An Agentic Approach to Phenotype Mapping from Rare Disease Surveys

Jipeng Di* MATT@GENEIAL.COM
Julie Renee Vaughn* JULIE@GENEIAL.COM
Joshua Proulx JOSH@GENEIAL.COM
Sadie Nordstrand SADIE@GENEIAL.COM
Bryce Daines BRYCE@GENEIAL.COM
Katrisa Madeline Ward KATRISA@GENEIAL.COM
Geneial, USA

Philip J. Lupo PHILIP.LUPO@EMORY.EDU
Emory University School of Medicine, USA

Jianhong Hu JIANHONG.HU@BCM.EDU
Mullai Murugan MULLAI.MURUGAN@BCM.EDU
Baylor College of Medicine, USA

Adam W. Hansen ADAM@GENEIAL.COM
Geneial, USA

Abstract

Rare disease patients worldwide often experience years-long diagnostic delays, in part due to fragmented and unstructured phenotypic information. Patient-reported surveys provide valuable insights but are typically unstructured and hard to integrate with structured data. We present GenOMA (Geneial Ontology Mapping Agent), a Large Language Model (LLM) agent built on the LangGraph framework and integrated with a Unified Medical Language System (UMLS) API for precise extraction and ontology mapping of phenotypic terms. Using a modular, node-based architecture for context-aware extraction, iterative refinement, candidate ranking, and semantic validation, GenOMA maps data to standardized Human Phenotype Ontology (HPO) codes without local ontology deployment. We evaluate GenOMA on the question fields of three rare disease surveys, mapping them to HPO terms, and compare its performance with other leading methods. On the Xia-Gibbs Syndrome (XGS) Registry, GenOMA achieved 0.92 accuracy, 0.94 precision, 0.97 recall, and 0.96 F1. On the Down Syndrome Phenotyping Acute Leukemia Study (DS-PALS) dataset, it obtained 0.92 accuracy, 0.93 precision, 0.98 recall, and 0.96 F1. Finally, on the GenomeConnect (GC) dataset, it obtained 0.91 accuracy, 0.91 precision, 1.0 recall, and 0.96 F1. In all tasks, GenOMA outperformed MetaMap, PhenoTagger, PhenoBERT, cTAKES, and GPT-5. These results show that GenOMA effectively converts unstructured survey data to structured phenotype information. To our knowledge, this is the first ontology mapping system specifically designed for patient-reported rare disease surveys, a critical but underexplored data modality.

Keywords: Rare disease; Agent; Phenotype extraction; Ontology mapping; Large language model; Unified Medical Language System; Human Phenotype Ontology

Data and Code Availability The GenomeConnect (GC) dataset comprises 187 simplified survey questions derived from the ClinGen GenomeConnect registry, provided as a small public example set for testing ontology mapping models. The Xia-Gibbs Syndrome (XGS) Registry and DS-PALS survey question datasets were obtained through agreements with their respective organizations and are not publicly available due to privacy restrictions. Access requests should be directed to the data owners.

The GenOMA agent code, including the LangGraph workflow, UMLS API integration, and Supplementary Materials is available on Github https://github.com/genealco/GenOMA-Geneial-Ontology-Mapping-Agent, along with the evaluation

* These authors contributed equally

scripts and documentation for reproducing the experiments.

Institutional Review Board (IRB) An IRB is not necessary because there is no use of human data. We only make use of generic survey questions rather than specific patient responses.

1. Introduction

Rare diseases in the United States are collectively defined as conditions affecting fewer than 200,000 individuals; more than 7,000 have been identified, and approximately 350 million people are affected worldwide (Chaudhary et al., 2025). Owing to the heterogeneity of these conditions and their largely nonspecific clinical presentations, rare diseases are easily mistaken for common disorders, leading to diagnostic delays that preclude timely access to effective treatment—one of the field's major global challenges (Garcelon et al., 2018). For each rare disease, comprehensive deployment of patient surveys and disease registries can efficiently capture high-quality patient information, offering a critical avenue for alleviating diagnostic difficulty (Cammel et al., 2020). However, to apply these data in downstream diagnostic and therapeutic workflows, the unstructured questionnaire text must first be mapped to standardized codes by biomedical informatics experts (Austin et al., 2018b; Liu et al., 2022). Among the available ontologies, the Human Phenotype Ontology (HPO) provides a comprehensive, hierarchical vocabulary that links observable patient traits to known disease–gene associations, enabling computational phenotype–genotype analysis and variant prioritization (Köhler et al., 2021). Mapping survey fields to HPO codes is therefore a critical step toward making patient-reported data interoperable with clinical and genomic databases, allowing integration with tools such as Exomiser and Phen2Gene that rely on standardized phenotype input (Javed et al., 2023). Given more than 7,000 rare diseases, and the rapid expansion of disease-specific questionnaires, a purely manual mapping pipeline is not scalable (Peng et al., 2021; Zhao et al., 2020). Moreover, because questionnaires are written in plain, patient-friendly language, existing batch term-extraction and mapping tools (e.g., MetaMap, PhenoTagger) were optimized for professional clinical narratives with explicitly stated single entities, so they perform suboptimally and cannot infer latent terms from whole-sentence context (Aronson and Lang, 2010; Luo et al., 2021). This gap underscores the need for more intelligent systems that can reason over context to recover implicit concepts (Schwab et al., 2023). Recent advances in large language models (LLMs) demonstrate impressive capabilities in contextual understanding and terminology recognition (Wu et al., 2024). Leveraging these strengths, we design prompt strategies that enable LLMs to extract latent phenotypic terms from complex survey text and, via our Unified Medical Language System (UMLS) database (Bodenreider, 2004) API wrapper, map the extracted terms to standardized codes for downstream diagnostic use. This approach substantially improves mapping efficiency, accelerates the operationalization of survey systems for rare-disease data capture and integration, and ultimately expedites diagnostic pathways for rare-disease patients.

Our main contributions are:

1. We built an LLM agent with LangGraph and the UMLS API, enabling strong understanding of survey text context, extraction of implicit terms, and conversion into standardized codes.

2. We evaluated the agent on three rare disease survey datasets: the GenomeConnect (GC) survey (Riggs et al., 2015), the Xia-Gibbs Syndrome (XGS) Registry (Jiang et al., 2018) and the DS-PALS Survey (Brown et al., 2019; Li et al., 2023), focusing on accurate and reliable term extraction and ontology mapping.

3. We compare our approach with MetaMap, PhenoBERT, PhenoTagger, cTAKES and GPT-5 prompted directly, without agent workflow or UMLS integration, demonstrating substantial improvements in context-aware mapping performance.

4. We introduce the first ontology mapping approach tailored specifically for rare disease survey data, showing that agentic LLM workflows enable reliable phenotype extraction from a neglected but critical data modality.

