

Large Language Models for Biomedicine: from PubMed Search to Gene Set Analysis

Zhiyong Lu, PhD FACMI FIAHSI
Senior Investigator, NCBI, NLM, NIH
Professor of Computer Science (Adjunct), UIUC

ISMB 2024 Tutorial
July 9, 2024

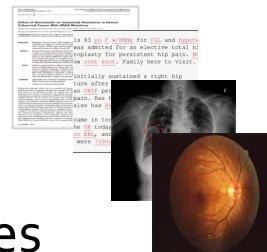
Our research at NLM IRP

Research Areas

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

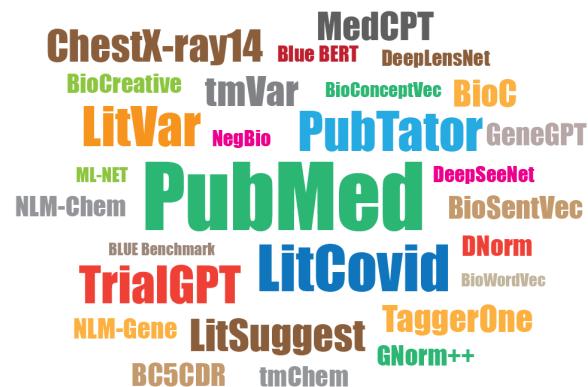
Text & Image data

Biomedical Literature
Clinical notes, EHRs
CT, CXR & retinal images



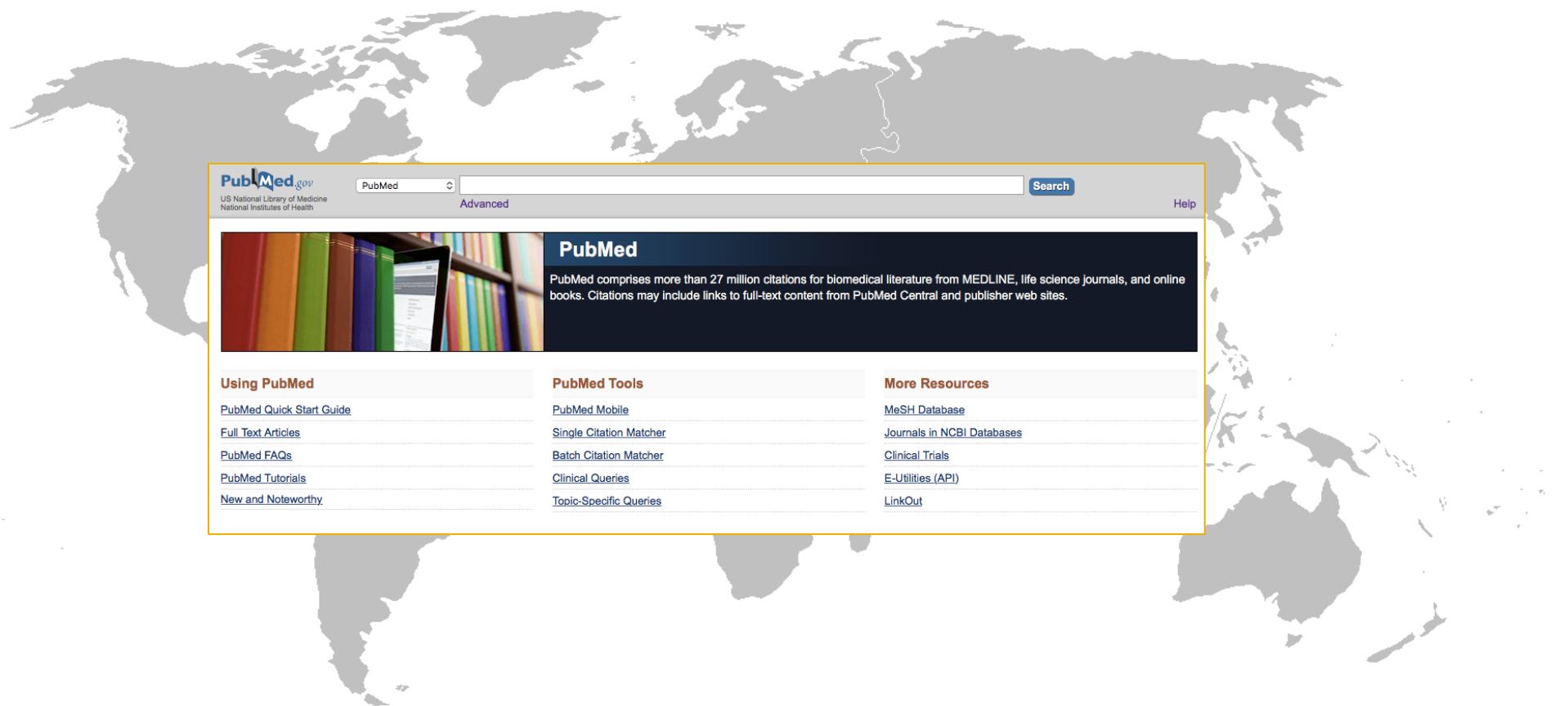
Application areas:

Literature Retrieval (e.g., PubMed Search; LitCovid)
Information Extraction/Curation (e.g., LitVar, PubTator)
AI/LLMs in Healthcare (e.g., diagnosis/prognosis; TrialGPT)



PubMed Daily Usage

- ~36+ million articles
- 2.5 million users daily
- ~ 3 million searches
- 9 million page views



AI/ML in PubMed

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



Perspective | Published: 01 October 2018

How user intelligence is improving PubMed

Nicolas Fiorini, Robert Leaman [...] Zhiyong Lu

Nature Biotechnology | Download Citation



COMMUNITY PAGE

Best Match: New relevance search for PubMed

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Khodolov, 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

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 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 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.



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): e2005943. <https://doi.org/10.1371/journal.pbio.2005943>

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.

2019 nCov infection

Unknown pneumonia

Novel coronavirus pneumonia

Coronavirus disease 2019

COVID-19

CoV 19 infection

Wuhan coronavirus pneumonia

Coronavirus 2 syndrome

2019 nCoV

SARS-CoV-2 associated ADRS

Novel coronavirus

SARS COVID 2

SARS-CoV-2

New Cov 19

HCoV-019

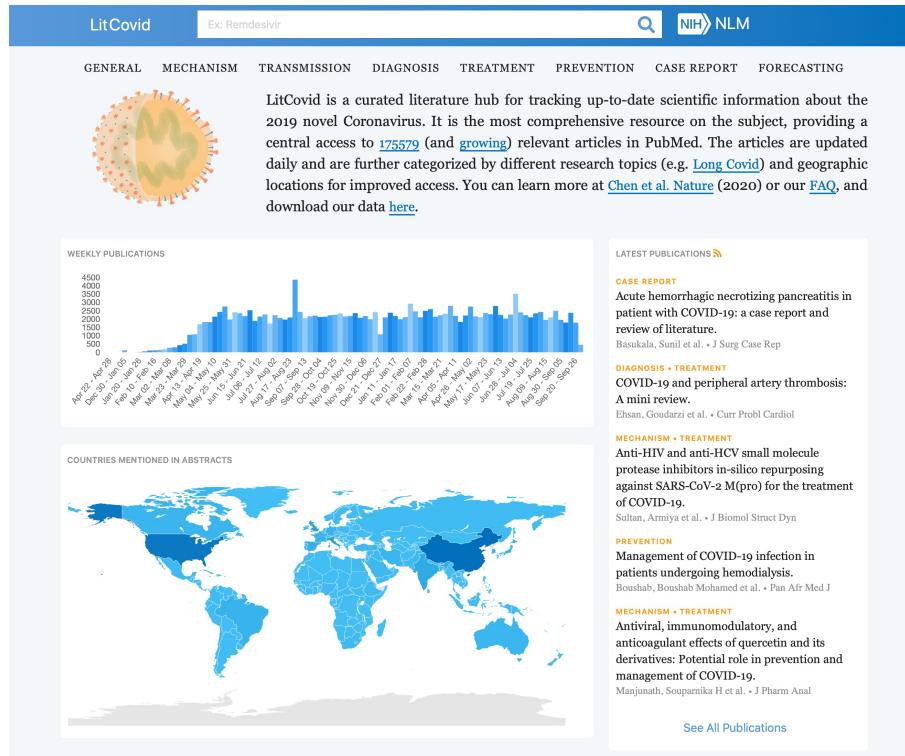
SC₂

WH-human 1' coronavirus

New CoV

Severe acute respiratory syndrome coronavirus 2

LitCovid: Research matters (est. 02/2020)



"I just wanted to say thank you for the LitCovid tool! It is an amazing resource."

"I just wanted to thank you for organising this informative information. I am incredibly grateful for your efforts."

