

Accelerating Variant Interpretation with Large Reasoning Models

Benchmarking Evidence Extraction Across Leading Models



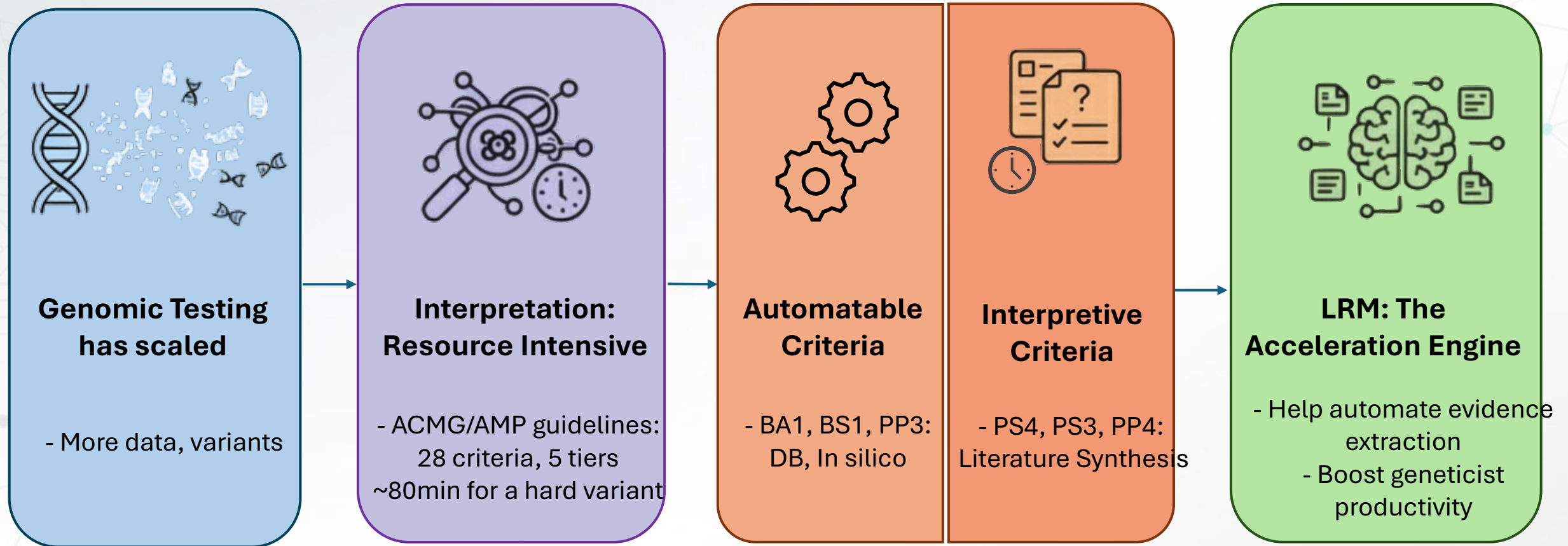
Mullai Murugan
Human Genome Sequencing Center
Baylor College of Medicine

