

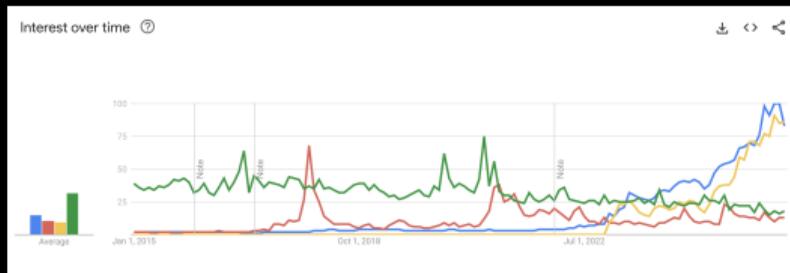
Transforming Precision Medicine with Smarter, Faster, and More Trustworthy AI

QCB Seminar

Hyun-Hwan Jeong, PhD
Baylor College of Medicine
February 4, 2026

The AI Revolution

Public Interest Shifts in the Age of AI: A New Era Entering Our Lives



AI Is Already Transforming Healthcare (and the World)

BUSINESS INSIDER

Subscribe | Newsletters

DOW JONES ▲ -1.65% NASDAQ ▲ -2.05% S&P 500 ▲ -1.66% AAPL ▲ +0.11% NVDA ▲ -0.5% MSFT ▲ +0.23% AMZN ▲ +0.34% META ▲ -0.2% TSLA ▲ -0.29%

NEWS

ChatGPT correctly diagnosed a 4-year-old's mysterious disease after 17 doctors failed

By Kate Hull



Svetlana Isakushova/Getty Images

AD



That's why we focus on immune science and advanced tech to drive better innovation

Beyond the Hype: The Limits and Environmental Impact of AI

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [news feature](#) > article

NEWS FEATURE | 21 January 2025

AI hallucinations can't be stopped – but these techniques can limit their damage

Developers have tricks to stop artificial intelligence from making things up, but large language models are still struggling to tell the truth, the whole truth and nothing but the truth.

By [Nicola Jones](#)

nature machine intelligence

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [nature machine intelligence](#) > comment > article

Comment | Published: 03 February 2025

Why the carbon footprint of generative large language models alone will not help us assess their sustainability

Leonie N. Bossert & Wulf Loh

Nature Machine Intelligence 7, 164–165 (2025) | [Cite this article](#)

1250 Accesses | 7 Citations | 2 Altmetric | [Metrics](#)

There is a growing awareness of the substantial environmental costs of large language models (LLMs), but discussing the sustainability of LLMs only in terms of CO₂ emissions is not enough. This Comment emphasizes the need to take into account the social and ecological costs and benefits of LLMs as well.

Beyond the Hype: Challenges of Using LLMs in Precision Medicine

Validity & Reliability

- LLMs can hallucinate, struggle with rare diseases, and may misinterpret subtle clinical signals.
- Ensuring consistent, safe, and evidence-based outputs is essential.



Data Privacy

- Patient data is protected, fragmented, and heterogeneous—making it difficult to obtain large, high-quality datasets for training or fine-tuning securely.

Ultra-High-Throughput Processing Requirements

- Precision Medicine involves genomic, clinical, imaging, and longitudinal data at massive scale, requiring LLMs to handle multimodal, high-volume inputs efficiently.

**How Do We Make AI Smarter,
Better, Faster... Greener?**

What It ⟨thinks⟩, It Becomes: o1 and DeepSeek-R1 in the Aha Era

OpenAI o1

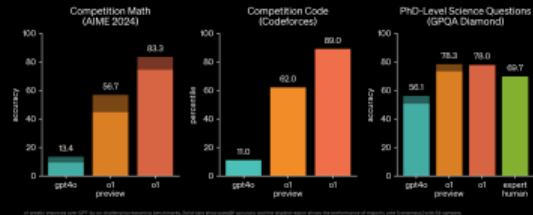


DeepSeek-R1

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both \cdots
 $(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2$.
Rearrange to isolate the inner square root term:
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
 \cdots
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be \cdots
We started with the equation:
 $\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
 $a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
Next, I could square both sides again, trusting the equation: \cdots
 \cdots

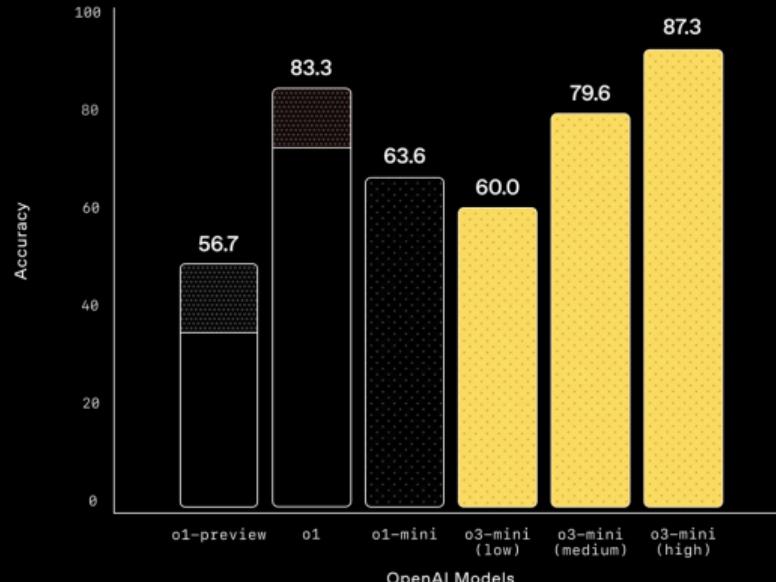
Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.



