Title

Evaluation of large language models for discovery of gene set function.

Journal

Nature Methods, 22(1)

Authors

Hu, Mengzhou

Alkhairy, Sahar

Lee, Ingoo

et al.

Publication Date

2025

DOI

10.1038/s41592-024-02525-x

Peer reviewed

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2025 January 13.

Published in final edited form as:

Nat Methods. 2025 January ; 22(1): 82–91. doi:10.1038/s41592-024-02525-x.

Evaluation of large language models for discovery of gene set function

Mengzhou Hu[1,4], **Sahar Alkhairy**[2,4], **Ingoo Lee**[1], **Rudolf T. Pillich**[1], **Dylan Fong**[1], **Kevin Smith**[3], **Robin Bachelder**[1], **Trey Ideker**[1,2,✉], **Dexter Pratt**[1,✉]

[1]Department of Medicine, University of California San Diego, La Jolla, CA, USA.

[2]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA.

[3]Department of Physics, University of California San Diego, La Jolla, CA, USA.

[4]These authors contributed equally: Mengzhou Hu, Sahar Alkhairy.

Abstract

Gene set enrichment is a mainstay of functional genomics, but it relies on gene function databases that are incomplete. Here we evaluate five large language models (LLMs) for their ability to discover the common functions represented by a gene set, supported by molecular rationale and a self-confidence assessment. For curated gene sets from Gene Ontology, GPT-4 suggests functions similar to the curated name in 73% of cases, with higher self-confidence predicting higher similarity. Conversely, random gene sets correctly yield zero confidence in 87% of cases. Other LLMs (GPT-3.5, Gemini Pro, Mixtral Instruct and Llama2 70b) vary in function recovery but are falsely confident for random sets. In gene clusters from omics data, GPT-4 identifies common

✉**Correspondence and requests for materials** should be addressed to Trey Ideker or Dexter Pratt. tideker@health.ucsd.edu; depratt@health.ucsd.edu.

Author contributions

M.H., S.A., T.I. and D.P. designed the study. M.H. and S.A. developed and implemented the automated LLM-based gene set interpretation pipeline, performed the data analysis and organized the GitHub repository. S.A. developed and assessed the semantic similarity calculation. I.L. and M.H. contributed to the development of the citation search and validation pipeline. D.P. contributed to the coding and the evaluation of the analysis. R.T.P. assisted in the study design, prompt engineering and the evaluation of the analysis. M.H., R.T.P., R.B. and D.P. conducted the scientific review of the LLM output. M.H. and D.P. contributed to the user interface design for the GSAI tool. D.F. built the web interface for the GSAI tool, and K.S. set up the server for accessing open-source LLMs. M.H., S.A., T.I. and D.P. wrote the paper with input from all authors. All authors approved the final version of this paper.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Competing interests

T.I. is a cofounder and member of the advisory board and has an equity interest in Data4Cure and Serinus Biosciences. T.I. is a consultant for and has an equity interest in Ideaya Biosciences. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict-of-interest policies. The other authors declare no competing interests.

Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02525-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02525-x.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02525-x.

functions for 45% of cases, fewer than functional enrichment but with higher specificity and gene coverage. Manual review of supporting rationale and citations finds these functions are largely verifiable. These results position LLMs as valuable omics assistants.

A fundamental goal of the omics sciences is to identify the groups of genes responsible for the distinct biological functions of life, health and disease. In this vein, numerous messenger RNA expression experiments over the past several decades have produced sets of genes that are differentially expressed across conditions or that cluster by common expression patterns. Similarly, proteomics experiments produce clusters of proteins that are coabundant, comodified or physically interacting, and gene knockout screens produce lists of genes required for fitness or a particular response. In all of these cases, the basic premise is that the identified genes work coherently toward the same biological process or function.

The usual approach to interpret the genes identified in omics experiments is through functional enrichment analysis[1-9]. This method seeks to identify similarities between a cluster of omics genes and those from a large predefined collection of gene sets organized by shared functions or pathway categories[10-15]. This predefined collection can come from literature-curated gene function databases, such as Gene Ontology (GO)[16,17], Kyoto Encyclopedia of Genes and Genomes (KEGG)[18-20] or Reactome[11,21,22]. Alternatively, one can perform enrichment analysis against databases of genes annotated from previous independent experiments, such as genes previously linked to the same disease in the Genome-Wide Association Studies Catalog[23], genes linked to the same mouse knockout phenotypes in the Mouse Genome Database[24,25], genes regulated by a common transcription factor[26,27] or genes that serve as canonical biomarkers for a given cell type[28-30].

