ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名： 世界最大級の高品質な日本語データを用いた日本文化、アイデンティティーに忠実なLLMの構築	実施時期：2025/04/15-2026/03/30 所属機関名：国立研究開発法人情報通信研究機構 代表者氏名：呉　鍾勲

成果概要：
NICTではこれまでに多くのLLMを構築した経験を有しており、本課題では、その経験を基に構築した世界最大規模の高品質な22.9TBの日本語学習データ等を活用して日本文化、アイデンティティーに忠実な2,080億パラメータのLLMを構築した。
また、株式会社Preferred Networks（以下、PFN）と共同研究を実施し、LLMを開発した 。

成果のポイント：
　NICTではこれまで、高性能な日本語特化型の大規模言語モデル（LLM）構築のため、長年に渡って収集してきた約381億ページの日本語Web文書から、22.9TBという日本語データとしては、我々の知る限り、世界最大の事前学習用データを構築している。なお、令和７年度には、700億ページの日本語Web文書から44.3TBの事前学習用データを構築している。本課題では、NICTで構築した事前学習用データや共同研究先であるPFNから提供された合成データ等も活用して12TBの日本語事前学習用データを整備し、それを用いて日本の文化やアイデンティティーに忠実な2,080億パラメータの日本語特化型LLMをスクラッチから構築した。
　また、日本の文化や社会等に留意した高品質かつ大量の学習データを用いた安全で高性能な国産LLMを目指して実施したPFNとの共同研究では、20億、80億、310億パラメータといったLLMが開発された。これらのLLMは、PFNのHuggingfaceレポジトリ（https://huggingface.co/pfnet）上にPLaMo-3-NICT（plamo-3-nict-31b-base等）の名称で一般公開されている。さらに、本課題の直接の成果ではないものの、PFNが、これらのLLMに基づいて独自に開発したPLaMo 3.0 Prime β 版 をリリースしている（2026年３月）。

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://www.preferred.jp/ja/news/pr20260319
https://tech.preferred.jp/ja/blog/plamo_3_8b_31b/