Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GeneResolver: A Traceable Hybrid Multi-Agent Pipeline for Etiology-Aware Gene Prioritization

Ivana Vichentijevikj

ivana@ireason.mk

iReason

Kostadin Mishev

Ss. Cyril and Methodius University in Skopje

Monika Simjanoska Misheva

Ss. Cyril and Methodius University in Skopje

Research Article

Keywords: rare disease, gene prioritization, Human Phenotype Ontology, clinical genomics, privacy-preserving machine learning, large language models

Posted Date: September 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10915520/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GeneResolver: A Traceable Hybrid Multi-Agent Pipeline for Etiology-Aware Gene Prioritization

Ivana Vichentijevikj[1*], Kostadin Mishev[2] and Monika Simjanoska Misheva[2]

[1*]iReason LLC, 3rd Macedonian Brigade 37, Skopje, 1000, North Macedonia.

[2]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, P.O. 393, Skopje, 1000, North Macedonia.

*Corresponding author(s). E-mail(s): ivana@ireason.mk;

Abstract

Background: Rare-disease diagnosis requires the integration of clinical phenotypes, genomic variants, inheritance patterns, and heterogeneous biological evidence. Existing gene-prioritization tools commonly assume that a disorder is caused by a single gene and that a curated Human Phenotype Ontology profile is already available. We developed GeneResolver as a traceable, privacy-oriented hybrid multi-agent pipeline that determines the likely etiology before applying multi-evidence gene prioritization. Cases are classified as single-gene, chromosomal, multifactorial, or non-genetic so that presentations unlikely to have a single-gene cause can enter mechanism-specific evidence branches instead of gene ranking. Language models support contextual interpretation, whereas ontology resolution, evidence retrieval, variant annotation, and numerical ranking remain deterministic and traceable.

Results: On 196 published case narratives, classification of raw narratives without few-shot examples achieved four-class accuracy of 77.0%, a macro-averaged F1 score of 0.767, and single-gene-versus-rest branch-separation accuracy of 86.2%. On 4,677 phenotype-only cases from the PhEval benchmark, GeneResolver achieved precision at rank one of 0.4368 and a mean reciprocal rank of 0.5154, compared with published Exomiser values of 0.391 and 0.458, respectively. On the LIRICAL spiked-exome benchmark, the percentage of cases ranked first increased from 45.80% with phenotype data alone to 54.57% when phenotype and variant call format data were combined; mean reciprocal rank increased from 0.5465 to 0.6421.

Conclusions: GeneResolver provides a reference architecture for integrating etiological routing, clinical-text interpretation, phenotype grounding, optional genomic-variant analysis, and traceable gene prioritization. The results support further evaluation of etiology-aware hybrid systems as clinical decision-support tools. GeneResolver remains a proof-of-concept intended to assist, rather than replace, expert clinical interpretation.

Keywords: rare disease, gene prioritization, Human Phenotype Ontology, clinical genomics, privacy-preserving machine learning, large language models

1 Background

Patients with rare genetic diseases frequently endure prolonged diagnostic journeys, even in the era of exome and genome sequencing. Although sequencing has accelerated variant discovery, clinical diagnosis still depends on prioritizing a small number of plausible causal variants from a large candidate set by integrating patient phenotype, population frequency, predicted molecular consequence, gene–disease knowledge and inheritance model. Consequently, the major bottleneck in many rare-disease workflows has shifted from generating variants to interpreting and prioritizing them. [1]

Phenotype-driven gene prioritization tools address part of this burden by ranking genes or diseases according to the similarity between a patient's phenotype profile, commonly encoded using the Human Phenotype Ontology (HPO) [2], and known disease or gene annotations. Established systems include Exomiser [3], which combines phenotype similarity with variant filtering and remains a widely used reference tool for exome-based prioritization, LIRICAL [4], which applies a likelihood-ratio pipeline to phenotype-driven diagnosis, and phenotype-only prioritizers such as Phen2Gene [5], together with embedding- and network-based methods. More recently, large language model based systems have begun to explore agentic reasoning for rare-disease interpretation. Knowledge-driven systems such as AI-MARRVEL [6] combine genotype, phenotype and curated disease knowledge to support Mendelian diagnosis, rare-disease reasoning agents such as DeepRare [7] orchestrate retrieval and multi-step reasoning over phenotype descriptions, and general biomedical agents such as Biomni [8] demonstrate how agentic systems can plan and execute bioinformatics analyses across large tool libraries. These approaches illustrate the promise of agentic reasoning, but most still assume that the patient has already been represented as a structured phenotype profile, or they rely on deployment settings in which sensitive clinical data may be processed outside the institution.

Two gaps motivate the present work. First, many prioritization pipelines begin from a curated HPO profile. The clinically important and time-consuming step of transforming an unstructured referral letter or clinical narrative into a structured, negation-aware phenotype profile is often treated as a preprocessing task outside the prioritization system itself. Second, privacy-preserving deployment remains insufficiently addressed in many LLM-assisted clinical genomics workflows. Clinical narratives and genomic data contain highly sensitive information, including health and

genetic data, and their processing is subject to strict data-protection and medical-AI governance requirements. For this reason, rare-disease decision-support systems need architectures that can operate locally or minimize the information exposed to external models.

The aim of this study was to develop and evaluate **GeneResolver**, a gene-prioritization decision-support system designed to address these gaps. GeneResolver integrates clinical-text processing, phenotype extraction, ontology grounding, optional variant call format based interpretation, etiological routing, and gene prioritization within a single traceable pipeline. We evaluated whether this architecture could distinguish single-gene cases from other etiological classes, prioritize causal genes from phenotype data, and improve prioritization when genomic-variant data were available.

The remainder of the paper is structured as follows. We first describe the GeneResolver pipeline and experimental setup (Section 2), then report the results (Section 4). We discuss the implications of these findings in Section 5, before turning to ethical and regulatory considerations (Section 6) and outlining limitations and directions for future work (Section 7). Finally, we conclude in Section 8.

2 Materials and methods

2.1 System overview

GeneResolver adapts its analysis to the clinical and genomic data available as input. Inputs may include any combination of a free-text clinical narrative, coded HPO terms, a genome supplied as a VCF or DRAGEN output directory, a single candidate variant, a named candidate gene, and a variant identifier (dbSNP rsID). The system selects one of five processing modes according to the available input fields:

(i) Combined clinical-genomic analysis is performed when both clinical findings and a genome are available.
(ii) Clinical-only analysis is performed when clinical findings are available without a genome.
(iii) Variant interpretation is performed when a genome or an individual candidate variant is provided without clinical findings.
(iv) A gene-disease query is performed when a named candidate gene is provided.
(v) A single-locus lookup is performed when a dbSNP rsID is provided without a clinical narrative or VCF path.