Chen et al., Keep up with the latest coronavirus research.
Nature. 2020, *NAR* 2021, 2023



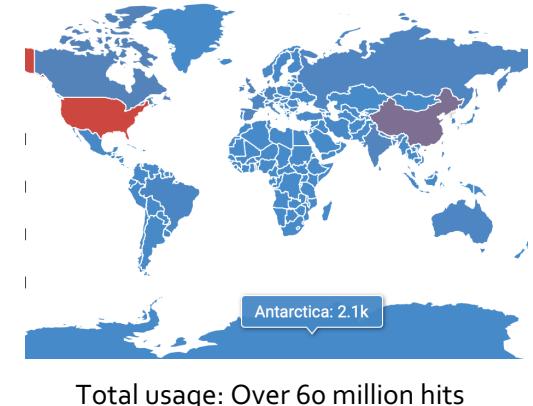
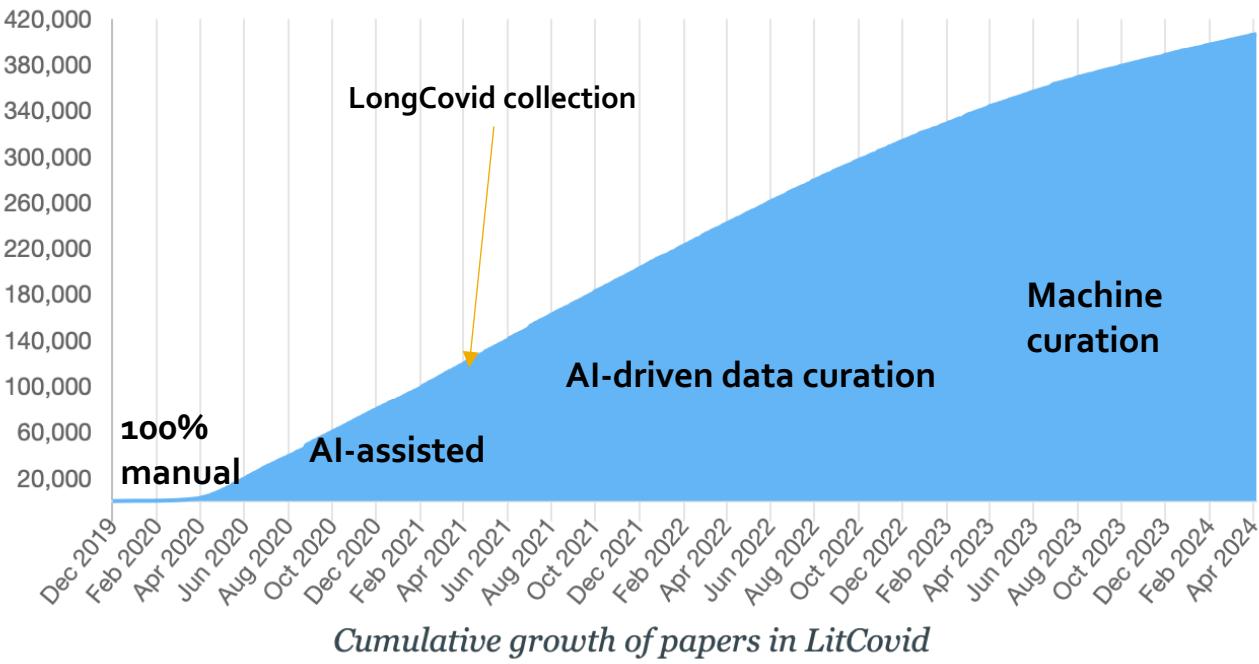
Qingyu Chen, postdoc
Assistant Professor, Yale

"First thing in the morning,
right after brushing your teeth,
you click on LitCovid"

"My favorite site for following the latest [#covid19](#) literature is NLM LitCovid. Organized, comprehensive and user friendly."

"Thank you for putting together such a wonderful resource."

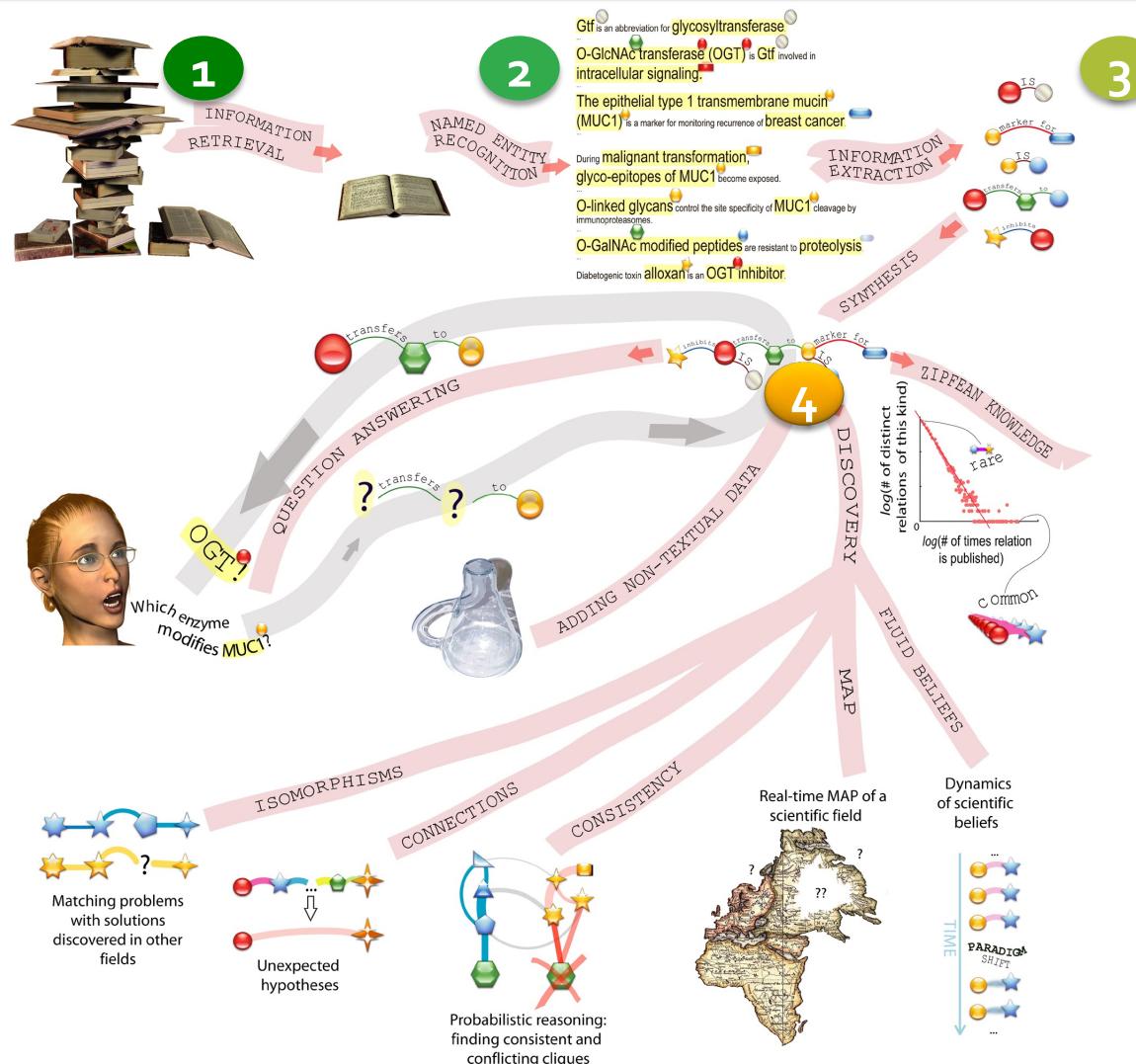
AI + Curator > Curator



Lilly Xu, 2022 summer intern
AMIA 2022 High School Scholar
Harvard, Class of 2027

Leaman et al. Comprehensively identifying Long Covid articles with **human-in-the-loop** machine learning." *Patterns*, 2023
Chen et al. LitMC-BERT: Transformer-Based Multi-Label Classification of COVID-19 Literature. *IEEE/ACM TCBB*, 2022

Information Extraction for Discovery



1. Articles

2. Entities

3. Relations

4. Integration

Extracting gene-disease-variant relations from free text

[J Alzheimers Dis.](#) 2012;32(2)

Disease

Gene

Variant

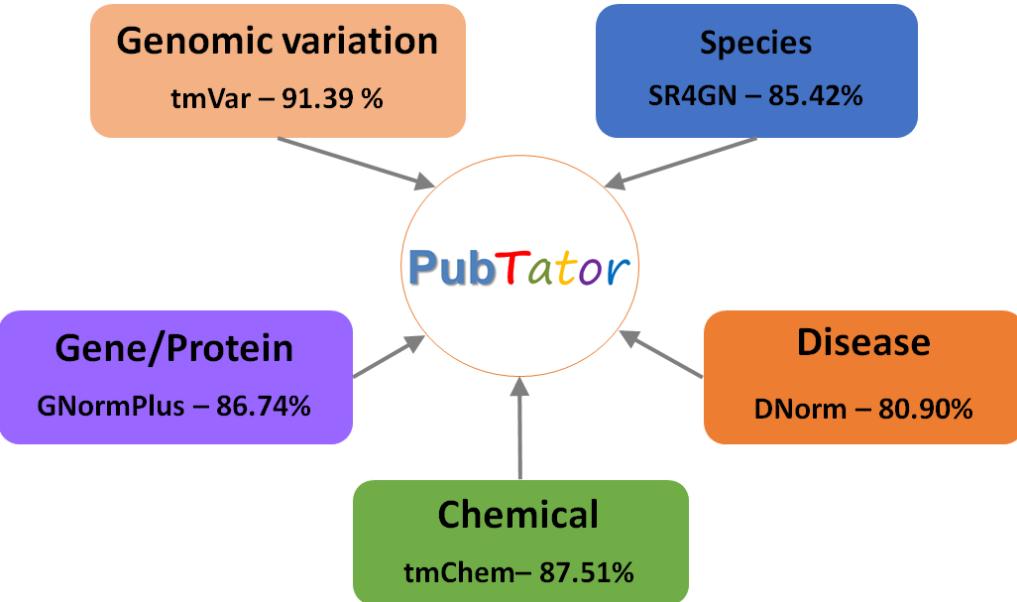
Highly pathogenic Alzheimer's disease presenilin 1 P117R mutation causes a specific increase in p53 and p21 protein levels and cell cycle dysregulation in human lymphocytes.

Bialopiotrowicz E¹, Szybinska A, Kuzniewska B, Buizza L, Uberti D, Kuznicki J, Wojda U.



