

Frequency-Stratified Benchmarking Reveals Strong Long-Tail Limitations of Large Language Models in Diagnosis of Monogenic Diseases

Jihao Cai

Peking Union Medical College Hospital

Jianle Yang

Peking Union Medical College Hospital

Maimaitiyibubaji Abudukadier

Peking Union Medical College Hospital

Aoxiang Luo

Peking Union Medical College Hospital

Lina Zhao

Peking Union Medical College Hospital

Guozhuang Li

Peking Union Medical College Hospital

Kexin Xu

Peking Union Medical College Hospital

Zhihong Wu

Peking Union Medical College Hospital

Terry Jianguo Zhang

Peking Union Medical College Hospital

Sen Zhao

Peking Union Medical College Hospital

Zefu Chen

jeffchenmed@gmail.com

Southern Medical University Nanfang Hospital

Nan Wu

dr.wunan@pumch.cn

Peking Union Medical College Hospital

Keywords: Large language model, Benchmark, Monogenic diseases, Mendelian disorders

Posted Date: June 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10158732/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Knowledge of monogenic disorders is long-tailed: a small number of canonical disease genes are extensively represented in the literature, whereas many rare, recently characterized genes remain sparsely reported. What remains less clearly characterized is how this uneven knowledge structure affects the diagnostic behavior of large language models (LLMs). Here we present MendelianBench-5K, a benchmark stratified by the literature reporting frequency of disease genes, comprising 5,000 molecularly diagnosed cases, and use it to evaluate nine LLMs on a phenotype-to-gene diagnostic task. The best-performing model achieved a Top-5 accuracy of 33.78%. Accuracy increased markedly with the literature frequency of the genes, whereas performance declined for recently published cases. False-positive predictions clustered on classical, well-studied genes. These findings reveal a systematic knowledge-frequency bias in LLM-based diagnosis of monogenic disorders and highlight the need for frequency-stratified benchmarks.

INTRODUCTION

Monogenic diseases, also known as Mendelian disorders, collectively account for approximately 80% of rare diseases^{1,2}. Although each individual disorder is typically uncommon, monogenic diseases as a group affect a substantial population and contribute to pediatric hospitalization, long-term disability, and early mortality^{3,4}. For affected individuals, accurate identification of the causative gene can not only end the prolonged “diagnostic odyssey”, but also directly inform clinical management, genetic counseling, and, in some cases, potential targeted therapies³. Molecular diagnosis is therefore a pivotal step in the clinical care of monogenic diseases.

The knowledge of monogenic diseases is not evenly distributed across the biomedical literature. Instead, it follows a characteristic long-tailed structure⁵. A small number of disease genes have been repeatedly reported and are supported by abundant patient cases, review articles, and database records, making them familiar to clinicians and researchers. In contrast, many causative genes are represented by only a few case reports, and some have only been described in recent years⁶. This long tail of knowledge implies that the most challenging diagnostic tasks are often concentrated in gene–phenotype correlations that are sparsely documented, insufficiently characterized, and less familiar in clinical practice.

This long-tailed distribution shapes not only human clinical knowledge but also the diagnostic behavior of artificial intelligence models^{7–10}. Clinicians and researchers learn gene–phenotype correlations largely from published cases and established knowledge resources. Similarly, the biomedical knowledge of large language models (LLMs) is derived to a substantial extent from large-scale text corpora and subsequent alignment processes^{11,12}. Accordingly, causative genes that are frequently reported in the literature are more likely to recur in model training data and to form stable statistical associations. By contrast, low-frequency or recently reported genes may be difficult for models to learn adequately

because of sparse supporting evidence^{13,14}. These considerations suggest that LLM performance in molecular diagnosis of monogenic diseases may not be uniformly distributed but may instead be shaped by the frequency structure of biomedical knowledge production itself.

This issue also poses an important challenge for the reliability assessment of medical LLMs. In a highly long-tailed field such as monogenic disease diagnosis, overall accuracy may obscure systematic failures on low-frequency and evidence-sparse genes. If a model performs well primarily on frequent, canonical, and long-established gene–phenotype correlations but degrades substantially on low-frequency or recently reported genes, its clinical utility and fairness in real-world monogenic disease diagnosis may be overestimated. It is therefore necessary to develop evaluation frameworks that explicitly capture gradients of literature frequency and test whether LLMs exhibit frequency-dependent diagnostic bias.

In this study, we constructed MendelianBench-5K, a frequency-stratified benchmark for evaluating LLM performance in molecular diagnosis of monogenic diseases. The benchmark was derived from the PhenoGemini Atlas, a literature-based patient-level phenotype–gene resource and includes 5,000 molecularly diagnosed cases¹⁵. We divided cases into five tiers based on the frequency of their causative genes reported in the literature and performed balanced sampling within each tier, enabling a direct assessment of whether model performance varies with gene-level literature frequency. Our analysis shows that, in the absence of external retrieval or structured genetic evidence, current LLMs cannot yet serve as reliable standalone diagnostic systems for this task, as they remain limited in their ability to perform molecular diagnosis of monogenic diseases, and their performance depends strongly on the literature reporting frequency of the causative gene. For rarely reported or recently characterized genes, model accuracy was substantially lower than for well-studied or long-established genes. Further error analysis revealed that LLM errors were not randomly distributed but were systematically structured by the frequency landscape of the biomedical literature: underrepresented causative genes were often missed, whereas familiar, high-frequency disease genes were repeatedly overpredicted. Taken together, these findings indicate that LLM-assisted molecular diagnosis of monogenic diseases is affected by knowledge-frequency bias, and that frequency-stratified benchmarks are essential for assessing the reliability of such systems in patients with long-tail rare diseases.

RESULTS

Construction and Characterization of MendelianBench-5K

PhenoGemini Atlas is a patient-level phenotype–gene database curated from published literature on monogenic diseases. To characterize the representation of genes in the literature, we defined “gene literature frequency” as the proportion of patients associated with each causative gene among all records. Analysis of 330,166 published patient records covering 4,735 distinct causative genes in the PhenoGemini Atlas revealed a highly imbalanced, long-tailed distribution of gene literature frequency (Fig. 1, **Fig. S1**).

