バイオデータサイエンス時代の統合化推進プログラム

伊藤 隆司（九州大学 医学研究院）

統合化推進プログラム（DICP）

- 2011年に第１期が開始
  - 過去３期で計31課題を支援

- 2022年から第４期が開始
  - 「つなぐ」から「使う」へ
  - 広範なユーザーの知識発見の支援
  - 国際的プレゼンス
  - 新しい動向への対応
  - 2023年から育成型も開始

研究アドバイザー

鎌田　真由美	北里大学　未来工学部　教授
坂井　寛章	農業・食品産業技術総合研究機構　高度分析研究センター　ユニット長
清水　佳奈	早稲田大学　理工学術院　教授
瀬々　潤	（株）ヒューマノーム研究所　代表取締役社長
馬場　健史	九州大学　生体防御医学研究所　教授
山本　一夫	お茶の水女子大学　ヒューマンライフサイエンス研究所　客員教授
吉田　哲郎	アクセリード（株）　経営企画部　シニアディレクター

（2024年9月現在；五十音順に掲載）

DICPがサポート中のDB #1

2022年度採択（本格型）

- バイオイメージングデータのグローバルなデータ共有システムの構築
  - 大浪 修一
- 統合的な転写制御データ基盤の構築
  - 粕川 雄也
- ヒトゲノム・病原体ゲノムと疾患・医薬品をつなぐ統合データベース
  - 金久 實
- 異分野融合を志向した糖鎖科学ポータルのデータ拡充と品質向上
  - 木下 聖子
- 蛋白質構造データバンクのデータ駆動型研究基盤への拡張
  - 栗栖 源嗣
- マイクロバイオーム研究を先導するハブを目指した微生物統合データベースの特化型開発
  - 森 宙史

SSBD:database

intrared

KEGG Kyoto Encyclopedia of Genes and Genomes

GlyCosmos

PDBj Protein Data Bank Japan

Microbiome Datahub

DICPがサポート中のDB #2

2023年度採択（本格型）

- jPOST prime：コミュニティ連携を基盤とするプロテオームデータベース環境の実現
  - 石濱 泰
- 次世代低分子マススペクトルデータベース　シン・マスバンクの構築
  - 松田 史生

2023年度採択（育成型）

- 非モデル植物のための遺伝子ネットワーク情報活用基盤
  - 大林 武
- 日本人塩基配列情報の公開可能なゲノム・オミクス情報基盤による双方向型研究教育データベース開発と国際連携
  - 長崎 正朗
- 空間オミックスデータ解析用データベースの開発
  - VANDENBON Alexis

DICPがサポート中のDB #3

2024年度採択（育成型）

- AI駆動型データキュレーションによる持続可能な中分子相互作用統合データベースの開発
  - 池田　和由
- 細胞レベルの機能・表現型と遺伝子発現を関連付ける「Cell IO」データベースの開発
  - 尾崎　遼
- 創発的再解析のためのメタボローム統合データベース
  - 早川　英介

2024年度応募課題の特徴
提案の60%が生成AI・LLMを何らかの形でDB構築に活用

統合化推進プログラム (DICP)

- 2011年に第１期が開始（トーゴーの日シンポジウムも）
  - 過去３期で計31課題を支援

- 2022年から第４期が開始
  - 「つなぐ」から「使う」へ
  - 広範なユーザーの知識発見の支援
  - 国際的プレゼンス
  - 新しい動向への対応　☞　新しいタイプのデータ　＋　AI活用
  - 2023年から育成型も開始

バイオデータサイエンス時代のDB構築

- DBはバイオサイエンスに必須の基盤
  AIの力を発揮させるには良質のDBが重要
- DB構築者はバイオサイエンスのエッセンシャルワーカー
  DB構築者にとってのAI = assistant intelligent
- DB構築の自動化？（Automated Research Workflow?）
  AIがバイオサイエンスのエッセンシャルワーカー？
- AIによるDB設計？
  AIの方が新規性の高い研究を提案可能？


nature › news › article

NEWS | 20 September 2024

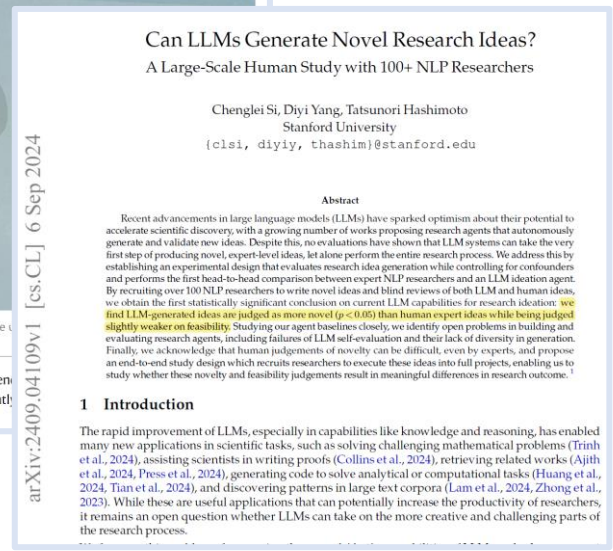
Do AI models produce more original ideas than researchers?

The concepts were judged by reviewers. They were not told who or what had created them.

