

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

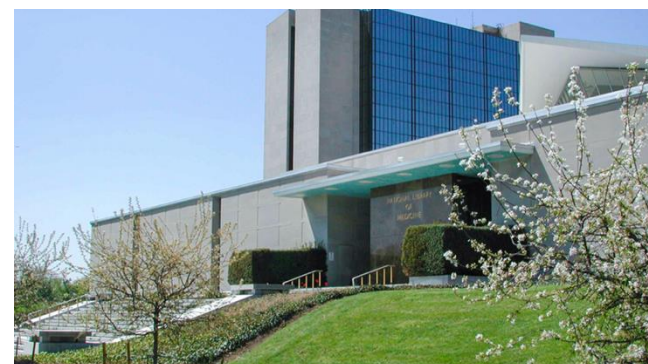
Our research at NIH IRP

➤ Research Areas

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

➤ 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
- ...



Biomedical Literature Mining

LitCovid (NIH NLM) is a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus. It is the most comprehensive resource on the subject, providing a

LitVar² (NIH NLM) Search for variants in more than 35 million biomedical publications. Variant (e.g. CFH R1210C) Optional Text (e.g. AMD)

PubTator³ (NIH NLM) Search entities & relations in 35+ million biomedical publications. Ex: Remdesivir Try: N-dimethylnitrosamine and Metformin COVID-19 and PON1

Entities Autocomplete
PubTator3 uses a high-performance entities search engine, to normalize different forms of the same entity into a unique standardized name to returned all matching articles.

Full Text Search
PubTator3 provides unified access to the entire 35+ million abstracts in PubMed and nearly 6 million full-text articles in the PMC Text Mining subset.

Relations
PubTator3 allows to filter results to only return publications containing specific relations between two entities, such as diseases, chemicals, genes or variants.

PMID36569030 · PMC6978007 Dec 15, 2022

Effect of Polymorphisms in the **FCN1**, **FCN2**, and **FCN3** Genes on the Susceptibility to Develop **Rheumatoid Arthritis**: A Systematic Review

Gil-Quiriones SR, Gutierrez-Castañeda L ... Tovar-Parra D • Int J Rheumatol

Save Share Cite Download Add To Playlist

Genetic association studies in **rheumatoid arthritis** conducted in various populations have yielded heterogeneous results. The present systematic review was conducted to synthesize the results of the studies in order to establish the impact of polymorphisms in the ficolin-coding genes **FCN1**, **FCN2**, and **FCN3** on the susceptibility to develop **rheumatoid arthritis**. A systematic literature review was performed using the following keywords "gene (**FCN1/FCN2/FCN3**)", "Polymorphism/Genetic Variant", and "**rheumatoid arthritis**" in different databases until January 2022. Authors assessed articles by title/abstract and then assessed by full text for data extraction. The risk of bias was assessed using the Newcastle-Ottawa scale. Data synthesis was performed qualitatively and quantitatively. A total of 1519 articles were eligible for inclusion in this review, 3 were identified as relevant for the quantitative synthesis with 670 **patients** and 1019 controls. For the **FCN1** gene, an association was found in the dominant and recessive genetic models of the variants **rs2989727** (genotype TT = OR: 0.577, 95% CI: 0.430-0.769) and **rs1071583** (genotype GG = OR: 1.537, 95% CI: 1.153-2.049, p = 0.0032) with the development of **rheumatoid arthritis** as a protective or susceptibility factor. **FCN2** and **FCN3** genes did not show association with disease development. The **FCN1** gene variants **rs2989727** and **rs1071583** are associated with the risk of developing **rheumatoid arthritis** in populations from Brazil and Belgium, but not in **FCN2** and **FCN3** gene variants.

1. Introduction

Rheumatoid arthritis is a chronic, **systemic autoimmune disease** accompanied by heterogeneous clinical manifestations and persistent **inflammation** of the **synovial membrane** leading to damage and destruction of both articular **cartilage** and adjacent bone. This disease affects between 0.5 and 1.0% of the adult population and is up to 3 times more frequent in **women**. Serologically, it is characterized by increased acute phase reactants and positivity for rheumatoid factor antibodies (antibodies against immunoglobulin G) and anticitrullinated peptides. Currently, the pathogenesis is not fully elucidated, and both the contributions of environmental factors such as smoking and the involvement of genes in different biological pathways that could confer susceptibility have been identified. In fact, up to 50% of the risk of developing **rheumatoid arthritis** can be attributed to genetic factors.

