

# データリポジトリとLLM勉強会 趣旨説明

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）



## LLM 勉強会

本勉強会では、自然言語処理および計算機システムの研究者が集まり大規模言語モデルの研究開発について定期的に情報共有を行っています。

具体的には、以下の目的で活動しています。

- オープンソースかつ日本語に強い大規模モデルの構築とそれに関連する研究開発の推進
- 上記に関心のある自然言語処理および関連分野の研究者によるモデル構築の知見や最近の研究の発展についての定期的な情報交換
- データ・計算資源等の共有を前提とした組織横断的な研究者間の連携の促進
- モデル・ツール・技術資料等の成果物の公開

News

<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



Explain the use of DIAS dataset in two paragraphs

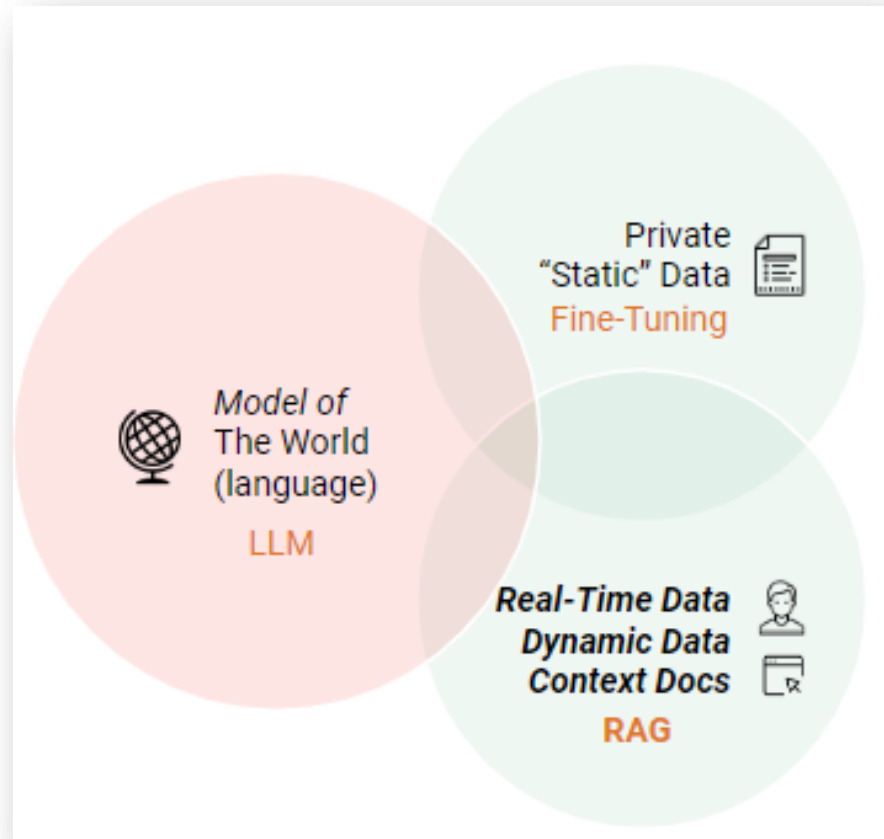


The DIAS dataset, specifically the Global Soil Wetness Project Phase 3 (GSWP3), plays a crucial role in the study as a source of climate forcing data required to run the Community Land Model (CLM). The GSWP3 dataset contains essential climate variables such as temperature, precipitation, and radiation. 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

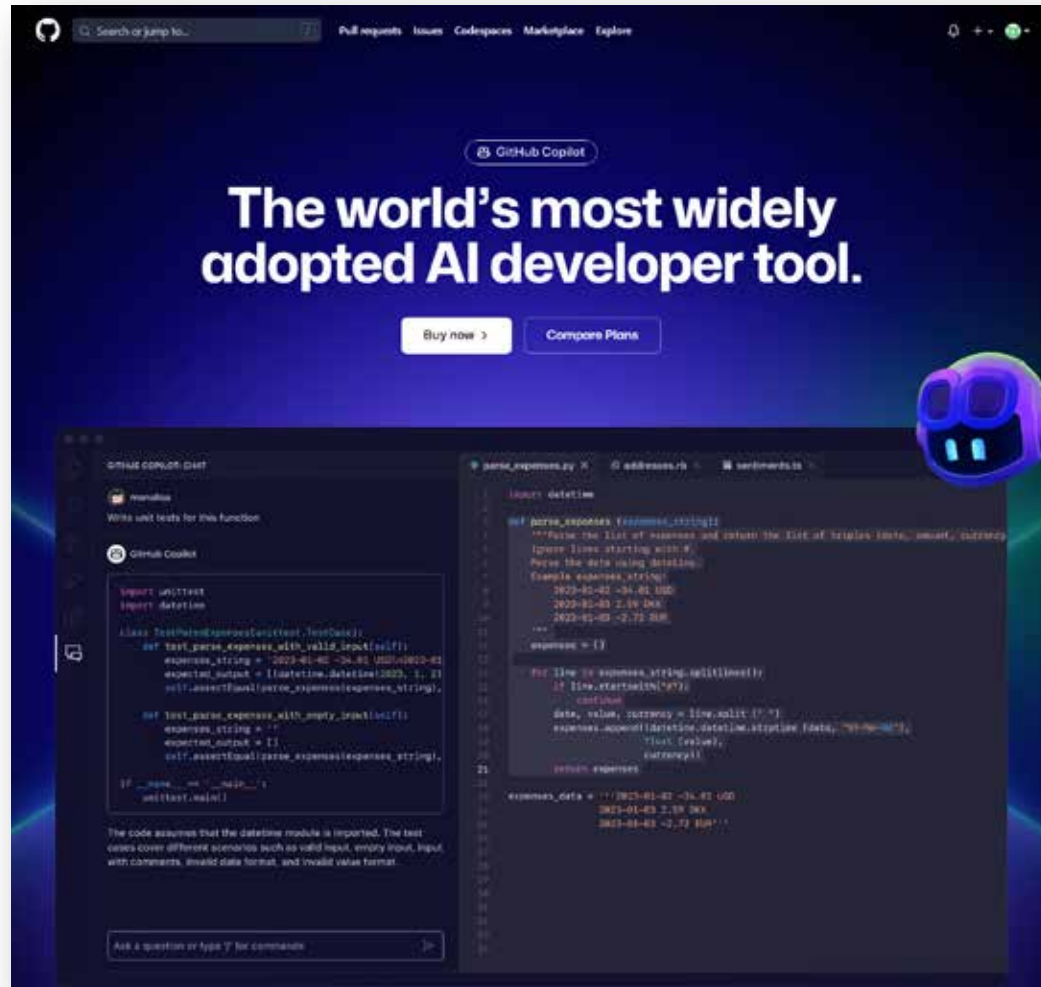
# 検索の高度化



<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

# 利活用の円滑化



1. LLM拡張用のメタデータの付与
2. データセットのスキーマを反映したコード生成
3. データセット分析・可視化のためのコード生成

# LLMの強みと弱み

1. LLMは「言語モデル」であり、**自然言語／人工言語で表現された「言語の世界」**を対象としている
2. 「言語の世界」で完結するタスクは強い（定型的なメール返信、一般的な悩み相談など）
3. **特定分野の正確な知識や、知識の更新**が求められるタスクでは、外部のサービスと連携する必要がある
4. **LLM単体で完結するタスクと、LLMを外部サービスで補強すべきタスク**とを区別する必要がある

# 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）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. 皆さんが興味をもつテーマについて、情報共有と気軽な議論が行えるようにしたい