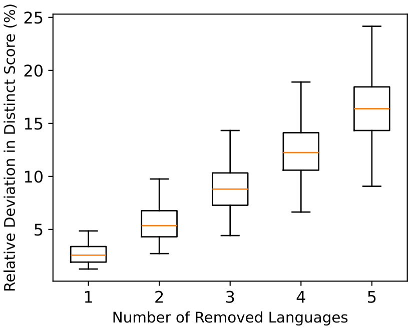

Figure 3: Relative deviation in Distinct Score under the removal of k languages in *Mixed-Language Sampling*.

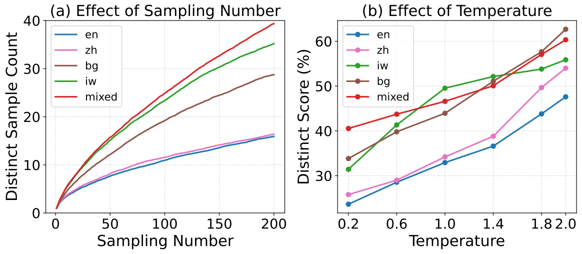

Figure 4: Effects of sampling parameters on output diversity. (a) Distinct sample count as a function of the sampling number M at a fixed temperature (0.6). (b) Distinct Score (%) under different temperatures with a fixed sampling number ($M = 15$).

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 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.

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 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.

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. Overreliance 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.