Southeast Asia LLMs: SEA-LION and Wangchan-LION

David Tat-Wee Ong,
AI Singapore

Peerat Limkonchotiwat
Vidyasirimedhi Institute of Science and Technology, Thailand

Abstract

SEA-LION (Southeast Asian Languages In One Network) (Singapore, 2023) is a family of multilingual LLMs that is specifically pre-trained and instruct-tuned for the Southeast Asian (SEA) region, incorporating a custom SEABPETokenizer which is specially tailored for SEA languages. The first part of this talk will cover our design philosophy and pre-training methodology for SEA-LION. The second part of this talk will cover PyThaiNLP's (Phatthiyaphaibun et al., 2023) work on Wangchan-LION, an instruct-tuned version of SEA-LION for the Thai community.

Biography

David Tat-Wee holds a M.Sc in Computer Science and a M.Sc in Financial Engineering from NUS. He started his career by spending a decade in tech as a software engineer, building early Internet applications and working in partnership with Apple, NEC, Siemens and the National Institute of Education. He spent the next decade in finance as a quantitative hedge fund manager with Octagon Capital Management, a Singapore quant fund, starting as a quant research analyst to managing global equities as a portfolio manager to becoming the Chief Investment Officer. He also developed Octagon's in-house financial applications and managed their IT systems.

He joined AI Singapore's (AISG) as an AI Engineer after graduating from its AIAP programme in 2021 and became the Head of the Computer Vsion Hub in 2022. David is presently the Head of Engineering in AISG's Products pillar, managing a team of software engineers to support AISG's Products research implementation.

Peerat Limkonchotiwat is pursuing a Ph.D. student in information science and technology (IST) at VISTEC, Thailand. His research experiences involve Natural Language Processing (NLP) and Information Retrieval (IR), including large language models, dense retrievals, semantic understanding, representation learning, question answering, and entity linking. He is currently working with Dr. Sarana Nutanong (VISTEC, Thailand) and Dr. Ekapol Chuangsuwanich (CU, Thailand).

Currently, he is a subject matter expert in a WangchanX project, a Thai NLP group developing applications based on research. His projects in WangchanX are Thai sentence embedding benchmarks, Thai text processing datasets (VISTEC-TP-TH-2021 and NNER-TH), and generative models, i.e. WangchanGLM (Polpanumas et al., 2023) and Wangchan-Sealion).

References

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai Natural Language Processing in Python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, Online. Association for Computational Linguistics.

Charin Polpanumas, Wannaphong Phatthiyaphaibun, Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Titipat Achakulwisut, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. WangChanGLM — The Multilingual Instruction-Following Model.

AI Singapore. 2023. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. https://github.com/aisingapore/sealion.

Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), pages 246–246