medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Large language models identify causal genes in complex trait GWAS

Authors: Suyash S. Shringarpure[1,*], Wei Wang[1], Sotiris Karagounis[1], Xin Wang[1], Anna C. Reisetter[2], Adam Auton[1], Aly A. Khan[1,3,4,*]

Affiliations:
[1] 23andMe Inc., 223 N Mathilda Ave, Sunnyvale, CA 94086, USA
[2] Therapeutics Division, 23andMe, 349 Oyster Point Blvd, South San Francisco, CA 94080, USA
[3] Departments of Pathology, and Family Medicine, University of Chicago, Chicago, IL 60637, USA
[4] Institute for Population and Precision Health, University of Chicago, Chicago, IL 60637, USA

[*] Corresponding authors: sshringarpure@23andme.com; aakhan@uchicago.edu

Abstract

Identifying underlying causal genes at significant loci from genome-wide association studies (GWAS) remains a challenging task. Literature evidence for disease-gene co-occurrence, whether through automated approaches or human expert annotation, is one way of nominating causal genes at GWAS loci. However, current automated approaches are limited in accuracy and generalizability, and expert annotation is not scalable to hundreds of thousands of significant findings. Here, we demonstrate that large language models (LLMs) can accurately identify genes likely to be causal at loci from GWAS. By evaluating the performance of GPT-3.5 and GPT-4 on datasets of GWAS loci with high-confidence causal gene annotations, we show that these models outperform state-of-the-art methods in identifying putative causal genes. These findings highlight the potential of LLMs to augment existing approaches to causal gene discovery.

Main

Genome-wide association studies (GWAS) have identified many regions of the genome associated with complex traits, enhancing our understanding of trait biology. However, a significant limitation of GWAS is the difficulty in pinpointing the underlying causal gene for a given association. Approaches to causal gene identification from GWAS loci use 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 potentially provides evidence for the causal role of the gene in the disease, and may recapitulate the knowledge that an expert biologist or clinician might use to identify the causal gene at a GWAS locus. However, current literature mining approaches

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

(Kafkas, Dunham, and McEntyre 2017; Tirunagari et al. 2024) for causal gene prioritization have been evaluated in limited settings or through related tasks such as drug/gene entity recognition and normalization, and their generalizability to all datasets is unclear.

Large language models (LLMs) are deep learning models trained on large text corpora, initially to predict masked/next words from a sentence, and then subsequently trained for a large number of tasks including text generation, summarization, and question-answering. Recent studies have demonstrated their capability to perform biomedical tasks (Sarwal et al. 2023), including summarizing gene function (Chen and Zou 2024), medical question answering (Singhal et al. 2023), cell-type annotation (Hou and Ji 2024), and identifying causal genetic factors from murine experimental data (Tu et al. 2023). We hypothesize that large language models like GPT-3.5 (Brown et al. 2020) and GPT-4 (OpenAI et al. 2024) offer a systematic way to mining literature and identifying causal genes at GWAS loci, as their training datasets include scientific literature and other sources of information about genetics. This approach would enable efficient and scalable annotation of likely causal genes at GWAS loci using literature evidence, which is impractical through expert human annotation.

We performed a systematic evaluation of GPT-3.5 and GPT-4 for the task of causal gene identification, and compared their performance to state-of-the-art computational methods (Supplementary Figure 1). Four evaluation datasets containing 641 to 1692 GWAS loci, with ground-truth annotations of causal genes based on different criteria, were used to test the generalizability of the LLM-based approach (Supplementary Table 1). Given that LLMs are trained on large undisclosed text datasets and it is hard to verify which datasets were used in their training, we selected several well-studied benchmark datasets, including a newly curated dataset created after the training period of GPT-3.5 and GPT-4, as well as a benchmark dataset that is not available on the internet (referred to as GWAS catalog and Weeks et al. respectively, Methods).

The input prompt to the LLMs contained a generic description of an expert geneticist seeking to identify the causal gene, followed by the name of the GWAS phenotype and a list of all genes within 500 kbp of the lead variant at the locus (Figure 2, Methods). Additionally, the LLM was instructed to output the name of the causal gene, a confidence score between 0 and 1, and a short reason for the choice. We queried the LLMs using available APIs to use specific model versions and ensure reproducibility of results (Methods).

