A Versatile Multi-Modal Agent for Rare Disease Diagnosis and Risk Gene Prioritization

Anonymous ACL submission

Abstract

Accurate and timely diagnosis is essential for effective treatment, particularly in the context of rare diseases. However, current diagnostic workflows often lead to prolonged assessment times and low accuracy. To address these limitations, we introduce Hygieia, an adaptive AI agent system designed to support precision disease diagnosis by integrating diverse data sources, including phenotypic features, genetic profiles, and clinical records. Hygieia features a router-based and knowledge-enhanced framework that mitigates hallucination and tailors diagnostic strategies to different disease categories. Notably, it prioritizes risk-related genomic factors for rare diseases and provides confidence scores to assist clinical decision-making. We conducted a comprehensive evaluation demonstrating that Hygieia achieves state-of-the-art performance across multiple diagnostic benchmarks. In collaboration with clinical experts from Yale School of Medicine and Duke-NUS Medical School, we further validated its practical utility by showing (1) Hygieia's superior diagnostic performance compared to physicians and (2) its effectiveness in assisting clinicians with medical records for handling real-world cases. Our findings indicate that Hygieia not only enhances diagnostic accuracy and interpretability but also significantly reduces clinician workload, highlighting its potential as a valuable tool in clinical decision support systems.

1 Introduction

Rare diseases are defined as conditions affecting fewer than 1 in 2,000 individuals, affecting over 300 million patients worldwide (Valdez et al., 2016; Nguengang Wakap et al., 2020; Paul, 2013; Jonker et al., 2024; Zhao et al., 2026). Moreover, the diagnosis and test recommendation of rare diseases are very challenging. Diagnosing rare diseases based on conventional medical methods typically takes 4 to 5 years (Ghosh, 2025), which is known as "diagnostic odyssey". One reason is that the phenotypes of rare diseases can sometimes be difficult to distinguish directly from common diseases. This also increases the likelihood of misdiagnosis by physicians (Dong et al., 2020).

To overcome the challenges mentioned above, researchers have begun to collect genotypes, phenotypes, and diagnosis plans for rare diseases, to enrich our understanding of these diseases and to design a better diagnosis and treatment plan. These multimodal datasets provide valuable research materials for disease diagnosis. Moreover, data-driven solutions, such as disease diagnosis methods based on data mining, deep learning, and advanced artificial intelligence (AI), have also garnered significant attention recently (Lee et al., 2022). Experts can train a model using the aforementioned diagnostic data to predict diseases or prioritize disease risk genes, thereby improving diagnostic accuracy. However, such models often face shortcomings in terms of generalization and deployment (Rossi and El-Sayed, 2025). To enhance the model's versatility and accessibility for physicians, rare disease diagnostic tools based on Foundation Models (such as Large Language Models (LLMs) (Thirunavukarasu et al., 2023) and Visual Language Models (VLMS) (Liu et al., 2025a)) have also been developed. LLMs are pre-trained with a large-scale text corpus and can generalize into different tasks in natural language processing (NLP) with techniques of post-training. LLMs can process electronic health records (EHRs) from patients and make diagnoses accordingly (Sarker et al., 2024; Liu et al., 2024b), and multiple LLMs with different roles can also work together as an agent system (Tran et al., 2025; Du et al., 2025). Such an AI agent can make diagnoses by simulating the real scenarios, leveraging prior knowledge, and providing recommendations and suggestions for physicians as medical AI assistants (Liu et al., 2026a, 2025b).

Several AI-based tools have been developed for both common and rare disease diagnosis. For example, a general baseline for disease diagnosis will be prompting LLMs (Chen et al., 2024b). Researchers also consider developing AI agents for medical usage based on techniques such as knowledge-enhanced retrieval (Wang et al., 2025) as well as multi-agent communication (Dhatterwal et al., 2023; Chen et al., 2025). Although these models are interesting in design and have some clinical significance, their shortcomings are still quite evident. First, even advanced AI-based models used for rare disease diagnosis cannot distinguish between common and rare diseases, which is the basic step to avoid misdiagnosis (Supplementary Figure 1 (a)). This finding severely limits the applicability of rare disease diagnostic models. Second, due to the randomness that exists in the model training and inference, these models might not give consistent outputs based on the same input with different random seeds, which is also harmful for the trustworthy output (Supplementary Figure 1 (b)). Third, current AI methods only focus on diagnosis, but lack the necessary steps and capacities for result interpretation and discovery of important causal genes of rare diseases. Finally, current studies (Lee et al., 2022; Zhao et al., 2026; Liu et al., 2026b) does not directly address how these AI models can be applied in diagnostic scenarios and collaborate with physicians. Moreover, recent studies have shown that over 80% of rare diseases are influenced by genetic factors and can be passed on to the next generation (Zhao et al., 2026), but how these genetic factors are incorporated into diagnosis and how to infer disease-related genes based on clinical presentation have not been well researched in AI-based methods. Therefore, there is a critical need to design an AI model that can simultaneously diagnose different types of diseases and provide explanations for the diagnostic results.

In this manuscript, we introduce Hygieia, which is an AI agent for disease diagnosis and interpretation. Our model breaks down the diagnostic process into multiple stages, first determining the disease type, then designing distinct diagnostic approaches for common and rare diseases. Diagnosing common diseases is based on prompting LLMs. Meanwhile, due to the complexity of rare disease diagnosis, our agent utilizes multiple tools (such as website searching as well as patient retrieval) to leverage prior knowledge and make a decision. We have two innovations in the design of this agent. First, to resolve the inconsistency, our agent has a verifier to monitor the outputs of the main body of Hygieia and ensure the results converge. Second, to improve the transparency of using AI agents for making clinical decisions, we implement a framework with a reasoning trajectory and confidence estimation to help users understand and trust this workflow. Hygieia can accept multiple modalities or types of data as inputs, such as phenotype information, gene information, medical history, and other information. We also provided an table in Appendix B to distinguish Hygieia versus other AI agents focusing on (rare) disease diagnosis, and the unique components (case router, confidence estimation, and multi-task capacity) of Hygieia further enhance its novelty.

Overall, Hygieia can also interpret for the diagnosis results with trackable reasoning paths and prioritize disease-associated risk genes, which provide more informative feedback as references for helping physicians in making diagnoses. We invite physicians from Yale School of Medicine and DUKE-NUS Medical School to evaluate the contribution of Hygieia as an effective medical AI assistant, and explore new directions for the diagnosis of rare diseases at the age of AI and digital health.

