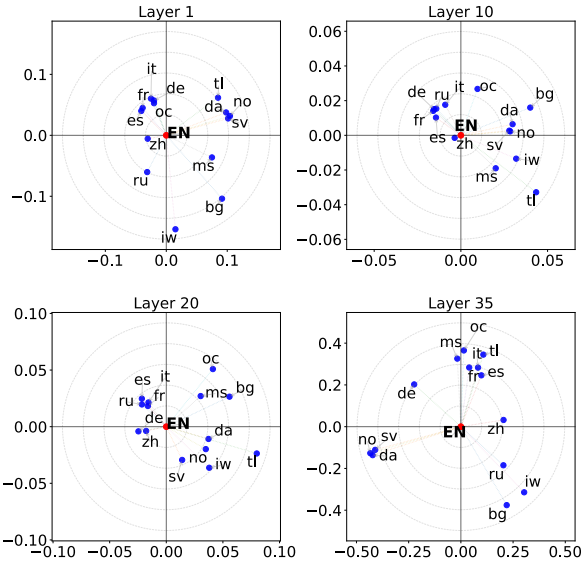

Figure 1: Language geometry of thinking space on Qwen3-8B, with different distance scales across layers for visualization purposes.

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.

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

	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.

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.

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.

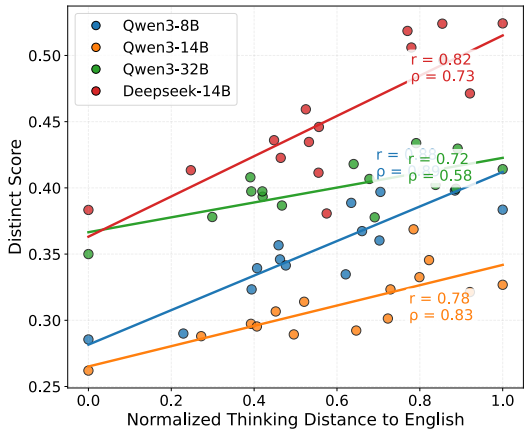

Figure 2: Correlation between the Distinct Score and the thinking distance to English across languages. Pearson's r and Spearman's ρ are reported for each model. Distinct Scores are obtained under *Single-Language Sampling* on NOVELTYBENCH. Thinking distances are normalized to the range $[0, 1]$ for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

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 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.

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.

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*,