2. Background

2.1. Named Entity Recognition in Healthcare

Named entity recognition (NER) identifies and classifies biomedical entities from unstructured text (Wang et al., 2018). In biomedicine, it enables clinical

decision support, literature mining, and ontology-based reasoning (Luo et al., 2019). Early rule- and dictionary-based systems performed well in narrow contexts but lacked scalability and robustness (Funk et al., 2014). Machine learning models better capture context and semantics (Habibi et al., 2017), yet many still struggle with complex syntax and ambiguity—especially in patient-generated texts such as questionnaires, where both prompts and responses must be interpreted in context (Zhao et al., 2022). For rare diseases, questionnaires and registries are key sources of phenotypic information (Cao et al., 2024), extracting and mapping entities from them and integrating results with genomic data are essential for precise diagnosis and treatment (Austin et al., 2018a). However, no models are tailored to questionnaire-based entity extraction and ontology mapping. Developing efficient and accurate NER and ontology mapping approaches tailored to such data is therefore of substantial practical importance.

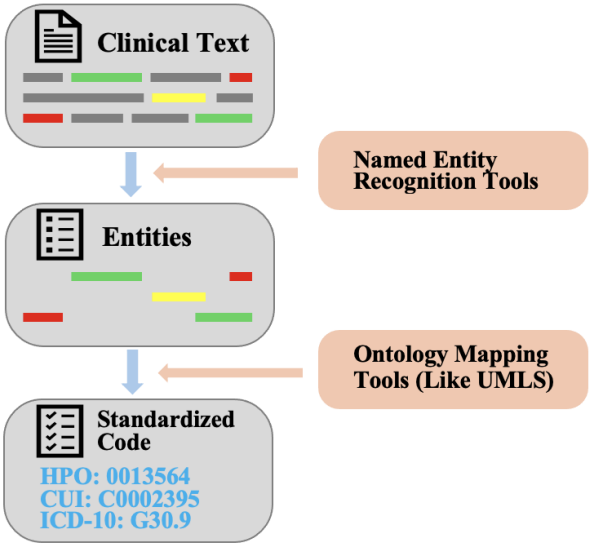

Figure 1: The main workflow of named entity recognition (NER) and ontology mapping.

2.2. Prior Work

NER and ontology mapping tools have been applied in the biomedical domain, spanning dictionary-based systems, machine learning models, and, more recently, LLMs. Most of these systems follow a two-stage pipeline: entities are identified or extracted from unstructured text, then entities are mapped to standardized vocabularies, including those in the UMLS Metathesaurus (Le et al., 2025), as shown in Figure 1.

Among these two stages, highly accurate entity recognition or extraction from unstructured text remains the most challenging component (Luo et al., 2020). There are several types of entity recognition or extraction and mapping methods that have been put into practical application. Dictionary-based approaches such as MetaMap and cTAKES use lexical resources and rules to map terms to the UMLS Metathesaurus; they require no labeled data and perform well on structured clinical narratives (Demner-Fushman et al., 2017; Lee et al., 2020). Machine learning–based systems—e.g., PhenoTagger, PhenoBERT—leverage contextual embeddings and supervised NLP to identify entities and map them to standardized vocabularies (e.g., HPO) via UMLS (Savova et al., 2010). For instance, PhenoBERT builds on BioBERT for biomedical NER with a subsequent UMLS mapping step (Feng et al., 2022). Yet these models often operate at the token/phrase level and still struggle with complex syntax and ambiguity in clinical narratives, especially patient-generated text. LLMs (e.g., GPT-5, Claude-3.5) offer stronger contextual reasoning and can extract terms from noisy or informal text; coupled with tools like the UMLS API, they combine flexible language understanding with standardized outputs (Chen et al., 2025). Remaining issues include computational cost, limited controllability, and prompt-sensitive variability (Andrew et al., 2024).

Overall, the field is shifting from static rules to more generalizable models, but current systems rely heavily on identifying and extracting individual entities within sentences. They are unable to fully understand the entire sentence and infer the underlying terms. Our agentic framework addresses this gap via context-aware extraction, iterative refinement, and semantic validation.

2.3. Our Approach

We built an intelligent agent with LangGraph, which represents LLM workflows as graphs of functional nodes, enabling fine-grained control over multi-step reasoning and tool use (LangChain, 2024). Key advantages:

- **Task specialization**: dedicated nodes for subtasks reduce prompt interference and improve accuracy/interpretability.
- **Workflow flexibility**: nodes can be added or rewired to meet task needs.
- **Prompt-level adaptability**: node-specific prompts tailor behavior to datasets/domains.
- **No training/deployment**: direct LLM API calls remove local hosting/fine-tuning, lowering operational and compute costs.

3. Methodology

3.1. Agent Architecture Overview

GenOMA is a modular LLM agent built with LangGraph, orchestrating GPT-5/4o/4–powered, task-specific nodes in a graph pipeline. It comprises three stages: (1) term extraction—medical NER on patient survey text; (2) ontology mapping—querying the UMLS API for standardized concept identifiers; and (3) structural validation—refining outputs to ensure consistency and correctness.

Based on LangGraph's modular scalability, we optimize agent performance by refining node-level functions. The resulting workflow is shown in Figure 2. Several components are particularly critical to performance, including:

1. `Retry_with_llm_rewrite_node`, which can rewrite the input and retry failed mappings to improve coverage on difficult cases.
2. `Rank_mappings_node`, which can score and rank candidate ontology mappings, prioritizing those with higher semantic similarity and clinical relevance.
3. `Validate_mapping_node`, which validates the selected mapping and assigns a confidence score to ensure reliability.

The text data of the survey questions are used as the input of this agent. After passing through these nodes, the extracted terms and corresponding standardized mapping codes (such as HPO codes) can be output. Further information about the agent architecture can be found in Appendix A, while an ablation study of the nodes can be found in Appendix H.

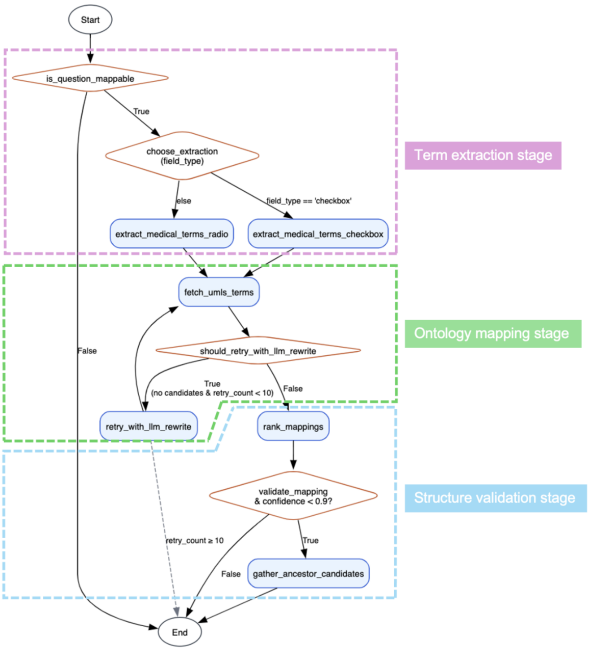

Figure 2: GenOMA agent workflow for phenotypic term extraction and ontology mapping from rare disease surveys. It includes three stages: term extraction (pink), ontology mapping with iterative retries (green), and structure validation with ranking and ancestor retrieval (blue).