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

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

8 Feb 2023

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

7 Feb 2023

Google trials its own AI chatbot Bard after success of 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

Colombian judge says he used ChatGPT in ruling

2 Feb 2023



Who said it: an Australian MP or ChatGPT?

10 Feb 2023

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

8 Feb 2023

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

7 Feb 2023

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

6 Feb 2023

Google poised to release chatbot technology after ChatGPT success

3 Feb 2023

ChatGPT reaches 100 million users two months after launch

2 Feb 2023

ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users



* one million backers ** one million nights booked *** one million downloads

Source: Company announcements via Business Insider/LinkedIn



statista



WIRED

Dr. ChatGPT Will See You Now

24 days ago

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



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

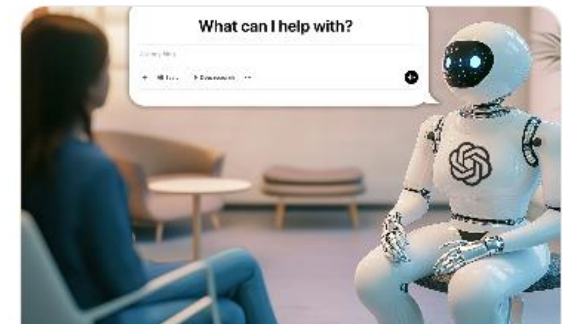
The research, conducted with 50 physicians last year, found that using ChatGPT did not significantly improve doctors' diagnostic reasoning.



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

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.



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.

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

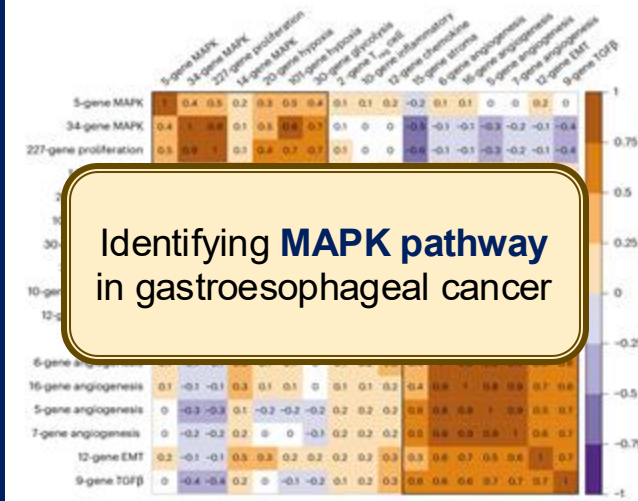
Intro: Gene Set Analysis

- 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



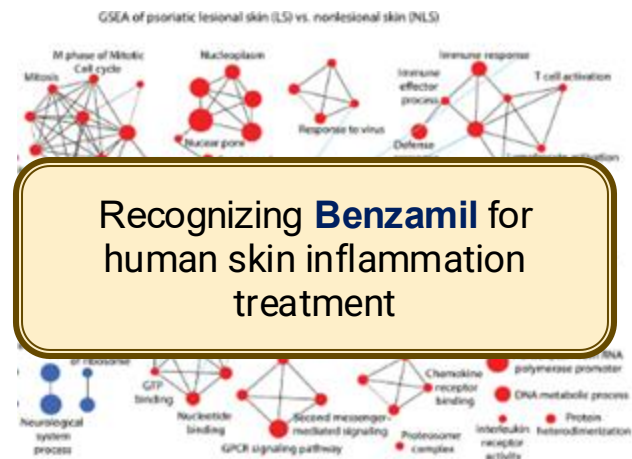
Disease Mechanisms



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).