Mode selection is determined by explicit rules applied to the populated input fields. In the combined clinical-genomic, clinical-only, variant-interpretation, and gene-disease-query modes, candidate diagnoses or gene-disease relationships are additionally weighed against retrieved literature, pathway, and gene-disease database evidence (Table 1). The single-locus lookup mode performs a more limited locus-centred retrieval, including GWAS Catalog evidence, rather than the full multi-source evidence workflow. An etiology classifier is run in the combined clinical-genomic, clinical-only, and variant-interpretation modes to determine the likely disorder class: single-gene, chromosomal, multifactorial, or non-genetic. This allows the case to be routed either to gene prioritization or to a mechanism-specific evidence branch, as described in

Section 2.4. Each mode invokes only the analyses relevant to its inputs: clinical processing runs only when clinical findings are present, and variant processing runs only when a genome is provided.

Within a mode, the case is analyzed by a sequence of specialist agents, each responsible for one well-defined sub-task and each contributing to a shared, structured representation of the case. Concretely, the orchestrator is a LangGraph over the case state, with each specialist agent as one node and a conditional edge deciding the next step from the case state alone. This modular organization keeps the pipeline traceable: every intermediate result is recorded and attributable to the agent that produced it, together with the evidence in which it was grounded.

Table 1 Specialist agents and components of the GeneResolver pipeline, grouped by stage

Stage	Agent / component	Type	Function
Ingest	De-identifier	Det. (NLP)	Masks PHI (Protected Health Information) spans in the clinical note
Ingest	Disease extractor	LLM + Det.	Disease-mention NER; each mention is grounded and normalized to Orphanet/OMIM/ICD-10/MONDO, and unresolved mentions are discarded
Ingest	Negation detector	LLM	Assigns present, absent, or uncertain status to disease and phenotype mentions from the surrounding context
Ingest	Phenotype mapper (FastHPOCR)	Det. + LLM	Extracts offset-tagged HPO terms; a supplementary LLM pass proposes missed phenotypes, each of which is grounded to an existing HPO term through ontology search
Ingest	Clinical synthesizer	LLM	Additive, validated enrichment of missed phenotypes and syndrome hypotheses
Variant	Variant annotator (VEP)	Det.	Retrieves consequence, HGVS, population-frequency, and clinical-significance annotations
Variant	ACMG classifier	Det.	ACMG/AMP 2015 pathogenicity classification
Variant	Inheritance agent	LLM + Det.	The locally deployed LLM infers the suspected Mendelian inheritance pattern from the unmasked family-history and pedigree context. Deterministic rules then evaluate candidate variants for compatibility with the inferred pattern using zygosity and sex-linkage information.
Etiology	Etiology classifier	LLM	Single call over the case's raw clinical narrative that states the specific causal mechanism, then assigns one of four etiology labels: single-gene, chromosomal, multifactorial, or non-genetic
Etiology	Evidence branch (×4)	LLM	Four mechanism-specific retrieve-reason branches that emit candidate conditions, the proposed mechanism, supporting evidence, and limitations
Candidate gen.	Gene aggregator	Det.	Assembles candidates from phenotype, variant and disease-mention streams
Grounding	Evidence grounder	Det.	Retrieves Open Targets [9], UniProt, Reactome, gnomAD, GWAS Catalog [10], PubMed and HPO-A LRs
Ranking	Multi-evidence gene ranker	Det.	Seven-component weighted fusion to a ranked, confidence-banded gene list
Differential	Biochemical-marker agent	Det.	Specificity-gated metabolic recall: seeds candidate diseases into the differential pool via the HPO-A inverted index
Differential	Disease reasoner (propose / ground / score)	LLM + Det.	Proposes disease names, resolves them against Orphanet/MONDO, discards unresolved candidates, and scores the remaining candidates using retrieved evidence
Interpretation	Interpretation agent	LLM	Mechanism summaries and suggested investigations
Output	Report exporter	Det.	Serializes results to JSON, CSV, FHIR R4 and Phenopacket v2

GeneResolver follows a multi-agent architecture in which each diagnostic function is assigned to a specialized agent with a clearly defined role. All LLM-assisted components use the locally deployed **MedGemma-27B-IT** model. Individual agents either execute deterministic procedures, including de-identification, phenotype extraction, VEP-based variant annotation, ACMG/AMP evidence assignment, evidence retrieval, and gene ranking, or perform LLM-assisted tasks requiring contextual interpretation, including clinical synthesis, etiological classification, mechanism-specific reasoning, and report generation. The architecture is summarized in Figure 1. Inputs are ingested through two parallel intake streams: clinical text is de-identified locally and mapped to standardized phenotypes, while variants are annotated and classified. Both feed an *etiology classifier*, which decides what kind of diagnostic problem the case is. That verdict selects between two downstream arms: monogenic cases proceed to multi-evidence gene ranking and a disease differential, while every other etiology takes an evidence branch that does not rank genes. Whichever arm runs, its result is integrated into a single report.

2.2 Intake, privacy filtering and phenotyping

The first phase converts raw clinical text into a structured, standardized profile. A *de-identifier* applies a local clinical de-identification model (selected empirically in Section 4.1.1) to the clinical note and replaces each detected personally identifying span with a typed placeholder. Nearly every subsequent component operates on the masked text. Two controlled exceptions require access to the original note: demographic extraction reads age, sex, and family-history context, while the inheritance agent interprets unmasked family-history and pedigree descriptions that may lose essential relationships during masking. Both operations are therefore restricted to the locally deployed MedGemma-27B-IT model, and the unmasked text is never transmitted to an external service. If de-identification fails, a safety flag is raised and the case is marked as requiring human review.

A *disease extractor* then identifies syndrome- and disease-level statements and normalizes them to Orphanet identifiers [11] (with OMIM, ICD-10 and MONDO cross-references), assigning each a present/absent/uncertain status via a negation detector. A *phenotype mapper* extracts HPO terms using FastHPOCR [12], a deterministic, dictionary-based concept recognizer that returns offset-tagged HPO identifiers with sub-second latency, each annotation is assigned a status by the same negation detector. Finally, a grounded *clinical synthesizer* performs an LLM-assisted enrichment pass that proposes phenotype findings and syndrome hypotheses the dictionary extractors missed.

2.3 Variant processing and ACMG classification

When a VCF is supplied, the pipeline performs quality-control checks and determines the genome assembly from the file header. Single-nucleotide variants and small insertions or deletions marked PASS by the upstream variant caller are then annotated locally with the Ensembl Variant Effect Predictor (VEP) [13]. The resulting annotations include the affected gene, predicted molecular consequence, HGVS nomenclature,

and population allele frequency. Variants are retained if they are rare or have a predicted consequence of high or moderate impact. This preliminary filter reduces the candidate set but does not determine pathogenicity, which is assessed subsequently by the ACMG/AMP classification engine.

An inheritance agent, run with the local MedGemma backend in the evaluated configuration (the same local-model restriction applied to demographics extraction, Section 2.2), uses the reported family history and pedigree context from the family-history field to infer a plausible Mendelian inheritance pattern. It then assesses whether the number and zygosity of variants in each candidate gene are compatible with that pattern, while also accounting for patient sex where relevant. If the available family history is insufficient to support an inheritance model, the pattern is recorded as unknown rather than inferred.