Rank	Gene or Protein ID	Gene SYM	WTAA	MTAA	POS	Disease	PMIDs
1	Q13131	PRKAA1	Q	R	16	Breast cancer	16959974
2	P31749	AKT1	E	K	17	Breast cancer	17611497 18954143 19713527 21793738
3	P10275	AR	H	Y	874	Prostate cancer	17591767

PubTator (est. 2012): integrating text-mined results at PubMed scale



Go back Bioconcepts
 Curatable Species Mutation Chemical Gene
 Not Curatable Disease Gene
 TBD Chemical

PubTator

PMID:26022131 Selection of a novel DNA thioaptamer against HER2 structure.

Publication: Clinical _ translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico; 2015 May 29 [Full text links]

Gene Chemical Disease Species Mutation Clear Reset ⓘ

TITLE:
Selection of a novel DNA thioaptamer against HER2 structure.

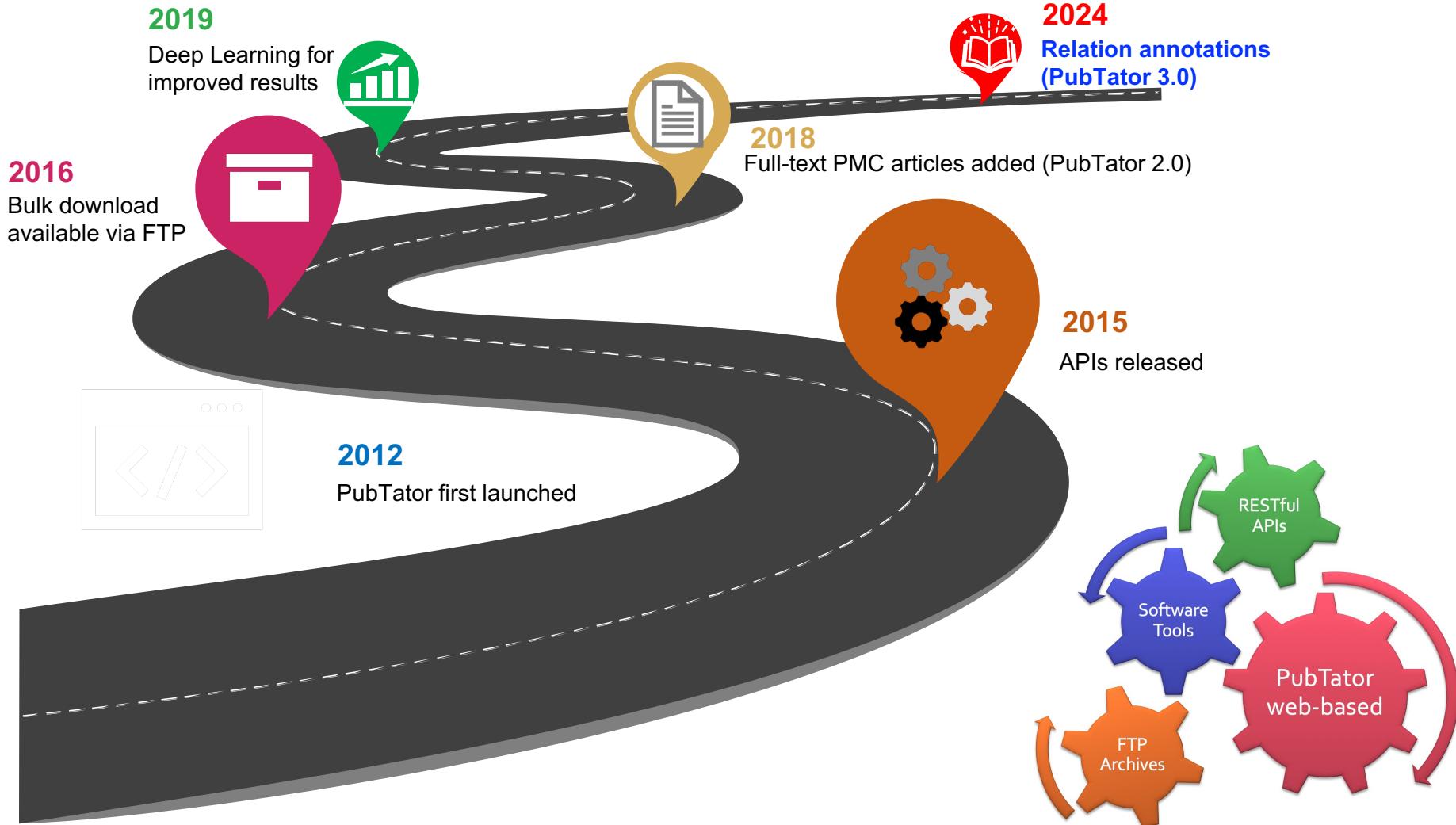
ABSTRACT:
PURPOSE: Human epithelial growth factor receptor 2 (HER2) is over-expressed in several malignancies and represents an important therapeutic target. Aptamers are oligonucleotides that may potentially serve as tumor-homing ligand with excellent affinity and specificity for target cancer therapy. However, aptamers need to have nucleic resistance in order to function in vivo. The aim of this study was to generate a novel HER2 thioaptamer with enhanced nuclease resistance. METHODS: The HER2 thioaptamer is selected in an evolutionary process called systematic evolution of ligands by exponential enrichment. RESULTS: The thioaptamer could bind to the extracellular domain of HER2 with a K d of 172 nM and had minimal cross reactivity to trypsin or IgG. Moreover, the thioaptamer was found capable of binding with the HER2-positive breast cancer cells SK-BR-3 and MDA-MB-453, but not the HER2-negative cells MDA-MB-231. Notably, the thioaptamer HY6 largely maintained its structural integrity facing the nucleases in serum, while regular DNA aptamers were mostly digested. Additionally, the thioaptamer retained the capability of binding with the HER2-positive cells in the presence of serum, whereas non-thiomated HER2 aptamer lost the binding function. CONCLUSION: The results indicated that the selected thioaptamer was more resistant to nuclease than regular DNA aptamers and might potentially function as a HER2-targeting ligand in complicated environment.

ⓘ Concept View ⓘ Mention View Add bio-relation annotation to the table below.

Entity type	Entity mention	Concept ID	Nomenclature	Delete
Disease	breast cancer	D001943	MEDIC	Delete
Disease	cancer	D009369	MEDIC	Delete
Gene	HIER2	2064	NCBI Gene	Delete
	Human epithelial growth factor receptor 2			

Save Annotation Results Save & Export Annotation Results ⓘ

PubTator milestones



PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge

Chih-Hsuan Wei  , Alexis Allot  , Po-Ting Lai , Robert Leaman , Shubo Tian , Ling Luo , Qiao Jin , Zhizheng Wang , Qingyu Chen , and Zhiyong Lu  *

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894, USA

* To whom correspondence should be addressed. Tel: +1 301 594 7089; Email: zhiyong.lu@nih.gov

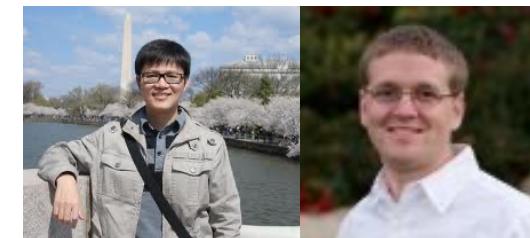
† The first two authors should be regarded as Joint First Authors.

Present addresses:
Alexis Allot, The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, Quebec H3A 2B4, Canada.
Ling Luo, School of Computer Science and Technology, Dalian University of Technology, 116024 Dalian, China.

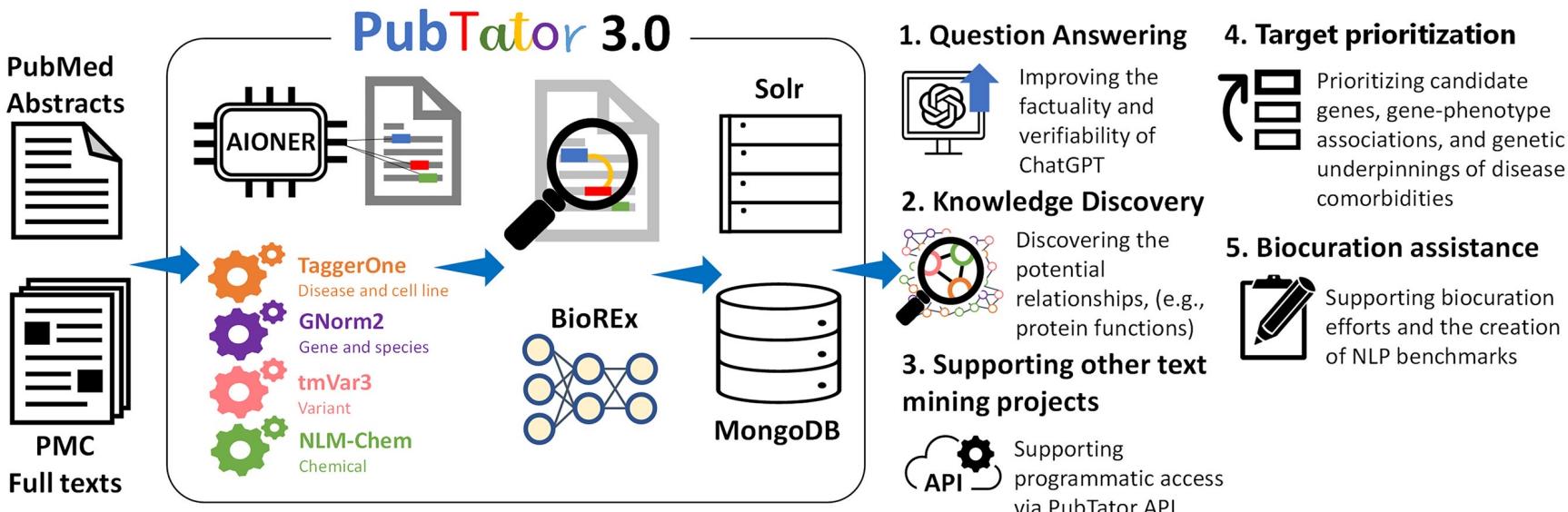
Qingyu Chen, Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT 06510, USA.