2 Methods

Problem definition. Here we intend to design a specific agnetic system for disease-relevant analysis, including disease diagnosis and risk factor (such as gene) prioritization. Our agent system $\mathcal{A}()$ can take the following information as inputs, including task-specific prompts R, phenotypes P, functional and/or genetic information of risked genes G, and context information C. For the problem of diagnosis, this system makes inference with $O_i = \mathcal{A}(R_d, P_i, G_i, C_i)$ for the patient i, and the diagnosis result will be the output. Similarly, for risk gene prioritization, we replace the task-specific prompt R_d with R_g, and the output will be the prioritized gene. The model output not only contains the expected diseases or genes (in string), but also the confidence of making such a decision (in number).

Workflow of Hygieia. Our workflow contains four main stages, including *task-specific planning*, *information retrieval and integration*, and *self-reflection-based validation*, and *confidence estimation*. By default, all agents are implemented using GPT-5-chat after considering the trade-off between model performances and protection of patient pri-

vacy. We have discussed the ablation studies in the Methods section.

Regarding *task-specific planning*, we can integrate information proposed by known biomedical databases and divide the main task into several steps, based on Biomni (Huang et al., 2025). The agent will parse the input information first, and search the current tool base developed based on both tools in Biomni and newly implemented searching functions. It will then determine the correct tool for addressing the given task. We do not map phenotypes with HPO terms (Robinson and Mundlos, 2010) as we assume that our agent (based on advanced LLMs) already knows related information. Regarding the difference in diagnosing common and rare diseases, we train a classifier-based router to make more precise diagnoses and reduce cost based on a KNN classifier (Pedregosa et al., 2011). This stage is performed by the parser and the router component.

Regarding *information retrieval and integration*, we search the related information of phenotypes and possible gene functions based on web-searching tools, and the data sources including Google, Google Scholar, and PubMed (White, 2020). We also allow the advanced searching tool in LLMs such as the web-searching tools in GPT-4o (Hurst et al., 2024) and GPT-5 (OpenAI, 2025). We also consider searching the top $K = 5$ patients from known databases with diagnosis information as references. After collecting the necessary information, we will integrate the prior knowledge as inputs for the next component in this system. This task is finished by the knowledge-manager, web-searcher, and patient-retriever components.

Regarding *self-reflection-based validation*, we provide the diagnosis decision as well as methods for validation. The summary agent in our system will utilize the information from the previous two stages and generate a clinical decision. After making the decision based on the summary agent, we introduce our verification agent, which takes the prior knowledge and the output of the summary agent as input and validates whether the result is correct or incorrect (Yao et al., 2022). If the result is incorrect, the prompt used for the summary agent will update, and the summary agent will make a decision again, until the verification agent agrees with the decision or we reach the limit of tries. The algorithm of this system is shown in Algorithm 1, with the disease diagnosis task as an example. This task is finished by the summary-agent and verification-agent components.

Regarding *confidence estimation*, we refer to the method introduced by (Xiong et al., 2024), and we ask the summary agent s times to get s answers as well as s paired confidence lists $c_1, c_2, ..., c_s$. We average the confidence levels and use the major voting result from these s answers as the final decision. Therefore, the final confidence is $c_f = \frac{\sum_{k=1}^{s} c_k}{s}$. We have tried other methods, including summarization of logprobs, self-evaluation (Kadavath et al., 2022), and thinking-twice-before-answering (Li et al., 2024), but their performances are not good enough to represent the confidence. Since most of the advanced LLMs are closed-source and black-box models, other approaches used for open-source LLMs are not applicable. We have performed statistical tests to demonstrate that our current settings can help us calibrate the model outputs.

Ablation studies. In Supplementary Figures 6 (a) and (b), we show the contributions of adjusting the base models and the input information for these tasks. In both cases, changing the base models from GPT-4o to GPT-5 makes an obvious improvement, and incorporating more context, such as detailed phenotype descriptions, as well as having a verifier, can help Hygieia determine diseases and gene sets. We also find that providing disease information can help Hygieia rank genes, however, since in real clinical cases, physicians must know the results of the genetic tests and then make a diagnosis, we do not use this information as input for the prioritization of risk genes. Details of our method are shown in the Appendix section.

3 Results

Method overview. For each patient, we collect the annotations from phenotypes, medical records, and genetic test records as the input of Hygieia, and then make a decision based on the prior knowledge in medical research retrieved from the internet and databases. To initiate the workflow, Hygieia will first compute the probability of disease type based on a router, and then determine the most suitable pipeline for either common disease or rare disease. The pipeline used for rare disease diagnosis is more complicated and involves several agents, such as a knowledge-retrieval agent, an information-extraction agent, a summary agent, and a verification agent. Hygieia also infers the risk gene based on patient-level phenotypes or medical records, as an extra function for medical geneticists. Finally,

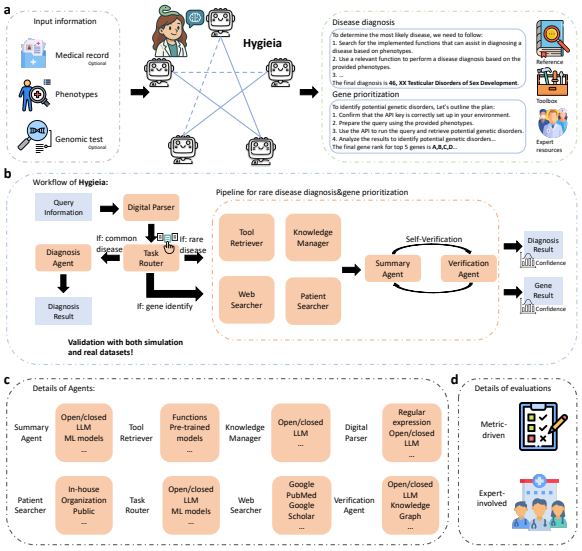


Figure 1: Overall pipeline of Hygieia. (a) Here we showcase how Hygieia can help physicians and clinicians working on two important problems in rare disease analysis, including diagnosis and risk gene prioritization. (b) The workflow of the AI Agent pipeline. We have multiple components, first routing the AI agents with correct models based on inferred disease type, and then providing diagnosis outcomes as well as confidence. (c) We provide the detailed information of each component in our AI agent. (d) Our evaluation criteria, including numerical evaluation and human evaluation, to support mimicking the scenario of clinical application.

