GeneAgent: self-verification language agent for gene-set analysis using domain databases

Zhiyong Lu, PhD FACMI FIAHSI
Senior Investigator, NLM, NIH
Adjunct Professor, CS UIUC

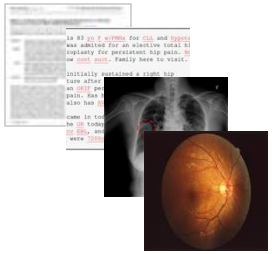
Bioinformatics Community Fair
NIH Research Festival
September 9, 2025

Our research at NIH IRP

- Research Areas
  - AI & Machine Learning, LLMs
  - Natural Language Processing (NLP)
  - Medical Image Analysis

NIH Intramural Research Program
Our Research Changes Lives

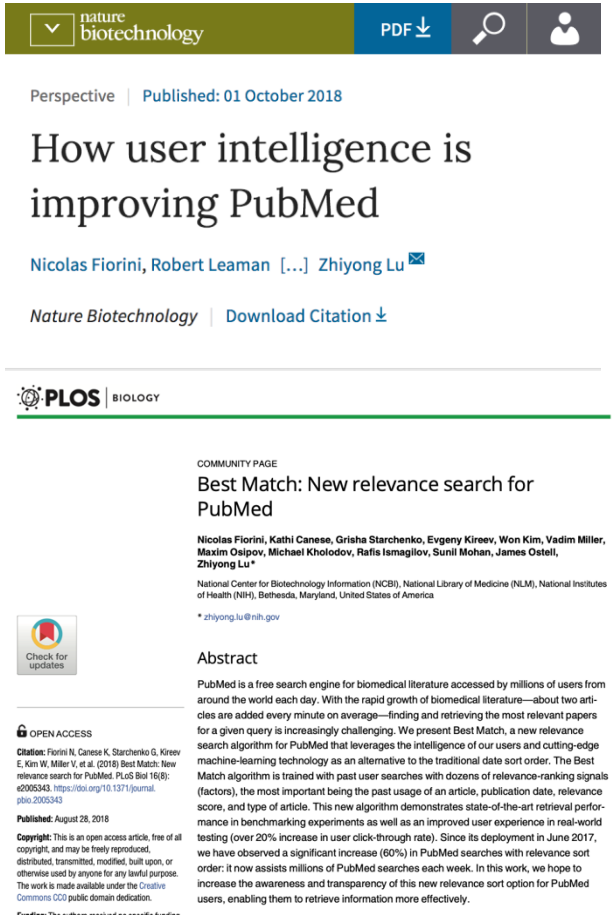
- Multimodal data analysis
  - Biomedical Literature
  - Clinical notes, EHRs
  - CT, CXR & retinal images



- LLM-powered applications in biomedicine
  1. TrialGPT: Assisting patient-to-trial matching with LLMs (2024)
  2. GeneAgent: AI agent for gene set analysis (2025)

AI in PubMed - serving millions each day

- Related articles
- Spell checker
- Query autosuggest
- Semantic query understanding
- Citation sensor
- Author name disambiguation
- Query expansion
- Best Match: Sort by Relevance
- …



nature biotechnology

PDF

Perspective | Published: 01 October 2018

How user intelligence is improving PubMed

Nicolas Fiorini, Robert Leaman […] Zhiyong Lu

Nature Biotechnology | Download Citation

PLOS BIOLOGY

COMMUNITY PAGE

Best Match: New relevance search for PubMed

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, James Ostell, Zhiyong Lu*

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland, United States of America

* zhiyong.lu@nih.gov

Check for updates

OPEN ACCESS

Citation: Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. (2018) Best Match: New relevance search for PubMed. PLoS Biol 16(8): e2005343. https://doi.org/10.1371/journal.pbio.2005343

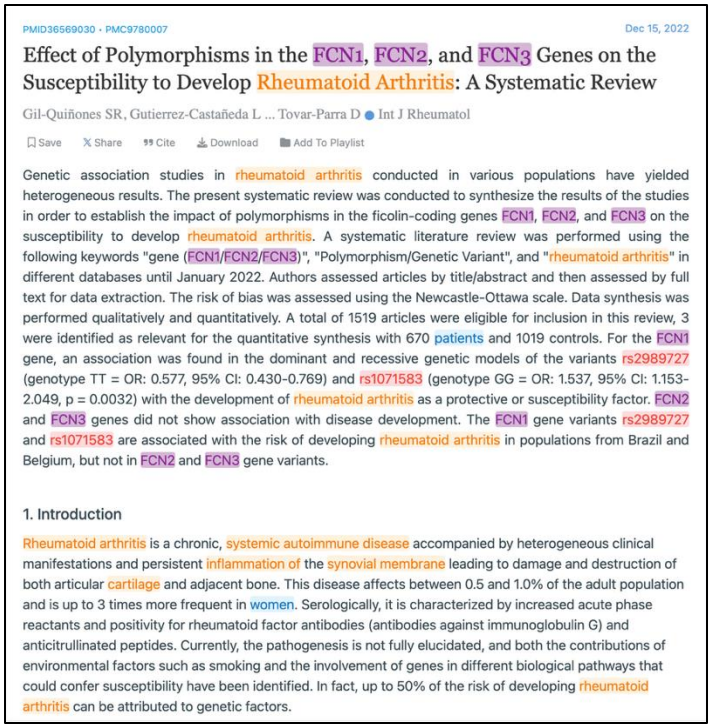
Published: August 28, 2018

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Abstract

PubMed is a free search engine for biomedical literature accessed by millions of users from around the world each day. With the rapid growth of biomedical literature—about two articles are added every minute on average—finding and retrieving the most relevant papers for a given query is increasingly challenging. We present Best Match, a new relevance search algorithm for PubMed that leverages the intelligence of our users and cutting-edge machine-learning technology as an alternative to the traditional date sort order. The Best Match algorithm is trained with past user searches with dozens of relevance-ranking signals (factors), the most important being the past usage of an article, publication date, relevance score, and type of article. This new algorithm demonstrates state-of-the-art retrieval performance in benchmarking experiments as well as an improved user experience in real-world testing (over 20% increase in user click-through rate). Since its deployment in June 2017, we have observed a significant increase (60%) in PubMed searches with relevance sort order: it now assists millions of PubMed searches each week. In this work, we hope to increase the awareness and transparency of this new relevance sort option for PubMed users, enabling them to retrieve information more effectively.

