報告タイトル（＊日本語と英語両方ご記入ください）

東南アジアにおける研究の実践から見た大規模言語モデルの利活用について

“Large Language Models in Southeast Asian Research Practice”

氏名（所属）

八木暢昭（京都大学大学院）

YAGI Nobuaki（Kyoto University）

要旨（800 字程度）

東南アジアにおいてもソーシャルメディアは急速に普及しており、インドネシアでは選挙ツールとしても重要な役割を果たしている(岡本ら 2024)。ソーシャルメディア上のやり取りは画像や動画などもあるが、中でもテキストは広く発信されているデータである。

テキストデータといったそのままでは機械学習のような処理では扱いづらいデータを非構造化データと呼ぶ(Ozdemir 2022/2023 田村ら訳)。近年では BERT(Devlin et al. 2018)などの大規模言語モデルが発展しており、構造化されていないテキストデータを機械的に扱うことができるようになった。また、大規模言語モデルに関する知見も体系的に整備されており、そうした技術を扱うための障壁も下がってきている(例えば Alammar & Grootendorst(2024/2025 中山訳)、山田ら(2023)などが挙げられる)。

一方で、社会科学においてテキストデータを扱った研究が少ないことが指摘されている(Müller-Hansen et al. 2020)。これはテキストデータの重要性に比して、こうしたデータの研究現場における利活用が追いついていない状況を反映していると考えられる。こうした課題意識から、本報告では大規模言語モデルを研究に適用するための実践的な知見について議論する。

本報告ではモデルの訓練に用いるデータの準備、モデルの構築における OpenAI API などの基盤モデルの利活用、より効率的で信頼性の高い研究に向けた各種技術(特に human-in-the-loop 機械学習によるアノテーション(Monarch 2021/2023 上田ら訳)や GPT を使ったデータの偏りの補正(Zhao et al. 2023)など)について議論したい。