To construct MendelianBench-5K, we stratified patients in PhenoGemini Atlas into five tiers according to literature frequency of their causative genes and randomly sampled 1,000 patients from each tier. MendelianBench-5K comprises 2,995 unique causative genes, among which 1,785 appear only once (**Table S1**). On average, each patient was annotated with 7.06 free-text phenotypic descriptions and 7.24 HPO terms mapped using Doc2Hpo. The 20 most frequent phenotypes are presented in **Table S2**, and their distribution across HPO organ systems is shown in **Table S3**. The median publication year of the source articles was 2016, with the temporal distribution summarized in **Table S4**. In addition, we report the coarse-grained Orphanet classification of MendelianBench-5K (**Supplementary Materials**).

Human Evaluation of MendelianBench-5K Quality

To evaluate the validity of cases included in MendelianBench-5K, two clinical physician reviewers with 1 year of experience in phenotype assessment independently performed a manual review of 500 randomly selected cases. Conflicting cases were subsequently adjudicated by a third physician with 5 years of experience in phenotype assessment. The manual assessment showed that 485 cases were classified as *supported*, accounting for 97.00% of the reviewed cases, whereas 15 cases were classified as *not supported*, accounting for 3.00% (Table 1). Among the 15 unsupported cases, the most common error type was misattribution (9 cases), including cases in which a phenotype was assigned to an incorrect gene, the phenotype was derived from an unaffected individual carrying a pathogenic variant, or demographic information was incorrectly treated as a phenotype. The next most frequent error type was model hallucination, defined as cases in which the reported gene–phenotype correlation was not mentioned in the source literature, accounting for 3 cases. Negative conclusions, including cases in which the source literature reported a negative finding for the association, accounted for 2 cases. One additional case was classified as other (Table 1). These results indicate that most cases recorded in MendelianBench-5K were consistent with the source literature. The small proportion of unsupported cases was mainly attributable to misattribution errors.

Table 1
Human evaluation of case quality in MendelianBench-5K.

	Category	Count
Human evaluation	Supported	485
	Not supported	15
	Total	500
Error type	Misattribution errors	9
	Hallucination	3
	Negative conclusion	2
	Other	1
	Total	15

Performance Comparison of Nine Mainstream Large Language Models in Molecular Diagnosis of Monogenic Diseases

We first evaluated the overall diagnostic performance of nine mainstream large language models on MendelianBench-5K using Top-1 accuracy, Top-5 accuracy, and mean reciprocal rank (MRR). Overall, the diagnostic performance of all models remained limited, indicating that molecular diagnosis of monogenic diseases using large language models alone still poses challenges (Fig. 2A, S2, and S3).

Among the evaluated models, GPT-5.5 achieved the best performance across all three metrics, with a Top-1 accuracy of 18.56%, a Top-5 accuracy of 33.78%, and an MRR of 0.24. Gemini-3.1-pro-preview ranked second, achieving a Top-1 accuracy of 17.22%, a Top-5 accuracy of 30.94%, and an MRR of 0.22, followed by gemini-3-flash-preview, claude-opus-4.7, and deepseek-v4-pro. In contrast, qwen3-235b-a22b-2507 showed the lowest performance, with a Top-1 accuracy of 8.18%, a Top-5 accuracy of 17.40%, and an MRR of 0.11 (Fig. 2A).

Pairwise McNemar's tests revealed statistically significant differences between most models, with GPT-5.5 significantly outperforming all other models across comparisons (all $P < 0.001$). However, differences were not statistically significant between some models with comparable performance, such as claude-sonnet-4.6 and deepseek-v4-flash, claude-opus-4.7 and gemini-3-flash-preview, and deepseek-v4-pro and gemini-3-flash-preview (Fig. 2B).

Together, these findings suggest a clear performance gradient among current LLMs, while also highlighting that even the strongest model correctly identified the causative gene as the top-ranked diagnosis in fewer than one-fifth of cases.

LLM Diagnostic Performance Exhibits a Pronounced Long-Tail Bias: Accurate on Genes Frequently Reported in the Literature but Systematically Failing on Rarely Reported Ones

Because MendelianBench-5K was explicitly stratified by literature frequency of genes, with 1,000 patients sampled from each of five frequency tiers, it enabled a direct assessment of whether LLM diagnostic accuracy was preserved across the long tail of rarely reported genes. Across all nine models, diagnostic performance showed a strong and monotonic dependence on literature frequency of genes. For Top-1 accuracy, the average accuracy increased from only 2.24% in the quintile of genes least frequently reported in the literature (Q1) to 42.48% in the quintile of genes most frequently reported in the literature (Q5), representing an approximately 19-fold difference. Even the strongest model, GPT-5.5, correctly ranked the causative gene first in only 5.20% of Q1 cases, compared with 46.70% of Q5 cases (Fig. 3A). A similar pattern was observed for Top-5 accuracy: the mean accuracy increased from 6.92% in Q1 to 66.10% in Q5, and GPT-5.5 improved from 10.80% to 71.30% across the same frequency range (Fig. 3B). This frequency-dependent gradient was highly consistent across models, with Spearman correlations between gene literature frequency quintile and Top-1 accuracy ranging from $\rho = 0.375$ to 0.418 , and correlations for Top-5 accuracy ranging from $\rho = 0.442$ to 0.485 ; all associations were highly significant (all $P < 0.001$, **Table S5**). These results indicate that current LLMs are not uniformly effective across the diagnostic space of monogenic diseases. Rather, their apparent diagnostic ability is disproportionately concentrated in relatively literature-enriched genes, while rarely reported long-tail genes remain systematically underrecognized.

Effects of Publication Year and Input Phenotypic Information Content on LLM Diagnostic Accuracy

We next examined whether diagnostic accuracy was associated with the publication year of the source report and the amount of phenotypic information provided for each case. Among cases with available publication-year information ($n = 4,991$), LLM performance declined markedly for more recently published cases. Across models, the mean Top-1 accuracy decreased from 28.47% for cases published in or before 2004 to 7.83% for cases published during 2020–2024, representing a 3.6-fold reduction (Fig. 4A). A similar trend was observed for mean Top-5 accuracy, peaking at 44.00% in 2005–2009 and declining to 17.20% in 2020–2024 (Fig. 4B). This temporal gradient was consistent across all nine models; Spearman correlations between publication year and Top-5 accuracy were uniformly negative, ranging from $\rho = -0.195$ to -0.221 , with all associations highly significant (all $P < 0.001$, **Table S6**). Even GPT-5.5, the strongest model, showed a substantial decline from 33.19% Top-1 accuracy in cases published before 2005 to 11.40% in cases published during 2020–2024. In contrast, the relationship between diagnostic performance and the amount of phenotypic information was non-linear (Fig. 4C, 4D). Cases with only one phenotype showed the lowest mean accuracy across models, with a Top-1 accuracy of 10.26% and a Top-5 accuracy of 22.14%. Performance improved most notably when 2–3 phenotypes were provided, reaching the highest mean Top-1 and Top-5 accuracy of 16.46% and 31.92%, respectively. However, additional phenotypic terms did not lead to further improvement; performance plateaued or

declined in cases with 4–6, 7–10, or ≥ 11 phenotypes. Together, these results suggest that LLMs perform best on long-established gene–phenotype correlations, while phenotypic input improves model performance only up to a certain point, beyond which additional information may not lead to better diagnostic ranking.