Biomedical Literature Mining



Chen et al., *Nature* 2020; Allot et al., *Nature Genetics*, 2023; Wei et al., *Nucleic Acids Research* 2024

ChatGPT: revolution or hype?

Google v Microsoft: who will win the AI chatbot race?

Bard's misfire on launch cost owner $160bn but experts believe ChatGPT is also prone to errors

10 Feb 2023

Who said it: an Australian MP or ChatGPT?

10 Feb 2023

'ChatGPT needs a huge amount of editing': users' views mixed on AI chatbot

8 Feb 2023

US experts warn AI likely to kill off jobs – and widen wealth inequality

8 Feb 2023

Microsoft to power Bing with AI as race with Google heats up

7 Feb 2023

How will Google and Microsoft AI chatbots affect us and how we work?

7 Feb 2023

Google trials its own AI chatbot Bard after success of ChatGPT

6 Feb 2023

MP tells Australia's parliament AI could be used for 'mass destruction' in speech part-written by ChatGPT

6 Feb 2023

The networker

"ChatGPT isn't a great leap forward, it's an expensive deal with the devil *John Naughton*

4 Feb 2023 312

Google poised to release chatbot technology after ChatGPT success

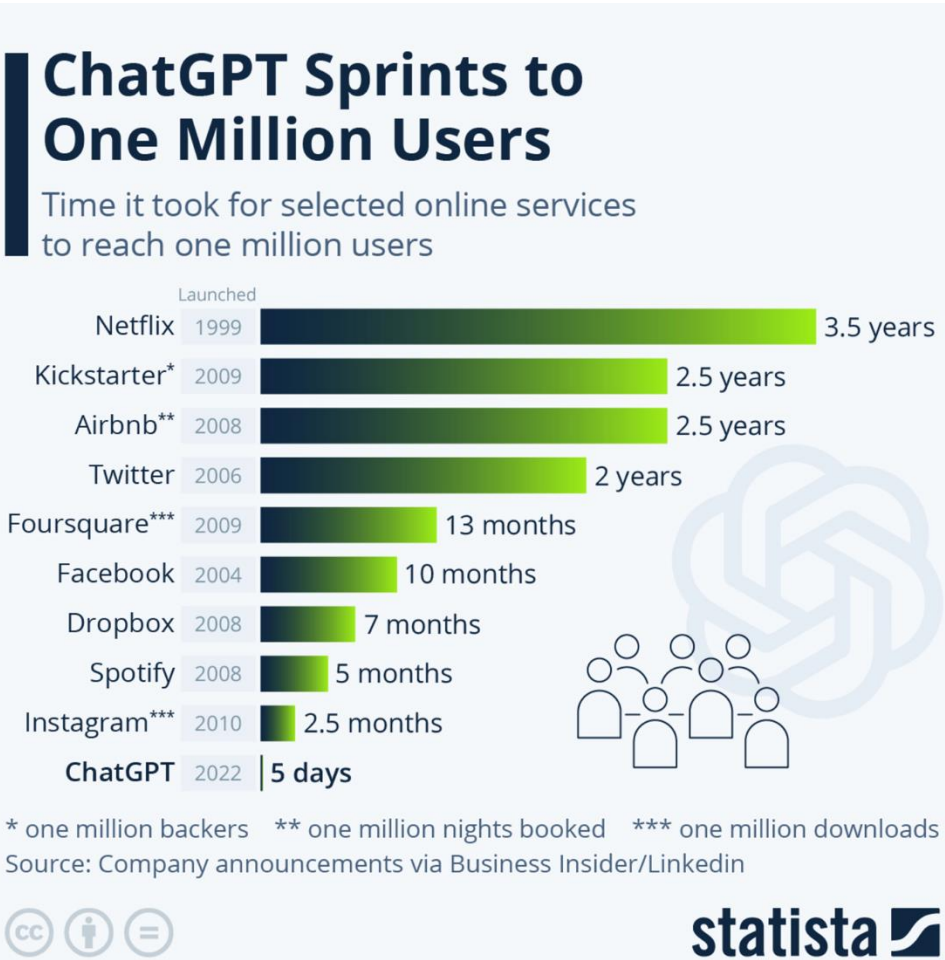
3 Feb 2023

Colombian judge says he used ChatGPT in ruling

2 Feb 2023

ChatGPT reaches 100 million users two months after launch

2 Feb 2023



WIRED

Dr. ChatGPT Will See You Now

24 days ago

CNN Health | Life, But Better | Fitness | Food | Sleep | Mindfulness | More

AI may be as effective as medical specialists at diagnosing disease

By Jack Guy, CNN

The Guardian

First NHS AI-run physio clinic in England halves back-pain waiting list

3 days ago

WSJ Wall Street Journal

Companies Bring AI Agents to Healthcare

Feb 27, 2025

The Washington Post

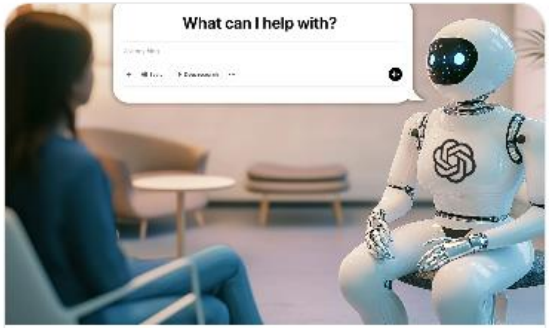
ChatGPT is little help for doctors in diagnosing diseases, study finds

The research, conducted with 50 physicians last year, found that using ChatGPT did not significantly improve doctors' diagnostic reasoning.