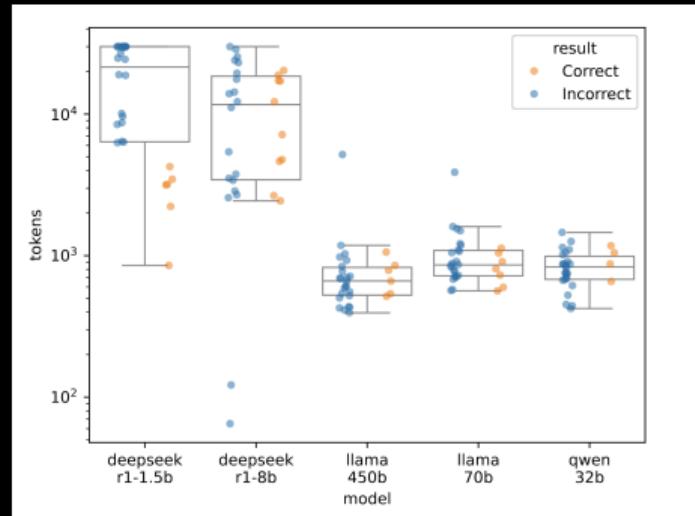
<https://arxiv.org/pdf/2501.12948>

The Hidden Cost of Accuracy: More Runs, More Thinking, More Compute

Competition Math (AIME 2024)



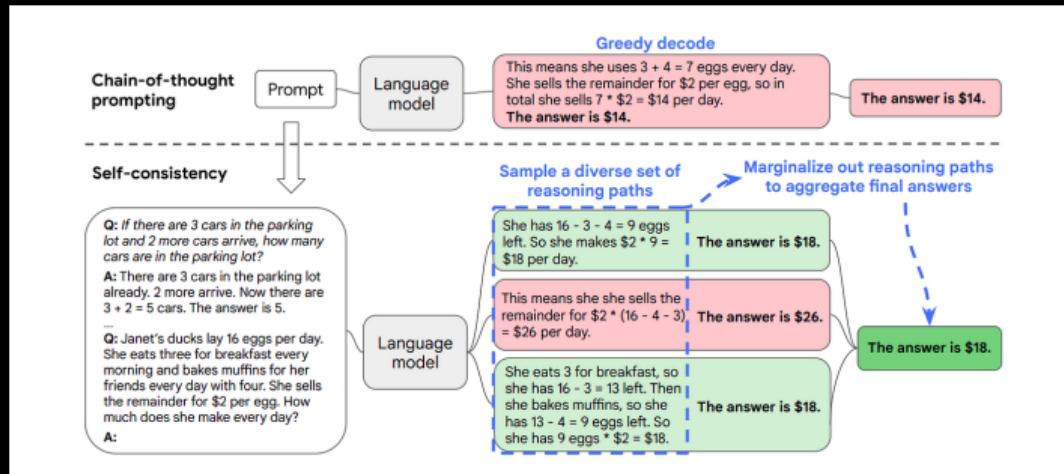
Mathematics: With low reasoning effort, OpenAI o3-mini achieves comparable performance with OpenAI o1-mini, while with medium effort, o3-mini achieves comparable performance with o1. Meanwhile, with high reasoning effort, o3-mini outperforms both OpenAI o1-mini and OpenAI o1, where the gray shaded regions show the performance of majority vote (consensus) with 64 samples.



Thinking LLMs Consume Massive Compute—More tokens (words) required

(Lee et al., ArXiv 2025)

Why? The Same Prompt—Different Reasoning Paths



(Wang et al., ArXiv 2022)

