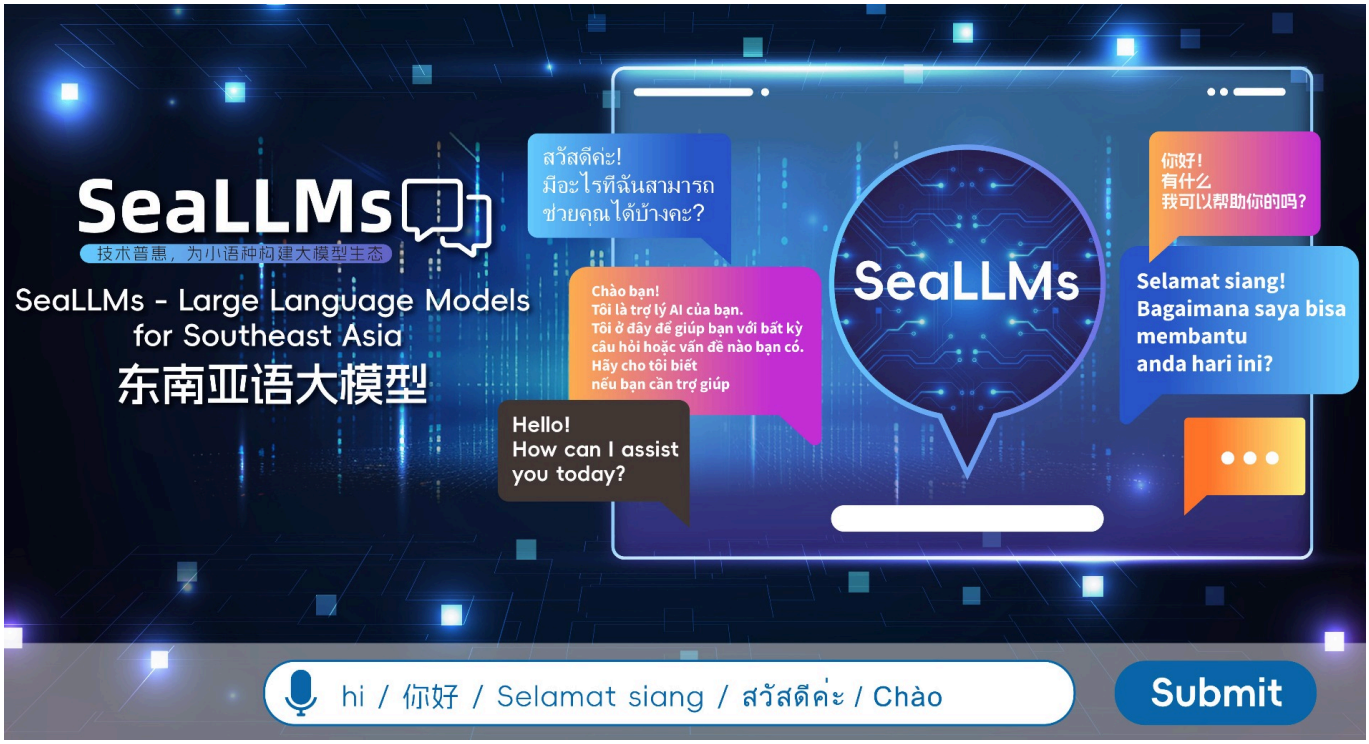
TECH OFFER

SeaLLMs - Large Language Models for Southeast Asia



KEY INFORMATION

TECHNOLOGY CATEGORY:
Infocomm - Artificial Intelligence

TECHNOLOGY READINESS LEVEL (TRL): TRL4
COUNTRY: SINGAPORE
ID NUMBER: TO175075

OVERVIEW

Despite the remarkable achievements of large language models (LLMs) in various tasks, there remains a linguistic bias that favors high-resource languages, such as English, often at the expense of low-resource and regional languages. To address this imbalance, we introduce SeaLLMs, an innovative series of language models that specifically focuses on Southeast Asian(SEA) languages. SeaLLMs are built upon the Llama-2 model and further advanced through continued pre-training with an extended vocabulary, specialized instruction and alignment tuning to better capture the intricacies of regional languages. This allows them to respect and reflect local cultural norms, customs, stylistic preferences, and legal considerations.