Disclosure

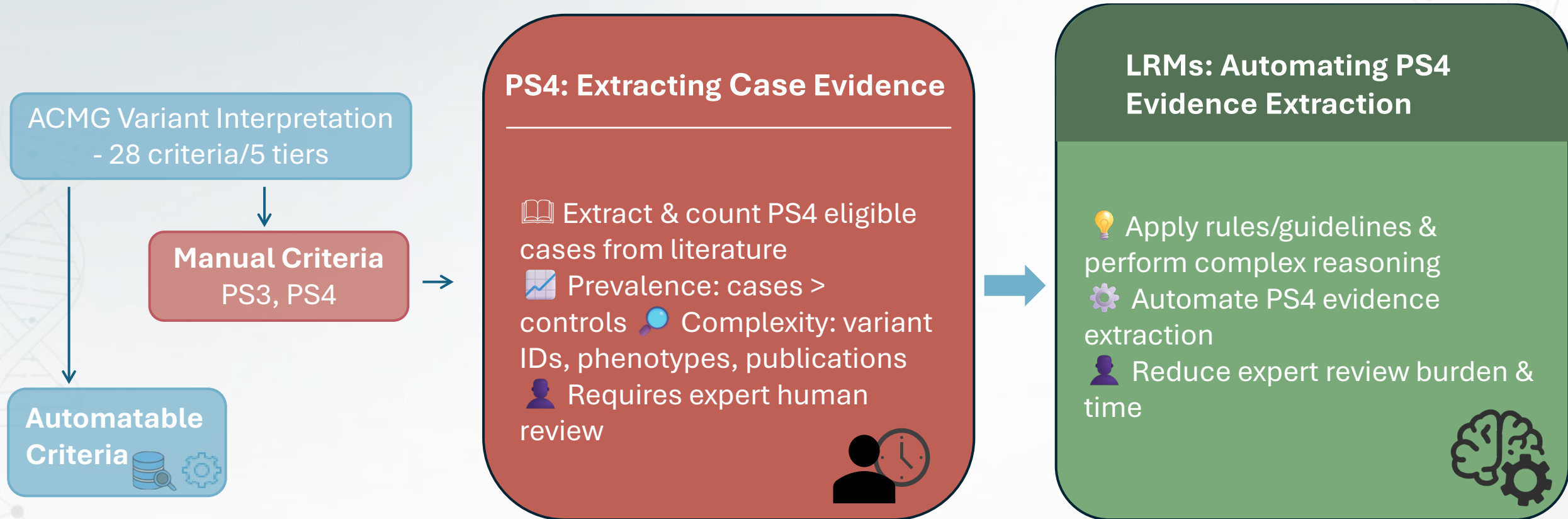
I do not have any disclosures.



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



