AI Model Card: DiagAI HPO v1.0

Compliance Framework: ISO/IEC 42001:2023

1. General Information

1.1. Basic Information

AI System Name	DiagAI HPO
Model identifier	AIMC-05
Model Version	V1.0
Release Date	15 December 2025 (SeqOne Platform 1.50)
Model Type	AI-Powered Clinical Phenotype Analysis and HPO Term Recommendation System
Development Team	SeqOne — GAA
Maintenance Status	Active
Related Products	SeqOne Platform Integrates Phenogenius, DiagAI Short List, DiagAI Score
EU AI Act Classification	High-risk AI system (Article 6 & Annex 3.5b) – Ai intended to be used as a safety component in the management and operation of critical digital infrastructure, or as medical device

1.2. Licensing and Access

License Type	Proprietary
Access Level	Restricted to authorized clinical and research users
Terms of Use	Subject to SeqOne Platform Terms of Service and Clinical Use Agreements

2. Intended Use & Constraints

2.1. Intended Purpose

DiagAI HPO is an AI-powered clinical decision support tool designed to assist healthcare professionals in translating clinical observations into standardized Human Phenotype Ontology (HPO) terms. The system uses natural language processing (NLP), Large Language Models, and machine learning to analyze clinical notes, patient descriptions, and medical records to suggest appropriate HPO terms, improving the accuracy and completeness of phenotype documentation for genetic diagnosis.

Primary Functions:

- Extract phenotypic information from free-text clinical notes
- Recommend appropriate HPO terms from clinical descriptions
- Identify missing phenotype dimensions to improve diagnostic accuracy
- Standardize phenotype documentation across users and institutions
- Interactive phenotype refinement and expansion
- Support differential diagnosis through phenotype analysis
- Handle several input languages
- Ensure HPO terms are returned in their canonical form

2.2. Target Users

Primary Users:

Healthcare professionals responsible for routine clinical genomic analysis: geneticists, bioinformaticians, biologists, engineers, and NGS platform technicians.

Indirect Users:

- Patients receiving diagnostic information
- Healthcare systems managing patient care
- Regulatory authorities overseeing device performance

2.3. Out-of-Scope Uses

- Standalone clinical diagnosis without expert review
- Automated phenotype-based diagnosis without genetic testing
- Direct patient-facing phenotype interpretation
- Replacement of clinical examination and assessment
- Billing or coding purposes (ICD codes should be assigned separately)
- Legal or forensic applications
- Population health screening without clinical context
- Symptom checker for self-diagnosis (consumer use)
- Automated triage or medical decision-making
- Phenotype extraction in other language than English or French

2.4. Deployment Constraints

Runtime Environment

Parameter	Value
Language	Python
Interface	HTTP API calls to biologic and AWS Bedrock
Upstream integration	SeqOne Platform — Feeds Phenogenius, DiagAI Short List, DiagAI Score
Model storage	AWS Bedrock

Input Format

Free-text clinical notes (plain text, RTF, PDF)

Output Format

- HPO term list with IDs (e.g. HP:0001250 — Seizures),
- Rationale (source text fragment),
- Presence/absence indication,
- Export in JSON format.

3. Model Architecture

3.1. Model Architecture

DiagAI HPO extracts Human Phenotype Ontology (HPO) terms from a free-text clinical note. The extraction logic lives in an internal library, text2hpo, called in the same process.

No model is trained, fine-tuned, or adapted. The system orchestrates a general-purpose language model, a pinned copy of the HPO ontology, and deterministic pre-processing and post-processing around both.

Parameters

Parameter	Value
Model type	Multi-component NLP pipeline combining a foundation model with deterministic ontology matching
Library / Runtime	AWS Bedrock, Converse API
Language	Python 3.11
Language model	Mistral Large (mistral.mistral-large-2402-v1:0)
Context window	32 000 tokens
Inference temperature	0.0001
Maximum output tokens	8 000, reduced to 300 for single-word notes
Reference ontology	HPO, release 2026-01-08, deployed as a fixed file asset
De-identification engine	Microsoft Presidio, French and English
Ontology matching	Exact and fuzzy string matching over HPO terms and their synonyms
Hyperparameter optimization	None
Probability calibration	None. The system produces no score to calibrate
Explainability	Each returned term carries the exact source fragment that produced it, an explicit present or absent flag, and an identifier that exists in the pinned ontology release

Service interface

The service exposes one operation. It accepts a single free-text clinical note and returns the phenotypes found in it. The note is limited to 4 000 characters and must not be empty.

