Large language models identify causal genes in complex trait GWAS

Suyash S. Shringarpure[1], Wei Wang[1], Sotiris Karagounis[1], Xin Wang[1], Anna C. Reisetter[1],
Adam Auton[1], and Aly A. Khan[1,2]

[1] *23andMe Inc., Palo Alto, CA, USA*
[2] *Departments of Family Medicine, and Pathology,*
and Institute for Population and Precision Health,
University of Chicago, Chicago, IL, USA
E-mail: suyashss@gmail.com; aakhan@uchicago.edu

Pinpointing causal genes at genome-wide association study (GWAS) loci remains a major bottleneck. Existing literature-mining approaches are often limited in accuracy and scalability. We show that large language models (LLMs) can accurately prioritize likely causal genes at GWAS loci. We systematically evaluated several widely available general-purpose LLMs against benchmark datasets of high-confidence causal genes, including a unique set from 23 unpublished GWAS. Our results demonstrate that LLMs outperform or match current state-of-the-art methods and, crucially, exhibit robust performance on novel loci not previously linked to traits, underscoring their generalizability. Moreover, when integrated with existing methods, LLMs substantially enhance overall performance. This work establishes LLMs as an accurate, scalable, and broadly generalizable approach to accelerate causal gene identification in complex traits.

Keywords: Large Language Models; GWAS; Causal Gene Identification; Gene Prioritization.

1. Introduction

Genome-wide association studies (GWAS) have identified numerous genomic regions associated with complex traits, enhancing our understanding of trait biology. However, pinpointing the exact causal genes within these regions remains a major hurdle. Approaches to causal gene identification from GWAS loci utilize a broad range of information including functional annotation, colocalization with quantitative trait loci (QTL) datasets, biological insights, and literature evidence. Literature mining for the co-occurrence of a (disease, gene) pair in a publication can provide evidence for the causal role of the gene, recapitulating knowledge an expert biologist might use. However, current literature mining approaches[1,2] have been evaluated in limited settings or through indirect tasks, and their generalizability to diverse phenotypes remains unclear.

Large language models (LLMs) are deep learning models trained on large text corpora for tasks including text generation, summarization, and question-answering. Recent studies have demonstrated their capability to perform biomedical tasks,[3] including summarizing gene function,[4] medical question answering,[5] cell-type annotation,[6] predicting CRISPR screening results,[7] and identifying causal genetic factors from murine experimental data.[8] We hypothesize that LLMs like GPT-4[9] and Claude 3.5,[10] with extensive training on scientific publications, provide a systematic and scalable approach to identify likely causal genes at GWAS loci, potentially overcoming the limitations of manual expert annotation.

We systematically evaluated several general-purpose LLMs, comparing their performance to state-of-the-art methods (Supplementary Figure S1, available online). We assembled three distinct evaluation sets with different ground-truth criteria (Supplementary Table S1, available online). The first set comprises recently published GWAS loci (after April 2023) (the "GWAS Catalog"). The second set contains loci from a benchmark not publicly available online[11] ("Weeks et al."). Finally, we created a new dataset from 23 previously unpublished GWAS from a 23andMe cohort to test performance on novel loci. Our results demonstrate that LLM-based predictions surpass existing methods, including the polygenic priority score (PoPS)[11] and the 'nearest gene' method,[12] in precision, recall, and F-score (Supplementary Figure S2 and Supplementary Table S2, available online). Furthermore, we show that integrating LLM predictions with existing methods can lead to substantial performance improvements.

2. Methods

2.1. *Evaluation Datasets*

We assembled three datasets to benchmark LLMs for causal gene prioritization. A key challenge in evaluating performance is potential data contamination, where evaluation data may have been part of the LLM training corpora. To mitigate this, our datasets were chosen for their varying levels of public availability and recency. Additionally, we tested for direct inclusion of these datasets in the LLMs via prompting and found no evidence of contamination (Supplementary Table S20, available online). For all datasets, we defined a standard locus window of 500 kbp on either side of the lead variant to generate candidate gene lists, a common distance used in causal gene prioritization methods.[11,12,21] We restricted our gene lists to only include protein-coding genes. Our evaluation framework assumes a single causal gene per locus for each prediction.

2.1.1. *Weeks et al. Dataset*

The Weeks et al. benchmark dataset[11] contains 1,348 causal gene-phenotype pairs, with ground truth established through genetic interpretation. The authors identified non-coding credible sets from UK Biobank GWAS that were located within 500 kbp of a high-confidence (Posterior Inclusion Probability $>$ 0.5) fine-mapped coding variant from the same GWAS. The gene containing the coding variant was designated the causal gene. As this dataset is not publicly available online and was obtained directly from the authors via email, it serves as a strong control against LLM training data contamination.

2.1.2. *GWAS Catalog Dataset*

To evaluate performance on recent discoveries, we created a dataset from the GWAS Catalog (v1.0.2, associations_e111_r2024-03-11, downloaded March 19, 2024). To minimize the possibility of these loci being included in the benchmark LLMs trainings, we selected only associations from manuscripts published after April 30, 2023. A causal gene was assigned using a methodology similar to the Weeks et al. dataset: if a non-coding lead variant was within 500 kbp of a coding lead variant in the same study, the gene corresponding to the coding variant was considered causal. This process resulted in a dataset of 641 loci. Candidate genes for each locus were identified using GENCODE release 43.