For comparison, we evaluated other state-of-the-art approaches for predicting the causal gene, including the polygenic priority score (PoPS, Weeks et al. 2023), locus-to-gene score (L2G, Mountjoy et al. 2021), OpenTargets text-mining (Tirunagari et al. 2024), and the ‘nearest gene’ method (Stacey et al. 2019), all applied on the same datasets. We evaluated all methods based on agreement of predictions with the ground-truth annotations of causal genes in the original dataset, using precision, recall, and F-score (the harmonic mean of precision and recall) as performance measures.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

We found that GPT-3.5 is competitive with existing methods, and GPT-4 outperforms existing methods on all four datasets in F-score (Figure 2a and 2b, Supplementary Table 2 and Supplementary Table 8, p = 8.1e-8, 3.84e-15, 8.72e-4, 1.04e-11 for a paired Wilcoxon signed-rank test on OpenTargets, Pharmaprojects, Weeks et al., and GWAS catalog respectively, when comparing GPT-4 with the best non-LLM approach). The performance of the LLM-based methods correlates with a number of factors related to the GWAS locus. We observed that prediction accuracy is negatively correlated with the number of genes in the locus, a relationship observed for all methods (Figure 2c, Supplementary Table 2). LLM-based methods showed a positive correlation between prediction accuracy and the number of publications for the causal gene, and a similar correlation was observed in other methods (Figure 2d, Supplementary Table 2). Additionally, we also noticed a small number of obvious hallucinations, where the LLMs reported a causal gene that was not included in the set of provided genes at a locus (fewer than 2% of all loci for GPT-3.5, fewer than 0.9% for GPT-4, Supplementary Table 3).

Next we assessed the LLM-based methods' ability to generate a confidence score associated with their prediction. We observed that LLMs were well-calibrated at higher confidence levels (>= 0.8), but overly optimistic in their confidence estimates at lower confidence levels (0.5-0.7), with GPT-4 showing better calibration than GPT-3.5 (Supplementary Figure 2, Supplementary Table 4). We found that focusing only on the high-confidence predictions from GPT-4 (confidence >= 0.8) allowed additional improvements to precision, with improvements from 10% to 43% across the datasets (Supplementary Figure 7).

We examined the purported reasoning underlying the LLMs' confidence and predictions, which might shed light on their ability to interpret complex prompts. The reasons provided by the LLMs for correct predictions include phrases describing the functions of genes (“is involved in”, “the regulation of”, “in immune response” etc.), or their association with the phenotype (“is associated with”, “is implicated in” etc.) among others (Supplementary Table 9). Since LLMs are sensitive to their input prompts, we tested the impact of prompt structure on our results. In our sensitivity analysis, we found that the LLM-based methods maintained their performance even when provided with only a minimal prompt containing the output format, task instruction, phenotype name, and gene names, without any additional context (Supplementary Figure 3). This suggests that the LLMs internally contained most of the information needed for the causal gene identification task without requiring substantial context.

To further explore the internal model representation of genetic and phenotypic associations, we examined the embeddings of phenotypes and genes using the “text-embedding-3-large” model (OpenAI 2024). LLMs represent words as points in a high-dimensional embedding space, where similarity in these representations can capture semantic relationships (Mikolov et al. 2013). In the context of causal gene identification, we hypothesized that causal genes are likely to be proximal to the phenotypes they influence in the embedding space. To test this hypothesis, we use pre-computed high-dimensional embeddings of LLM-generated gene and phenotype descriptions at a GWAS locus. The predicted causal gene at the locus is the gene which is most similar to the phenotype in the embedding space. We found that

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

using similarity of pre-computed embeddings alone achieves about 75-90% of the performance of GPT-3.5 (Supplementary Figure 4).

To illustrate the role of embeddings in prediction, we present t-SNE projections for a locus associated with LDL cholesterol from the Weeks et al. dataset in Figure 3a. For this locus, which has 12 candidates for the causal gene, the text embeddings of the *PCSK9* gene are closest to those of LDL cholesterol (as measured through cosine similarity). Consequently, the LLM-based approaches correctly nominate *PCSK9* as the causal gene at the locus. Supplementary Figure 5 shows a barplot of the gene-phenotype similarities for this locus. Figure 3b quantifies the similarity between the causal gene and phenotype in the embedding space for all loci in our evaluation datasets. We observe that the causal gene is most similar to the phenotype for 40-70% of all examples, depending on the evaluation set, among all genes at the locus. Extending this further, we find that the causal gene is among the top 5 most similar genes to the phenotype in the embedding space for 75%-93% of all examples (Supplementary Figure 6).

Although similarity in high-dimensional embedding space explains a large proportion of the performance of LLM-based approaches, we find that LLMs improve on these phenotype embeddings. For instance, the GWAS catalog dataset contains 250 loci for a phenotype related to sex differences that is only described as “Multi-trait sex score” (originally defined as the sum of multiple quantitative traits, weighted by their respective sex-difference effect sizes). The embedding-based approach makes incorrect predictions for most of these loci (precision for phenotype = 0.05) while GPT-4 achieves a precision of 0.65 for the same phenotype (Supplementary Table 5). This suggests that with the additional context of the task description and the gene information, the LLM is correctly able to infer that the short phenotype description refers to sexual dimorphism and sex-specific traits.

In considering the broad applicability of this approach for causal gene prediction beyond currently published phenotypes, we found a couple of notable failure modes when examining phenotypes where GPT-4 had very low precision. First, GPT-4 had a precision of 0.08 for the “Total protein” phenotype. The reasons provided by the LLM suggest an incomplete understanding of the short phenotype description, interpreting it only in terms of broad protein levels rather than specifically about protein levels in blood (Supplementary Table 6). This indicates that more specific phenotype descriptions could improve the performance of LLMs.