Automating PS4 Evidence Extraction with LRMs

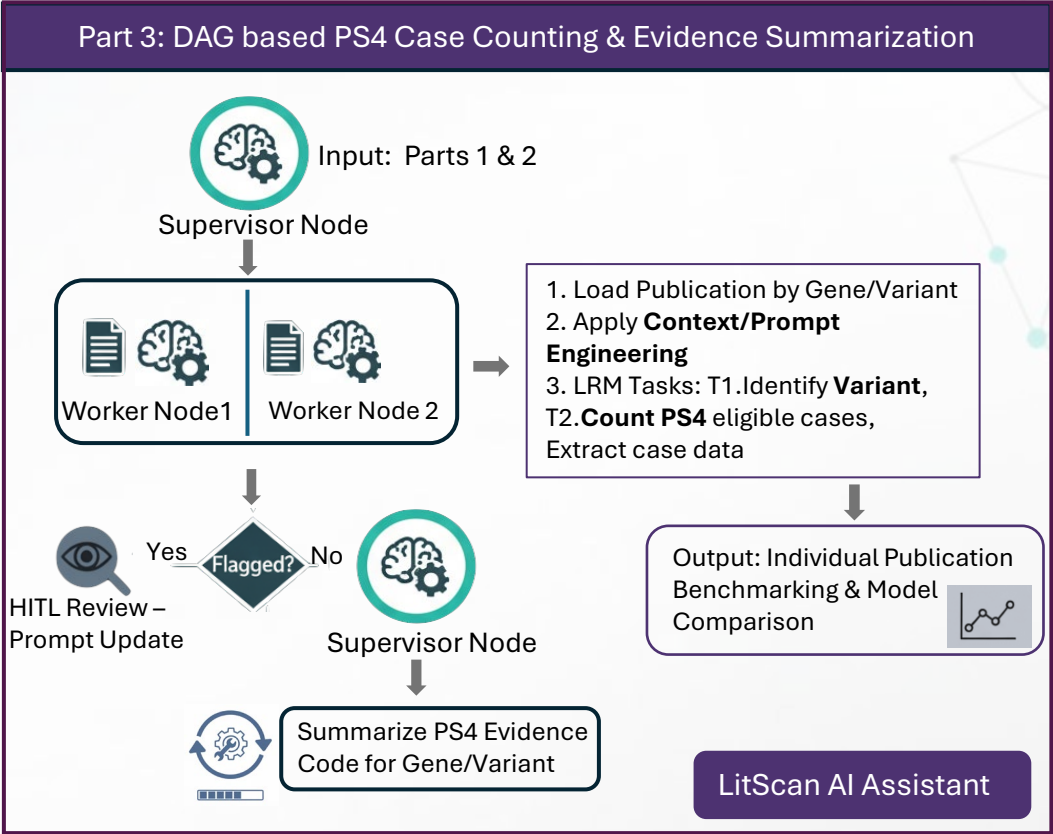
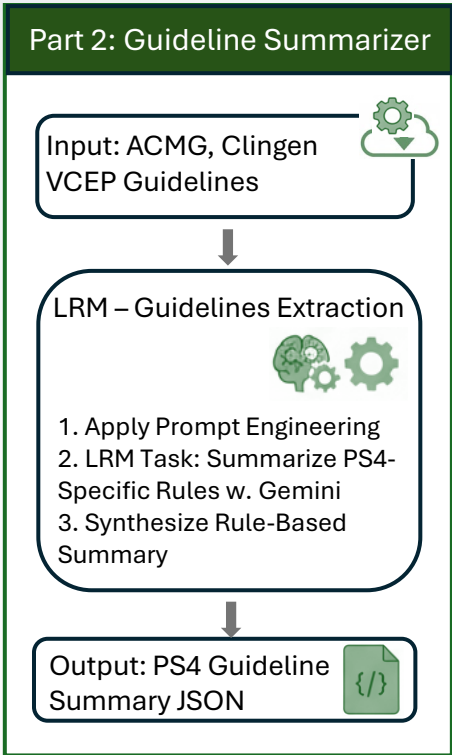
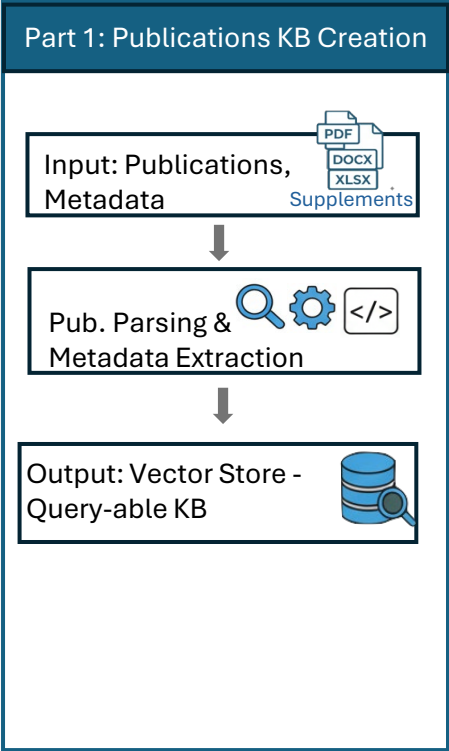