CNN · 11d

FDA's artificial intelligence is supposed to revolutionize drug approvals. It's making up studies

Insiders tell CNN the FDA's AI is "hallucinating" studies and can't access ...



New York Post

Harmful AI therapy: Chatbots endanger users with suicidal thoughts, delusions, researchers warn

Jun 28, 2025

NIH National Institutes of Health
Turning Discovery Into Health

MENU

News & Events

Recent News Releases

NIH researchers develop AI agent that improves accuracy of gene set analysis by leveraging expert-curated databases

July 28, 2025 — The AI agent could help lead to a better understanding of how different diseases and conditions affect groups of genes individually and together.

NIH National Institutes of Health
Turning Discovery Into Health

MENU

Home › News & Events › News Releases

Tuesday, July 23, 2024

NIH findings shed light on risks and benefits of integrating AI into medical decision-making

Researchers at the National Institutes of Health (NIH) found that an artificial intelligence (AI) model solved medical quiz questions—designed to test health professionals' ability to diagnose patients based on clinical images and a brief text summary—with high accuracy. However, physician-graders found the AI model made mistakes when describing images and explaining how its decision-making led to the correct answer. The findings, which shed light on AI's potential in the clinical setting, were published in *npj Digital Medicine*. The study was led by researchers from NIH's National Library of Medicine (NLM) and Weill Cornell Medicine, New York City.

"Integration of AI into health care holds great promise as a tool to help medical professionals diagnose

GPT-4V, an AI model, often made mistakes when describing the medical image and explaining its reasoning behind the diagnosis—even in cases where it made the correct final choice. *NIH/NLM*

NIH National Institutes of Health

NEWS RELEASES

Monday, November 18, 2024

NIH-developed AI algorithm matches potential volunteers to clinical trials

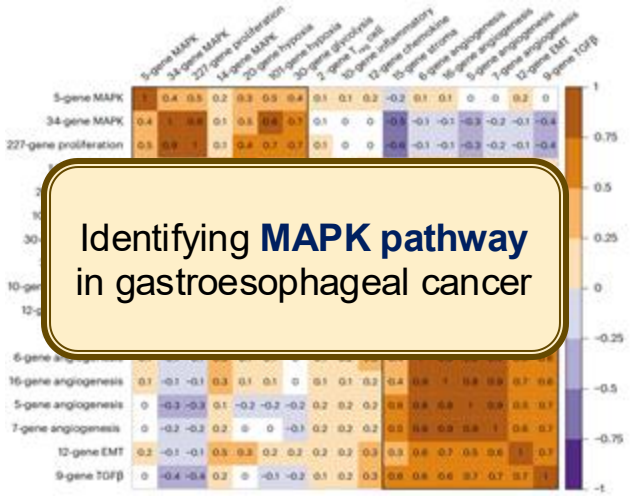
Such an algorithm may save clinicians time and accelerate clinical enrollment and research.

Intro: Gene Set Analysis

- Gene sets from high-throughput experiments
- Differentially expressed genes under different conditions
- Data analysis goal: determine the collective functions by a group of genes

GSA applications in drug discovery

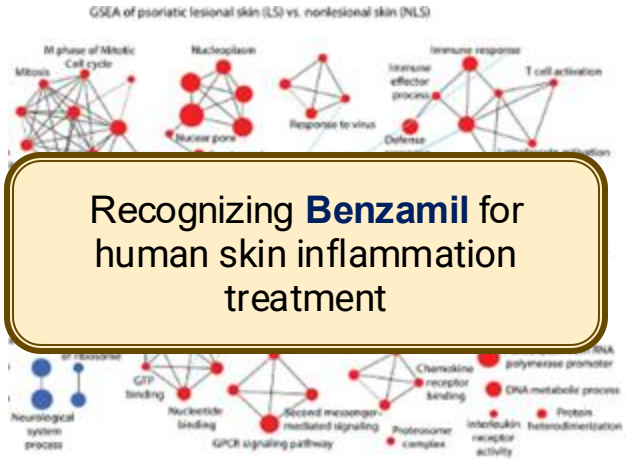
Disease Mechanisms



Identifying **MAPK pathway** in gastroesophageal cancer

Shitara, K., Janjigian, Y.Y., Ajani, J. et al. **Nivolumab plus chemotherapy or ipilimumab in gastroesophageal cancer: exploratory biomarker analyses of a randomized phase 3 trial**. Nat Med 31, 1519–1530 (2025).

Drug Repositioning

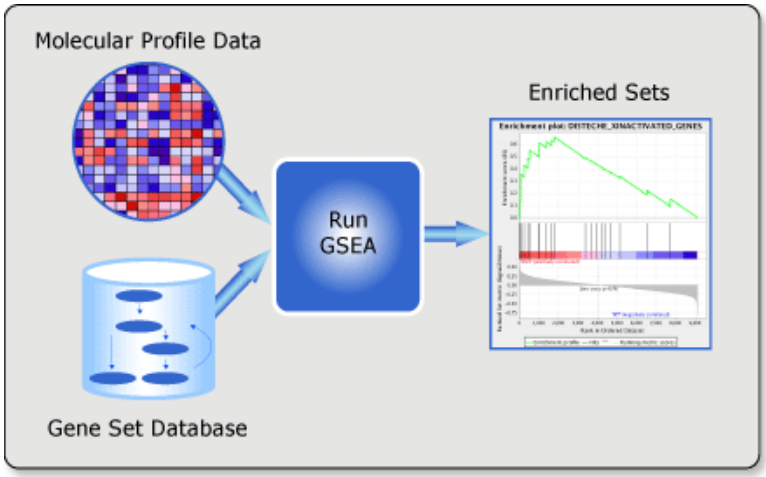


Recognizing **Benzamil** for human skin inflammation treatment

Mårten C. G. Winge et al. , **Repurposing an epithelial sodium channel inhibitor as a therapy for murine and human skin inflammation**. Sci. Transl. Med. 16, eade5915 (2024).