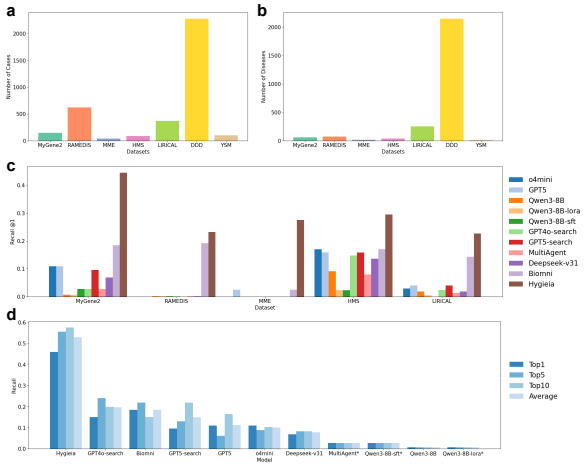


Figure 2: Benchmarking results for Hygieia in rare disease diagnosis. (a) Number of cases in our testing datasets. (b) Number of diseases in our testing datasets. (c) Top1 recall rate across different datasets. (d) Comparisons of different models with different recall rates in MyGene2.

we provide the confidence level of the model output through a majority voting approach. The input data types, workflow, and application scenarios of Hygieia are summarized in Figure 1.

Hygieia serves as a strong medical agent for disease diagnosis. We first demonstrate the strong capacity of Hygieia serving as a virtual physician, supported by validations conducted from various clinical datasets with different sources. The accuracy of disease classification is much higher than using random guess or prompting LLMs (Supplementary Figure 2 (a)), which shows the advantages of having a router to set up a simpler pipeline for common disease diagnosis. Our classifier also predicts accurately for the correct disease types not only in our held-out testing split, but also external validation datasets (MyGene2 and RareBench), shown in Supplementary Figure 2 (b). We also visualize the distribution of embeddings with Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) in Supplementary Figure 2 (c), colored by disease types, and we see a clear difference between samples with rare diseases and samples with common diseases, which works as an explanation for our contribution. The diagnosis of rare diseases is more difficult, and we collect seven datasets in our evaluation pipeline. Figure 2 (a) shows the distribution of case numbers across all datasets, while Figure 2 (b) shows the unique number of diseases in different datasets. These figures show that our selected datasets have obvious differences in data distribution, which helps us simulate the clinical usage in the real-world setting. Among these datasets, RAMEDIS, MME, HMS, and LIRICAL are extracted from RareBench, which is a public benchmark framework for evaluating the performance of models for rare disease diagnosis. MyGene2 (Rodrigues et al., 2022) contains family-level information for patients with rare diseases, which is also publicly available. Undiagnosed Diseases Network (UDN) (Ramoni et al., 2017) covers diseases that are hard to diagnose and it is not directly accessible by public researchers to protect personal information.

Regarding baseline methods, we also consider a group with strong diversity. Our baselines include LLMs with/without reasoning and searching capacities, agents for biomedical research, and LLMs finetuned with simulated patient-level data from (Alsentzer et al., 2025) with the open-source Qwen3 model (Yang et al., 2025a). Details of our baseline methods can be found in the Methods section.

During evaluation, we test the Recall rate by

comparing the observed diseases with the predicted diseases from different methods. Figure 2 (c) shows that Hygieia outperforms various baseline methods across datasets from different resources. Fine-tuning LLMs for disease diagnosis also does not have strong generalization ability, and thus cannot outperform most of the agent-based solutions. Moreover, since some rare diseases have similar phenotypes or can be treated similarly (Griggs et al., 2009), we also test if increasing the size of predicted targets (e.g., top 5 and top 10 diseases) can improve the Recall rate. Figure 2 (d) shows our results based on the MyGene2 dataset, where Hygieia had a higher recall rate when we increased the testing size. At the same time, not all methods exhibit this phenomenon, which is counterintuitive. This might indicate that the approach of some models taken to clinical decision-making may still be primarily based on guesswork, with reasoning playing a secondary role. However, by integrating the prior knowledge and introducing the self-reflection design, Hygieia achieves a better result in diagnosing rare diseases.

We also test the reliability of estimated confidence, shown in Supplementary Figure 3 (a). Our expectation is that Hygieia has higher confidence for questions with correct answers. Our results demonstrate that 1. the confidence proposed by Hygieia is reliable as the answer group with higher confidence also has higher recall rate and 2. other confidence estimation methods are worse than the current design, shown in Supplementary Figure 3 (b) as the rest of three methods cannot produce significant differences between these two groups.

Case study of Hygieia for disease diagnosis shows its unique and informative decision-making process. To understand the decision-making process of Hygieia and make an in-depth comparison with other baseline methods, we provide a case study with MyGene2. In this sample, we provide phenotypes and detected genes as inputs; the diagnosis is *Distal arthrogryposis, type 10*. The full reasoning path of Hygieia can be found in Supplementary Figure 4 (a).

Based on Figure 3 and among all evaluated methods, Hygieia is the only model that correctly made the definitive diagnosis *Distal arthrogryposis, type 10*, fully matching both the annotated answer and the genetic etiology (*TTN* mutation). In contrast, alternative baselines either provide incorrect diagnoses (e.g., *Bethlem myopathy, Emery–Dreifuss muscular dystrophy, Flexion contracture, Arthrogryposis multiplex congenita*) or produce overly broad, nonspecific conclusions. This demonstrates that Hygieia not only retrieves the correct disease category but is capable of fine-grained subtype resolution, as an essential requirement in precision medicine.

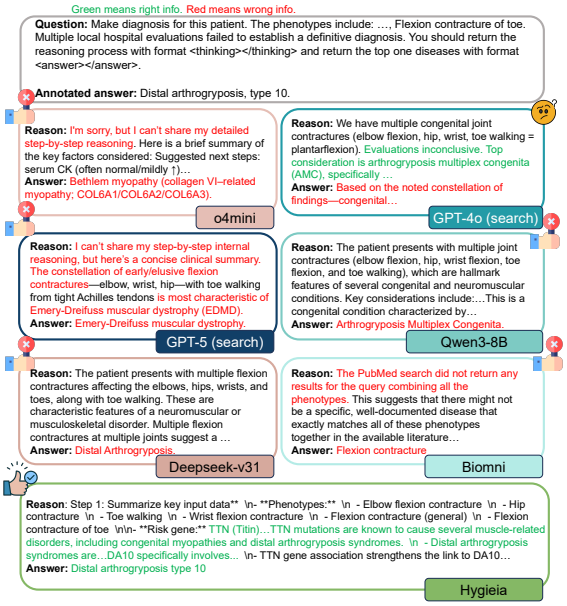


Figure 3: Case study of Hygieia and other baselines in disease diagnosis. We mask some phenotypes to protect personal information.