Using LRMs for Literature-Based Case Evidence Extraction (PS4 Criterion)

- 📄 **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

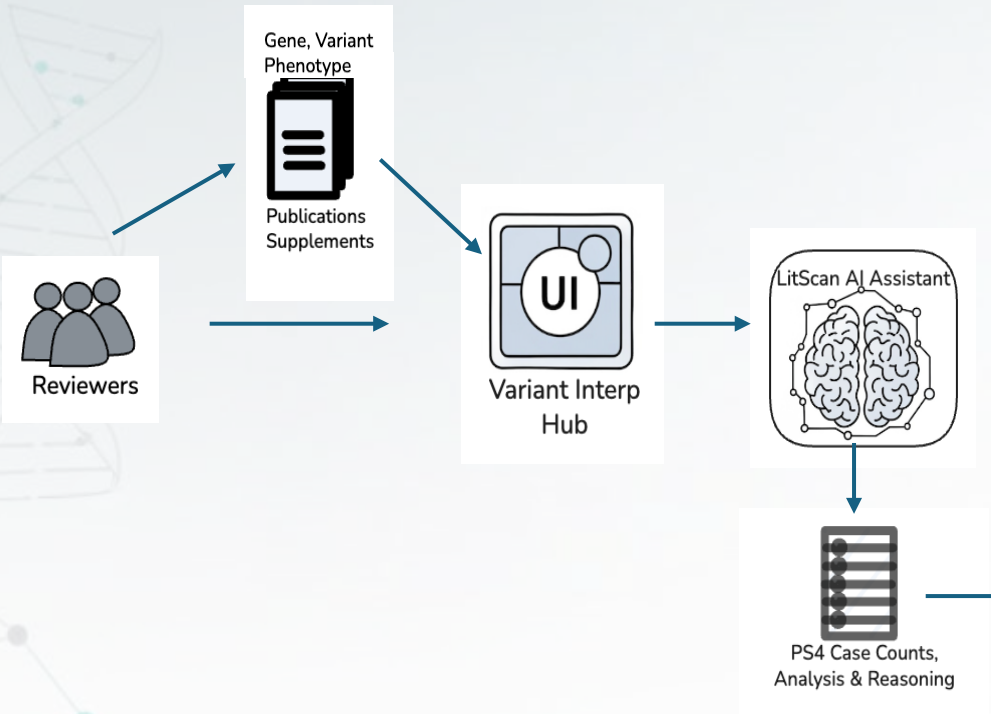


Empowering personalized pharmacogenomics with generative AI solutions

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



Run History

+ New request

Previous runs
9285787_MTM1_P_A...
Load selected

Variant Interpretation Hub Hub

Settings

Sign out

PS4 Evidence — Case Counts
Purpose: Count PS4-eligible cases to support determination of the PS4 evidence code.

Upload & Inputs

Publication (PDF, primary)
Drag and drop file here
Limit 200MB per file • PDF
Browse files

Supplements (PDF/DOCX/PPTX/XLS/XLSX)
Drag and drop files here
Limit 200MB per file • PDF, DOC, DOCX, PPT, PPTX, XLS, XLSX
Browse files

PMID
9285787

Gene
MTM1

Transcript (optional)

variant_p
p.Arg69Cys

variant_c
c.205C>T

variant_g

Phenotype (optional; merged into phenotype set)

Required: PMID, Gene, and at least one of variant_p / variant_c / variant_g.

Run PS4

Approve · selected

Needs correction

Feedback: Approved

variant_found
True

total_ps4_case_counts
3

flag_for_review
No

variants input
2

supplements
0

Counting justification

The publication is a primary case series. The variant of interest, MTM1 p.Arg69Cys (c.205C>T), was identified in the document as '259C>T' leading to 'R69C' (Table 1). The cDNA numbering '259' corresponds to the standard HGVS 'c.205' based on the CDS start at position 55 of the canonical transcript (NM_000252.2), as 205+54=259. The study cohort consists of '41 independent males with biopsy-proven myotubular/centronuclear myopathy', which matches the allowed phenotypes ('X-linked myotubular myopathy', 'centronuclear myopathy'). The variant was found in three of these patients: J.F., S.S., and A.V. (Table 1). The authors explicitly state they analyzed '41 unrelated, male MTM1 patients' and that the R69C mutation was 'observed in multiple unrelated patients'. Therefore, these three individuals are counted as three independent probands. The total number of affected individuals in the study is 41. The number of affected individuals with the variant (a) is 3. The number of affected individuals without the variant (b) is 41 - 3 = 38. No unaffected controls were included in the study (c=0, d=0). As the PS4 mode is 'proband_count', the total PS4 case count is 3.

Model reasoning

Initial analysis followed the structured workflow. The PS4 mode was set to 'proband_count' based on the GUIDE. A list of variant aliases was generated for MTM1 p.Arg69Cys (c.205C>T), including the legacy cDNA numbering '259C>T' found in the paper, which was reconciled with standard HGVS nomenclature. The document was searched, and the variant was found in Table 1 in three patients (J.F., S.S., A.V.). The allowed phenotypes were defined from the GUIDE and input dictionary, and the study cohort of '41 independent males with biopsy-proven myotubular/centronuclear myopathy' was confirmed to match. The three patients with the variant were counted as affected with the variant (a=3). The remaining 38 patients in the cohort were counted as affected without the variant (b=38). No controls were present (c=0, d=0). For the proband count, the paper's explicit statement that the 41 patients were 'unrelated' and 'independent' was used to justify counting all three patients as independent probands, resulting in a total_ps4_case_counts of 3. Zygosity was inferred as hemizygous as all cases are males with an X-linked condition. A self-critique was performed, confirming the variant identification, phenotype mapping, count integrity, and independence assessment. The draft passed all checks, and no corrections were needed.

Reviewer Feedback
Notes / feedback (optional)

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 publication p.Asp482His (c.1444G>C) was identified in the 1998 study as a mutation in the LDLR gene (Table 1, page 312). The variant is reported as a single family (No. column = 1), and the base change is 1443 G/C (Table 1, page 312). 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').

