Accelerating Variant Interpretation with Large Reasoning Models

Benchmarking Evidence Extraction Across Leading Models

Mullai Murugan

Human Genome Sequencing Center

Baylor College of Medicine

Disclosure

I do not have any disclosures.

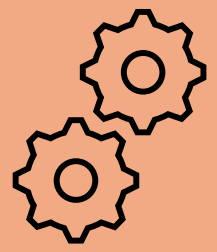
Accelerating Variant Interpretation: The Challenge & Large Reasoning Model (LRM) Opportunity

Genomic Testing has scaled

- More data, variants

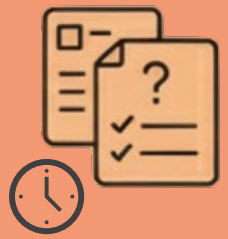
Interpretation: Resource Intensive

- ACMG/AMP guidelines: 28 criteria, 5 tiers
~80min for a hard variant



Automatable Criteria

- BA1, BS1, PP3: DB, In silico



Interpretive Criteria

- PS4, PS3, PP4: Literature Synthesis

LRM: The Acceleration Engine

- Help automate evidence extraction
- Boost geneticist productivity

Automating PS4 Evidence Extraction with LRMs

ACMG Variant Interpretation
- 28 criteria/5 tiers

Manual Criteria
PS3, PS4

Automatable Criteria

PS4: Extracting Case Evidence

📖 Extract & count PS4 eligible cases from literature
📈 Prevalence: cases > controls 🔎 Complexity: variant IDs, phenotypes, publications
👤 Requires expert human review

LRMs: Automating PS4 Evidence Extraction

💡 Apply rules/guidelines & perform complex reasoning
⚙️ Automate PS4 evidence extraction
👤 Reduce expert review burden & time

Using LRMs for Literature-Based Case Evidence Extraction (PS4 Criterion)

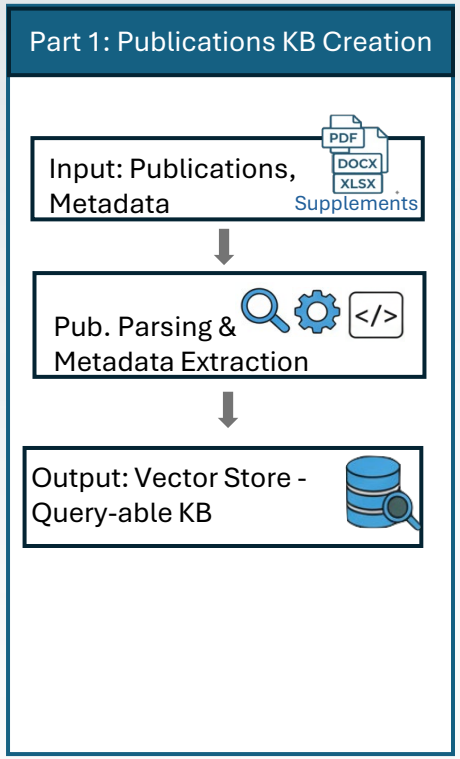
- **Task 1** : Detect variant in publications
- **Task 2** : Count cases (variant + phenotype match)
  - Extract case data, citations & reasoning

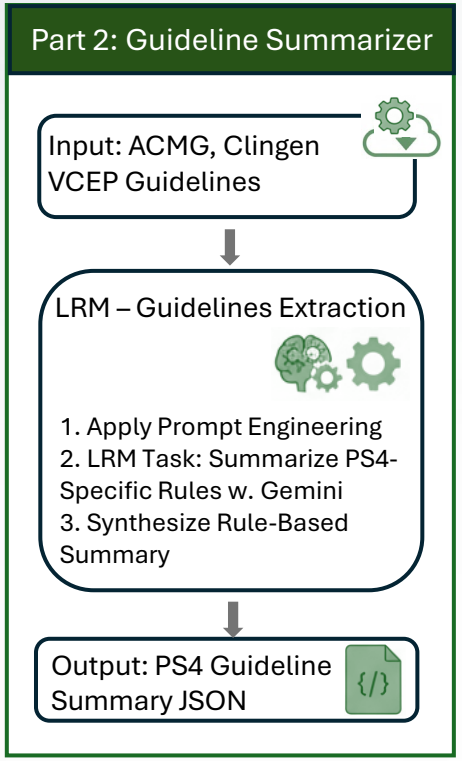
- Benchmarking against *expert-curated ground truth*
  - Models: *Large* - OpenAI GPT-5, OpenAI o3, Google Gemini 2.5 Pro
    Small - Anthropic Claude Sonnet 4, OpenAI o4-mini