Existing GESA methods & their limitations

- Limited to curated knowledge and/or predefined gene sets
- No explanation to support predictions



Subramanian, Tamayo, et al. (2005, PNAS); Mootha, Lindgren, et al. (2003, Nature Genetics).

Related works

[Submitted on 21 May 2023 (v1), last revised 25 May 2023 (this version, v2)]

Gene Set Summarization using Large Language Models

Marcin P. Joachimiak, J. Harry Caufield, Nomi L. Harris, Hyeongsik Kim, Christopher J. Mungall

Molecular biologists frequently interpret gene lists derived from high-throughput experiments and computational analysis. This is typically done as a statistical enrichment analysis that measures the over- or under-representation of biological function terms associated with genes or their properties, based on curated assertions from a knowledge base (KB) such as the Gene Ontology (GO). Interpreting gene lists can also be framed as a textual summarization task, enabling the use of Large Language Models (LLMs), potentially utilizing scientific texts directly and avoiding reliance on a KB.
We developed SPINDOCTOR (Structured Prompt Interpolation of Natural Language Descriptions of Controlled Terms for Ontology Reporting), a method tl
summarization as a complement to standard enrichme
sources of gene functional information: (1) structured
annotations, (2) ontology-free narrative gene summari
We demonstrate that these methods are able to genera
term lists for gene sets. However, GPT-based approacl
values and often return terms that are not statistically
able to recapitulate the most precise and informative t
inability to generalize and reason using an ontology. R
variations in prompt resulting in radically different terr
based methods are unsuitable as a replacement for sta
curation of ontological assertions remains necessary.

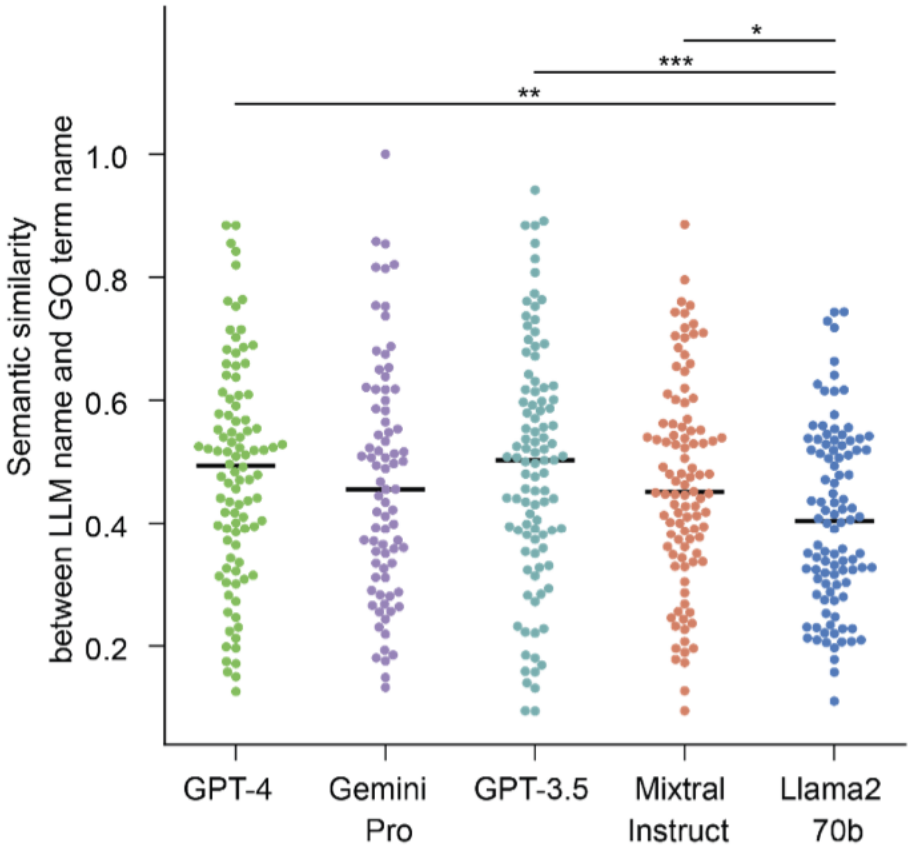
[Submitted on 7 Sep 2023 (v1), last revised 1 Apr 2024 (this version, v2)]

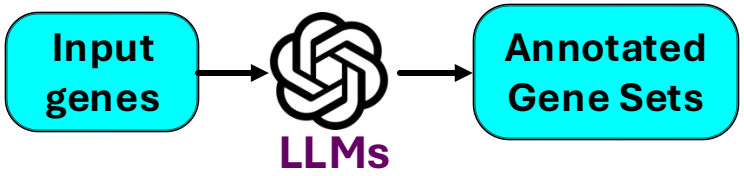
Evaluation of large language models for discovery of gene set function