Paradoxically, an omics gene cluster that is highly similar to gene sets in a reference database may be of lesser interest, since the cluster and its function have already been well characterized. Of greater interest are clusters of genes that have not been previously implicated, because it is precisely in these cases that new biological insights emerge. These less-studied cases may either show no statistically significant enrichment in the reference database, or they may return enrichments that are significant in terms of *P* value but not substantial in terms of gene set overlap. Here, an immediate next step is to explore the biological literature, as well as complementary datasets, to learn as much as possible about the genes in question. The goal is to mine knowledge pertinent to each gene and then use this knowledge to synthesize mechanistic hypotheses for a function that might be held in common by all or many genes in the set. This protracted process of discerning relevant findings from data and literature, then reasoning on this information to synthesize functional hypotheses, has not yet been widely automated but is one of the central tasks performed by a genome scientist.

The advent of generative artificial intelligence (AI) models and, specifically, large language models (LLMs) is highly relevant to these tasks. At its core, generative AI is an approach to machine learning by which a model is trained to recognize underlying patterns in data in a manner that allows it to generate new results with properties similar to the training data. The underlying technology behind LLMs is the transformer architecture[31-33], which

uses a self-attention mechanism to understand context and handle long-range dependencies in text, delivering notable advancements in tasks such as text translation, summarization and generation. Recent AI research has produced a flurry of general-purpose LLMs, such as Generative Pre-trained Transformer 4 (GPT-4)[34] by OpenAI, Llama2[35] by Meta, Mixtral[36] by MistralAI, and Gemini[37] by Google, which incorporate information from an enormous corpus of sources, including the biomedical literature. Based on these developments, LLMs present major opportunities to assist in the interpretation of gene sets derived from omics experiments[38].

Here, we evaluate the degree to which LLMs provide insightful functional analyses of gene sets based on their embedded biological knowledge and text-generation capabilities. First, we develop a gene set analysis pipeline based on queries to a panel of current LLMs. We then test the ability of each LLM to propose succinct names describing the functions of gene sets of interest, as well as to support this choice by referenced text and an overall assessment of confidence. Finally, we discuss our findings and their implications for the general use of LLMs in functional genomics.

Results

Development of an LLM functional genomics pipeline

We designed a pipeline in which an LLM is instructed to analyze a gene set and then generate a short biologically descriptive name, a supporting analysis essay and a score reflecting the LLM's 'confidence' in these results (Fig. 1a and Methods). A separate LLM instruction was used to validate statements made in the analysis essay with pertinent literature citations (Extended Data Fig. 1 and Methods). The instruction to an LLM is called a 'prompt' and can include data and examples to guide the response. Best practices for formulating this prompt are the subject of ongoing experimentation[39-42]; here, our prompt was engineered to capture desired properties of the results to be generated, including guiding phrases such as "After completing your analysis, propose a short descriptive name for the most prominent biological process performed by the system". The engineered prompt also included a single (one-shot) example to help the LLM imitate the desired format and thought process (Fig. 1a, Extended Data Fig. 1b and Extended Data Table 1). This LLM functional genomics pipeline is available for general use via the Gene Set AI (GSAI) web portal (https://idekerlab.ucsd.edu/gsai/).

We sought to evaluate this LLM pipeline using reference gene sets derived from two primary sources. The first source was literature curation, for which we evaluated sets of genes drawn from GO terms[16,17] (Fig. 1b, evaluation task 1). The second data source was 'omics analysis, for which we evaluated clusters of genes identified by various 'omics platforms, including transcriptomics and proteomics (Fig. 1c, evaluation task 2). The goal of the first task was to benchmark how well LLMs recover gene set functions previously documented by a human-curated reference database, while the goal of the second task was to explore the extent to which LLMs provide complementary insights beyond what is obtained from such databases.

