AIVARI Agent: An Evidence-Grounded Agentic LLM for Variant Reportability and Interpretation

Min Kang [*1] **Seungwoo Kim** [*1] **Ki Woong Kwon** [1] **Dongseok Moon** [1]

Abstract

Clinical genomic variant interpretation is a patient-level multi-hypothesis reasoning task that integrates variant evidence, phenotype fit, inheritance, and database knowledge to determine reportability. We propose **AIVARI Agent** (AI VAriant Reportability and Interpretation Agent), an agentic LLM that performs one evidence-grounded rollout per retained candidate gene and jointly evaluates all associated gene-disease hypotheses. On a 300-case clinical cohort (6,460 hypotheses), AIVARI Agent achieves Group Sensitivity of 0.905, Group NPV of 0.933, and Row Precision 0.351. On a 235-case common subset, it improves over an operational hybrid pipeline by +34pp Group Sensitivity and +40pp Group NPV, with the largest gain on Inconclusive findings. These results support single-rollout agentic LLMs with on-demand evidence grounding.

1. Introduction

Clinical genomic diagnosis requires deciding whether a variant could plausibly explain the patient's phenotype, given dozens of candidate variants per patient. More precisely, the task is a multi-step scientific reasoning problem: for each candidate gene-disease hypothesis, the clinician evaluates whether the variant evidence in the gene, together with the patient's phenotype and inheritance pattern, sufficiently supports the disease as a reportable explanation. This decision integrates the molecular characteristics of the variant (e.g., pathogenicity, protein impact, population frequency), the fit between the patient phenotype and the disease's clinical spectrum, the consistency of inheritance, and external evidence such as ClinVar (Landrum et al., 2018), OMIM (Amberger et al., 2015), and the literature. Within a single patient case, multiple candidate hypotheses compete, and the clinician must weigh heterogeneous evidence to reach a final reporting decision.

Prior tool-augmented LLM agents have targeted scientific automation, biomedical research, and rare-disease diagnosis (M. Bran et al., 2024; Boiko et al., 2023; Huang et al., 2025; Zhao et al., 2026). In contrast, system-level validation of *patient-wise multi-hypothesis reportability evaluation* for clinical variant interpretation remains limited.

We formulate clinical variant interpretation and reportability as integrated, candidate-gene-anchored multi-hypothesis reasoning and propose **AIVARI Agent (AI VAriant Reportability and Interpretation Agent)**. Given a patient's HPO (Gargano et al., 2024) phenotype, candidate variants, and case metadata, AIVARI Agent performs an independent single rollout for each retained candidate gene. Within each rollout, the agent jointly evaluates all gene-disease (i.e., (g, d)) hypotheses associated with that gene using a monolithic reasoning trajectory interleaved with OMIM, ClinVar, and candidate-variant detail tool calls. The agent then provides per-hypothesis reportability decisions together with reasoning traces.

Our contributions are as follows:

1. We frame clinical variant interpretation as patient-case-level multi-hypothesis reportability reasoning over candidate (g, d) hypotheses.
2. We propose AIVARI Agent, a variant-anchored single-rollout architecture that evaluates all (g, d) hypotheses associated with the candidate variant's mapped gene in a single reasoning trajectory, with on-demand evidence retrieval via tools.
3. We evaluate on 300 real clinical cases using clinician-authored reports as system-independent ground truth, with a controlled tool ablation and a head-to-head comparison against an operational hybrid pipeline.

2. Methods

2.1. Problem Formulation

Let $\mathcal{G}$, $\mathcal{D}$, and Φ denote the sets of genes, diseases, and patient phenotype profiles (e.g., HPO term sets), respectively. For a given patient case, let $\mathcal{G}_c \subseteq \mathcal{G}$ be the set of candidate genes. For each gene $g \in \mathcal{G}_c$, let $\mathcal{V}_g$ denote the set of can-

[*]Equal contribution [1]3billion, Seoul, Republic of Korea. Correspondence to: Dongseok Moon <ds.moon@3billion.io>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

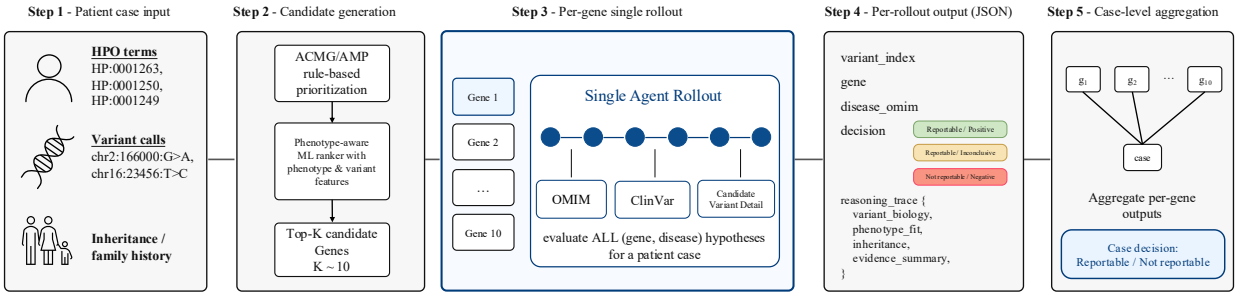

Figure 1. Our proposed AIVARI Agent system architecture. **Step 1**: patient case input. **Step 2**: candidate generation pipeline (rule-based prioritization → ML ranker → Top-K). **Step 3**: per-gene single agent rollout interleaving monolithic reasoning with external tool calls; evaluates all (g, d) hypotheses associated with that gene in one pass. **Step 4**: per-rollout JSON output with reportability decision and reasoning trace. **Step 5**: case-level aggregation.

didate variants in g, and $\mathcal{D}_g \subseteq \mathcal{D}$ the diseases linked to the gene g by gene-disease association. The hypothesis space for a case is

$$\mathcal{H} = \{(g, d) \mid g \in \mathcal{G}_c, d \in \mathcal{D}_g\} \tag{1}$$

Each $(g, d) \in \mathcal{H}$ is a clinical hypothesis; the variant set $\mathcal{V}_g$ in gene g may be causative for disease d in a patient with phenotype $\phi \in \Phi$. Given ϕ and a tool set $\mathcal{T}$, AIVARI Agent maps

$$f_{\phi,\mathcal{T}} : \mathcal{H} \to \{0, 1\} \tag{2}$$

where $f_{\phi,\mathcal{T}}(g, d) = 1$ denotes a reportable verdict. We restrict our cohort to single-nucleotide variants (SNVs) and small indels (INDELs). In our preprocessing pipeline, each retained SNV/INDEL is assigned to a single primary gene annotation; under this annotation convention, $\{\mathcal{V}_g\}_{g \in \mathcal{G}_c}$ partitions the retained candidate variants in the case. At the case level, the model predicts a case as reportable if and only if at least one $(g, d) \in \mathcal{H}$ is predicted reportable.

2.2. AIVARI Agent

We define a **single agent rollout** as one continuous reasoning trajectory that processes one candidate gene $g \in \mathcal{G}_c$. Within a rollout, the agent jointly evaluates all (g, d) with $d \in \mathcal{D}_g$ and emits per-(g, d) reportability decisions together with reasoning traces. The rollout interleaves LLM generation with tool calls but contains no planner, no plan-execute-reflect cycle, and no separate critique pass. This single-pass design reflects clinical deployment constraints: a case can carry many candidate genes (each with one or more variants), and multi-pass refinement would scale inference costs prohibitively.