2.1.3. *Unpublished 23andMe Novel Loci Dataset*

To test the generalizability of LLMs to novel unseen data, we utilized unpublished GWAS data from 23andMe for 23 phenotypes (Supplementary Table S13, available online). We first performed fine-mapping using SuSiE to identify credible sets containing a protein-coding variant (PIP > 0.5). We then identified independent non-coding credible sets within 500 kbp of these. A locus was defined as “novel” if its lead variant had a linkage disequilibrium (LD) $r^2 < 0.5$ with any variant in the entire GWAS catalog. This process resulted in a test set of 403 loci to assess performance on genetic associations not previously documented in the GWAS catalog.

2.2. ***LLM Execution and Prompting Strategy***

2.2.1. *Models and Prompting*

We evaluated version-controlled models from major developers: OpenAI (gpt-3.5-turbo-0125, gpt-4-0613, gpt-4o-2024-08-06), Anthropic (Claude 3.5 Sonnet), Google (Gemini 1.5 Pro), and Meta (Llama 3.1-405b). To ensure more consistent behavior, all models were queried with temperature set to 0. OpenAI and Google models were queried via their native APIs, while Anthropic and Meta models were accessed via AWS Bedrock using the LangChain[23] framework. To prevent any potential for positional bias, all gene lists provided to the models were lexicographically sorted.

We used a two-part prompt structure: a general system prompt outlining the task, and a locus-specific user prompt providing the data. As a sanity check, we confirmed that model performance degraded considerably when provided with randomly shuffled phenotypes instead of the correct ones, validating that the models were using the phenotype information for their predictions (Supplementary Table S21, available online).

System Prompt: LLM Instructions

```
You are an expert in biology and genetics.
Your task is to identify likely causal genes within a locus for a
given GWAS phenotype based on literature evidence.
From the list, provide the likely causal gene (matching one of the
given genes), confidence (0: very unsure to 1: very confident),
and a brief reason (50 words or less) for your choice.
Return your response in JSON format, excluding the GWAS phenotype
name and gene list in the locus. JSON keys should be
'causal_gene','confidence','reason'.
```

```
Your response must start with '{' and end with '}'.
```

User Prompt: Locus-Specific Data

```
Identify the causal gene.
GWAS phenotype: {Morning person}
Genes in locus: {A},{B},{C},{D}
```

2.3. *Benchmark Methods for Comparison*

2.3.1. *Nearest Gene Method*

This baseline method assigns causality to the gene whose body (defined by GENCODE release 43 on the hg38 reference genome) has the smallest physical distance to the lead GWAS variant. For the Weeks et al. dataset, we used the nearest gene predictions provided directly by the authors. In the approximately 4% of cases where a lead variant was located within the bodies of multiple genes, one gene was chosen at random to simplify the evaluation.

2.3.2. *Polygenic Priority Score (PoPS)*

PoPS is a state-of-the-art gene prioritization method that integrates gene-level features with polygenic enrichments to generate a prioritization score for a given phenotype.[11] We obtained pre-computed PoPS scores for the 1,348 loci in the Weeks et al. dataset directly from the original authors. For each locus, the gene with the highest PoPS score was selected as the predicted causal gene.

We did not include other methods, such as ones that leverage expression quantitative trait loci (eQTLs) in our comparison, since previous work has shown that they are outperformed by the nearest gene and PoPS approaches.[11]

2.4. *Performance Evaluation and Statistical Analysis*

Model performance was quantified using standard metrics: Precision (the proportion of correct predictions among all predictions made), Recall (the proportion of true causal genes correctly identified), and the F-score (the harmonic mean of precision and recall). We computed 95% confidence intervals for all metrics using bootstrapping with 1,000 samples. Statistical significance for performance differences between methods was assessed using a Wilcoxon signed-rank test.

To understand factors influencing performance, we analyzed the impact of locus complexity (measured as the number of genes in the locus) and publication bias (measured by the publication count per gene from NCBI's gene2pubmed database) using Spearman correlation. To assess the impact of multiple independent signals at the same locus, which can complicate evaluation, we also analyzed performance on deduplicated datasets where only one signal per unique gene window was retained for each phenotype (Supplementary Figure S4 and Supplementary Table S5, available online).

2.5. *Analysis of Model Behavior and Robustness*

We conducted several analyses to probe the behavior of the LLMs. A prediction was classified as a "hallucination" if the predicted gene was not present in the input list provided to the model; these were penalized as incorrect predictions in our evaluation. We assessed the calibration of the LLM-provided confidence scores by comparing them to the empirical precision at different score levels.

Robustness was tested through multiple experiments. We tested sensitivity to the input gene list by removing the true causal gene and observing the change in prediction confidence. We also randomly shuffled the gene order to check for sensitivity to input structure. To understand if a more complex prompting strategy would improve results, we tested minimal "ablation" prompts and "chain-of-thought" prompting, neither of which improved performance (Supplementary Figure S6, S10 and Supplementary Table S9, available online). Finally, to probe the models' internal representations of biological concepts, we generated gene and phenotype descriptions with GPT-3.5, embedded these descriptions into a high-dimensional space using OpenAI's text-embedding-3-large model,[19] and calculated their cosine similarity.