Abstract

PubTator 3.0 (<https://www.ncbi.nlm.nih.gov/research/pubtator3/>) is a biomedical literature resource using state-of-the-art AI techniques to offer semantic and relation searches for key concepts like proteins, genetic variants, diseases and chemicals. It currently provides over one billion entity and relation annotations across approximately 36 million PubMed abstracts and 6 million full-text articles from the PMC open access subset, updated weekly. PubTator 3.0's online interface and API utilize these precomputed entity relations and synonyms to provide advanced search capabilities and enable large-scale analyses, streamlining many complex information needs. We showcase the retrieval quality of PubTator 3.0 using a series of entity pair queries, demonstrating that PubTator 3.0 retrieves a greater number of articles than either PubMed or Google Scholar, with higher precision in the top 20 results. We further show that integrating ChatGPT (GPT-4) with PubTator APIs dramatically improves the factuality and verifiability of its responses. In summary, PubTator 3.0 offers a comprehensive set of features and tools that allow researchers to navigate the ever-expanding wealth of biomedical literature, expediting research and unlocking valuable insights for scientific discovery.



Staff Scientists: C. Wei, R. Leaman



API usage: over 1 billion requests



Tracking genetics variants in literature

LitVar²



FAQ

API

Search for variants in more than 35 million biomedical publications.

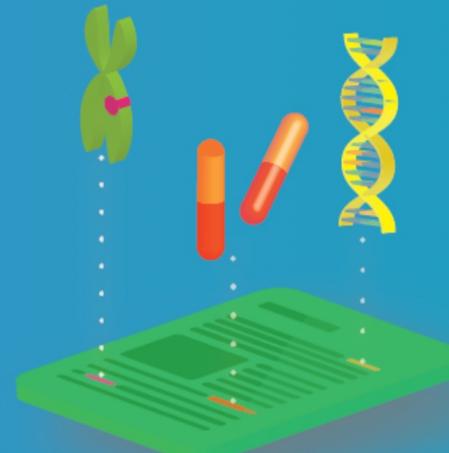
Variant (e.g. CFH R1210C)

Optional Text (e.g. AMD)

SEARCH

Try: KRAS A146T CFH R1210C rs121913527 CA123643

LitVar allows the search and retrieval of variant specific information from relevant studies in the literature, with related concept (e.g., diseases) annotations. By normalizing variant names, LitVar returns the same results regardless of which name of a variant (e.g. BRCA1 p.P871L or c.2612C>T) is used in the query. Read more here.



Lon Phan, dbSNP



Melissa Landrum, ClinVar



Heidi Rehm, Broad

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 • 312

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

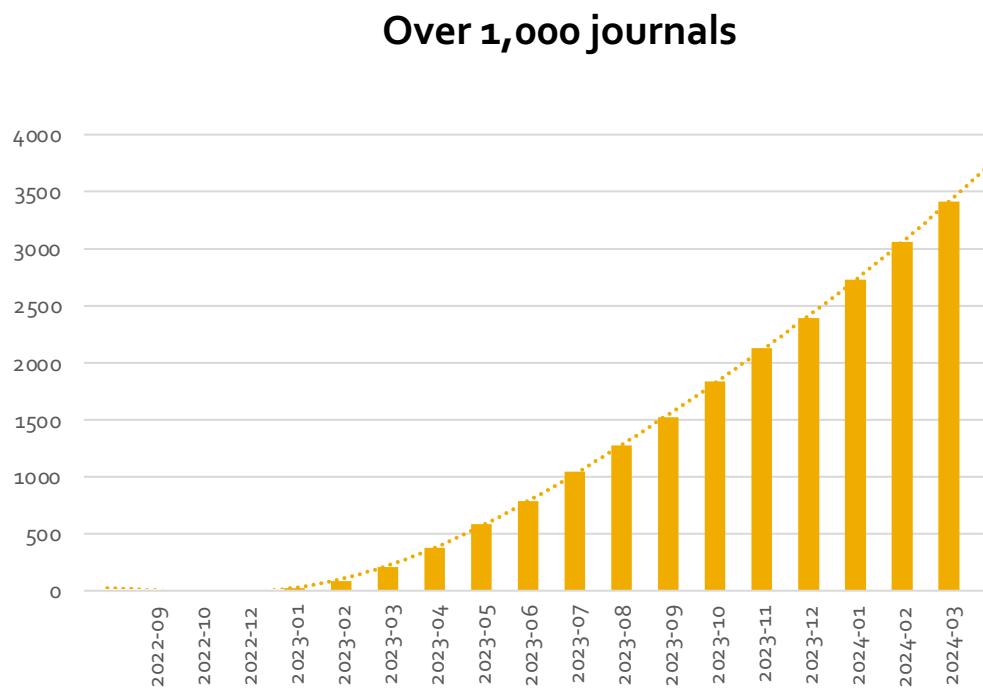


statista



LLMs for Biomedicine

PubMed: “large language models” OR chatgpt



~170 in ophthalmology

Large language models in medicine.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. *Nat Med.* 2023 Aug;29(8):1930–1940. doi: 10.1038/s41591-023-02448-8. Epub 2023 Jul 17. PMID: 37460753 Review.

Large language models (LLMs) can respond to free-text queries without being specifically trained in the task in question, causing excitement and concern about their use in healthcare settings. ...Here we outline how LLM applications such as **ChatGPT** are ...

Large language models and their impact in ophthalmology.

Betzler BK, Chen H, Cheng CY, Lee CS, Ning G, Song SJ, Lee AY, Kawasaki R, van Wijngaarden P, Grzybowski A, He M, Li D, Ran Ran A, Ting DSW, Teo K, Ruamviboonsuk P, Sivaprasad S, Chaudhary V, Tadayoni R, Wang X, Cheung CY, Zheng Y, Wang YX, Tham YC, Wong TY. *Lancet Digit Health.* 2023 Dec;5(12):e917–e924. doi: 10.1016/S2589-7500(23)00201-7. PMID: 38000875 Free article. Review.

The advent of generative artificial intelligence and **large language models** has ushered in transformative applications within medicine. ...This Viewpoint seeks to stimulate broader discourse on the potential of **large language models** in ...

Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard.

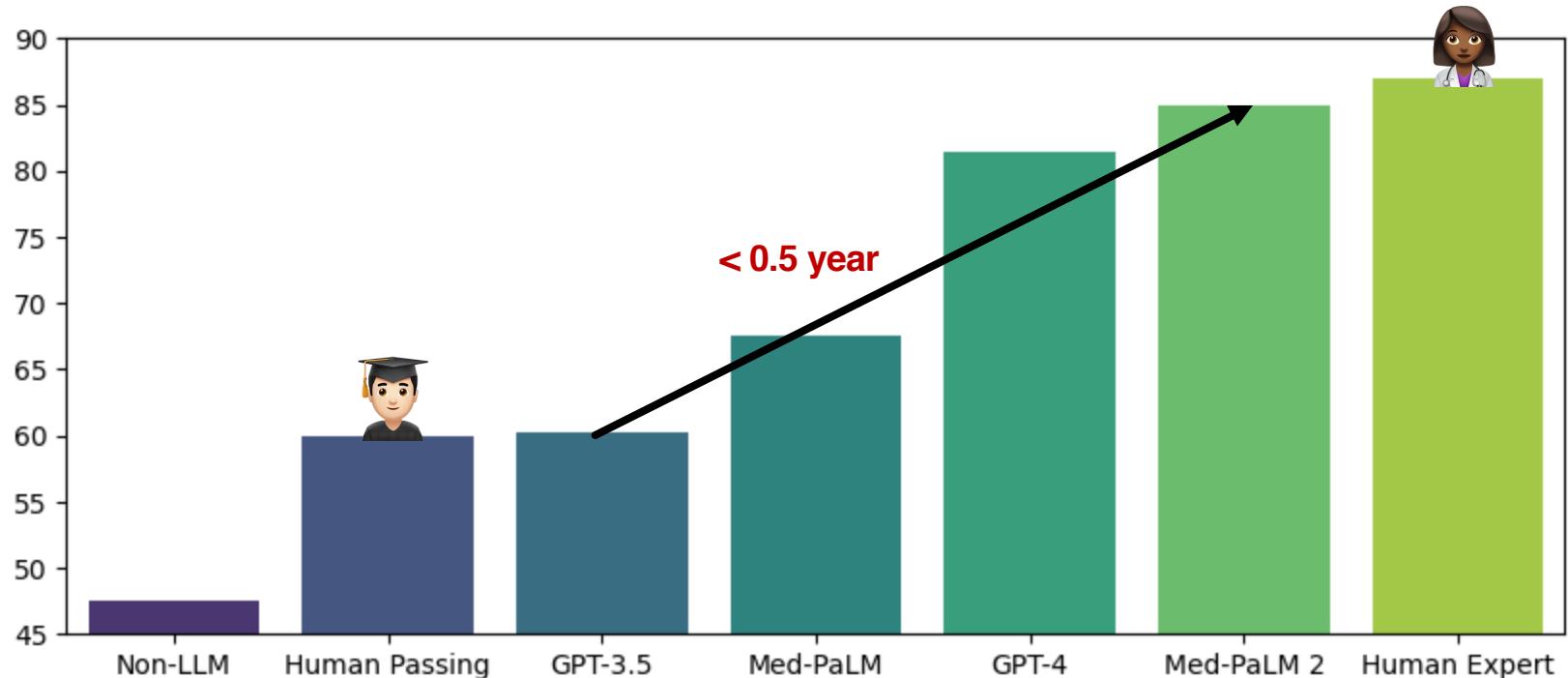
Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, Chen DZ, Goh JHL, Tan MCJ, Sheng B, Cheng CY, Koh VTC, Tham YC. *EBioMedicine.* 2023 Sep;95:104770. doi: 10.1016/j.ebiom.2023.104770. Epub 2023 Aug 23. PMID: 37625267 Free PMC article.