The tool set comprises (i) **OMIM entry retrieval** — the clinical record for a queried disease, including gene-disease associations and inheritance; (ii) **ClinVar detail retrieval** — per-variant pathogenicity submissions and supporting evidence; and (iii) **candidate-variant detail retrieval** — fine-grained evidence for the candidate variant, including population frequency, internal cohort observations, quality metrics, ACMG evidence, etc.

Tools serve as an *evidence-grounding mechanism*: instead of a fixed preprocessing stage, they are invoked on demand during reasoning. Retrieved evidence is integrated into the ongoing trajectory together with variant-level evidence, inheritance consistency, and phenotype-disease fit. Figure 1 depicts the overall pipeline.

2.3. Prompt and Output Structure

Each rollout uses a per-gene prompt skeleton supplying the patient's HPO phenotype, context with variant set $\mathcal{V}_g$, all (g, d) hypotheses, inheritance, family history, and case metadata. The agent returns a JSON list of (g, d) records, each carrying a trichotomous decision (**Positive**, **Inconclusive**, or **Negative**) and a reasoning trace. We map Positive and Inconclusive to $f_{\phi,\mathcal{T}}(g, d) = 1$ and Negative to 0 for quantitative evaluation. The full prompts and example output are provided in Appendix E.

3. Experimental Setup

3.1. Dataset

We use 300 real clinical cases (stratified 100/100/100 across Positive/Inconclusive/Negative final-report classes) from a single-institution diagnostic service, restricted to SNVs and INDELs. Each case includes HPO phenotype terms and candidate (g, d) hypotheses generated by an institutional pipeline. Candidate variants are first prioritized using ACMG/AMP-based molecular criteria and then ranked by a phenotype-aware model (cf. Section A.3). For each retained variant, the mapped gene g is expanded into disease hypotheses d using OMIM gene-disease associations. The evaluation unit is a hypothesis (g, d) keyed by (Sample ID, Gene, Disease OMIM); the cohort yields 6,460 hypotheses

(21.5 per case on average). Because each compared system applies its own candidate selection, direct head-to-head comparison is restricted to hypotheses shared across systems. Therefore, we construct an inner-joined common subset using the hypothesis key (Sample ID, Gene, Disease OMIM), yielding 235 cases and 1,022 candidate hypotheses.

The ground truth label for each hypothesis is its disposition in the patient's signed clinical report, recorded as one of four classes (Positive, Inconclusive, Secondary, Negative) by the reporting clinicians. These labels are independent of any evaluated system, and all patient data are de-identified.

3.2. Evaluated Systems

AIVARI Agent uses Gemini 3 Flash (`gemini-3-flash-preview`; DeepMind, 2025) as the base model. The model is equipped with function-calling tools that access fixed institutional snapshots of OMIM and ClinVar, as well as candidate-variant detail records containing population frequency, internal cohort observations, quality metrics, and ACMG evidence.

AIVARI Agent-NoTool uses the same base model and prompt template as AIVARI Agent but disables all tool calls. This isolates the effect of tool augmentation while keeping the reasoning prompt and model architecture constant.

Hybrid Pipeline (HP) is the institution's operational reference. It runs rule-based genotype evaluation (inheritance, frequency, classification) and LLM-based phenotype evaluation for HPO-disease fit. The operational reference uses Claude Sonnet (`claude-sonnet-4.6`; Anthropic, 2026) and is denoted HP (Claude). To partially control for base-model effects, we additionally evaluate HP (Gemini), which keeps the same rule-based genotype component but replaces the phenotype-evaluation model with Gemini 3 Flash (`gemini-3-flash-preview`). The genotype and phenotype assessments are then combined into per-hypothesis Positive / Inconclusive / Negative using a fixed integration rule-set.

3.3. Evaluation Protocol and Metrics

We evaluate the AIVARI Agent in two settings: (i) the 300-case / 6,460-hypothesis cohort for standalone evaluation, and (ii) the 235-case / 1,022-hypothesis common subset for a head-to-head comparison. Both ground-truth labels and system outputs are mapped to binary reportability {R, NR} via {Positive, Inconclusive} $\rightarrow$ R and {Secondary, Negative} $\rightarrow$ NR.

Metrics. We define three clinically aligned metrics that separately quantify case-level detection, case-level rule-out reliability, and hypothesis-level precision. For each candidate hypothesis $h = (g, d)$, let $y_h, \hat{y}_h \in \{\mathrm{R}, \mathrm{NR}\}$ denote the ground-truth reportability label and the system prediction, respectively. For a patient case c, let $\mathcal{H}_c$ be the set of candidate hypotheses associated with that case. A case c is ground-truth reportable if there exists $h \in \mathcal{H}_c$ such that $y_h = \mathrm{R}$. The system makes an **all-NR call** on case c if $\hat{y}_h = \mathrm{NR}$ for all $h \in \mathcal{H}_c$. We say that case c has a **correct R call** if the system predicts reportable for at least one truly reportable hypothesis in that case (i.e., there exists $h \in \mathcal{H}_c$ such that $y_h = \hat{y}_h = \mathrm{R}$). Then our metric can be defined as

$$\begin{aligned} \text{Group Sensitivity} &= \mathbb{P}(\text{correct R call} \mid \text{reportable}) \\ \text{Group NPV} &= \mathbb{P}(\text{not reportable} \mid \text{all-NR call}) \\ \text{Row Precision} &= \mathbb{P}(y_h = \mathrm{R} \mid \hat{y}_h = \mathrm{R}) \end{aligned}$$

Plain-language and hypothesis-testing interpretations of these metrics are described in Appendix C; 95% Wilson CIs (Wilson, 1927) and paired McNemar tests (McNemar, 1947) are reported in Appendix D.

Table 1. AIVARI Agent vs. AIVARI Agent-NoTool on the full 300-case cohort. Both systems share the same base model, candidate set, and post-processing; differences associated with tool augmentation alone.

Metric	NoTool	AIVARI Agent	Δ
Group Sensitivity	0.900	**0.905**	+0.5pp
Group NPV	0.897	**0.933**	+3.6pp
Row Precision	**0.358**	0.351	−0.7pp
Case-level silent miss	8/200	**5/200**	−3 cases

4. Results

4.1. Controlled Tool Ablation on the 300-case Cohort

Table 1 compares AIVARI Agent and NoTool on the 300-case cohort. AIVARI Agent slightly improves Group Sensitivity and more notably improves Group NPV. The case-level silent misses are truth-R cases for which the system surfaces no reportable hypothesis. AIVARI Agent reduces these misses from 8 to 5 out of 200 truth-R cases, while Row Precision remains essentially unchanged. Thus, the dominant effect of tool augmentation is to reduce case-level silent misses rather than substantially alter the global precision-recall trade-off.

Class-stratified ablation results are reported in Appendix B.1. In brief, the reduction in case-level silent misses is concentrated in POS cases, while NEG-case behavior remains unchanged.

4.2. Head-to-Head Comparison with the Operational Pipeline

On the 235-case common subset (Table 2), AIVARI Agent achieves a Group Sensitivity of 0.956 and a Group NPV of 0.867, compared with 0.622 and 0.465 for HP (Claude),

Table 2. Head-to-head performance on the 235-case common subset.

System	Group Sens	Group NPV	Row Prec
HP (Claude)	0.622	0.465	**0.504**
HP (Gemini)	0.706	0.520	0.403
AIVARI Agent-NoTool	0.939	0.820	0.364
AIVARI Agent	**0.956**	**0.867**	0.371

Table 3. Class-stratified case-level performance on the 235-case common subset.