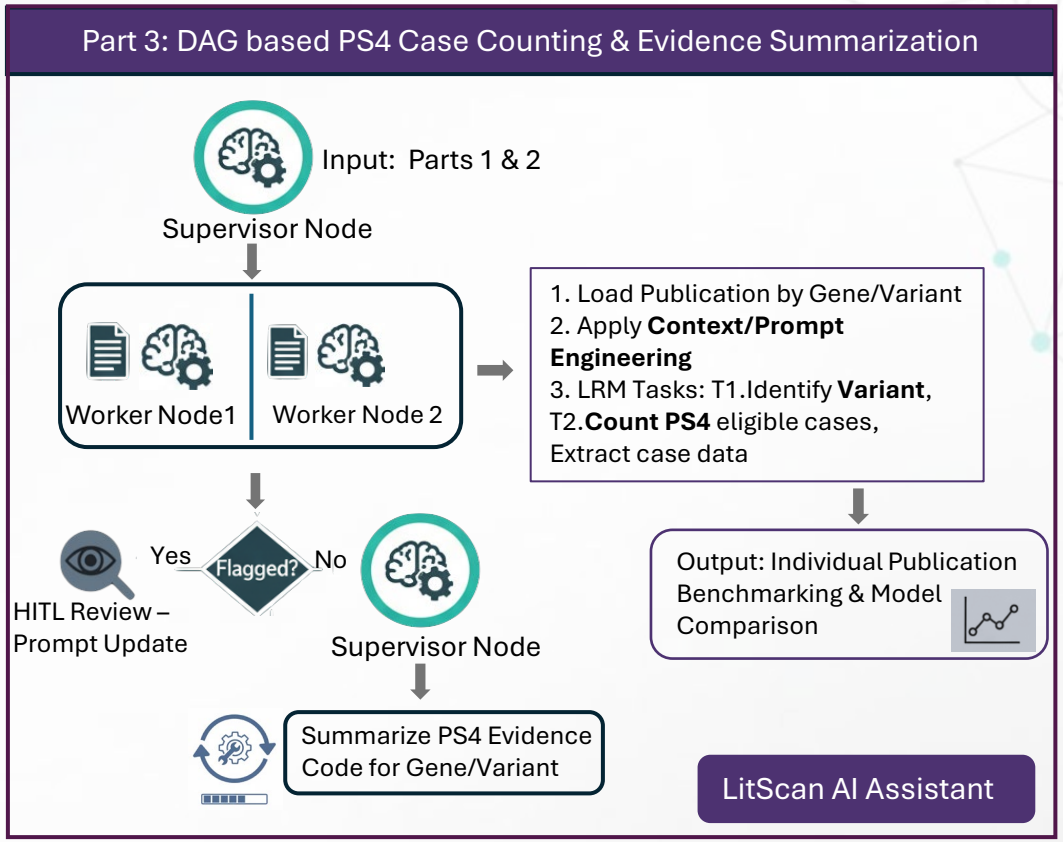
- **Goal:** Evaluate LRM utility & reliability for ACMG-guided, literature-based variant evidence extraction.