Evaluation task 1

Recovery of literature-curated functions.—For the first task, we randomly sampled a representative corpus of terms from the GO biological process branch (GO-BP 2023-11-15 release; Extended Data Fig. 2 and Methods). The gene set annotated to each term was used to prompt five different LLMs (GPT-4, Gemini Pro, GPT-3.5, Mixtral Instruct and Llama2 70b; Fig. 1b), after which the names suggested by the LLMs were compared with the term names assigned by the GO curators. In each case, performance was measured by the semantic similarity of the LLM name to the GO name. Semantic similarity[43] is a quantitative score (range 0–1) that measures the closeness in meaning of two words or phrases, regardless of whether those phrases involve different words or expressions (Methods). For example, the word ‘socks’ is semantically closer to the word ‘shoes’ than it is to ‘airplane’.

The five LLMs required 7.9 s (Gemini Pro) to 61.8 s (Llama2 70b) to process a gene set and return a proposed concise name, a confidence score and supporting analysis text (Extended Data Table 2). Semantic similarity scores ranged from values as high as 1.0, in cases where the LLM name exactly matched the GO name (for example, Gemini Pro: ‘synaptic vesicle exocytosis’, GO:0016079), to values below 0.1, in cases where the names were not intuitively similar (for example, GPT-3.5: ‘regulation of ion transport and cellular homeostasis’ versus GO: ‘negative regulation of CD8-positive, alpha–beta T cell differentiation’, GO:0043377) (Table 1 and Supplementary Table 1). We found that GPT-4, Gemini Pro, GPT-3.5 and Mixtral Instruct showed roughly equivalent performance in proposing a name that was similar to the GO name (median similarities in the range 0.45–0.50), whereas the performance of Llama2 70b was significantly worse (median similarity 0.40; Fig. 2a).

To interpret these similarity scores, we calibrated them against background semantic similarity distributions, defined by comparing each LLM-proposed name against the entire set of 11,943 term names documented in GO-BP (Methods). For example, the GPT-4 name (‘DNA damage response and repair’) had semantic similarity of 0.54 to the GO name (‘response to X-ray’), a score that was higher than 99% of semantic similarities between the GPT-4 name and every other term name in GO-BP (Fig. 2b and Supplementary Table 2). Using this scoring approach, we found that 60% of gene set names proposed by GPT-4 were close matches to the corresponding GO term names, with semantic similarities ranking above the 95th percentile (Fig. 2c,d). In approximately one-third of remaining cases, the LLM proposed a name matching a broader concept (Fig. 2d and Methods). For example, the gene set corresponding to the GO term ‘negative regulation of triglyceride catabolic process’ resulted in the GPT-4 name ‘lipid metabolism and trafficking’ with a semantic similarity of 0.41 ranking in the 89th percentile. The GPT-4 name matched most closely to the GO term ‘lipid metabolic process’, a less specific category higher in the ontology and annotated by a larger set of genes (Fig. 2e). Qualitatively similar results were observed when analyzing gene sets from the cellular component and molecular function branches (Extended Data Fig. 3 and Supplementary Table 2).

Assessment of LLM confidence.—We next focused on the self-confidence reported by each LLM. As noted above (Fig. 1a), we asked each LLM to provide a continuous confidence score[44,45] for each gene set analysis, in the range 0–1. For gene sets for which the LLM assigned a confidence of '0', we requested the LLM to return 'system of unrelated proteins' rather than a proposed name, since it could not confidently determine a collective functional description. We observed that these quantitative confidence scores were not uniformly distributed but clustered around distinct modes; hence, we further thresholded scores to high/medium/low confidence outcomes based on this distribution (Extended Data Fig. 4a).

To gain insight into whether the LLM self-confidence assessments were informative and useful, we introduced in our evaluation the concept of contaminated gene sets. Specifically, each of the GO terms used previously ('real GO term'; Fig. 3a) was substituted with a synthetic gene set containing 50% of genes randomly selected from that GO term and 50% of genes randomly selected from the background pool of all genes with GO annotations ('50/50 mix'). We also examined a fully random variant whereby 100% of genes were randomly selected from background ('random').