The response separates phenotypes asserted as present from phenotypes explicitly negated in the text. Each entry carries three fields: the HPO identifier, the official HPO term, and the fragment of the source note that produced the match.

The service is stateless. Neither the clinical note nor the result is persisted. The ontology, the de-identification engine, and the model client are loaded once when the application starts, then shared by every request.

The endpoint returns a complete result or an error. It never returns a partial list. Three failure classes are distinguished: input rejected by validation, text left without usable content after de-identification, and any other internal failure. The last class returns a generic message and exposes no internal detail.

Processing pipeline

Stage 1, preparation. The note is lowercased, then de-identified. Person names and email addresses are detected and replaced. If de-identification leaves no alphabetic content, the request is rejected rather than sent to the model. A single-word note is expanded into a short sentence and the output budget is reduced, because short inputs otherwise produce verbose and unreliable answers.

Stage 2, phenotype extraction. Accents and special characters are normalized. The de-identified note is submitted to the language model with a fixed instruction set and one worked French example. The model returns candidate phenotypes, each made of a source fragment, an English HPO term name, and a present or absent judgement. No identifier is requested at this stage.

Stage 3, candidate grounding. This stage uses no model. For each term the model proposed, the pinned ontology is searched for plausible identifiers. A term matching an official HPO name or synonym exactly is resolved immediately, without any further model call. A term with several plausible candidates carries that candidate list into stage 4. A term with no candidate at all is marked for removal.

Stage 4, identifier confirmation. Terms with several candidates are grouped into small batches and submitted to the language model a second time. The model does not generate identifiers freely. It selects one identifier per term from the candidate list supplied with that term, and must answer in the order received. Batches are capped by term count and by total candidate volume, so the request stays within the model's reliable working range. A term with no acceptable candidate receives a reserved sentinel identifier. If the model returns the wrong number of answers, the response is padded or truncated so that answers stay aligned with their terms.

Stage 5, verification and normalization. Every confirmed identifier is checked against the ontology. The official name and synonyms of that identifier are compared with the term the model proposed, and the entry is discarded when the two do not resemble each other closely enough. This step removes hallucinated and sentinel identifiers. Surviving entries have their term replaced by the official HPO label, so the output uses ontology wording rather than model wording. Entries sharing an identifier are reduced to one.

Control over model output and explainability

The language model is used twice, with two different levels of freedom. The first call is generative and proposes terms. The second call is constrained and only chooses among candidates that the ontology itself produced.

Every identifier returned to the caller therefore exists in the pinned ontology release, and has passed a similarity check against that ontology. A term the model invented, or mapped to an unrelated identifier, is dropped rather than returned.

In the graphical User interface, the raw text that led to the extraction of an HPO term is highlighted to explain and help the user verify that the presence of the term is valid. This is done using the annotations filled by the extraction object and sent to the frontend.

Determinism and versioning

Inference runs at a temperature close to zero, so the same note yields the same terms across calls, subject to the variability of the hosted model. The ontology is a version-pinned file deployed with the service, so a change of HPO release is an explicit deployment event.

The system does not cache results, does not learn from usage, and has no feedback loop. Behaviour changes only through a new library version, a new ontology file, or a new model identifier.

4. Training Data & Provenance

⚠ *No custom training was performed for DiagAI HPO. The system relies on Mistral Large, a pre-trained foundational Large Language Model provided by Mistral AI. SeqOne does not possess the training data for this base model. Augmentation and few-shot prompting are applied at inference time.*

4.1. Training Data

Not applicable.

4.2. Validation Datasets

DiagAI HPO is validated on one external clinical cohort. The cohort was shared by Hopital Tenon for the evaluation of phenotype-driven gene ranking in nephrology.

Characteristic	Value
Patients	44, one record each
Clinical domain	Nephrology, genetic kidney disease
Diagnostic outcome	One confirmed causal gene per patient. One patient carries a second gene
Distinct genes	18: COL4A4 (10 patients), PKD1 (6), COL4A3 (5), MUC1 (4), COL4A5 (4), HNF1B (2), TTR (2), TTC21B (2), and ten genes with one patient each
Distinct diseases	17
Zygosity	Heterozygous 34, homozygous 4, hemizygous 1, not recorded 5
Clinical note used for validation	The free-text note recorded in REDCap. Present for all 44 patients. 224 to 1 365 characters, median 612
Manual HPO annotation	A clinician annotated each patient. The final set holds 202 terms, 80 distinct, 1 to 13 per patient, median 4
Reference file	ef915476c222b0f3c7965d85e1422734960b5adf6bc506e68733ec4ab1e6a6f9 patient_notes.tsv

The cohort is not used for training, prompt design, or ontology selection. It is used once, to measure the effect of the extracted terms on gene ranking.

