

データリポジトリにおける大規模言語モデルの利活用を考える

北本 朝展（国立情報学研究所）



<https://dias.ex.nii.ac.jp/>



大規模言語モデル (LLM)



<https://llm-jp.nii.ac.jp/>

1. ChatGPT等の大規模言語モデルが社会に広く浸透
2. GPT (General-Purpose Technology) は全分野に影響
3. 日本の学术界でもLLMを構築するプロジェクト開始
4. データリポジトリにも大きな影響を与えるのは確実

タスクの効率化

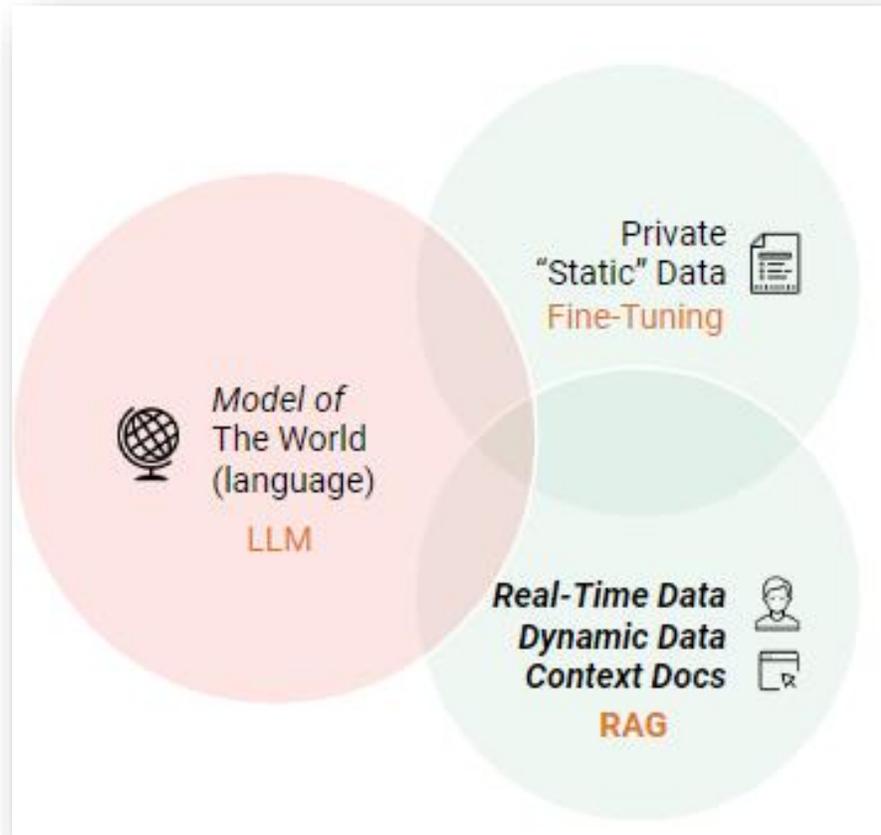
KE Explain the use of DIAS dataset in two paragraphs

 The DIAS dataset, specifically the Global Soil Wetness Project Phase 3 (GSWP3), plays a crucial role in the study as a source of climate forcing data required to run the Community Land Model (CLM). The GSWP3 dataset contains essential climate variables such as temperature, precipitation, and radiation. These climate variables are vital for simulating ecosystem processes and understanding the interactions between climate and ecosystem dynamics in the model.

By using the DIAS dataset in their model simulations, the researchers can establish initial conditions, perform spin-up runs, and investigate the impacts of historical changes in climate and environmental factors on ecosystems under different nitrogen cycle configurations. This approach enables a more accurate representation of ecosystem behavior and responses, taking into account the influence of climate variables on the simulation results, and ultimately leading to a better understanding of ecosystem dynamics and the factors that drive them.