2.6. *Ensemble Modeling Framework*

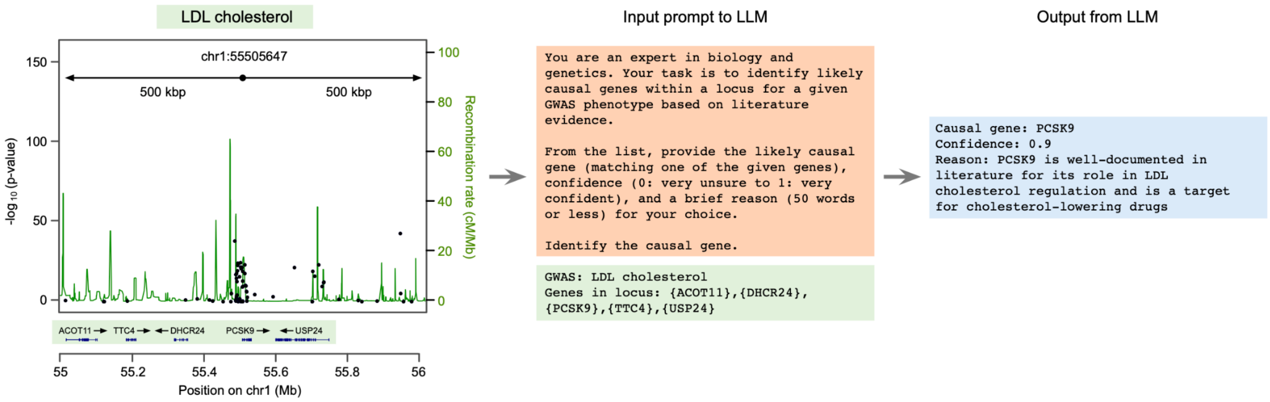
To explore whether combining the distinct information captured by LLMs and traditional methods could yield superior performance, we developed an ensemble learning framework. The problem was framed as a classification task to predict which method (e.g., LLM or nearest gene) would be correct for a given locus. We trained a decision tree classifier using the scikit-learn[25] library with two primary features: (1) a binary indicator of whether the LLM and a non-LLM method agreed on the prediction, and (2) the confidence score provided by the LLM. To prevent data leakage during training, we employed a nested, chromosome-based cross-validation scheme, where all loci from a given chromosome were held out for the test set while the model was trained on the remaining chromosomes. Model features are listed in Supplementary Table S19 (available online).

3. Results

Our study systematically evaluated the ability of LLMs to identify causal genes at GWAS loci. We compared their performance against established methods, analyzed their robustness and reasoning, and developed an ensemble framework to enhance prediction accuracy. The overall methodology is depicted in Figure 1.

3.1. *LLMs Significantly Outperform Existing Methods on Public Benchmarks*

To establish a performance baseline, we first compared a suite of LLMs against established methods on two public benchmark datasets. Our evaluation revealed that recent LLMs, particularly Claude 3.5 Sonnet[10] and OpenAI's GPT-4o,[14] deliver a substantial performance improvement. On the GWAS Catalog dataset, Claude 3.5 Sonnet achieved an F-score of 0.66, representing a 49% improvement over the "nearest gene" method (F-score = 0.44).[12] Similarly, on the Weeks et al. dataset, Claude 3.5 Sonnet (F-score = 0.60) and GPT-4o (F-score



```
You are an expert in biology and
genetics. Your task is to identify likely
causal genes within a locus for a given
GWAS phenotype based on literature
evidence.

From the list, provide the likely causal
gene (matching one of the given genes),
confidence (0: very unsure to 1: very
confident), and a brief reason (50 words
or less) for your choice.

Identify the causal gene.
```

```
GWAS: LDL cholesterol
Genes in locus: {ACOT11},{DHCR24},
{PCSK9},{TTC4},{USP24}
```

```
Causal gene: PCSK9
Confidence: 0.9
Reason: PCSK9 is well-documented in
literature for its role in LDL
cholesterol regulation and is a target
for cholesterol-lowering drugs
```

Fig. 1. Schematic of the LLM-based approach for identifying causal genes at GWAS loci. For a given locus, a 500 kbp window around the lead variant defines the candidate gene set. The LLM receives the phenotype name and an alphabetical list of candidate genes. The model is instructed to return the most likely causal gene, a confidence score, and a brief justification. The orange-colored text in the prompt is fixed, while the green-colored text is locus-specific. The example shows a locus for LDL cholesterol containing the well-established causal gene PCSK9.

= 0.58) showed a marked improvement over PoPS, the best non-LLM method (F-score = 0.50)[11] (Figure 2a-b). The performance difference between Claude 3.5 Sonnet and GPT-4o was not statistically significant, establishing these two models as a clear top tier for this task (Supplementary Table S3, available online).

To understand the drivers of this performance advantage, we examined how performance varied with locus complexity and the extent of existing literature. While all methods performed worse in loci with a higher number of candidate genes, this performance drop was less pronounced for LLMs. This amplified their advantage in gene-dense regions, where LLMs achieved up to a 67% improvement on the GWAS catalog dataset and a 38% improvement on the Weeks et al. dataset relative to the next-best method (Figure 2c; Supplementary Figure S3 and Supplementary Table S4, available online). Furthermore, LLM accuracy positively correlated with the number of publications for the causal gene, indicating these models effectively leverage the scientific literature to inform their predictions (Figure 2d; Supplementary Table S2, available online).