5. Performance Metrics & Evaluation

5.1. Performance Metrics

Evaluation Datasets

Dataset	Description
Test set DS-16	44 patients from Hopital Tenon with a confirmed diagnostic gene and a free-text clinical note. The set measures the effect of the extracted phenotypes on the rank of the true gene.

Evaluation method

DiagAI HPO extracts the phenotypes of each note. The phenotypes asserted as present are submitted to PhenoGenius, which ranks all genes in its phenotype-gene matrix. The metric is the rank of the confirmed diagnostic gene. For the one patient with two genes, the better of the two ranks is kept. A rank of 1 means the causal gene was the first proposal.

HPO Term Extraction and Gene Ranking Metrics

All 44 notes were processed and all 44 diagnostic genes received a rank.

Metric	Test Set
Top-1 rank	2 / 44
Top-10 rank	15 / 44
Top-50 rank	35 / 44
Top-100 rank	39 / 44
Top-200 rank	42 / 44
Median rank	17
Mean rank	40
Phenotypes asserted present per note	Mean 8.3, range 4 to 16
Phenotypes asserted absent per note	Mean 1.32 range 0 to 8
Extraction latency per note	Mean 13.9 s

Rank distribution: 2 patients at rank 1, 13 at ranks 2 to 10, 20 at ranks 11 to 50, 4 at ranks 51 to 100, 3 at ranks 101 to 200, 2 beyond 200.

6. AI Risk Assessment & Impact

6.1. Identified Risks

ID	Hazard Description	Consequence / Harm	P	S	Risk Level	Mitigation Strategy	Residual Risk
RSK-76	Explainability Transparency - User interface is not understood / Misunderstanding of scores, rankings, or evidence. - AI-powered "black box" logic provides clinical scores without clear biological reasoning.	Incorrect results - Misinterpretation of data	2	2	Low	Mitigated by comprehensive training and clear UI design (in-product tooltips and contextual help on each score component; user manual section dedicated to the interpretation of evidence and rankings.) The system shall display a DiagAI score summary, with the different score sub-components : UP2, Phenogenius, Family Pattern, Inheritance compatibility, Quality Checks, PanelApp. (SR-616)	Low
RSK-81	Bias in predictions / ancestry disparities Reduced performance for patients from ancestry groups underrepresented in training data Bias in predictions: underperformance on under-represented genomic populations or variant types	Disparate performance across populations; health equity concerns; potential missed diagnoses in certain groups	3	2	Medium	Training data limitation documented; performance stratified by ancestry where possible; user guidance provided Fairness subgroup evaluation ; ethics review prior to deployment ; Direct ancestry-based fairness evaluation is not feasible due to absence of ancestry metadata in raw data source (ClinVar) Addition of new datasets to improve the AI model	Low
RSK-84	Over-reliance on AI / clinical misinterpretation Predictions used without appropriate clinical context or expert review	Human fails to catch AI errors — incorrect decisions acted upon uncritically; missed pathogenic variant; erroneous clinical decision; potential patient harm	3	2	Medium	Mandatory human-in-the-loop validation; clear UI warnings; system labeled as decision support; predictions accompanied by confidence scores	Low
RSK-90	Feature extraction Missing Critical Phenotypes DiagAI HPO: Missing Critical Phenotypes: incomplete extraction from clinical notes	Incomplete phenotype reduces downstream diagnostic accuracy; important discriminating features missed	2	2	Low	Detected terms are highlighted in the original input text which allows the verification of the output. A manual validation of the exact terms is mandatory, if the user is missing terms, a manual addition interface is available.	Low
RSK-91	Model Drift / Knowledge Base Obsolescence Performance degradation as new genes, diseases, and variants are discovered after model training, and as genomic databases and evidence evolve post-training	Reduced accuracy; outdated pathogenicity predictions; missed clinically relevant variants	2	2	Low	Major knowledge base updates in our bioresources. Some DB (ClinVar, OMIM) are also plugged to directly use the latest information from the pipeline, minimizing the risk of missing an important annotation. Models are retrained at least annually	Low