3.2. Dataset Description

To evaluate the model's ability to understand text, we use three datasets of rare disease survey questions with increasing difficulty: the GenomeConnect (GC) dataset, the Xia-Gibbs Syndrome (XGS) Registry Survey Dataset, and the DS-PALS Survey Dataset. All datasets have been manually annotated by bioinformatics experts and serve as gold standards. GC contains 187 simplified mappable medical phrases, XGS contains 119 fully mappable phenotype question sentences, while DS-PALS contains 166 sentences, 50 of which are non-mappable questions, enabling evaluation of extraction and filtering in ontology mapping. Because our agent is designed primarily for survey question mapping, each mappable question in all datasets corresponds to exactly one standardized HPO term, reflecting the one-to-one design used in

most registry instruments. Further dataset information can be found in Appendix E.

3.3. Performance Evaluation

Using UMLS CUI to provide a unique semantic identifier for each concept, which enables accurate and standardized cross-model comparison, we unify the model output into CUI codes and evaluate performance against human-labeled data. We compute precision, recall, and F1 using the standard confusion-matrix formulation, and analyze the composition of mispredictions. Further details of the evaluation methods are included in Appendix F. We compare GenOMA to the strongest baseline per dataset using McNemar's test on item-level correctness. For accuracy, precision, and recall, we report 95% Wilson confidence intervals.

4. Results

4.1. Overall Mapping Accuracy

To evaluate model performance, we measured the overall mapping accuracy of six ontology mapping systems (GenOMA, PhenoTagger, PhenoBERT, MetaMap, cTAKES, and GPT-5) on the GC, XGS and DS-PALS survey datasets. Across all surveys, GenOMA achieved consistent accuracy above 0.91, outperforming dictionary-based (MetaMap, cTAKES) and machine learning baselines (PhenoTagger, PhenoBERT) (Figure 3). While other models exhibited significant performance fluctuations when faced with datasets with varying levels of input text complexity, GenOMA remained stable, demonstrating its robustness to dataset variations.

4.2. Precision, Recall and F1 Score

To provide a comprehensive assessment of model performance, we evaluated precision, recall, and F1 score on these three datasets. As shown in Figure 4, across the three datasets (GC, XGS, and DS-PALS), GenOMA consistently achieved the best performance, with precision, recall, and F1 scores all above 0.93 and reaching up to 1.00 on GC, demonstrating strong robustness and stability across evaluation settings. In comparison, PhenoTagger and PhenoBERT showed relatively high precision but lower recall, indicating that they tend to miss valid terms. By contrast, MetaMap and GPT-5 favored recall over precision—both achieving high recall (up to 0.91–0.90) but suffering from much lower precision (0.33–0.62), leading to frequent false positives. cTAKES consistently underperformed, with moderate recall but low precision, resulting in the weakest overall F1 scores. Taken together, these results demonstrate that GenOMA provides a superior balance between sensitivity and specificity across diverse datasets, while other models exhibit dataset-dependent weaknesses—either conservative behavior that sacrifices recall (PhenoTagger, PhenoBERT) or overly permissive outputs that compromise precision (MetaMap, GPT-5).

4.3. Term vs. No Term Mapping Accuracy

To better interpret the practical differences behind similar precision and recall scores, we analyzed the error patterns of each model's outputs. Specifically, we categorized the errors into three types:

- Type IA: Wrong mapping (FP, gold-standard term present)—The item has a gold-standard ontology term, but the model outputs a different term/code.
- Type IB: Spurious mapping (FP, gold-standard term absent)—The item has no gold-standard ontology term, but the model outputs a term/code.
- Type II: Omission (FN, gold-standard term present)—The item has a gold-standard ontology term, but the model produces no output.

Across all three datasets (GC, XGS, and DS-PALS), as shown in Figure 5, GenOMA consistently demonstrated the highest proportion of correct mappings with minimal errors. On the GC dataset, GenOMA correctly mapped 91% of items, with only 9% errors. On the XGS dataset, it reached 92% correct predictions, with just 2% omissions and 6% mispredictions, while on the DS-PALS dataset it achieved 93% accuracy on items requiring mapping and 96% on those requiring no mapping. This indicates that GenOMA not only excels in recognizing valid terms but also effectively avoids unnecessary mappings when no terms are present, while baseline models show systematic weaknesses—either omitting too many valid terms (PhenoTagger, PhenoBERT) or over-generating false mappings (MetaMap, GPT-5), with cTAKES falling in between, further error analysis can be found in Appendix G.

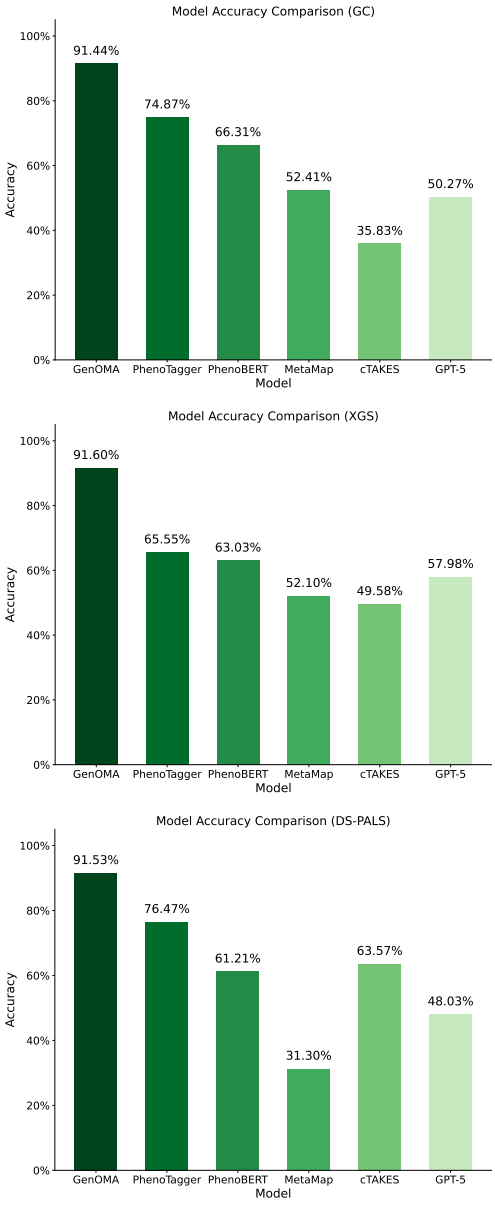

Figure 3: Overall mapping accuracy of six ontology mapping systems on the GC (top), XGS (middle), and DS-PALS (bottom) survey datasets.