Following initial gene ranking, variants belonging to ranked candidate genes are evaluated in their gene-disease context by a deterministic rule engine implementing the ACMG/AMP 2015 evidence framework [14]. The engine combines the criteria that can be assessed from the available evidence to assign one of five categories: pathogenic, likely pathogenic, uncertain significance, likely benign, or benign. Criteria for which the required evidence is unavailable are explicitly recorded as unevaluable and are not treated as satisfied. The resulting ACMG evidence is then folded back into the variant component of the final gene score, after which the candidate genes are re-ranked.

The engine reports 21 criterion codes. Pathogenic evidence includes: PVS1 for MANE-gated, NMD-relevant loss of function, PS1 for an identical known pathogenic amino-acid change, PS2/PM6 for confirmed/unconfirmed de novo occurrence, PM1 for pathogenic clustering in a functional region, PM2 for rarity, PM3 for recessive phase-unknown evidence, PM4 for in-frame indels or stop-loss outside repeats, PM5 for another pathogenic change at the same residue, PP1 for segregation, PP2 for missense constraint, PP3 for calibrated REVEL or SpliceAI predictions, and PP4 for phenotype match [15–18].

Benign evidence includes BA1 for coverage-confirmed gnomAD popmax frequency above 5% after exception checking, BS1 unevaluable without disease-specific prevalence data to set its expected-frequency threshold, BS2 for an unaffected adult relative reported to carry the same genotype, BS4 for non-segregation, BP1 for missense variants in genes whose known pathogenic variants are established as primarily truncating, BP3 for in-frame indels within repeats, BP4 for benign REVEL or SpliceAI predictions and BP7 for non-splice-relevant, poorly conserved synonymous variants.

Together, variant annotation, inheritance assessment, and ACMG re-ranking provide the genomic evidence layer for cases directed to the single-gene branch by the etiological routing stage described next.

2.4 Etiology classification and branching

Phenotype-driven gene prioritization assumes that the patient's disorder has a monogenic basis. This assumption is inappropriate for chromosomal abnormalities, multifactorial disorders, and non-genetic conditions. Applying gene ranking to such cases may still produce a plausible list of candidates, but the result addresses

Table 2 Evidence components and default weights used for gene prioritization when variant evidence is available

Evidence component	Weight	Source
Phenotype evidence	30%	HPO likelihood ratio
Variant pathogenicity	26%	ACMG/AMP, impact and rarity
Gene-disease association	17%	Open Targets association score
Pathway evidence	9%	Reactome Jaccard overlap
Loss-of-function constraint	8%	gnomAD pLI/LOEUF
Literature support	7%	PubMed support level
GWAS evidence	3%	GWAS Catalog best p-value

Note: In phenotype-only analysis, phenotype, constraint, gene-disease, and pathway evidence receive weights of 0.80, 0.10, 0.05, and 0.05, respectively. The remaining channels receive zero weight.

an inappropriate biological mechanism. GeneResolver therefore performs etiological classification before initiating gene prioritization.

When a clinical narrative is available, the de-identified narrative itself is retained as the primary classifier input. Classification is performed in a single LLM call. The model first identifies and explicitly states the single mechanism that best explains the case. The supported mechanisms include a causal single-gene variant, a chromosomal abnormality, a polygenic or gene-environment multifactorial pattern, and a non-genetic cause. The model then provides its reasoning and commits to the corresponding etiological label as the final step. The classifier returns a structured record containing one of the four labels: single-gene, chromosomal, multifactorial, or non-genetic or an unresolved outcome. An unresolved case is retained as uncertain rather than being forced into an etiological category.

The orchestrator uses the committed decision for downstream routing. A single-gene classification proceeds to gene prioritization, whereas chromosomal, multifactorial, and non-genetic classifications are directed to their corresponding mechanism-specific evidence branches (Section 2.6).

2.5 Gene prioritization by multi-evidence fusion

The evidence channels and their default weights are summarized in Table 2.

Gene prioritization is performed for cases assigned to the single-gene pathway. Candidate genes are assembled from multiple sources: phenotype evidence derived from HPO terms, variant evidence when a VCF is available, genes associated in Orphanet with disease mentions extracted from the case, Open Targets disease-associated genes.

The principal ranking signal is a disease-aware likelihood-ratio score. For each candidate disease D, the contributions of the patient's observed HPO terms are calculated as:

$$S_{\text{phen}}(D) = \sum_{t \in H_{\text{obs}}} \log \frac{P(t \mid D)}{P(t \mid \neg D)}, \tag{1}$$

where $P(t \mid D)$ represents the frequency of phenotype term t in disease D, and $P(t \mid \neg D)$ represents its background frequency across other diseases. Common and

nonspecific findings therefore contribute little, whereas rare and disease-specific findings provide stronger evidence. Explicitly absent findings introduce a frequency-scaled penalty, bounded so that one missing feature cannot dominate the complete phenotype. Disease-level scores are subsequently transferred to their associated genes to obtain a gene-level phenotype score.

This likelihood-ratio formulation reduces the tendency of overlap-based semantic similarity to favour extensively annotated genes that share many nonspecific terms with the patient.

When a VCF is available, candidate variants are first annotated using VEP. Before ACMG/AMP classification is performed, every candidate variant receives a generic triage score based on its predicted molecular consequence and population frequency:

$$S_{\text{triage}}(v) = w_{\text{impact}}(v) \times w_{\text{rarity}}(v), \tag{2}$$

where the impact weight ranges from 0.85 for high-impact consequences to 0.30 for modifier consequences. The rarity weight ranges from 1.00 for variants with a gnomAD allele frequency below 10^{-4} to 0.55 for variants with an allele frequency of at least 10^{-2}. This score provides an initial, classification-independent prioritization of all candidate variants rather than a refinement applied specifically to variants of uncertain significance.

The ACMG/AMP engine is applied after the initial gene ranking to variants of ranked candidate genes for which a gene-disease context can be constructed. When it returns a classification, the corresponding fixed classification score supersedes the variant's triage score. In particular, all variants classified as being of uncertain significance receive a score of 0.5, irrespective of their predicted consequence or population frequency. The triage score is retained when no downstream ACMG/AMP classification score is available. The updated variant score is folded back into the gene's final score and the candidate list is re-ranked.

Functional annotations are retrieved from UniProt [19] and the corresponding identifiers are also used to retrieve Reactome pathways. The evidence channels supplementing phenotype and variant evidence comprise gene-disease associations from Open Targets [9], pathway evidence from Reactome [20], loss-of-function constraint from gnomAD [21], literature support from PubMed, and association evidence from the GWAS Catalog [10]. PubMed evidence is represented as a categorical support level, mapped to 1.0 for strong, 0.6 for moderate, 0.3 for weak, and 0 for no support.

