２Ｂ２３

アジア・太平洋発 AI モデルの研究開発動向と課題
―東南アジア諸国を中心に―

○斎藤　至（JST アジア・太平洋総合研究センター）
itaru.saito@jst.go.jp

1. はじめに：アジア・太平洋発 AI モデルの台頭

人工知能（AI: Artificial Intelligence）は社会実装の速さから世界の経済発展の基盤となり，かつ安全保障を規定する重要技術と認識され，政策的な振興が強化されている。その隆盛を支える大規模言語モデル（LLM: Large Language Model）は，機械学習を基盤としてデータの入出力がなされる。

アジア・太平洋発 AI モデルで広く注目を集めるのは，中国発 AI モデルである。中国の新興企業ディープシークは，2024 年 12 月に対話型モデル DeepSeekV3 を，2025 年 1 月に専門化モデル DeepSeek R1 を発表した。従来の米国発 AI モデルに比べ 10 分の 1 程度の研究開発費で ChatGPT-4o や Claude 3.5 に匹敵する性能を発揮し，世界に衝撃を与えた。学習データの質の懸念や開発プロセスの不透明性が露見し，同社モデルの利用を禁止する国も現れたものの，各モデルの改良版が 2025 年前半に同社から相次いで発表された。また，ディープシーク社に続く中国 AI 企業も台頭しつつある。

本稿では，東南アジア諸国で開発の進む新興 AI モデルに注目する。応用面を中心に科学技術力の高まる各国で，ヘルスケア・先端技術での専門的利用やイノベーションへの活用[1]を促すには，話者が多く利用局面に浸透した現地語[2]に対応する LLM の開発・普及が急務とされる。しかし東南アジア諸語の多くは，注釈付きデータセット，計算機資源，オンラインコンテンツの相対的に乏しい「低リソース言語」である。地域内の言語的多様性もしばしば高く，インドネシアでは，共通語であるインドネシア語のほかに 700 を超える地域語が用いられ，シンガポールでも，英語（シングリッシュ）のほか，中国語（北京語），マレー語，タミル語の 4 つが公用語とされる。こうした言語事情から，東南アジア諸国では単一モデルで多言語のプロンプトに対応する必要性がとりわけ高い。

以下ではまず，シンガポールを中心とした東南アジア諸国連合（ASEAN）主要加盟国の政策とガバナンス（利用促進・規制枠組み）を概観する。次に，各国の代表的な低リソース言語対応 LLM を紹介し，研究開発動向と課題を検討する。これらを踏まえ，アジア・太平洋発 AI モデルの可能性を述べる。

2. 東南アジア諸国の AI に関する政策・ガバナンス

ASEAN 主要加盟国において，AI の研究開発推進はシンガポールで顕著に発展している。またインドネシアやタイはビジネス上の有望な参入市場とみられている[3]。

シンガポールでは，2015 年発表の国家 5 カ年計画「研究・イノベーション・企業 2025 年計画」で AI の革新性を先見した。のち，二度の AI 国家戦略策定を経て，「スマート・ネイション構想 2.0」（2024 年 10 月発表）では AI をスマート国家建設の中核に位置付け，「信頼」（安全性）「成長」（経済発展への寄与）「連帯」（包摂性）を鍵概念として提示した。これらビジョンの実現と推進に際しては，2019 年，スマート・ネイションおよびデジタル政府局（SNDGO）の中に国家 AI 局が設置され，翌 2020 年に情報メディア開発庁（IMDA）内へ設置された「国家デジタル局」との連携により，体制を強化した。この下で，科学技術研究庁（A*STAR）が基礎研究と産業応用の架橋を，自治大学など 6 研究機関から成る AI シンガポールが AI 人材育成・活用促進を，シンガポール国立大学（NUS）と南洋理工大学（NTU）の両研究型大学が AI 基盤技術の研究を，それぞれ担っている。近年，上述の各機関は AI モデル MERALiON（3.で後述）の開発で協力を緊密化させている。2025 年 5 月， A*STAR の情報通信技術研究所（I²R）と IMDA は，AI を諸科学分野や産業領域に応用し実践的開発を促すべく，主要な政府系組織のコンソーシアム設立を発表している[4]。