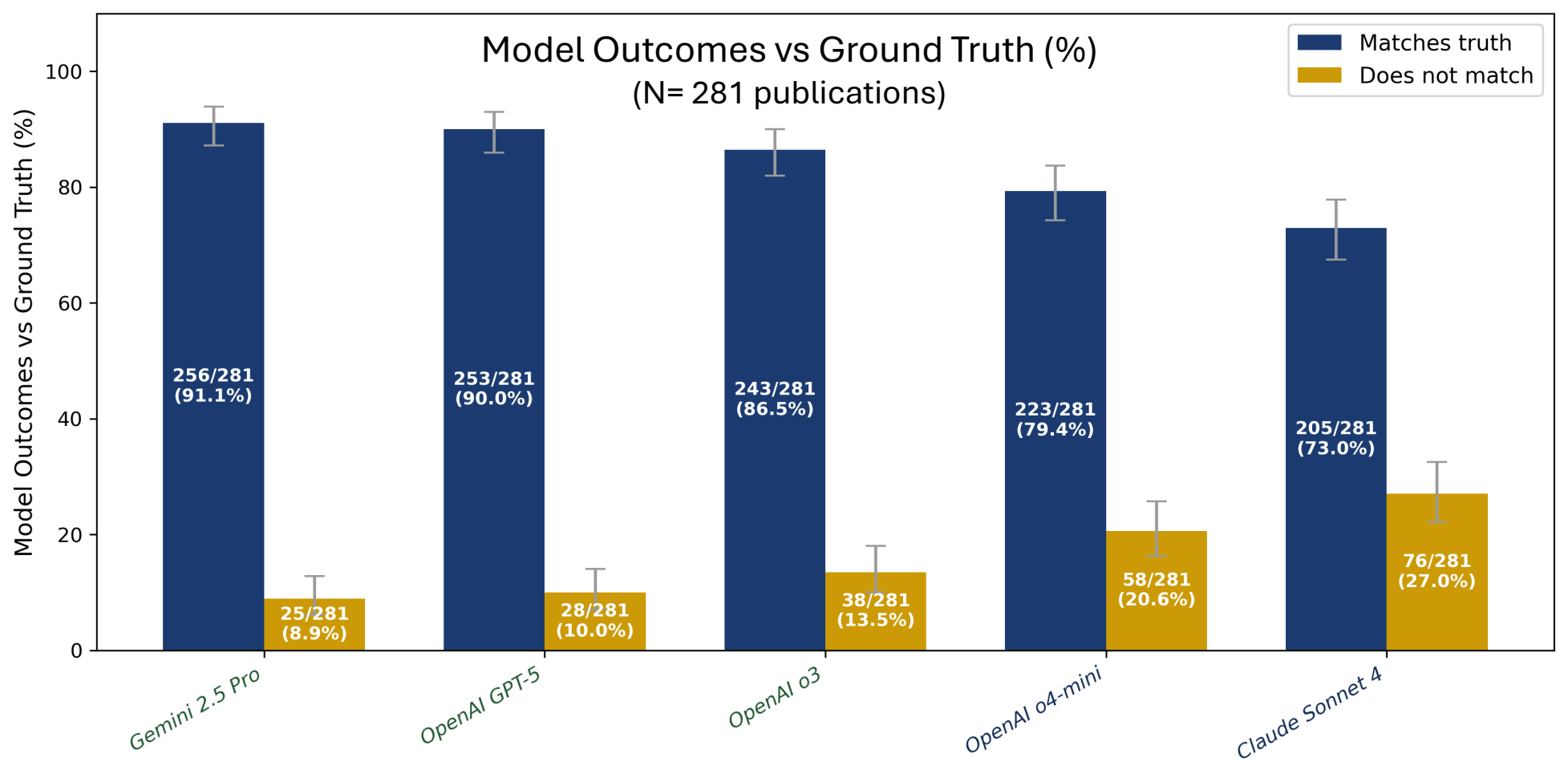
312

C.A. Graham et al. / Ather

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]

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)...

