Developing Language Models and Resources for Communication across Diverse Languages and Cultures

Recent progress in natural language processing (NLP) technologies has been largely English-centric, often overlooking variations across languages, cultures, and values. This leads to poor performance of large language models (LLMs) in non-English languages, excluding a wide array of communities from the benefits of these technologies. To address this challenge, my research aims to **reduce communication disparities across languages and cultures via human-AI interaction**. In pursuit of inclusive LLMs, I have worked on developing multilingual language models (§1) and multicultural language resources (§2), especially in Korean and Hanja (§3), and applying these advances to English as a Foreign Language (EFL) education (§4).

1. Multilingual Language Modeling

Code-switching is a common linguistic phenomenon in which two or more languages are interleaved within a single conversational context. Inspired by human pattern, **I treat code-switching not as noise but as a signal to improve robustness, efficiency, and effectiveness in multilingual LLMs.** I proposed a coherent pipeline for multilingual LLMs employing code-switching [1, 2, 3].

Evaluation. I revealed that LLMs are vulnerable when prompts interleave many or low-resource languages and introduced **code-switching red-teaming (CSRT)** [1] as both an LLM attack method and an evaluation framework for both safety and multilingual understanding.

Training. To mitigate such failures, I proposed **code-switching curriculum learning (CSCL)** [2] to enhance cross-lingual transfer of English-centric LLMs. CSCL continually pretrains LLMs via a curriculum inspired by human second language acquisition: progressively training on (1) intra-sentential code-switching, (2) inter-sentential code-switching, and (3) monolingual corpora. Using less data, CSCL reduces catastrophic forgetting in English, improves target language performance, and mitigates spurious correlations between language resources and safety alignment.

Inference. Furthermore, I introduced **code-switching in-context learning (CSICL)** [3], an inference-time technique that explicitly scaffolds LLMs' latent reasoning on multilingual inputs through controlled, progressive code-switching. CSICL serves as a linguistic bridge, guiding LLMs to align cross-lingual representations directly instead of relying solely on latent translation, without requiring additional training.

2. Multicultural Language Resources

I introduced data construction frameworks for non-English, low-resource languages [4, 5] and presented language resources for multilingual, multicultural LLMs [6, 7, 8, 9, 10, 11].

Data Construction Frameworks. It is challenging to construct low-resource and domain-specific language datasets due to the limited availability of native speakers and experts. **I proposed a pipeline that employs language learners as NLP data annotators** with support from translations and dictionaries [4]. I demonstrated that language learners can produce reliable labels while learning the language, offering a scalable alternative when recruiting native speakers is challenging in low-resource languages. I also released a web platform that incorporates both the public and experts in annotating Hanja historical documents, with the support of LLM and glossary [5].
In addition, I showed that naïve translation of English benchmarks often overlooks cultural and linguistic differences, misleading LLM evaluations. **I proposed a general pipeline for culturally adapting English benchmarks** and showcased it by presenting a Korean Bias Benchmark for Question-Answering [8]. This pipeline was adopted by eight subsequent works adapting BBQ datasets into 15 languages.

LLM Evaluation Benchmarks. To provide reliable assessments of multilingual LLMs, **I constructed six benchmarks: two multilingual [6, 7], two Korean [8, 9], one Hanja [10], and one English [11]**. These datasets have received wide attention (**100+ citations in total**) and become standard practice in LLM evaluations, adopted by two representative Korean-specialized LLMs (*i.e.*, HyperCLOVA X from NAVER and MIDM from KT). Furthermore, I revealed that existing LLM benchmarks are highly skewed and introduced BenchHub [6], a unified benchmark suite in 10 languages. BenchHub enables NLP practitioners to dynamically customize evaluation sets to their needs via a web platform and code utilities.

3. Language-Specialized Modeling and Resources in Korean and Hanja

Beyond generic multilinguality, I built specialized models and data for Korean and Hanja [8, 9, 10, 12, 13, 14, 15]. These efforts show my full-stack track record from data curation to modeling and evaluation for non-English languages.

Korean. I contributed to the development of HyperCLOVA X [12], a Korean-specialized LLM, and released 5B tokens of Korean historical corpora spanning from the 7th century to the present [13]. I introduced KoBBQ [8] for social bias and CLIcK [9] for cultural and linguistic knowledge in Korean.

