the sampling number M from 1 to 200 while keeping the temperature fixed at 0.6. For *Mixed-Language Sampling*, we utilize the full language pool supported by Qwen3 (approximately 100 languages) and randomly select one language as the thinking language for each sampling. Rather than *Distinct Score* C/M, Figure 4(a) directly reports the number of distinct samples C.

Across all settings, we observe that the growth of C slows down as M increases, suggesting the existence of an upper bound on achievable output diversity. However, *Mixed-Language Sampling* exhibits a much slower saturation rate compared to *Single-Language Sampling*. As M increases, its advantage over all *Single-Language Sampling* settings continues to widen.

This behavior indicates that *Mixed-Language Sampling* effectively expands the model's diversity ceiling. Such an expansion arises from the increased coverage of distinct thinking regions enabled by linguistic heterogeneity. Although we explore over 100 languages, further unlocking the benefits of linguistic diversity remains an interesting direction for future work.

5.4.2 Varying Temperatures

We next fix the sampling number M at 15 and vary the temperature over $\{0.2, 0.6, 1.0, 1.4, 1.8, 2.0\}$. The results are shown in Figure 4(b).

We observe a compositional effect between the language of thought and temperature scaling: while switching the language of thought from English to other languages already improves output diversity, increasing the temperature further yields additional gains. Moreover, the advantages of non-English and mixed-language sampling become especially evident. For instance, *Mixed-Language Sampling* at temperature 1.0 achieves a level of diversity comparable to English sampling at temperature 2.0.

6 Application: Pluralistic Alignment

In this section, we further investigate the practical utility of *Mixed-Language Sampling*, given its distinct advantages. Specifically, we focus on pluralistic alignment scenarios, where model responses are expected to reflect cultural pluralism.

Model	Method	Blend	WVS
Qwen3-8B	ES	67.9	40.0
	HT	68.0 (+0.1)	39.0 (-1.0)
	RD	73.3 (+5.4)	52.7 (+12.7)
	MP	76.1 (+9.2)	52.0 (+12.0)
	MLS	**76.7 (+8.8)**	**59.0 (+19.0)**
Qwen3-14B	ES	66.7	31.6
	HT	67.1 (+0.4)	32.7 (+1.1)
	RD	68.4 (+1.7)	38.0 (+6.4)
	MP	72.7 (+6.0)	45.1 (+13.5)
	MLS	**74.0 (+7.3)**	**48.4 (+16.8)**
Qwen3-32B	ES	67.5	40.1
	HT	69.2 (+1.7)	43.6 (+3.5)
	RD	72.8 (+5.3)	**53.4 (+13.3)**
	MP	73.4 (+5.9)	46.1 (+6.0)
	MLS	**74.6 (+7.1)**	50.4 (+10.3)
DeepSeek-8B	ES	78.6	52.3
	HT	80.7 (+2.1)	60.1 (+7.8)
	RD	78.6 (+0.0)	54.7 (+2.4)
	MP	80.6 (+2.0)	67.2 (+14.9)
	MLS	**83.0 (+4.4)**	**73.3 (+21.0)**

Table 3: Cultural pluralism performance (entropy normalized to 0–100). Methods: ES (English Sampling), HT (High Temperature), RD (Request Diversity), MP (Multilingual Prompting), MLS (Mixed-Language Sampling). Parentheses show absolute gains/losses relative to ES within each model and benchmark. **Bold** indicates the best-performing setting per model and benchmark.

6.1 Settings

Data We consider two types of cultural pluralism: *cultural knowledge* and *cultural values*, evaluated using the BLEND (Myung et al., 2024) and WVS (Haerpfer et al., 2022) datasets, respectively. Both datasets consist of multiple-choice questions.

Evaluation Following Wang et al. (2025a), for each cultural question, we perform repeated sampling to obtain M responses and measure cultural pluralism based on the resulting output distribution. For BLEND, where each option is associated with one or more countries, we map the sampled outputs to countries and compute the entropy over the country distribution. For WVS, we directly compute the entropy over the output distribution, which characterizes the diversity of value orientations reflected in the model responses.