インドネシアでは，約 2.8 億人超の人口が支えるデジタル経済規模の拡大に対し，情報インフラの立ち後れが国家的課題である。2025 年 8 月に中央政府は「国家 AI ロードマップ白書」を発表し，情報インフラ整備に加えて AI 人材育成とイノベーションの促進を柱とした短・中・長期の行動計画を示した。

また，関連の特別規則と倫理・ガイドライン整備にも 2023 年後半から着手している。

タイでは，シンガポールに比べ体系性を欠くものの，戦略・行動計画に基づき投資を強化している。高等教育科学研究イノベーション省(MHESI)がデジタル経済社会省と共同で「国家 AI 5 カ年戦略・行動計画 2022~27」を発表し，①ELSI 面の整備，②国家計算機インフラの整備，③人的能力・教育の強化，④技術開発の推進，⑤官民活用の推進，を掲げている。鍵となる 10 の産業セクターを措定し，2027 年までに，関連法案の施行，3 万人の AI 人材育成，600 機関での AI 活用を目指している。同戦略・行動計画は，①AI 倫理ガイドラインの設置，②エヌビディアの A100 GPU を搭載した高性能コンピュータ LANTA の稼働など，順調な進捗をみせている。

AI の開発促進と共に，ガバナンスの構築も広域的な共通課題と認識され，シンガポールを中心に ASEAN での整備がみられる。OECD 原則（2017）や AI 倫理に関する UNESCO 勧告（2021）を範として，ASEAN の共通ガイドライン（2024.2，増補版 2025.2）が発表され，加盟国のガバナンス強化を推奨し， 生成 AI のリスク対応指針と，AI の構造的リスクの検討を表明している。特に安全性（safety）は世界的な共通課題化を受けて，東南・南アジア発の非政府的な枠組み AI Safety Asia が発足し，国家連合としての安全性ネットワーク ASEAN AI Safe の策定も準備中である。国でみると，シンガポールでは，安全性研究所の設置，検証ツール「AI Verify」の導入，諸外国関係者を集めた模擬サイバーセキュリティ攻撃訓練の実施など，多様な実践が機動的に進んでいる。

3. 東南アジア発 AI モデルの開発アプローチ

以下では，低リソース言語対応 LLM の開発アプローチの類型と，開発上の留意点を示す。

3.1. LLM の代表的な要素技術と、低リソース言語対応 LLM の問題点

今日の AI モデルの主流である LLM は，様々なタイプの機械学習（AI が人間の思考や判断を模倣するためのプロセス）や深層学習を通じて，情報処理や推論を行う。処理する情報は多岐にわたり，自然言語処理および音声処理（機械翻訳，音声認識，自然言語生成〔文章生成〕），画像処理（画像中の情報の正確な認識）やコンピュータ・ビジョン（AI が自ら見たものを理解したタスクの実行）等が代表的である。また，自然言語の文章・音声や画像など，複数のモダリティや異なる種類のデータを組み合わせて複合的に処理・生成する「マルチモーダル AI」も現れている。

LLM に代表される基盤モデルは，AI の開発潮流において「第三世代」とされ，普及が進んできた。一方で，課題としては大きく (1)大量の教師データや計算資源を必要とし，(2)実世界状況への臨機応変な対応がとれず，(3)意味理解・説明等の高次処理を不得意とする，の 3 つが指摘されてきた。

加えて，低リソース言語はデータセットに対する注釈の誤りがあり，非ローマ字で表記される言語が多く，口語・日常語の代表的リソース（母集団を代表する言語情報）が web 上に乏しい傾向にある。そのため，LLM の学習データとして用いた際，アウトプットの精度が著しく下がる懸念を抱えている。