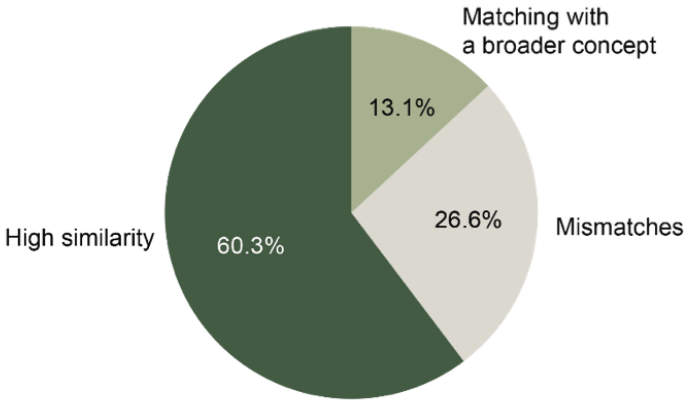
Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, Dexter Pratt

Gene set analysis is a mainstay of functional genomics, but it relies on curated databases of gene functions that are incomplete. Here we evaluate five Large Language Models (LLMs) for their ability to discover the common biological functions represented by a gene set, substantiated by supporting rationale, citations and a confidence assessment. Benchmarking against canonical gene sets from the Gene Ontology, GPT-4 confidently recovered the curated name or a more general concept (73% of cases), while benchmarking against random gene sets correctly yielded zero confidence. Gemini-Pro and Mixtral-Instruct showed ability in naming but were falsely confident for random sets, whereas Llama2-70b had poor performance overall. In gene sets derived from 'omics data, GPT-4 identified novel functions not reported by classical functional enrichment (32% of cases), which independent review indicated were largely verifiable and not hallucinations. The ability to rapidly synthesize common gene functions positions LLMs as valuable 'omics assistants.

Key results in Hu et al.,







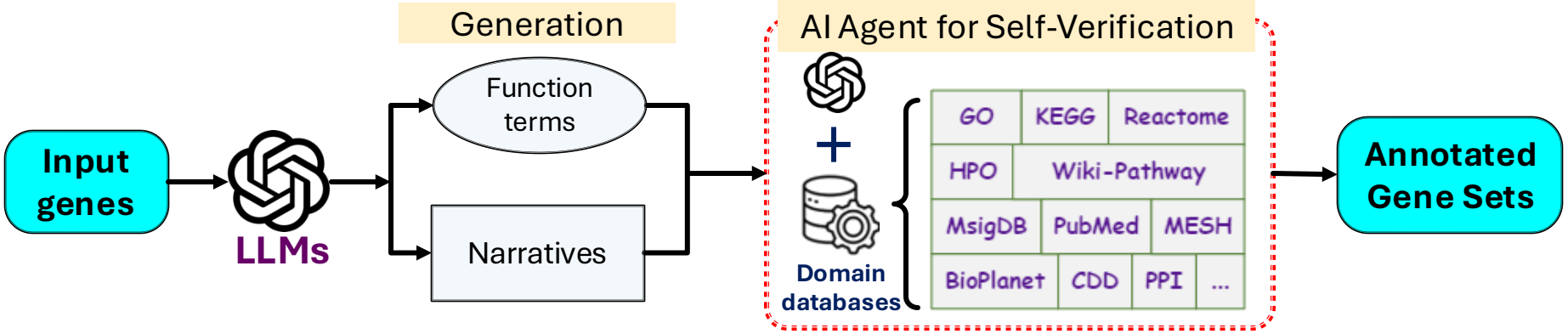
LLM as a judge

- Pros: scalable; consistent
- Cons: Self-preference; limited factual checking
- We propose an **AI agent** that performs automatic self-verification grounded in domain knowledge

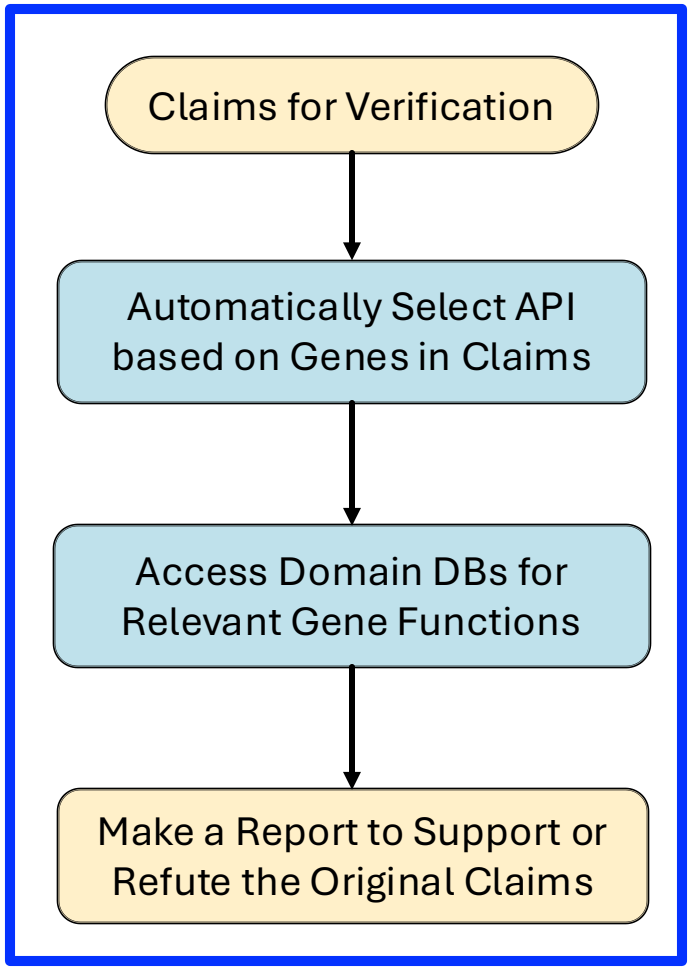
Gene-centric information in expert-curated biological databases