LLMs Experiments are conducted on Qwen3-8B, Qwen3-14B, Qwen3-32B, and DeepSeek-R1-Distill-Llama-8B (DeepSeek-8B), with temperature set to 0.6 by default.

Sampling Strategies We compare the following sampling strategies: (1) *English Sampling*, where the language of thought is English; (2) *High Temperature*, where the temperature is increased to 1.0 while keeping English as the thinking language; (3) *Request Diversity*, where the model is explicitly instructed to generate novel responses; (4) *Multilingual Prompting* (Wang et al., 2025a), where each cultural question is translated into the same 15 languages used in previous experiments; and (5) *Mixed-Language Sampling*, where the language of thought varies across the same 15 languages used in previous experiments.

The sampling number M is set to 15 for all strategies. For *Multilingual Prompting* and *Mixed-Language Sampling*, each language is sampled once.

Additional details on the datasets, evaluation protocols, and baselines are provided in Appendix A.5.

6.2 Results

The results in Table 3 clearly demonstrate the practical advantage of *Mixed-Language Sampling* for pluralistic alignment. Across benchmarks and models, *Mixed-Language Sampling* consistently achieves the highest cultural pluralism performance, enabling LLMs to reflect more diverse cultural knowledge and value orientations.

In contrast, simply increasing the temperature, explicitly requesting diversity, or using multilingual inputs does not yield improvements comparable to *Mixed-Language Sampling*. These results highlight the practical value of diversifying the language of thought as a means of more fully exploiting the model's thinking space for pluralistic alignment.

7 Conclusion

In this paper, we establish that controlling the *language of thought* provides a structural source of output diversity in LLMs. We find that switching the thinking language from English to non-English languages consistently increases output diversity, with stronger gains observed for languages farther from English in the thinking space. We further demonstrate that aggregating samples across multiple thinking languages yields additional diversity improvements through their compositional effects, and that scaling the sampling number with linguistic heterogeneity effectively expands the model's diversity ceiling. Finally, we show that these findings translate into broader coverage of cultural knowledge and values of LLMs in pluralistic alignment.

8 Limitations

This work has two main limitations.

First, while we observe a positive correlation between the geometric distance of non-English thinking languages from English and the output diversity achieved under repeated sampling, there are still several open questions that are not addressed in this work. For example, many cross-lingual alignment methods explicitly aim to align non-English representations toward English. An important question is whether such alignment procedures may inadvertently reduce the output diversity associated with aligned non-English languages, and if so, what mechanisms or strategies could mitigate this effect. Addressing these questions would require controlled interventions or additional training on the model, which we leave for future work.

Second, although we demonstrate the practical utility of our findings in pluralistic alignment settings, our evaluation relies on output entropy as a proxy for cultural pluralism. This experimental setup remains an abstraction of real-world deployment scenarios. In practice, pluralistic alignment often requires models to align with multiple specific and context-dependent cultural values under explicit constraints. The sampling strategies studied in this work would likely need to be further adapted—e.g., by incorporating culturally contextualized language-of-thought routing—to be effective in such settings, which we leave for future investigation.

References

Sanchit Ahuja, Praneetha Vaddamanu, and Barun Patra. 2025. Efficientxlang: Towards improving token efficiency through cross-lingual reasoning. *CoRR*, abs/2507.00246.

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai Alkhamissi, and Mona T. Diab. 2024. Investigating cultural alignment of 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 12404–12422. Association for Computational Linguistics.

Prasoon Bajpai and Tanmoy Chakraborty. 2025. Multilingual test-time scaling via initial thought transfer. *CoRR*, abs/2505.15508.

Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.