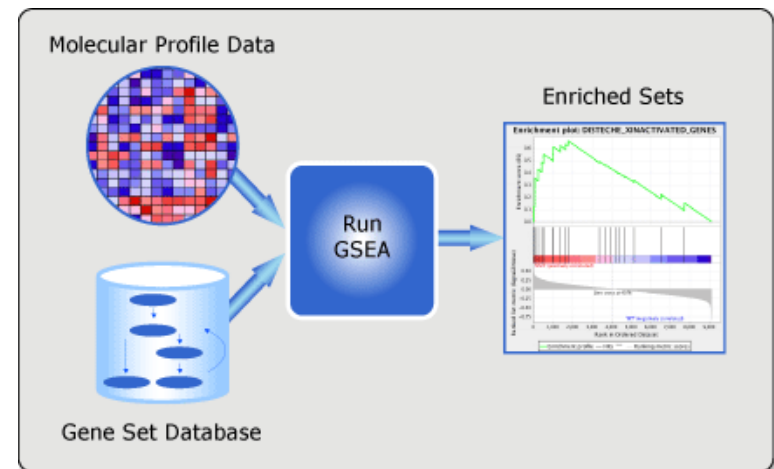
Drug Repositioning



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 GSEA methods & their limitations

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



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

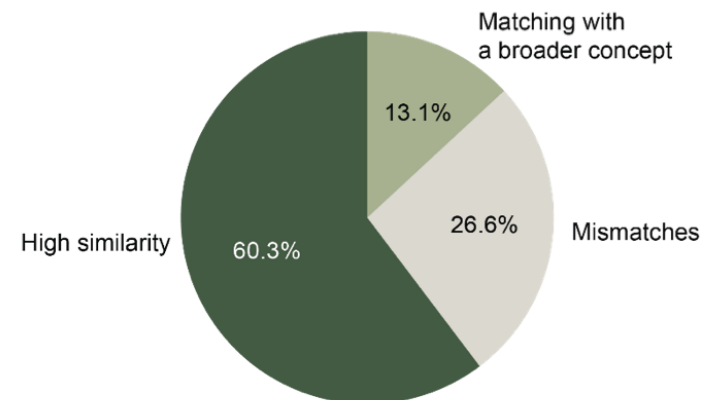
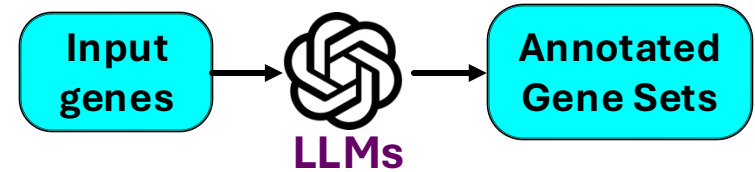
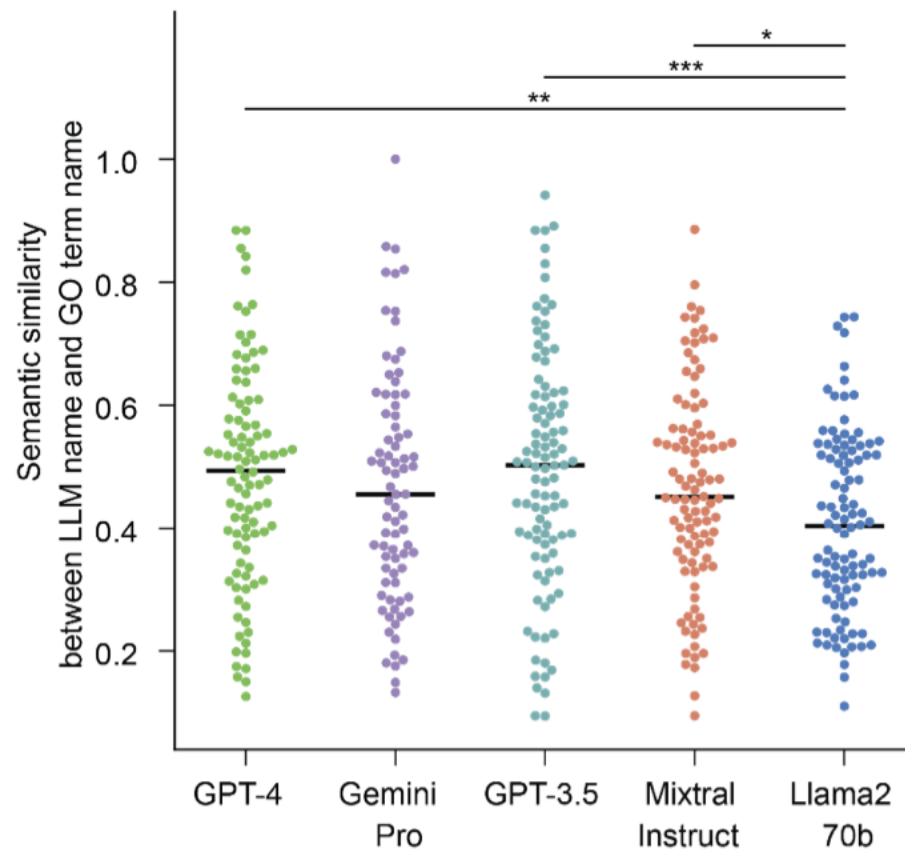
Controlled Terms for Ontology Reporting), a method for gene set summarization as a complement to standard enrichment analysis. We evaluated SPINDOCTOR on two sources of gene functional information: (1) structured annotations, (2) ontology-free narrative gene summaries. We demonstrate that these methods are able to generate term lists for gene sets. However, GPT-based approaches often return terms that are not statistically significant and are often unable to recapitulate the most precise and informative terms. Variations in prompt resulting in radically different term sets suggest that GPT-based methods are unsuitable as a replacement for statistical enrichment analysis. Curation of ontological assertions remains necessary.