After min-max normalization of the phenotype scores across the candidate pool, the seven scored evidence channels are combined into a weighted score. The default weights used when variant evidence is available are presented in Table 2:

$$\begin{aligned} S_{\text{fusion}}(g) = {} & 0.30 S_{\text{phen}}(g) + 0.26 S_{\text{variant}}(g) + 0.17 S_{\text{gene-disease}}(g) \\ & + 0.09 S_{\text{pathway}}(g) + 0.08 S_{\text{constraint}}(g) \\ & + 0.07 S_{\text{literature}}(g) + 0.03 S_{\text{GWAS}}(g). \end{aligned} \tag{3}$$

The weighting scheme is adapted to the available evidence. When both phenotype and VCF data are present, phenotype and variant scores jointly influence the ranking,

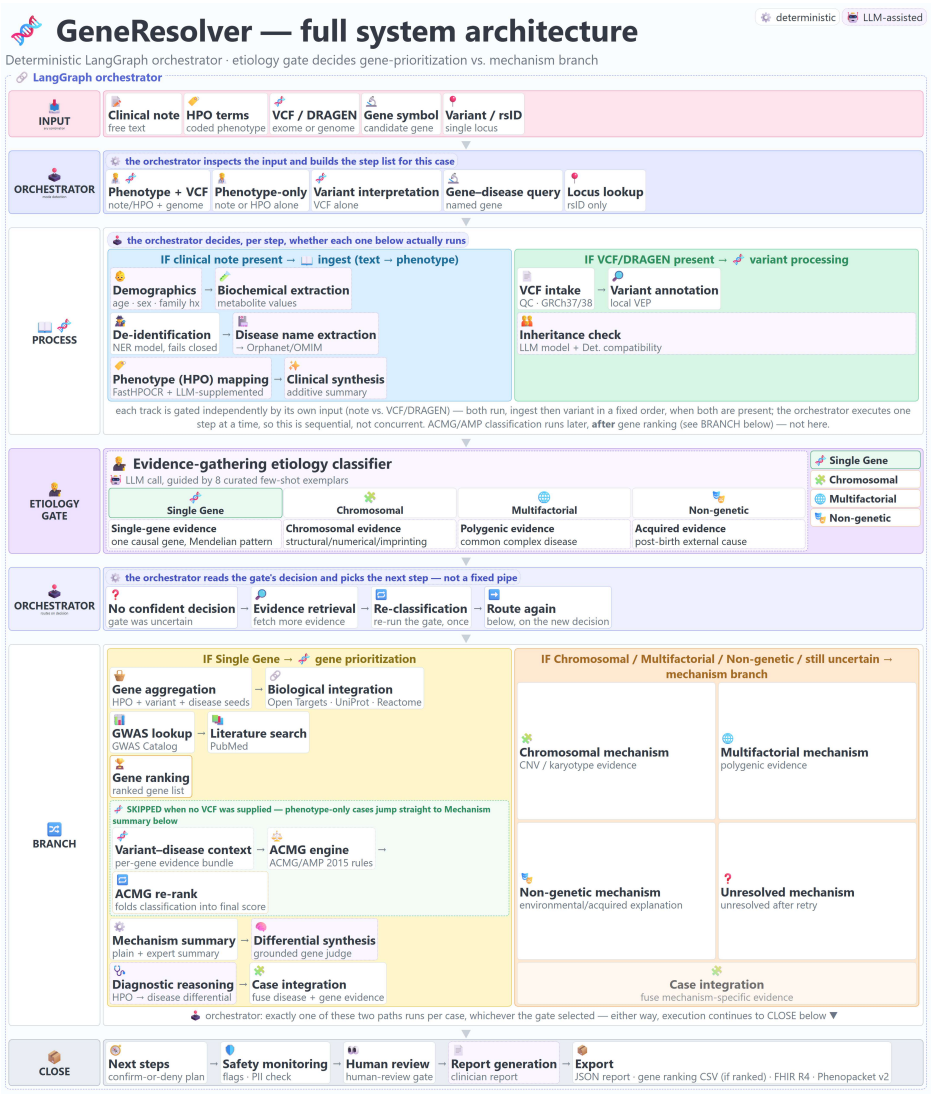


Fig. 1 GeneResolver architecture. A rule-based orchestrator coordinates agents through a shared case state. Clinical intake performs local de-identification, disease normalization, FastHPOCR phenotype extraction, and optional LLM enrichment; genomic intake performs VCF quality control, VEP annotation, rarity filtering, inheritance assessment, and ACMG/AMP classification. Etiological routing sends single-gene cases to candidate aggregation, evidence fusion, re-ranking, and disease interpretation, while other cases enter mechanism-specific branches. Dashed outlines denote on-premise MedGemma-27B-IT components; identifier resolution, evidence retrieval, variant processing, and numerical ranking remain deterministic.

allowing a compatible pathogenic variant to strengthen the position of the corresponding gene. In phenotype-only analysis, the inactive variant channel receives zero weight, and the score is calculated as

$$S_{\text{phenotype-only}}(g) = 0.80S_{\text{phen}}(g) + 0.10S_{\text{constraint}}(g) + 0.05S_{\text{gene-disease}}(g) + 0.05S_{\text{pathway}}(g). \tag{4}$$

Following evidence fusion, bounded adjustments account for contradictions with explicitly absent HPO findings, incompatibility between the patient's sex and the expected inheritance pattern, and agreement with a disease already supported by the case. The maximum adjustments are -0.05 for negated-phenotype evidence, -0.20 for Y-linked genes and -0.05 for X-linked-recessive genes in female patients, and $+0.10$ for an Orphanet disease match. These adjustments act as tie-breakers and cannot override substantially stronger phenotype or variant evidence. The resulting score is clamped to $[0, 1]$, and genes are assigned high (≥ 0.70), moderate (≥ 0.50), low (≥ 0.30) or insufficient confidence.

The final ranking is presented as an evidence-supported set of candidate genes for expert review. For the leading candidates, the report summarizes phenotypic compatibility, relevant variants, biological mechanisms, pathway relationships, supporting evidence and identified limitations. The ranked list is intended to guide confirmatory genetic, biochemical and inheritance-based investigations rather than to represent a definitive diagnosis.

2.6 Mechanism-specific evidence branches

Gene ranking answers "which gene?", a question that is ill-posed when the case is not monogenic. Cases the classifier (Section 2.4) assigns to a chromosomal, multifactorial, or non-genetic etiology, together with cases that remain unresolved after retry, are therefore routed away from the gene-ranking engine and into a mechanism-specific evidence branch. The four branches share a single implementation and differ only through a descriptor that specifies which databases to consult, how to steer the queries, what the reasoning step should establish, and which databases are unavailable for that etiology.

Each branch retrieves class-appropriate evidence: literature from Europe PMC [22], gene-disease associations from Open Targets, and, for the multifactorial and uncertain branches, trait associations from the GWAS Catalog. Queries are seeded with disease and gene hypotheses when available, together with phenotype terms. To minimize the information transmitted online, no more than two phenotype terms are included, and only terms sufficiently common to avoid acting as quasi-identifiers are permitted. An HPO term associated with very few diseases in HPO-A is therefore excluded from external queries. Genomic findings, where a VCF is available, are provided to every branch, so a non-monogenic case can still reference variant evidence. A reasoning step then operates strictly over the retrieved evidence to produce a set of candidate conditions with supporting and opposing evidence, an explanation of the mechanism by which the assigned etiology produces the observed findings, and the reasoning connecting the two.