In a second instance, GPT-4 had a precision of 0.0 for the phenotype "Neonatal circulating Complement Component 4 (C4) protein concentration". For this phenotype, all the GWAS loci in the evaluation dataset include the *C4A* gene in the 500 kbp window around the lead variant, though they have different causal genes based on coding variant signals. Since the *C4A* gene is a major part of the complement system, the LLM-based approach always predicts *C4A* as the causal gene. This suggests that combining LLMs with functional annotation data could improve causal gene prioritization. Supplementary Table 5 shows the precision and recall for all methods stratified by phenotype.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

While we have demonstrated the potential of LLMs for causal gene identification, it is important to consider the limitations and challenges associated with their use in our study. The lack of disclosure regarding the training datasets for most LLMs makes it challenging to verify whether our evaluation datasets were somehow included in their training. To mitigate this risk, we introduced a study design using data curated after the LLM training period as well as a benchmark dataset not available on the internet. Additionally, while LLMs return plausible reasons along with their predictions, it is challenging to pinpoint the exact information the LLM used to make each prediction. This is a current shortcoming, but it also points to an area of active research in the field. As reasoning capabilities continue to improve in LLMs, we anticipate that their ability to provide transparent and verifiable explanations will evolve rapidly.

Overall, our study demonstrates for the first time that LLMs can significantly enhance causal gene identification and GWAS interpretation by systematically incorporating literature evidence. LLM-based approaches, requiring only the location of the lead variant, can be applied to any GWAS locus. With improved prompting, these methods could become valuable tools for causal gene identification. Furthermore, combining LLM-based approaches with functional data holds the potential to create even more accurate and robust methods for identifying causal genes, advancing our understanding of genetic contributions to complex traits.

References

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Chen, Yiqun, and James Zou. 2024. “GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT.” bioRxiv. https://doi.org/10.1101/2023.10.16.562533.

Hou, Wenpin, and Zhicheng Ji. 2024. “Assessing GPT-4 for Cell Type Annotation in Single-Cell RNA-Seq Analysis.” *Nature Methods*, March, 1–4. https://doi.org/10.1038/s41592-024-02235-4.

Kafkas, Şenay, Ian Dunham, and Johanna McEntyre. 2017. “Literature Evidence in Open Targets - a Target Validation Platform.” *Journal of Biomedical Semantics* 8 (1): 20. https://doi.org/10.1186/s13326-017-0131-3.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” arXiv. http://arxiv.org/abs/1301.3781.

Minikel, Eric Vallabh, Jeffery L. Painter, Coco Chengliang Dong, and Matthew R. Nelson. 2024. “Refining the Impact of Genetic Evidence on Clinical Success.” *Nature*, April. https://doi.org/10.1038/s41586-024-07316-0.

Mountjoy, Edward, Ellen M. Schmidt, Miguel Carmona, Jeremy Schwartzentruber, Gareth Peat, Alfredo Miranda, Luca Fumis, et al. 2021. “An Open Approach to Systematically Prioritize Causal Variants and Genes at All Published Human GWAS Trait-Associated Loci.” *Nature Genetics* 53 (11): 1527–33. https://doi.org/10.1038/s41588-021-00945-5.

OpenAI. 2024. “New Embedding Models and API Updates.” *New Embedding Models and API Updates* (blog). January 2024. https://openai.com/index/new-embedding-models-and-

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

api-updates/.
OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. “GPT-4 Technical Report.” arXiv. https://doi.org/10.48550/arXiv.2303.08774.
Sarwal, Varuni, Viorel Munteanu, Timur Suhodolschi, Dumitru Ciorba, Eleazar Eskin, Wei Wang, and Serghei Mangul. 2023. “BioLLMBench: A Comprehensive Benchmarking of Large Language Models in Bioinformatics.” bioRxiv. https://doi.org/10.1101/2023.12.19.572483.
Singhal, Karan, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. 2023. “Large Language Models Encode Clinical Knowledge.” *Nature* 620 (7972): 172–80. https://doi.org/10.1038/s41586-023-06291-2.
Stacey, David, Eric B Fauman, Daniel Ziemek, Benjamin B Sun, Eric L Harshfield, Angela M Wood, Adam S Butterworth, Karsten Suhre, and Dirk S Paul. 2019. “ProGeM: A Framework for the Prioritization of Candidate Causal Genes at Molecular Quantitative Trait Loci.” *Nucleic Acids Research* 47 (1): e3–e3. https://doi.org/10.1093/nar/gky837.
Tirunagari, Santosh, Shyamasree Saha, Aravind Venkatesan, Daniel Suveges, Annalisa Buniello, David Ochoa, Johanna McEntyre, Ellen McDonagh, and Melissa Harrison. 2024. “Lit-OTAR Framework for Extracting Biological Evidences from Literature.” bioRxiv. https://doi.org/10.1101/2024.03.06.583722.
Tu, Tao, Zhouqing Fang, Zhuanfen Cheng, Svetolik Spasic, Anil Palepu, Konstantina M. Stankovic, Vivek Natarajan, and Gary Peltz. 2023. “Genetic Discovery Enabled by A Large Language Model.” bioRxiv. https://doi.org/10.1101/2023.11.09.566468.
Weeks, Elle M., Jacob C. Ulirsch, Nathan Y. Cheng, Brian L. Trippe, Rebecca S. Fine, Jenkai Miao, Tejal A. Patwardhan, et al. 2023. “Leveraging Polygenic Enrichments of Gene Features to Predict Genes Underlying Complex Traits and Diseases.” *Nature Genetics* 55 (8): 1267–76. https://doi.org/10.1038/s41588-023-01443-6.