Universal Failure Cases and False-Positive Genes Reveal Knowledge Bias and Diagnostic Behavioral Patterns in LLMs

To characterize cases that were universally difficult for current LLMs, we classified each case by the number of models that ranked the causative gene within their Top-5 predictions: Hard (0/9 models correct), Medium (1–4/9), and Easy (5–9/9). Among the 5,000 cases, 2,912 (58.24%) were classified as Hard, 697 (13.94%) as Medium, and 1,391 (27.82%) as Easy, indicating that universal Top-5 failure was common across current LLMs. Hard cases were strongly enriched for disorders caused by low literature frequency gene: 54.98% of Hard cases came from the two lowest literature frequency quintiles (Q1–Q2), whereas only 6.90% came from the highest quintile (Q5). In contrast, Easy cases showed the opposite pattern, with 50.04% originating from Q5 and only 11.79% from Q1–Q2 (Fig. 5A). This gradient was not explained simply by the amount of phenotypic information, as Hard and Easy cases had the same median phenotype count of 4 terms per case, with broadly overlapping distributions (**Fig. S4**). However, Hard cases tended to be more recently reported, with a median publication year of 2017 compared with 2013 for Easy cases, consistent with the temporal decline in LLM diagnostic accuracy described above (Fig. 5B).

Gene-level analysis of universal failures showed that recurrent Hard-case genes were predominantly low literature frequency genes in PhenoGemini Atlas, including *PACS2*, *NUP93*, *TNRC6B*, *HSD3B7*, *RINT1*, *S1PR2*, and *RAX2*, whereas only a small number of recurrently missed genes, such as *AR* and *LPL*, were relatively common (Fig. 5C). In parallel, false-positive predictions were disproportionately concentrated in well-known, high-frequency disease genes. *MECP2* was the most frequent false-positive gene across models, appearing 2,002 times when the true causative gene was absent from the Top-5, followed by *FMR1*, *ARID1B*, *POLG*, *CDKL5*, *KMT2D*, *CHD7*, *SCN1A*, *STXBP1*, and *GJB2* (Fig. 5D). Among the top 20 false-positive genes, the median PhenoGemini Atlas literature frequency rank was 106.5, and 14 of 20 ranked within the top 200 genes, indicating that models repeatedly defaulted to familiar, literature-rich genes when diagnostic uncertainty was high (**Table S7**). Together, these findings suggest that LLM diagnostic errors are not random: universal failures preferentially occur in rare, recently characterized, or underrepresented gene–phenotype correlations, whereas false-positive predictions are biased toward canonical and highly represented genes, revealing a systematic knowledge-frequency bias in LLM-based molecular diagnosis.

DISCUSSION

In this study, we constructed MendelianBench-5K, a literature-frequency-stratified benchmark designed to evaluate the ability of LLMs to perform phenotype-to-gene diagnostic task in monogenic diseases. By sampling 5,000 cases across five tiers of causative-gene literature reporting frequency, we were able to assess not only overall diagnostic accuracy but also whether model performance remained stable across the long tail of rarely reported genes. Our results show that the diagnostic capability of current large language models in monogenic disease molecular diagnosis remains substantially limited. Even the best-performing model ranked the correct causative gene first in fewer than one-fifth of cases, indicating that, in the absence of external retrieval or structured genetic evidence, current LLMs cannot yet serve as reliable standalone diagnostic systems for this task.

The central finding of this study is that the diagnostic performance of current mainstream LLMs is strongly shaped by the literature reporting frequency of the causative gene. Across all evaluated models, diagnostic accuracy increased progressively from the lowest to the highest gene-frequency tier. This trend was observed for both Top-1 and Top-5 accuracy and was highly consistent across models. These findings indicate that LLMs perform better for genes that are well represented in the biomedical literature, whereas rarely reported long-tail genes are systematically underrecognized and missed.

The publication-year analysis supports this interpretation. The LLMs achieved the highest accuracy in cases reported in earlier publications, whereas performance declined substantially for more recently published cases. This pattern may reflect the fact that earlier cases and genes are more likely to have been repeatedly reported, extensively studied, and incorporated into established biomedical knowledge. In contrast, recently reported cases may have had insufficient dissemination and may more often involve ultra-rare or atypical gene–phenotype correlations.

Our analysis of phenotype counts further revealed how LLMs respond to clinical input. Cases containing only a single phenotype had the lowest accuracy, suggesting that minimal phenotypic input is insufficient for reliable gene prioritization. Model performance improved when a small number of phenotypic features were provided, indicating that concise but diagnostically informative clinical descriptions can help narrow the diagnostic space. However, accuracy did not continue to improve as the number of phenotypes increased further. In cases with more extensive phenotypic descriptions, model performance plateaued or even declined, with a more pronounced effect among weaker models. A similar pattern was also observed by Deisseroth et al., who found that limiting the number of phenotypes could improve the performance of pathogenic gene-ranking algorithms¹⁶. This may be because cases with many phenotypes often represent more complex syndromic presentations, broader differential diagnoses, or noisier clinical descriptions. These results suggest that LLMs are most likely to benefit from curated, diagnostically meaningful phenotypes rather than unfiltered accumulation of phenotype terms.

Analyses of universally failed cases and false-positive predictions showed that LLM errors were not random, clarifying both their underlying mechanisms and clinical risks. Failure cases were enriched for low literature frequency and recently reported genes, whereas false-positive predictions were disproportionately concentrated in canonical and literature-rich “star” genes, such as *MECP2*, *FMR1*, and

POLG. This asymmetric error pattern indicates a knowledge-frequency bias: under diagnostic uncertainty, LLMs tend to default to familiar, high-attention genes instead of prioritizing rare or poorly documented ones. In clinical settings, this behavior could be risky, as it may lead models toward well-known diagnoses while overlooking patients affected by less well-studied genes.