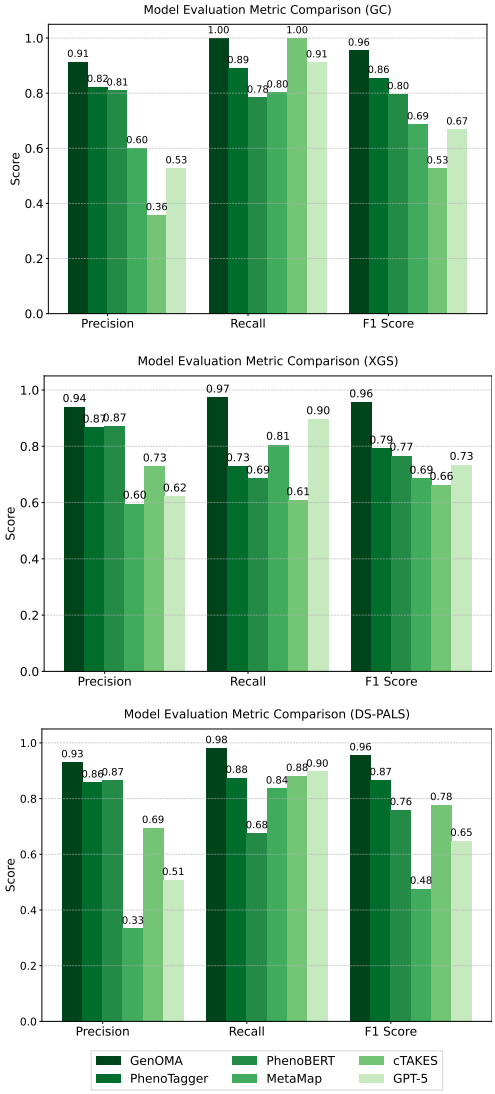

Figure 4: Precision, recall, and F1 scores of six models on the GC (top), XGS (middle), and DS-PALS (bottom) survey datasets.

4.4. Statistical Analysis of Results

We assess pairwise significance using McNemar's test on item-level correctness (GenOMA vs. PhenoTagger—the highest performing baseline) for each dataset. On the GC dataset, GenOMA achieved a net accuracy gain of 16.6% over PhenoTagger ($p<0.0001$). On the XGS dataset, the gain was 26.1% ($p<0.0001$), and on the DS-PALS dataset, the gain

Figure 5: Prediction breakdown by model on the GC (top), XGS (middle), and DS-PALS (bottom) datasets, showing the proportions of correct predictions, wrong predictions (Type I), and omissions (Type II).

was 14.7% (p = 0.0009). These results demonstrate that GenOMA provides statistically significant improvements across all all datasets, as shown in Table 1.

Table 1: Pairwise McNemar's test comparing GenOMA with PhenoTagger on item-level correctness for GC, XGS and DS-PALS. The table reports N, discordant counts (b, c), net accuracy change ((b–c)/N), and p-values.

Dataset	N	b	c	**Δ acc.**	**McNemar** p
GC	187	38	7	+0.166	< 0.0001
XGS	119	34	3	+0.261	< 0.0001
DS-PALS	116	21	4	+0.147	0.0009

We also report 95% Wilson confidence intervals for proportions (accuracy, precision, recall) computed on item counts, shown in Table 2. Across all datasets, GenOMA consistently outperforms PhenoTagger on accuracy, precision, and recall. The confidence intervals for accuracy and recall show clear separation, indicating robust improvements rather than chance variation, while precision also favors GenOMA with only minor overlap.

Table 2: 95% Wilson confidence intervals (CI) for accuracy, precision, and recall by dataset and model.

Dataset	**Model**	**Acc. (95% CI)**	**Prec. (95% CI)**	**Rec. (95% CI)**
GC	GenOMA	0.91 (0.87–0.95)	0.91 (0.87–0.95)	1.00 (0.98–1.00)
	PhenoTagger	0.75 (0.68–0.81)	0.82 (0.76–0.87)	0.89 (0.83–0.93)
XGS	GenOMA	0.92 (0.85–0.95)	0.93 (0.87–0.97)	0.98 (0.94–1.00)
	PhenoTagger	0.66 (0.57–0.74)	0.86 (0.78–0.91)	0.88 (0.80–0.93)
DS-PALS	GenOMA	0.92 (0.85–0.95)	0.93 (0.87–0.97)	0.98 (0.94–1.00)
	PhenoTagger	0.77 (0.68–0.83)	0.86 (0.78–0.91)	0.88 (0.80–0.93)

Taken together, GenOMA recovers more true positives with fewer misses without a corresponding rise in false positives, and paired comparisons (McNemar's test) further support the significance of these gains across cohorts.

4.5. Semantic Similarity Analysis

Even when model predictions do not exactly match the gold-standard terms, it is important to assess

how semantically close these "errors" are, as this reflects the model's conceptual understanding of phenotype meaning. To this end, we incorporated a GPT-based semantic evaluation function that assigns similarity scores to each prediction, providing a complementary measure of semantic appropriateness for ambiguous or partially correct cases. A score of 10 represents a perfect match. As shown in Figure 6, GenOMA's incorrect predictions were, on average, semantically closer to the gold-standard terms than those produced by other baseline models, demonstrating stronger ontology-level reasoning and contextual alignment.