We observed that all LLM models, except for Llama2, showed a significant reduction in self-confidence when asked to generate names for the 50/50 mix and random gene sets (Fig. 3b). GPT-4 was the most likely of the five LLMs to correctly associate lower confidence with contaminated gene sets, and it gave zero confidence for (refusing to name) most of the gene sets that were fully random (87%). In contrast, GPT-4 rated nearly all analyses involving real gene sets as medium confidence or above (96%; Fig. 3b), with quantitative confidence scores that were predictive of the accuracy of name recovery (Extended Data Fig. 4b). These GPT-4 confidence assessments approximately agreed with those from manual independent review, in which a human reviewer rated 25 real gene sets for the degree to which the GPT-4 analysis essay supported its proposed gene set name (Extended Data Table 3).

Lastly, we compared these results with those of classic functional enrichment analysis run on the same real, contaminated and random gene sets (g:Profiler[46], Benjamini–Hochberg (BH)-adjusted $P \leq 0.05$; Methods). As expected, enrichment analysis always returned the correct GO term for the real gene set, while for most random gene sets, it failed to meet the significance cutoff (73%; Fig. 3b). In contrast, enrichment analysis nearly always returned significant GO terms for 50/50 mix contaminated gene sets, indicating that in this respect it was less conservative than a GPT-4 confidence assessment.

Evaluation task 2

Exploration of omics gene clusters.—The second major task we evaluated was in naming gene sets that had been identified experimentally, via clustering of omics data. Here, we focused on GPT-4, given its good performance in task 1, especially its superiority in refusing to name noncoherent gene sets. The omics gene clusters included the following: (1) genes differentially expressed in transcriptomic profiles collected in response to a panel of drug treatments ($n = 126$ gene clusters, Integrated Network-Based Cellular Signatures (LINCS) L1000 Connectivity Map Signatures)[47,48], (2) genes differentially expressed upon

infection by a panel of viruses (n = 48 clusters, Gene Expression Omnibus (GEO) Signatures of Differentially Expressed Genes for Viral Infections)[49] and (3) genes encoding complexes of interacting proteins in cancer proteomics data (n = 126 clusters, Nested Systems in Tumors (NeST))[50] (Methods). Together, these sources comprised 300 gene clusters of sizes ranging from 3 to 100 genes (Extended Data Fig. 5).

When prompted with each of these omics clusters, we found that GPT-4 proposed a name in 135 cases (45%) and otherwise deferred with zero confidence. As a comparative benchmark, we also subjected each cluster to functional enrichment analysis against the GO biological process database, yielding significant GO term names for 229 clusters (g:Profiler[46], BH-adjusted $P \leq 0.05$; Methods). Preliminary inspection of these naming results suggested that both GPT-4 and GO enrichment could produce names with low specificity, that is, that apply to only a small fraction of genes in the cluster or, conversely, apply broadly to numerous genes outside. To quantify this specificity, for each cluster we characterized the degree of overlap between the set of genes comprising the cluster and the set of all human genes associated with the proposed name (Jaccard index; Methods). Indeed, even a modest specificity requirement eliminated the majority of proposed cluster names. For example, requiring a minimum specificity of 10% left 42 clusters named by GPT-4 and 33 named by functional enrichment; increasing the minimum to 20% left 21 clusters named by GPT-4 and 4 named by functional enrichment (Fig. 4a and Extended Data Table 4). In general, GO enrichment was more likely to name a cluster, whereas GPT-4 tended to yield names supported by a greater number of cluster genes (Fig. 4b).

Direct comparison of the GPT-4 versus GO names across clusters revealed that the GPT-4 name was often semantically similar to one of the functionally enriched terms but with additional implicated genes (65% of clusters; Fig. 4b). An example was protein interaction cluster NeST:2–105, which yielded the GPT-4 name ‘regulation of cullin–RING ubiquitin ligase (CRL) complexes’ (Figs. 4c and 5). GPT-4 analysis text and citations connected ubiquitin ligase complexes to all 16 proteins in the cluster, whereas the most relevant GO term match, ‘protein ubiquitination’, covered 8 of the 16 proteins. Both analyses associated ubiquitination with members of the potassium channel tetramerization domain (KCTD) and Kelch-like (KLHL) gene families, which have been implicated as substrate adaptors for E3 ubiquitin ligases[51,52], and they also both implicated WNK1[53]. Among the additional cluster members covered by the GPT-4 analysis were RHOBTB proteins, which have also been studied as E3 adaptors[54], the additional KCTD member SHKBP1[51], additional WNK family members[55,56] and the understudied protein ANKRD39 based on its predicted ubiquitin transferase activity[57]. Notably, the term ‘protein ubiquitination’ was neither the most significantly enriched nor the highest overlap, since it broadly covers many genes; rather, the top match was the unrelated concept ‘negative regulation of pancreatic juice secretion’ based on inclusion in the cluster of three of five genes annotated to this term. Furthermore, the association of WNK1 with protein ubiquitination (by both methods, GPT-4 and functional enrichment) is speculative and needs further study to determine whether WNK proteins are merely a target of ubiquitin ligation or integral to the mechanism.