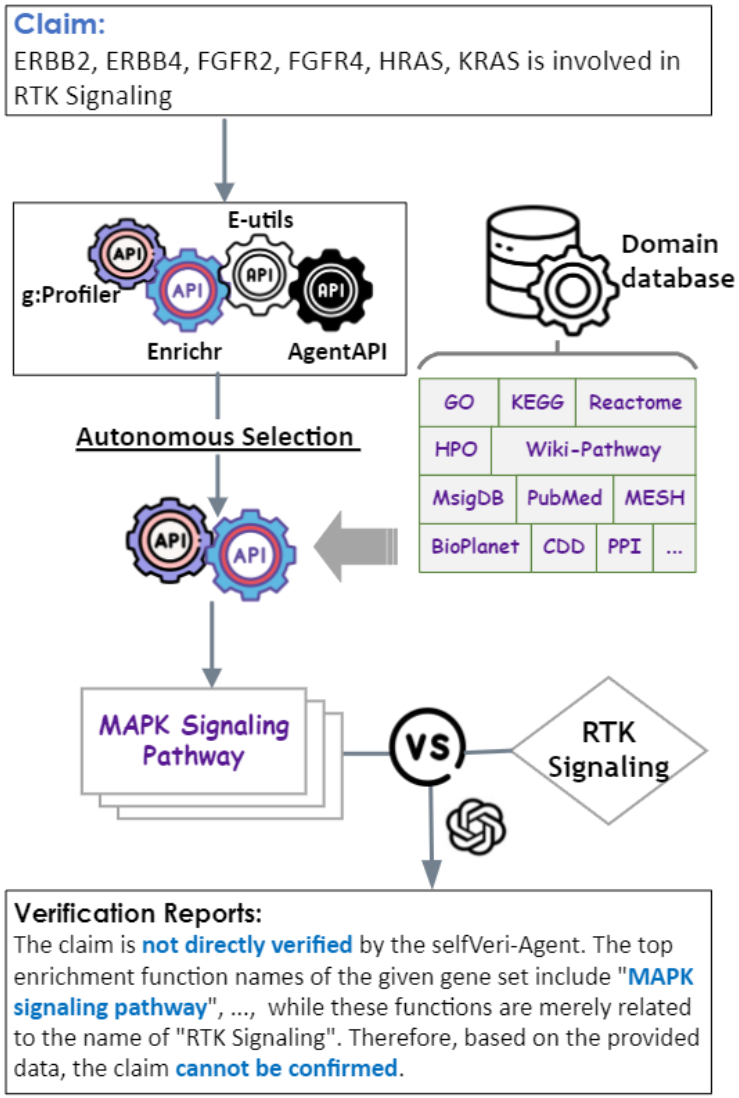
Our GeneAgent Approach

Zhizheng Wang, PhD



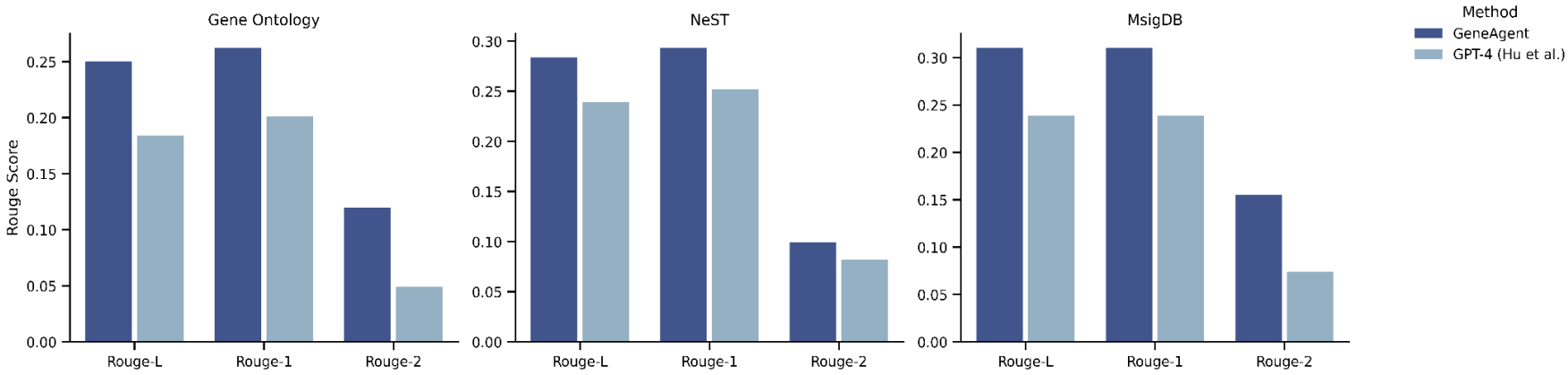
AI Agent for self-Verification





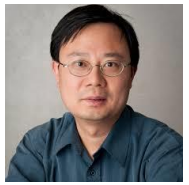
GeneAgent vs. standard GPT-4

Dataset	#gene sets	#genes	Avg. genes
Gene Ontology	1,000	3 to 456	48.32
NeST	50	5 to 323	18.96
MsigDB	56	4 to 200	112.00
All	1,106	3 to 456	50.67



Pilot study with novel gene sets

- To assess its potential utility in real-world applications
- Worked with domain experts from NCI/NLM
- Novel gene sets from mouse B2905 melanoma cell line



Chi-Ping Day PhD
Lab of Cancer Biology and Genetics,
Cancer Data Science Lab, NCI