ConSol: Consistency Solver

CONSOL: SEQUENTIAL PROBABILITY RATIO TESTING TO FIND CONSISTENT LLM REASONING PATHS EFFICIENTLY

Jaeyeon Lee, Guantong Qi, Matthew Brady Neeley, Zhandong Liu*, Hyun-Hwan Jeong*

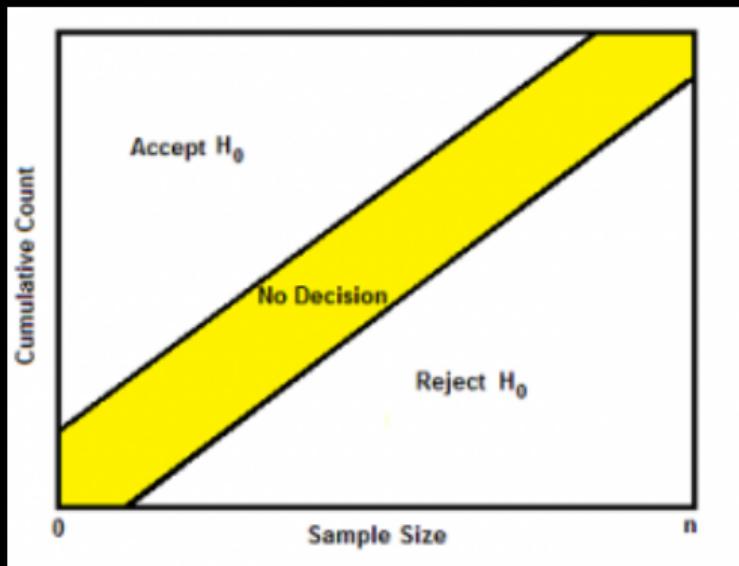
Baylor College of Medicine

Houston, Texas, USA

{jaeyeon.lee, guantong.qi, matthew.neeley, zhandonl, hyun-hwan.jeong}@bcm.edu

<https://arxiv.org/abs/2503.17587>

ConSol (Consistency Solver)—An Overview



Answer	Probability
1	0.0
2	0.1
3	0.7
4	0.0
5	0.2

Based on Wald's Sequential Probability Ratio Test (SPRT).

A. Wald. Sequential Tests of Statistical Hypotheses.

Ann. Math. Statist. 16(2), 117–186. June 1945.

(Courtesy of Jaeyeon Lee)

ConSol: Comparable Accuracy to Best-Known Methods

Table 1: A Comparison of Accuracy (Top) and Token Reduction (Bottom)

Method Benchmark / LLM	Self-Cons@40	p-value ^[2]	Ada-Cons	SPRT (Ours)	mSPRT (Ours)
AIME24 / o3-mini-low	70.0%	70.0%	70.0%	80.0%	76.7%
AIME24 / o3-mini-medium	90.0%	93.3%	90.0%	90.0%	90.0%
AIME24 / o3-mini-high	96.7%	96.7%	96.7%	96.7%	96.7%
GPQA Diamond / o3-mini-low	68.2%	69.2%	67.2%	66.7%	70.2%
GSM / gpt-4o-mini	94.2%	94.2%	94.2%	93.5%	93.9%
<hr/>					
AIME24 / o3-mini-low	0.0%	49.3%	47.9%	63.9%	57.3%
AIME24 / o3-mini-medium	0.0%	60.2%	62.8%	84.8%	76.8%
AIME24 / o3-mini-high	0.0%	54.0%	64.6%	84.5%	82.2%
GPQA Diamond / o3-mini-low	0.0%	63.2%	70.1%	86.2%	83.6%
GSM / gpt-4o-mini	0.0%	79.2%	82.8%	88.6%	88.7%

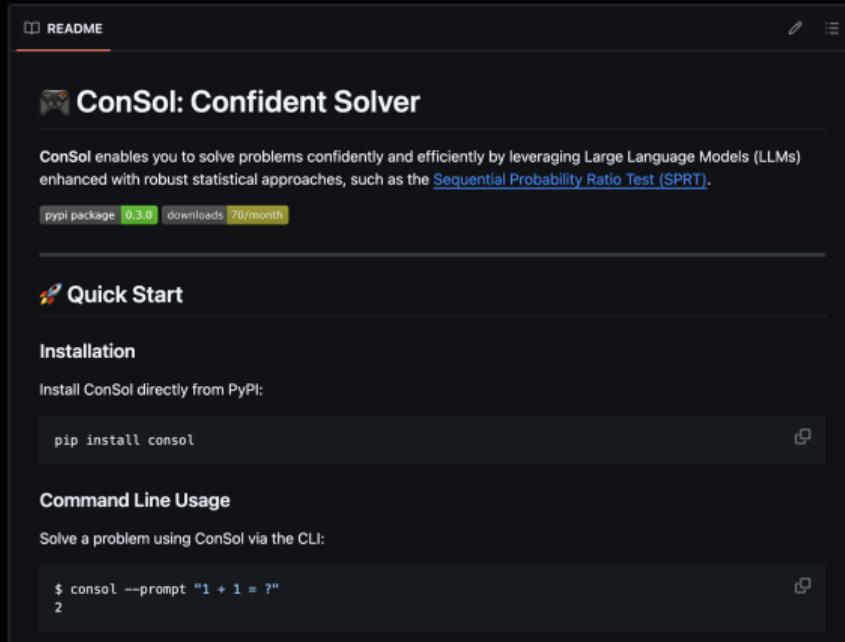
ConSol: Far Fewer Tokens ($\sim 80\%$ Token Reduction)

Table 1: A Comparison of Accuracy (Top) and Token Reduction (Bottom)

Method Benchmark / LLM	Self-Cons@40	p-value ^[2]	Ada-Cons	SPRT (Ours)	mSPRT (Ours)
AIME24 / o3-mini-low	70.0%	70.0%	70.0%	80.0%	76.7%
AIME24 / o3-mini-medium	90.0%	93.3%	90.0%	90.0%	90.0%
AIME24 / o3-mini-high	96.7%	96.7%	96.7%	96.7%	96.7%
GPQA Diamond / o3-mini-low	68.2%	69.2%	67.2%	66.7%	70.2%
GSM / gpt-4o-mini	94.2%	94.2%	94.2%	93.5%	93.9%
<hr/>					
AIME24 / o3-mini-low	0.0%	49.3%	47.9%	63.9%	57.3%
AIME24 / o3-mini-medium	0.0%	60.2%	62.8%	84.8%	76.8%
AIME24 / o3-mini-high	0.0%	54.0%	64.6%	84.5%	82.2%
GPQA Diamond / o3-mini-low	0.0%	63.2%	70.1%	86.2%	83.6%
GSM / gpt-4o-mini	0.0%	79.2%	82.8%	88.6%	88.7%

$$\text{Token Reduction} = \frac{T_{\text{Self-Cons@40}} - T}{T_{\text{Self-Cons@40}}} \times 100,$$