These findings have several implications for the evaluation and deployment of medical LLMs. First, in biomedical domains with pronounced long-tailed knowledge structures, such as rare diseases and monogenic disorders, overall accuracy is insufficient for assessing diagnostic reliability. If an evaluation set is dominated by common, canonical, or long-established diseases, a model may appear clinically promising while still performing poorly on the rare and underrepresented cases for which diagnostic assistance is most needed. Frequency-stratified evaluation should therefore become a standard component of benchmark design in long-tailed biomedical settings. Second, clinical use of LLMs should emphasize decision support rather than autonomous diagnosis. LLM outputs must be integrated with expert assessment, genomic variant interpretation, pedigree analysis, and evidence from prior literature.

Our results also point to potential directions for improving LLM-assisted diagnosis of monogenic diseases. Agentic systems can equip LLMs with capabilities such as retrieval, memory, and reflection. Recent agentic diagnostic systems, such as DeepRare, have been reported to outperform humans in rare-disease diagnosis, suggesting that, compared with standalone LLMs, agentic systems may reduce reliance on static parametric knowledge by retrieving external data, maintaining memory, and performing iterative reflection, thereby substantially improving diagnostic performance¹⁷. In addition, future benchmarks should include long-tail, recently reported, and low literature frequency genes, so that model development does not optimize performance only on common diagnostic knowledge but also addresses the underrepresented cases that most urgently need diagnostic support.

This study has several limitations. First, although we evaluated multiple mainstream LLMs under zero-shot prompting, model performance may differ under alternative prompting strategies, chain-of-thought prompting, retrieval augmentation, or agentic workflows. Therefore, our results primarily characterize the independent parametric knowledge of current LLMs and likely represent a lower-bound estimate of what LLM-based diagnostic systems may achieve. Second, MendelianBench-5K evaluates phenotype-to-gene prioritization among individuals with established molecular diagnoses and therefore does not assess screening or case-finding tasks, such as determining whether a patient has a monogenic disorder. Finally, because commercial LLMs are updated rapidly, our findings should be interpreted as a time-specific assessment of current models, rather than as a permanent ranking of model capability.

In conclusion, current LLMs exhibit pronounced long-tail limitations in the molecular diagnosis of monogenic diseases: they are better at recognizing well-studied, canonical disease genes, but their performance declines substantially for rare or recently characterized genes. Their diagnostic outputs are influenced by knowledge frequency, making them prone to missing low-attention genes while overpredicting highly represented genes. Therefore, LLMs should be evaluated using frequency-stratified benchmarks and applied with caution in monogenic disease diagnosis. Future systems should move

beyond memorized associations and instead rely on continuously updated, evidence-driven, and clinically accountable diagnostic reasoning.

MATERIALS AND METHODS

PhenoGemini Atlas Data Source

We used the PhenoGemini Atlas as the source database for constructing MendelianBench-5K. The PhenoGemini Atlas is a literature-derived, patient-level phenotype–gene resource for monogenic diseases. Briefly, the atlas was generated by systematically mining the biomedical literature using OMIM-based disease and gene queries, followed by retrieval of candidate publications, identification of high-confidence case reports, and extraction of individual patient records from published clinical descriptions. Each record links a patient to the reported causative gene and the corresponding phenotypic descriptions. The PhenoGemini Atlas contains 347,046 molecularly diagnosed individuals, 604,259 phenotype–genotype associations, and 4,739 causative genes. In the present study, the PhenoGemini Atlas served as the source database for MendelianBench-5K and as the basis for estimating literature frequency of causative genes. Each patient record provided the benchmark input phenotypes and the corresponding causative gene label.

Construction of MendelianBench-5K

To construct MendelianBench-5K, we first excluded cases from the PhenoGemini Atlas that lacked a valid causative gene or phenotype annotation. The literature frequency of each causative gene was defined as the number of patient records in the full PhenoGemini Atlas associated with each causative gene, with Entrez Gene identifiers used to ensure consistent gene-level counting. Genes were then ranked by their literature frequency and divided into five frequency tiers, with Q1 containing the genes with the lowest frequency and Q5 those with the highest. Each patient record was assigned to a tier according to the frequency tier of its causative gene, and 1,000 patients were randomly sampled from each tier to generate a balanced 5,000-case benchmark. This design enabled frequency-stratified evaluation of large language model performance across both common and long-tail genes in the literature-frequency distribution.

In addition, we used Doc2Hpo to convert the free-text phenotype descriptions of each case into HPO terms. Specifically, phenotype annotation was performed using the Doc2HPO v2.0, which maps free-text phenotype descriptions to HPO terms via the Aho–Corasick string-matching algorithm.

Manual Evaluation

We randomly selected 500 cases from MendelianBench-5K using a fixed random seed to ensure reproducibility. Because the MendelianBench-5K was constructed by uniformly sampling cases across five gene literature frequency tiers, this random subset retained approximately balanced representation across the frequency tiers.

Two clinician reviewers, each with 1 year of experience in phenotype assessment, independently evaluated each case. The review focused on whether the patient phenotypes and causative gene reported in the original literature were consistent with the corresponding records in the dataset. Each case was assigned a binary judgment of either *supported* or *not supported*. For cases judged as *not supported*, reviewers additionally recorded the error type. The two reviewers worked independently and were blinded to each other's assessments. Cases with discordant judgments between the two reviewers were adjudicated by a third clinician with 5 years of experience in phenotype assessment, who determined the final review outcome and error type.

Error types were classified into four categories. The first category was misattribution error, including cases in which the phenotype was real but assigned to an incorrect gene, the phenotype came from an unaffected individual or a group-level description, or the extracted content represented a demographic characteristic rather than a patient phenotype. The second category was negative conclusion, referring to cases in which the original literature did not support the phenotype–gene association or reported a conclusion in the opposite direction from the dataset record. The third category was model hallucination, defined as cases in which the corresponding content was not mentioned in the source literature. The fourth category was other error, which was used for cases that did not fall into the above three categories.

Benchmark Characterization

We characterized MendelianBench-5K at multiple levels, including cases, genes, phenotypes, and publication years. Gene coverage was assessed by counting the number of unique Entrez Gene identifier in the benchmark. For phenotype characterization, we summarized both the free-text phenotype descriptions and the HPO terms mapped by Doc2Hpo, including the number of phenotypes per case and the most frequent phenotype terms. To describe the organ-system distribution of the phenotypes, HPO terms were mapped to major organ-system categories using the direct child terms of *Phenotypic abnormality* (HP:0000118). For each patient, all mapped HPO terms were assigned to their corresponding system-level ancestors, and the predominant organ-system category was determined by majority vote. Publication year was extracted from the source literature metadata in the PhenoGemini Atlas and summarized both overall and across predefined time intervals.