2.7 Clinical interpretation and structured output

For the top-ranked candidate genes in the single-gene arm, an *interpretation agent* produces a human-readable rationale: plain-language and expert-level mechanism summaries covering each gene's known molecular function, pathways, and disease associations. Like every other LLM-assisted step, it is constrained to the retrieved-evidence bundle. For the non-monogenic branches, the equivalent mechanism explanation is not a separate step: it is produced within the branch's own evidence-gathering call.

A unified reporting component consolidates the outputs of the selected etiological pathway into a clinically interpretable and machine-readable format. It performs one LLM-assisted synthesis step using only evidence generated upstream. For the single-gene branch, this includes ranked genes, prioritized variants, and variant classifications. For the chromosomal, multifactorial, and non-genetic branches, it includes the mechanism-specific interpretation, candidate conditions, supporting evidence, and stated limitations. The component structures these findings without introducing new candidates or modifying branch conclusions. Results are exported as JSON, FHIR R4, and GA4GH Phenopacket v2, with an additional CSV gene-ranking file for single-gene cases. Representative summaries from the phenotype-only, mechanism-specific, and phenotype-plus-VCF pathways are shown in Figure 2. Example output reports are available in the GitHub repository (https://github.com/ChatMED/GeneResolver).

3 Datasets and evaluation protocol

We evaluate the pipeline stage by stage on task-specific datasets so that the contribution of each component can be assessed independently. The de-identification and phenotype-extraction components are compared with alternative methods to justify their selection, while the etiology classifier is evaluated both as a four-class prediction task and as the routing decision that separates single-gene cases from all other etiologies.

De-identification is evaluated on a held-out set of 100 synthetic clinical notes generated specifically for this study and annotated with PHI spans. Performance is reported as token-level PHI-removal recall with a 95% bootstrap confidence interval and the number of leaked PHI tokens. The silver dataset-generation procedure and prompt are provided in Supplementary Material 1.

Phenotype extraction is evaluated on the GSC+ gold-standard concept-recognition corpus [23], which contains 228 documents. Performance is reported using precision, recall, and F1 score.

Etiology classification is evaluated using a silver-standard dataset derived from published adult case reports. To our knowledge, no public case-level benchmark supports etiological classification from clinical narratives across single-gene, chromosomal, multifactorial, and non-genetic disorders. We therefore constructed a 220-case silver-standard corpus from published case reports, as described in Supplementary Material 2. A fixed eight-case pool used to construct the few-shot condition was excluded from scoring to preserve a paired comparison. The evaluated conditions are raw narrative without few-shot examples, raw narrative with few-shot examples, RAG-enriched narrative without few-shot examples, and RAG-enriched narrative with

few-shot examples. Although the dataset retains somatic as a provenance label, the deployed classifier uses a four-class output space: single-gene, chromosomal, multifactorial, and non-genetic. Before scoring, each somatic case is mapped to its recorded single-gene or multifactorial subgroup, ensuring that all cases are evaluated within the four-class space without discarding somatic presentations. Performance is assessed using four-class accuracy, macro-F1, and *branch separation*, defined as the accuracy of the binary single-gene-versus-rest routing decision. Accuracy and branch-separation estimates are reported with 95% bootstrap confidence intervals.

Phenotype-only gene ranking, the primary task, is measured on the PhEval corpus [24] derived from the Phenopacket Store [25] (4,677 cases). The reported metrics are precision@k, mean average precision (MAP@k), F_β@k, normalized discounted cumulative gain (NDCG@k) and mean reciprocal rank (MRR). GeneResolver is scored locally with the PhEval framework, whereas the Exomiser, PhenoGenius [26], Phen2Gene, and GADO [27] values are reproduced from the published PhEval statistics.

Genome-level evaluation uses the real causal variant from each case in the LIRICAL benchmark corpus, spiked into a shared public GRCh37 template exome using the PhEval spiked-VCF procedure to produce reproducible whole-exome VCFs. PhEval percentage@k, precision@k, MAP@k, F_β@k, NDCG@k, and MRR are reported for the phenotype-only and phenotype+VCF arms.

4 Results

We first report three component evaluations (de-identification, phenotype extraction, and etiology classification) then the end-to-end phenotype-only gene-ranking benchmark against the published baselines under the PhEval framework, and finally the genome-level (HPO+VCF) ablation. Gene ranking is the primary task; the components are evaluated separately so that each production choice is justified on its own task-specific dataset rather than only end-to-end.

4.1 Component evaluation

We report the three component evaluations in turn, de-identification, phenotype extraction, and etiology classification, each on its task-specific dataset (Section 3).

4.1.1 De-identification

We compared three locally deployable de-identification models on a held-out set of 100 synthetic clinical notes, with the dataset construction protocol and generation prompt provided in Supplementary Material 1. The evaluated models were the Stanford clinical de-identifier (StanfordAIMI/stanford-deidentifier-base) [28], a RoBERTa de-identifier trained on the i2b2/UTHealth 2014 corpus (obi/deid-roberta-i2b2) [29], and the OpenMed privacy filter (OpenMed/privacy-filter-multilingual) [30–33]. OpenMed achieved the highest recall (0.772), followed by the Stanford de-identifier (0.661) and the i2b2-trained RoBERTa model (0.635) (Table 3). Based on its superior PHI-removal performance while maintaining fully local deployment, OpenMed was selected as the GeneResolver production component.