ConSol: Open Source

A screenshot of the GitHub README page for the ConSol project. The page features a dark-themed header with the title "ConSol: Confident Solver". Below the title, a brief description states: "ConSol enables you to solve problems confidently and efficiently by leveraging Large Language Models (LLMs) enhanced with robust statistical approaches, such as the [Sequential Probability Ratio Test \(SPRT\)](#)". A green button indicates the latest version is "0.3.0" with "70/month" downloads. The "Quick Start" section includes instructions for installation via PyPI ("pip install consol") and command-line usage ("\$ consol --prompt "1 + 1 = ?"").

ConSol: Confident Solver

ConSol enables you to solve problems confidently and efficiently by leveraging Large Language Models (LLMs) enhanced with robust statistical approaches, such as the [Sequential Probability Ratio Test \(SPRT\)](#).

pip package 0.3.0 downloads 70/month

Quick Start

Installation

Install ConSol directly from PyPI:

```
pip install consol
```

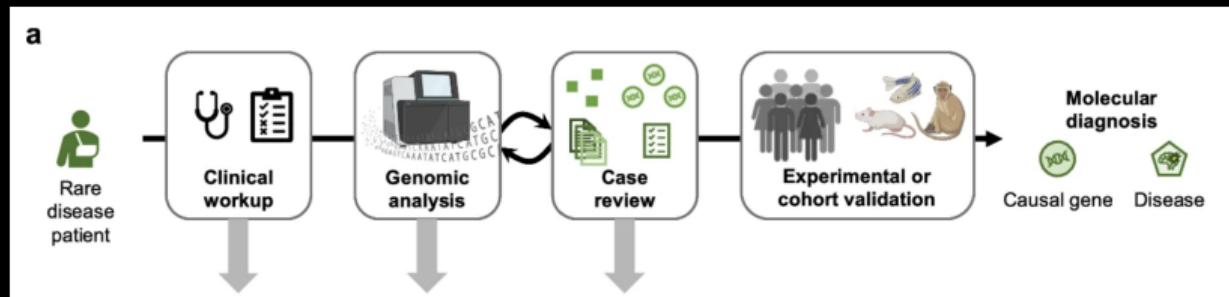
Command Line Usage

Solve a problem using ConSol via the CLI:

```
$ consol --prompt "1 + 1 = ?"
2
```

<https://github.com/LiuzLab/consol>

Rare Disease Diagnosis: A Needle-in-a-Haystack Problem



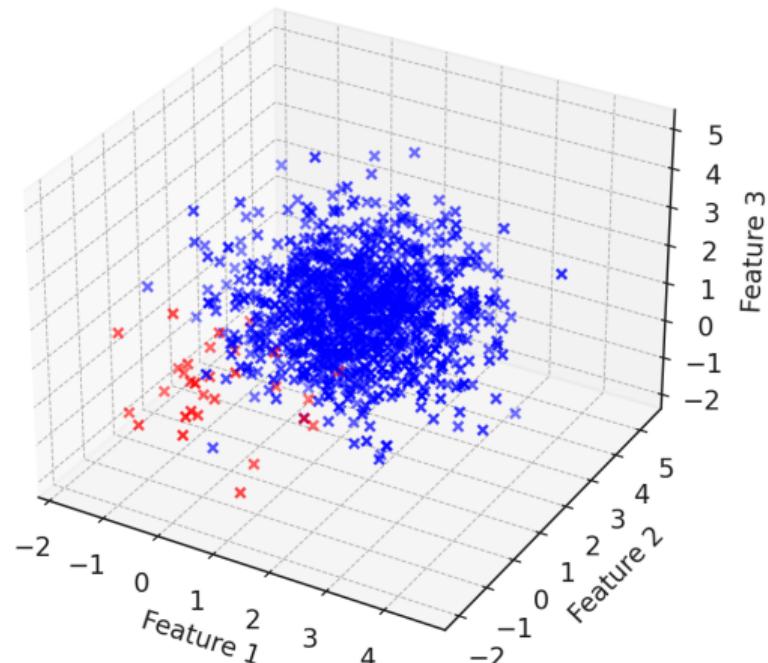
(Alsentzer et al., *npj Digital Medicine*, 2025)