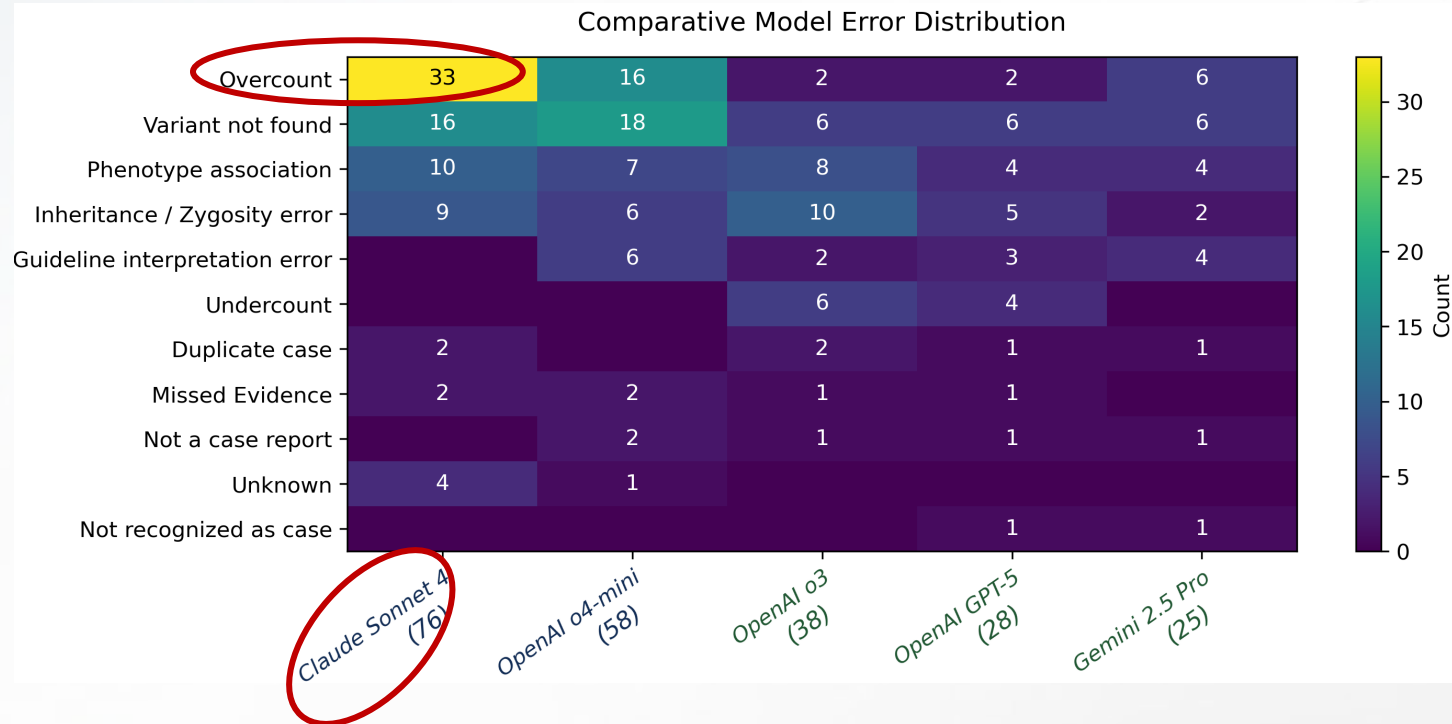
5. **PS4 Eligibility Application:** The crucial step was applying the GUIDE's rules for this autosomal recessive condition... Consequently, despite finding 14 affected carriers, none are eligible for PS4 counting according to the specific rules in the GUIDE. The **`total\_ps4\_case\_counts`** is set 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.