Kang Chen, Mengdi Zhang, and Yixin Cao. 2025a. Less data less tokens: Multilingual unification learning for efficient test-time reasoning in llms. *CoRR*, abs/2506.18341.

Xiaoyang Chen, Xinan Dai, Yu Du, Qian Feng, Naixu Guo, Tingshuo Gu, Yuting Gao, Yingyi Gao, Xudong Han, Xiang Jiang, Yilin Jin, Hongyi Lin, Shisheng Lin, Xiangnan Li, Yuante Li, Yixing Li, Zhentao Lai, Zilu Ma, Yingrong Peng, and 12 others. 2025b. Deepmath-creative: A benchmark for evaluating mathematical creativity of large language models. *CoRR*, abs/2505.08744.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025. Could thinking multilingually empower LLM reasoning? *CoRR*, abs/2504.11833.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle H. Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 7174–7188. Association for Computational Linguistics.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. 2025a. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 5888–5899. ACM.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025b. Benchmarking linguistic diversity of large language models. *Trans. Assoc. Comput. Linguistics*, 13:1507–1526.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic

diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3589–3604. Association for Computational Linguistics.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bjorn Puranen. 2022. World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 12(10):8.

Simeng Han, Stephen Xia, Grant Zhang, Howard Dai, Chen Liu, Lichang Chen, Hoang Huy Nguyen, Hongyuan Mei, Jiayuan Mao, and R. Thomas McCoy. 2025. Creativity or brute force? using brainteasers as a window into the problem-solving abilities of large language models. *CoRR*, abs/2505.10844.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *CoRR*, abs/2510.22954.

Ammar Khairi, Daniel D'souza, Ye Shen, Julia Kreutzer, and Sara Hooker. 2025. When life gives you samples: The benefits of scaling up inference compute for multilingual llms. *CoRR*, abs/2506.20544.

Anatoliy V. Kharkhurin, Valeriya Koncha, and Morteza Charkhabi. 2023. The effects of multilingual and multicultural practices on divergent thinking. implications for plurilingual creativity paradigm. *Bilingualism: Language and cognition*, 26(3):592–609.

Arash Lagzian, Srinivas Anumasa, and Dianbo Liu. 2025. Multi-novelty: Improve the diversity and novelty of contents generated by large language models via inference-time multi-views brainstorming. *CoRR*, abs/2502.12700.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. 2025a. The impact of language mixing on bilingual llm reasoning. *CoRR*, abs/2507.15849.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025b. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024a. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17889–17904. Association for Computational Linguistics.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2024b. Mapping the increasing use of llms in scientific papers. *CoRR*, abs/2404.01268.

Jiaming Luo, Colin Cherry, and George F. Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Trans. Assoc. Comput. Linguistics*, 12:355–371.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *CoRR*, abs/2501.19393.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21, 2024*, pages 226–235. Association for Computational Creativity (ACC).

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2025. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models. *CoRR*, abs/2507.20956.

Antoine Bellemare Pépin, François Lespinasse, Philipp Thölke, Yann Harel, Kory W. Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *CoRR*, abs/2405.13012.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001. Association for Computational Linguistics.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S. Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy. *CoRR*, abs/2505.22888.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2024. Liveideabench: Evaluating llms' divergent thinking for scientific idea generation with minimal context. *CoRR*, abs/2412.17596.

Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim, and Yonatan Belinkov. 2024. Growing a tail: Increasing output diversity in large language models. *CoRR*, abs/2411.02989.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14333–14368. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Haoran Sun, Yekun Chai, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. 2025. Curiosity-driven reinforcement learning from human feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23517–23534. Association for Computational Linguistics.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *CoRR*, abs/2505.17407.

Selim F. Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. LLM-TOPLA: efficient LLM ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 11951–11966. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 326–346. Association for Computational Linguistics.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2024. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5303–5324. Association for Computational Linguistics.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025a. Multilingual prompting for improving LLM generation diversity. *CoRR*, abs/2505.15229.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in 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 6349–6384. Association for Computational Linguistics.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Polymath: Evaluating mathematical reasoning in multilingual contexts. *CoRR*, abs/2504.18428.