Methods

Evaluation datasets

We used four datasets for evaluation (1) the OpenTargets gold-standard dataset, (2) the Citeline Pharmaprojects dataset of drug targets and their approved indications, (3) an evaluation set created by Weeks et al. based on proximity to fine-mapped coding variant associations, and (4) an evaluation set we created using associations added to the GWAS catalog from manuscripts published after April 2023.

The OpenTargets gold-standard dataset is based on GWAS loci for which there is high confidence (through functional criteria) in the causal gene. We downloaded the OpenTargets gold-standard dataset from https://github.com/opentargets/genetics-gold-standards/ (file: https://github.com/opentargets/genetics-gold-standards/blob/master/gold_standards/processed/gwas_gold_standards.191108.tsv). We subsetted the dataset to only rows that had a high-confidence annotation (highest_confidence = "High"). We added gene symbol annotations to the dataset with GENCODE release 43, and

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

excluded rows where a gene symbol was not found. This resulted in a dataset with 851 rows. To create descriptions of the phenotypes, we combined trait information from standard trait names and reported trait names. We appended reported trait names to standard trait names if the reported trait names are non-redundant, and we also removed irrelevant text from reported trait names, such as “[EA]”, “non-cancer illness code, self-reported”, “GWAS/Metabochip 2012” and “conditional on rs7709212”.

The Pharmaprojects dataset contains drug targets and indications, as well as their drug development stage. We used the Pharmaprojects dataset released by Minikel et al. (Minikel et al. 2024) and created additional mappings of disease indications to EFO (Supplementary Notes). We subsetted the dataset to only rows that have a mapped MeSH or EFO term and that correspond to launched drugs (hcat = “Launched”). We added gene symbol annotations to the dataset with GENCODE release 43, and excluded rows where a gene symbol was not found. This resulted in a dataset with 1692 causal gene - phenotype pairs. Since the Pharmaprojects data does not contain GWAS information, we created synthetic GWAS hits near the target gene for each row to mimic real GWAS hits in terms of the distance between the GWAS hit and the underlying causal gene. More details about the methods for creation of the synthetic GWAS hits can be found in the Supplementary Note.

The Weeks et al. evaluation set contains non-coding credible sets that are within 500 kbp of a high-confidence (posterior inclusion probability > 0.5) fine-mapped coding association in the same GWAS in UK Biobank (Weeks et al. 2023). We obtained this dataset by emailing the authors. Since this dataset is not available directly on the internet, it is the least likely to have been directly used for training the LLMs evaluated in our study (though the underlying GWAS summary statistics and publication could be part of the LLM training data). We used the gene symbol annotations provided by the authors. This resulted in a dataset with 1348 causal gene - phenotype pairs.

The GWAS catalog dataset contains non-coding lead variants that are within 500 kbp of a coding lead variant in the same GWAS study. We downloaded the GWAS catalog lead variant file version “gwas_catalog_v1.0.2-associations_e111_r2024-03-11.tsv” from https://www.ebi.ac.uk/gwas/docs/file-downloads (download date: 03/19/2024). To avoid the possibility of these results being included in LLM training datasets, we subset the data to only include associations added to the GWAS catalog for manuscripts published after April 30, 2023, the latest reported training cutoff date among the LLMs we evaluated in our study. This resulted in a dataset with 641 causal gene - phenotype pairs.

Generating LLM input from datasets

For input to the LLMs, we converted each (phenotype, lead variant, causal gene) triplet corresponding to a GWAS locus to a (phenotype, list of genes in locus) pair.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

For OpenTargets, Pharmaprojects, and the GWAS catalog datasets, we identified all genes within 500 kbp of the lead variant (using GENCODE release 43) based on the smallest distance between the gene body and the lead variant. For the Weeks et al. dataset, the dataset provided by the authors included all genes within 500 kbp of the lead variant, so we used that list of genes without any additional processing.

To avoid leaking information about the lead variant location or causal gene position through the ordering of gene symbols by physical position, we sorted all gene symbols lexicographically before including them in the prompt to the LLMs.

As an additional sanity check to verify that the LLM relies on phenotype information and not just on gene information (such as location) for predictions, we randomly sampled phenotypes at the evaluation loci and provided those as input to GPT-4. As expected, we found that LLM performance is considerably degraded from an average precision of 63% across all datasets to an average precision of 32%, along with an increase in number of obvious hallucinations from 8.25 to 34 (Supplementary Table 10).