Benchmarking Large Language Models Performance

We evaluated large language models on MendelianBench-5K using a phenotype-to-gene diagnostic task. For each patient case, the model was given the free-text phenotype terms and asked to return the five most likely causal genes, ordered from most to least likely. To reduce variability across runs, all models were evaluated under a deterministic setting with temperature set to zero. The same zero-shot prompt format was used for all models (**Supplementary Materials**). Model outputs were constrained to gene symbols approved by the HUGO Gene Nomenclature Committee (HGNC), with one predicted gene per line. After generation, predictions were parsed into ranked gene lists of up to five candidates. To account

for aliases and symbol normalization, each predicted gene symbol was mapped to its corresponding Entrez Gene identifier before comparison with the ground-truth causal gene.

Performance was assessed using three ranking-based metrics. Top-1 accuracy measured the fraction of cases in which the highest-ranked prediction matched the causative gene. Top-5 accuracy measured the fraction of cases in which the causative gene appeared anywhere among the five predictions. Mean reciprocal rank measured how highly the correct gene was ranked, assigning a reciprocal score according to the first correct position and zero when the causative gene was absent from the prediction list.

Case-Difficulty Classification

Case difficulty was defined according to the number of models that included the causative gene within their Top-5 predictions. Cases were classified as Hard if 0 of 9 models were correct, Medium if 1–4 models were correct, and Easy if 5–9 models were correct. False-positive genes were summarized by counting predicted genes in cases where the true causative gene was absent from the Top-5 prediction list.

Statistical Analysis

Pairwise comparisons of Top-1 accuracy between models were conducted using McNemar's test for paired binary outcomes, as all models were evaluated on the same benchmark cases. To evaluate monotonic associations between benchmark characteristics and model performance, Spearman rank correlation analyses were performed at the case level. Specifically, the literature frequency of each ground-truth causative gene in the PhenoGemini Atlas was correlated with Top-1 and Top-5 correctness. Publication year was also correlated with diagnostic correctness among cases with available year information. Phenotype-count analyses were conducted descriptively after grouping cases into predefined categories: 1, 2–3, 4–6, 7–10, and ≥ 11 phenotype terms. All statistical analyses were performed in Python using pandas and SciPy. All statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

Declarations

Ethics Statement

This study did not involve any direct interaction with human participants or the collection of new clinical data. All analyses used de-identified information extracted from published literature, and no identifiable private information was accessed. The study protocol was reviewed and granted an exemption by the Ethics Committee of Peking Union Medical College Hospital. The need for informed consent was waived accordingly.

Data Availability

The MendelianBench-5K dataset and manual review records have been deposited on Hugging Face and is publicly available at <https://huggingface.co/datasets/DISCOStudy/MendelianBench-5K>.

Code Availability

The code used in this study has been deposited on GitHub and is publicly available at <https://github.com/Marker0707/MendelianBench-5K>.

Conflict of Interest

The authors declare no competing interests.

Acknowledgments

This study is part of the Deciphering Disorders Involving Scoliosis and COmorbidities (DISCO) Study Group. This research was funded in part by the National Natural Science Foundation of China (82572698 to J.Z. and 82402760 to L.Z.); National High Level Hospital Clinical Research Funding (2025-PUMCH-C-003 to J.Z., 2025-PUMCH-D-001 to Z.W., 2025-PUMCH-C-002 to Z.W., 2025-PUMCH-A-115 to L.Z.); Peking Union Medical College Hospital Talent Cultivation Program (Category A) No. ULJ11240 to N.W. ; CAMS Innovation Fund for Medical Sciences (CIFMS, 2023-I2M-C&T-A-003 to J.Z., 2023-I2M-2-002 to S.Z., 2024-I2M-TS-002 to N.W., 2025-I2M-XHJC-002 to N.W.; 2025-I2M-XHXX-020 to L.Z.).

References

1. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* 2018; **19**(5): 253–68.
2. Nurchis MC, Altamura G, Riccardi MT, et al. Whole genome sequencing diagnostic yield for paediatric patients with suspected genetic disorders: systematic review, meta-analysis, and GRADE assessment. *Arch Public Health* 2023; **81**(1): 93.
3. Boycott KM, Rath A, Chong JX, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet* 2017; **100**(5): 695–705.
4. Wojcik MH, Schwartz TS, Yamin I, et al. Genetic disorders and mortality in infancy and early childhood: delayed diagnoses and missed opportunities. *Genet Med* 2018; **20**(11): 1396–404.
5. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol* 2018; **16**(9): e2006643.
6. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* 2015; **97**(2): 199–215.
7. Kim J, Wang K, Weng C, Liu C. Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease. *Am J Hum Genet* 2024; **111**(10): 2190–202.

8. Chen H, Zhao Z, Zhou S, et al. RareArena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *Lancet Digit Health* 2026; **8**(2): 100953.
9. Kafkas Ş, Abdelhakim M, Althagafi A, et al. The application of Large Language Models to the phenotype-based prioritization of causative genes in rare disease patients. *Sci Rep* 2025; **15**(1): 15093.
10. Reese JT, Chimirri L, Bridges Y, et al. Systematic benchmarking demonstrates large language models have not reached the diagnostic accuracy of traditional rare-disease decision support tools. *Eur J Hum Genet* 2026; **34**(4): 498–504.
11. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates Inc.; 2020. p. Article 159.
12. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, LA, USA: Curran Associates Inc.; 2022. p. Article 2011.
13. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA: JMLR.org; 2023. p. Article 641.
14. Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. 2023 July; Toronto, Canada: Association for Computational Linguistics; 2023. p. 9802–22.
15. Chen Z, Cai J, Yang Y, et al. Large language models enhance molecular diagnoses of mendelian disorders via a novel logic. In Review; 2025.
16. Deisseroth CA, Birgmeier J, Bodle EE, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med* 2019; **21**(7): 1585–93.
17. Zhao W, Wu C, Fan Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature* 2026; **651**(8106): 775–84.