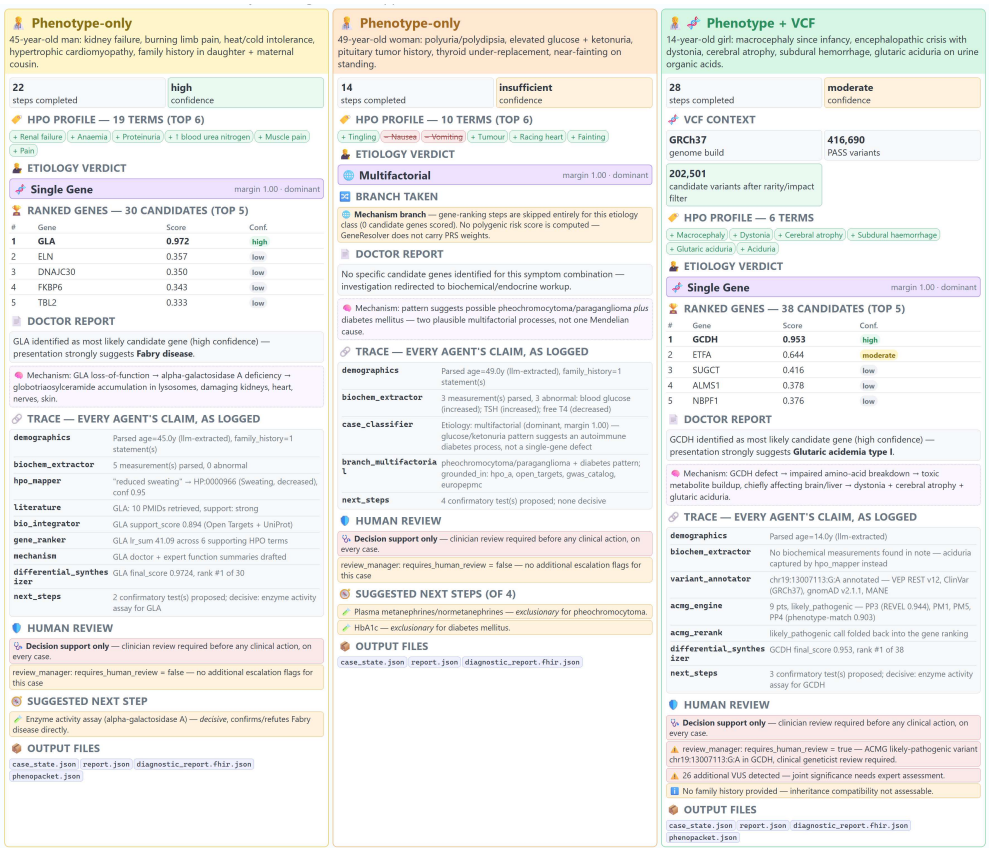


Fig. 2 Representative GeneResolver output summaries for three demonstration cases. The left panel shows a phenotype-only case routed to the single-gene pathway, including the extracted HPO profile, etiological verdict, ranked candidate genes, mechanism summary, and suggested next investigation. The centre panel shows a multifactorial verdict for which gene ranking is bypassed and the report instead presents a mechanism-specific interpretation and suggested investigations. The right panel shows a combined phenotype-plus-VCF analysis, including variant-processing context, gene ranking, provisional variant evidence, and explicit human-review flags. These panels show only excerpts from the complete reports, which are available in the GitHub repository (see the Availability of data and materials section).

Table 3 De-identification model comparison on a held-out set of 100 synthetic clinical notes (3,632 PHI tokens). Recall is the fraction of PHI tokens removed (the safety-critical metric), leaks are tokens that survived masking

Model	Recall (95% CI)	Leaks
OpenMed privacy filter	**0.772** (0.7642-0.7800)	**828/3,632**
Stanford de-identifier	0.661 (0.6458-0.6752)	1,233/3,632
deid_roberta (i2b2-2014)	0.635 (0.6189-0.6525)	1,324/3,632

Table 4 Phenotype-extraction methods on the GSC+ corpus (228 documents), document-level micro-averaged.

Method	Precision	Recall	F1 (95% CI)
FastHPOCR (selected)	**0.777**	**0.686**	**0.729 (0.7069-0.7506)**
ClinPhen	0.753	0.422	0.541 (0.5048-0.5764)

Table 5 Etiology classification ablation across 196 dataset case narratives: few-shot guidance and RAG-enriched context, independently varied.

Few-shot	Context	4-way acc. (95% CI)	Macro F1	Branch sep. (95% CI)
No	**Raw**	**77.0% (70.9-82.7%)**	**0.767**	**86.2% (81.6-91.3%)**
Yes	Raw	72.4% (66.3-78.6%)	0.715	83.2% (78.1-88.3%)
Yes	RAG	72.4% (65.8-78.6%)	0.713	84.2% (79.1-89.3%)
No	RAG	72.4% (65.8-78.6%)	0.710	82.1% (76.5-87.2%)

4.1.2 Phenotype (HPO) extraction

Phenotype extraction was evaluated on the GSC+ corpus (228 documents). The selected deterministic concept recognizer (FastHPOCR [12]) achieves F1 = 0.729 (precision 0.777, recall 0.686), substantially exceeding ClinPhen [34] (F1 = 0.541), whose lower F1 is driven by markedly lower recall (0.422 versus 0.686) (Table 4).

4.1.3 Etiology classification

The etiology classifier (Section 2.4) was evaluated on 196 cases from a 220-case corpus under all four combinations of two design choices: with and without few-shot guidance and, independently, with and without RAG-enriched context (Table 5). Because the classifier's operational role is to gate entry to the gene ranker, we report both overall four-way accuracy and *branch separation*: the accuracy of the single-gene-versus-rest routing decision.

Raw-narrative classification with no few-shot examples numerically outperformed every other condition on accuracy, macro F1, and branch separation. We therefore selected raw-narrative classification with *no* few-shot guidance and *no* RAG-enriched context as the production configuration: the classifier operates directly on the case representation generated by the upstream components, with neither curated examples nor an additional extraction-and-retrieval pass providing a measurable benefit at this sample size.

4.2 Phenotype-only gene ranking on the PhEval benchmark

On the PhEval 4,677-case corpus [24] (Section 3), GeneResolver achieved precision@1 of 0.4368, MRR of 0.5154, MAP@10 of 0.5113, and NDCG@10 of 0.5281. Table 6 compares these results with the published PhEval statistics for Exomiser, PhenoGenius, Phen2Gene, and GADO. The baseline systems were not independently rerun in this study, so the table provides benchmark context rather than a paired local evaluation.

Table 6 Phenotype-only ranking and binary-classification metrics on the PhEval corpus (4677 cases)

Metric	GeneResolver	Exomiser	PhenoGenius	Phen2Gene	GADO
MRR	0.5154	0.458	0.263	0.180	0.023
Precision@1	0.4368	0.391	0.162	0.096	0.006
Precision@3	0.1883	0.164	0.097	0.071	0.005
Precision@5	0.1226	0.104	0.075	0.052	0.004
Precision@10	0.0669	0.057	0.049	0.033	0.004
MAP@1	0.4368	0.391	0.162	0.096	0.006
MAP@3	0.4932	0.436	0.217	0.150	0.010
MAP@5	0.5040	0.442	0.237	0.160	0.012
MAP@10	0.5113	0.448	0.252	0.169	0.014
F_β@1	0.4365	0.391	0.162	0.096	0.006
F_β@3	0.2823	0.231	0.121	0.082	0.006
F_β@5	0.2043	0.164	0.102	0.067	0.005
F_β@10	0.1215	0.099	0.075	0.049	0.005
NDCG@3	0.5092	0.260	0.136	0.151	0.015
NDCG@5	0.5226	0.249	0.165	0.163	0.016
NDCG@10	0.5281	0.221	0.184	0.172	0.022

4.3 Gene ranking with genomes: VCF ablation

The phenotype+VCF arm achieved higher arm-level ranking metrics than the phenotype-only arm across all reported metrics (Table 7). Percentage@1 increased from 45.80% to 54.57%, percentage@5 from 65.39% to 76.65%, percentage@10 from 72.01% to 80.96%, and MRR from 0.5465 to 0.6421. Thus, the VCF-augmented arm showed an 8.77-percentage-point increase at rank 1 and an 11.26-point increase at rank 5.

5 Discussion

This study tested whether a privacy-oriented gene-prioritization pipeline using a local language model could achieve accuracy comparable to established systems. GeneResolver combines etiological routing and gene prioritization while keeping sensitive clinical and genomic processing local in the evaluated configuration. Optional online retrieval used only limited, non-identifying query terms.