LLM prompts for identifying causal genes

To query LLMs to identify the causal genes for a (phenotype, list of genes in locus) pair, we used a prompt describing the task for the LLM along with the phenotype and gene list. The basic prompt we used had two components, a system prompt (for general behavior), and a user prompt (pair-specific). For an example pair (“Morning person”,[A,B,C,D]), the prompts are described below.

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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

LLMs evaluated

For our experiments, we evaluated LLMs from OpenAI, GPT-3.5 and GPT-4. For GPT-3.5, we used the model version “gpt-3.5-turbo-0125”, which has been trained on data up to September 2021. For GPT-4, we tested the model version “gpt-4-1106-preview”, which has been trained on data up to April 2023, and model version “gpt-4-0613”, which has been trained on data up to September 2021. We found that the “gpt-4-0613” model performed comparably or better than the “gpt-4-1106-preview” model and had fewer obvious hallucinations (Supplementary Table 7), hence we report the results for “gpt-4-0613” as GPT-4 in the main text.

To make outputs reproducible, we queried all LLMs with temperature set to 0.

Post-processing of LLM results

We consider an LLM prediction to be an obvious hallucination if the predicted gene was not in the list of genes at the locus provided to the LLM. These are easily detectable, and we set such predictions to NA, excluding them from downstream evaluation.

Comparison to other methods

We compared our results to several state-of-the-art methods for each dataset. First, we evaluated the “nearest gene” method for all datasets. We evaluated text-mining on the OpenTargets and Pharmaprojects dataset, where there was sufficient overlap between the phenotypes in the dataset and phenotypes included in the text-mining scores. We also added prediction based on “locus-to-gene” (L2G) scores to the OpenTargets gold-standard dataset, and prediction based on polygenic priority score (PoPS) to the Weeks et al. evaluation dataset. Both these methods were previously evaluated to have good performance on their respective datasets.

To obtain the nearest gene prediction, we computed the distance between the genes in each locus and the lead variant (using GENCODE release 43 and reference genome hg38 for the OpenTargets, Pharmaprojects, and GWAS catalog datasets), and defined the nearest gene as the gene with the least distance from the lead variant based on gene body. For the Weeks et al. data, we directly used the nearest gene predictions provided by the authors (downloaded from https://www.finucanelab.org/data, file: https://www.dropbox.com/sh/o6t5jprvxb8b500/AACqCux_jJbF9F56ozhzzkpia/results/UKB_AllMethods_GenePrioritization.txt.gz?dl=0). In the Weeks et al. data, we found that “nearest gene” nominated multiple genes at about 4% of loci (51 of 1348) due to the lead variant position being within multiple gene bodies. To simplify evaluation, we randomly chose a single gene out of all nominated genes as the prediction at such loci. We found that this had a minor impact on

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

precision and recall compared to that reported in the original publication (precision = 0.47 vs previously reported 0.46, recall = 0.47 vs previously reported 0.48).

To get the text mining prediction, we first downloaded (gene,disease) co-occurrence information and scores (Supplementary Note) from OpenTargets. We aggregated the scores of each gene/disease pair by summing over all scores for the pair from all publications. We defined the predicted causal gene from text mining at a locus as the gene with the largest aggregated score in each locus for the GWAS phenotype.

The L2G score is a machine learning approach trained using fine-mapped genetics and functional genomics data on 445 gold-standard curated GWAS loci (Mountjoy et al. 2021). The predicted causal gene based on L2G score was defined as the gene with maximal L2G score in each locus (indexed by the phenotype and the lead variant). We downloaded the L2G scores from https://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/latest/l2g/ (download date: 03/26/2024).

PoPS is a similarity-based gene prioritization method that leverages gene-level summary statistics and incorporates data about genes from a variety of sources to produce a phenotype-gene level prioritization score. We obtained PoPS scores for the 1348 evaluation loci by emailing the authors of the original publication (Weeks et al. 2023). The predicted causal gene based on the PoPS score was defined as the gene with the highest PoPS score in the locus.

Evaluation of predictions

For evaluation of predictions, we computed precision (proportion of predicted causal genes that were annotated as causal in the dataset), recall (proportion of annotated causal genes identified among predictions), and F-score (harmonic mean of precision and recall). Some methods (nearest gene, PoPS, LLM-based approaches) made a single causal gene prediction for each example, while others (text mining, L2G score) only made predictions at a subset of all examples. To compute these metrics, the predictions from each method were converted to 0/1/NA, with 1 assigned if the method made a prediction and it matched the annotated causal gene, 0 assigned if the method made a prediction but it did not match the annotated causal gene, and NA assigned if the method did not make a prediction for an example. For all methods except text mining, we found that NAs were only a small proportion of the predictions (Supplementary Table 4).

Precision = (Number of predictions scored 1) / (Number of non-NA predictions)

Recall = (Number of predictions scored 1) / (Number of examples in dataset)

F-score = 2 * Recall * Precision / (Recall + Precision)

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

For each metric, we computed 95% confidence intervals using the bootstrap with 1,000 samples.