Figures

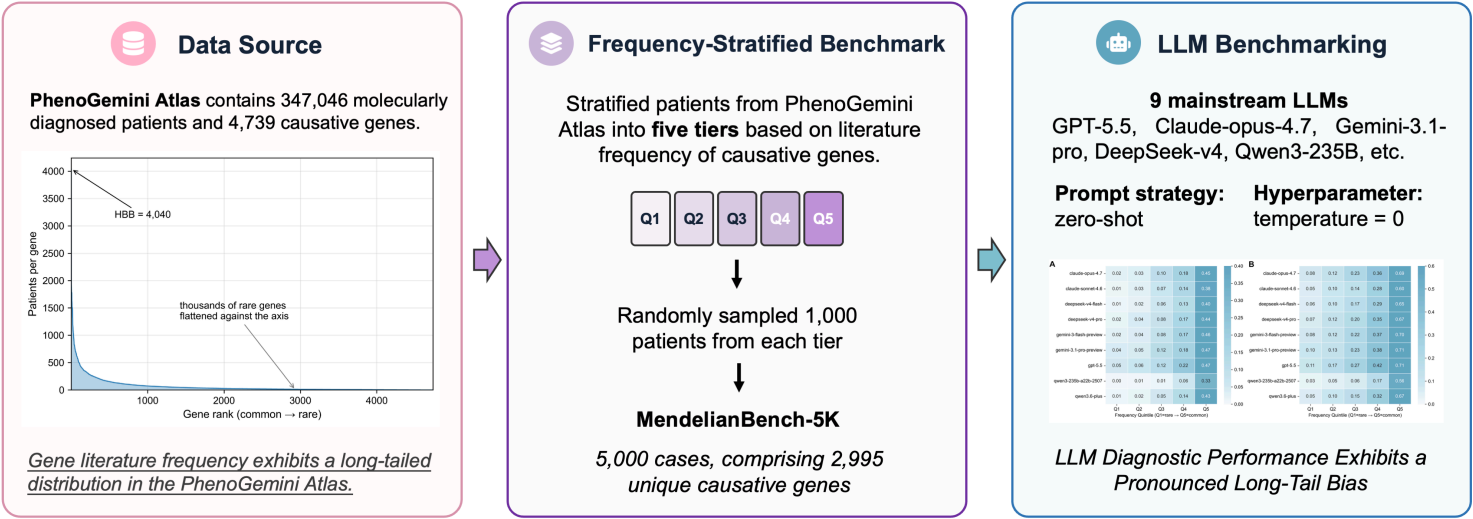


Figure 1

Overview of the construction and evaluation framework of MendelianBench-5K.

PhenoGemini Atlas was used as the source dataset, in which causative genes showed a long-tailed distribution by literature frequency: a small number of highly represented genes, led by *HBB* with 4,040 patients, dominated the head of the distribution, whereas thousands of rare genes were compressed near the x-axis. Patients were stratified into five tiers according to literature frequency, and 1,000 patients were randomly sampled from each tier to construct MendelianBench-5K. Nine mainstream LLMs were then evaluated on a zero-shot phenotype-to-gene diagnostic task, enabling frequency-stratified assessment across both common and rare disease genes.

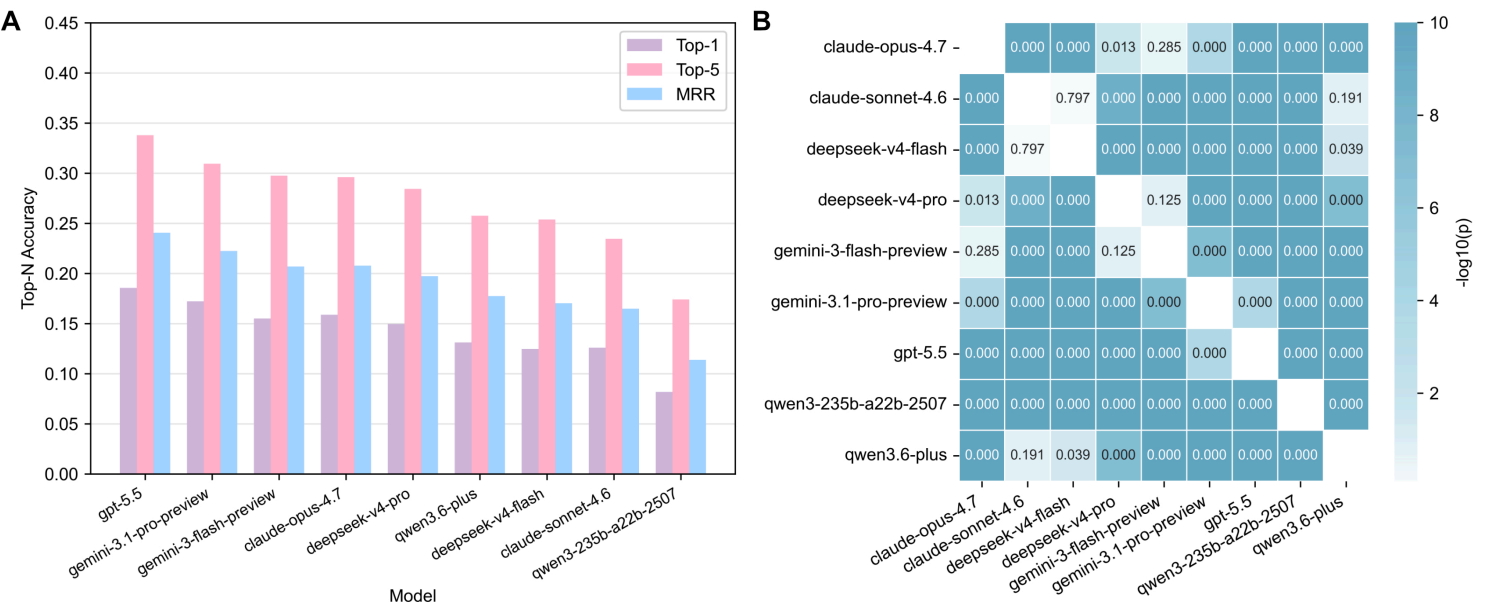


Figure 2

Overall diagnostic performance of large language models on MendelianBench-5K.

(A) Overall Top-1 accuracy, Top-5 accuracy, and mean reciprocal rank (MRR) of nine large language models evaluated on MendelianBench-5K. Models were prompted to return up to five candidate causative genes for each case under zero-shot setting. (B) Pairwise McNemar tests comparing Top-1 correctness between models on the same benchmark cases; color intensity indicates statistical significance. GPT-5.5 achieved the highest overall performance; however, even the best-performing model ranked the causative gene first in fewer than one-fifth of cases.