Design

Ground Truth Benchmark Dataset	
Publications	281
Genes	58
Variants	128





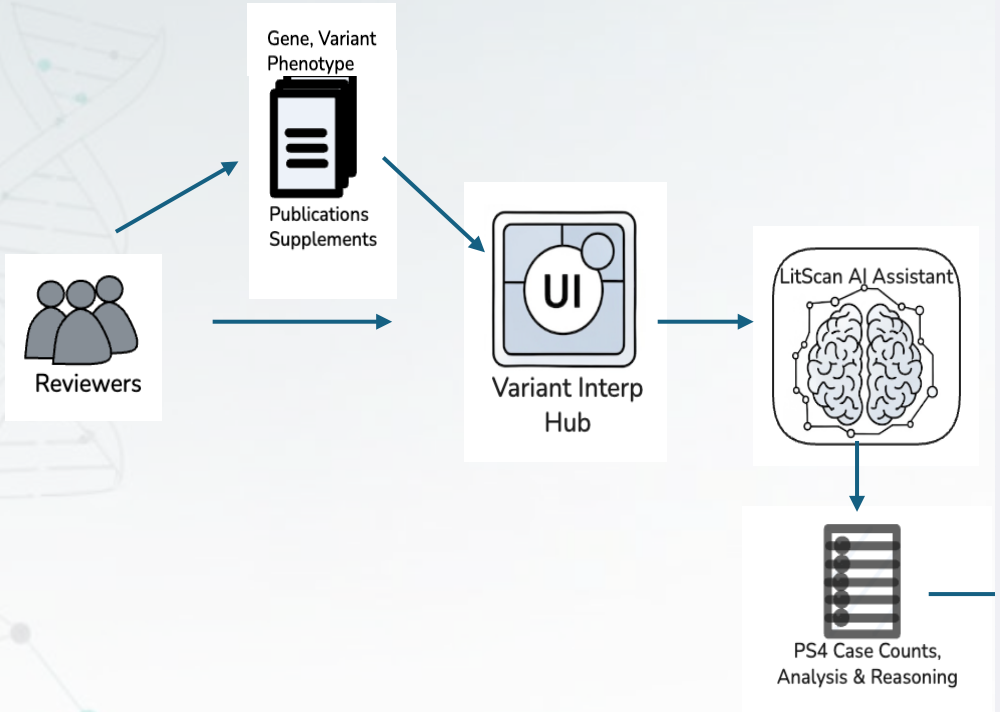


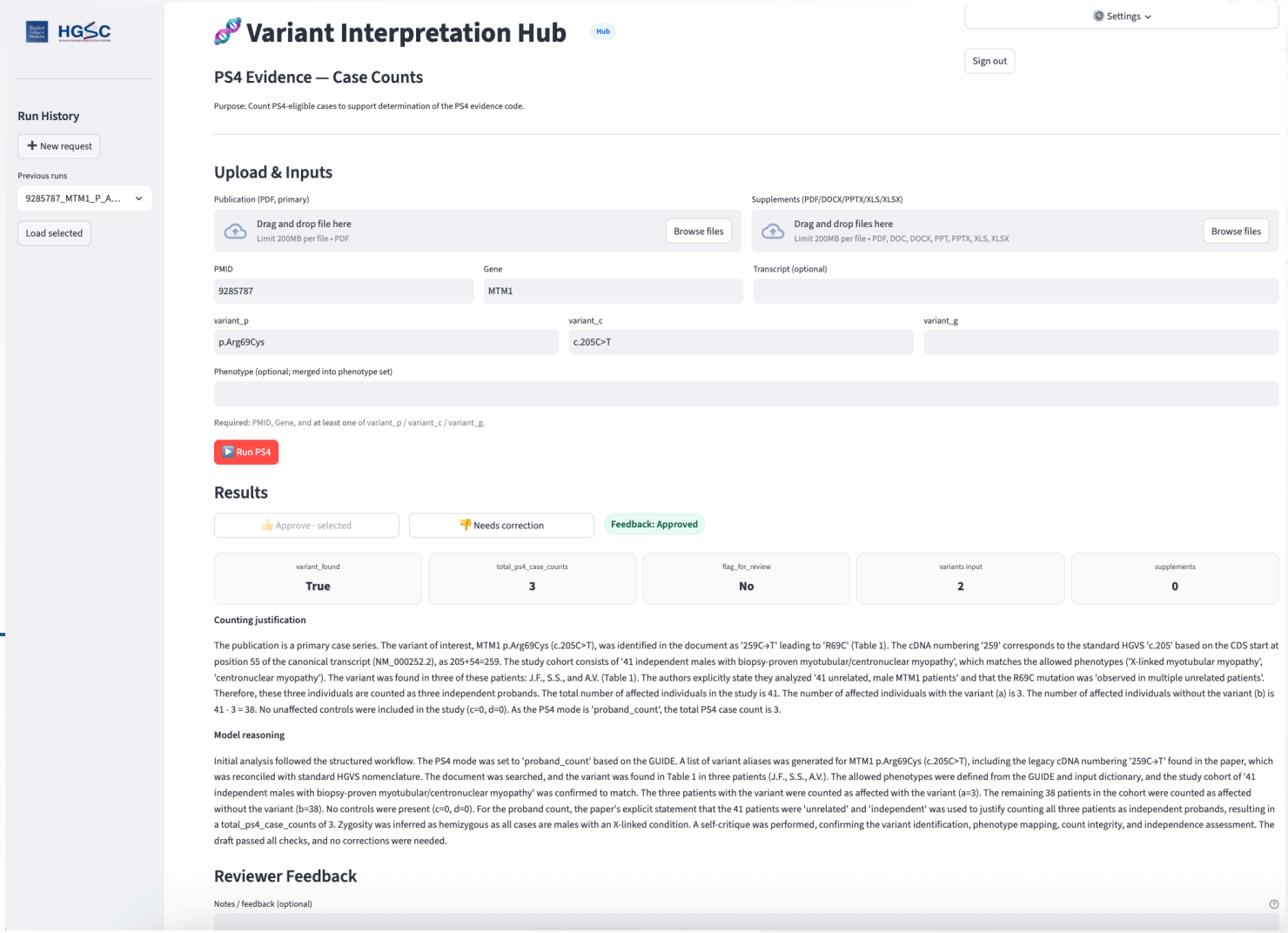
Empowering personalized pharmacogenomics with generative AI solutions