Unlike other models, Hygieia has a multi-step reasoning process that integrates phenotypic patterns, risk gene associations, and syndrome specificity. This demonstrates a higher level of biomedical causal interpretability. While several baselines refuse to show internal reasoning or default to vague clinical summaries, Hygieia transparently links phenotype, genotype, and nosology, exhibiting a cognitively valid chain of inference.

This specific case intentionally includes multiple phenotypes designed to confound rule-based or pattern-matching systems. Models like o4-mini, GPT-4o, and Qwen3-8B fail to integrate the constellation of findings, instead overfitting to a single clinical feature (e.g., *toe walking, wrist contracture*) and outputting unrelated diagnoses. Hygieia, however, successfully recognizes that the multi-joint congenital contracture pattern represents a diagnostic signature of distal arthrogryposis, showing resilience to feature redundancy and phenotypic noise.

Where competing models either offer no next steps or suggest non-specific investigations, Hygieia explicitly ties the diagnosis to a recognized

molecular driver and its associated disease spectrum. This strengthens the translational value of its output, enabling downstream steps such as confirmatory genetic testing, family counseling, and prognosis stratification. The output is not merely a label, but clinically operational knowledge, surpassing the diagnostic passivity of baseline systems. Taken together, the evidence indicates that Hygieia demonstrates a substantially higher standard of diagnostic precision, biomedical reasoning depth, and clinical applicability compared to both traditional LLM baselines and search-augmented systems. Its ability to unify phenotypic complexity with molecular knowledge exemplifies the next generation of AI-assisted medical decision systems, thereby positioning Hygieia as the most reliable and clinically aligned model in this evaluation.

Utilizing Hygieia as a medical assistant for physicians in solving complicated cases. One major objective of developing and deploying Hygieia is to make a virtual assistant (Copilot) for physicians and clinicians working on rare disease diagnosis and treatment development, and thus, matching user requirements with the functionality of Hygieia is a crucial step. Previous medical AI agent development has not explored this area extensively, thereby limiting their specific deployment capabilities. In this work, we collaborated with physicians from Yale School of Medicine (YSM) and Yale New Haven Hospitals (YNHH) to define key stakeholders in disease diagnosis and chart a blueprint for human-AI collaboration, thereby providing a guidance framework for the concrete implementation of Hygieia.

Figure 4 (a) shows two expected functions of Hygieia from physicians to improve the efficiency and reduce the effort, including direct diagnosis based on physician input, and verification and refinement of physician judgments. The commonality between these two tasks is that both require interaction between the physician and the AI agent, where the physician articulates the character's needs, and the agent fulfills them. We also provide some examples here to showcase how we can use Hygieia to accomplish these tasks. In Figure 4 (b), we illustrate that Hygieia can take the physicians' instructions with patient phenotype information (free text format) and other relevant information as inputs, and integrate its different components to create a pipeline, call different tools, and communicate with different AI agents to produce a diagnosis result. In Figure 4 (c), we showcase how Hygieia can take the physicians' diagnosis results as well as phenotype information as inputs, and by using an alternative best pipeline, first determine the True/False or original diagnosis, and then perform reasoning and correction, to report a new diagnosis answer.

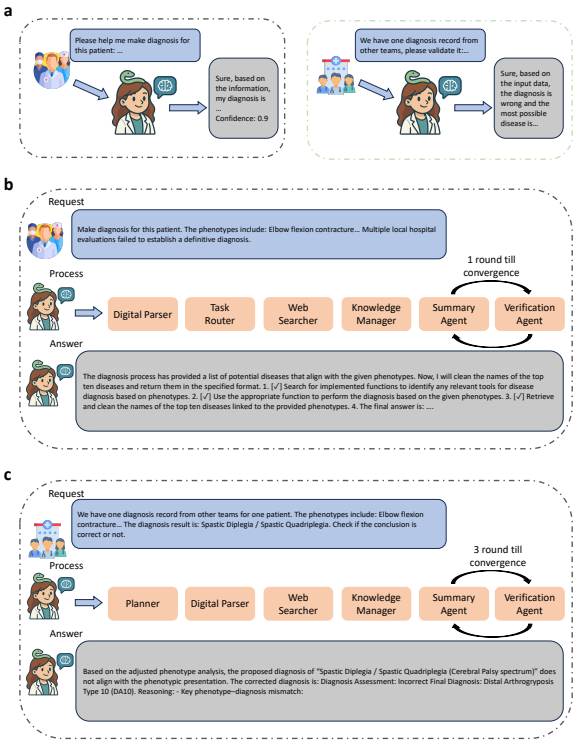


Figure 4: Illustration of Human-AI collaboration for disease diagnosis and decision correction based on physicians and Hygieia. We mask phenotype information to protect patients' privacy. (a) Explanations of two selected tasks; (b) Illustration of the diagnosis of diseases based on physician input; (c) Illustration of verification and refinement of physician judgments.

The successful operation of Hygieia (its correct reasoning process and outcomes) demonstrates its potential to translate into tangible medical value for healthcare teams, further highlighting its dual contributions at both the algorithmic and application levels. Moving forward, we will engage physicians as human evaluators to directly compare Hygieia with doctors in tasks such as disease diagnosis and key gene identification, thereby further extending Hygieia.

Hygieia successfully prioritizes genes with higher disease risks from individual-level data. To enhance diagnostic interpretability and confidence while providing additional therapeutic insights, modern medicine often analyzes causative factors (Sanchez et al., 2022). In rare disease diagnosis, some physicians may recommend exome-

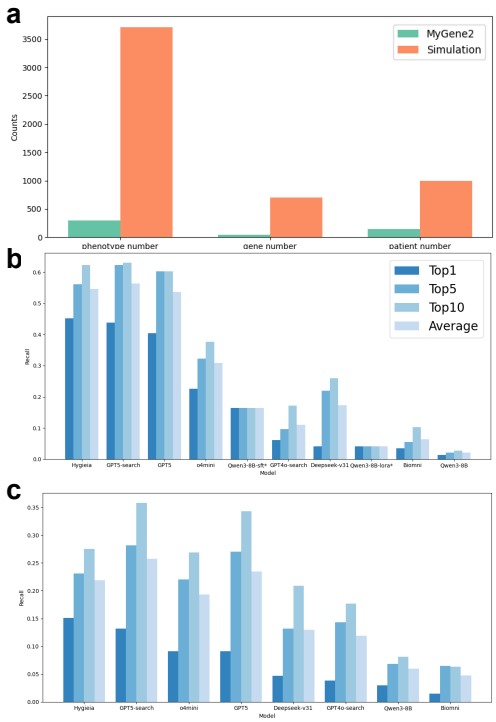


Figure 5: Benchmarking results for Hygieia in risk gene prioritization. (a) Statistics in our testing datasets. (b) Top1 recall rate across different datasets. (c) Comparisons of different models with different recall rates in MyGene2.

