

ABCI 3.0開発加速利用 (2025年度) 成果概要 (公開用)

<p>課題名： 日本語に強い大規模言語モデルの構築およびチューニング並びに透明性・信頼性の確保 に向けた研究開発</p>	<p>実施時期：2025/4 – 2026/3 所属機関名：国立情報学研究所 代表者氏名：黒橋 禎夫</p>
<p>成果概要： 国産オープンソースLLM「LLM-jp-4」シリーズ(8B denseおよび32B-A3B MoEのbase/thinking計4モデル)、および視覚言語モデル「LLM-jp-4-VL 9B beta」を公開した。日本語MT-BenchでGPT-4oやQwen3-8Bを上回る性能を達成し、Apache 2.0ライセンスで商用利用も可能。学習コーパス・トークナイザー・開発レシピまで公開した、透明性の高いソブリンAI基盤として高く評価されている。</p>	
<p>成果のポイント： ① 言語モデル LLM-jp-4 シリーズ 合計19.5兆トークン規模の学習コーパス(LLM-jp-3.1比 約6倍)を新たに構築し、うち事前学習に約10.5兆トークン、中間学習にさらに約1.26兆トークンを投入し、合計約11.74兆トークンでフルスクラッチ学習を実施した。コンテキスト長は最終的に64Kトークンまで拡張されている。事後学習では、約130万件の指示から合成した5.4Bトークンの教師あり学習(SFT)データと、約15万件の選好学習データを用いてチューニングを行った。その結果、日本語MT-Benchにおいて8Bモデルは7.54、32B-A3Bモデルは7.82を達成し、GPT-4o(7.29)やgpt-oss-20b(7.33)、Qwen3-8B(7.14)を上回る性能を示した。llm-jp-eval v2.1.3の総合評価でもQwen3-8Bやgpt-oss-20bとほぼ同等の平均スコアを記録している。これらのモデルはApache 2.0ライセンスで公開され、商用利用やファインチューニングも自由で、OSAI(Open Source AI Definition)に準拠した高い透明性を備えている。 ② 視覚言語モデル LLM-jp-4-VL 9B beta LLM-jp-4 8B instructモデルをベースに、インターネット上の公開データや政府・国会の文書などから抽出した画像・言語データと合成データを組み合わせた、日英合計約3,340万事例(約180Bトークン)の学習データを用いて構築した。このマルチモーダル事後学習における学習トークン数はQwen3-VLの約1/10に抑えられているにもかかわらず、日本語タスクの平均性能ではQwen3-VL-8Bと同等の水準を達成しており、特に日本の文化・常識の知識を問うタスクで高い性能を発揮している。</p>	
<p>成果についてより詳細な情報を提供しているWebページ、発表論文などの情報： https://llm-jp.nii.ac.jp</p>	