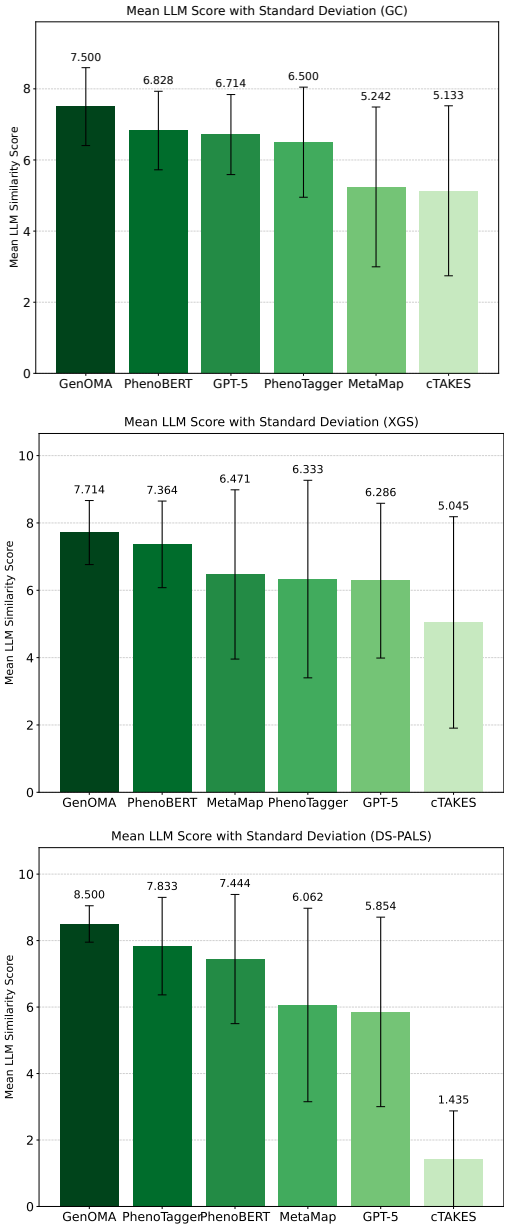

Figure 6: LLM similarity scores for incorrectly mapped terms across six models on the GC, XGS and DS-PALS dataset, with error bars representing standard deviation. Higher scores indicate that errors were semantically closer to the ground truth.

5. Discussion

5.1. Performance Comparison

On all datasets, our GenOMA agent consistently outperformed all baseline systems (MetaMap, PhenoTagger, PhenoBERT, cTAKES, and GPT-5) across all evaluation metrics, including overall mapping accuracy, precision, recall, and F1 score. Further error analysis revealed that GenOMA had the lowest proportions of both omissions ("no predictions") and wrong mapping, even rare spurious mapping, underscoring its strong ability to understand context and accurately extract terminology. Other models (MetaMap, PhenoTagger, PhenoBERT, cTAKES) focus narrowly on surface entities, often missing implied or context-dependent terms. They are not suitable for questionnaire-type texts that require a full understanding of the context. Although GPT-5 and other state-of-the-art LLMs have powerful semantic understanding capabilities, they need to be combined with a standardized ontology tool (e.g., the UMLS API), otherwise they will generate erroneous information due to hallucinations.

5.2. Advantages of the Agentic Approach

The superior performance of the GenOMA agent across all datasets can be attributed to several key design advantages inherent in its node-based architecture:

1. **Context-Aware Term Extraction:** The agent processes both question stems and associated labels/options, enabling inference of intended concepts even when implicit. This is particularly valuable for patient-facing surveys, where complex medical terms are often replaced with simpler language.

2. **Iterative Refinement via Retry Mechanism:** The `retry_with_llm_rewrite_node` reformulates unmapped terms iteratively, avoiding missing term information.

3. **Integrated Ranking and Validation:** The `rank_mappings_node` and `validate_mapping_node` jointly score, filter, and confirm candidate mappings, enhancing the precision.

4. **Flexible Task Adaptation via LLM Node:** The LLM-based node design allows the agent to adapt to different terminology extraction tasks simply by adjusting prompts, providing high flexibility without retraining.

Collectively, these design advantages allow the agent to achieve both high recall and high precision in ontology mapping for rare disease studies, outperforming existing baseline systems while maintaining flexibility, scalability, and ease of deployment.

5.3. Limitations and Challenges

Despite achieving high accuracy, our approach has several limitations:

- **Output Variability** – The quality of term extraction is dependent on the LLM's performance and prompt design (Andrew et al., 2024; Hu et al., 2024). The same input may not always yield identical results, leading to potential inconsistencies.

- **Computational Cost** – Although local model deployment is not required, LLM API calls are slower and more costly than lightweight dictionary-based methods, which may impact scalability for large-scale datasets.

- **Data Security Considerations** – While local deployment of LLMs can mitigate privacy concerns, the highest-quality outputs typically require calls to the latest cloud-based models, which introduces potential risks of data leakage.

- **Dataset Accessibility and Scale** – Two of the three evaluation datasets cannot yet be made public due to privacy and institutional collaboration constraints. We are actively working with data partners to make these datasets available and will update our public GitHub repository accordingly. In addition, the relatively small size of the test datasets may lead to slightly inflated recall scores and should be interpreted with caution.

- **Future Extension** – The current architecture focuses on single-term mapping from rare disease survey questions. Extending this framework to handle long-form medical narratives with multiple phenotypic mentions represents a key direction for future work.

6. Conclusion

Our study presents an LLM-based ontology-mapping agent built on LangGraph that uses node-based contextual reasoning, iterative refinement, and semantic ranking to extract and standardize phenotype terms from rare-disease survey data. Across three representative datasets—the GC, XGS and DS-PALS Survey—our agent consistently outperformed established systems (MetaMap, PhenoTagger, PhenoBERT, cTAKES, GPT-5), achieving over 91% accuracy and balanced precision–recall profiles with F1 scores around 0.95 across all tasks. Gains in mapping accuracy commonly exceeded 15–40 percentage points. Error analyses showed fewer omissions and mis-mappings, minimizing redundant outputs without sacrificing coverage.

The modular LangGraph architecture affords flexibility for rapid adaptation to new datasets, ontologies, and clinical domains; nevertheless, challenges remain, including variability in LLM outputs, computational cost, ontology disambiguation under ambiguity, and privacy considerations for cloud-based inference. Future work will target improved prompting and controllability, hybrid symbolic–neural inference, efficient local deployment, and privacy-preserving mechanisms. Our system is specifically applied to patient-facing rare disease surveys and leverages contextual understanding to extract and map terms, showing consistently stronger performance than dictionary- or conventional NLP-based tools in our evaluations. By enabling accurate, scalable, and context-aware phenotype extraction, GenOMA can accelerate rare disease diagnostics and strengthen the integration of patient-reported outcomes into research and care.

Acknowledgments

We'd like to acknowledge the support of the Xia Gibbs Research Foundation and Emory University, as well as funding from ARPA-H Biomedical Data Fabric and NHLBI BioData Catalyst (Agreement No. 1OT3HL147154) with RENCI.

References

J. J. Andrew, M. Vincent, A. Burgun, and N. Garcelon. Evaluating LLMs for temporal entity extraction from pediatric clinical text in rare diseases context. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 145–152, 2024.

A. R. Aronson and F. M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

C. P. Austin, C. M. Cutillo, L. P. L. Lau, A. H. Jonker, A. Rath, D. Julkowska, D. Thomson, S. F. Terry, B. de Montleau, D. Ardigo, V. Hivert, J. Harris, H. Lochmuller, G. Baynam, P. Kaufmann, and D. Taruscio. Future of rare diseases research 2017–2027: An IRDiRC perspective. *Clinical and Translational Science*, 11(1):21–27, 2018a.