or whole-genome sequencing or targeted gene sequencing to identify disease-causing variants, enabling more reliable conclusions (Boycott et al., 2013). Therefore, determining how to provide patients with a list of potential genes for sequencing represents a critical task in rare disease diagnosis. Previously, few AI agent frameworks have considered this task. However, our analysis indicates that disease diagnosis shares similarities with risk factor prioritization, suggesting it can be addressed using a unified framework. Here, our input data are still phenotypes or EHR data, while the output will be a gene or a list of genes. Our selected baselines are similar to the candidates used for evaluating disease diagnosis functions. Details can be found in the Methods section.

Due to the scarcity of datasets containing both phenotypes and true disease-causing genes, this section employs MyGene2 and simulation data provided by SHEPHERD (Alsentzer et al., 2025) for model evaluation. The statistics of selected datasets are summarized in Figure 5 (a). The scale of simulation data is larger than MyGene2, and thus our assessment also took various scenarios into account. We still computed the Recall rate based on gene lists of different sizes and observed gene labels. Figures 5 (b) and (c) show that Hygieia has a high recall rate, especially under the top 1 setting versus other baselines. However, as we increase the pool of candidates, the advantages of Hygieia have also diminished. When comparing the recall rate of this task to disease diagnosis, we observe that gene prioritization is a relatively simpler task. Consequently, as the pool of candidates expands, the benefits of employing agents (e.g. additional verifier) diminish proportionally. Considering that the more genes that need to be tested, the higher the cost for patients, our approach balances accuracy and expense. We also compared the costs of Hygieia and GPT-5-search, revealing that Hygieia holds an advantage in token consumption as well, shown in Supplementary Figure 5. In Figure 5 (c), since we can create a training dataset from the large simulation data, we can also create an oracle model (Qwen3-8B-sft, score is 0.724). However, the performance of this model for recommending based on MyGene2 is poor, demonstrating that AI Agents have better generalization ability than traditional SFT approaches.

Case study of Hygieia for gene ranking shows its unique and informative decision-making process. To understand the decision-making process of Hygieia and make an in-depth comparison versus other baseline methods, we provided a case study with one sample from MyGene2, but for risk gene prioritization. In this sample, we provided phenotypes as inputs; the observed risk gene is *NALCN*. The full reasoning path of Hygieia can be found in Supplementary Figure 4 (b).

According to Figure 6, in this evaluation, Hygieia is the only system that successfully identifies NALCN, which is the correct gene associated with the patient's constellation of phenotypes, fully aligning with the annotated ground truth. All other baseline models, including o4mini, GPT-4o (search), GPT-5 (search), Qwen3-8B, and DeepSeek-v31, either returned incorrect candidates such as *MYH3* or *PIEZO2*, or failed to provide a usable response altogether. This example demonstrates that Hygieia consistently outperforms both general-purpose large language models and search-augmented tools in high-stakes biomedical inference tasks requiring precise gene–phenotype matching.

Unlike competing models that default to *MYH3* based on superficial resemblance to Freeman–Sheldon syndrome, Hygieia distinguishes

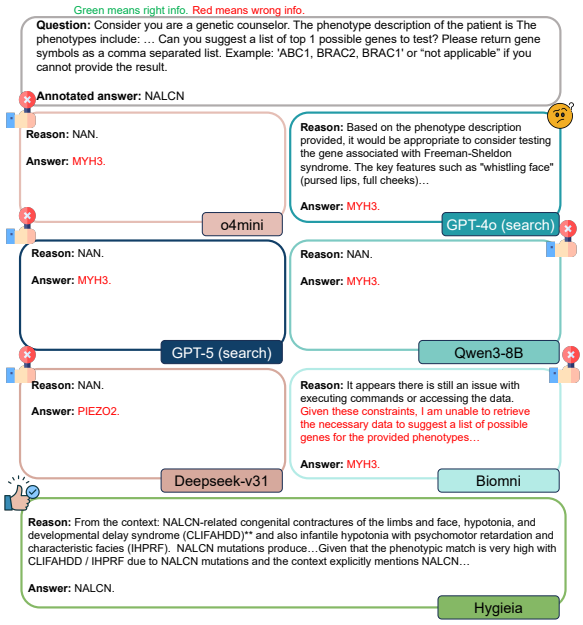


Figure 6: Case study for Hygieia and other baselines in risk gene prioritization. We mask some phenotypes to protect personal information.

clinically overlapping yet genetically distinct disorders by integrating phenotypic, molecular, and nosological evidence. The model explicitly links the observed features—including congenital contractures, hypotonia, neurodevelopmental delay, and characteristic facies, which are NALCN-associated CLIFAHDD/IHPRF syndromes. This indicates that Hygieia does not rely solely on phenotypic pattern matching, but instead performs multi-layered biomedical reasoning consistent with genetic counseling practice.

Several baselines fail due to inability to retrieve or interpret data, producing unusable outputs ("NAN", "unable to retrieve data"). Hygieia remains fully functional even under incomplete signal, reflecting robustness to real-world clinical constraints, where patient phenotypes may be sparsely documented, noisy, or partially overlapping. This resilience is essential for deployment in clinical decision support, where diagnostic completeness is rarely guaranteed.

Hygieia not only outputs the correct target gene but also contextualizes it within a clinically actionable diagnostic category. This stands in sharp contrast to competing models that provide unsubstantiated gene names without justification, which would be unacceptable in a clinical genetics workflow where gene testing decisions have financial, ethical, and prognostic implications. Hygieia's interpretability and biological validity, therefore, make it a more trustworthy candidate for integration into precision medicine pipelines.

The comparison clearly illustrates that Hygieia surpasses existing LLM-based and search-augmented baselines in terms of accuracy, reasoning validity, and clinical relevance. Its ability to discriminate among phenotypically similar developmental syndromes and return a gene with direct translational value underscores its potential as a next-generation AI system for genetic factor prioritization.