BACKGROUND: Large language models (LLMs) are garnering wide interest due to their human-like and contextually relevant responses. ...All three LLM-Chatbots showed high mean comprehensiveness scores (Google Bard: 4.35; ChatGPT-4.0: 4.23; ChatGPT- ...

Generative Artificial Intelligence Through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges.

Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, Kalpathy-Cramer J, Lum F, Kim JE, Baxter SL, Ting DSW. *Ophthalmol Sci.* 2023 Sep 9;3(4):100394. doi: 10.1016/j.xops.2023.100394. eCollection 2023

LLMs for Biomedical QA: USMLE Performance



A boy saw 17
doctors over 3 years
for chronic pain.
ChatGPT found the
diagnosis



<https://rb.gy/tueqe7>

Hallucination in LLMs

QI

Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source article title and PMID.

fabricated



Sure, here are some of the most common symptoms associated with long-term sequelae of COVID-19, along with their corresponding source article title and PMID:

1. Fatigue and Muscle Weakness: "Post-acute COVID-19 syndrome: a case series and multidisciplinary approach to treatment" (PMID: 33571309)
2. Shortness of Breath: "Symptomatic Outcomes in Young Adults With COVID-19" (PMID: 32936950)
3. Joint Pain: "Musculoskeletal symptoms in COVID-19 patients" (PMID: 32767232)
4. Chest Pain: "Chest CT Findings in Long-COVID Patients with Dyspnea and Chest Pain" (PMID: 33828099)
5. Cognitive Dysfunction: "Neuropsychiatric and Cognitive Sequelae of COVID-19" (PMID: 33402831)

(truncated response)

Retrieval Augmented Generation

Q1

Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source PMID.

Review > *Nat Immunol.* 2022 Feb;23(2):194-202. doi: 10.1038/s41590-021-01104-y. Epub 2022 Feb 1.

Pathological sequelae of long COVID

Saurabh Mehndru¹ ², Miriam Merad³ ⁴

Affiliations + expand

PMID: 35105985 PMCID: PMC9127978 DOI: 10.1038/s41590-021-01104-y Free PMC article

Abstract

The world continues to contend with successive waves fuelled by the emergence of viral variants. At the same time, debilitating sequelae are increasingly recognized in 'long COVID'. Clinical symptom defects in memory and concentration, and a variety of manifestations, and several organ systems can be involved. These mechanisms are poorly understood at present. This review summarizes the post-COVID syndromes and examines the underlying mechanisms. Finally, we propose diagnostic strategies to better understand these syndromes and continue to afflict millions of people worldwide.

© 2022, Springer Nature America, Inc.

Meta-Analysis > *JAMA.* 2022 Oct 25;328(16):1604-1615. doi: 10.1001/jama.2022.29392. PMID: 35105985 PMCID: PMC9127978 DOI: 10.1001/jama.2022.29392 Free PMC article

Estimated Global Proportions of Individuals with Persistent Fatigue, Cognitive, and Physical Symptom Clusters Following Symptom Onset in 2020 and 2021

Global Burden of Disease Long COVID Collaborators; Sarah Wulf¹, Joachim G Aerts³, Ziyad Al-Aly⁴ ⁵, Charlie Ashbaugh¹, Tali Bialik⁶, Polina Bobkova⁸, Goulee Bonello¹⁰, Svetlana Borzakova¹¹ ¹², Daniel Butruju⁹, Austin Carter¹, Helen Chu¹⁶, Cristina De Rosa¹³, Mohamed Mustafa Diab¹⁷ ¹⁸, Emri Ekboom¹⁹, Maha El Tantawy²⁰, Robert Frittich²², Aysulu Gamirova²³, Petr V Glybochko²⁴, Justyna Hach²⁵, Shaghayegh Haghjooy Javandard²⁶, Erin B Hamilton¹, Gabriel Majurka H Heijnenbroek-Kal²⁸ ²⁹, Raimund Helbok³⁰, Meen E Hwang³¹, Susanne M Hüttig³², Michael Hultström²² ³³, Wasila Jassar³⁴, Ing-Marie Larsson²², Miklós Lipcsay²², Chelsea Liu³⁷, Callan Andrei Malinovschi³⁸, Werhui Ma¹⁷ ³⁹, Lyudmila Mazankova⁴⁰, Dominik Menges⁶, Noushin Mohammadifard⁴¹, Daniel Munro⁴², Osundu Ogbojuji³⁹, Ismail M Osmanov¹¹ ⁴³, José L Peñalvo⁴⁴ ⁴⁵,

> *BMJ.* 2023 Jan 1;380:e072529. doi: 10.1136/bmj-2022-072529.

Long covid outcomes at one year after mild SARS-CoV-2 infection: nationwide cohort study

Barak Mizrahi¹, Tamir Sudry¹, Natalie Flaks-Manor¹, Yaniv Yerushalmi¹, Nir Pinchas Akiva¹, Anat Elka-Zohar², Shirley Shapiro Ben David³, Uri Lerner⁴, Maytal Bivas-Benita¹, Shira Greenfeld²

Affiliations + expand

PMID: 36631153 PMCID: PMC9832503 DOI: 10.1136/bmj-2022-072529 Free PMC article

> *Lancet.* 2022 Aug 6;400(10350):452-4. doi: 10.1016/S0140-6736(22)01214-4.

Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study

Anranka V Ballering¹, Sander K R van Zon², Tim C Olde Hartman³, Lifelines Corona Research Initiative

Affiliations + expand

PMID: 35934007 PMCID: PMC9352274 DOI: 10.1016/S0140-6736(22)01214-4 Free PMC article

> *Nat Med.* 2022 Nov;28(11):2406-2415. doi: 10.1038/s41591-022-02001-z. Epub 2022 Sep 22.

Long-term neurologic outcomes of COVID-19

Evan Xu¹, Yan Xie¹ ² ³, Ziyad Al-Aly⁴ ⁵ ⁶ ⁷ ⁸

Affiliations + expand

PMID: 36138154 PMCID: PMC9671811 DOI: 10.1038/s41591-022-02001-z Free PMC article

Abstract

The neurologic manifestations of acute COVID-19 are well characterized, but a comprehensive evaluation of postacute neurologic sequelae at 1 year has not been undertaken. Here we use the national healthcare databases of the US Department of Veterans Affairs to build a cohort of 154,068 individuals with COVID-19, 5,638,795 contemporary controls and 5,859,621 historical controls; we use inverse probability weighting to balance the cohorts, and estimate risks and burdens of incident neurologic disorders at 12 months following acute SARS-CoV-2 infection. Our results show that in the postacute phase of COVID-19, there was increased risk of an array of incident neurologic sequelae including ischemic and hemorrhagic stroke, cognition and memory disorders, peripheral nervous system disorders, episodic disorders (for example, migraine and seizures), extrapyramidal and movement disorders, mental health disorders, musculoskeletal disorders, sensory disorders, Guillain-Barré syndrome, and encephalitis or encephalopathy. We

Providing relevant articles to ChatGPT

Retrieval → Summarization → Verification

Q1

Please list the most common symptoms associated with the long-term sequelae of COVID-19. Each symptom should list the source PMID.

5 x [PMID, title, abstract]

{35105985, 36215063, 36631153, 35934007, 36138154}



But there are still problems

combination error	not a symptom
missing citation	incorrect citation



Based on the sources provided, some common symptoms associated with the long-term sequelae of COVID-19 include:

1. Fatigue and malaise (PMID: 35105985; PMID: 36215063)
2. Dyspnea (difficulty breathing) (PMID: 35105985; PMID: 36215063)
3. Defects in memory and concentration (PMID: 35105985)
4. Cognitive problems (PMID: 36215063; PMID: 35934007)
5. Neuropsychiatric syndromes (PMID: 35105985)
6. Persistent inflammation (PMID: 35105985)

(truncated response)