Benjamin Lee Whorf. 2012. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Grace Guiling Wang. 2025. Assessing the creativity of llms in proposing novel solutions to mathematical problems. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25687–25696. AAAI Press.

Zheng-Xin Yong, Muhammad Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji.

Language	Prefix inserted after <think> token
English (en)	Okay, the user is asking
Italian (it)	Va bene, l'utente sta chiedendo
Malay (ms)	Baiklah, pengguna sedang bertanya
Chinese (zh)	好的，用户在问
Russian (ru)	Хорошо, пользователь спрашивает
German (de)	Okay, der Benutzer fragt
Hebrew (iw)	בסדר, המשתמש שואל
Bulgarian (bg)	Добре, потребителят пита
Danish (da)	Okay, brugeren spørger
Norwegian (no)	Greit, brukeren spør
Swedish (sv)	Okej, användaren frågar
Spanish (es)	De acuerdo, el usuario pregunta
Tagalog (tl)	Sige, nagtatanong ang gumagamit
Occitan (oc)	Bon, l'utilizaire demanda
French (fr)	D'accord, l'utilisateur demande

Figure 5: Prefix translations used for Thinking Language Control.

2025. Crosslingual reasoning through test-time scaling. *CoRR*, abs/2505.05408.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4651–4665. Association for Computational Linguistics.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. Noveltybench: Evaluating language models for humanlike diversity. *CoRR*, abs/2504.05228.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

A Appendix

A.1 Language Control Details

Figure 5 presents the translated prefixes used for Thinking Language Control across 15 languages. By inserting the corresponding prefix immediately after the <think> token, the model is guided to conduct its intermediate thinking in the target language.

Model	Lang	Think-Target (%)	Output-EN (%)
Qwen3-8B	en	100.00	98.29
	non-en	99.88 ± 0.25	98.28 ± 1.31
Qwen3-14B	en	100.00	98.37
	non-en	99.57 ± 1.45	99.50 ± 0.35
Qwen3-32B	en	100.00	100.00
	non-en	99.54 ± 1.47	98.61 ± 0.69
DeepSeek-14B	en	100.00	96.10
	non-en	98.70 ± 2.57	95.32 ± 1.51

Table 4: Sanity-check verification of thinking and output language control. Results for English thinking are reported individually, while results for non-English thinking are averaged over multiple languages and reported as mean ± standard deviation.

Combined with Output Language Control, the model is guided to thinking in a specified language while producing English responses. As a sanity check, we apply an off-the-shelf language identification tool[1] to the thinking content within the <think> ... </think> span, as well as to the final output following </think>.

Table 4 summarizes the averaged results on NOVELTYBENCH and INFINITY-CHAT. Across models, the thinking segments are predominantly detected as the target thinking language, and the output segments are predominantly detected as English. Although language identification may introduce some noise, these results indicate that the intended language control signals are largely reflected in the generated text.

A.2 Output Quality Evaluation Details

Table 5 shows the complete prompt used for output quality evaluation. The total quality score is computed as the sum of the two evaluation dimensions. For each task instance, all sampled responses are evaluated independently, and we report the average quality score across samples.

A.3 Additional Results on Single-Language Sampling

Table 6 reports the results of *Single-Language Sampling* on INFINITY-CHAT. Overall, we observe several consistent trends that align with the main findings. First, switching the language of thought from English to non-English languages generally leads to higher output diversity across models, as

[1] https://github.com/pemistahl/lingua-py

Output Quality Evaluation Prompt

You are an evaluator assessing the quality of a single response to a task instruction.
You will be given:
(1) A task instruction
(2) A response

Evaluate the response along the following two dimensions:

1. Instruction Adherence (0–50)
To what extent does the response follow the task instruction?
Note that if the response explicitly refuses to perform the task, this should NOT be penalized.
You only need to judge the degree to which the response is relevant to the task instruction.
2. Response Quality (0–50)
Assess the overall quality of the response in terms of clarity, fluency, and grammatical correctness.
Scoring:
- Each dimension should be scored from 0 to 50 (integer only).
- Total Score = sum of the two dimensions (0–100).