Mullai Murugan, MS ✉, Bo Yuan, PhD, Eric Venner, PhD, Christie M Ballantyne, MD, Katherine M Robinson, PharmD, James C Coons, PharmD, Liwen Wang, PhD, Philip E Empey, PharmD, PhD, Richard A Gibbs, PhD

Journal of the American Medical Informatics Association, Volume 31, Issue 6, June 2024, Pages 1356–1366, https://doi.org/10.1093/jamia/ocae039

Variant Interpretation Hub





Benchmarking Results: Task 1 - Variant Detection

	Model	N	TP	TN	FP	FN	Accuracy	F1
Large	**Google Gemini 2.5 Pro**	281	262	13	0	6	**0.979**	0.988
	OpenAI GPT-5	281	263	12	1	5	**0.979**	0.988
	OpenAI o3	281	262	13	0	6	**0.979**	0.988
Small	**Anthropic Claude Sonnet 4**	281	256	9	4	12	**0.943**	0.969
	OpenAI o4-mini	281	250	13	0	18	**0.936**	0.965

Performance Metrics - Model Outcomes vs Ground Truth
(N=281 publications)

Legend	
N – Total Publications	Benchmarked publications in ground truth dataset
TP – True Positive	Variant found in publication **and matches** ground truth
TN – True Negative	Variant **not found** in publication **and absent** in ground truth
FP – False Positive	Variant found in publication **but not** in ground truth
FN – False Negative	Variant **not found** in publication **but present** in ground truth

Task 1: Variant Detection - Large vs Small Model

PMID: 10559517(1998) Variant: LDLR, c.1444G>C, p.Asp482His In Table1 as D461H

Gemini (Small)

The variant of interest, p.Asp482His (c.1444G>C), is identified in the publication using a legacy protein nomenclature, D461H, which accounts for the cleavage of the 21-amino acid signal peptide (482 - 21 = 461). The paper reports this variant as '1443 G/C', which is treated as the same molecular event. ..Table 1 (page 312) lists the 'D461H' variant as being identified in 1 family ('No.' column = 1).

Gemini (Large)