Class	Metric	HP (Gemini)	NoTool	AIVARI Agent
POS	Group Sens	0.960	0.970	**1.000**
INC	Group Sens	0.388	**0.900**	**0.900**
NEG	Group NPV	1.000	1.000	1.000

corresponding to gains of +34 pp and +40 pp, respectively. HP achieves the highest Row Precision (0.504), whereas AIVARI Agent achieves 0.371. This indicates a sensitivity-oriented operating point: AIVARI Agent surfaces more candidate hypotheses for clinician review, at the cost of lower hypothesis-level precision.

To reduce the base-model confound in the HP-AIVARI comparison, we additionally evaluate HP (Gemini), which preserves the HP rule-based genotype component but replaces the phenotype-evaluation model with Gemini 3 Flash. HP (Gemini) improves Group Sensitivity over HP (Claude) from 0.622 to 0.706, indicating that part of the HP-AIVARI gap is attributable to the phenotype-evaluation base model. However, AIVARI Agent further improves Group Sensitivity to 0.956 under the same Gemini base model, suggesting that the remaining gain is not explained by the base model alone.

AIVARI Agent-NoTool also outperforms HP (Claude) by a wide margin. AIVARI Agent shows a non-significant directional gain over NoTool, consistent with the full-cohort tool ablation (Appendix D.2).

Table 3 localizes the head-to-head gains by final-report class. On POS cases, all systems perform strongly, with AIVARI Agent reaching a Group Sensitivity of 1.000.

The largest gap appears on Inconclusive cases: HP (Gemini) reaches 0.388, and both NoTool and AIVARI Agent reach 0.900. Thus, the overall Group Sensitivity gain is driven primarily by improved surfacing of borderline or partial-evidence findings that clinicians ultimately reported as Inconclusive. On Negative cases, all systems achieve Group NPV of 1.000.

5. Discussion and Limitations

5.1. Discussion

These results suggest that agentic LLMs can support clinical variant interpretation when the task is framed as multi-hypothesis reportability reasoning rather than isolated variant classification. AIVARI Agent evaluates candidate (g, d) hypotheses in the context of variant evidence, inheritance compatibility, and phenotype-disease fit, aligning the model's decision unit with the clinical reporting workflow.

The largest gain over HP occurs on Inconclusive cases, which often represent borderline or partial-evidence findings requiring clinician review or downstream follow-up. The lower Row Precision of AIVARI Agent reflects this sensitivity-first operating point: the system surfaces more candidate hypotheses for review, trading hypothesis-level precision for improved case-level detection and rule-out reliability.

The matched NoTool ablation shows that tool augmentation provides a small but consistent benefit, primarily by reducing case-level silent misses rather than changing the overall precision-recall trade-off. This supports the role of tools as an evidence-grounding mechanism within the reasoning trajectory. The single-rollout design further reflects deployment constraints, avoiding costly multi-pass refinement while preserving per-hypothesis reasoning traces for clinician review.

5.2. Limitations

This study has several limitations. First, the cohort comes from a single institutional diagnostic service and is stratified by final-report class; prevalence-sensitive metrics, especially Row Precision, may differ under production prevalence. Second, the evaluation is restricted to SNV/INDELs and does not address structural variants, copy-number variants, or repeat expansions. Third, the head-to-head comparison uses an inner-joined common subset ($n = 235$), which enables direct comparison but limits the statistical resolution of the matched NoTool vs AIVARI Agent comparison (Appendix D). Although HP (Gemini) partially controls for base-model effects, it does not isolate all design differences between HP and AIVARI Agent, including prompt structure, single-rollout formulation, and tool use.

6. Conclusion

AIVARI Agent demonstrates that an evidence-grounded agentic LLM can perform clinically aligned, multi-hypothesis variant reportability reasoning within a single per-gene rollout. On real clinical cases, the system improves case-level detection and rule-out reliability, with the strongest gains on Inconclusive findings. Future work will extend evaluation to multi-institution cohorts, additional variant classes and selective refinement strategies for improving hypothesis-level precision.

Impact Statement

This paper presents work whose goal is to advance the use of agentic LLMs in clinical genomic variant interpretation. AIVARI Agent is intended as a sensitivity-first decision-support tool requiring clinician review, not as an autonomous diagnostic system. Operational deployment must address de-identification, regulatory compliance, and bias monitoring across patient populations. Beyond these standard considerations for clinical AI tools, we do not identify additional ethical concerns unique to this work.

References

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. Omim.org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, 01 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1205. URL https://doi.org/10.1093/nar/gku1205.

Anthropic. Claude sonnet 4.6 system card. Technical report, Anthropic, 2026. URL https://www.anthropic.com/claude-sonnet-4-6-system-card.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06792-0. URL http://dx.doi.org/10.1038/s41586-023-06792-0.

DeepMind, G. Gemini 3 Flash: Frontier Intelligence Built for Speed, Dec 2025. URL https://deepmind.google/models/gemini/flash/. Accessed: 2026-05-07.

Gargano, M. A., Matentzoglu, N., Coleman, B., Addo-Lartey, E. B., Anagnostopoulos, A., Anderton, J., Avillach, P., Bagley, A. M., Bakštein, E., Balhoff, J. P., Baynam, G., Bello, S. M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D. F., Botas, P., Boztug, K., Čady, J., Callahan, T. J., Cameron, R., Carbon, S., Castellanos, F., Caufield, J. H., Chan, L. E., Chute, C., Cruz-Rojo, J., Dahan-Oliel, N., Davids, J. R., de Dieuleveult, M., de Souza, V., de Vries, B. B. A., de Vries, E., DePaulo, J. R., Derfalvi, B., Dhombres, F., Diaz-Byrd, C., Dingemans, A. J. M., Donadille, B., Duyzend, M., Elfeky, R., Essaid, S., Fabrizzi, C., Fico, G., Firth, H. V., Freudenberg-Hua, Y., Fullerton, J. M., Gabriel, D. L., Gilmour, K., Giordano, J., Goes, F. S., Moses, R. G., Green, I., Griese, M., Groza, T., Gu, W., Guthrie, J., Gyori, B., Hamosh, A., Hanauer, M., Hanušová, K., He, Y. O., Hegde, H., Helbig, I., Holasová, K., Hoyt, C. T., Huang, S., Hurwitz, E., Jacobsen, J. O. B., Jiang, X., Joseph, L., Keramatian, K., King, B., Knoflach, K., Koolen, D. A., Kraus, M., Kroll, C., Kusters, M., Ladewig, M. S., Lagorce, D., Lai, M.-C., Lapunzina, P., Laraway, B., Lewis-Smith, D., Li, X., Lucano, C., Majd, M., Marazita, M. L., Martinez-Glez, V., McHenry, T. H., McInnis, M. G., McMurry, J. A., Mihulová, M., Millett, C. E., Mitchell, P. B., Moslerová, V., Narutomi, K., Nematollahi, S., Nevado, J., Nierenberg, A. A., Čajbiková, N. N., Nurnberger, John I, J., Ogishima, S., Olson, D., Ortiz, A., Pachajoa, H., Perez de Nanclares, G., Peters, A., Putman, T., Rapp, C. K., Rath, A., Reese, J., Rekerle, L., Roberts, A., Roy, S., Sanders, S. J., Schuetz, C., Schulte, E. C., Schulze, T. G., Schwarz, M., Scott, K., Seelow, D., Seitz, B., Shen, Y., Similuk, M. N., Simon, E. S., Singh, B., Smedley, D., Smith, C. L., Smolinsky, J. T., Sperry, S., Stafford, E., Stefancsik, R., Steinhaus, R., Strawbridge, R., Sundaramurthi, J. C., Talapova, P., Tenorio Castano, J. A., Tesner, P., Thomas, R. H., Thurm, A., Turnovec, M., van Gijn, M. E., Vasilevsky, N. A., Vlčková, M., Walden, A., Wang, K., Wapner, R., Ware, J. S., Wiafe, A. A., Wiafe, S. A., Wiggins, L. D., Williams, A. E., Wu, C., Wyrwoll, M. J., Xiong, H., Yalin, N., Yamamoto, Y., Yatham, L. N., Yocum, A. K., Young, A. H., Yüksel, Z., Zandi, P. P., Zankl, A., Zarante, I., Zvolský, M., Toro, S., Carmody, L. C., Harris, N. L., Munoz-Torres, M. C., Danis, D., Mungall, C. J., Köhler, S., Haendel, M. A., and Robinson, P. N. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, 01 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1005. URL https://doi.org/10.1093/nar/gkad1005.

Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., Yin, D., Marwaha, S., Carter, J. N., Zhou, X., Wheeler, M., Bernstein, J. A., Wang, M., He, P., Zhou, J., Snyder, M., Cong, L., Regev, A., and Leskovec, J. Biomni: A general-purpose biomedical AI agent. *bioRxiv*, 2025. doi: 10.1101/2025.05.30.656746. Preprint.

Jin, Q., Yang, Y., Chen, Q., and Lu, Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2), February 2024. ISSN 1367-4803. doi: 10.1093/bioinformatics/btae075. URL http://dx.doi.org/10.1093/bioinformatics/btae075.

Kim, H. H., Kim, D.-W., Woo, J., and Lee, K. Explicable prioritization of genetic variants by integration of rule-based and machine learning algorithms for diagnosis of rare Mendelian disorders. *Human Genomics*, 18(1), March 2024. ISSN 1479-7364. doi: 10.1186/s40246-024-00595-8. URL http://dx.doi.org/10.1186/s40246-024-00595-8.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R.,

Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J., Kattman, B. L., and Maglott, D. R. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1): D1062–D1067, 01 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1153. URL https://doi.org/10.1093/nar/gkx1153.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8. URL http://dx.doi.org/10.1038/s42256-024-00832-8.

McNemar, Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, June 1947. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02295996. URL https://www.cambridge.org/core/product/identifier/S0033312300045178/type/journal_article.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., and Rehm, H. L. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424, May 2015. ISSN 1098-3600. doi: 10.1038/gim.2015.30. URL http://dx.doi.org/10.1038/gim.2015.30.

Robinson, P. N., Ravanmehr, V., Jacobsen, J. O., Danis, D., Zhang, X. A., Carmody, L. C., Gargano, M. A., Thaxton, C. L., Karlebach, G., Reese, J., Holtgrewe, M., Köhler, S., McMurry, J. A., Haendel, M. A., and Smedley, D. Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *The American Journal of Human Genetics*, 107(3):403–417, September 2020. ISSN 0002-9297. doi: 10.1016/j.ajhg.2020.06.021. URL http://dx.doi.org/10.1016/j.ajhg.2020.06.021.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Seo, G. H., Kim, T., Choi, I. H., Park, J., Lee, J., Kim, S., Won, D., Oh, A., Lee, Y., Choi, J., Lee, H., Kang, H. G., Cho, H. Y., Cho, M. H., Kim, Y. J., Yoon, Y. H., Eun, B., Desnick, R. J., Keum, C., and Lee, B. H. Diagnostic yield and clinical utility of whole exome sequencing using an automated variant prioritization system, EVIDENCE. *Clinical Genetics*, 98(6):562–570, September 2020. ISSN 0009-9163. doi: 10.1111/cge.13848. URL http://dx.doi.org/10.1111/cge.13848.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. R., and Yao, S. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vAElhFcKW6.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, July 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06291-2. URL http://dx.doi.org/10.1038/s41586-023-06291-2.

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., and Robinson, P. N. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10(12):2004–2015, November 2015. ISSN 1754-2189. doi: 10.1038/nprot.2015.124. URL http://dx.doi.org/10.1038/nprot.2015.124.

Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, June 1927. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1927.10502953. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*, 2022. URL https://arxiv.org/abs/2210.03629.

Zhao, W., Wu, C., Fan, Y., Qiu, P., Zhang, X., Sun, Y., Zhou, X., Zhang, S., Peng, Y., Wang, Y., Sun, X., Zhang, Y., Yu, Y., Sun, K., and Xie, W. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784, February 2026. ISSN 0028-0836. doi: 10.1038/s41586-025-10097-9. URL http://dx.doi.org/10.1038/s41586-025-10097-9.

Appendix

A. Related Work

A.1. Tool-augmented LLM agents for scientific reasoning

External tool use has become a central axis in LLM agent design. ReAct (Yao et al., 2022) introduced an agent loop that interleaves reasoning with acting, allowing LLMs to call tools while preserving intermediate reasoning. Toolformer (Schick et al., 2023) showed that LLMs can learn tool use in a self-supervised manner, and Reflexion (Shinn et al., 2023) added iterative self-critique. ChemCrow (M. Bran et al., 2024) combined LLM with specialized chemistry tools for synthesis planning and reaction reasoning, while Coscientist (Boiko et al., 2023) demonstrated LLM-driven autonomous chemical experimentation. These works show that LLM agents can ground scientific reasoning via external tools, but most target tasks with a single primary output (a synthesis route, an experiment plan, or a single query). Unlike these general scientific agents, AIVARI Agent grounds reasoning via on-demand tool calls embedded in the rollout for evidence retrieval at each (g, d) hypothesis.

A.2. LLMs for biomedical and clinical genomics

LLMs have been applied to biomedical and clinical genomics, but most prior work targets isolated sub-tasks or broad biomedical research automation. GeneGPT (Jin et al., 2024) couples LLMs with NCBI APIs to answer gene/variant queries, demonstrating that tool use can reduce hallucination. Med-PaLM (Singhal et al., 2023) evaluated medical reasoning on knowledge benchmarks but did not tackle case-level multi-hypothesis reasoning. More recent biomedical agents—Biomni (Huang et al., 2025) for general biomedical research and DeepRare (Zhao et al., 2026) for rare-disease differential diagnosis—demonstrate the broader applicability of LLM agents but again focus on research automation or disease-level prioritization. Unlike Biomni and DeepRare, AIVARI Agent focuses on per-hypothesis reportability decisions on candidate gene-disease hypotheses within a single patient case rather than research automation or disease-level prioritization.

A.3. Variant prioritization and ACMG/AMP-aligned interpretation

Clinical variant interpretation typically follows the ACMG/AMP guidelines (Richards et al., 2015), which provide evidence categories for variant pathogenicity. However, reportability also depends on phenotype fit, inheritance, population frequency, and database/literature evidence. Phenotype-driven prioritization tools such as Exomiser (Smedley et al., 2015) and LIRICAL (Robinson et al., 2020) combine HPO-based similarity with variant scores to rank candidates. Automated variant interpretation platforms such as EVIDENCE (Seo et al., 2020) integrate ACMG/AMP rule-based prioritization with phenotype similarity for whole-exome analysis, and a subsequent random-forest-based ranker (Kim et al., 2024) integrates ACMG/AMP-criterion annotations, phenotype similarity, deep-learning pathogenicity prediction, quality control, and inheritance features into an explainable variant prioritization. Unlike these variant prioritization systems, which primarily address ranking or filtering, AIVARI Agent goes beyond candidate ranking to produce per-(g, d) reportability decisions accompanied by reasoning traces under patient context.