ID	Hazard Description	Consequence / Harm	P	S	Risk Level	Mitigation Strategy	Residual Risk
RSK-92	Feature extraction Negation errors DiagAI HPO: Negation errors: present vs. absent features misinterpreted	Opposite phenotype recorded; downstream tools receive incorrect information; may lead to incorrect diagnosis	2	2	Low	Model is explicitly asked to detect negative phenotypes. A manual validation of the exact terms is mandatory, if the user is missing terms, a manual addition interface is available.	Low
RSK-93	Phenotype Description Bias / Performance varies with clinical documentation quality or language	Inconsistent results across institutions	2	2	Low	User training on HPO term selection ; guided phenotype entry interface Additionnaly, DiagAI HPO extracts hpo terms by including "lay person terms" from hpo.jax Interface displayed best matching HPO terms.	Low
RSK-94	Out-of-distribution (OOD) input / scope creep AI model used outside validated use cases	Unreliable outputs on unseen data patterns without detection	2	2	Low	Models are gatekept in variant specific pipelines.	Low
RSK-104	Data management & data quality Training data used to build or retrain a model is maliciously altered (data poisoning), a model artifact/weights file is substituted or tampered with without authorization, or crafted ("adversarial") input files are used to deliberately manipulate a model's output.	Incorrect results - Misinterpretation of data	3	2	Medium	Model artifacts are version-controlled and integrity-checked before deployment; access to training pipelines and datasets is restricted and logged in accordance with the ISMS (ISO/IEC 27001:2022/A1:2024); the release process requires validation of model provenance before a new version reaches production (see RSK-089).	Low
RSK-108	Post-production monitoring & performance degradation over time No structured post-market monitoring of ML-specific performance indicators (data drift, concept drift) is in place beyond general PMS activities. Gradual degradation of classification/scoring accuracy in the field is not detected in a timely manner	Incorrect results - Misinterpretation of data	2	2	Low	Post-market surveillance activities are documented in the PMS Procedure, and will be followed on a quarterly basis.	Low

P = Probability (1=Improbable — 1 in 10 000 000; 2=Occasional — 1 in 1 000; 3=Probable — 1 in 100; 4=Frequent — 1 in 1) | S = Severity (1=Negligible — no/negligible patient risk; 2=Minor — temporary impairment, incorrect/missing variant; 3=Major — injury requiring medical intervention, false diagnostic; 4=Critical — death, permanent impairment) | Risk Level derived from SEQ-011 Risk Acceptability Matrix.

6.2. Impact Assessment Result

Question	Answer
High-impact individual decisions?	Yes — clinical decision support for genetic diagnosis
Sensitive data processed?	Yes — clinical notes containing patient health information (PHI); de-identified in training
Overall risk level	High (clinical diagnostic aid; patient safety implications)
Referenced AIIA document	AIIA-001_DiagAI

7. Ethical Considerations

7.1. Fairness and Bias

Identified Biases

- Rare phenotype under-representation: Features with <10 training examples have reduced accuracy → ultra-rare diseases may be under-served.
- Language and cultural bias: Training data primarily English from Western medical settings → reduced performance for non-English or non-Western terminology.
- Medical specialty bias: Clinical genetics notes vs. other specialties → better performance on genetics notes than general paediatrics.

Mitigation Strategies

- Language and cultural adaptation: require adequate training data per language; cultural consultation for terminology; transparent documentation of limitations.
- Performance monitoring by subgroup: track accuracy by practice setting; user satisfaction by specialty; language performance if multilingual; regular equity audits.
- Transparent communication: clear documentation of limitations; performance characteristics by context; appropriate expectations setting.

7.2. Transparency and Explainability

Interpretability Methods (XAI)

XAI Method	Scope	Availability in Production	Limitations
Extraction Highlighting (Text-to-Term)	Text segments that triggered each HPO suggestion highlighted	Included in every output — surfaced in UI	Only as reliable as the LLM extraction quality

Human Oversight

Oversight Principle	Implementation
Oversight Model	Human-in-the-loop: all AI-suggested HPO terms must be explicitly reviewed and confirmed by a qualified clinical operator before being used in genetic analysis
Decision Authority	The AI system provides ranked HPO term suggestions with confidence scores. Final phenotype selection and clinical accountability remain with the human operator at all times
Override Mechanism	Operators can accept, modify, or reject any AI suggestion. Manual addition of HPO terms is always possible. All decisions are logged with timestamp and user ID
Uncertainty Communication	Low-confidence suggestions are visually flagged; ambiguous negations require explicit confirmation; rare-term alerts notify the operator of reduced model reliability
Operator Notification	Completeness gaps (missing body systems) and quality scores are surfaced proactively to prompt comprehensive phenotype documentation
Audit Trail	All AI outputs, confidence scores, and subsequent human decisions are logged in the SeqOne Platform
Failure Mode Handling	If the AI system is unavailable or returns an error, manual HPO term entry remains available — no silent degraded mode