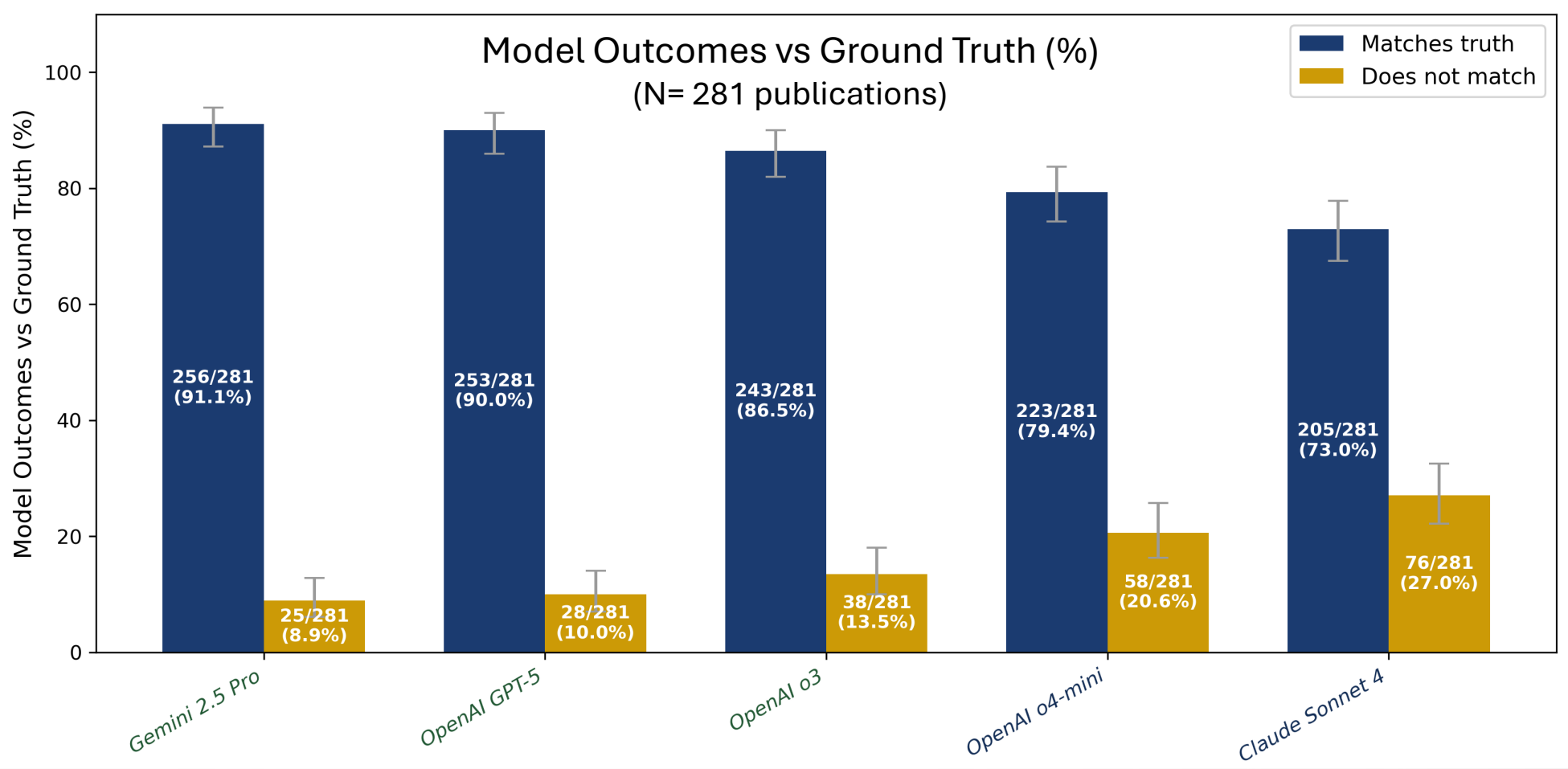
The variant of interest, p.Asp482His (c.1444G>C), is identified in the publication using a legacy protein nomenclature, D461H, which accounts for the cleavage of the 21-amino acid signal peptide (482 - 21 = 461). The paper reports this variant as '1443 G/C', which is treated as the same molecular event. ..Table 1 (page 312) lists the 'D461H' variant as being identified in 1 family ('No.' column = 1).

Table 1
LDLR mutations identified in this study of Northern Irish families with familial hypercholesterolaemia

Mutation	Base change	Exon-LDLR	No.	Reference
E80K	301 G/A	3	2	[5]
211delG	211delG	3	1	[12]
W66G	259 T/G	3	1	[5]
C68Y	266 G/A	3	1	[22] abs.
313+1G/A	313+1G/A	3	1	[23]
C210X	693 C/A	4	5	[24]
C163Y	551 G/A	4	4	[22] abs.
319insA	319insA	4	1	This study
C152X	518 C/A	4	1	[22] abs.
D200G	662 A/G	4	1	[5]
680delAC	680delAC	4	1	[24]
932delA	932delA	6	5	[22] abs.
K369X	1168 A/T	8	2	[25]
1185delG	1185delG	8	1	This study
D461N	1443 G/A	10	7	[25]
G457R	1431 G/A	10	1	[5]
D461H	1443 G/C	10	1	[22] abs.
W462R	1446 T/C	10	1	[25]
1778delG	1778delG	12	1	This study
C656R	2029 T/C	14	1	[5]
P664L	2054 C/T	14	1	[5]
2292delA	2292delA	15	3	[26]

