Towards Inclusive AI: Advancing Multilingual Large Language Models

Wenxuan Zhang

Information Systems Technology and Design Pillar
Singapore University of Technology and Design
wxzhang@sutd.edu.sg

Large language models (LLMs) have achieved remarkable progress in recent years, yet their development remains heavily centered on a handful of high-resource languages such as English and Chinese. This imbalance raises fundamental scientific and societal questions: *How multilingual are today's models? What challenges arise when applying or extending them to a broad range of other languages?* And *how can we design systems that respect both linguistic and cultural diversity?* My work approaches these questions systematically, treating **multilingual LLMs** as a central lens through which to explore evaluation, safety, enhancement, and the cognitive frontiers of language models.

The first part of the talk will focus on evaluation. Through projects such as M3Exam (Zhang et al. 2023) and subsequent studies, we have shown that current LLMs often underperform in low-resource and structurally diverse languages, revealing significant performance disparities. More recently, our work on disentangling language and culture (Ying et al. 2025) provides new tools to probe the extent to which models capture core linguistic competence versus culturally specific knowledge. These studies point to a deeper need for evaluation that recognizes multilingualism as a multidimensional phenomenon rather than a simple extension of monolingual capabilities.

The second part examines safety in multilingual settings. Our research identified one of the first multilingual jailbreak phenomena (Deng et al. 2024), showing that adversarial behaviors are more easily triggered in underrepresented languages. We further proposed MPO (Zhao et al. 2025), a framework for aligning models consistently across languages through minimizing reward discrepancies. Together with recent work on uncovering safety-specific neurons within models, these efforts chart a pathway to understand multilingual alignment mechanisms and the vulnerabilities that emerge when safety protocols are uneven.

The third part highlights methods for enhancing multilingual capabilities. Our work on analyzing the internal mechanisms of multilingual LLMs (Zhao et al. 2024) reveals language-specific parameters and offers new strategies for efficient enhancement. Alongside these methodological advances, the SeaLLMs family (Zhang* et al. 2024) provides one of the first large-scale open-source LLMs for Southeast Asian languages, later expanded into SeaLLMs-Audio and Babel as broader open-source efforts to support multilingual, multimodal, and globally inclusive AI.

Finally, I will discuss emerging directions beyond language. Our recent study on abstract thought (Chen et al. 2025) shows that LLMs exhibit thinking abilities beyond specific linguistic forms. This raises the prospect of moving from multilingual to multicultural models: systems that not only process language but also adapt to social and cultural contexts. Looking ahead, this trajectory envisions a shift from surface-level text handling to deeper, contextually grounded, and socially-aware intelligence.

References

Chen, Y.; Zhao, Y.; Zhang, Y.; Zhang, A.; Kawaguchi, K.; Joty, S.; Li, J.; Chua, T.; Shieh, M. Q.; and Zhang, W. 2025. The Emergence of Abstract Thought in Large Language Models Beyond Any Language. *CoRR*, abs/2506.09890.

Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2024. Multilingual Jailbreak Challenges in Large Language Models. In *ICLR 2024*.

Ying, J.; Tang, W.; Zhao, Y.; Cao, Y.; Rong, Y.; and Zhang, W. 2025. Disentangling Language and Culture for Evaluating Multilingual Large Language Models. In *ACL 2025*.

Zhang, W.; Aljunied, M.; Gao, C.; Chia, Y. K.; and Bing, L. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *NeurIPS 2023*.

Zhang*, W.; Chan*, H. P.; Zhao*, Y.; Aljunied*, M.; Wang*, J.; Liu, C.; Deng, Y.; Hu, Z.; Xu, W.; Chia, Y. K.; Li, X.; and Bing, L. 2024. SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages. *CoRR*, abs/2407.19672.

Zhao, W.; Hu, Y.; Deng, Y.; Wu, T.; Zhang, W.; Guo, J.; Zhang, A.; Zhao, Y.; Qin, B.; Chua, T.; and Liu, T. 2025. MPO: Multilingual Safety Alignment via Reward Gap Optimization. In *ACL 2025*.

Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024. How do Large Language Models Handle Multilingualism? In *NeurIPS 2024*.