3.2. *LLM Predictions are Robust, Calibrated, and Reasoned*

Beyond raw accuracy, a critical aspect of any predictive model is reliability. We therefore conducted a series of experiments to assess the robustness, calibration, and reasoning capabilities of the top-performing LLMs. We found that LLM-generated confidence scores were well-calibrated at high levels (≥ 0.9) but tended toward overconfidence at lower scores (0.6 to 0.8) (Supplementary Figure S5 and Supplementary Table S6, available online). Importantly, these confidence scores were highly reproducible, with identical inputs yielding identical scores in 96% of test cases (Supplementary Table S7, available online). An examination of the justifications for correct predictions revealed that the models frequently provided valid rationales based on gene function (e.g., “is involved in,” “key regulator of”) or established phenotype

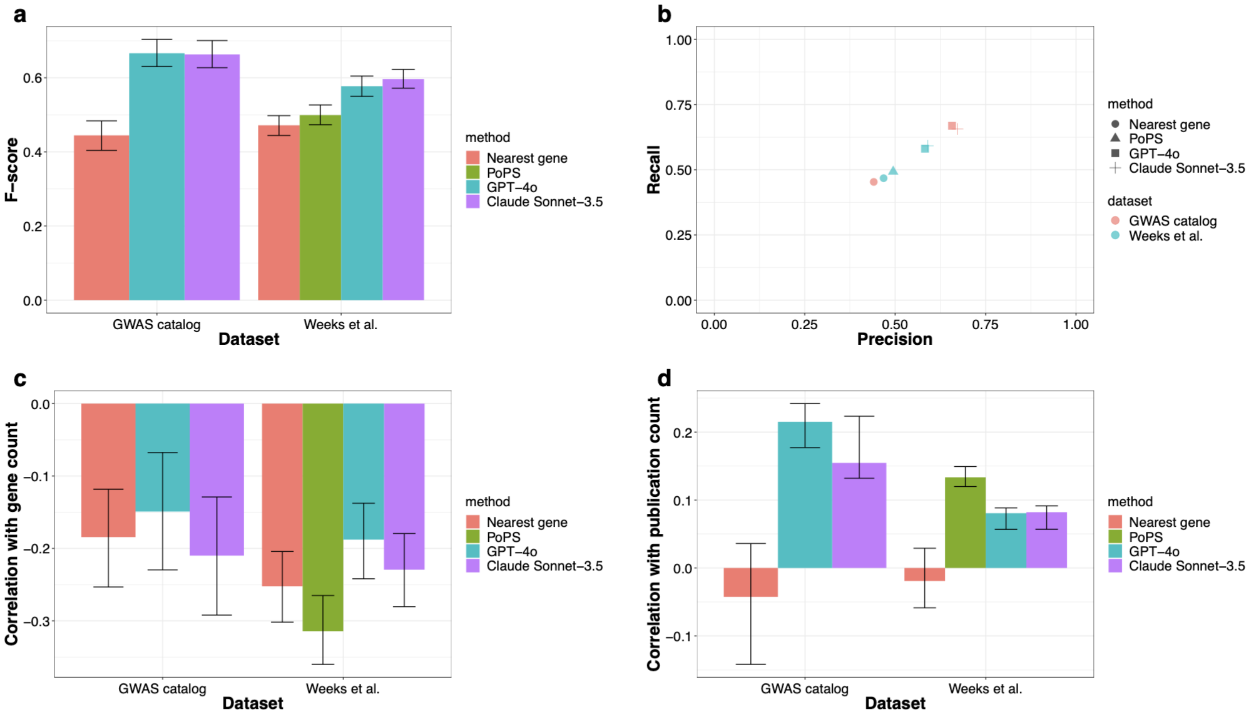


Fig. 2. Performance comparison of leading LLMs and benchmark methods. (a) F-scores show that Claude 3.5 Sonnet and GPT-4o outperform Nearest Gene and PoPS on both evaluation datasets. (b) The precision-recall plot highlights the superior balance of LLM performance. (c) Prediction accuracy for all methods negatively correlates with the number of genes at a locus. (d) LLM and PoPS accuracy positively correlates with the number of publications for the causal gene, unlike the nearest gene method.

associations (Supplementary Table S8, available online).

Robustness of the model was further confirmed through perturbation experiments.[18] Removing the true causal gene from the input list not only caused a significant drop in confidence but also led to a 77% reduction in high-confidence predictions, demonstrating that the model relies on the correct gene's presence (Supplementary Table S11, available online). Conversely, shuffling the input gene order had a minimal impact, resulting in a prediction match 87% of the time, with mismatches concentrated in lower-confidence predictions (Supplementary Table S12, available online). Finally, obvious hallucinations (predicting a gene not provided in the input) were rare, occurring in fewer than 4% of loci for Claude 3.5 Sonnet (Supplementary Table S15, available online).

3.3. *Impact of deduplication on performance*

To assess potential performance inflation due to multiple signals within the same locus, we deduplicated loci with overlapping gene windows. Dataset sizes were reduced (Weeks: 1,348 to 965 loci; GWAS Catalog: 641 to 336 loci), and performance dropped by 10–20% across all methods. However, Claude 3.5 Sonnet and GPT-4o continued to outperform non-LLM methods (Supplementary Figure S4 and Table S5). We therefore retained the original datasets for subsequent analyses to mirror typical GWAS post-processing pipelines.