Output format (strict JSON only):

```
{
  "Instruction Adherence": <score>,
  "Response Quality": <score>,
  "Total Score": <score>
}
```

Table 5: Prompt template used for output quality evaluation with `gpt-4o-mini`.

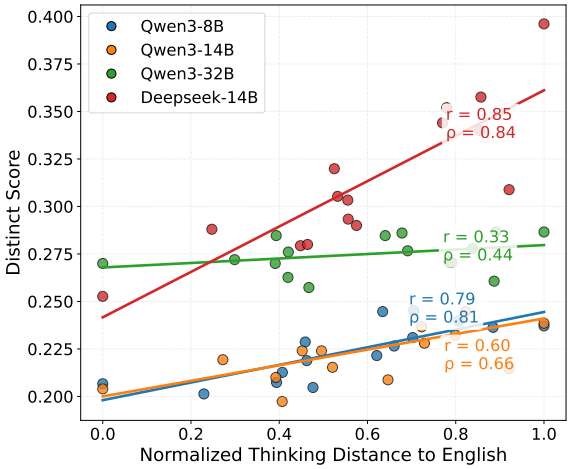

Figure 6: Correlation between the Distinct Score and the thinking distance to English across languages. Pearson's r and Spearman's ρ are reported for each model. Distinct Scores are obtained under *Single-Language Sampling* on INFINITY-CHAT. Thinking distances are normalized to the range $[0, 1]$ for visualization.

reflected by higher *Distinct Score* and lower *Similarity Score*. Second, there exists notable variation across thinking languages: languages such as `en`, `ru`, and `fr` tend to exhibit lower diversity, whereas others, including `iw`, `tl`, and `oc`, consistently achieve higher diversity. Finally, we do not observe a clear or systematic trade-off between output diversity and quality across languages. Several non-English languages achieve improved diversity while maintaining comparable output quality.

Figure 6 further reports the correlation between output diversity and the thinking distance to English across languages on INFINITY-CHAT. Consistent with our main results, we observe a strong positive correlation for most models. This result further corroborates that repeated sampling within thinking regions farther from English is associated with higher output diversity.

A.4 Additional Results on Mixed-Language Sampling

Table 7 compares *Mixed-Language Sampling* with three *Single-Language Sampling* settings using the *Similarity Score*. Consistent with the main results, *Mixed-Language Sampling* consistently outperforms S-en and S-non-en avg, and in several cases matches or exceeds the S-best setting. This shows that its advantage lies in improving diversity without requiring the selection of a single best-performing language.

A.5 Culture Evaluation Details

Datasets For BLEND, we extract the set of unique questions from the original large-scale dataset and merge all answer options into each question, resulting in a multiple-choice dataset with 402 questions. For WVS, the original dataset contains 290 questions. We remove 8 questions without predefined options, yielding a final set of 282 multiple-choice questions.

Evaluation Protocols In BLEND, each answer option is associated with one or more countries. For each sampled response, we extract the selected option and increment the count of its associated country (or countries). Let $p(c)$ denote the empirical distribution over countries aggregated from M samples. Cultural pluralism is measured as the normalized entropy:

$$H_{\text{Blend}} = \frac{-\sum_c p(c) \log p(c)}{\log |C|}$$

where C denotes the set of all countries appearing in the answer options for the question. The reported results are averaged over all questions.