To compare the performance of GPT-4 to the best non-LLM approach, we used a Wilcoxon signed rank test with continuity correction, and we report p-values for the alternative hypothesis that the GPT-4 performance is higher than the highest non-LLM approach performance. We also report a p-value from a McNemar test for the same comparison with the alternative hypothesis of a difference between performance for the two methods.

Factors affecting prediction accuracy

To assess the impact of locus complexity, we used the number of genes in the locus as a measure of its complexity, and computed the Pearson correlation between the number of genes in the locus and whether the prediction was correct.

To assess the impact of publication count, we used publication counts for causal genes to evaluate correlations with prediction performance. We downloaded the “gene2pubmed” file from https://ftp.ncbi.nlm.nih.gov/gene/DATA/ (download date: January 26, 2024). We subset the file to only include human genes (tax_id = 9606), and count the number of publications per gene. Since many examples might share the same causal gene, for each causal gene, we computed the proportion of loci predicted correctly. To account for the heavy-tailed distribution of publication count by gene, we computed the Spearman correlation between the number of publications for the causal gene and the proportion of correct predictions at loci with the specified causal gene.

We used the bootstrap with 1,000 samples to compute 95% confidence intervals for both correlations. We used the same resampled dataset for confidence interval calculations for precision, recall, F-score and both correlations.

Calibration analysis

To assess calibration, we used the confidence scores provided by the LLM. We computed precision for all predictions at a given confidence score, for any confidence score with at least 5 predictions. Supplementary Table 4 shows all unique confidence scores predicted by the LLMs, with their counts and precision estimates. We computed a standard error for each precision by assuming it to be the mean of a binomial distribution with size given by the number of predictions with that score. 95% confidence intervals were calculated as precision +/- 1.96*se.

Summarizing reasons provided by the LLM for correct predictions

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Across all 4 datasets, the LLMs make thousands of correct predictions. To summarize these for easy inspection, we concatenated all reason strings provided by GPT-4 for predictions evaluated to be correct. We then identified trigrams in the concatenated string using the ngram package in R, and sorted them in decreasing order of frequency of occurrence to identify the most common ones. We report the top 200 most frequent trigrams in Supplementary Table 9.

Sensitivity analysis for prompt structure

To examine how the prompt affects the prediction accuracy of the LLMs, we experimented with replacing the prompt by a minimal alternative containing only the output format, task description, and the locus information. We conducted these experiments with GPT-3.5 (model version “gpt-3.5-turbo-0125”). The detailed prompt is described below, using the same example locus as earlier.

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

Embeddings for genes and phenotypes

We generated descriptions for genes and phenotypes using GPT-3.5 (model version “gpt-3.5-turbo-0125”) by using the prompt below.

"You are an expert in biology and genetics.
Your task is to provide biologically relevant information about the query below in 300 words or less.

Query: {text}."

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The "text" value was generated by adding the entity type (gene/phenotype) to the beginning of the entity. So the value of text for the *BRCA1* gene would be "gene *BRCA1*", and that for the breast cancer phenotype would be "phenotype breast cancer".

We provided the generated descriptions as input text to get embeddings using the OpenAI text embedding model "text-embedding-3-large". This produced embeddings with 3,072 dimensions.

Embedding-based causal gene prediction

Using the computed embeddings, we provided the embedding-based causal gene prediction. We calculated the dot product (identical to cosine similarity since the embeddings are normalized to length 1) between the phenotype embedding and the gene embedding, and we defined the predicted causal gene as the gene with the largest dot product of phenotype and gene embeddings among all the genes in each loci.

To create the t-SNE plot for the *PCSK9* locus, we ran t-SNE on the 13 data points (1 phenotype + 12 genes in the locus) using the Rtsne package with a perplexity of 4.