スタンフォード大学人間中心 AI 研究所（Stanford HAI）では，第三世代 AI の課題と，低リソース言語データの傾向を踏まえ，これらの克服を試みる 3 つのアプローチを提示している。すなわち，(1)できるだけ多くの多言語で訓練したモデル（極大多言語モデル），(2) 少数の多言語で訓練したモデル（地域多言語モデル），(3) 単一言語で訓練したモデル（単一言語文化モデル）の 3 つである。表 1 では，グーグル の双方向言語モデル BERT[5]に加え，本稿で触れる各国発のモデルも例示した。

表 1　低リソース言語対応 LLM に対する 3 つの開発アプローチと留意点

	(1) 極大多言語モデル	(2) 地域多言語モデル	(3) 単一言語/単一文化モデル
言語の種類	100超	10~20	1~2
AIモデルの例	BERTシリーズ，mT5，XLM-R，Aya	BERTシリーズ，SEA-LION，MERaLiON，Sailor	BERTシリーズ，Suraシリーズ，URA-LLaMA，Typhoon，Pathumma
肯定的評価	言語間の転移学習・スケール化に適 複数の単一言語モデル併用より機能的	言語間の転移学習・スケール化に適 複数の単一言語モデル併用より機能的	多言語性の弊害を回避できる
否定的評価	多言語性の弊害を回避できない アウトプットが曖昧化しやすい	ドメイン特定的な自然言語処理が低劣 包括的ベンチマークの不足	言語間の移転学習が不足 最初からの訓練がコスト高

出典：Pava，J.N. et al.（2025），p.10 を基に著者作成

上表(1)(2)の多言語モデルは，言語間の転移学習（transfer learning，あるタスクで学習済みのモデル

を別の関連タスクに応用し，パフォーマンスと汎用性を高める技術）の効果で，少量のデータからでも高精度なモデル開発を実現し，複数の単一言語モデルを併用するよりも機能的である（スケールメリットがある）点で，一定の支持を得ている。だが対応言語の種類が多いほど，多言語性の弊害（複数言語を学習データに含むことで，個別言語によるアウトプットの質が劣化する問題）が顕在化するため，否定的評価もある[6]。モデルの開発では，転移学習効果と弊害のトレードオフに留意せねばならない。

3.2. 東南アジア発 LLM の代表例と特質

表 2 に一覧するシンガポール，インドネシア，タイの代表的な AI モデルの特質を，以下で整理する。

表 2　代表的な東南アジア発 AI モデル

呼称	シーライオン SEA LiON Southeast Asian Languages in One Network	マーライオン/ MERaLiON Multimodal Empathetic Reasoning and Learning in One Network	サハバットAI / Sahabat AI	パトゥマ / Pathumma
開発主体	AIシンガポール* *シンガポールの主要大学・研究機関による共同研究プログラム	A*STAR情報通信技術研究所（I²R），米マイクロソフト	ゴートゥ・ゴジェック・トコペディア，Indosat，米エヌビディア	BDI，国立電子コンピュータ技術研究センター (NECTEC)，チュラロンコン大，マヒドン大
資金提供主体	シンガポール国家研究基金（NRF）	NRFおよび情報メディア開発庁（IMDA）	インドネシア中央政府	タイ中央政府
v.1の発表	2023年12月	2024年12月 (オーディオLLM)	2024年11月	2025年9月リリース予定
アーキテクチャ	Meta LLaMA 3 (v.2) Google Gemma 3 (v.4)	SEA LiON (文章) Whisper (音声)	Google Gemma2	Qwen 2.0 (文章・画像) Whisper (音声)
対応言語	英語，中国語，インドネシア語，マレー語，タイ語，ベトナム語，フィリピン語，タミル語，ビルマ語，クメール語，ラオ語，ジャワ語，スンダ語	英語（米語・シングリッシュ），北京語 オーディオLLM v.2 (2025.2) では，対応言語がマレー語・タミル語・インドネシア語・ベトナム語に拡張され，1つの会話で多言語が入り交じる際の処理も可に	英語，インドネシア語，ジャワ語，スンダ語	タイ語
主要機能	基本の自然言語処理	反訳・翻訳・要約・質疑応答	基本の自然言語処理	文章・音声・画像生成