AI-MARRVEL: A Rare-Disease Diagnosis Tool Taking a Wholistic View

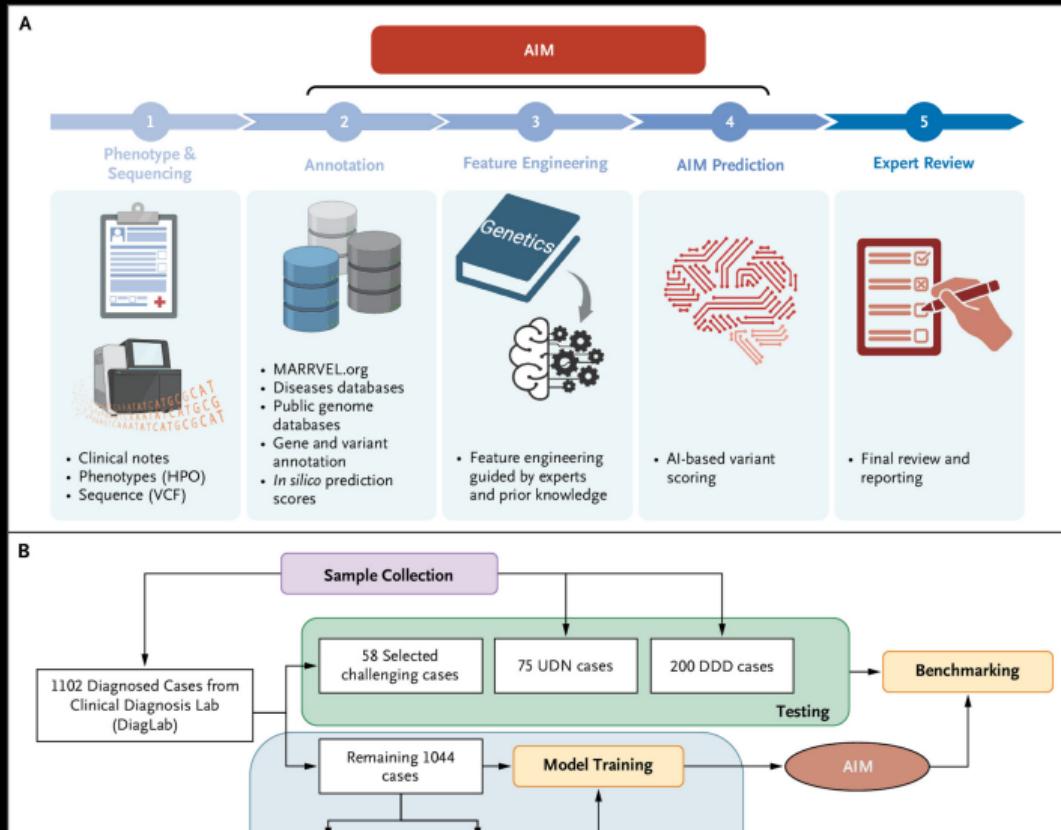
Pipeline Overview:

- Data collection for building AI-MARRVEL
- Sequencing & clinical annotation
- Feature engineering (data projection)
- Classifier
- Hyperparameter tuning
- Evaluation
- Data collection for validation

3D Scatter Plot of Pathogenic and Benign Variants



AI-MARRVEL: A State-of-the-Art Method



AI-MARRVEL

Mao D, Liu C, Wang L, et al.

AI-MARRVEL—A Knowledge-Driven AI System for Diagnosing Mendelian Disorders. *NEJM AI*. 2024;1(5).

LA-MARRVEL: LLM-Augmented MARRVEL

LA-MARRVEL: A Knowledge-Grounded and Language-Aware LLM Reranker for AI-MARRVEL in Rare Disease Diagnosis

Jaeyeon Lee,^{1,2} Hyun-Hwan Jeong,^{1,2*} Zhandong Liu^{1,2,3,*}

¹ Department of Pediatrics, Baylor College of Medicine, Houston, Texas, 77030

² Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, Texas, 77030

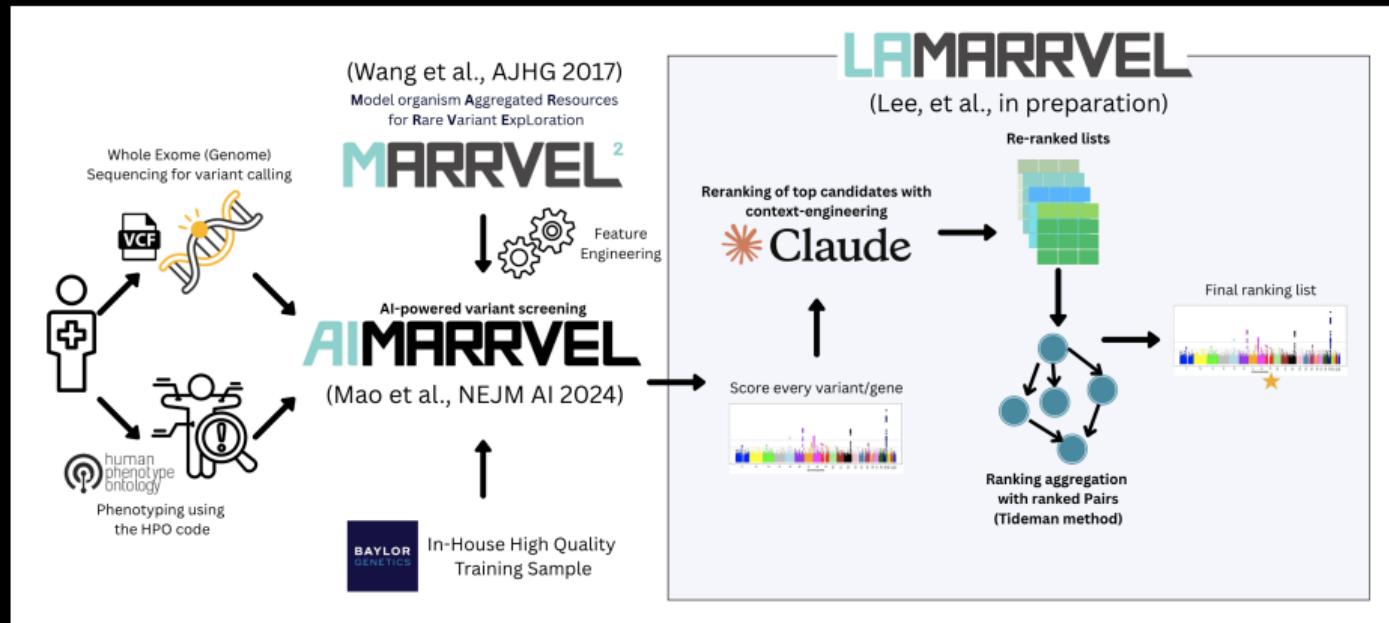
³ Quantitative and Computational Biosciences program, Baylor College of Medicine, Houston, Texas, 77030