Highlights:

- The models' attunement to local norms and legal stipulations—validated by human evaluations—establishes SeaLLMs as not only a technical breakthrough but also a socially responsiveinnovation.
- SeaLLM-13b models exhibit superior performance across a wide spectrum of linguistic tasks and assistant-style

instruction-following capabilities relative to comparable open-source models.

- SeaLLMs outperform mainstream commercialized models for some tasks in non-Latin languages spoken in the region, meanwhile, SeaLLMs are efficient, faster, and cost-effective compared to commercialized models.

TECHNOLOGY FEATURES & SPECIFICATIONS

The SeaLLMs went supervised finetuning (SFT) and specialized self-preferencing alignment usinga mix of public instruction data and a small number of queries used by SEA language native speakers in natural settings, which adapt to the local cultural norms, customs, styles and laws inthese areas. SeaLLM-13b models exhibit superior performance across a wide spectrum of linguistic tasks and assistant-style instruction-following capabilities relative to comparable open source models. Moreover, they also outperform other mainstream commercialized models in tasks involving very low-resource non-Latin languages spoken in the region, such as Thai, Khmer, Lao,and Burmese.

Training Process

Our pre-training data consists of more balanced mix of unlabeled free-text data across all SEA languages. We conduct pre-training in multiple stages. Each stage serves a different specific objective and involves dynamic control of (unsupervised and supervised) data mixture, as well as data specification and categorization. We also employ novel sequence construction and masking techniques during these stages.Our supervised finetuning (SFT) data consists of many categories. The largest and most dominantof them are public and open-source. As the aforementioned are English only, we employed several established automatic techniques to gather more instruction data for SEA languages through synthetic means. For a small number of SFT data, we engaged native speakers to vet, verify and modify SFT responses so that they adapt to the local cultural customs, norms, and laws. We also adopted safety tuning with data for each of these SEA countries, which helps to address many culturally and legally sensitive topics more appropriately - such tuning data tend to be ignored, or may even appear in conflict with the safety-tuning data of other mainstream models. Therefore, we believe that our models are more local-friendly and abide by local rules to a higher degree. We conduct SFT with a relatively balanced mix of SFT data from different categories. We make use of the system prompt during training, as we found it helps induce a prior which conditions the model to a behavioral distribution that focuses on safety and usefulness.

POTENTIAL APPLICATIONS

Through rigorous pre-training enhancements and culturally tailored fine-tuning processes,SeaLLMs have demonstrated exceptional proficiency in language understanding and generation tasks, challenging the performance of dominant commercial players in SEA languages, especially non-Latin ones. The models' attunement to local norms and legal stipulations—validated by human evaluations—establishes SeaLLMs as not only a technical breakthrough but a socially responsive innovation, poised to democratize access to high-quality AI language tools across linguistically diverse regions. This work lays a foundation for further research into language models that respect and uphold the rich tapestry of human languages and cultures, ultimately driving the AI community towards a more inclusive future.

UNIQUE VALUE PROPOSITION

One of the most reliable ways to compare chatbot models is peer comparison. With the help ofnative speakers, we built an

instruction test set, called Sea-bench that focuses on various aspects expected in a user-facing chatbot, namely: (1) task-solving (e.g. translation & comprehension), (2)math-reasoning (e.g., math and logical reasoning questions), (3) general-instruction (e.g.,instructions in general domains), (4) natural-questions (e.g., questions about local context often written informally), and (5) safety- related questions. The test set also covers all languages that we are concerned with. AI model candidates' responses to the test set's instructions may be judged and compared by human evaluators or more powerful large and commercialized AI models to derive a reliable performance metric. Through this process, we demonstrate that our SeaLLM-13b model is able to perform on-par or supasses other open-source or private state-of-the-art models across many linguistic and writing tasks.