C. P. Austin et al. Future of rare diseases research 2017–2027: An IRDiRC perspective. *Clinical and Translational Science*, 11(1):21–27, 2018b.

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue): D267–D270, 2004. doi: 10.1093/nar/gkh061. URL https://doi.org/10.1093/nar/gkh061.

A. L. Brown, A. J. de Smith, V. U. Gant, W. Yang, M. E. Scheurer, K. M. Walsh, et al. Inherited genetic susceptibility to acute lymphoblastic leukemia in down syndrome. *Blood*, 134(15):1227–1237, 2019.

S. A. Cammel, M. S. De Vos, D. van Soest, et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC Medical Informatics and Decision Making*, 2020.

L. Cao et al. AutoRD: an automatic end-to-end system for constructing rare disease knowledge graphs. *JMIR Medical Informatics*, 12:e60665, 2024.

A. Chaudhary et al. Rare diseases: a comprehensive literature review and future directions. *Journal of Rare Diseases*, 4:33, 2025. doi: 10.1007/s44162-025-00099-6. URL https://doi.org/10.1007/s44162-025-00099-6.

Q. Chen et al. Benchmarking large language models for biomedical natural language processing. *Nature Communications*, 16:56989, 2025.

D. Demner-Fushman, W. J. Rogers, and A. R. Aronson. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844, 2017.

Y. Feng, L. Qi, and W. Tian. PhenoBERT: A combined deep-learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1269–1277, 2022.

C. Funk, W. A. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, and K. Verspoor. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15:59, 2014.

N. Garcelon et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. *Orphanet Journal of Rare Diseases*, 13:85, 2018.

M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

Y. Hu, J. Zeng, L. Wang, and H. Xu. Improving large language models for clinical named entity recognition. *Journal of the American Medical Informatics Association*, 31(9):1812–1822, 2024.

Zeeshan Javed, Ahmed Malik, Ali Khan, Naif Alghamdi, Ahmad A. Alghamdi, Haroon Ahmed, Hafiza Sadia, Anwar Khan, Masood Lali, Xiang Zhang, Xia Li, Hong Xu, and Xiaoqiang

Liu. Human phenotype ontology-based deep phenotyping for rare disease diagnosis: Recent advances and future directions. *Frontiers in Genetics*, 14:1112345, 2023. doi: 10.3389/fgene.2023.1112345. URL https://doi.org/10.3389/fgene.2023.1112345.

Y. Jiang, M. F. Wangler, A. L. McGuire, J. R. Lupski, J. E. Posey, M. M. Khayat, D. R. Murdock, L. Sanchez-Pulido, C. P. Ponting, F. Xia, J. V. Hunter, Q. Meng, M. Murugan, R. A. Gibbs, et al. The phenotypic spectrum of xia-gibbs syndrome. *American Journal of Medical Genetics Part A*, 176 (8):1703–1715, 2018.

Sebastian Köhler, Michael Gargano, and Matentzoglu. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 2021. doi: 10.1093/nar/gkaa1043. URL https://doi.org/10.1093/nar/gkaa1043.

LangChain. Langgraph: Build stateful, multi-actor applications with LLMs. Available at: https://www.langchain.com/langgraph (accessed 16 August 2025), 2024.

T. Le, M. Phan, and P. Do. Leveraging semantic type dependencies for clinical named entity recognition. arXiv preprint, 2025.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Z. Li, T.-C. Chang, J. J. Junco, M. Devidas, Y. Li, W. Yang, et al. Genomic landscape of down syndrome–associated acute lymphoblastic leukemia. *Blood*, 142(2):172–184, 2023.

C. Liu et al. OARD: Open annotations for rare diseases and their genes. *American Journal of Human Genetics*, 2022.

L. Luo, X. Ji, J. Yang, Y. Li, F. Liu, and J. Xu. A review of biomedical named entity recognition. *Frontiers in Genetics*, 11:612, 2020.

Y. Luo, Y. Xin, and H. Xu. Biomedical named entity recognition: A survey of approaches and challenges. *Information*, 10(3):109, 2019.

Y. Luo et al. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *BMC Bioinformatics*, 22(1):402, 2021.

Y. Peng et al. Artificial intelligence in rare diseases: a systematic review. *Orphanet Journal of Rare Diseases*, 16:138, 2021.

Erin R. Riggs, Karen E. Wain, Diana Riethmaier, Michelle Savage, Barbara Smith-Packard, W. Andrew Faucett, Jamie Hinton, David T. Miller, Swaroop Aradhya, John Belmont, and Heidi L. Rehm. Genomeconnect: engaging patients and families to improve genomic knowledge. *Genetics in Medicine*, 19(7):854–862, 2015. doi: 10.1038/gim.2015.166. URL https://pubmed.ncbi.nlm.nih.gov/26178529/.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

J. D. Schwab, S. D. Werle, R. Hühne, H. Spohn, U. X. Kaisers, and H. A. Kestler. The necessity of interoperability to uncover the full potential of digital health devices. *JMIR Medical Informatics*, 11:e49301, 2023.

Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, and P. He. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. *Journal of Biomedical Informatics*, 92: 103133, 2018.

J. Wu et al. Hybrid frameworks combining dictionary-based nlp and large language models for rare disease identification. *BMC Medical Informatics and Decision Making*, 24:87, 2024.

S. Zhao, H. Liu, Y. Guo, X. Jiang, and Y. Peng. Challenges and opportunities in patient-generated health data: a survey. *Artificial Intelligence in Medicine*, 127:102276, 2022.

Y. Zhao et al. Challenges and opportunities in machine learning for rare diseases. *Frontiers in Genetics*, 11:638, 2020.

Appendix A. GenOMA Ontology Mapping Pipeline Details

Here we describe the implementation details of the GenOMA agent, focusing on the workflow diagram and the role of each LangGraph node and its contribution to the ontology mapping workflow. Each node performs a distinct subtask, and together they form a modular pipeline for robust phenotype term extraction and mapping.

- `is_question_mappable_node`: Filters out questions that do not require ontology mapping, ensuring that irrelevant items are excluded early in the workflow.
- `extract_medical_terms_node`: Extracts candidate medical terms directly from the input text using the LLM.
- `fetch_umls_terms_node`: Maps extracted terms to standardized ontology codes (e.g., HPO) via the UMLS API.
- `retry_with_llm_rewrite_node`: Rewrites the input and retries failed mappings to improve coverage on difficult cases.
- `rank_mappings_node`: Scores and ranks candidate ontology mappings, prioritizing those with higher semantic similarity and clinical relevance.
- `validate_mapping_node`: Validates the selected mapping and assigns a confidence score to ensure reliability.
- `gather_ancestor_candidates_node`: Retrieves ancestor terms of a candidate mapping from the ontology hierarchy, enabling more generalized or fallback mappings when exact matches are not available.

