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.

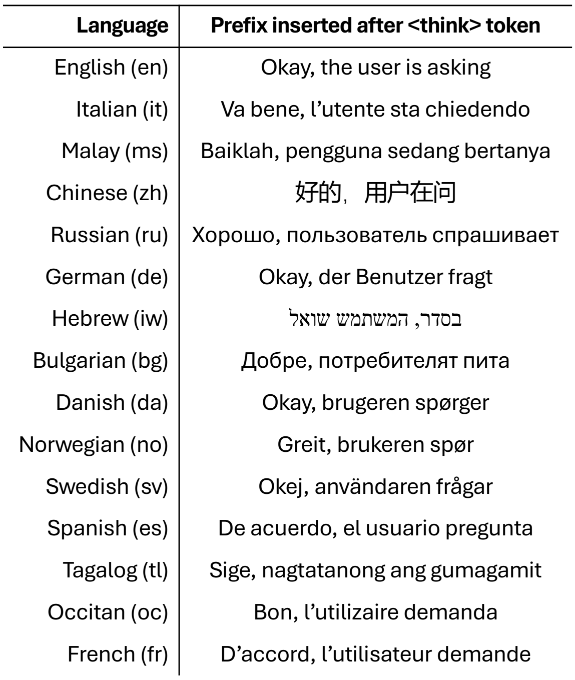
Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 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,*



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.

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. 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.

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.

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.

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.

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