doi:10.1016/s0021-9150(99)00201-4

Benchmarking Results: Task 2 - PS4 Case Counts



Tasks 1 & 2 – Detailed Model Reasoning

PMID: 24444654 Variant: MUTYH c.821G>A, p.Arg274Gln 0 PS4 cases

Gemini 2.5 Pro: Case Count – 0 ✓

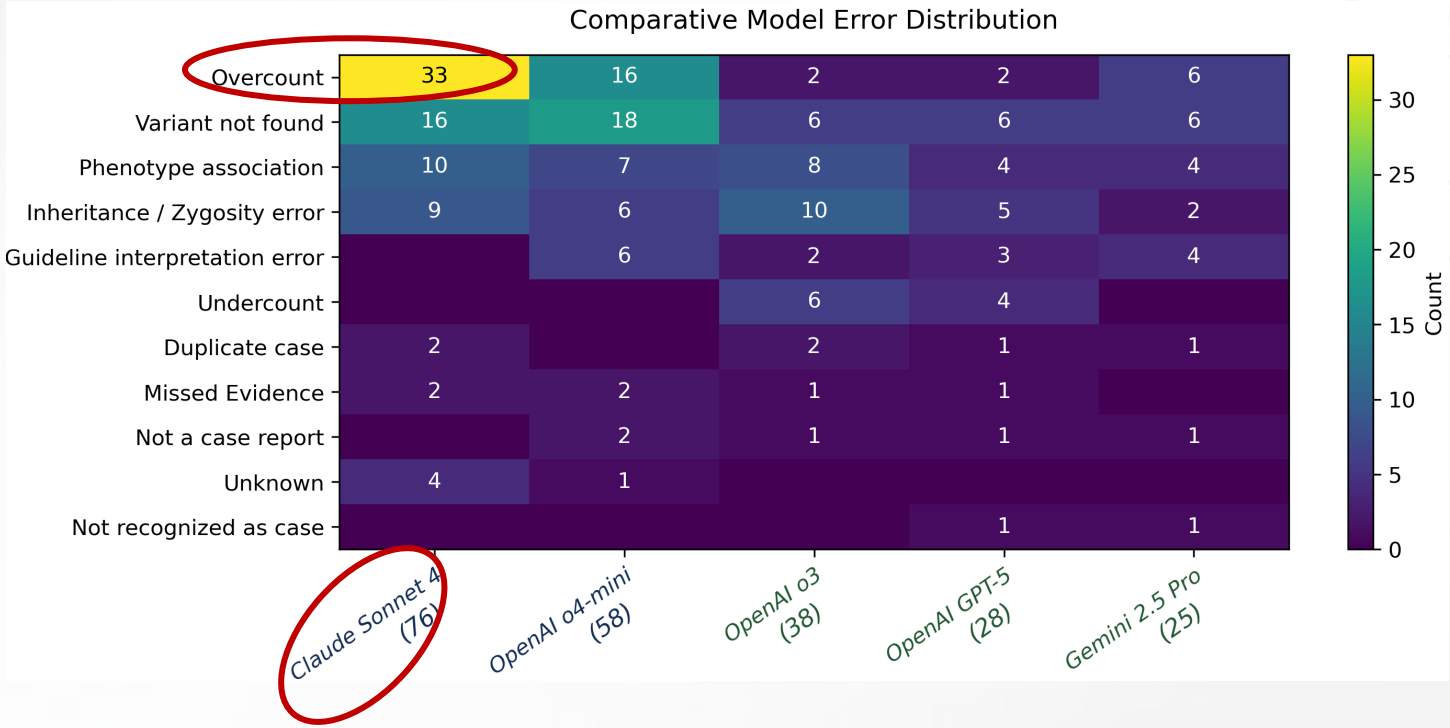
The user wants me to perform a PS4 data extraction for MUTYH p.Arg274Gln from PMID:24444654....
2. **Variant Identification:** I created aliases for p.Arg274Gln (R274Q, c.821G>A) and found it listed as '821G>A, CGG>CAG, R274Q' in Supplementary Table 2 of the document.
3. **Phenotype Matching:** The probands in the study have colorectal cancer (CRC), which is the primary clinical outcome of 'familial adenomatous polyposis 2' (MUTYH-associated polyposis), the allowed phenotype....
4. **Raw Count Tally:** Supplementary Table 2 lists 1 biallelic (compound heterozygous with G396D) and 13 monoallelic probands with the R274Q variant, for a total of 14 affected individuals (a=14)...
5. **PS4 Eligibility Application:** The crucial step was applying the GUIDE's rules for this autosomal recessive condition... Consequently, despite finding 14 affected carriers, none are eligible for PS4 counting according to the specific rules in the GUIDE. The `total_ps4_case_counts` is set to 0.
6. **Finalization: ...**A flag for review is set to `true` because the GUIDE's rules effectively prevent the application of PS4 in this context, which is a significant finding that may require expert review.