1. メタデータ作成の効率化
2. データセット論文からのメタデータ自動生成
3. メタデータの翻訳
4. 関連論文の要約
5. データセット利用方法の整理

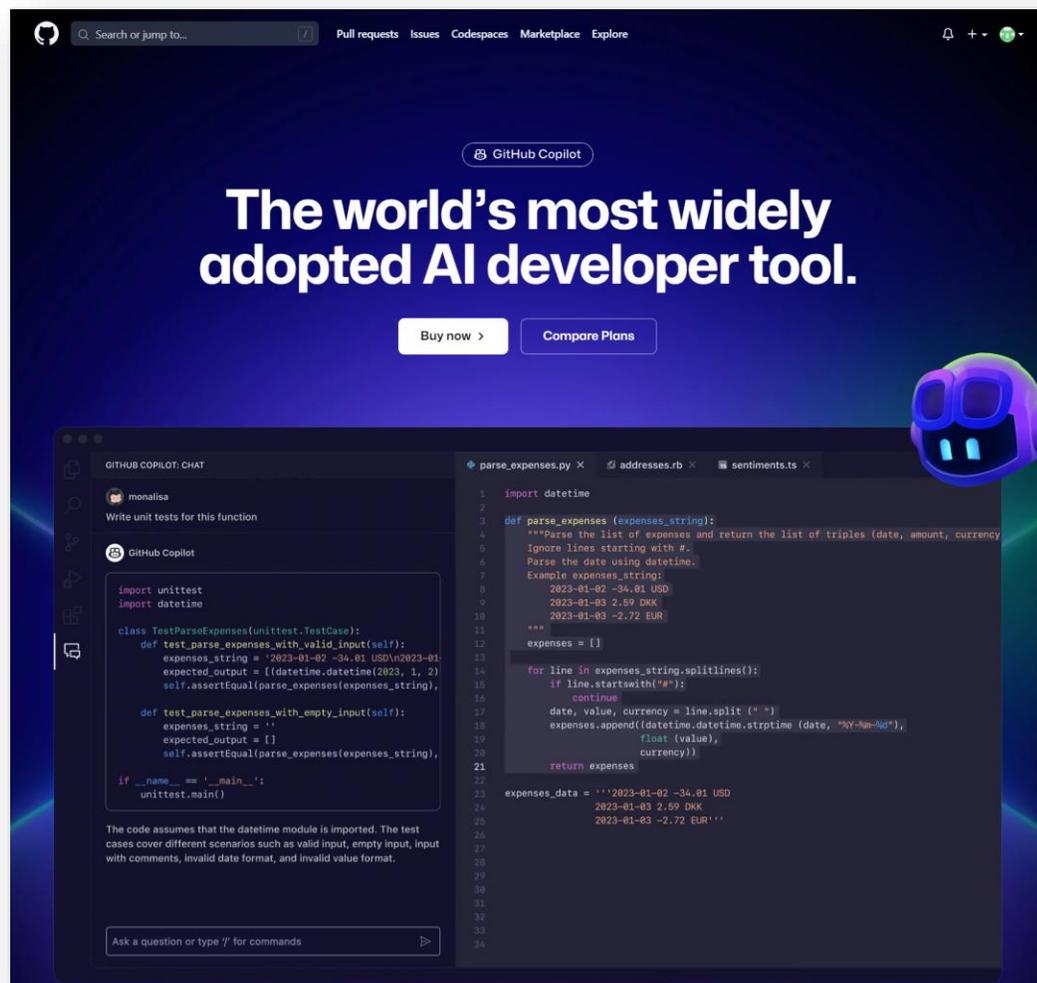
検索の高度化



<https://www.hopsworks.ai/dictionary/retrieval-augmented-generation-llm>

1. 自然言語文の直接入力による意味的な検索
2. 自然言語文をクエリ言語（例：SPARQL）に変換
3. 検索結果の整理や要約、複数DBを統合した検索結果表示
4. ユーザのレベルに応じた検索結果生成

コード生成



1. データセットの構造を反映したコードの自動生成
2. LLMコード生成用のメタデータの準備
3. データセット分析・可視化のためのコード生成

データリポジトリとLLM勉強会

<https://dias.ex.nii.ac.jp/llm/>

1. データリポジトリを対象に、LLM（生成AI）をどのように活用するかを考える勉強会
2. DIAS（Data Integration and Analysis System）での取り組みに閉じず、みんなでアイデアを共有する
3. 気軽に情報交換できる場とし、実験的な試みやコードについても共有する（参考：LLM勉強会）
4. **どんな勉強会になるかわかりませんが、関心のある方はウェブサイトへ！**