出典：各 AI モデルの web サイトから著者作成

シンガポールは，非西洋高度産業社会（WIRED と総称）からの情報発信水準の向上をミッションに，東南アジア主要諸語を網羅する「地域多言語モデル」の開発を牽引している。その旗艦モデルが，SEA LiON（シーライオン，Southeast Asian Languages in One Network）と MERaLiON（マーライオン，Multimodal Empathetic Reasoning and Learning in One Network）である。

SEA LiON は，シンガポール国家研究基金（NRF）が支援し，国の主要大学・研究機関の共同研究プログラムである AI シンガポールで開発された。AI シンガポールの検証によれば，インドネシア諸語によるタスク実行では LLaMA 2 をはじめ主要な大規模モデルに比べて高い成果を示した一方，英語によるタスク実行では中庸の成果に留まったことから，東南アジア諸語の利用に特化した局面で，最適なパフォーマンスを発揮するモデルと想定できる。

MERaLiON は，SEA LiON の言語処理機能を組み込み，A*STAR I²R が米マイクロソフトと共同開発を進めているマルチモーダル LLM である。2024 年 12 月発表のオーディオ LLM v1 では，従来の小規模モデル AI が不得意とした，文脈的理解とマルチタスクの汎用性を高めるべく，言語・画像・映像などの多様なデータソースを統合し，方言を含む高度な言語処理にも対応している。2.で述べた研究開発コンソーシアム設置という制度的後押しもあり，欧米発の AI モデルに代替する東南アジア発モデルとして開発が進展すると見込まれる。

このほか，インドネシアでは主要な現地語を含む「地域多言語モデル」が，タイでは国語に対応した「単一言語モデル」の開発が進展している。タイではタイフーン(Typhoon) が代表例となってきたが，近年は 2022 年発表の第 2 次 AI 国家戦略で示されたマルチモーダル型 LLM「パトゥマ(Pathumma)」[7]の開発が進んでいる。政府は 2024 年 10 月の v.1 発表に続き，ヘルスケア・観光・環境などの産業領域での活用を目指し，2025 年 9 月中のリリースを目指している。研究開発チームには，BDI，国立電子コンピュータ技術研究センター(NECTEC)と共に，チュラロンコン大学とマヒドン大学，2 つの AI 産業団体が参加し，約 340 万米ドル相当の政府予算が投じられている[8]。インドネシアでは，2024 年にエヌビディアの大型クラウド GPU Merdeka の導入で，英語と国内 3 言語に対応の「サハバット AI」が発表された[9]。各種 AI モデルは COVID-19 パンデミックで導入された統合医療情報アプリ・サツセハッ

ト（SatuSehat）等に組み込まれ，群島国家である同国のヘルスケア DX にも寄与している。

4. おわりに：東南アジア発 AI モデルの可能性

上述のように，シンガポールをはじめ，AI の研究開発や産業応用が進む東南アジア諸国では「地域多言語モデル」と「単一言語モデル」の開発が進められている。また，インドネシアやタイが比較的少数の国内言語を対象にモデル開発を進めるのに対し，各種の先端技術で高い科学技術力を示すシンガポールは，東南アジアの主要諸語を網羅した汎東南アジア・モデルの開発を牽引している。

低リソース言語対応 AI モデルの開発では，多様性（マイナーな言語文化）への配慮や，公平性（社会正義）などの社会的意義が強調されやすい。だが情報科学的側面から見ると，多言語に対応したモデル設計は，ユーザーが増えるほどフィードバックをより多く得られ，予測の精度を高める「データネットワーク効果」[10]をも持つことになる。低リソース言語の制約を踏まえた転移学習効果を活用した開発を進めれば，データネットワーク効果は一層高まり，AI モデルの精度を高めると考えられる。