Christina Ross, Ph.D. NCI
Lab of Cancer Biology and Genetics

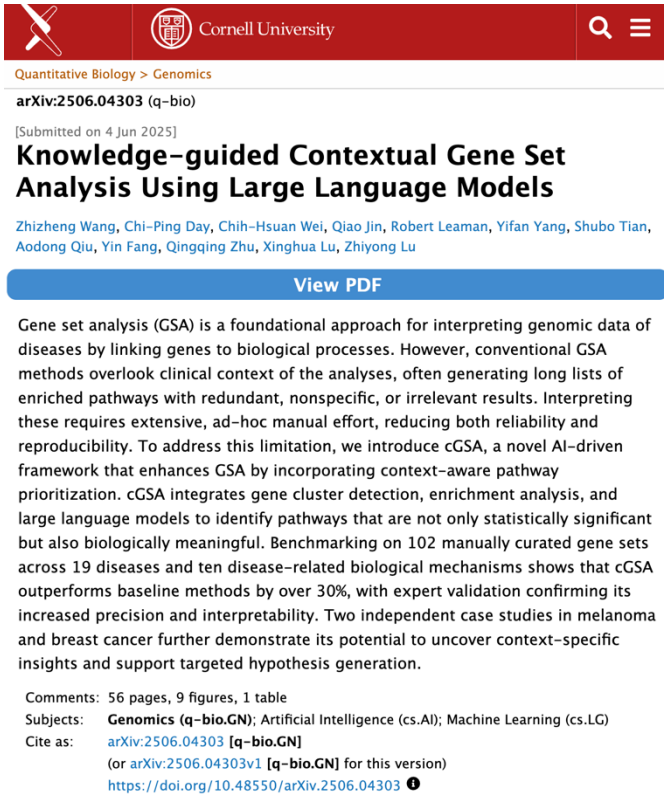
Wang, et al. GeneAgent: Self-verification Language Agent for Gene Set Knowledge Discovery using Domain Databases. *Nature Methods*, 2025.

Evaluation results by 2 domain experts

ID	Generated by GPT-4	Generated by GeneAgent	Gene Coverage	Better Output Annotated by Genomic Experts									
				Relevance		Readability		Consistency		Comprehensive		Final Decision	
				GPT-4	GeneAgent	GPT-4	GeneAgent	GPT-4	GeneAgent	GPT-4	GeneAgent	GPT-4	GeneAgent
mmu05171 (HA-R)	Ribosomal Protein Synthesis	Cytosolic Ribosome and Protein Synthesis	33/36		○	○	○	○	○	○	○		✓
mmu03010 (HA-R)	Ribosomal Protein Synthesis and Assembly	Cytosolic Ribosome	34/35		○	○	○	○	○	○	○		✓
mmu03010 (HA-S)	Ribosomal Protein Synthesis	Cytosolic Ribosome	13/49									×	×
mmu05171 (HA-S)	Ribosomal Protein Synthesis	Cytosolic Ribosome Assembly and Protein Synthesis	47/47		○	○	○		○		○		✓
mmu04015 (HA-S)	MAPK/ERK Pathway Regulation	Rap1 Signaling Pathway	27/27		○	○	○	○	○	○	○		✓
mmu05100 (HA-S)	Caveolae-Mediated Endocytosis and Actin Remodeling	Bacterial Invasion of Epithelial Cells	19/19	○		○	○	○		○		✓	
mmu05022 (LA-S)	Oxidative Phosphorylation and Neurodegeneration	Neurodegeneration and Respiratory Chain Complex	23/24		○	○		○	○		○		✓

Limitations & Future Directions

- Human & mouse genes only
- Very large gene sets
- Add contextual information (study designs, experimental conditions)



Cornell University

Quantitative Biology > Genomics

arXiv:2506.04303 (q-bio)

[Submitted on 4 Jun 2025]

Knowledge-guided Contextual Gene Set Analysis Using Large Language Models

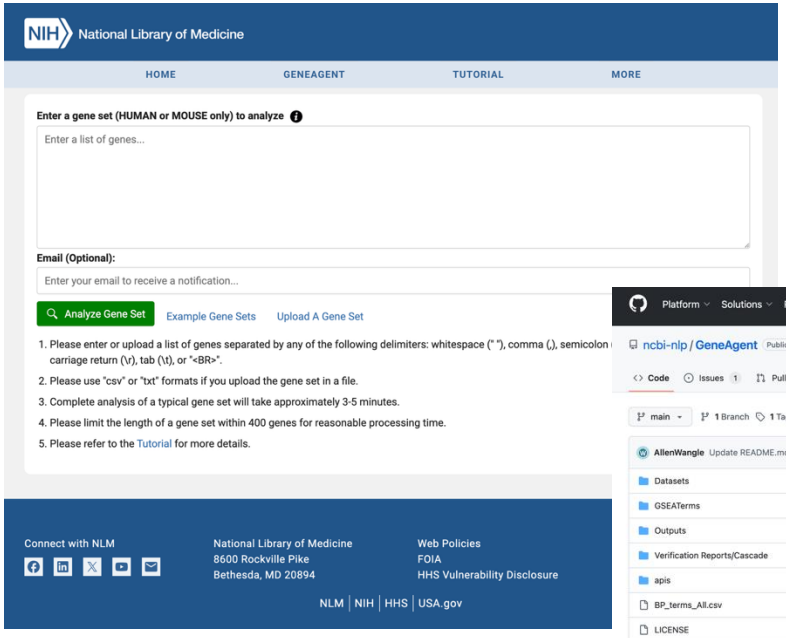
Zhizheng Wang, Chi-Ping Day, Chih-Hsuan Wei, Qiao Jin, Robert Leaman, Yifan Yang, Shubo Tian, Aodong Qiu, Yin Fang, Qingqing Zhu, Xinghua Lu, Zhiyong Lu