Figures

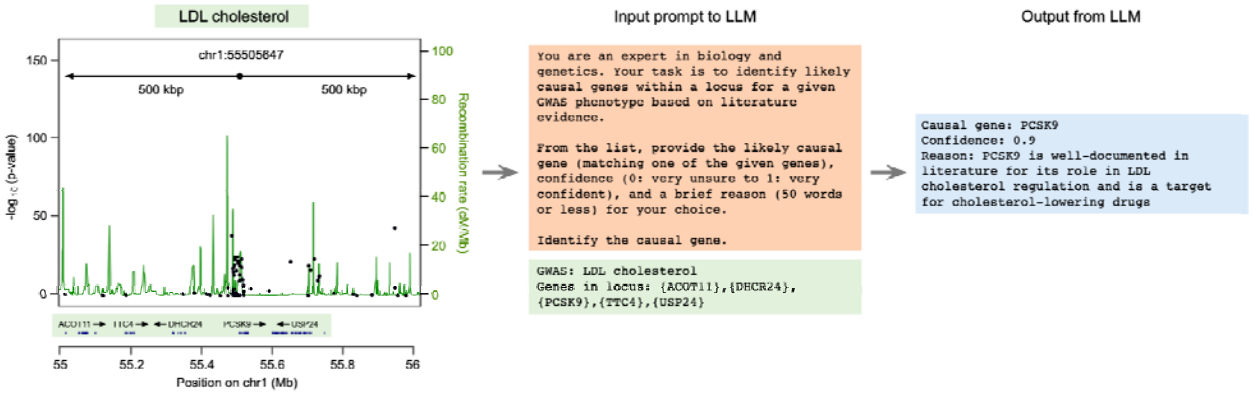

Figure 1: Schematic illustrating the use of LLMs for identifying causal genes at GWAS loci, with an example of an association for LDL cholesterol near the PCSK9 gene. For a GWAS locus, a 500 kbp window is extended on either side of the lead variant (indicated by the black dot), and all genes within this window are considered as candidates. The LLM is then provided with the phenotype name and the alphabetical list of genes and is instructed to predict the causal gene, its confidence, and the reason for its choice. The part of the input prompt colored in orange is the same for all loci, while the part in green is modified for each locus. For readability, output formatting instructions are excluded from the figure.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

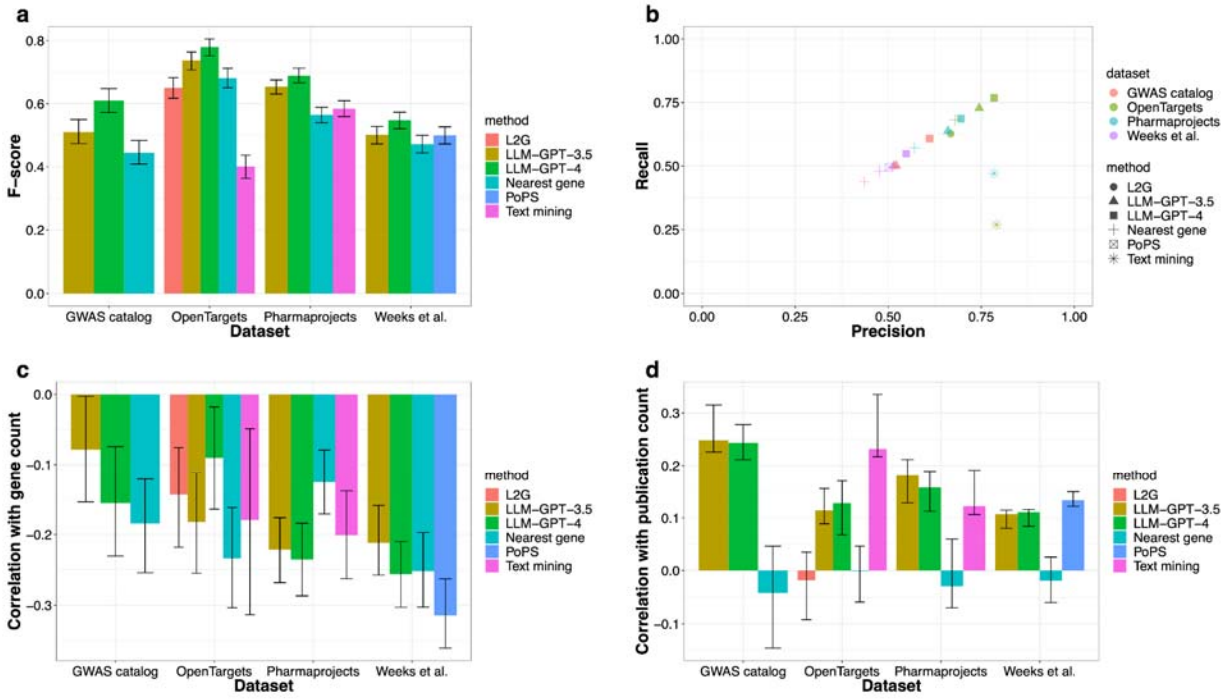

Figure 2: Performance of LLMs and other methods on evaluation datasets, and the factors affecting performance

a. Performance of all methods on evaluation datasets as measured by F-score (a combination of precision and recall)
b. Precision-recall plot showing performance of all methods on each dataset
c. Impact of locus complexity on performance for all methods, measured by correlation of prediction accuracy with number of genes at locus. All methods, including LLMs, perform worse at loci with more candidate genes.
d. Impact of number of publications for the causal gene on performance for all methods, measured by correlation of prediction accuracy with number of publications for causal gene. Most methods show a positive correlation of performance with how well-studied the causal gene is, except the nearest gene and L2G score methods.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