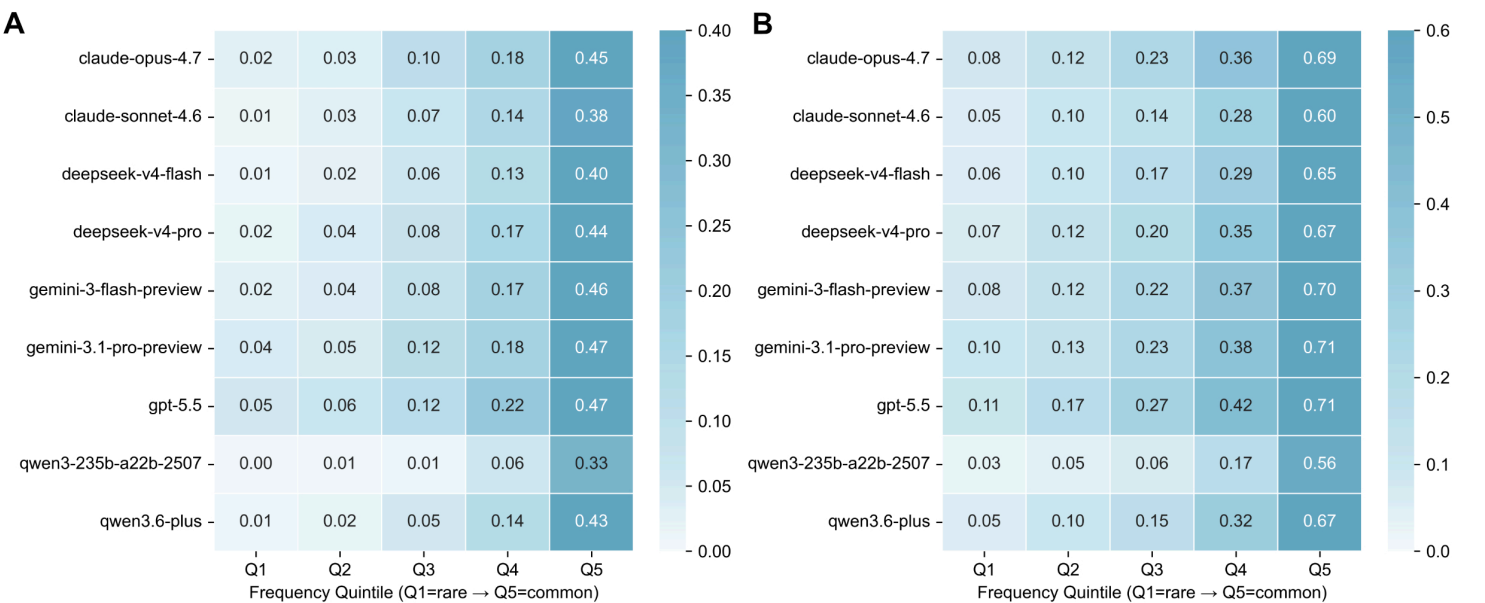


Figure 3

Literature-frequency stratification reveals a pronounced long-tail bias in LLM-based molecular diagnosis.

(A) Top-1 accuracy stratified by causative-gene literature-frequency quintile. (B) Top-5 accuracy stratified by the same quintiles. Q1 denotes the least frequently reported genes and Q5 the most frequently reported genes in the PhenoGemini Atlas; each quintile contained 1,000 benchmark cases. Across all models, diagnostic accuracy increased monotonically from Q1 to Q5. Mean Top-1 accuracy increased from 2.24% in Q1 to 42.48% in Q5, and mean Top-5 accuracy increased from 6.92% to 66.10%. These results indicate that model performance was concentrated among literature-rich genes and degraded substantially for long-tail genes.