Our work on medical LLMs

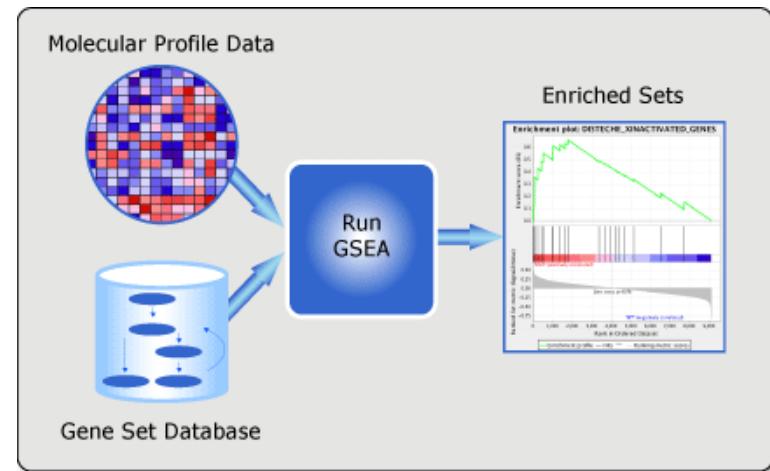
- Evaluating LLMs on medical benchmarks
 - Scientific communications (*J Am Soc Nephrol.*, 2023)
 - BioNLP tasks (*nature communications*, under revision)
 - Hidden flaws in multi-modal GPT4V (*npj Digital Medicine*, under revision)
- Enhancing standard LLMs
 - Retrieval augmented generation (MedRAG, *ACL*, under review)
 - Tool learning (GeneGPT, *bioinformatics*, 2024)
 - Incorporating domain knowledge (GeneAgent for gene set analysis)
- Novel applications/AI agents
 - TrialGPT: patient-trial matching (*nature communications*, under revision)
 - AgentMD: medical risk calculation
- AI Safety
 - Trustworthiness in clinical evidence synthesis (*J Biomed Inform*, 2024)
 - Bias & fairness (*communications medicine*, under revision)
 - Adversarial attacks & risks

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

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

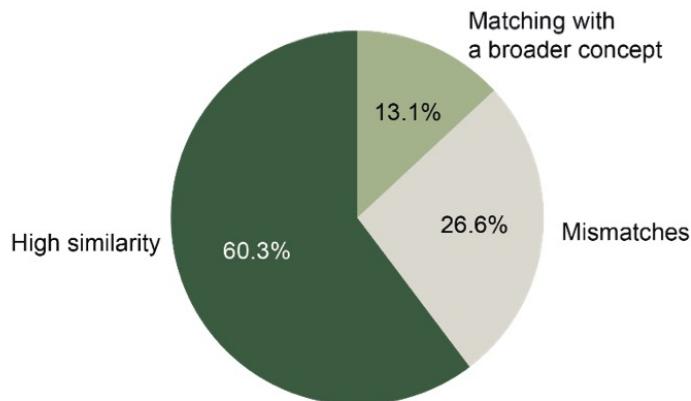
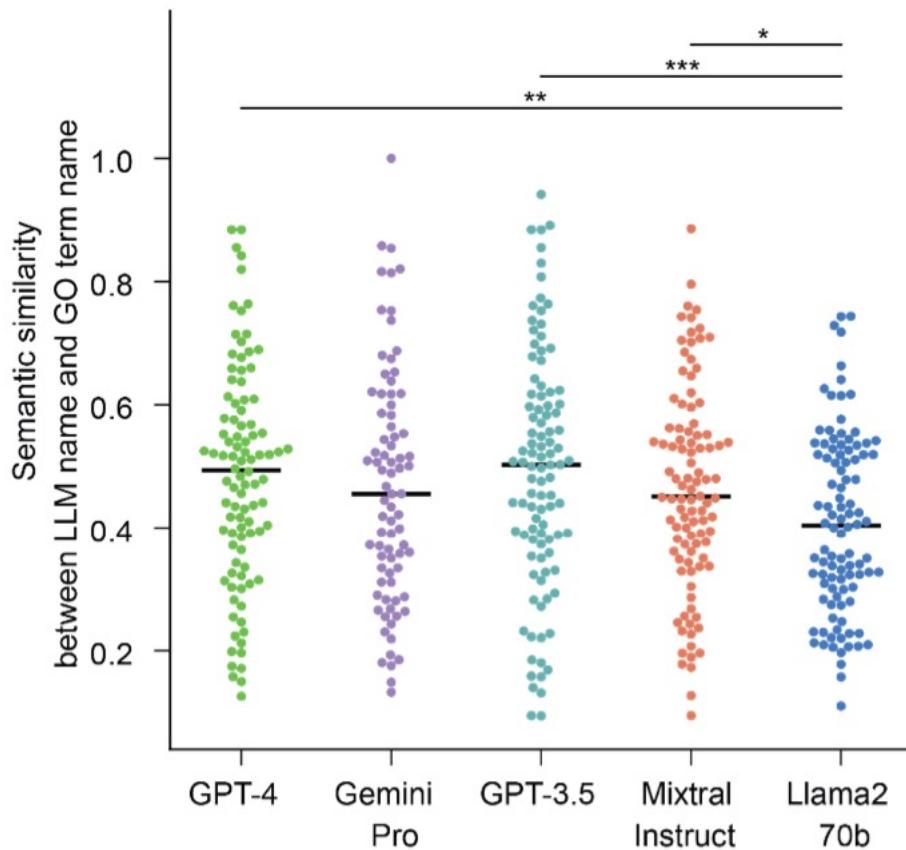
Controlled Terms for Ontology Reporting), a method to [Submitted on 7 Sep 2023 (v1), last revised 1 Apr 2024 (this version, v2)]

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 Bacheader, 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.