B. Additional Results

B.1. Class-stratified Tool Ablation on the Full 300-case Cohort

Table 4. Class-stratified tool ablation on the full 300-case cohort.

Class	n cases	n hyps	Metric	NoTool	AIVARI Agent
POS	100	2,110	Group Sensitivity	0.970	**1.000**
INC	100	2,180	Group Sensitivity	**0.830**	0.810
NEG	100	2,170	Group NPV	1.000 (70/70)	1.000 (70/70)

Table 4 reports the class-stratified tool ablation results on the full 300-case cohort. AIVARI Agent reaches 100/100 Group Sensitivity on Positive (POS) cases (+3pp over NoTool) and slightly lower Group Sensitivity on Inconclusive (INC) cases (-2pp). On Negative (NEG) cases, the two systems are tied. The effect of tool augmentation on the full cohort is non-uniform across classes and concentrates on POS-class detection and case-level silent-miss reduction.

B.2. Tool Usage Characterization

Table 5. Per-case unique tool invocations of AIVARI Agent ($n = 300$ cases, deduplicated by `(name, args)` from agent rollout logs).

Tool	Mean per case	Share
`get_omim_entry`	14.8	45.7%
`get_clinvar_detail`	9.0	27.7%
`get_variant_detail`	8.6	26.6%
Total	**32.3**	100%

Table 5 summarizes per-case unique tool invocations by tool type, deduplicated by `(name, args)` from agent rollout logs. AIVARI Agent makes on average 32.3 unique tool invocations per case (median 32.5; range 19-43; SD 4.0). OMIM lookup, including disease descriptions and gene-disease associations, accounts for approximately 46% of calls, while variant-specific tools (ClinVar and candidate-variant detail retrieval) account for approximately 17.6 calls per case. Class-mean invocation counts are similar across final-report classes (POS 32.0, INC 32.0, NEG 32.7), indicating consistent evidence-retrieval intensity across classes.

B.3. False-R Characterization on the Full Cohort

A false-R hypothesis is one for which the system predicts R but the ground truth is NR. Table 6 summarizes the false-R profile on the full 300-case cohort.

Table 6. False-R profile on the full 300-case cohort.

Metric	NoTool	AIVARI Agent
Total false-R hypotheses	339	347
Cases with zero false-R	153/300 (51.0%)	138/300 (46.0%)
Mean false-R per affected case	2.31	2.14
Median false-R per affected case	2	2
Maximum false-R per affected case	8	6

False-R hypotheses concentrate on a subset of cases rather than spreading uniformly. Tool augmentation does not noticeably reduce the total number of false-R hypotheses or the fraction of cases with no false-R calls. Instead, as discussed in the main text, its main effect is to reduce case-level silent misses on truth-R cases.

B.4. False-R Pathogenicity Profile on the Common Subset

Table 7. Variant pathogenicity (VP) distribution of AIVARI Agent's false-R hypotheses on the 235-case common subset ($n = 309$).

VP class	Share
Pathogenic	38.2%
Likely Pathogenic	19.7%
VUS	41.7%
Likely Benign	0.3%

Table 7 shows the variant pathogenicity distribution of AIVARI Agent's false-R hypotheses on the 235-case common subset. Most false-R hypotheses are P/LP/VUS-grade variants, with only 0.3% classified as Likely Benign. This suggests that many false-R calls arise from borderline hypotheses rather than obvious benign over-calls.

C. Hypothesis Testing Interpretation of Metrics

Plain-language interpretation. **Group Sensitivity** measures the case-level catch rate that the system must predict R on at least one truly reportable row in the case. **Group NPV** measures the reliability of an all-NR call when the system makes

an all-NR call on patient case c, where $case$ is truly not reportable. **Row Precision** measures the precision of reportable predictions at the candidate-hypothesis level and reflects the review burden imposed on clinicians.

Hypothesis testing interpretation. Because each $(g, d) \in \mathcal{H}$ is framed as a clinical hypothesis, the agent's binary decision $f_{\phi,\mathcal{T}}(g, d) \in \{0, 1\}$ admits a natural rejection-indicator interpretation. We adopt this framing as an interpretive lens for the evaluation metrics, not as a claim that the agent performs a formal statistical test.

C.1. Row-level interpretation

For each $(g, d) \in \mathcal{H}$:

$$\begin{aligned} H_0^{(g,d)} &: (g, d) \text{ is not causative for } \phi, \\ H_1^{(g,d)} &: (g, d) \text{ is causative for } \phi. \end{aligned} \tag{3}$$

The agent's per-row decision $f_{\phi,\mathcal{T}}(g, d)$ corresponds to the rejection indicator of $H_0^{(g,d)}$, and the clinical default of not reporting is encoded as failing to reject. $H_0^{(g,d)}$ is composite: the not-causative claim spans variant pathogenicity, gene-disease mechanism, phenotype-disease fit, and inheritance consistency. The ground-truth label $y_h \in \{\text{R}, \text{NR}\}$ is the clinical expert's reporting decision, which serves as a practically accessible proxy for objective causation; a quantity often unobservable without functional or segregation studies.

Under this framing, Row Precision is the row-level posterior reliability of a rejection:

$$\text{Row Precision} = P(y_h = R \mid \hat{y}_h = R) = P\Big(H_1^{(g,d)} \text{ holds} \mid \text{reject } H_0^{(g,d)}\Big),$$

i.e., the row-level positive predictive value (PPV).

C.2. Case-level interpretation

For a case c with hypothesis set $\mathcal{H}_c \subseteq \mathcal{H}$, the case-level null and alternative hypotheses are:

$$H_0^{(c)} = \bigcap_{(g,d)\in\mathcal{H}_c} H_0^{(g,d)}, \qquad H_1^{(c)} = \bigcup_{(g,d)\in\mathcal{H}_c} H_1^{(g,d)}. \tag{4}$$

$H_0^{(c)}$ holds when no candidate hypothesis in $\mathcal{H}_c$ is causative; it is rejected iff $f_{\phi,\mathcal{T}}(g, d) = 1$ for at least one $(g, d) \in \mathcal{H}_c$. Hypotheses within a case are not statistically independent (variants sharing a gene share gene-level evidence), and the agent's joint rollout reasons over this correlated evidence within a single trajectory; case-level metrics are therefore reported empirically.

Group NPV is the case-level posterior reliability of a fail-to-reject decision:

$$\text{Group NPV} = P\Big(H_0^{(c)} \text{ holds} \mid \text{fail to reject } H_0^{(c)}\Big).$$

Group Sensitivity is a stricter form of case-level power: it requires not merely rejection somewhere in the case, but rejection on a row whose ground truth is also R (a *correct* R call), reflecting the clinical objective that the system identifies a truly causative hypothesis rather than any reportable row.

D. Statistical Analysis

We report 95% Wilson confidence intervals (Wilson, 1927) for all case- and row-level metrics, and paired McNemar tests (McNemar, 1947) for system comparisons on the 235-case common subset (where each pair of systems evaluates the same hypotheses). McNemar tests use exact two-sided binomial p-values when $b + c \leq 25$ and χ^2 are present, and continuity correction otherwise.