4 Discussion

Overall, Hygieia introduces several conceptual and technical advances that collectively address the above challenges. First, it formalizes diagnosis as a multi-stage agnetic workflow, beginning with task parsing and disease-type routing, followed by tailored pipelines for common and rare diseases. By explicitly separating these pathways, Hygieia avoids over-generalized reasoning and reduces both computational cost and diagnostic error for common conditions, while allocating more sophisticated reasoning resources to rare disease cases. Second, the system incorporates a self-verification mechanism that iteratively evaluates and corrects intermediate diagnostic outputs until convergence. This verifier–corrector design substantially improves consistency and robustness, mitigating the well-documented randomness of LLM-based inference. Empirically, this design leads to more stable predictions, and we also provide better-calibrated confidence estimates with multiple queries. Third, Hygieia extends beyond diagnosis to risk gene prioritization, framing it as a closely related inference task that benefits from the same phenotypic and contextual reasoning. By unifying these tasks in a single framework, we produce clinically actionable outputs that can directly empower several associated tasks. Fourth, Hygieia also emphasizes interpretability and clinical alignment. Rather than producing opaque predictions or generic summaries, our system explicitly links phenotypes, disease entities, and gene-level evidence through multi-step reasoning trajectories.

Despite its strong performance, Hygieia also has several limitations that warrant discussion. In the future, we will focus on extending Hygieia via more model selection as well as constructing more datasets for the system training and validation.

References

Emily Alsentzer, Michelle M Li, Shilpa N Kobren, Ayush Noori, Undiagnosed Diseases Network, Isaac S Kohane, and Marinka Zitnik. 2025. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):380.

Kym M Boycott, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–691.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2024a. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *arXiv preprint arXiv:2412.12475*.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024b. Rarebench: can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4850–4861.

Jagjit Singh Dhatterwal, Mahaveer Singh Naruka, and Kuldeep Singh Kaswan. 2023. Multi-agent system based medical diagnosis using particle swarm optimization in healthcare. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pages 889–893. IEEE.

Dong Dong, Roger Yat-Nork Chung, Rufina HW Chan, Shiwei Gong, and Richard Huan Xu. 2020. Why is misdiagnosis more likely among some people with rare diseases than others? insights from a population-based cross-sectional study in china. *Orphanet journal of rare diseases*, 15(1):307.

Yuanqi Du, Botao Yu, Tianyu Liu, Tony Shen, Junwu Chen, Jan G Rittig, Kunyang Sun, Yikun Zhang, Zhangde Song, Bo Zhou, and 1 others. 2025. Accelerating scientific discovery with autonomous goal-evolving agents. *arXiv preprint arXiv:2512.21782*.

Tapan Ghosh. 2025. Artificial intelligence in rare disease diagnostics: Shortening the path to early detection.

Robert C Griggs, Mark Batshaw, Mary Dunkle, Rashmi Gopal-Srivastava, Edward Kaye, Jeffrey Krischer, Tan Nguyen, Kathleen Paulus, Peter A Merkel, and 1 others. 2009. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular genetics and metabolism*, 96(1):20–26.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Anneliene H Jonker, Maria Cavaller-Bellaubi, Yukiko Nishimura, and David A Pearce. 2024. Access in the rare diseases landscape. *The Lancet Global Health*, 12(10):e1587.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Junghwan Lee, Cong Liu, Junyoung Kim, Zhehuan Chen, Yingcheng Sun, James R Rogers, Wendy K Chung, and Chunhua Weng. 2022. Deep learning for rare disease: A scoping review. *Journal of biomedical informatics*, 135:104227.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. *CoRR*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chunyu Liu, Yixiao Jin, Zhouyu Guan, Tingyao Li, Yiming Qin, Bo Qian, Zehua Jiang, Yilan Wu, Xiangning Wang, Ying Feng Zheng, and 1 others. 2025a. Visual–language foundation models in medicine. *The Visual Computer*, 41(4):2953–2972.

Tianyu Liu, Tinglin Huang, Tong Ding, Hao Wu, Peter Humphrey, Sudhir Perincheri, Kurt Schalper, Rex Ying, Hua Xu, James Zou, and 1 others. 2026a. Leveraging multi-modal foundation models for analysing spatial multi-omic and histopathology data. *Nature Biomedical Engineering*, pages 1–18.

Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, Wenjin Zheng, and Hongyu Zhao. 2024b. Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 4819–4836.

Tianyu Liu, Weihao Xuan, Hao Wu, Peter Humphrey, Marcello DiStasio, Heli Qi, Rui Yang, Simeng Han, Tinglin Huang, Fang Wu, and 1 others. 2025b. Teampath: Building multimodal pathology experts with reasoning ai copilots. *arXiv preprint arXiv:2511.17652*.

Yang Liu, Honglei Li, Peng Jiang, Lizhen Wu, Zhi Xie, Chao Ning, Xiangya Kong, Yayun Wang, Xinlei Zhang, and Zechi Huang. 2026b. Vc-rdagent: An efficient rare disease diagnosis agent via virtual case construction informed by hybrid statistical-metric and hyperbolic-semantic prioritization. *bioRxiv*, pages 2026–02.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Stéphanie Nguengang Wakap, Deborah M Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European journal of human genetics*, 28(2):165–173.

OpenAI. 2025. Gpt-5 system card. Available at: https://openai.com/index/introducing-gpt-5/.

OpenAI. 2025. Openai o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card/. System card. Accessed 2025-11-07.

Friedemann Paul. 2013. Hope for a rare disease: eculizumab in neuromyelitis optica. *The Lancet Neurology*, 12(6):529–531.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rachel B Ramoni, John J Mulvihill, David R Adams, Patrick Allard, Euan A Ashley, Jonathan A Bernstein, William A Gahl, Rizwan Hamid, Joseph Loscalzo, Alexa T McCray, and 1 others. 2017. The undiagnosed diseases network: accelerating discovery about health and disease. *The American Journal of Human Genetics*, 100(2):185–192.

Peter N Robinson and Stefan Mundlos. 2010. The human phenotype ontology. *Clinical genetics*, 77(6):525–534.

Eliete da S Rodrigues, Sean Griffith, Renan Martin, Corina Antonescu, Jennifer E Posey, Zeynep Coban-Akdemir, Shalini N Jhangiani, Kimberly F Doheny, James R Lupski, David Valle, and 1 others. 2022. Variant-level matching for diagnosis and discovery: Challenges and opportunities. *Human mutation*, 43(6):782–790.