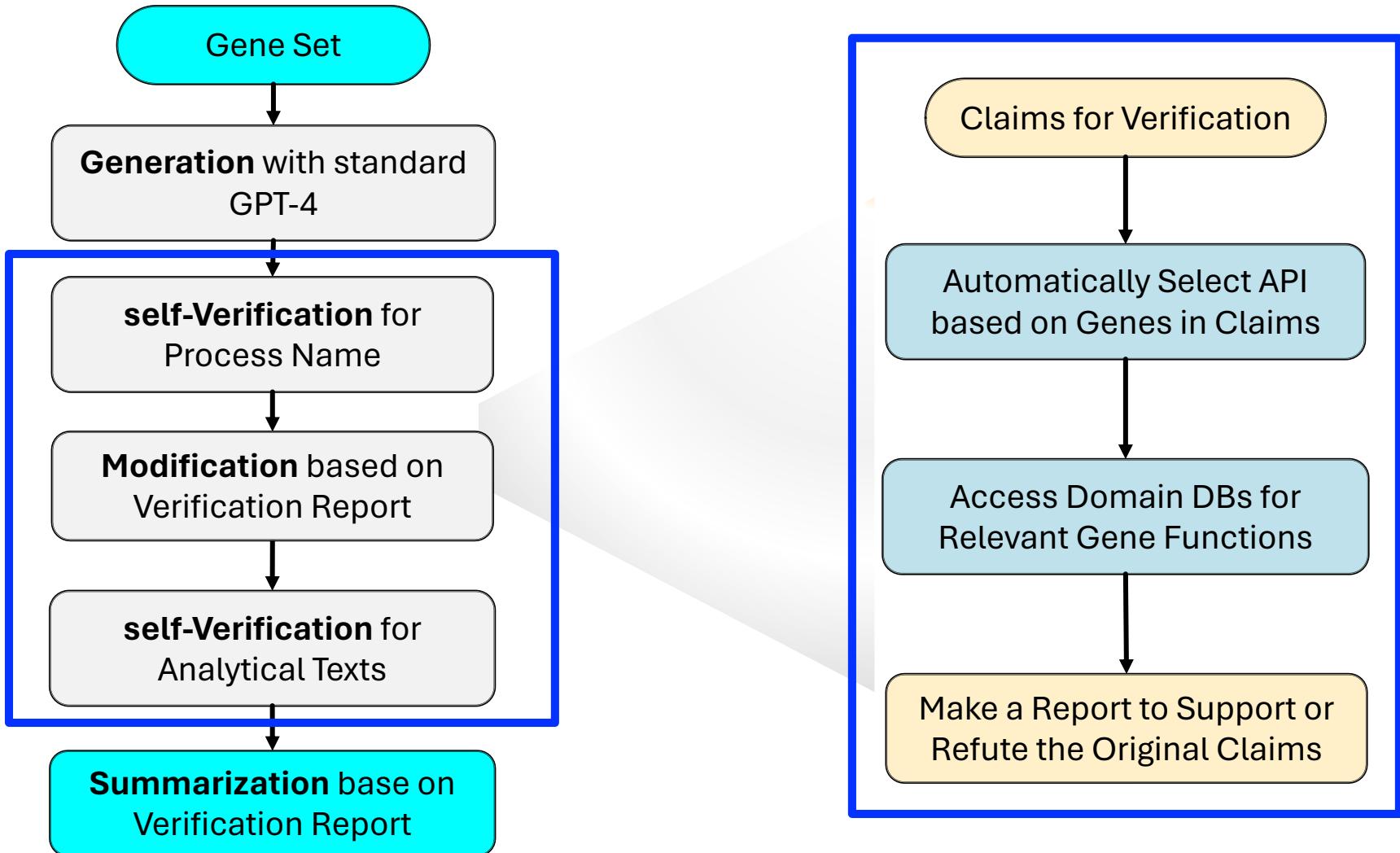
Key results in Hu et al.,



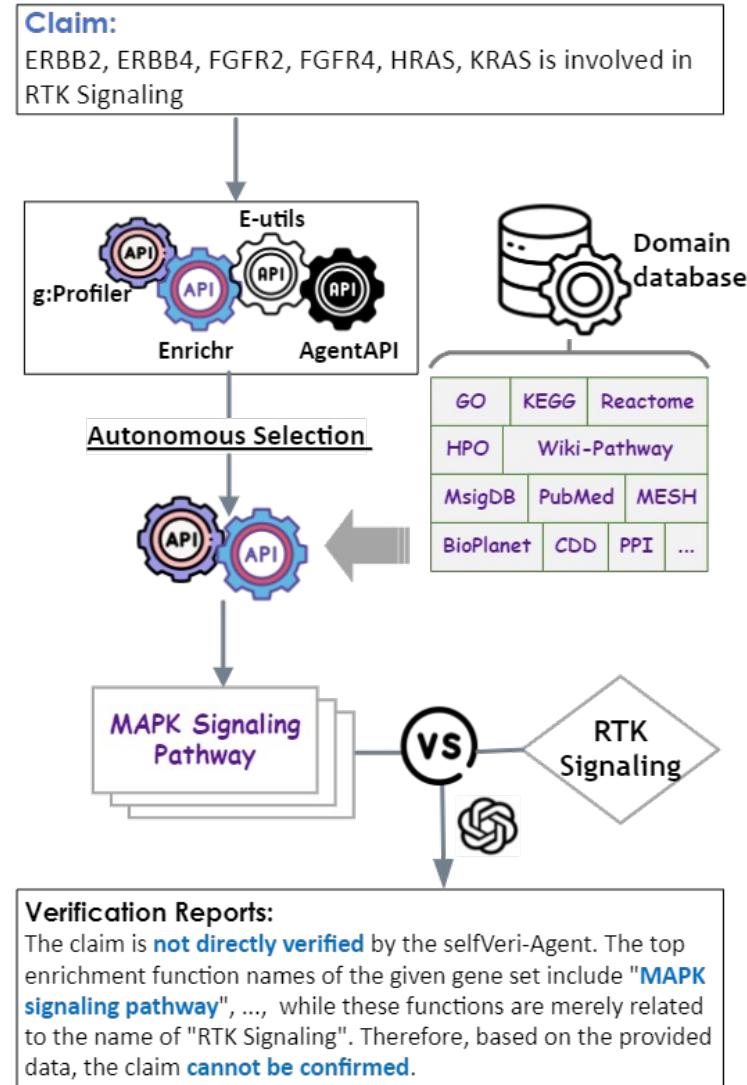
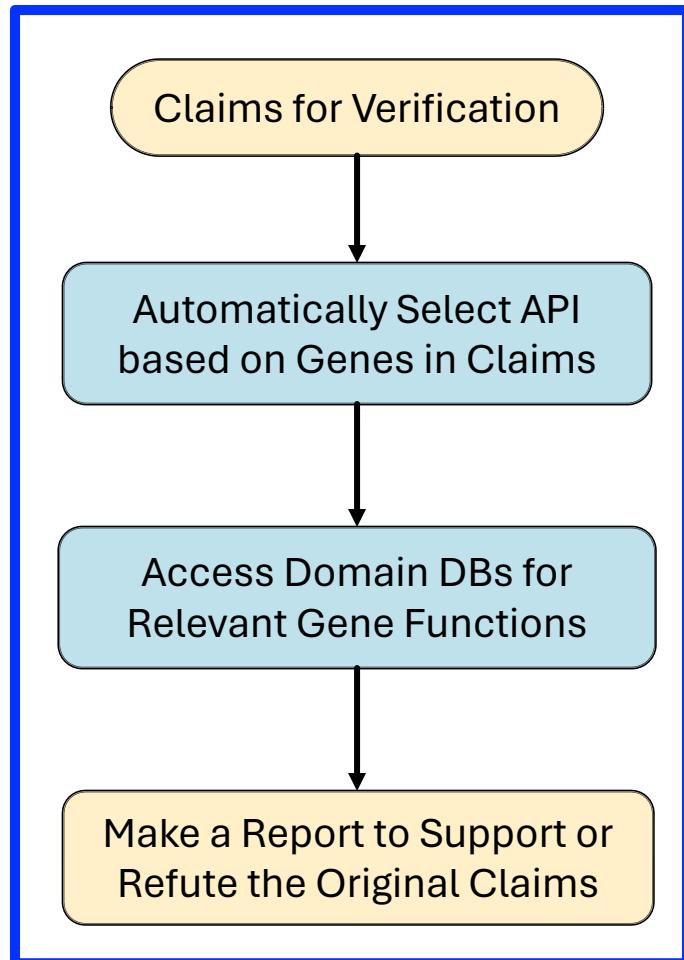
Gene-centric information in expert-curated biological databases



GeneAgent: overall workflow

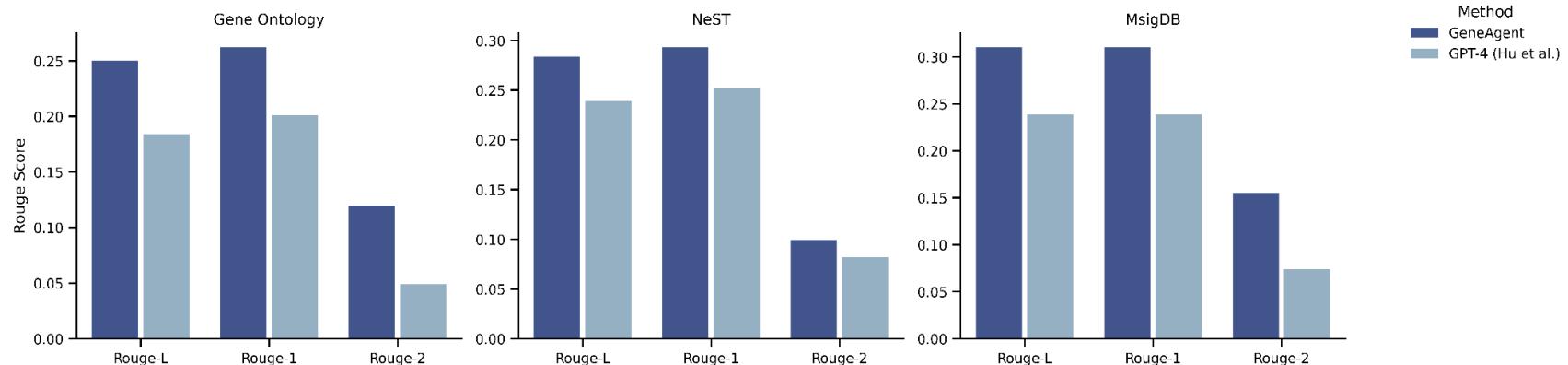


Example of self-Verification



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



Human review of LLM-generated verification reports

- Self-verification report (partially) supports or refutes original claims

Claim:

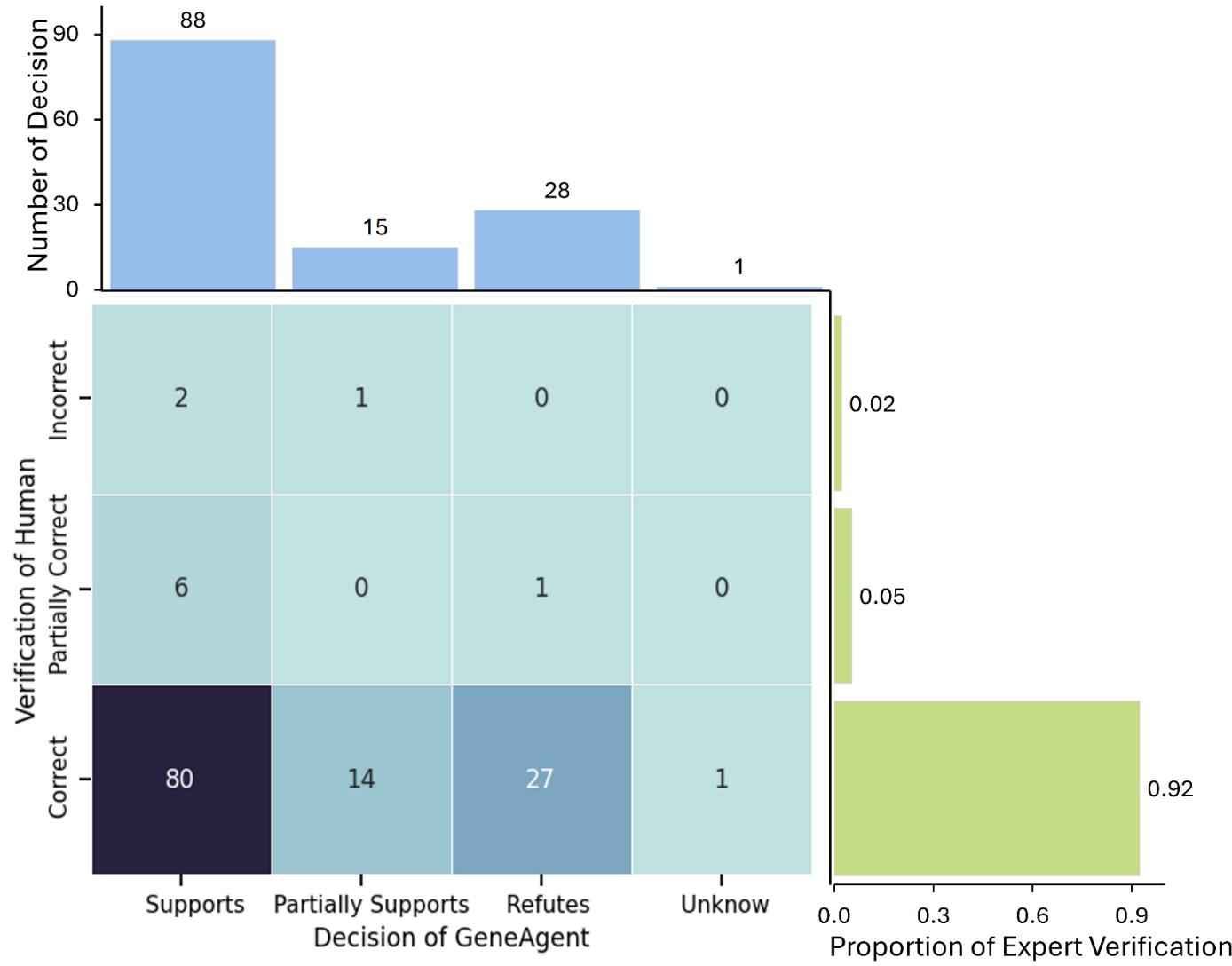
ERBB2, ERBB4, FGFR2, FGFR4, HRAS, KRAS is involved in RTK Signaling