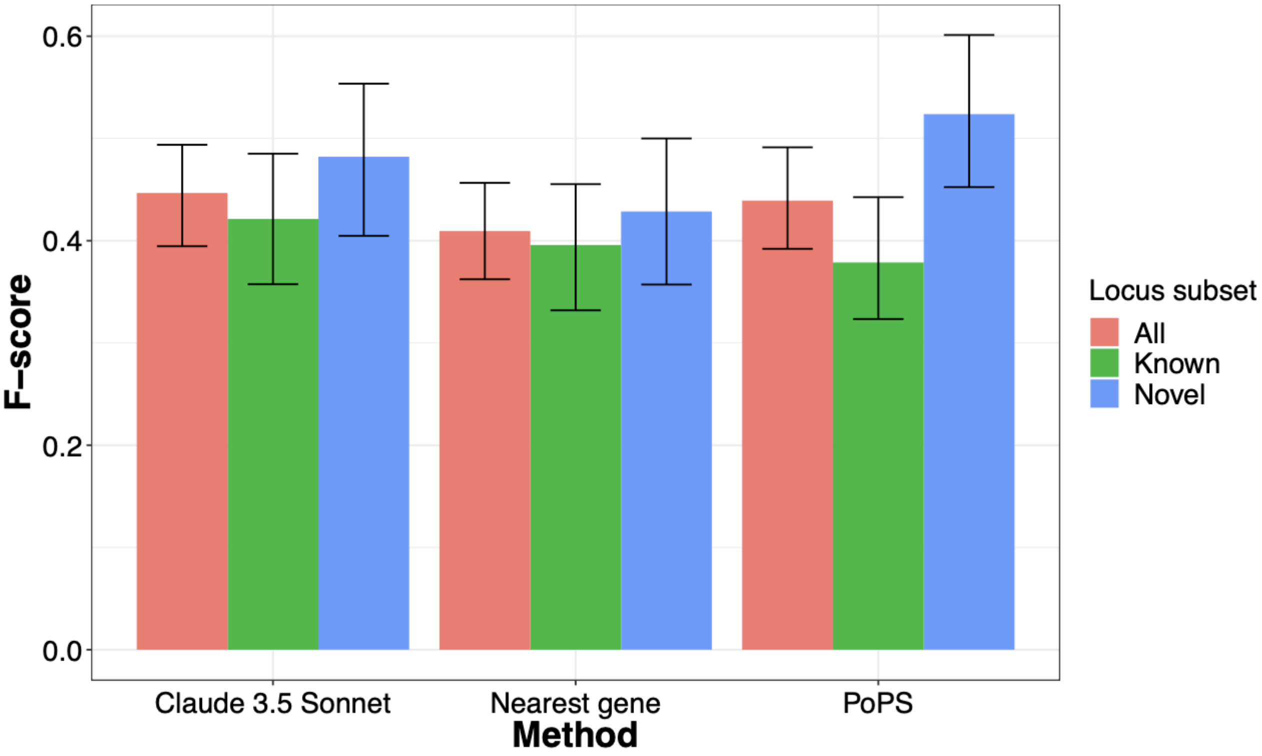


Fig. 3. Performance on the unpublished 23andMe dataset containing novel loci. F-scores are shown for the full dataset (All), the subset of known loci, and the subset of novel loci. Novel loci are defined as having an index variant with LD r-squared < 0.5 with any variant in the GWAS catalog. LLM performance remains stable on novel loci.

3.4. *LLMs Generalize to Novel Loci and Leverage Semantic Similarity*

A key question is whether LLMs simply retrieve known associations or can generalize to novel biological inferences. To test this, we evaluated their performance on an unpublished 23andMe dataset containing loci absent from the GWAS catalog. On this dataset, Claude 3.5 Sonnet again performed best (F-score = 0.45), slightly ahead of PoPS (0.44) and 'nearest gene' (0.41). Crucially, LLM performance did not degrade on these novel loci compared to known loci (F-score 0.48 on novel vs. 0.42 on known). In contrast, the performance of PoPS was significantly better on novel loci (F-score 0.52 vs. 0.38), suggesting LLMs may generalize more consistently across varying levels of prior evidence (Figure 4; Supplementary Table S14, available online). A logistic regression model confirmed that locus complexity (number of genes) was a much stronger predictor of LLM accuracy (McFadden's pseudo-R-squared = 3.6%) than novelty status (pseudo-R-squared = 0.26%).

This generalization capability appears to be driven by a semantic understanding of biology, a concept foundational to modern language models.[20] An approach based solely on the cosine similarity of gene and phenotype text embeddings (generated using the 'text-embedding-3-

large‘ model[19]) achieved 58-68% of the performance of Claude 3.5 Sonnet (Supplementary Figure S7, available online). For a locus associated with LDL cholesterol, the causal gene PCSK9 is closest to the phenotype in the embedding space (Figure 4a; Supplementary Figure S8, available online). Across our datasets, the true causal gene was among the top 5 most semantically similar genes to the corresponding phenotype for 75-93% of loci. This observation indicates that while semantic similarity is a key driver, LLMs utilize additional context from the prompt to refine predictions (Figure 4b; Supplementary Figure S9, available online).