Assessment and validation of supporting analysis text.—An important concern with LLM outputs is the potential to ‘hallucinate’, that is, to generate plausible but

unverifiable or nonfactual statements[34]. We therefore evaluated the analysis essays generated by GPT-4 in support of its proposed gene cluster names to determine the degree to which hallucination might influence its analyses. For this purpose, four human scientists participated in a structured review process for 403 sentences generated in the analysis of 20 omics gene sets (Methods). As a conservative criterion, we considered a sentence 'verified' only if the reviewer found evidence in the literature for every stated fact. Of the 403 sentences evaluated, we found 354 to be fully verifiable (88%; Supplementary Table 4). Examination of the 49 remaining sentences revealed two major types of unverified facts: (1) miscategorization of gene functions (n = 15, 4%) and (2) speculation of gene functions (n = 34, 8%). In one case relevant to type 1, GPT-4 stated that WDTC1 "is involved in the regulation of the cell cycle and apoptosis…" when in fact, it is an E3 ubiquitin ligase and is involved in adipogenesis and obesity[58] (Supplementary Table 4). Relevant to type 2, GPT-4's speculation that REN "may be affected by vesicular trafficking processes" could not be verified (Supplementary Table 4).

To facilitate statement verification, we developed a separate GPT-4-based system to add citations to the analysis essay in support of key statements made (Extended Data Fig. 1, Supplementary Table 4 and Methods). In formulating the engineered prompt for this task, we did not stipulate that the title or abstract of a publication must be primarily about the statement; it was sufficient that a supporting fact was present. The 403 previously reviewed sentences returned 489 citations through this automated system. In 383/489 cases, the paper title or abstract provided clear evidence for the cited statement. For example, the statement that RHOBTB2 and RHOBTB3 "have been implicated in the regulation of CRL3 complexes" was supported in the title of Berthold et al. (2008)[54] and the abstract of Ji and Rivero (2016)[59] (see analysis paragraph 4 and its citations in Fig. 5). The remaining 106 citations (22%) did not verifiably support their corresponding LLM statements, although we reviewed titles and abstracts only without a systematic review of the main manuscript text. These results suggest that most but not all citations found by this procedure are reliable, such that they may be viewed as useful guidance for further study but not unquestioned facts.

Discussion

The evaluations performed here suggest that LLMs have notable potential as automated assistants for understanding the collective functions of gene sets. In the analysis of gene sets from GO, four out of five LLMs performed comparably in proposing names similar to the names assigned by the GO curators, producing highly similar names for most gene sets. The accompanying analysis text was found to be largely factual, although GPT-4's occasional generation of unverifiable statements shows that even current state-of-the-art LLMs should be coupled to fact-checking and/or reference validation, whether automated or manual.

It is somewhat unexpected that GPT-3.5 performs equally well to GPT-4 in the recovery of GO gene set names (Fig. 2a). In other applications, GPT-3.5 typically shows a 10–30% performance decrease relative to GPT-4[34,60,61]. This comparable performance is important because GPT-3.5 is faster and less costly to execute than GPT-4 (Extended Data Table 2). However, while GPT-3.5 performed well in gene set naming, it struggled to assess the confidence of its answers (Fig. 3b). Here, GPT-4 demonstrated a clear ability to assess

confidence, particularly in refusing to name incoherent gene sets. As LLMS continue to evolve, advances in speed, cost and output quality will probably impact the preferred model for gene set analysis.