Verification Reports:

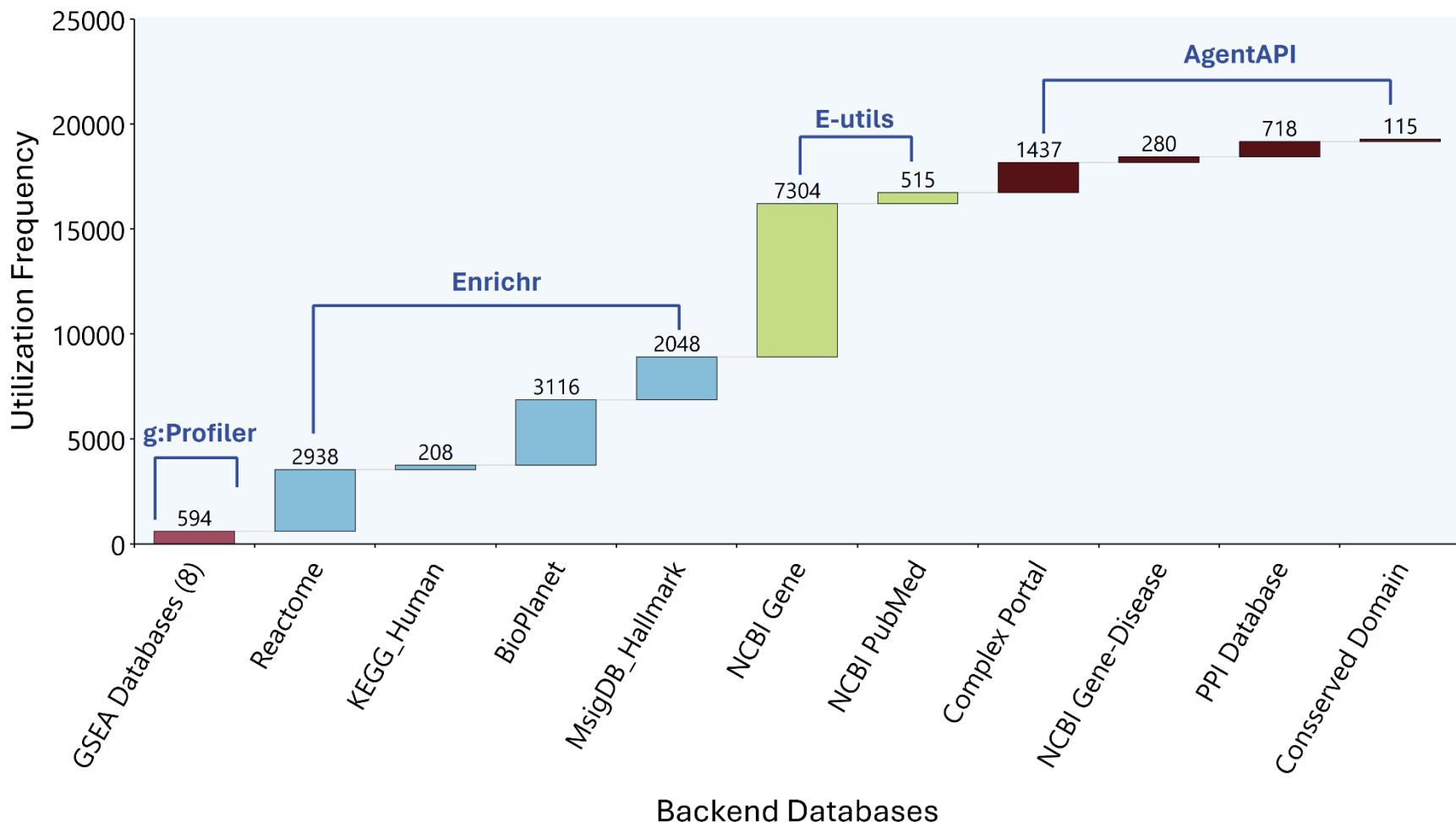
The claim is **not directly verified** by the selfVeri-Agent. The top enrichment function names of the given gene set include "**MAPK signaling pathway**", ..., while these functions are merely related to the name of "RTK Signaling". Therefore, based on the provided data, the claim **cannot be confirmed**.

- 132 manually reviewed
 - Double annotated
 - High inter-rater agreement

Self-verification is highly accurate



Utilization of different KBs



Pilot study with novel gene sets

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



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

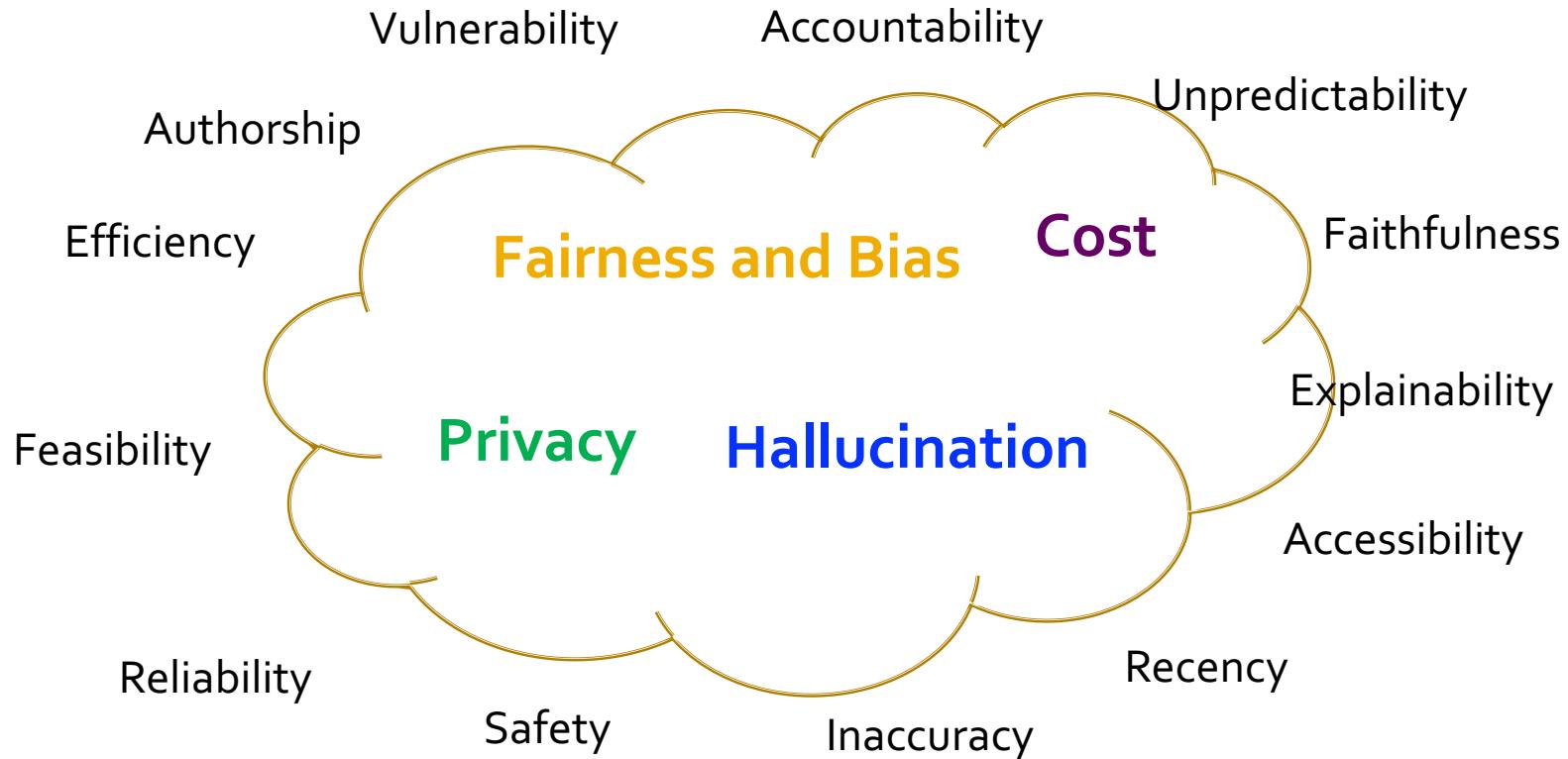


Christina Ross, Ph.D. NCI
Lab of Cancer Biology and Genetics

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 & Risks of LLMs in Medicine



Acknowledgments

Lu lab at NLM/NCBI: Don Comeau, Rezarta Dogan, Amr Elsawy, Won Kim, Qiao Jin, Robert Leaman, Po-Ting Lai, Ashley Shin, Shubo Tian, Zhizheng Wang, Chih-Hsuan Wei, John Wilbur, Natalie Xie, Yifan Yang, Lana Yeganova, Qingqing Zhu

NEI: Emily Chew, Michael Chiang, Tiarnan Keenan

CC: Ron Summers, Le Lu, Xiaosong Wang, Ke Yan, etc.

NCATS: Tyler Beck, Christine Colvis

NCI: Harris Floudas, James Gulley

PubMed: Grisha Starchenko, Jessica Chan, Kathi Canese, etc.

dbSNP: Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon

ClinVar: Melissa Landrum

BioCreative: Cecilia Arighi, Graciela Gonzalez-Hernandez, Martin Krallinger, Lynette Hirschman

UniProt: Alex Bateman, Alan Bridge, Livia Famiglietti, Michele Magrane, Sylvain Poux, Cathy Wu

GWAS Catalog: Jackie MacArthur, Fiona Cunningham, Helen Parkinson