*Co-corresponding authors: hyun-hwan.jeong@bcm.edu, zhandonl@bcm.edu

ABSTRACT

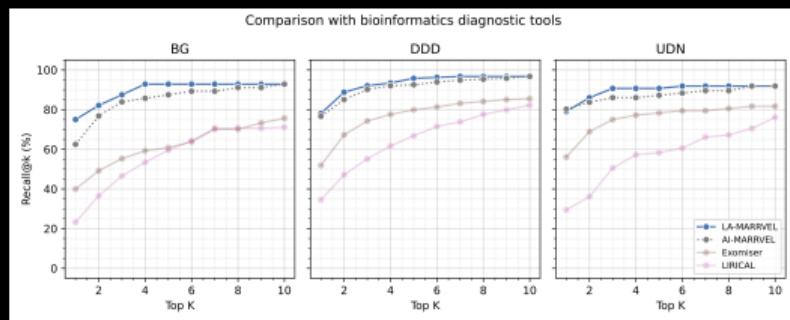
Diagnosing rare diseases often requires connecting variant-bearing genes to evidence that is written as unstructured clinical prose, which the current established pipelines still leave for clinicians to reconcile manually. To this end, we introduce LA-MARRVEL, a knowledge-grounded and language-aware reranking layer that operates on top of AI-MARRVEL: it supplies expert-engineered context, queries a large language model multiple times, and aggregates the resulting partial rankings with a ranked voting method to produce a stable, explainable gene ranking. Evaluated on three real-world cohorts (BG, DDD, UDN), LA-MARRVEL consistently improves Recall@K over AI-MARRVEL and established phenotype-driven tools such as Exomiser and LICAL, with especially large gains on cases where the first-stage ranker placed the causal gene lower. Each ranked gene is accompanied by LLM-generated reasoning that integrates phenotypic, inheritance, and variant-level evidence, thereby making the output more interpretable and facilitating clinical review.

AI-MARRVEL + LA-MARRVEL = Smarter, Context-Aware Genomic Insights



When LLMs Join Forces with AI-MARRVEL: Smarter Variant Discovery

Comparative Performance Analysis of LA-MARRVEL



(Lee, Jeong, and Liu, ArXiv 2025)

Performance Comparison of LA-MARRVEL and Vanilla LLM Models

[See WMF figure—convert manually if needed]

Context Engineering: The Key Factor Behind the Performance Boost

tobi lutke @tobi - Jun 18

I really like the term "context engineering" over prompt engineering.

It describes the core skill better: the art of providing all the context for the task to be plausibly solvable by the LLM.

Andrei Karpathy @karpathy

+1 for "context engineering" over "prompt engineering".

So context engineering is just one small piece of an emerging thick layer of non-trivial software that coordinates individual LLM calls (and a lot more) into full LLM apps. The term "ChatGPT wrapper" is tired and really, really wrong.



(Lee, Jeong, and Liu, ArXiv 2025)

Case Study: LLM Correctly Re-Ranks SGP7 Through Clinical Context

LLM Explainer Trace for Case Study 1

...]

CLDN16

ACMG Classification: Likely Benign

- Molecular evidence:**
 - Very low CADD (0.019) suggests benign
 - Single heterozygous variant for AR condition
 - ClinVar: Uncertain significance
- Phenotype matching: POOR (3/10)**
 - Primarily renal/metabolic disorder
 - Some overlap: nystagmus, muscle weakness, hypertension