既に触れたように，低リソース言語対応 LLM は代表的データが僅少であり，その開発も着手されたばかりである。他のアジア諸国に目を転じると，インドには公用語だけで 22，その他の地域語を含めると 121 言語が存在するといわれ，9 言語に対応した地域多言語モデルのマルチモーダル LLM バーラット・ジェン（BharatGen）の開発が 2024 年 9 月から始まっている。また，LLM のモデルとしての過大さがもたらす技術的な困難から，一定の用途に特化した小規模言語モデル（SLM）の開発も並行して進んでいる。今後は本稿の知見も起点の一つに，様々な国・地域における事例の蓄積と検証が望まれる。

参考文献

科学技術振興機構 アジア・太平洋総合研究センター（JST/APRC），シンガポールの科学技術人材育成・確保に関する調査，APRC-FY2022-RR-02, 40p. (2023). https://spap.jst.go.jp/investigation/downloads/2022_rr_02.pdf

――，アジア・太平洋主要国における人工知能（AI）の政策と研究開発動向，APRC-FY2024-RR-06, 152p. (2025). https://spap.jst.go.jp/investigation/downloads/2024_rr_06.pdf

――，シンガポール，東南アジア諸語対応の大規模言語モデル(LLM)開発とその将来，サイエンスポータル・アジアパシフィック, 2025 年 6 月 30 日. https://spap.jst.go.jp/asean/experience/2025/topic_ea_16.html

Noor, E. and B. Kanitroj, Speaking in Code: Contextualizing Large Language Models in Southeast Asia, Washington, DC: Carnegie Endowment for International Peace, 60+vip. (2025). https://carnegieendowment.org/research/2025/01/speaking-in-code-contextualizing-large-language-models-in-southeast-asia?lang=en

Pava, J.N. et al., Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts, Stanford: Stanford University Human-Centered Artificial Intelligence, 25p. (2025). https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts

---

[1] JST/APRC (2023). また東・東南アジアでの事業経験に基づく経営分析からの概観として，東南アジアにおける生成 AI ビジネスの潮流と要諦. 山田コンサルティンググループ, 2025 年 8 月 2 日 https://www.ycg-advisory.jp/learning/oversea_218/

[2] 今日の AI モデルにおける使用言語は，英語を中心とする欧米語や中国語が主流であるが，一般に世界の諸言語のうち約 3 分の 2 は，アジアとアフリカで用いられているという。

[3] シンガポールは，米 AI シンクタンク Thundermark の研究力ランキング（2022）で世界 11 位に位置した。JST/APRC (2025) 「はじめに」も参照のこと。

[4] 2025 年 9 月現在の参加機関は，公衆衛生や国民の保健に携わる保健省（MOHT），国民の安全を担うホームチーム科学技術庁（HTX），大規模計算機 ASPIRE を所掌する国立スパコンセンター（NSCC）である。

[5] BERT が多用された背景としては，このモデルが自然言語処理でラベル未付与データを大量に用いて事前学習を実行できる特徴を有し，データ不足の著しい低リソース言語を扱うモデルに適したためと考えられる。

[6] 多言語性の弊害は，人間を対象とした言語教育の観点でも，多言語話者（マルチリンガル）の抱える課題（複数言語の混同や不充分な習得）として，つとに指摘される。

[7] パトゥマ LLM: タイの文脈・文化に合わせた AI 技術. NSTDA, 2025 年 3 月 25 日 https://www.nstda.or.th/en/news/news-years-2025/pathumma-llm-ai-technology-tailored-to-thai-context-and-culture.html

[8] 投資額等の詳細は，国家 AI 戦略の一環で 6 事業を計画，タイ. itnews, 2024 年 3 月 12 日 https://www.itnews.asia/news/thailand-plans-six-new-projects-as-part-of-its-national-ai-strategy-606000

[9] サバハット AI の概要を参照。 https://sahabat-ai.com/en#cloudeka

[10] AI 戦略を左右する新しいネットワーク効果とは何か. DIAMOND ハーバード・ビジネス・レビュー, 2023 年 5 月 16 日 https://dhbr.diamond.jp/articles/-/9518