In WVS, each sampled response corresponds to a discrete value option. Let $p(o)$ denote the

	en	it	ms	zh	ru	de	iw	bg	da	no	sv	es	tl	oc	fr	avg (non-en)
Distinct Score ↑																
Qwen3-8B	20.67	21.89	22.15	20.13	20.47	22.87	23.98	23.64	23.10	24.51	22.65	20.73	23.71	24.47	21.27	22.54
Qwen3-14B	20.40	22.40	20.88	21.93	21.53	22.40	27.07	21.47	23.67	24.47	22.80	21.00	23.85	23.23	19.73	22.60
Qwen3-32B	27.00	27.60	27.67	27.20	25.73	26.27	27.05	26.07	28.60	27.78	28.47	28.47	28.66	28.66	27.00	27.52
DeepSeek-14B	25.27	30.53	29.00	28.80	29.33	30.33	35.76	30.88	34.40	34.00	35.20	27.93	39.61	31.99	28.00	31.84
Similarity Score ↓																
Qwen3-8B	89.05	88.69	88.80	88.80	89.30	87.83	87.36	88.09	88.12	87.47	88.30	88.75	88.26	86.78	88.64	88.23
Qwen3-14B	89.53	88.89	89.13	88.50	89.36	89.12	87.77	88.83	88.53	88.18	88.60	89.36	88.81	88.37	89.58	88.79
Qwen3-32B	85.24	81.97	84.98	82.89	84.27	76.49	86.22	85.52	82.54	84.10	79.24	80.83	85.72	83.77	82.31	82.92
DeepSeek-14B	85.97	83.16	85.52	85.74	84.09	83.06	79.11	83.31	80.85	80.15	82.64	85.46	79.30	83.11	85.19	82.91
Output Quality ↑																
Qwen3-8B	96.82	95.86	95.72	95.53	96.11	96.69	95.53	96.04	95.09	95.00	96.82	95.72	95.70	95.59	95.40	95.77
Qwen3-14B	96.93	94.94	95.48	95.03	94.70	96.03	96.50	96.00	96.10	96.78	96.16	95.79	95.49	95.87	95.75	95.76
Qwen3-32B	97.36	96.08	95.85	96.22	95.36	94.47	95.57	97.07	95.52	96.87	95.96	94.97	96.04	96.19	94.26	95.74
DeepSeek-14B	88.46	89.45	88.99	89.44	90.71	86.79	86.51	80.12	87.24	82.13	85.06	87.52	87.13	83.99	90.07	86.80

Table 6: Distinct Score (%), Similarity Score (%), and Output Quality across models and thinking languages under *Single-Language Sampling* on INFINITY-CHAT. For each row, the best and worst language results are highlighted.

Model	S-en	S-non-en avg	S-best	Mixed
NOVELTYBENCH				
Qwen3-8B	87.28	84.72	**80.79**	82.84
Qwen3-14B	87.82	86.78	**85.04**	85.29
Qwen3-32B	82.10	79.99	**77.65**	79.44
DeepSeek-14B	81.15	79.86	**76.16**	77.64
INFINITY-CHAT				
Qwen3-8B	89.05	88.23	86.78	**86.47**
Qwen3-14B	89.53	88.79	**87.77**	87.87
Qwen3-32B	85.24	82.92	**76.49**	80.29
DeepSeek-14B	85.97	82.91	**79.11**	82.15

Table 7: Similarity score (%) comparison of *Mixed-Language Sampling* and *Single-Language Sampling* on NOVELTYBENCH and INFINITY-CHAT. **Bold** indicates the best-performing sampling setting for each model and benchmark.

empirical distribution over predicted options across M samples. Cultural pluralism is defined as the normalized entropy:

$$H_{\text{WVS}} = \frac{-\sum_o p(o) \log p(o)}{\log |O|}$$

where O denotes the set of possible value options for the question. The reported results are averaged over all questions.

Baselines The *Request Diversity* baseline appends the following sentence to the original instruction: *"Please try to provide a novel answer."*

For *Multilingual Prompting*, we use Google Translate to translate each original question from English into the same set of 14 non-English languages used in the main experiments.