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

nature methods

[View all journals](#)

[Search](#)

[Log in](#)

[Explore content](#)

[About the journal](#)

[Publish with us](#)

[nature](#) > [nature methods](#) > [articles](#) > [article](#)

[Download PDF](#)

Article | [Open access](#) | Published: 28 July 2025

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

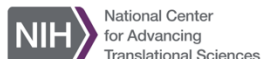
[Zhizheng Wang](#), [Qiao Jin](#), [Chih-Hsuan Wei](#), [Shubo Tian](#), [Po-Ting Lai](#), [Qingqing Zhu](#), [Chi-Ping Day](#), [Christina Ross](#), [Robert Leaman](#) & [Zhiyong Lu](#) [✉](#)

[Nature Methods](#) (2025) | [Cite this article](#)

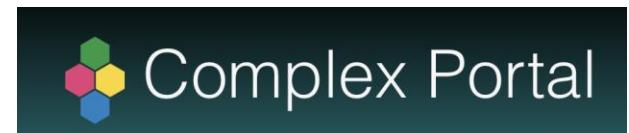
Gene-centric information in expert-curated biological databases



GENEONTOLOGY
Unifying Biology



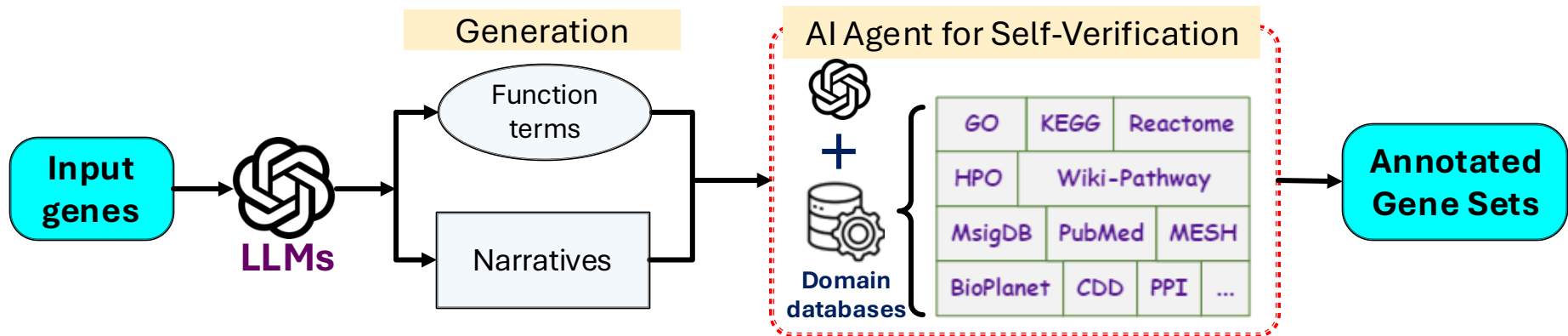
BioPlanet



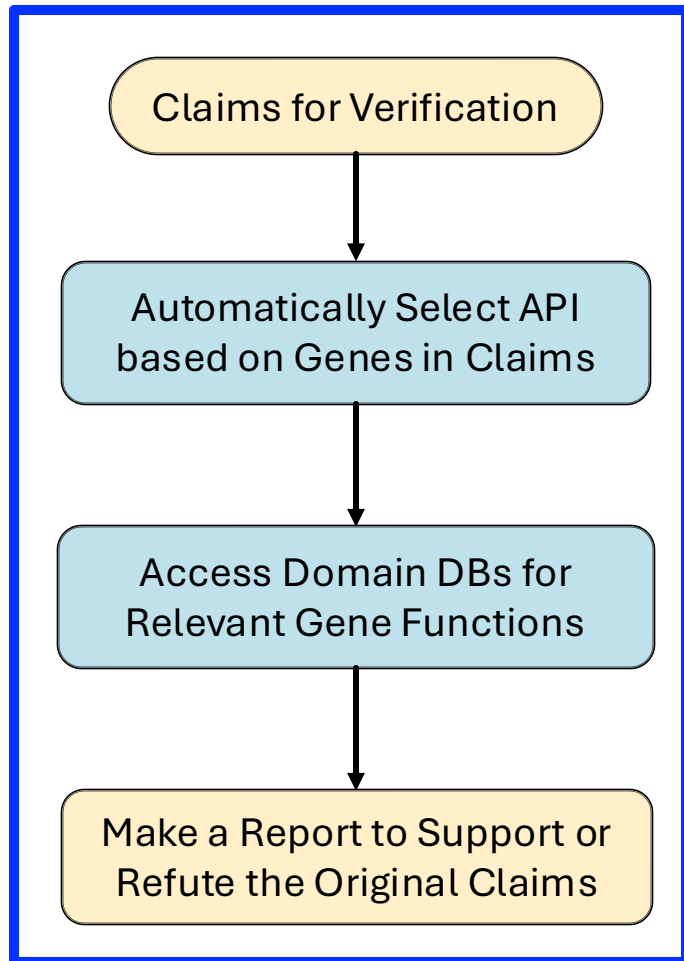
Our GeneAgent Approach



Zhizheng Wang, PhD

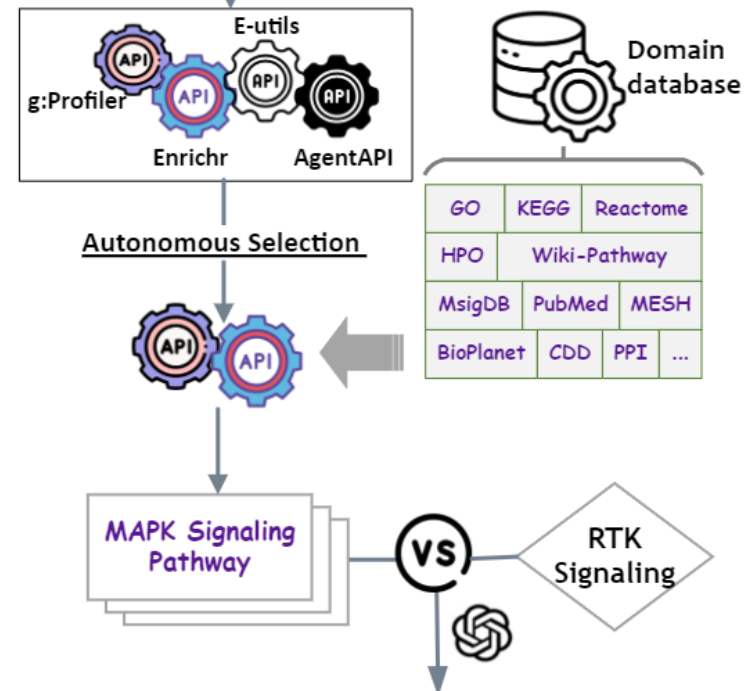


AI Agent for self-Verification



Claim:

ERBB2, ERBB4, FGFR2, FGFR4, HRAS, KRAS is involved in RTK Signaling

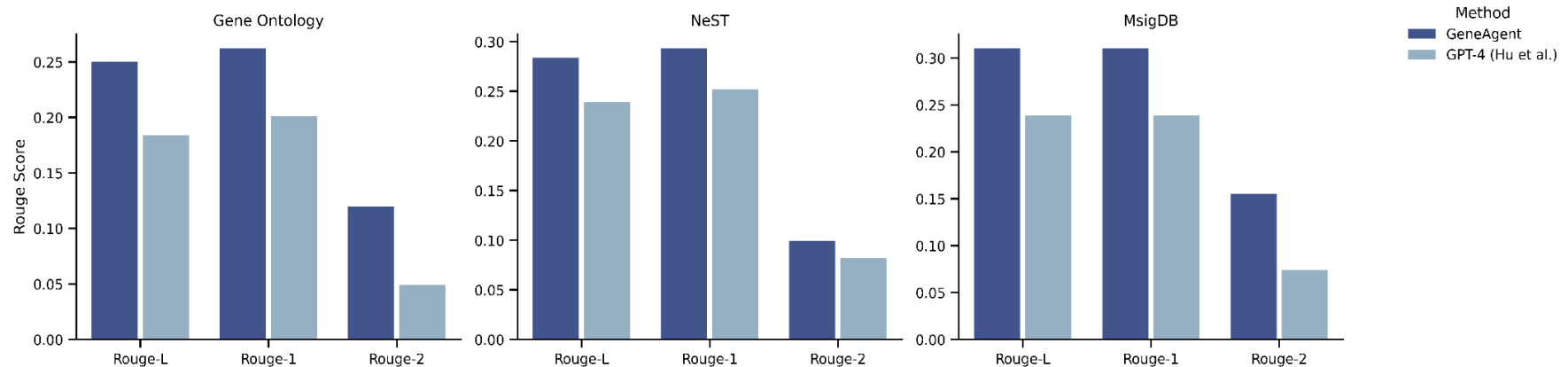


Verification Reports:

The claim is **not directly verified** by the selfVeri-Agent. The top enrichment function names of the given gene set include "**MAPK signaling pathway**", ..., while these functions are merely related to the name of "RTK Signaling". Therefore, based on the provided data, the claim **cannot be confirmed**.

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



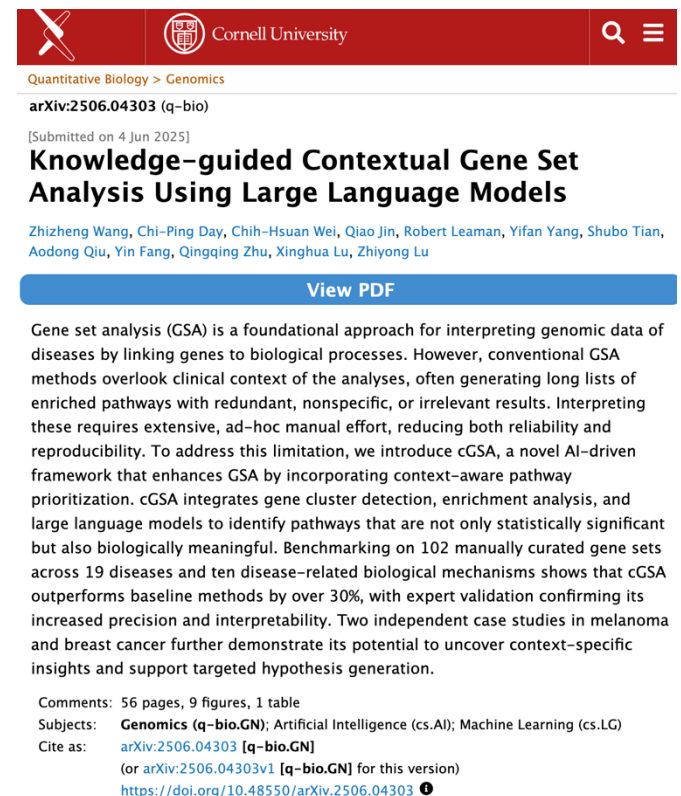
Christina Ross, Ph.D. NCI
Lab of Cancer Biology and Genetics

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)



The screenshot shows the top portion of an arXiv preprint page. At the top is a red navigation bar with the Cornell University logo and a search icon. Below this, the page is categorized under 'Quantitative Biology > Genomics'. The preprint ID 'arXiv:2506.04303 (q-bio)' is displayed, along with the submission date '[Submitted on 4 Jun 2025]'. The title 'Knowledge-guided Contextual Gene Set Analysis Using Large Language Models' is prominently featured. Below the title, the authors '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' are listed. A blue button labeled 'View PDF' is visible. The abstract text begins with '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.' At the bottom, it provides details: 'Comments: 56 pages, 9 figures, 1 table', 'Subjects: Genomics (q-bio.GN); Artificial Intelligence (cs.AI); Machine Learning (cs.LG)', and 'Cite as: arXiv:2506.04303 [q-bio.GN] (or arXiv:2506.04303v1 [q-bio.GN] for this version)'. A DOI link 'https://doi.org/10.48550/arXiv.2506.04303' is also present.