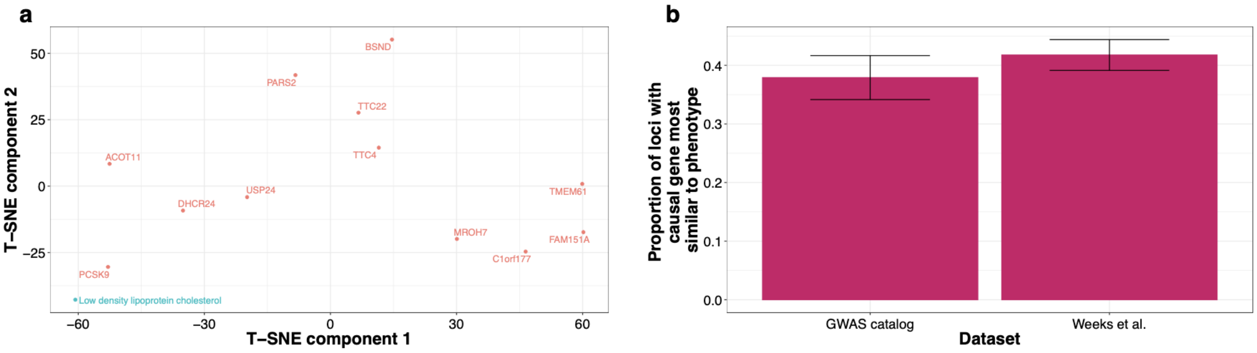


Fig. 4. Text embeddings of genes and phenotypes partially explain LLM performance. (a) A t-SNE plot visualizing text embeddings for a locus associated with LDL cholesterol shows the causal gene PCSK9 is closest to the phenotype in the embedding space. (b) Across both datasets, the causal gene is the most semantically similar to the phenotype in over 35% of cases.

3.5. *Failure Modes of LLMs*

LLMs occasionally struggled with ambiguous phenotype descriptions or overly broad trait interpretation. For example, gpt-4-0613 achieved only 0.08 precision on the “Total protein” phenotype, likely due to misinterpreting the term as general protein synthesis rather than serum protein levels. Providing more specific phenotype descriptions improved performance, and newer models such as Claude 3.5 Sonnet and GPT-4o did not show this issue (precision = 0.52). Another failure mode occurred when highly studied genes dominated predictions. For “Neonatal circulating Complement Component 4 (C4) protein concentration,” Claude 3.5 Sonnet always predicted C4A as causal (precision = 0.0), despite coding variant evidence pointing to other genes. This suggests that LLMs over-rely on well-known gene–phenotype associations, underscoring the need to combine LLMs with functional annotations for improved causal gene prioritization. Precision and recall values for all phenotypes are provided in Supplementary Table S10 (available online).

3.6. *Ensemble Framework Boosts Performance by Integrating Methods*

We sought to investigate whether LLMs provide orthogonal information to existing gene prioritization methods and whether combining them would enhance causal gene prediction. Previous work has shown that combining multiple gene prioritization methods improves performance.[11,12,21] We hypothesized this would also apply to LLM-based approaches. We began by examining the concordance of predictions across different methods. We found that LLM-based methods showed the highest agreement with other LLM-based methods, but only moderate agreement with the polygenic priority score (PoPS) and the ’nearest gene’ methods (Supplementary Figure S12 and Supplementary Table S16, available online). These findings suggest that LLMs and existing methods capture distinct aspects of the data, implying a potential for improved performance through combined approaches.