When the GPT-4 name for a GO gene set was not similar to the curated name, in roughly a third of those cases, it was conceptually broader (Fig. 2d). For the remaining gene sets with discrepant naming, the mismatch could reflect a failure of GPT-4 to recover a well-documented function or an indication the GO term no longer reflects the up-to-date literature. Alternatively, it is possible that both GPT-4 and GO offer valid, but alternate, interpretations. We indeed find evidence for this last possibility: for example, dendritic cell dendrite assembly (GO:0097026) is annotated with two chemokines (CCL19 and CCL21) and their receptor (CCR7), but these proteins are also critical to the related process of lymphocyte homing, consistent with the GPT-4-proposed name ‘lymphocyte homing and immune response regulation’ (Supplementary Table 2).

In the analysis of gene clusters derived from omics studies, GPT-4 proposed gene set functions in 135 out of 300 cases. As such clusters reflect patterns in molecular data that may be noisy or include less-studied genes, it is perhaps not surprising that not all clusters are assigned confident names. When there is no predominant theme, the LLM’s text-based analysis will, nevertheless, discuss the range of biological processes characterizing the cluster. Functional enrichment analysis named more clusters (229 out of 300 cases) but typically with low specificity or coverage (Fig. 4b), and it was also more likely to name random gene sets (Fig. 3b). That said, functional enrichment could be given access to a broader range of candidate names when using databases such as Reactome, KEGG or the Phenotype Ontology; here, we chose to focus on the GO-BP branch as a universally accepted, comprehensive collection. An exciting possibility would be to integrate the best of both worlds, combining the statistical transparency of enrichment analysis with the up-to-date literature knowledge and reasoning of LLMs.

This work has some relation to a recent preprint[38] that used GPT to extract terms from the GO database that best describe a gene set. Here, we have provided a broad selection of LLMs with the open-ended task of describing gene set functions without explicit reference to predefined databases. We also introduced a new metric, the LLM self-confidence score, to rate the functional coherence of a gene set and the quality of its functional summary. Via its self-confidence assessment, an LLM can potentially alert biologists to cases in which they should be skeptical of a simple ‘best match’ function proposal.

It is important to stress that the goal of this study was to assess the baseline capability of LLMs in functional genomics, using single queries and prompts developed by informal experimentation. Given this baseline, future studies might seek to build capability in several ways. A first major direction would be to further boost LLM accuracy and interpretability, for which recent techniques such as fine-tuning[31] and retrieval-augmented generation[62] are showing considerable promise. A second would be to systematically investigate LLM prompting strategies, including prompts that directly integrate LLMs with complementary tools[63-68] such as gene set enrichment and literature searches. Future prompting strategies might also evaluate and include descriptions of the biological and experimental context in

which a gene set was discovered, information that seems likely to improve the specificity, depth and quality of the analysis. Such prior context has been difficult to capture using gene set functional enrichment tools, since their preexisting mapping of gene sets to functional terms is static and does not attempt to encode the practically infinite space of biological conditions.

Methods

LLM installation

Five LLMs were selected for the evaluation, including GPT-3.5 and GPT-4 from OpenAI, Gemini Pro from Google, Mixtral Instruct from MistralAI, and Llama2 70B from Meta. We used the 'gpt-4-1106-preview' and 'gpt-3.5-turbo-1106' versions of the OpenAI GPT-4 and GPT-3.5 LLMs and the 'Gemini Pro' version of the Google Gemini model using their well-defined Application Programming Interfaces (APIs). Mixtral Instruct and Llama2 were downloaded from Ollama (https://ollama.com/) and queried through the API endpoint of Ollama.

Controlling the variability of LLM responses

Each LLM enables queries to set a 'temperature' parameter that controls the variability of the generated response, with lower temperatures producing more reproducible and reliable responses[69,70]. Exploring the effect of temperature on LLM analyses is outside the scope of this study, and therefore our queries used the lowest, most conservative/reproducible temperature value (0.0). In a manual inspection of repeated queries at temperature 0.0, we found that LLM names and analyses were conceptually equivalent but that the specific text could vary, from near identity to considerable differences in phrasing. The 'seed' parameter was set to 42 for all models and all runs. In addition, we made our manual review process manageable by forcing the responses to be concise. For this purpose, we set the maximum number of tokens (roughly corresponding to words) in each response to be 1,000.