DOI: 10.1053/j.gastro.2014.01.022

Benchmarking Results: Comparative Analysis of Model Error Patterns

Key Error Categories:

- Variant Identification
  - Difficulty with correctly identifying the variant in the publication
- Proband Counts
  - Unable to count unrelated probands only, either **overcounting** or **undercounting** cases
- Phenotype Association
  - Unable to identify or correctly link cases to phenotypes
- Guideline Interpretation
  - Incorrect interpretation or misapplication of guidelines
- Publication Complexity
  - Difficulty with processing complex/large literature
- Preexisting Literature/DB
  - Unable to identify if literature is primary case series or duplicate
- Unknown Errors



Stochasticity - Variability in LRM Outputs

Quantify effects of stochasticity on:

- 🧬 Task 1: Variant Detection
- 📊 Task 2: PS4 Case Counts
- 📝 Semantic Reasoning

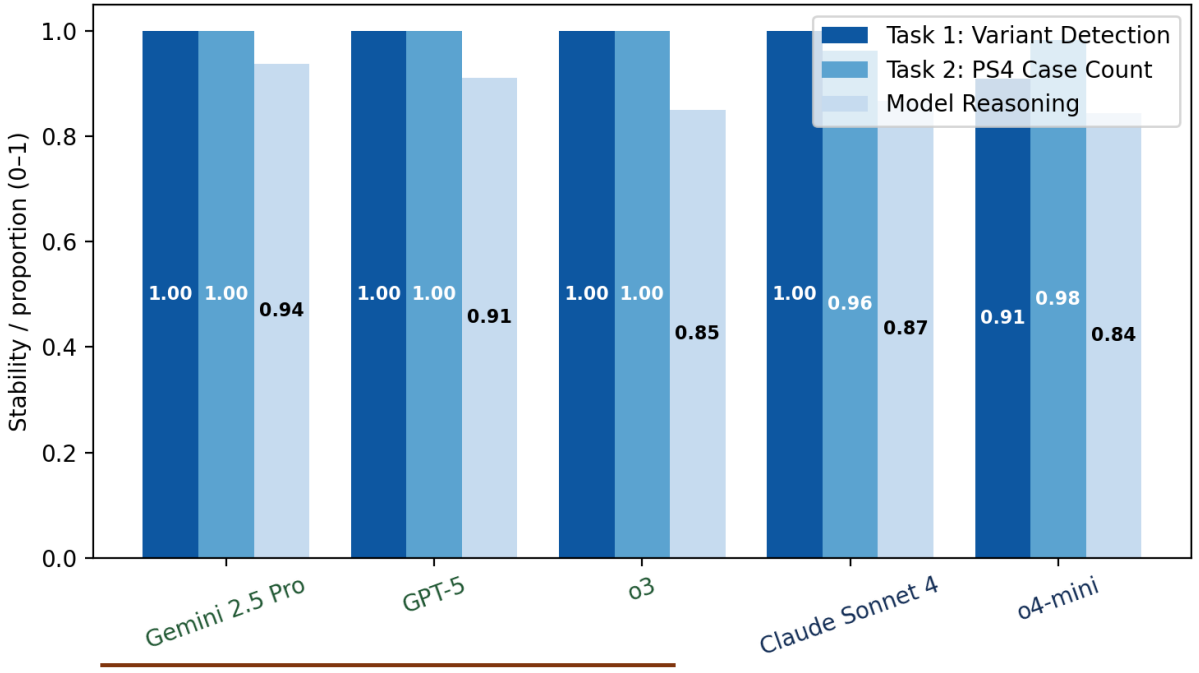
Goal

- Quantify **output variability** under different publication conditions
- Compare models on their **reproducibility and consistency**
- Identify where **safeguards** are needed to manage variability

