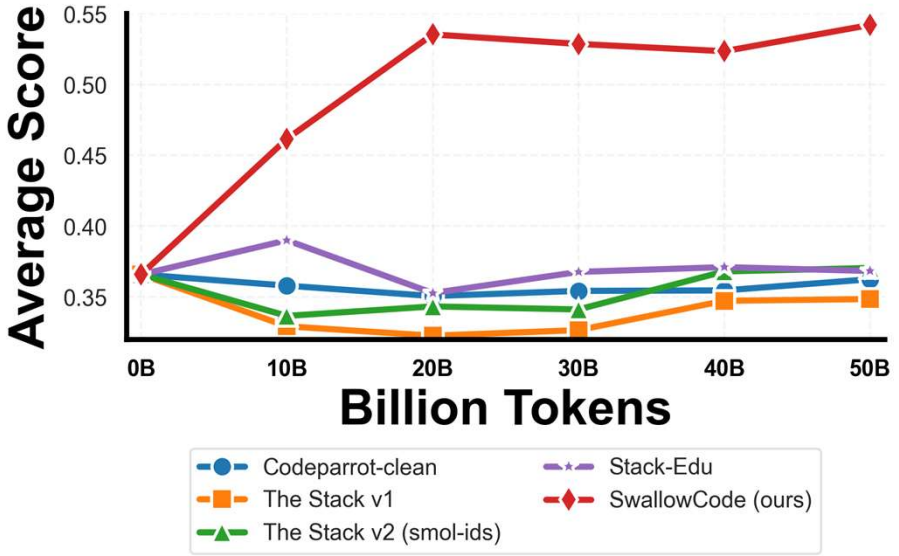
ABCI 3.0開発加速利用 （2024年度）成果概要（公開用）

課題名： 継続事前学習に基づく汎用性および有用性の高い日本語大規模言語モデルの構築	実施時期：2025年1月から3月 所属機関名：産業技術総合研究所 代表者氏名：高村大也

成果概要：数学とコードに強いLlama-3.1-Swallow-8B-v0.5構築のためのコードデータセット作成を行った。オープンなコードデータセットであるThe Stack v2にはSyntax Errorやコーディング規約を守らないコードが多数存在している。そこでLlama-3.3-70B-Instructを利用し、高品質なコードデータセットに作り直すデータ合成を行った。これにより、高品質なコードデータセットを構築することができた。

成果のポイント：
NeurIPS 2025 Dataset trackに投稿予定の論文でも報告するように、作成したデータセットで学習を行うことで、他の主要なオープンコードデータセットで学習したLLMを著しく上回る性能を発揮するLLMを構築することができた。
このデータセットは、Llama-3.3 Licenseのもとでhuggingfaceにて公開する予定である。



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：