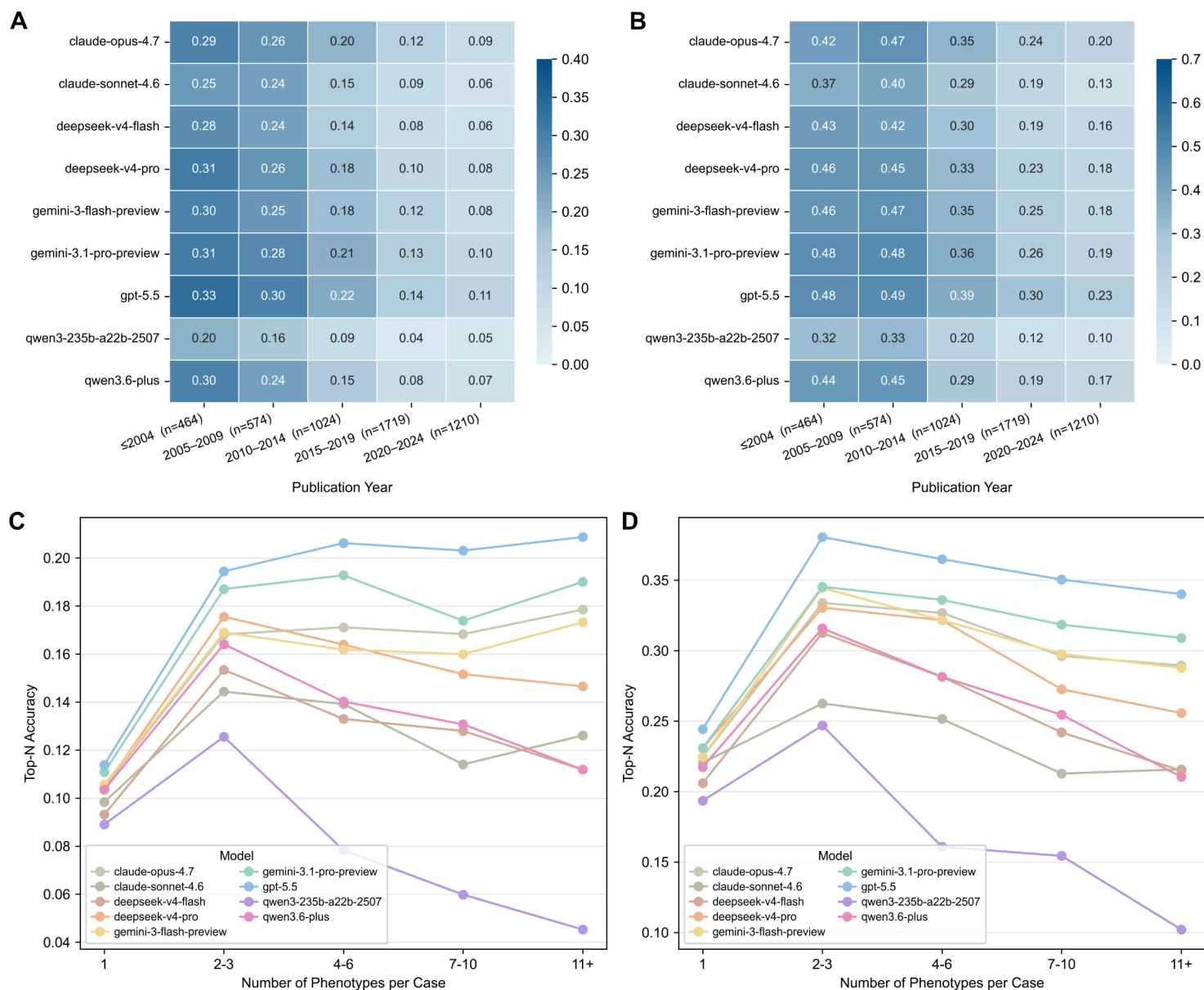


Figure 4

Effects of publication year and phenotypic information content on diagnostic accuracy.

(**A, B**) Model performance stratified by the publication year for cases with available year information ($n = 4,991$), shown for Top-1 and Top-5 accuracy, respectively. Diagnostic accuracy declined for more recently published cases, with mean Top-1 accuracy decreasing from 28.47% for cases published in or before 2004 to 7.83% for cases published during 2020–2024. (**C, D**) Model performance stratified by the number of phenotypes per case, shown for Top-1 and Top-5 accuracy, respectively. Accuracy was lowest for cases with a single phenotype, improved most clearly for cases with 2–3 phenotypes, and then plateaued or declined as additional phenotypic terms were provided.

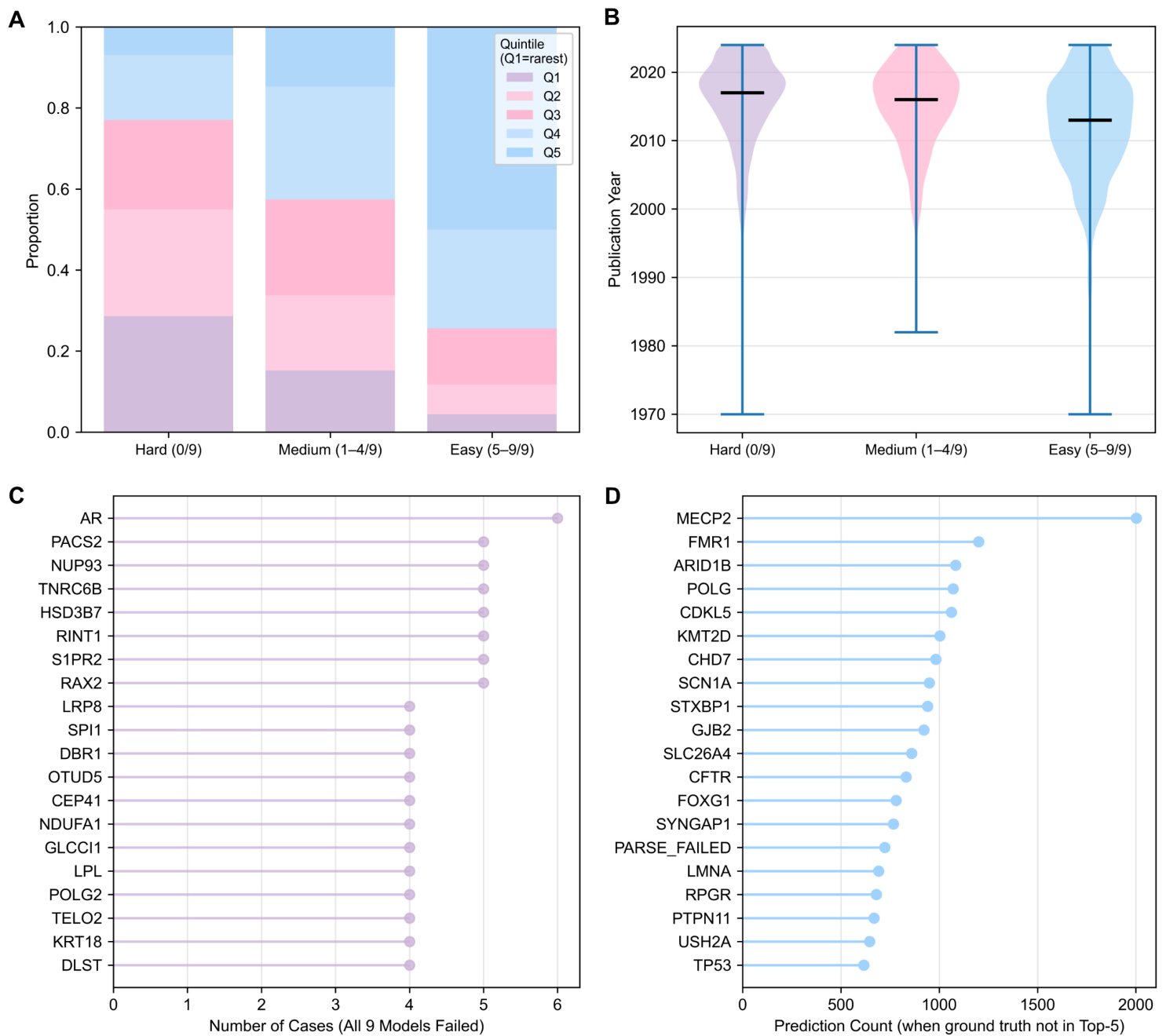


Figure 5

Universal failure cases and recurrent false-positive genes reveal structured diagnostic errors.

(A) Distribution of literature frequency quintiles across difficulty groups. Hard cases were enriched for rarely reported genes, whereas Easy cases were enriched for well-studied genes. **(B)** Publication-year distributions across difficulty groups, showing that Hard cases tended to originate from more recent reports. **(C)** Top-20 genes in Hard cases missed by all nine models. **(D)** Most frequent false-positive genes predicted when the true causative gene was absent from the Top-5 list. Recurrent false-positive predictions were concentrated in canonical, literature-rich genes, suggesting that models tend to default to familiar genes under diagnostic uncertainty. Benchmark cases were classified as Hard, Medium, or

Easy according to the number of models that included the causative gene within their Top-5 predictions:
Hard, 0 of 9 models correct; Medium, 1–4 models correct; Easy, 5–9 models correct.