7.3. Privacy & Security

Legal Framework	GDPR (EU); ISO 27001; ISO 27701; HIPAA (US);
Encryption	AES-256 at rest; TLS 1.3 in transit
Security Certification	ISO 27001 certified; Annual penetration testing
Access Control	Role-based access control (RBAC); multi-factor authentication (MFA) required ; principle of least privilege
Data Pseudonymization	All patient data pseudonymized before use in model development and validation
Audit Logging	All data access events logged (HIPAA compliant)
Data Retention	GDPR Article 17 right to erasure implemented; automated deletion after retention period; secure disposal procedures

7.4. Clinical Ethics

- System designed to augment — not replace — clinical judgment; no automated clinical decisions without human oversight
- Clear delineation between AI suggestions and clinical determinations
- Informed consent required: clear communication of AI involvement, explanation of limitations,
- Genetic counseling must be involved in result disclosure to patients

8. Monitoring and Maintenance

8.1. Known Limitations

Limitation	Impact	Communication Required?
Complex compound descriptions: multiple phenotypes in a single sentence may be challenging	Co-occurring vs. causally related features may not be correctly separated	Yes — user guidance to review compound sentences
Negation edge cases: double negatives, ambiguous phrasing, family history vs. patient features	Opposite phenotype may be recorded if negation misinterpreted	Yes — explicit visual confirmation required for negated features
Context-dependent terms: age-appropriate vs. abnormal; severity modifiers	Severity/temporal modifiers may be missed; macrocephaly (relative to age) may be misinterpreted	Yes — inform operators; prompt manual verification
Temporal aspects: age of onset and progressive vs. static distinction	Age-of-onset extraction accuracy limited; developmental milestone timing challenging	Yes — temporal information flagged for user review
Language and cultural bias: performance best for English	Non-English clinical notes yield reduced accuracy; cultural variations in phenotype description	Yes — explicitly flagged; route to careful human review

8.2. Documentation Update Triggers

Document	Update Trigger
Model Card (this document)	Within 30 days of any model changes
User Manual	Within 14 days of UI or functionality changes
Risk Assessment	Quarterly review; updated as needed
Performance Reports	Monthly generation

Retraining trigger criteria:

- Performance drop > 5% vs. validated baseline on reference cohort
- Major ClinVar update with significant change in training label distribution
- New gnomAD release with substantially revised population frequencies
- Serious incident linked to UP² output contributing to a significant adverse outcome
- Regulatory or policy framework change affecting this AI system
- Scheduled annual review — regardless of drift or incident

8.3. Decommissioning Plan

- Formal decommissioning decision approved by AI System Owner and top management
- All model artifacts (weights, training data, configurations) archived securely per data retention policy
- All integrations and API endpoints disabled; operators notified with 5 days' notice
- Final version of Model Card and all AIMS records archived in the document management system
- Post-decommissioning review: lessons learned captured for future model development

9. Governance & Accountability

9.1. Regulatory Compliance

AI Management	ISO/IEC 42001:2023 — AI Management System
Medical Device	EU IVDR (2017/746) — In Vitro Diagnostic Regulation
Data Protection	GDPR (EU); HIPAA (US); PIPEDA (Canada); national data protection laws as applicable
Clinical Standards	ACMG/AMP Guidelines for Sequence Variant Interpretation; ACMG Recommendations for Reporting Secondary Findings
Clinical AI	EU AI Act — High-risk AI system classification assessment ongoing
Digital Health	NICE Evidence Standards Framework for Digital Health Technologies
US FDA	FDA guidance on Clinical Decision Support Software

9.2. Approval & Accountability

Role	Responsibility
Model Developer / AI Team	Technical development, matrix construction, internal validation, maintenance
Data Scientist	Algorithm design, data source integration, performance analysis
Product Manager	Minor update approval; user requirement management
Compliance Officer	Regulatory compliance oversight; GDPR/IVDR coordination
Medical/Scientific Advisor	Clinical performance interpretation; guideline alignment

10. Version History

Version	Description of changes	Approval date
1	Initial Model Card creation	2025-12-15
2	Update to risks (§6), the datasets (§4) and performance metrics (§5)	See approval date

Table of signatures	Name / Function	Date	Signature
Written by	Nicolas Duforet Datascience manager	2026-09-25	Nicolas Duforet
Reviewed by	Nicolas Philippe Chief Scientific Officer	2026-09-25	Nicolas PHILIPPE
Approved by	Amélie Martinez AIGO	2026-09-25	Amélie Martinez