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

 National Library of Medicine

HOME GENEAGENT TUTORIAL MORE

Enter a gene set (HUMAN or MOUSE only) to analyze

Email (Optional):

[Analyze Gene Set](#) [Example Gene Sets](#) [Upload A Gene Set](#)

1. Please enter or upload a list of genes separated by any of the following delimiters: whitespace (" "), comma (,), semicolon (;), carriage return (\n), tab (\t), or "
".
2. Please use "csv" or "txt" formats if you upload the gene set in a file.
3. Complete analysis of a typical gene set will take approximately 3-5 minutes.
4. Please limit the length of a gene set within 400 genes for reasonable processing time.
5. Please refer to the [Tutorial](#) for more details.

Connect with NLM

National Library of Medicine
8500 Rockville Pike
Bethesda, MD 20894

Web Policies
FOIA
HHS Vulnerability Disclosure

NLM | NIH | HHS | USA.gov

Platform Solutions Resources Open Source Enterprise Pricing

ncbi-nlp / GeneAgent Public

Code Issues 1 Pull requests 1 Actions Projects Security Insights

main 1 Branch 1 Tag

AllenWangle	Update README.md	2107935 · last month	40 Commits
Datasets	Rename README.md to README.md	6 months ago	
GSEATerms	Add files via upload	6 months ago	
Outputs	Update and rename README.md to READ...	6 months ago	
Verification Reports/Cascade	Update and rename README.md to READ...	6 months ago	
apis	Create README.md	6 months ago	
BP_terms_All.csv	Add files via upload	6 months ago	
LICENSE	Add files via upload	6 months ago	
README.md	Update README.md	last month	
code structure.pdf	Add files via upload	6 months ago	
evaluate.ipynb	Add files via upload	6 months ago	
gene_narrative_synopsis_spindoc...	Add files via upload	6 months ago	
gene_ontological_synopsis_spindoc...	Add files via upload	6 months ago	
homepage.geneagent.jpg	Add files via upload	6 months ago	
main_Col.py	Add files via upload	6 months ago	
main_cascade.py	Add files via upload	6 months ago	
main_summary.py	Add files via upload	6 months ago	
topic.py	Add files via upload	6 months ago	
worker.py	Add files via upload	6 months ago	
workflow.geneagent.svg	Add files via upload	6 months ago	

zenodo

Published March 11, 2025 | Version v1.1.0

Software Open

GeneAgent: Self-verification Language Agent for Gene Set Analysis using Domain Databases

Wangle ·

Show affiliations

Recent work in gene set analysis has shown promising performance utilizing large language models (LLMs). Nonetheless, their results are subject to limitations common in LLMs, such as hallucinations. In response, we develop GeneAgent, a language agent for gene set analysis that self-verifies by autonomously interacting with biological databases, reducing hallucinations and enhancing accuracy. GeneAgent generates novel function names or aligns with notable enriched terms for input gene sets. Benchmarking on 1,106 gene sets from different sources, GeneAgent consistently outperforms vanilla GPT-4 by a significant margin. A detailed manual review confirms the effectiveness of the self-verification module in minimizing hallucinations and generating a more reliable customer review. We also apply GeneAgent to seven novel gene sets derived from mouse B2305 melanoma cells, showing that GeneAgent offers valuable insights into gene functions and pathways.

ent-v1.1.0.zip

9.2 MB

Intology

GO_terms.csv 346.1 kB

GO_toy.csv 373 Bytes

ReadME.MD 123 Bytes

MsigDB.csv 39.6 kB

MsigDB_toy.csv 105 Bytes

ReadME.MD 113 Bytes

Find out more on how we use cookies

Accept all cookies

Accept only essential cookies

10.2 KB

Python 63.6%

Jupyter Notebook 36.4%

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)