Hanja. I also pioneered Hanja modeling, an ancient and archaic Korean writing system using Chinese characters. For instance, I introduced Hanja Understanding Evaluation (HUE) [10], the first NLP benchmark comprising four tasks, and the first pretrained language models in Hanja. I demonstrated that reckless use of Classical Chinese corpora in existing Hanja language modeling approaches hinders Hanja performance [14]. In addition, I developed a neural machine translation model for translating Hanja historical documents into contemporary English and Korean [15].

4. Real-World Testbed: NLP Applications for English Education

I deployed multilingual LLMs as NLP applications for English as a Foreign Language (EFL) education and analyzed real-world human-AI interactions [16, 17, 18, 19, 20, 21].

NLP Tools for EFL Education. I introduced an interactive platform using ChatGPT for EFL writing education [16]. I deployed it to 200+ college students for a semester, publicly releasing logs and editing histories [17]. I further built the interactive platform for oral presentation practice [18] and designed a prompt analytics dashboard (PAD) for instructors [19]. These platforms have been widely adopted in subsequent works (**200+ citations in total**).

NLP Techniques for EFL Education. I collected a rubric-based automated essay scoring (AES) dataset for four years from EFL writing courses and developed AES models by fine-tuning LLMs [20]. This is the first large-scale, publicly available AES dataset, which has received wide attention (**150+ data requests**) and encouraged future studies and applications. Using this AES dataset and models, I proposed a score-based feedback generation that produces more detailed, accurate, relevant, and helpful feedback to students [21].

5. Future Directions

AI systems should provide equitable access to all populations; however, many languages, cultural contexts, and social values outside the U.S. and Western world remain underrepresented. My postdoctoral research will address current challenges and push the boundaries of inclusive LLMs in multimodal and responsible models, tackling problems that current models are far from solving. This research will directly contribute to the development of multilingual, multicultural, multimodal (M^3) LLMs.

Post-Training LLMs for Multilingual Adaptation. I will develop post-training techniques—continual learning, reinforcement learning, and fine-tuning—to enable more efficient adaptation of multilingual LLMs.

- Beyond one-to-one language transfer [2], I will introduce a **simultaneous cross-lingual transfer** approach, structurally related languages are interleaved in a unified continual learning curriculum reflecting shared morphology, syntax, and script. Drawing on my prior work on code-switching across 10 languages [1], I will design linguistically informed switching schedules to reduce data and compute demand in multilingual training.
- To handle unseen languages and tasks at inference, I will develop **multilingual test-time scaling** using reinforcement learning to regulate language-switching by task difficulty. Bilingual speakers often associate certain concepts and terms with specific languages. Combined with my findings that latent code-switching improves multilingual reasoning [3], I will design a verifiable reward model that enhance multilingual reasoning without further training.
- I will also fine-tune multilingual LLMs for **neural machine translation** (NMT) across 15 endangered and low-resource languages, constructing a new corpora of 10k+ parallel sentences. I will leverage my own guided annotation frameworks for language learners [4] and non-experts [5] to improve annotator accessibility and community participation in endangered languages.

Expanding Modality of Multilingual, Multicultural LLMs. While text-based LLMs have recently achieved some extent of multilinguality, multimodal models still exhibit extremely limited linguistic diversity, especially in the speech domain. I will extend continual multilingual learning [2] to align **text and speech representations** using parallel data and synthesized code-switching audio. This will enable robust multilingual spoken interaction and reduce acoustic-textual drift for low-resource languages.

Developing Responsible and Safe LLMs. Multilingual LLMs often exhibit inconsistent knowledge and biased value alignment depending on the input language and culture [8, 6, 9]. To understand the roots of such failures, I will extend my mechanistic interpretability analysis using Logit Lens [22] toward multilingual, multicultural inputs and use my AI safety benchmarks (*e.g.*, CSRT [1] and MAQA [11]), systemically probing layer-wise inference dynamics. To mitigate such an imbalance, I will introduce a **reversed knowledge distillation**: training a balanced, smaller model on culturally diverse data, then aligning larger models via KL-based distribution matching.

References

[1] **H. Yoo**, Y. Yang, and H. Lee, "Code-switching red-teaming: LLM evaluation for safety and multilingual understanding," in ***ACL***, 2025.

[2] **H. Yoo**, C. Park, S. Yun, A. Oh, and H. Lee, "Code-switching curriculum learning for multilingual transfer in LLMs," in ***Findings of ACL***, 2025.