Appendix B. Output Consistency and Validation

To ensure consistent and reproducible outputs across all nodes involving LLM calls, we set the temperature parameter to 0 for relatively deterministic decoding and enforce strict output schemas. In addition, the `fetch_umls_terms_node` leverages the standardized UMLS ontology service for term normalization, reducing potential semantic drift and improving cross-dataset stability. These design choices collectively enhance output reliability while maintaining transparency regarding any residual model variability.

The `validate_mapping_node` performs the final semantic verification of each candidate mapping. It evaluates whether the ontology term is clinically appropriate and semantically consistent with the question context, then assigns a confidence score ranging from 50% to 100% (90–100%: accurate match; 70–85%: acceptable match; 50–65%: partial match). If the score falls below 0.9, the workflow automatically triggers the `gather_ancestor_candidates_node` to refine the output using broader ancestor terms. This confidence-driven feedback mechanism, derived from explicit LLM reasoning rules, ensures that final mappings are both semantically valid and clinically meaningful.

Appendix C. From Questions to Clinical Text: Toward Multi-Term and Long-Text Mapping

While this study focuses on instrument harmonization—mapping survey question fields to Human Phenotype Ontology (HPO) codes to improve data sharing across rare disease registries—the proposed agentic pipeline is input-agnostic. It can also be applied to other forms of text, not just structured survey questions. In future work, we plan to extend this framework to handle free-text patient responses and other unstructured narratives such as clinical summaries and online community posts, which often include diverse and variable phenotype descriptions.

For multi-term extraction tasks, the architecture can be expanded by adding a multi-term extraction head to the existing agent. The extracted terms can then be processed one by one to obtain their standardized ontology terms and corresponding HPO codes. This minimal-change pathway—preprocessing → existing agent → aggregation—shows how the current single-term workflow can be adapted for long-text and multi-phenotype inputs, supporting broader applications such as registry integration and patient-generated data analysis.

Appendix D. Prompt Design and Optimization

Prompt design played a central role in achieving consistent and high-quality outputs across heterogeneous survey datasets. Our design process followed two guiding principles: robustness and constraint. Questionnaire structures can vary widely across registries, so prompts needed to generalize across diverse syntactic and semantic patterns. At the same time, healthcare text requires strict compliance with defined output formats to ensure interpretability and reproducibility, even when the LLM operates deterministically (temperature = 0).

Each prompt was constructed using a modular template that includes: (i) a clear role specification to define the model's function (e.g., "You are an ontology-mapping assistant"), (ii) an explicit input format outlining which parts of the question text are provided, (iii) focus guidance describing which linguistic features to attend to (e.g., disease symptoms vs. demographic context), (iv) a catalog of common edge cases with short examples and handling rules, and (v) a strict output schema ensuring structured and machine-readable results.

Prompt optimization was performed iteratively through error-driven refinement. We analyzed logical inconsistencies or hallucinated outputs, identified failure patterns (e.g., multi-term conjunctions joined by "and" or "/"), and refined instructions to encourage concept-level reasoning rather than surface extraction. Workflow-level adjustments—such as segmenting input text into labeled spans (title, question stem, options, notes)—further improved precision and stability.

All finalized prompt templates used in the GenOMA pipeline, including node-specific examples, are publicly available in our GitHub repository.

Appendix E. Dataset Descriptions

We evaluate and compare ontology mapping systems using three representative rare disease surveys: the GenomeConnect (GC) dataset, the Xia-Gibbs Syndrome (XGS) Registry Survey Dataset and the DS-PALS Survey Dataset. The latter two datasets are not publicly available, so we present their description details in Table 3.

A key difference between the datasets is that the DS-PALS dataset includes 50 items that require no ontology mapping. These non-mappable items serve to evaluate the models' capacity to accurately identify irrelevant content, thereby preventing redundant or spurious mappings.

Table 3: Comparison of the GenomeConnect (GC), Xia-Gibbs Syndrome (XGS), and DS-PALS survey datasets.

Aspect	Description
Source (GC)	ClinGen GenomeConnect registry
Source (XGS)	Xia-Gibbs Syndrome advocacy group
Source (DS-PALS)	Down Syndrome organization
Annotation (GC)	Published annotated with HPO codes
Annotation (XGS)	Manually annotated with HPO codes
Annotation (DS-PALS)	Manually annotated with HPO codes (after filtering)
# Questions (GC)	187 simplified survey questions
# Questions (XGS)	119 survey questions
# Questions (DS-PALS)	166 items (50 non-mappable)
Ground Truth (GC)	Each phrase linked to an HPO term
Ground Truth (XGS)	Each question linked to an HPO term
Ground Truth (DS-PALS)	116 with HPO mappings; 50 require no mapping
Evaluation Role (GC)	Provides simple testbed for ontology mapping
Evaluation Role (XGS)	Tests capturing of critical phenotypes
Evaluation Role (DS-PALS)	Tests extraction and filtering of non-mappable items

These datasets were selected for their diverse symptom descriptions and relevance to real-world patient-reported outcomes. They provide a realistic and efficient testbed for evaluating the robustness and adaptability of biomedical concept extraction and ontology mapping tools.

Appendix F. Definition Details of Evaluation and Comparison Methods

To evaluate the performance of our ontology mapping agent, we compared it with five other widely recognized medical terminology ontology mapping models (PhenoTagger, PhenoBERT, MetaMap, cTAKES, and GPT-5) on the XGS Registry Survey and the DS-PALS Survey.

We employed standard named entity recognition (NER) evaluation metrics:

- **True Positive (TP):** The model predicts the correct HPO code when a gold-standard code exists.
- **False Positive (FP):** (1) The model predicts an incorrect code when a gold-standard code exists, or (2) the model predicts any code when the gold standard is empty.

- **False Negative (FN):** The model predicts nothing when a gold-standard code exists.
- **True Negative (TN):** The model predicts nothing when the gold standard is empty. TNs are excluded from precision, recall, F1, and mapping accuracy, as they are not informative in ontology extraction tasks.
- **Precision:** Precision $= \frac{TP}{TP+FP}$.
- **Recall:** Recall $= \frac{TP}{TP+FN}$.
- **F1 score:** Harmonic mean of precision and recall, F1 $= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$.
- **Mapping accuracy:** Accuracy $= \frac{TP}{TP+FP+FN}$, i.e., the proportion of correct predictions among all cases where either a gold-standard code exists or the model makes a prediction.

Together, these strategies enabled both quantitative and qualitative comparisons across systems, capturing accuracy as well as semantic fidelity and adaptability.

