

ABCI 3.0開発加速利用（2024年度）成果概要（公開用）

課題名： 大規模言語モデル支援による細胞分化法のデータベース構築及び培養条件の最適	実施時期：2024年度 所属機関名：東京大学 代表者氏名：西川 昌輝
成果概要： 本研究は、細胞培養、その中でも特に細胞分化手法を、大規模言語モデルを用いて最適化することを目的としている。ABCIの計算資源を利用し、DeepSeekやLlamaに代表されるローカル言語モデルを稼働している。ジャーナルが提供するAPIを利用し、高速に文献情報を抽出する手法の提案を試みた。結果として、文献全体の内容を適切に処理し、必要な情報を正確に抽出するためのワークフローを構築した。	
成果のポイント： 報告される文献は非構造的なデータの形式をとっている。よって、データ駆動的なアプローチの導入を妨げる大きな要因と言える。そこで、文献データを言語モデルによって処理し、構造的なデータを抽出するワークフローを開発した（右図）。 一般に言語モデルは、入力が冗長になればなるほどその性能が低下する。論文一報分のテキストは、この性能の低下を引き起こすのに十分に長く、工夫が必要であった。そこで、論文内のテキストを分割し、各テキストセグメントから情報を抽出する。略語や、事前に設定された条件等をこぼさないために、過去に抽出したデータを統合していくことで、全体のデータの整合性を担保している。 The diagram illustrates the workflow for extracting structured data from unstructured document text. It starts with data sources like Elsevier and Springer, which feed into a 'Document Data' box. This data is then processed by a 'Language Model' (Llama, DeepSeek, etc.) to perform 'Data Extraction'. The extracted data is then consolidated ('データ統合') into a final 'Structured Data' box. An example of the final structured data is shown as a JSON object: <pre>{ "consideration": "GSK3 inhibition", "method": { "cell/gene": [{ "source": "defined culture medium", "cell": "human pluripotent stem cells (hPSCs)", "gene": ["not mentioned"] }], "factors": [{ "factor": "GSK3 inhibition", "concentration": "not mentioned" }, { "factor": "defined culture medium", "concentration": "not mentioned" }], "culture": { "medium": "defined culture medium", "medium_change_frequency": "not mentioned" } } }</pre> 図: 言語モデルによる文献情報の抽出フロー	
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： 特になし。	