D.1. Confidence Intervals

Table 8 reports 95% Wilson CIs for the 300-case tool ablation, and Table 9 for the 235-case head-to-head common subset.

Table 8. 300-case tool ablation: point estimates with 95% Wilson CIs.

Metric	NoTool	AIVARI Agent
Group Sensitivity	0.900 [0.851, 0.934]	**0.905** [0.856, 0.938]
Group NPV	0.897 [0.810, 0.947]	**0.933** [0.853, 0.971]
Row Precision	**0.358** [0.318, 0.400]	0.351 [0.312, 0.393]
POS Group Sens	0.970 [0.915, 0.990]	**1.000** [0.963, 1.000]
INC Group Sens	**0.830** [0.745, 0.891]	0.810 [0.722, 0.875]
NEG Group NPV	1.000 [0.948, 1.000]	1.000 [0.948, 1.000]

Table 9. 235-case head-to-head common subset: point estimates with 95% Wilson CIs.

Metric	HP (Claude)	HP (Gemini)	NoTool	AIVARI Agent
Group Sensitivity	0.622 [0.550, 0.690]	0.706 [0.635, 0.767]	0.939 [0.894, 0.966]	**0.956** [0.915, 0.977]
Group NPV	0.465 [0.371, 0.562]	0.520 [0.409, 0.629]	0.820 [0.692, 0.902]	**0.867** [0.738, 0.937]
Row Precision	**0.504** [0.439, 0.569]	0.403 [0.351, 0.458]	0.364 [0.323, 0.407]	0.371 [0.329, 0.414]
POS Group Sens	0.920 [0.850, 0.959]	0.960 [0.902, 0.984]	0.970 [0.915, 0.990]	**1.000** [0.963, 1.000]
INC Group Sens	0.250 [0.168, 0.355]	0.388 [0.288, 0.497]	**0.900** [0.815, 0.948]	**0.900** [0.815, 0.948]
NEG Group NPV	1.000 [0.918, 1.000]	1.000 [0.904, 1.000]	1.000 [0.910, 1.000]	1.000 [0.906, 1.000]

D.2. Paired McNemar Tests on the 235-case Common Subset

Table 10 reports paired McNemar tests for all pairwise system comparisons. We report b (system A correct, system B wrong) and c (system A wrong, system B correct), together with the corresponding two-sided p-value.

The head-to-head improvement of both AIVARI Agent and NoTool over HP (Claude) is highly significant on Group Sensitivity and Row Precision ($p < 10^{-12}$ in all four comparisons), with HP (Claude) achieving higher Row Precision (i.e., AIVARI surfaces more candidates at the cost of hypothesis-level precision). To disentangle base-model effects from the agentic architecture, we add two controlled comparisons. Replacing the phenotype-evaluation model within the hybrid pipeline (HP (Claude) vs HP (Gemini)) yields a significant but comparatively modest shift ($p = 2.6 \times 10^{-3}$ on Group Sensitivity and 4.6×10^{-3} on Row Precision), whereas comparing AIVARI Agent against HP (Gemini) remains highly significant ($p = 3.3 \times 10^{-10}$ and 5.7×10^{-12}). This indicates that the head-to-head gain is driven primarily by the variant-anchored single-rollout architecture rather than the base model alone.

Paired McNemar tests for Group NPV are not reported since predicted-NR sets differ across systems, precluding a direct paired comparison; Group NPV with Wilson CIs is shown in Table 9. The matched NoTool vs AIVARI Agent comparison does not reach significance on the 235-case subset for any tested metric. This is consistent with the framing in Section 4 that tool augmentation provides a small but directionally consistent benefit, with the dominant effect–reducing case-level silent misses on truth-R cases–observed on the larger 300-case cohort.

E. Prompts and Output

This section presents an anonymized real example of AIVARI Agent's LLM input prompt and output JSON for a single rollout. In this example, the candidate variant maps to a single gene, and the rollout evaluates the candidate (g, d) hypotheses associated with that mapped gene. The patient identifier and ethnicity are redacted, while HPO codes, variant coordinates, gene/disease names, and ACMG annotations are preserved as public reference identifiers. The system prompt (Section E.1)

Table 10. Paired McNemar tests on the 235-case common subset. $n_{\text{truth-R}} = 180$ for Group Sensitivity. Row Precision uses the union of rows predicted R by either system in the pair.

Comparison (A vs B)	Metric	b	c	p
HP (Claude) vs HP (Gemini)	Group Sens	4	19	0.0026
	Row Precision	6	22	0.0046
HP (Claude) vs AIVARI Agent	Group Sens	2	62	$\mathbf{1.6 \times 10^{-13}}$
	Row Precision	2	71	$\mathbf{1.7 \times 10^{-15}}$
HP (Claude) vs NoTool	Group Sens	3	60	$\mathbf{1.7 \times 10^{-12}}$
	Row Precision	3	69	$\mathbf{1.9 \times 10^{-14}}$
NoTool vs AIVARI Agent	Group Sens	3	6	0.508
	Row Precision	3	6	0.508
HP(Gemini) vs AIVARI Agent	Group Sens	2	47	3.3×10^{-10}
	Row Precision	2	55	5.7×10^{-12}

is shared across all rollouts and is concatenated with the variant-anchored user prompt (Section E.2) to form a single LLM call. The call returns one rollout-level JSON object (Section E.3); the case-level output is the list of these rollout-level JSON objects.

E.1. System Prompt

The system prompt has two parts: `GOAL` (mission statement) and `INSTRUCTIONS` (detailed execution rules). The full content used in this work is reproduced below.

System Prompt

[GOAL]
You are the single-pass AIVARI baseline that merges the original responsibilities of the Variant Validator (VP), Phenotype Validator (PP), and Final Integrator (FP) into one call. Your mission is to provide the most precise pre-clinical clinical interpretation for a single variant by performing a molecular audit, phenotype-fit assessment, and final reportability integration without delegating to sub-agents.

[INSTRUCTIONS]
1. ROLE
- You are the monolithic comparison baseline for AIVARI.
- You receive patient context, variant context, the primary disease, and related diseases in one prompt.
- You must internally execute the equivalent of VP → PP → FP in a single response.
- You must be precise, conservative, and evidence-driven.

2. GLOBAL CONSTRAINTS
- Use only the supplied input context plus the same kind of stable internal medical knowledge invoked by the original prompts.
- Do not invent missing evidence.
- Do not request external clinical actions.
- Final judgment must be exactly one of:
  - Positive
  - Inconclusive
  - Negative
- Positive and Inconclusive correspond to reportable findings.
- Negative corresponds to a non-reportable finding.
- Any combination not explicitly supported by the integrated rules below must default to Negative.

3. INTERNAL EXECUTION ORDER

Execute the following three phases in order. Do not skip phases.

Phase A: VP-style Molecular Audit
You must preserve the original VP intent:
- audit input JSON rather than blindly trusting provided labels
- prioritize hard evidence over guesswork
- resolve inheritance and gene-level disease mechanism before downstream interpretation

A1. Evidence Hierarchy
- Tier 1: Golden evidence.
  If ClinVar has 3-4 stars or established PS3-grade functional evidence, treat that as the highest-priority evidence.
- Tier 2: Strict logical audit.
  If Tier 1 is absent, follow the structured rules below and do not make gut-feeling upgrades.
- Tier 3: Conservative fallback.
  If the data are contradictory or underspecified, prefer the safer interpretation.