Conclusion — Insufficient evidence for pathogenicity

SPG7

ACMG Classification: Likely Pathogenic

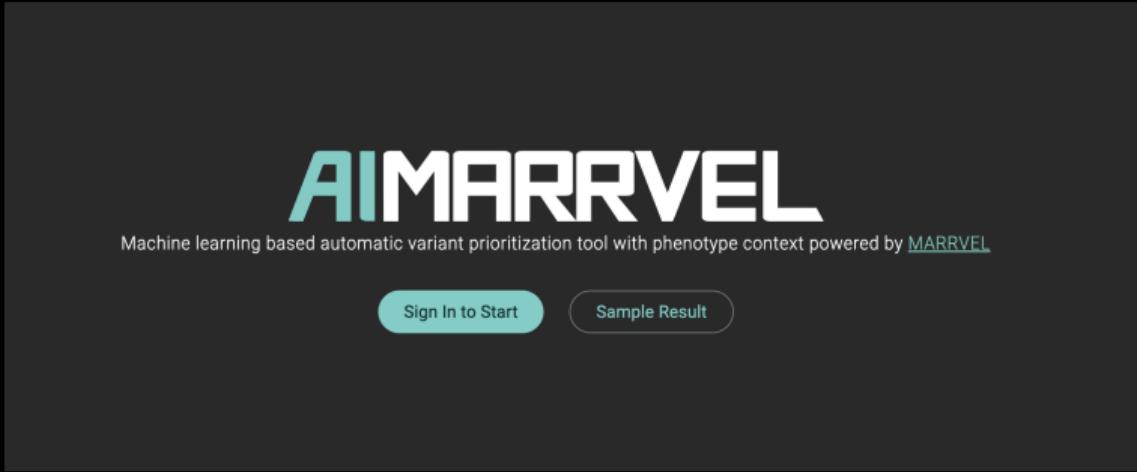
- Molecular evidence:**
 - High impact missense (CADD 27.8 — strong computational evidence)
 - ClinVar: Conflicting interpretations suggests some pathogenic evidence
 - Concern:** High population frequency (0.002923) reduces pathogenicity likelihood
- Phenotype matching: EXCELLENT (9/10)**
 - HPO matches: Nystagmus, Muscle weakness, Gait ataxia, Dysarthria, Dysdiadochokinesis, Spasticity, Unsteady gait
 - Disease inheritance: AD/AR [heterozygous variant can be causative in AD form]

Conclusion — Despite population frequency concerns, phenotype match is exceptional

{additional genes omitted for brevity}

(Lee, Jeong, and Liu, ArXiv 2025)

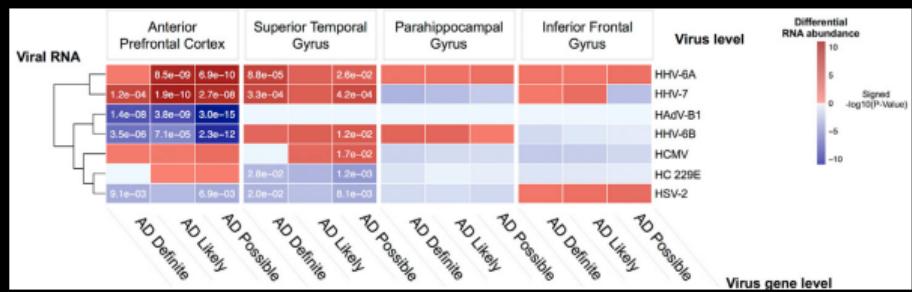
AI-MARRVEL: Try It Now



<https://ai.marrvel.org>

Trustworthy AI

Replication is Not Optional: Lessons from the HHV–Alzheimer's Debate

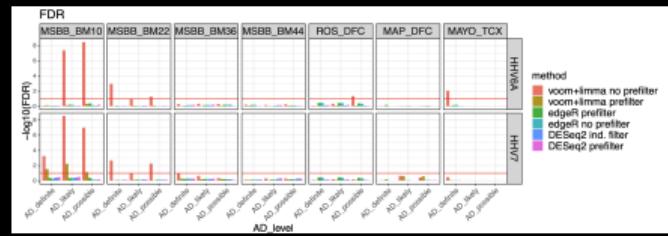


TREATMENTS
Researchers Find Herpes Viruses In Brains
Marked By Alzheimer's Disease
JUNE 21, 2010 · 11:10 AM ET
HEARD ON ALL THINGS CONSIDERED
by Jon Hamilton