Prompt engineering

The LLM prompt was organized into seven sections (Fig. 1a; see full prompt in Extended Data Table 1). (1) System content section: System content tells the role of the LLM when to process the prompt. Here, our analysis was associated with molecular biology; thus, we set the role to be 'assistant of a molecular biologist'. (2) Task instruction section: The instructions were engineered to meet multiple criteria. Notably, the LLMs were guided to first perform the analysis before proposing a process name, encouraging a structured 'chain of thought'. (3) Confidence score assignment section: This prompt section instructed the LLM to generate a 'confidence score' expressing its confidence in its choice of name, taking into account the fraction of genes that participate in the corresponding biological process(es). The coherence score was specified to be between 0.00 and 1.00. The prompt was also engineered to handle situations where the genes in a set are not sufficiently related to warrant a name. In particular, the prompt instructed the LLM to output a zero confidence score and the name 'system of unrelated proteins' in these cases. (4) Format instruction section: We asked the LLM to place the name as a title in the final analysis for easy

extraction. (5) Analytical approach section: The instructions in this section guided the LLM to be succinct, factual and focused on finding commonalities and relationships. (6) One-shot example section: This section contained an example of a gene set and the corresponding name, confidence score and analysis text. This format follows the 'in-context learning' approach, in which examples provide a template to help the LLM generate outputs consistent with the desired behavior and format. After substantial manual testing, we determined that the quality of the output was no different when using one example versus several examples; thus, we chose to use a 'one-shot' single example strategy, minimizing both prompt size and associated costs. (7) User input of genes/proteins section: The last section is the user's input of the gene or protein list.

Download and parsing of GO

GO (2023-11-15 version) was obtained from the geneontology.org website in the Open Biomedical Ontologies[71,72] format. The ontology file was subsequently divided into its three constituent branches: biological processes (BP), cellular component (CC) and molecular function (MF). The gene set corresponding to each GO term was determined by aggregating the genes with which it was directly annotated with those of all its ontological descendants. We randomly drew 1,000 human gene sets from terms in each branch (sampling terms from 3 to 100 genes) for evaluation task 1. We found that 1,000 gene sets were sufficient to achieve statistical significance via a representative distribution (Extended Data Fig. 2). We limited the size to 100 gene sets for the comparison of confidence scores between five LLMs owing to the cost of LLM queries as well as the required computing time (Extended Data Table 2; relevant to Figs. 2 and 3, Extended Data Figs. 3 and 4 and evaluation task 1).

Calculation of semantic similarity

Semantic similarity between names was determined using the SapBERT model[73] from huggingface (cambridgeltl/SapBERT-from-PubMedBERT-fulltext) via the transformers package[74] (version 4.29.2). SapBERT produces embeddings of each name and then computes the cosine similarity between the embeddings, yielding a similarity score ranging from 0 (no similarity) to 1 (identical). SapBERT is a domain-specific language representation model pretrained on large-scale biomedical data, including Unified Medical Language System, a massive collection of biomedical ontologies with 4M+ concepts. Since models like Bidirectional Encoder Representations from Transformers (BERT)[31] are trained on vast amounts of textual data, they can learn general patterns and relationships and capture context by considering surrounding words, providing a measure of similarity based on semantics rather than lexical matching. Although both SapBERT and GPT-4 are LLMs, they are separate models with different purposes, model architecture, training objectives and data. SapBERT therefore provides an independent evaluation of similarity.

Calibrating the similarity between GPT-4 names and GO names

To evaluate the performance of the GPT-4 model in recapitulating GO names, we computed the semantic similarity between the GPT-4 name and the assigned name of the GO term query, using SapBERT as described above. We then performed this semantic similarity calculation for the same GPT-4 name against every other GO term name in the biological

process branch (GO-BP), yielding a background distribution of semantic similarity scores for each GO term query. The actual and background similarities were then concatenated into a single list, sorted in descending order (largest to smallest), and the rank of the actual similarity was recorded and expressed as a percentile. This percentile score is thus the percentage of GO-BP term names that are less similar to the GPT-4 name than to the assigned name of the GO term query.

Definition of 'broader concepts'

A proposed name is said to capture a 'broader concept' than that represented by a query gene set, as follows:

Q: target gene set for analysis;

$N_i(X)$: name of gene set X proposed by method $i = \{\mathsf{LLM}, \mathrm{GO}\}$; $P_{i=\mathsf{LLM}}$: complete gene set annotated to GO term that is closest to the name proposed by the LLM, that is, maximizing $\mathrm{sim}(N_{\mathrm{GO}}(Q), N_{\mathsf{LLM}}(Q))$;

$P_{i=\mathrm{GO}}$: complete gene set annotated to name proposed by GO term enrichment.