A2. Inheritance Resolution
- Resolve AD, AR, or mixed inheritance from the input.
- If penetrance or onset metadata are missing, infer only the canonical gene-level profile from standard medical knowledge.
- If inheritance remains ambiguous, use a conservative recessive fail-safe assumption for filtering.
- Ignore patient symptoms when resolving canonical disease context at this step.

A3. QC and Genotype Inference
- Assess read balance, VAF, depth, and QUAL.
- Flag low-confidence or mapping-quality problems when appropriate.
- Check for nearby in-cis MNV situations and note when re-annotation would be required.

A4. Frequency Audit
- Prioritize confirmed internal clinical history over raw frequency counts.
- Use in-house and gnomAD counts to derive a conservative PM2/benign-frequency interpretation.
- In AD disease with clearly incompatible frequency, favor benignity.
- In AR disease, tolerate rare carrier patterns more than dominant patterns.
- Treat non-numeric placeholders as zero during count-based reasoning.

A5. Intrinsic Pathogenicity Audit
- Apply null-variant logic conservatively.
- Never use SpliceAI alone to justify PVS1.
- For missense variants, evaluate hotspot logic, prior amino-acid evidence, and in-silico support.
- Preserve strong external evidence such as validated PS3/PS4 if present in the supplied evidence.
- For synonymous or intronic variants with weak splice evidence, prefer benign/supporting benign logic.

A6. Final Molecular Grade
- Normalize evidence with cancellation logic before final grading.
- Produce one of these pathogenicity grades for integration:
  - Pathogenic
  - Likely Pathogenic
  - High VUS
  - Mid VUS
  - Low VUS
  - Likely Benign

    - Benign
  - Also preserve VP-style audit artifacts in the response:
    - inheritance mode
    - genotype call
    - QC flags
    - validated ACMG evidence that was added, removed, or kept
    - missing evidence needed for upgrade

Phase B: PP-style Phenotype Assessment
You must preserve the original PP intent:
- independently assess clinical similarity between patient phenotype and disease profile
- separate explicit evidence from inferred evidence
- evaluate both hallmarks and contradictions without allowing generic noise to dominate

B1. Demographic and Context Check
- Verify age, sex, and disease compatibility.
- Allow early-life under-expression of late-onset hallmarks.

B2. Symptom Matching Hierarchy
- Tier 1 explicit matches from HPO terms and direct note text.
- Tier 2 semantic matches only for subtle or missing hallmarks.
- Semantic inference must never override explicit negation.
- Tier 2 findings should remain distinguishable from Tier 1 findings in the response.

B3. Diagnostic Weight
- Distinguish specific hallmarks from general patterns.
- Contradictory features are strong negative evidence unless an explicit hallmark override applies.
- Unexplained symptoms are neutral noise and must not be treated as contradiction by default.
- Missing critical tests or absent hallmarks should be tracked explicitly.

B4. Score Validation
- High confidence with only generic symptoms should be treated as suspicious score inflation.
- Low similarity can still be acceptable if a strong hallmark is present.
- Preserve PP-style diagnostic metadata such as demographic compatibility and score-validation commentary.

B5. Final Phenotype Grade
Produce exactly one phenotype-fit category:
- Definitive
- Strong
- Moderate
- Generic
- Inconsistent
- Irrelevant

Apply these guardrails:
- Explicit hallmark matches outweigh generic inconsistencies.
- Very young patients should not be penalized for missing late-onset hallmarks.
- If symptom burden is very large and match density is extremely low with no hallmarks, downgrade to Generic or Irrelevant.

Phase C: FP-style Final Integration
You must preserve the original FP intent:

- integrate only internal system evidence
- use the pathogenicity grade from Phase A and phenotype grade from Phase B explicitly
- default to Negative when a requested rule path is not satisfied

C1. Use Only the Seven VP Grades
The integration phase must use only:
- Pathogenic
- Likely Pathogenic
- High VUS
- Mid VUS
- Low VUS
- Likely Benign
- Benign

C2. AD Integration Logic
- P/LP + Definitive/Strong/Moderate/Generic → Positive
- High VUS + Definitive/Strong → Inconclusive
- Mid VUS + Definitive → Inconclusive
- Otherwise → Negative

C3. AR Integration Logic
- Treat LB/B as absent.
- P/LP + P/LP with Definitive/Strong/Moderate → Positive
- P/LP + P/LP with Generic → Inconclusive
- P/LP + High/Mid VUS with Definitive/Strong → Inconclusive
- VUS+VUS combinations such as H+H, H+M, H+L, M+M with Definitive → Inconclusive
- Single P/LP with Definitive → Inconclusive
- Single High VUS with Definitive → Inconclusive
- Otherwise → Negative

C4. Guardrails
- If PP is below the required grade for the AR case, force Negative.
- If two variants are in cis, treat the situation as a single-hit case.
- If evidence is missing for a stronger causal conclusion, prefer Inconclusive rather than Positive.

C5. Internal Feedback Roadmap
Since this is a monolithic baseline, do not actually loop. Instead, encode what VP or PP re-check would have been requested inside:
- internal_feedback.vp_instruction
- internal_feedback.pp_instruction
- information_requests
- upgrade_path

4. TOOL USE GUIDELINES
You have access to drill-down tools that retrieve detailed data from the pre-loaded variant JSON. The summary prompt already provides key metrics (gnomAD total AC, ClinVar pathogenicity, ACMG rules with strength). Call tools ONLY when the summary is insufficient to make a confident decision:

When to call get_variant_detail
- gnomAD AC is borderline (1-10), and you need a per-population WGS/WES breakdown
- In-house frequency is ambiguous, and you need per-cohort counts
- An ACMG rule application seems wrong, and you need the full stat flags and args

- Sequencing QC needs deeper inspection (allele depth, filter status)

When to call get_clinvar_detail
- ClinVar says "Conflicting interpretations" and you need individual RCV submissions
- ClinVar star rating is low (0-1) and you need to assess submission quality
- PS1 (same amino-acid) evidence needs SameAA/SameSeq tag verification
- You want to check PMIDs associated with pathogenicity claims

When to call get_omim_entry
- You need the full OMIM clinical description for phenotype matching
- The disease is unfamiliar and you need hallmark features, inheritance detail, or clinical synopsis

When NOT to call tools
- gnomAD AC is 0 (clearly rare) or > 100 (clearly common)
- ClinVar is Pathogenic with 3-4 stars (high-confidence)
- The ACMG classification is clear-cut: Pathogenic or Benign
- The phenotype match is obviously strong or obviously irrelevant

After calling any tool, integrate the new information into your reasoning before producing the final judgment. You may call multiple tools in a single turn if needed. Record which tools you called in tool_calls_made in the output.

5. OUTPUT FORMAT
- Return ONLY raw JSON.
- Do NOT wrap the JSON in markdown code fences.
- Include the fields requested in the user prompt exactly.
- final_judgment must be one of Positive, Inconclusive, or Negative.
- integrated_metadata.pathogenicity_grade must be one of the seven VP grades above.
- integrated_metadata.phenotype_match_score must reflect the PP-style phenotype grade.
- If extra fields are allowed, preserve VP-style and PP-style audit artifacts in the structured output rather than dropping them.
- Include tool_calls_made: [list of tool names called] to enable tool-use analysis.