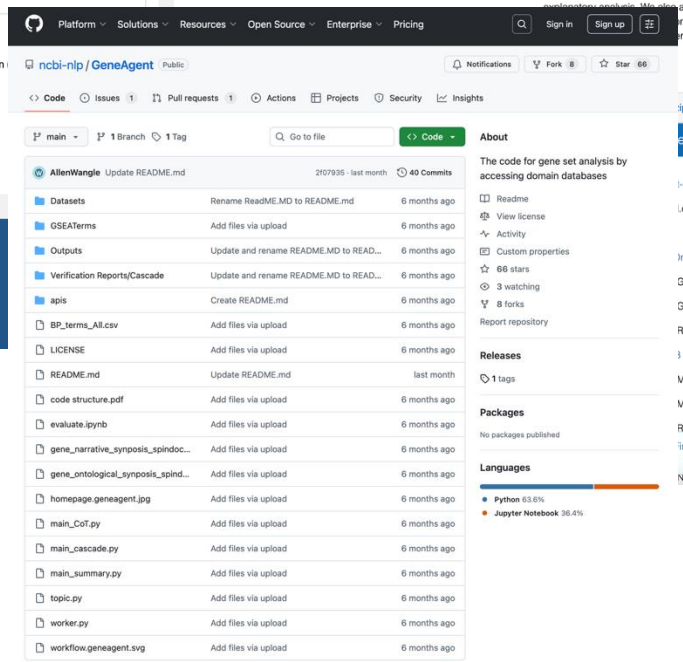
View PDF

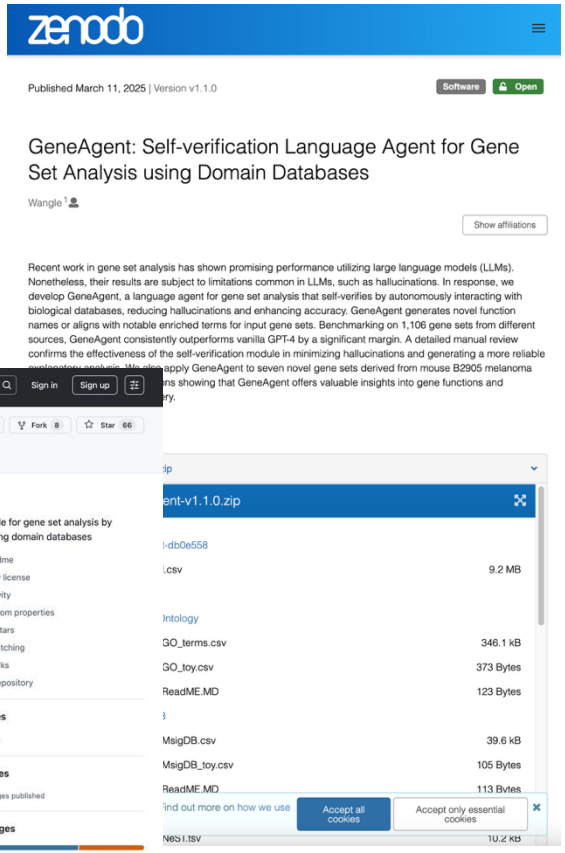
Gene set analysis (GSA) is a foundational approach for interpreting genomic data of diseases by linking genes to biological processes. However, conventional GSA methods overlook clinical context of the analyses, often generating long lists of enriched pathways with redundant, nonspecific, or irrelevant results. Interpreting these requires extensive, ad-hoc manual effort, reducing both reliability and reproducibility. To address this limitation, we introduce cGSA, a novel AI-driven framework that enhances GSA by incorporating context-aware pathway prioritization. cGSA integrates gene cluster detection, enrichment analysis, and large language models to identify pathways that are not only statistically significant but also biologically meaningful. Benchmarking on 102 manually curated gene sets across 19 diseases and ten disease-related biological mechanisms shows that cGSA outperforms baseline methods by over 30%, with expert validation confirming its increased precision and interpretability. Two independent case studies in melanoma and breast cancer further demonstrate its potential to uncover context-specific insights and support targeted hypothesis generation.

Comments: 56 pages, 9 figures, 1 table
Subjects: **Genomics (q-bio.GN)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: arXiv:2506.04303 **[q-bio.GN]**
(or arXiv:2506.04303v1 **[q-bio.GN]** for this version)
https://doi.org/10.48550/arXiv.2506.04303

We welcome your feedback







Current & Future Work

zhiyong.lu@nih.gov

- **Benchmarking stock models on medical tasks**
  - MedCal-Bench: evaluating LLMs for medical calculations (*NeurIPS oral*, 24)
  - Evaluating LLMs on various BioNLP tasks (*Nature Communications*. 2025)
- **Enhancing standard LLMs with domain-specific data or tools**
  - GeneGPT: domain tool learning (*Bioinformatics*, 2024)
  - MedRAG: Best practices with retrieval augmented generation (*ACL 2024, PSB 2025)*
- **Multimodal LLMs in Healthcare**
  - Assessing Progress of Multimodal LLM Performance on Clinical Cases (*Radiology*, 2025)
  - Developing MLLMs with CT images and clinical notes (*Nature Biomedical Engineering*, in press)
- **Agentic AI systems:**
  - TrialGPT: patient-trial matching (*Nature Communications*, 2024)
  - ***GeneAgent*: Gene set analysis with domain knowledge, *Nature Methods*, 2025**
  - *AgentMD:* automating medical risk calculation *(Nature Communications, to appear)*
- **Challenges & Limitations:**
  - Hidden flaws in multi-modal GPT4-V (npj Digital Medicine, 2024)
  - Evaluation beyond Multiple-Choice Accuracy (*Annual Review of Biomedical Data Science*, 2025)
  - Risks of AI Scientists: Prioritizing Safeguarding Over Autonomy (*Nature Communications*, in press)
  - AI model vulnerability under adversarial attacks (*Nature Communications*, to appear)