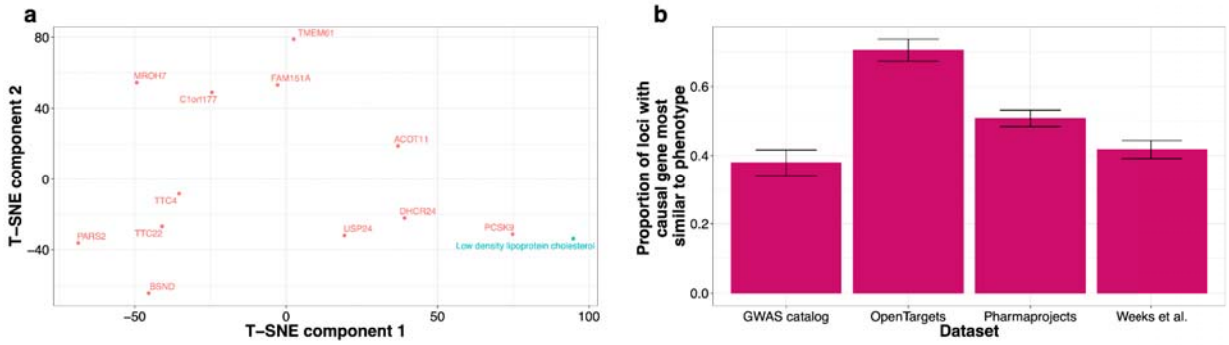

Figure 3: Text embeddings of genes and phenotypes partially explain performance of LLMs

a) T-SNE plot to visualize text embeddings of the genes and phenotype at a locus for LDL cholesterol in the Weeks et al. data. The causal gene *PCSK9* is closest to the LDL cholesterol phenotype in the text embedding space as measured by cosine similarity.
b) Proportion of examples in the evaluation datasets where the causal gene is most similar to the phenotype in the text embedding space as measured by cosine similarity.

Supplementary Tables

1 - Information about datasets
2 - Performance metrics for all methods and datasets
3 - Number of hallucinations per dataset for the LLM approaches
4 - Confidence scores predicted by LLMs and precision estimates for each score value
5 - Precision and recall stratified by phenotype for all methods and datasets
6 - GPT-4 provided reasons for examples about the phenotype 'Total protein' from the Weeks et al. dataset
7 - Comparison of two GPT-4 version in number of hallucinations and performance metrics
8 - P-values from paired Wilcoxon's signed rank test and McNemar test comparing performance of the LLM-based approach to the best non-LLM approach.
9 - Trigrams from the reasons reported by GPT-4 for correct predictions, along with their counts, for the 200 most frequent trigrams
10 - Performance of LLMs on datasets with scrambled phenotypes

Acknowledgements

We would like to thank the employees of 23andMe for making this work possible. We would like to acknowledge Elle Weeks, Hilary Finucane and Jacob Ulirsch for sharing the evaluation dataset from their publication. We would also like to acknowledge David Hinds, Steve Pitts,

medRxiv preprint doi: https://doi.org/10.1101/2024.05.30.24308179; this version posted May 31, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Bertram Koelsch, Michael Holmes, Stella Aslibekyan, Nick Eriksson for comments on the manuscript.

The following members of the 23andMe Research Team contributed to this study:
Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Ninad S. Chaudhary, Zayn Cochinwala, Sayantan Das, Emily DelloRusso, Payam Dibaeinia, Sarah L. Elson, Nicholas Eriksson, Chris Eijsbouts, Teresa Filshtein, Pierre Fontanillas, Davide Foletti, Will Freyman, Zach Fuller, Julie M. Granka, Chris German, Éadaoin Harney, Alejandro Hernandez, Barry Hicks, David A. Hinds, M. Reza Jabalameli, Ethan M. Jewett, Yunxuan Jiang, Sotiris Karagounis, Lucy Kaufmann, Matt Kmiecik, Katelyn Kukar, Alan Kwong, Keng-Han Lin, Yanyu Liang, Bianca A. Llamas, Aly Khan, Steven J. Micheletti, Matthew H. McIntyre, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Jared O'Connell, Steve Pitts, G. David Poznik, Alexandra Reynoso, Shubham Saini, Morgan Schumacher, Leah Selcer, Anjali J. Shastri, Jingchunzi Shi, Suyash Shringarpure, Keaton Stagaman, Teague Sterling, Qiaojuan Jane Su, Joyce Y. Tung, Susana A. Tat, Vinh Tran, Xin Wang, Wei Wang, Catherine H. Weldon, Amy L. Williams, Peter Wilton.

S.S., W.W., S.K., X.W., A.R., A.A., A.K., are employed by and hold stock or stock options in 23andMe, Inc.

Data availability

We will share all processed datasets used in our analysis, as well as the prediction results from all methods on all datasets, intermediate outputs like gene and phenotype embeddings using Zenodo (doi: 10.5281/zenodo.11391053).

All source data were openly available. Download links:
1) OpenTargets - https://github.com/opentargets/genetics-gold-standards/
2) Pharmaprojects - https://github.com/ericminikel/genetic_support
3) Weeks et al. - https://www.finucanelab.org/data
4) GWAS Catalog - https://www.ebi.ac.uk/gwas/docs/file-downloads

Code availability

We will share the scripts we used to query the LLM, as well as the scripts we use to compute our evaluation metrics using Zenodo (doi: 10.5281/zenodo.11391053). All prompts we used are included in the manuscript or supplementary materials.