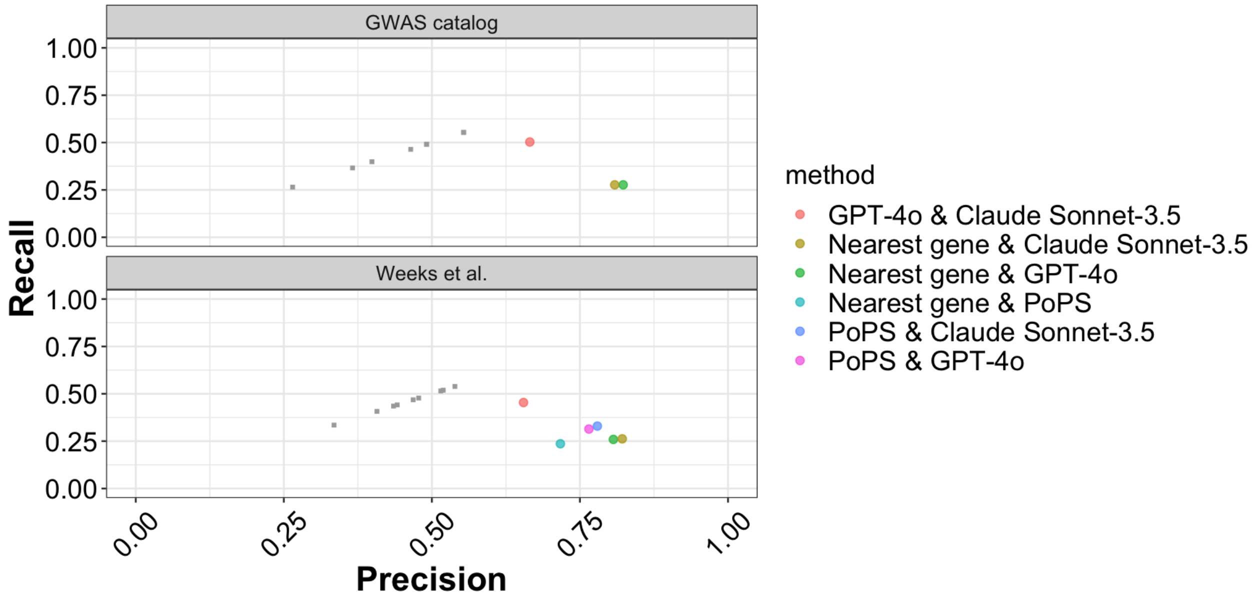


Fig. 5. Performance of consensus prediction methods. Gray squares represent the performance of individual methods. Consensus methods, which only make a prediction when two methods agree, consistently achieve higher precision but lower recall.

To leverage this, we first tested a simple consensus approach, which improved precision to 0.86-0.89 but did so at a significant cost to recall (0.33-0.36), as this approach discards all discordant predictions (Figure 5). To overcome this trade-off, we developed a trained ensemble decision tree framework to intelligently arbitrate between methods. This ensemble approach significantly enhanced overall performance. By combining Claude 3.5 Sonnet with the ‘nearest gene’ method, the F-score on the GWAS catalog dataset improved by 48% (from 0.45 to 0.67) (Supplementary Table S17, available online). A similar integration with PoPS on the Weeks et al. dataset increased the F-score by 17% (from 0.50 to 0.59) (Supplementary Table S18, available online). These results demonstrate that integrating the unique, literature-based intelligence of LLMs with traditional methods offers a powerful and robust strategy to advance causal gene prediction (Figure 6).

4. Discussion

This work establishes that LLMs are a powerful, scalable, and cost-effective new tool for causal gene prioritization. We show that LLMs can accurately synthesize vast scientific literature to

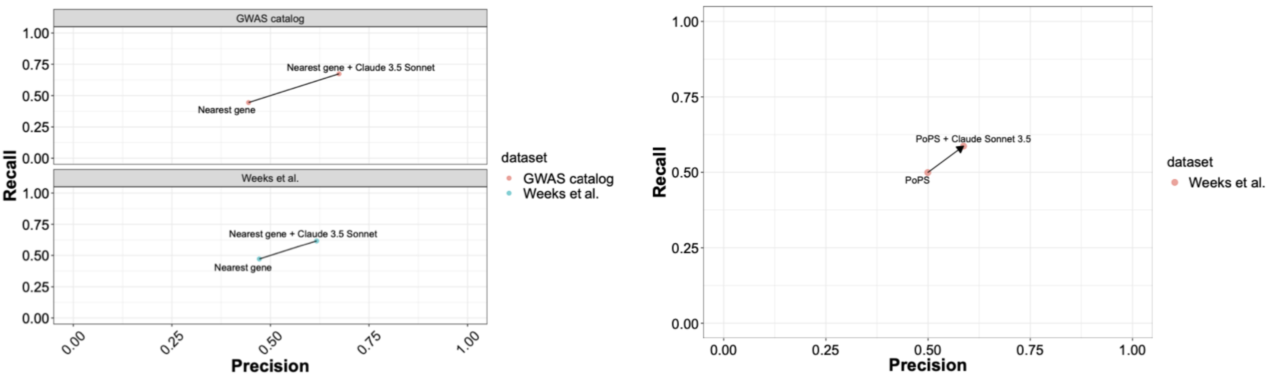


Fig. 6. Performance of the ensemble framework compared to individual methods. The ensemble classifier (red points), which combines predictions from Claude Sonnet-3.5 and the Nearest Gene method, significantly outperforms the individual methods on both datasets.

identify high-probability candidate genes. To address the critical issue of data circularity in AI benchmarking, we introduce a rigorous evaluation framework that includes a novel benchmark built from 23 unpublished GWAS and systematic comparisons with state-of-the-art methods. This framework provides a transparent and unbiased assessment of LLM performance, setting a foundation for best practices in the field.

Our results reveal two key insights. First, on public datasets, advanced LLMs such as Claude 3.5 Sonnet and GPT-4o substantially outperform established methods, with the largest gains observed in gene-dense regions, underscoring their ability to exploit the existing knowledge base. Second, on truly novel loci, including the unpublished 23andMe GWAS dataset, LLMs perform comparably to state-of-the-art tools, confirming their capacity for generalization rather than memorization. Finally, our ensemble approach, which integrates LLM predictions with established methods, significantly improves causal gene identification, positioning LLMs as a complementary and transformative component of next-generation gene prioritization pipelines.

Broader Implications and Potential Applications

The LLM-based framework has broad implications for genomics research. It enables rapid annotation and prioritization of genes from any GWAS, even when full summary statistics are unavailable, a common limitation of approaches like PoPS. This capability can accelerate hypothesis generation by highlighting likely causal genes for functional follow-up. In drug discovery, it could help identify new therapeutic targets by linking disease-associated loci to druggable genes. Overall, this work provides a blueprint for leveraging the exponential growth of scientific literature, transforming unstructured text into structured, actionable biological insights.