[3] **H. Yoo**, J. Jin, K. Cho, and A. Oh, "Code-switching in-context learning for cross-lingual transfer of large language models," 2025 (Submitted to ARR).

[4] **H. Yoo**, R. A. Putri, C. Lee, Y. Lee, S. Ahn, D. Kang, and A. Oh, "Rethinking annotation: Can language learners contribute?" in ***ACL***, 2023.

[5] S. Song, **H. Yoo**, J. Jin, K. Cho, and A. Oh, "HERITage: An end-to-end web platform for processing Korean historical documents in Hanja," 2025 (Submitted to ARR).

[6] E. Kim*, **H. Yoo***, G. Son, H. Patel, A. Agarwal, and A. Oh, "BenchHub: A unified benchmark suite for holistic and customizable LLM evaluation," 2025 (Submitted to ICLR).

[7] J. Oh, **H. Yoo**, and A. Oh, "Evaluating LLMs' language confusion in code-switching context," in *NeurIPS Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.

[8] J. Jin*, J. Kim*, N. Lee*, **H. Yoo***, A. Oh, and H. Lee, "KoBBQ: Korean bias benchmark for question answering," ***TACL***, 2024.

[9] E. Kim, J. Suk, P. Oh, **H. Yoo**, J. Thorne, and A. Oh, "CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean," in ***LREC-COLING***, 2024.

[10] **H. Yoo**, J. Jin, J. Son, J. Bak, K. Cho, and A. Oh, "HUE: Pretrained model and dataset for understanding Hanja documents of Ancient Korea," in ***Findings of NAACL***, 2022.

[11] Y. Yang, **H. Yoo**, and H. Lee, "MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty," in ***Findings of NAACL***, 2025.

[12] HyperCLOVA X Team (**H. Yoo** as Contributor), "HyperCLOVA X technical report," 2024.

[13] S. Song, N. Kim, S. Chae, K. Park, J. Jin, **H. Yoo**, K. Cho, and A. Oh, "Open Korean historical corpus: A millennial-scale collection of public domain texts," 2025 (Submitted to LREC).

[14] S. Song, **H. Yoo**, J. Jin, K. Cho, and A. Oh, "Shared heritage, distinct writing: Rethinking resource selection for East Asian historical documents," in ***Findings of IJCNLP-AACL***, 2025.

[15] J. Son*, J. Jin*, **H. Yoo**, J. Bak, K. Cho, and A. Oh, "Translating Hanja historical documents to contemporary Korean and English," in ***Findings of EMNLP***, 2022.

[16] J. Han*, **H. Yoo***, Y. Kim, J. Myung, M. Kim, H. Lim, J. Kim, T. Y. Lee, H. Hong, S. Ahn, and A. Oh, "RECIPE: How to Integrate ChatGPT into EFL Writing Education," in ***ACM L@S***, 2023.

[17] J. Han*, **H. Yoo***, J. Myung, M. Kim, T. Y. Lee, S. Ahn, and A. Oh, "RECIPE4U: Student-ChatGPT interaction dataset in EFL writing education," in ***LREC-COLING***, 2024.

[18] J. Cha, J. Han, **H. Yoo**, and A. Oh, "CHOP: Integrating ChatGPT into EFL oral presentation practice," in *EDM Workshop on Leveraging LLMs for Next Generation Educational Technologies*, 2024.

[19] M. Kim, S. Kim, S. Lee, Y. Yoon, J. Myung, **H. Yoo**, H. Lim, J. Han, Y. Kim, S. Ahn, J. Kim, A. Oh, H. Hong, and T. Y. Lee, "Designing prompt analytics dashboards to analyze student-ChatGPT interactions in EFL writing," in *EMNLP Workshop on Customizable NLP*, 2024.

[20] **H. Yoo**, J. Han, S. Ahn, and A. Oh, "DREsS: Dataset for rubric-based essay scoring on EFL writing," in ***ACL***, 2025.

[21] J. Han, **H. Yoo**, J. Myung, M. Kim, H. Lim, Y. Kim, T. Y. Lee, H. Hong, J. Kim, S. Ahn, and A. Oh, "LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction," in *EMNLP Workshop on Customizable NLP*, 2024.

[22] S. Kim, **H. Yoo**, and A. Oh, "On the effect of uncertainty on layer-wise inference dynamics," in *ICML Workshop on Actionable Interpretability*, 2025.