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

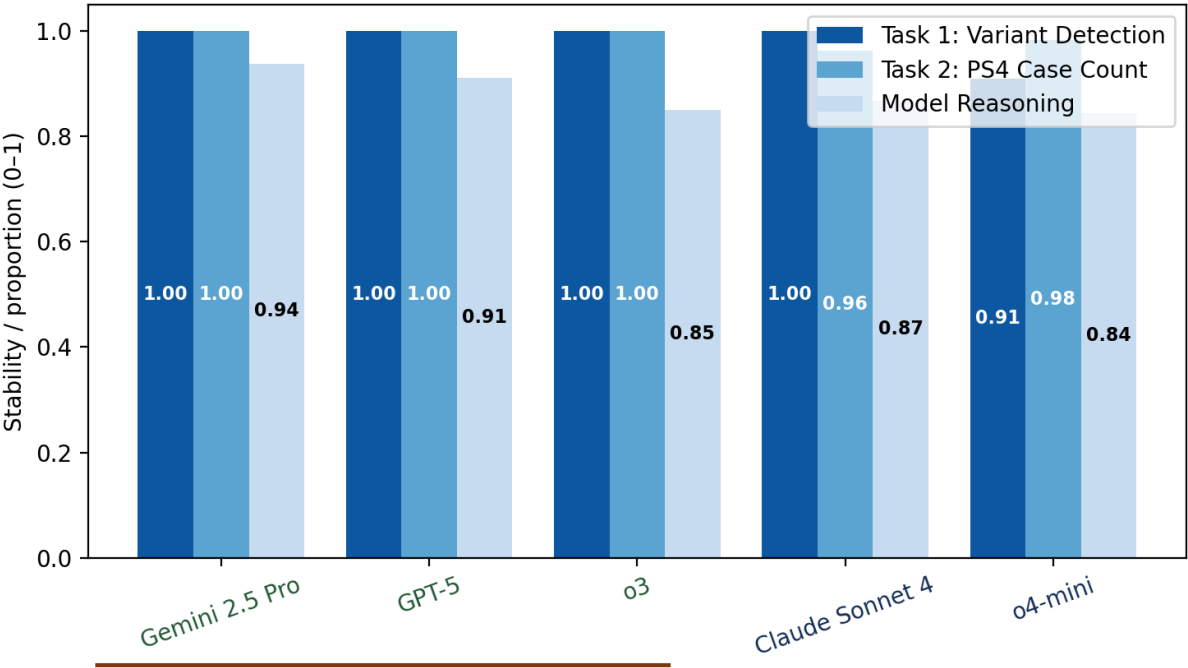
Goal

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

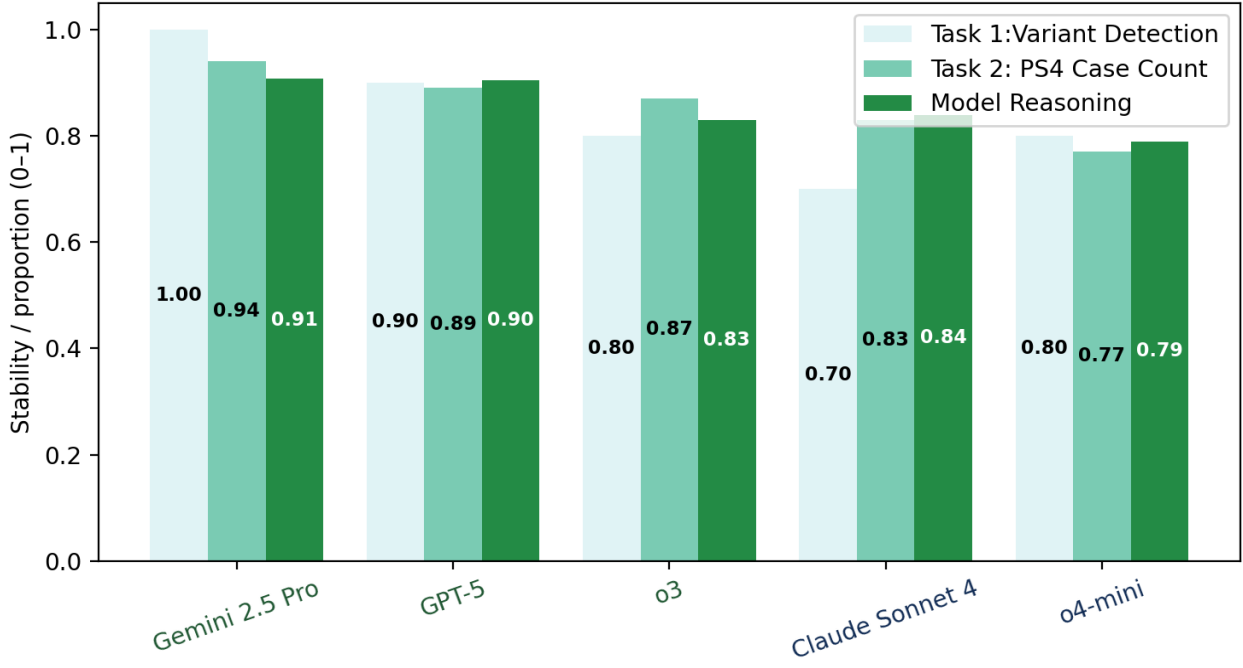
Quantifying Stochastic Variation across Publication Complexity

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)



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