Appendix G. Error Analysis

Missed detections often occur when identifying the target concept requires broader contextual understanding rather than relying solely on explicit keywords. In patient surveys, many phenotypes are implied rather than directly stated. Models that rely primarily on surface-level matching, such as MetaMap and cTAKES, frequently overlook these cases. For example, the question "Difficulty making friends outside of the immediate family" implies the phenotype "Lack of peer relationships", which was successfully extracted by GenOMA but missed by most baselines. GPT-5 sometimes generated related but imprecise alternatives (e.g., "Impaired social interactions"), reflecting partial contextual understanding but insufficient clinical alignment. See Table 4 for more examples.

Incorrect extractions arise when descriptive biomedical terms are misinterpreted as target concepts. In many survey items, explanatory context introduces simplified terms (e.g., "duodenal atresia" or "stenosis") to help respondents understand the question, while the intended clinical target is broader (e.g., "Abnormal duodenum morphology"). As shown in Table 5, GenOMA correctly captured the intended diagnostic targets, while other models often returned descriptive surrogates. PhenoTagger and PhenoBERT occasionally conflated descriptions with targets, MetaMap and cTAKES tended to default to surface terms, and ChatGPT5 frequently produced literal but incomplete matches (e.g., "Anal atresia" instead of "Abnormal anus morphology").

Appendix H. Key Nodes Ablation Study

To assess the contribution of key components, we performed node ablations of four key nodes: the `Retry_with_llm_rewrite_node`, `Rank_mappings_node`, `Validate_mapping_node` and `Gather_ancestor_candidates_node` modules on the XGS dataset. As shown in Figure 7, removing any module degraded performance, with the most pronounced drop when the rank node was omitted (accuracy decreased to 0.56), reflecting failure to disambiguate among multiple UMLS candidates and a resulting loss in precision. Eliminating the retry node chiefly reduced recall: after an initial mapping failure the agent produced more null predictions; with retry enabled, the agent tends to emit a single candidate even when uncertain, improving recall at the expense of precision. When the ancestor-gathering node was removed, accuracy dropped to 0.65, indicating its importance in supplying semantically related hierarchical candidates that aid disambiguation and validation. Overall, the results highlight complementary roles—rank for precise disambiguation, retry for coverage, and validate for filtering—that together sustain accuracy.

Appendix I. Computation and Cost Estimation

To assess the practical efficiency of the GenOMA agent, we estimated token consumption and corresponding API costs across datasets of different sizes. The computation cost depends primarily on the number of survey questions processed and the model tiering strategy used within the LangGraph architecture. Only the node requiring deep semantic understanding and ontology refinement uses the most recent model (GPT-5), while non-critical nodes (e.g., mapping-necessity checks) use a lighter model (GPT-4). This approach maintains consistent accuracy while significantly reducing total token usage.

Table 4: Examples of implied phenotypes in survey questions. GenOMA can capture both explicit and implicit terms, while other models often miss implicit cases or return partial matches.

Question	True	GenOMA	PhenoTagger	PhenoBERT	MetaMap	cTAKES	GPT-5
Problems with heartbeat rhythm requiring pacemaker	Arrhythmia	Arrhythmia	None	None	Rhythm heartbeat	Cardiac arrhythmia	–
Bone marrow transplant?	History of bone marrow transplant	History of bone marrow transplant	None	None	None	None	Bone marrow transplantation
Difficulty making friends outside family	Lack of peer relationships	Lack of peer relationships	None	None	None	None	Impaired social interactions
Glasses or contacts?	Abnormality of vision	Abnormality of vision	None	None	None	None	Refractive error

Table 5: Examples where explanatory context in questions led some models to misidentify the target term. GenOMA correctly maps to the intended clinical entity, while others often match the descriptive context instead.

Question	True	GenOMA	PhenoTagger	PhenoBERT	MetaMap	cTAKES	GPT-5
Normal EEG?	Abnormal EEG	Abnormal EEG	None	None	EEG normal	Normal EEG	Abnormal EEG
Gross motor skill issues	Gross motor impairment	Gross motor impairment	Gross motor skills	Difficulties with gross motor skills	Has difficulty doing	None	Delayed gross motor development
Duodenal atresia/stenosis/web	Abnormal duodenum morphology	Abnormal duodenum morphology	Duodenal atresia	Duodenal atresia	Stenosis	Duodenal atresia	Duodenal atresia
Anal atresia/stenosis	Abnormal anus morphology	Abnormal anus morphology	Anal atresia	Anal atresia/stenosis	Stenosis/atresia	Anal atresia	Anal atresia

Table 6: Estimated token usage and processing costs for different dataset sizes under the GPT-5/GPT-4 tiered setup.

Dataset Size	Estimated Tokens	Approximate Cost (USD)	Notes
1 question	2.6 K – 3.0 K	0.07 – 0.09	Single inference run
100 question	260 K – 300	7 – 9	Small-scale evaluation
250 question	650 K – 750 K	17 – 22	Medium-sized benchmark
500 question	1.3 M – 1.5 M	35 – 45	Large-scale deployment test

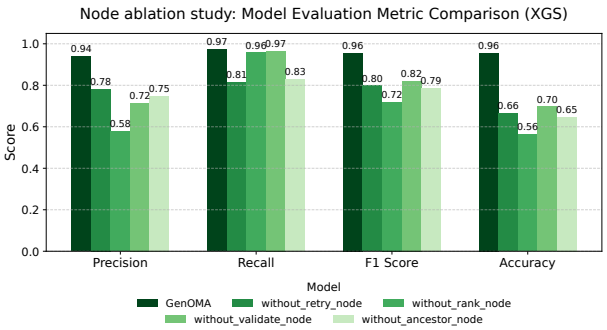

Figure 7: Node ablation on XGS, the precision, recall, and F1 score of the model after ablation of different nodes

For each input question, the agent typically consumes 2.6K – 3.0K tokens, corresponding to an average cost of \$0.07 – \$0.09 USD per question. Table 6 summarizes the expected computation cost as a function of dataset size.

Looking ahead, we plan to support on-premise and local deployments for environments with strict data-privacy requirements. The current trade-off is clear: the strongest local models generally require high-end hardware, whereas lighter open-weight models do not yet match commercial-grade performance. As open biomedical LLMs continue to improve, we will explore hybrid or fully local inference modes to complement the current API-based path.

Appendix J. MCP Server Integration

To support seamless interaction between the language model and external services, our system optionally supports integration with the Model Context Pro-

tocol (MCP)—a structured tool interface layer developed to standardize how LLMs invoke server-side APIs. The MCP layer enables organic interaction by exposing tools such as the UMLS mapping API through formalized input/output schemas and functional descriptions. We choose not to integrate the MCP layer into our workflow due to inconsistent tool calling and the need for a more structured, modular system. However, other teams have integrated it successfully to enhance agent reasoning and modularity.