Matteo Rossi and Aisha El-Sayed. 2025. Meta-learning driven few-shot diagnostics: Addressing rare disease classification in medical ai. *International Journal of Advanced Artificial Intelligence Research*, 2(05):7–14.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. 2022. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638.

Abeed Sarker, Rui Zhang, Yanshan Wang, Yunyu Xiao, Sudeshna Das, Dalton Schutte, David Oniani, Qianqian Xie, and Hua Xu. 2024. Natural language processing for digital health in the era of large language models. *Yearbook of Medical Informatics*, 33(01):229–240.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Rodolfo Valdez, Lijing Ouyang, and Julie Bolen. 2016. Public health and rare diseases: oxymoron no more. *Preventing chronic disease*, 13:E05.

Hengchang Wang, Li Liu, Huaxiang Zhang, Lei Zhu, Xiaojun Chang, and Hao Du. 2025. Visualrag: Knowledge-guided retrieval augmentation for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *ICLR*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jian Yang, Liqi Shu, Huilong Duan, and Haomin Li. 2025b. Rdguru: An intelligent agent for rare diseases. In *AMIA Annual Symposium Proceedings*, volume 2024, page 1275.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Ya Zhang, Yongguo Yu, and 1 others. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410.

A Details of human-in-the-loop design

Human-in-the-loop design. We also consider the collaboration between physicians and AI models in making decisions. We have invited 3 genetic physicians from YSM, YNHH, and Duke-NUS Medical School, and consider two scenarios.

In the first experimental setting, we assign 20 questions (10 for disease diagnosis, and 10 for gene prioritization) for physicians and Hygieia for disease diagnosis, and evaluate the performances.

In the second experimental setting, we allow the physicians to access Hygieia, utilize it for handling in-house cases and physicians will evaluate the outputs of Hygieia and provide feedback from three perspectives:

- Completeness: Which report more completely captures important information?
- Correctness: Which report includes less false information?
- Conciseness: Which report contains less non-important information?

We collect the reports made by physicians and evaluate the Hygieia as an AI for healthcare accordingly.

Case study investigation. To better demonstrate the advantages of Hygieia, we provide several case studies from different baseline methods and include the reasoning steps of Hygieia. In these questions, only Hygieia makes the correct decision, while the rest of the methods do not give us either correct reasoning paths or correct answers.

Explanations of baseline methods. For closed-source models, we consider LLMs with reasoning capacities including o4-mini (OpenAI, 2025), GPT-5 (OpenAI, 2025), and GPT-5 (search) (OpenAI, 2025), advanced LLMs including GPT-4o (Hurst et al., 2024) and GPT-4o (search) (Hurst et al., 2024). For open-source models, we consider Qwen3-8B (Yang et al., 2025a) and Deepseek-v3.1 (Liu et al., 2024a). For domain-expert models, we consider a multi-agent diagnosis system (base model GPT-4o as the recommended setting), as well as DeepRare (Zhao et al., 2026). However, DeepRare's API version has bugs related to data processing and we cannot find their released testing datasets and it is hard for us to retreive the data format. For biomedical agents, we consider Biomni (base model GPT-4o as the recommended setting. For Qwen3-8B, we also consider fine-tuning the base model with both Low-rank adaptation (LoRA) and full parameters (Full) (Zheng et al., 2024) based on the provided simulation data. The prompts used to query LRM and LLMs are documented in Appendix C. For other models, details can be found in our code base.

Metrics. We follow the settings discussed in (Chen et al., 2024b; Zhao et al., 2026), where the Recall@K is the main metric used in evaluating the generated disease diagnosis and gene rank results. We consider the top 1,5,10 candidates and compute Recall@1, Recall@5, Recall@10, accordingly. We perform metric computation after disease name/-gene name normalization.

B Comparison between Hygieia and other AI agents

In Supplementary Table , we showcase the unique contribution of Hygieia by comparing it with other disease diagnosis agents across multiple dimensions.

C Prompts

The prompts used for disease diagnosis: "Make diagnosis for this patient. Known phenotypes include: {phenotype_list}. Multiple local hospital evaluations failed to establish a definitive diagnosis."

The prompts used for gene prioritization: "Consider you are a genetic counselor. The phenotype description of the patient is {phenotype_list}. Can you suggest a list of top 1 possible genes to test?"

The prompts used for error detection and correction:

You are a board-certified clinical geneticist and neurologist with expertise in rare neuromuscular and congenital disorders. You reason step-by-step using established diagnostic criteria, genotype–phenotype correlations, and differential di-

System	Distinguish	Human	Confidence	Risk Gene	Flexible Input	Open Source	Evaluation Setting	Domain
MCA (Chen et al., 2025)	✗	✗	✗	✗	✓	✓	Benchmark	Rare
MDAgent (Kim et al., 2024)	✓	✗	✗	✗	✓	✓	Benchmark	General
DeepRare (Zhao et al., 2026)	✗	✗	✗	✗	✗	✓	Benchmark	Rare
Biomni (Huang et al., 2025)	✓	✗	✗	✓	✓	✓	Benchmark & Human Eval	General
RDguru (Yang et al., 2025b)	✗	✗	✗	✗	✗	✗	Benchmark	Rare
RareAgent (Chen et al., 2024a)	✗	✗	✗	✗	✗	✗	Benchmark	Rare
Hygieia (Ours)	✓	✓	✓	✓	✓	✓	Benchmark & Human Eval	Rare

Table 1: Comparison of representative AI Agents for disease-related tasks across different dimensions. Here we consider Distinguish (whether the agent can distinguish common and rare diseases), Human (whether the agent supports human-in-the-loop), Confidence (whether the agent can produce confidence), Risk Gene (whether the agent can also infer risk gene), Flexible Input (whether the agent supports input other than HPO terms), and Open Source (whether the codes are accessible). We also compare their evaluation settings as well as focused domains.

agnosis logic.

Below is a patient's clinical phenotype and a proposed diagnosis.

Your task is to determine whether the proposed diagnosis is correct.

Instructions: 1. Carefully assess whether the phenotype is consistent with the proposed diagnosis. 2. If the diagnosis is correct, explicitly state that it is correct and explain why. 3. If the diagnosis is incorrect or incomplete, clearly state that it is incorrect and: - Provide the most likely corrected diagnosis - Briefly justify the correction using key phenotype–disease matches 4. Do not provide multiple diagnoses—return one best diagnosis only. 5. Be concise, clinically precise, and avoid speculation beyond the given phenotype.