The selected raw-narrative classifier achieved accuracy of 0.770, macro-F1 of 0.767, and single-gene-versus-rest accuracy of 0.862 on 196 paired cases. Neither few-shot guidance nor RAG improved performance. Routing chromosomal, multifactorial, and non-genetic presentations to dedicated branches reduces the risk of producing biologically inappropriate gene lists.

On the 4,677-case PhEval benchmark, GeneResolver achieved precision@1 of 0.4368 and MRR of 0.5154, numerically exceeding the published Exomiser values of 0.391 and 0.458.

The deterministic ranker extends phenotype likelihood ratios [4] with constraint, gene-disease, pathway, literature, association, and variant evidence. On the LIRICAL

Table 7 Gene ranking on the LIRICAL spiked-exome benchmark

Metric	Phenotype-only	Phenotype + VCF
Percentage@1 (%)	45.80	**54.57**
Percentage@3 (%)	60.05	**69.04**
Percentage@5 (%)	65.39	**76.65**
Percentage@10 (%)	72.01	**80.96**
Precision@1	0.4675	**0.5584**
Precision@3	0.2043	**0.2355**
Precision@5	0.1335	**0.1569**
Precision@10	0.0735	**0.0829**
MAP@1	0.4675	**0.5584**
MAP@3	0.5316	**0.6234**
MAP@5	0.5442	**0.6414**
MAP@10	0.5530	**0.6475**
F_β@1	0.4627	**0.5520**
F_β@3	0.3049	**0.3512**
F_β@5	0.2217	**0.2605**
F_β@10	0.1334	**0.1503**
NDCG@3	0.5498	**0.6422**
NDCG@5	0.5650	**0.6685**
NDCG@10	0.5729	**0.6648**
MRR	0.5465	**0.6421**

benchmark, phenotype+VCF analysis increased percentage@1 from 45.80% to 54.57%, percentage@5 from 65.39% to 76.65%, and MRR from 0.5465 to 0.6421.

LLMs perform contextual interpretation, whereas de-identification, ontology resolution, candidate generation, annotation, retrieval, and scoring remain deterministic. This separation supports reproducibility, provenance, privacy, and expert review while retaining the coordination benefits demonstrated by recent biomedical agents [6–8].

The findings support the feasibility of privacy-oriented, etiologically aware genomic decision support, but do not establish clinical readiness or regulatory compliance. Prospective validation, governance, and meaningful human oversight remain essential.

6 Ethical and regulatory considerations

GeneResolver is a research prototype, not an approved clinical system. Its classifications, ranked genes, and proposed investigations require review by a qualified clinician who can reject or override every output.

This patient-level focus differs from our earlier Prompt-to-Pill research platform [35] and motivates explicit human oversight and regulatory-aware design [36].

Clinical narratives and variants constitute health and genetic data under the GDPR. In the evaluated configuration, local de-identification, inference, and variant processing support privacy by design, standardized exports and evidence provenance partly address broader interoperability and provenance gaps [37]. These features neither establish compliance nor authorize clinical use.

This study used only publicly available published case reports, publicly accessible benchmark datasets, and fully synthetic clinical notes. No new human participants were recruited, and the authors did not access private or directly identifiable patient data. Accordingly, institutional ethics approval and additional informed consent were not required for the analyses reported here.

7 Limitations and future work

At this proof-of-concept stage, GeneResolver is presented as a modular pipeline with initial evaluations of selected components rather than comprehensive end-to-end validation. Several experiments relied on synthetic or silver-standard data, which cannot fully represent the complexity and variability of clinical practice.

De-identification was evaluated on synthetic notes using publicly available models that showed limited performance and are not ready for clinical deployment. Future work should evaluate stronger, clinically adapted models on governed clinical notes. Etiological routing used a 220-case, LLM-assisted silver-standard corpus because no established reference benchmark exists. Its taxonomy was author-defined, it lacked independent adjudication by multiple clinical geneticists, and paired analysis was limited to 196 complete cases. Larger, balanced, and independently reviewed datasets are therefore needed.

The primary task evaluated was gene ranking. Although the pipeline also supports diagnostic reasoning, mechanism interpretation, inheritance assessment, and explanations, these outputs were not independently evaluated. Future studies should validate them using confirmed cases, branch-specific reference standards, and expert review.

GeneResolver has not undergone prospective clinical validation, usability testing, formal risk management, cybersecurity evaluation, or regulatory assessment. Future work should prioritize governed clinical notes and VCFs, stronger local models, comprehensive evaluation, and clinician-supervised prospective studies.

8 Conclusion

GeneResolver introduces a modular pipeline that uses etiological routing to determine when gene prioritization is biologically appropriate and integrates phenotype, variant, and biological evidence when applicable. Initial evaluations demonstrated competitive phenotype-only ranking, with improved aggregate metrics when valid VCF analyses were available. By restricting language models to contextual tasks while retaining deterministic annotation, retrieval, and scoring, GeneResolver provides a proof of concept for privacy-oriented, traceable, and etiology-aware genomic decision support. Independent end-to-end and prospective clinical validation remain necessary before clinical deployment.

Declarations

Ethics approval and consent to participate

Not applicable. This study used only publicly available published case reports, publicly accessible benchmark datasets, and fully synthetic clinical notes. No new human participants were recruited, and the authors did not access private or directly identifiable patient data.

Consent for publication

Not applicable. The manuscript contains no new identifiable data relating to an individual person.

Availability of data and materials

The datasets and code supporting this study are available in the GeneResolver GitHub repository: https://github.com/ChatMED/GeneResolver. The publicly available external benchmark datasets used in the evaluation are identified and cited in the Methods section.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the European Union's Horizon Europe programme under the ChatMED project (Grant Agreement No. 101159214). The funder had no role in the study design, data collection, analysis, interpretation of the results, or preparation of the manuscript.

Authors' contributions

I.V. conceived and implemented GeneResolver, curated the data, conducted the experiments, analyzed the results, and drafted the manuscript. K.M. contributed to the methodology, system design, interpretation of the results, and critical revision of the manuscript. M.S.M. contributed to the conceptualization, supervision, project administration, funding acquisition, interpretation of the results, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The figures were prepared with visual-design assistance from Claude Design. The authors independently verified all data, labels, relationships, and scientific interpretations shown in the figures and take full responsibility for the final content.

Supplementary Material

File name: Supplementary Material 1
File format: PDF (.pdf)
Title: Synthetic de-identification test-set construction.
Description: Detailed procedure and generation prompt used to construct the 100-note synthetic clinical de-identification test set.

File name: Supplementary Material 2
File format: PDF (.pdf)
Title: Silver-standard etiology-classification dataset construction.
Description: Detailed procedures, retrieval strategy, screening criteria, and prompts used to construct and verify the 220-case silver-standard etiology-classification dataset from published case reports.