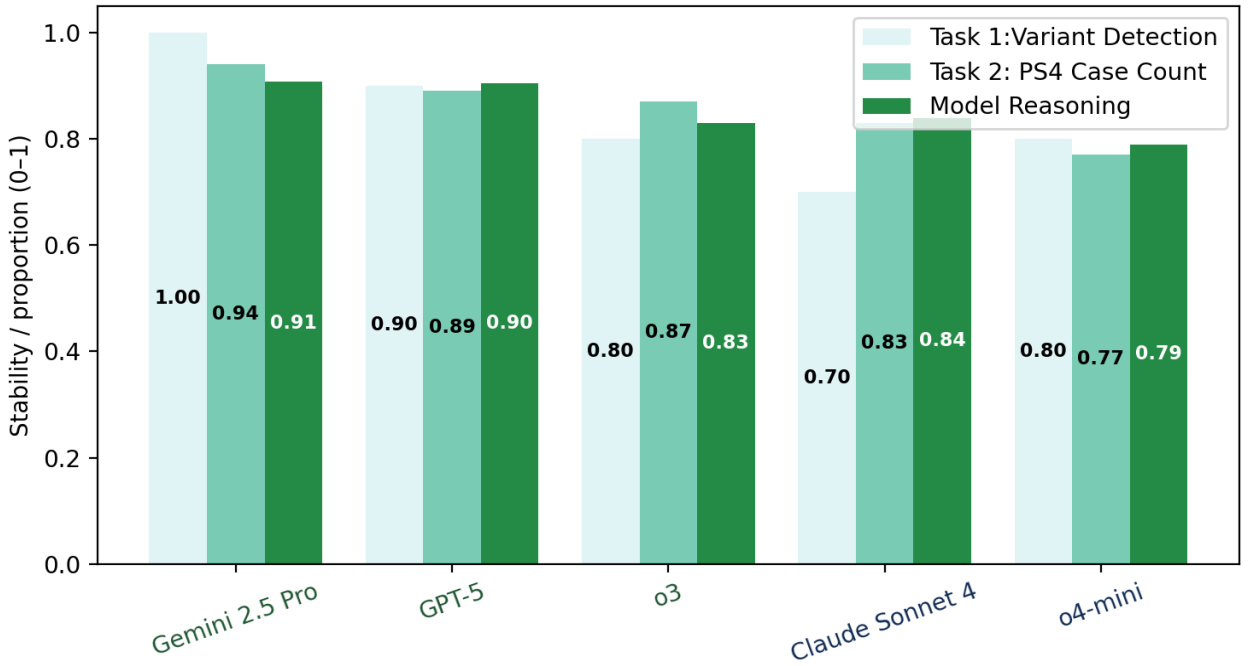
Quantifying Stochastic Variation across Publication Complexity

10 low-complexity and 10 high-complexity publications, 10 runs each

Reproducibility w. Low-Complexity Publications
(higher values = less stochastic variation)



Reproducibility w. High-Complexity Publications
(higher values = less stochastic variation)



Unpublished

Key Takeaways

🧭 **Scope**

- Benchmarked LRM-assisted evidence extraction for ACMG/ClinGen PS4 criterion

📊 **Findings**

- Frontier reasoning models (GPT-5, Gemini 2.5 Pro, o3) achieved ~85–90% concordance with expert-curated ground truth
- Revealed distinct model-specific error profiles, highlighting the need for model-aware prompting, reproducibility safeguards, and human-in-the-loop guardrails

🚀 **Next Steps**

- Refine design & extend scope to other ACMG criteria (e.g., PS3, PP1) and broaden ground truth datasets

🌎 **Broader Impact**

- Framework can generalize to reasoning over biomedical literature beyond variant interpretation

Acknowledgements

Baylor College of Medicine

Human Genome Sequencing Center

Dr. Richard Gibbs, PhD
Dr. Eric Venner, PhD
Dr. Bo Yuan, PhD
Dr. Joshi Stephen, PhD
Dr. Charul Gijavanekar, PhD
Dr. Senkot Kadirvel, PhD

Loka

Dr. Jorge Sampaio, PhD
Vitor Manita
Francisco Delca Pereira
John Restrepo

Amazon Web Services

Patrick Johnson
Alvin Richardson

murugan@bcm.edu

Questions/Discussion

Appendix A- Models, Context Window, Cost

	OpenAI GPT-5	OpenAI o3	OpenAI o4-mini	Google Gemini 2.5 Pro	Anthropic Claude Sonnet 4
Context Window (tokens)	400,000	200,000	200,000	1,000,000	200,000
Knowledge cutoff	Sep 30, 2024	June 01, 2024	June 01, 2024	Jan 2025	Jan 2025
Reasoning Support	Yes	Yes	Yes	Yes	Yes
Pricing (per 1M tokens)	Input: $1.25 Output: $10.00	Input: $2.00 Output: $8.00	Input: $1.10 Output: $4.00	Input: $1.25 Output: $10.00	Input: $3 Output: $15.00