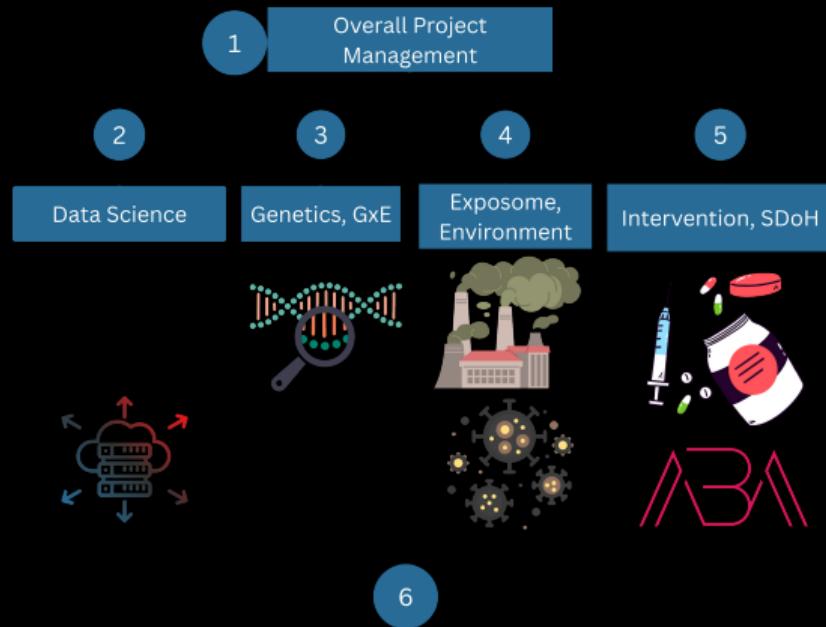
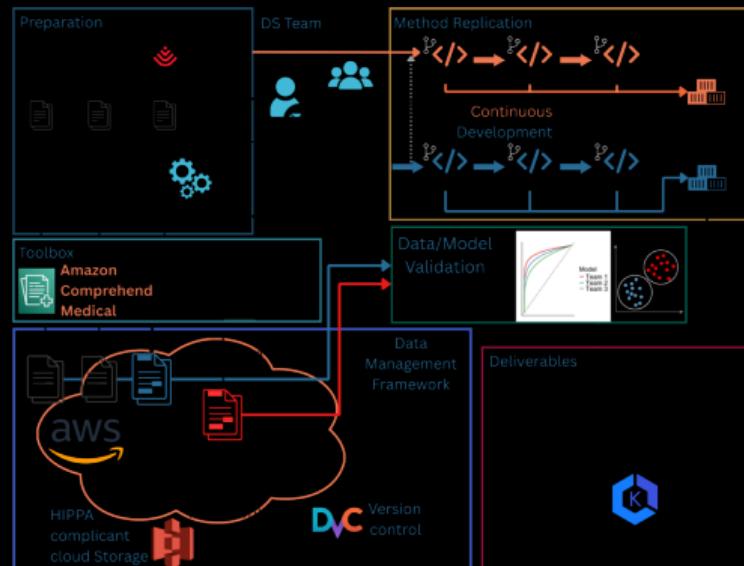


Replication is Not Optional: Further Evidence

Neuron Letter



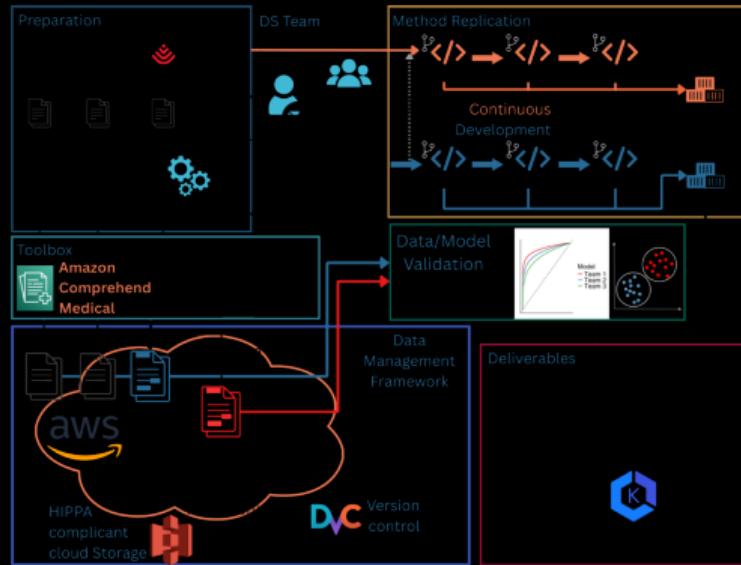
VALIDATE ASD: Strengthening Autism Data Science Through Rigorous AI Validation



Creating a Secured/Reproducible Workflow That Makes ASD AI

Compute, Security,
Infrastructure

VALIDATE ASD: Key Results



Project ID	Project Name	Year	Site	Principal Investigator(s)/Project Leader(s)	Organization	Fiscal Year	Admin ID	Funding ID	PV Total	Cost by ID	Similar Projects
1-GZL000000005-01	VANIJETTE, KIRI (INDEPENDENT MULTIMODAL REPLICATION AND VALIDATION OF AUTISM DATA SCIENCE MODELS)	2025	08	80	\$485,000	View					

Conclusion

Epilōgus: Moving Toward Trustworthy and Effective AI in Precision Medicine

Smarter:

- Thinking LLMs (o1, DeepSeek-R1) bring new reasoning capabilities
- ConSol: comparable accuracy with ~80% fewer tokens

Better:

- AI-MARRVEL: state-of-the-art rare disease diagnosis
- LA-MARRVEL: LLM-augmented context-aware genomic insights

Faster & Greener:

- Efficient inference through sequential testing
- Context engineering as a key performance multiplier

Trustworthy:

- Replication is not optional—rigorous validation matters
- VALIDATE ASD: building reliable, reproducible AI workflows

Thank You

Collaborators:

- Zhandong Liu, PhD
- Jaeyeon Lee
- Johnathan Jia
- Seon Young Kim

Funding Support:

- Chao Endowment
- The Huffington Foundation
- Jan and Dan Duncan Neurological Research Institute
- IDDRC
- Grant #2023-332162
- CPRIT RP240131
- NIH 1R01HG011795-01
- NIH 1OT2OD040565-01

Questions?