References

[1] Kim S, Zhou Y, Guo Y, Xiao C, Zheng K. Applying natural language processing and large language models to clinical notes for phenotyping and diagnosing rare diseases: a systematic review. Journal of the American Medical Informatics Association. 2026;33(6):1225–1235. https://doi.org/10.1093/jamia/ocag045.

[2] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics. 2008;83(5):610–615.

[3] Smedley D, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature Protocols. 2015;10(12):2004–2015.

[4] Robinson PN, et al. Interpretable clinical genomics with a likelihood ratio paradigm (LIRICAL). The American Journal of Human Genetics. 2020;107(3):403–417.

[5] Zhao M, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. NAR Genomics and Bioinformatics. 2020;2(2):lqaa032.

[6] Mao D, Liu C, Wang L, Ai-Ouran R, Deisseroth C, Pasupuleti S, et al. AI-MARRVEL – A Knowledge-Driven AI System for Diagnosing Mendelian Disorders. NEJM AI. 2024;1(5). https://doi.org/10.1056/aioa2300009.

[7] Zhao W, Wu C, Fan Y, Qiu P, Zhang X, Sun Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature. 2026;651:775–784.

[8] Huang K, Zhang S, Wang H, Qu Y, Lu Y, Roohani Y, et al. Biomni: A General-Purpose Biomedical AI Agent. bioRxiv. 2025;Preprint. https://doi.org/10.1101/2025.05.30.656746.

[9] Buniello A, Suveges D, Cruz-Castillo C, Bernal Llinares M, Cornu H, Lopez I, et al. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. Nucleic Acids Research. 2025;53(D1):D1467–D1475. https://doi.org/10.1093/nar/gkae1128.

[10] Cerezo M, Sollis E, Ji Y, Lewis E, Abid A, Bircan KO, et al. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Research. 2025;53(D1):D998–D1005. https://doi.org/10.1093/nar/gkae1070.

[11] Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Human Mutation. 2012;33(5):803–808. https://doi.org/10.1002/humu.22078.

[12] Groza T, Gration D, Baynam G, Robinson PN. FastHPOCR: pragmatic, fast, and accurate concept recognition using the human phenotype ontology. Bioinformatics. 2024 07;40(7):btae406. https://doi.org/10.1093/bioinformatics/btae406. https://academic.oup.com/bioinformatics/article-pdf/40/7/btae406/58463767/btae406.pdf.

[13] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17(1):122. https://doi.org/10.1186/s13059-016-0974-4.

[14] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. 2015;17(5):405–423. https://doi.org/10.1038/gim.2015.30.

[15] Abou Tayoun AN, Pesaran T, DiStefano MT, Oza A, Rehm HL, Biesecker LG, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Human Mutation. 2018;39(11):1517–1524. https://doi.org/10.1002/humu.23626.

[16] Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. Nature Genetics. 2014;46(9):944–950. https://doi.org/10.1038/ng.3050.

[17] Pejaver V, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. The American Journal of Human Genetics. 2022;109(12):2163–2177. https://doi.org/10.1016/j.ajhg.2022.10.013.

[18] Jaganathan K, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535–548.e24. https://doi.org/10.1016/j.cell.2018.12.0

15.

[19] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Research. 2025;53(D1):D609–D617. https://doi.org/10.1093/nar/gkae1010.

[20] Milacic M, et al. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Research. 2024;52(D1):D672–D678. https://doi.org/10.1093/nar/gkad1025.

[21] Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–443. https://doi.org/10.1038/s41586-020-2308-7.

[22] Consortium TEP. Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Research. 2015 01;43(D1):D1042–D1048. https://doi.org/10.1093/nar/gku1061. https://academic.oup.com/nar/article-pdf/43/D1/D1042/7316204/gku1061.pdf.

[23] Lobo M, Lamurias A, Couto FM. Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules. BioMed Research International. 2017;2017:8565739. https://doi.org/10.1155/2017/8565739.

[24] Bridges Y, de Souza V, Cortes KG, Haendel M, Harris NL, Korn DR, et al. Towards a standard benchmark for phenotype-driven variant and gene prioritisation algorithms: PhEval – Phenotypic inference Evaluation framework. bioRxiv. 2025;Preprint. https://doi.org/10.1101/2024.06.13.598672.

[25] Danis D, et al. A corpus of GA4GH phenopackets: case-level phenotyping for genomic diagnostics and discovery. Human Genetics and Genomics Advances. 2024;6(1):100371. https://doi.org/10.1016/j.xhgg.2024.100371.

[26] Yauy K, Duforet-Frebourg N, Testard Q, Beaumeunier S, Audoux J, Simard B, et al. Learning phenotypic patterns in genetic diseases by symptom interaction modeling. medRxiv. 2022;Preprint. https://doi.org/10.1101/2022.07.29.22278181.

[27] Deelen P, et al. Improving the diagnostic yield of exome- and genome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis (GADO). Nature Communications. 2019;10:2837.

[28] Chambon PJ, Wu C, Steinkamp JM, Adleberg J, Cook TS, Langlotz CP. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. Journal of the American Medical Informatics Association. 2022 11;Ocac219. https://doi.org/10.1093/jamia/ocac219. https://academic.oup.com/jamia/advance-article-pdf/doi/10.1093/jamia/ocac219/47220191/ocac219.pdf.

[29] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:190711692. 2019;https://doi.org/10.48550/arXiv.1907.11692.

[30] OpenAI.: openai/privacy-filter. Hugging Face. https://huggingface.co/openai/privacy-filter.

[31] OpenMed.: OpenMed: open models and resources for healthcare NLP. Hugging Face. https://huggingface.co/OpenMed.

[32] OpenMed.: OpenMed/privacy-filter-multilingual: multilingual fine-grained PII extraction across 16 languages and 54 categories. Hugging Face. https://huggingface.co/OpenMed/privacy-filter-multilingual.

[33] AI4Privacy.: AI4Privacy PII Masking Datasets. Hugging Face. https://huggingface.co/ai4privacy.

[34] Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine. 2019;21(7):1585–1593. https://doi.org/10.1038/s41436-018-0381-1.

[35] Vichentijevikj I, Mishev K, Simjanoska Misheva M. Prompt-to-Pill: Multi-Agent Drug Discovery and Clinical Simulation Pipeline. Bioinformatics Advances. 2026 01;6(1):vbaf323. https://doi.org/10.1093/bioadv/vbaf323. https://academic.oup.com/bioinformaticsadvances/article-pdf/6/1/vbaf323/68211589/vbaf323.pdf.

[36] Simjanoska Misheva M, Shahpaski D, Dobreva J, Bukovec D, Gjorgjioski B, Nikolov M, et al. AI Act Compliance Within the MyHealth@EU Framework: Tutorial. Journal of Medical Internet Research. 2025;27:e81184. https://doi.org/10.2196/81184.

[37] Gjorgjioski B, Bukovec D, Vichentijevikj I, Kitanovski I, Mishev K, Simjanoska Misheva M. Multi-Agent Readiness Scoring Methodology in Bioinformatics Domain. Future Internet. 2026;18(8):409. https://doi.org/10.3390/fi18080409.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial1.pdf
- SupplementaryMaterial2.pdf