A practical advantage of our approach is its scalability and cost-effectiveness. The LLMs used are accessible via paid APIs with costs determined by the number of processed tokens. For example, the average per-locus cost for Claude 3.5 Sonnet was approximately $0.0022 USD, meaning that annotating a GWAS with 300 significant loci can be completed for less than $1 USD. This cost profile makes LLM-based prioritization feasible for routine, large-scale analyses.

Limitations and Future Directions

This study provides a foundational baseline for applying LLMs to causal gene prioritization. Recognizing the risk of data contamination from public training corpora, we employed a robust evaluation design, testing across multiple datasets, including recent publications and unpublished GWAS, to provide a reliable assessment of model generalization.

Several challenges remain before LLMs can be fully integrated into scientific workflows. The "black box" nature of current LLMs limits the ability to trace predictions to specific sources. A clear next step is the development of Retrieval-Augmented Generation (RAG) frameworks that cite verifiable references from curated corpora like PubMed, improving both trust and interpretability.

It is also important to note that LLMs learn statistical associations from text rather than performing formal causal inference. The "causal" labels in our benchmarks are proxies derived from genetic evidence rather than functional validation. Future work should incorporate experimentally validated datasets to further refine these models. Moreover, our evaluation primarily focused on data from individuals of European ancestry. Expanding benchmarks to include diverse populations is essential to ensure equitable and generalizable applications. In the benchmark creation, we focused only on examples where the causal gene is a protein-coding gene. For some published loci, noncoding RNAs have been implicated as causal. The performance of our approach at such loci remains to be explored.

In conclusion, this work demonstrates that LLMs are powerful and scalable tools for causal gene prioritization. By synthesizing the scientific literature, they dramatically improve candidate gene identification for well-documented associations and perform comparably to state-of-the-art methods on truly novel loci. The future of this field lies in improving transparency and mitigating data circularity, likely through RAG-based systems and more robust, prospective benchmarks. By integrating the reasoning capabilities of LLMs with other data modalities, we can accelerate the translation of GWAS discoveries into deeper biological understanding.

Acknowledgments

The authors thank the employees and research participants of 23andMe for making this work possible. We acknowledge Elle Weeks, Hilary Finucane, and Jacob Ulirsch for sharing the evaluation dataset from their publication. We also thank our colleagues at 23andMe for their helpful comments on the manuscript. The authors gratefully acknowledge support from AWS for GPU computing and credits.

Data and Supplementary Materials Availability

Processed datasets, prediction results, and intermediate outputs are available via Zenodo (doi: 10.5281/zenodo.15594712). All supplementary materials, figures, and tables are publicly available online: `https://github.com/akds/llm-gwas-causal-genes`.

References

1. Kafkas, S., Dunham, I. & McEntyre, J. Literature evidence in open targets - a target validation platform. *J. Biomed. Semant.* **8**, 20 (2017).
2. Tirunagari, S. et al. Lit-OTAR Framework for Extracting Biological Evidences from Literature. Preprint at https://doi.org/10.1101/2024.03.06.583722 (2024).
3. Sarwal, V. et al. BioLLMBench: A Comprehensive Benchmarking of Large Language Models in Bioinformatics. Preprint at https://doi.org/10.1101/2023.12.19.572483 (2023).
4. Chen, Y. & Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. Preprint at https://doi.org/10.1101/2023.10.16.562533 (2024).
5. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
6. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods* 1–4 (2024) doi:10.1038/s41592-024-02235-4.
7. Song, S. et al. Virtual CRISPR: Can LLMs Predict CRISPR Screen Results? *Proc. 24th Workshop on Biomedical Language Processing*, 354–364 (Association for Computational Linguistics, 2025). https://aclanthology.org/2025.bionlp-1.30/
8. Tu, T. et al. Genetic Discovery Enabled by A Large Language Model. Preprint at https://doi.org/10.1101/2023.11.09.566468 (2023).
9. OpenAI et al. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
10. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Preprint at https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (2024).
11. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
12. Stacey, D. et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* **47**, e3–e3 (2019).
13. Brown, T. B. et al. Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).
14. OpenAI et al. GPT-4o System Card. Preprint at https://doi.org/10.48550/arXiv.2410.21276 (2024).
15. OpenAI et al. OpenAI o1 System Card. Preprint at https://doi.org/10.48550/arXiv.2412.16720 (2024).
16. Team, G. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint at https://doi.org/10.48550/arXiv.2403.05530 (2024).
17. Grattafiori, A. et al. The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
18. Sclar, M., Choi, Y., Tsvetkov, Y. & Suhr, A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. Preprint at https://doi.org/10.48550/arXiv.2310.11324 (2024).
19. OpenAI. New embedding models and API updates. New embedding models and API updates https://openai.com/index/new-embedding-models-and-api-updates/ (2024).
20. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at http://arxiv.org/abs/1301.3781 (2013).
21. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
22. Bordt, S., Nori, H., Rodrigues, V., Nushi, B. & Caruana, R. Elephants Never Forget: Memorization and Learning of Tabular Data in Large Language Models. Preprint at

https://doi.org/10.48550/arXiv.2404.06209 (2024).

23. Chase, H. LangChain. (2022).
24. Khattab, O. et al. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. Preprint at https://doi.org/10.48550/arXiv.2310.03714 (2023).
25. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).