Patient Phenotype: {PHENOTYPE_LIST}

Proposed Diagnosis: {PROPOSED_DIAGNOSIS}

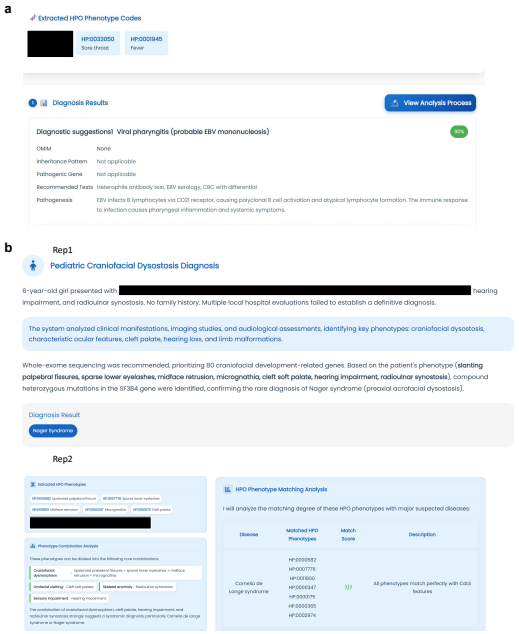
Output Format (strict):

Diagnosis Assessment: Correct / Incorrect

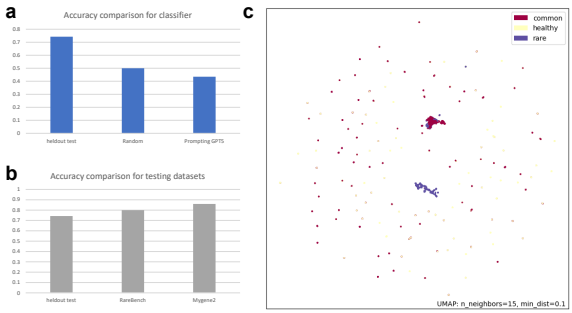
Final Diagnosis: <single diagnosis name>

Reasoning: - Key phenotype–diagnosis alignment (or mismatch) - Critical features supporting the final diagnosis

D Supplementary Figures

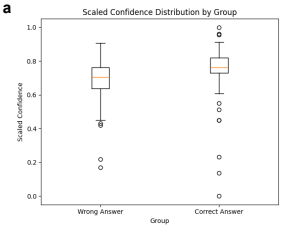


Supplementary Fig. 1: Failure cases of AI agent in disease diagnosis, by using DeepRare as an example. We mask some phenotypes to protect patient information. (a) Failure of diagnosis. (b) Failure of reproducing results.



Supplementary Fig. 2: Classification results of router. (a) Accuracy comparisons of different classification methods. (b) Accuracy comparisons of different testing sets. (c) UMAP colored by sample labels (common diseases, rare diseases, and healthy people).

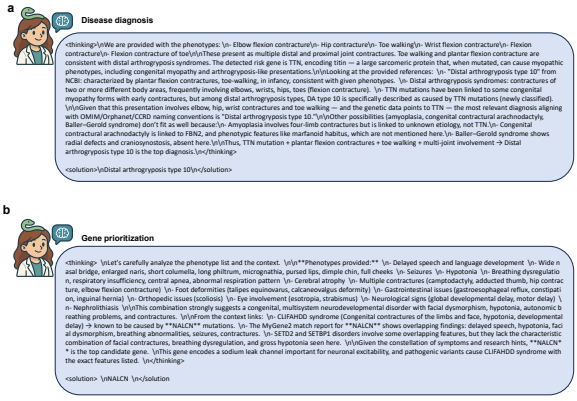
a



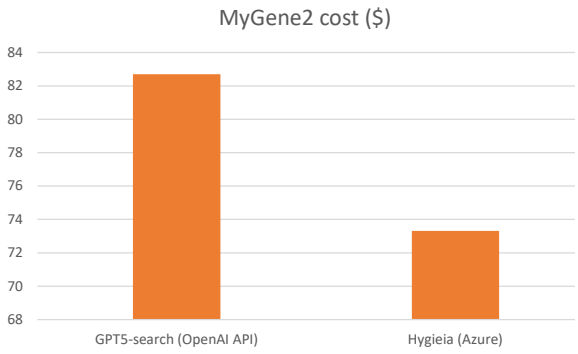
b

	p-value
LogProbs	0.294
Think-Twice	1
True/False Confidence	1
Voting Average	9.02E-06

Supplementary Fig. 3: Estimation of confidence levels based on Hygieia. (a) Comparison of scaled confidence scores tested with MyGene2 dataset. (b) Ablation studies for different estimation methods. The p-value is computed based on two-sided Mann-Whitney U test.

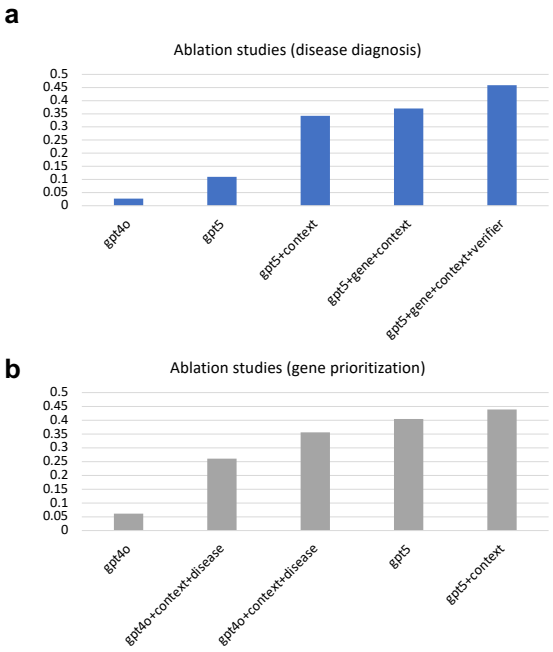


Supplementary Fig. 4: Full reasoning paths in two tasks by Hygieia. (a) is for disease diagnosis and (b) is for risk gene prioritization.



Supplementary Fig. 5: Cost analysis between GPT-5 Search and Hygieia in risk gene prioritization based on MyGene2.

This is an appendix.



Supplementary Fig. 6: Ablation studies of Hygieia. (a) represents the results for disease diagnosis and (b) represents the results for risk gene prioritization.