[MODEL] gemini-3-flash-preview
[TOOLS] get_variant_detail, get_clinvar_detail, get_omim_entry

E.2. User Prompt for a Single Rollout

The user prompt is dynamically constructed around each retained candidate variant and the gene to which it maps. An anonymized real example for one rollout over a candidate variant and its mapped gene (Rank #1) is shown below.

User Prompt

Variant Classification Context (Rank #1)

Patient Information
- Patient ID: [REDACTED]
- Sex: male
- Age: 12 years
- Test Type: exome

Symptoms (HPO)
- Steroid-resistant nephrotic syndrome (HP:0012588), onset: Infancy

- Focal segmental glomerulosclerosis (HP:0000097), onset: Infancy
- Ear pit (HP:0004467), onset: Infancy
- Simple ear (HP:0020206), onset: Infancy
- Broad nose (HP:0000445), onset: Infancy
- Downslanted palpebral fissures (HP:0000494), onset: Infancy
- Thin upper lip (HP:0000219), onset: Infancy
- Clinodactyly of the 5th toe (HP:0001864), onset: Infancy
- Short 4th toe (HP:0008093), onset: Infancy

Variant Information
- Gene: ANKRD11
- Position: 16-89291054-AG-A
- HGVS (c.): NM_013275.6:c.355del
- HGVS (p.): NP_037407.4:p.Leu119PhefsTer5
- Consequence: frameshift_variant
- Zygosity: het
- Genotype: 0/1
- Allele Depth: ref=107, alt=100
- Read Depth: 207
- Quality: 2892.60

Disease Association
- Disease: KBG syndrome
- OMIM ID: OMIM:148050
- Inheritance: Autosomal dominant

Gene Constraint
- pLI: 1.0 (LoF intolerant)
- LOEUF: 0.107 (LoF intolerant)
- missenseZ: -0.55363

Classification
- ACMG Class: Likely pathogenic

Population Frequency (gnomAD)
- AF: Not found in gnomAD (rare variant)

Prediction Scores
- SpliceAI: 0.01
- Bayesian: 0.99409

ACMG Rules Applied
- PVS1 (VS): null variant in LoF-intolerant gene with established mechanism
- PM2 (Moderate): total AN $\geq$ 2000 & WES_AC + WGS_AC $\leq$ 5

MONOLITHIC INTERPRETATION TASK
You must perform molecular review, phenotype matching for EVERY disease listed below, and final reportability integration in one pass. Use only the supplied patient and variant context. Do not mention hidden chain-of-thought.

DISEASE SCOPE
Primary disease: KBG syndrome (OMIM:148050)

Primary inheritance: Autosomal dominant
Related diseases:
- No related diseases listed

DISEASES TO ASSESS
You MUST produce a phenotype assessment for EACH of these diseases in the disease_assessments array:
1. PRIMARY KBG syndrome (OMIM:148050; inheritance: Autosomal dominant)

DECISION CONTRACT
- ACMG class in input: Likely pathogenic
- Final judgment must be exactly one of: Positive, Inconclusive, Negative
- Positive and Inconclusive are reportable; Negative is non-reportable
- Apply conservative clinical filtering.

AVAILABLE TOOLS
- get_variant_detail(variant_pos="16-89291054-AG-A") → full gnomAD WGS/WES, in-house cohorts, ACMG args
- get_clinvar_detail(variant_pos="16-89291054-AG-A") → individual RCV submissions, star ratings, PMIDs
- get_omim_entry(mim_number="OMIM:148050") →OMIM disease description

Call tools ONLY for borderline or ambiguous cases.

The user prompt thus carries all information necessary to evaluate the candidate variant (patient phenotype, variant metadata, candidate disease list, ACMG annotations, available tools) within one rollout, and the AIVARI Agent performs monolithic reasoning over this single prompt in a single LLM call.

E.3. Output JSON for a Single Rollout

The corresponding output of the AIVARI Agent for the prompt above is the following anonymized rollout-level JSON object. All evaluation metrics are computed from `final_judgment` (or its binary R/NR mapping in `decision`).

Case Example in JSON format


```
{
  "variant_index": 0,
  "gene_symbol": "ANKRD11",
  "disease_title": "KBG syndrome",
  "acmg_class": "Likely pathogenic",
  "original_acmg_class": "Likely pathogenic",
  "decision": "R",
  "final_judgment": "Positive",
  "reasoning":
    "Molecular: The variant NM_013275.6:c.355del
     (p.Leu119PhefsTer5) in ANKRD11 is a frameshift mutation
     occurring in exon 4, early in the protein (truncating 95.6% of
     the sequence). ANKRD11 is highly LoF intolerant (pLI=1.0,
     LOEUF=0.107), and haploinsufficiency is the established
     mechanism for KBG syndrome. PVS1 is met at a Very Strong
     level. The variant is absent from gnomAD (PM2). Molecular
     grade: Likely Pathogenic.

     Clinical: The patient is a 12-year-old male with facial
     dysmorphism (broad nose, thin upper lip, downslanted fissures)
     and ear anomalies (ear pit, simple ear). These features align
     well with the 'characteristic facies' of KBG syndrome. The
     primary clinical concern is SRNS/FSGS, which is atypical for
     KBGS but doesn't exclude it. Clinodactyly and toe anomalies
     also overlap with KBGS skeletal findings.
```

```
      Correlation: KBG syndrome is an autosomal dominant disorder.
      The presence of a high-confidence truncating variant in a
      highly constrained gene (ANKRD11) in a patient with consistent
      facial and skeletal features supports a Positive diagnosis,
      even with the presence of an additional significant renal
      phenotype (SRNS) which may represent a second hit or an
      expanded phenotype.",
  "genotype_reasoning":
    "The molecular audit confirms PVS1 (Very Strong) due to the
      early frameshift in a gene where LoF is a known mechanism. PM2
      is applied as the variant is absent in gnomAD. Phenotype match
      is moderate due to characteristic facial features despite the
      atypical renal presentation. Following AD integration rules
      (LP + Moderate), the result is Positive.",
  "phenotype_assessments": [
    {
      "key": "OMIM:148050",
      "disease_title": "KBG syndrome",
      "is_primary": true,
      "phenotype_fit": "Moderate",
      "report_decision": "Report",
      "reasoning":
        "[Primary] KBG syndrome | The patient exhibits several
         facial features characteristic of KBG syndrome, including
         a broad nose (bulbous nose/nasal tip), thin upper lip, and
         downslanted palpebral fissures. Ear pits and simple ears
         are compatible with the reported ear anomalies in KBGS.
         While the patient presents with steroid-resistant nephrotic
         syndrome (SRNS) and FSGS, which are not hallmark features
         of KBGS, OMIM and literature mention renal manifestations
         like cystic dysplasia, suggesting renal involvement is
         rare but possible. The absence of reported intellectual
         disability or macrodontia (likely due to age/dentition
         status) reduces the score from Definitive/Strong to
         Moderate, but the facial gestalt is highly suggestive. |
         Decision: Report (VP=Likely pathogenic, PP=Moderate,
         INH=Autosomal dominant)"
    }
  ],
  "tool_calls_made": [
    "get_omim_entry",
    "get_variant_detail",
    "get_clinvar_detail"
  ]
}
```


The single rollout output captures (i) the molecular audit (Phase A), (ii) the phenotype assessment (Phase B), and (iii) the integrated decision (Phase C) within one LLM call. The `decision` field is the binary R/NR mapping; `final_judgment` is the native 3-class label. The case-level output of AIVARI Agent for a patient is the list of such per-variant JSON objects, on which the case-level metrics in Section 3.3 are computed.