By Gemma Conroy

Researchers built an artificial intelligence tool that came up with ideas in hours. Credit: Malte Mueller/Getty

An ideas generator powered by artificial intelligence came up with more novel ideas than did 50 scientists working independently on the same topics, this month[1].


Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

Abstract

Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation. Finally, we acknowledge that human judgements of novelty can be difficult, even by experts, and propose an end-to-end study design which recruits researchers to execute these ideas into full projects, enabling us to study whether these novelty and feasibility judgements result in meaningful differences in research outcome.[1]

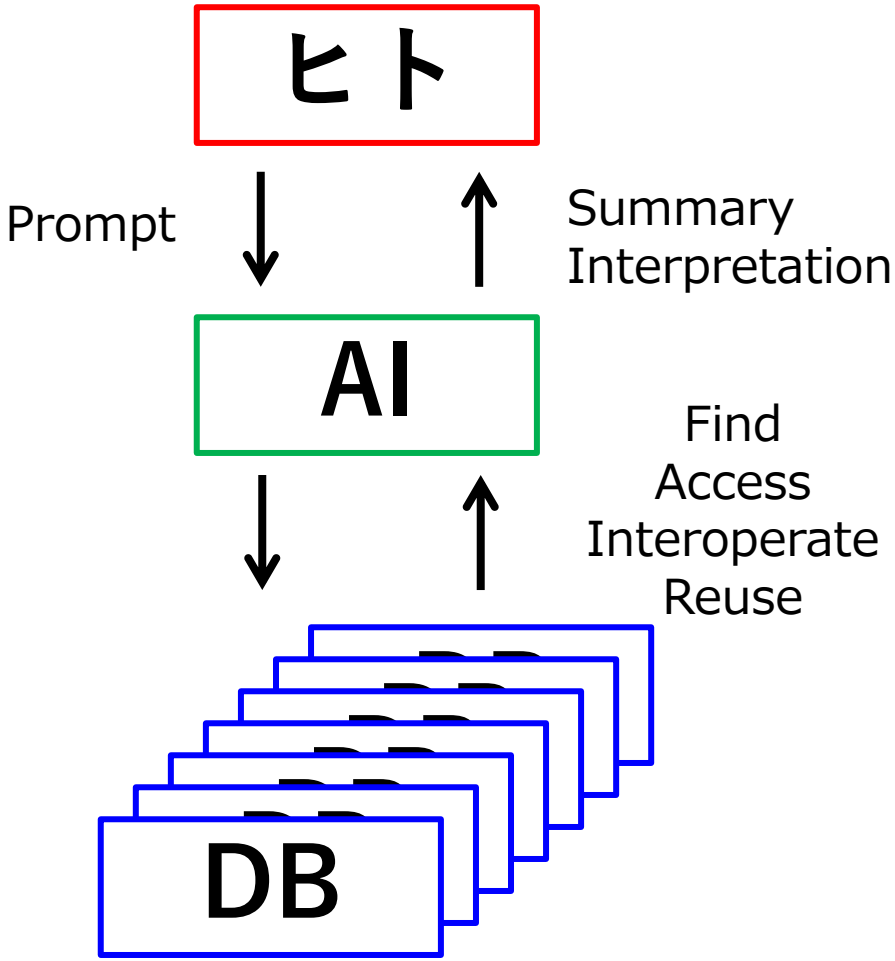
1 Introduction

The rapid improvement of LLMs, especially in capabilities like knowledge and reasoning, has enabled many new applications in scientific tasks, such as solving challenging mathematical problems (Trinh et al., 2024), assisting scientists in writing proofs (Collins et al., 2024), retrieving related works (Ajith et al., 2024, Press et al., 2024), generating code to solve analytical or computational tasks (Huang et al., 2024, Tian et al., 2024), and discovering patterns in large text corpora (Lam et al., 2024, Zhong et al., 2023). While these are useful applications that can potentially increase the productivity of researchers, it remains an open question whether LLMs can take on the more creative and challenging parts of the research process.

https://doi.org/10.1038/d41586-024-03070-5
https://arxiv.org/abs/2409.04109

バイオデータサイエンス時代のDB利用

- クロスボーダーを目指すDB利用者の戸惑い
  - 統合検索の困難
  - 膨大な検索結果
- 求められるものは統合検索の代行と結果の要約解釈？
  - DB利用者にとってのAI = agent intelligent
- DBの利用者はヒトとAI？
  - Find, Access, Interoperate, Reuseの主語は誰？



統合化推進プログラム (DICP)

- 2011年に第１期が開始
  - 過去３期で計31課題を支援

- 2022年から第４期が開始
  - 「つなぐ」から「使う」へ
  - 広範なユーザーの知識発見の支援
  - 国際的プレゼンス
  - 新しい動向への対応
  - 2023年から育成型も開始

研究アドバイザー

鎌田　真由美	北里大学　未来工学部　教授
坂井　寛章	農業・食品産業技術総合研究機構　高度分析研究センター　ユニット長
清水　佳奈	早稲田大学　理工学術院　教授
瀬々　潤	（株）ヒューマノーム研究所　代表取締役社長
馬場　健史	九州大学　生体防御医学研究所　教授
山本　一夫	お茶の水女子大学　ヒューマンライフサイエンス研究所　客員教授
吉田　哲郎	アクセリード（株）　経営企画部　シニアディレクター

☞　育成型３課題のトーク　＋　ポスター　＋　リーフレット