The proposed name $N_i(Q)$ expresses a 'broader concept' if $|P_i| > |Q|$ and $|P_i \cap Q| \geq 0.5|Q|$, that is, P_i is larger than Q and contains at least half of it. We selected 0.5 as the threshold on the grounds that concepts (P_i) that apply to a majority of the genes in Q can reasonably be considered as related.

Omics data processing

NeST data are raw files from a previous study of cancer protein clusters[50] obtained through personal communication with M. R. Kelly. The L1000 data and viral infection data were downloaded from the Harmonizome platform[75] (https://maayanlab.cloud/Harmonizome/; LINCS L1000 CMAP Signatures of Differentially Expressed Genes for Small Molecules and GEO Signatures of Differentially Expressed Genes for Viral Infections). For each omics source, we selected gene sets with a size between 3 and 100 genes. Furthermore, in the L1000 dataset, we selected the context with the greatest number of observations (cell line 'MCF7', duration 6.0 h, dosage 10.0 μm). For the viral disease perturbations dataset, we used a z-score cutoff of 2.

Gene set enrichment analysis

We used the g:Profiler[46] API service to perform gene set enrichment analysis for both task 1 and task 2. We used an adjusted P value ≤ 0.05 calculated by Benjamini–Hochberg false discovery rate to determine significantly enriched GO-BP terms. In task 2, we updated the omics gene set gene symbols, downloaded on 27 June 2024 from the HUGO Gene Nomenclature Committee database[76] (https://www.genenames.org/). We then computed the Jaccard index for enriched GO terms as $\frac{|P_{i=\mathrm{GO}} \cap Q|}{|P_{i=\mathrm{GO}} \cup Q|}$ (related to Fig. 4a and Extended Data Table 4).

Evaluation of specificity of naming for omics gene sets

For a given LLM-proposed name and analysis essay, we prompted GPT-4 to analyze the essay to identify all genes mentioned in support of the name and return them as a list G_{LLM}. This prompt (Extended Data Table 5) included the instruction to only consider definite assertions about the gene rather than conjectures. We found via manual inspection of approximately 20 essays that GPT-4 was able to reliably perform this task. We computed the semantic similarities between the LLM-proposed name and all GO-BP term names (2023-11-15 version) to extract the gene set from the closest GO-BP term name ($P_{i=\text{LLM}}$). The specificity was computed on the basis of the Jaccard index:

$$\frac{|\ G_{\text{LLM}}\ |}{|\ P_{i=\text{LLM}}\ \cup\ Q\ |},$$

where Q is the omics gene set used as the query.

Identification and validation of relevant references (citation module)

We followed a five-step process to identify and evaluate references for statements made in the LLM-generated analysis text. For each paragraph in the analysis text, we performed the following (Extended Data Fig. 1):

1. Prompt the LLM to extract two types of keyword from the analysis paragraph: (1) gene symbols explicitly mentioned in the paragraph and (2) up to three keywords associated with gene functions or biological processes, ordered by their importance. Paragraphs that do not yield at least one gene symbol and one functional keyword are skipped, returning ‘unknown’. The prompt incorporates a one-shot example of a paragraph and corresponding keywords.
2. Assemble a PubMed query expression to find scientific publications in which either the title or abstract contains one or more of the gene symbols and one or more of the function keywords.
3. Query PubMed via its web API, sorting the returned publication list by relevance.
4. Further prioritize the publications on the basis of the number of matching genes in the abstract. We prefer publications that provide information on the most genes.
5. For each of the top three publications, prompt the LLM to assess whether the title and abstract provide evidence for one or more statements of fact in the analysis paragraph. Return the publication as a reference if the LLM considers that it satisfies that requirement.

Reviewer fact-checking of GPT-4 analysis text

We performed a structured review of 403 sentences from the analysis text generated by GPT-4 based on 20 selected omics gene sets (Supplementary Table 3). In this review, each of the four reviewers recorded the number of unverified statements of fact for each analysis

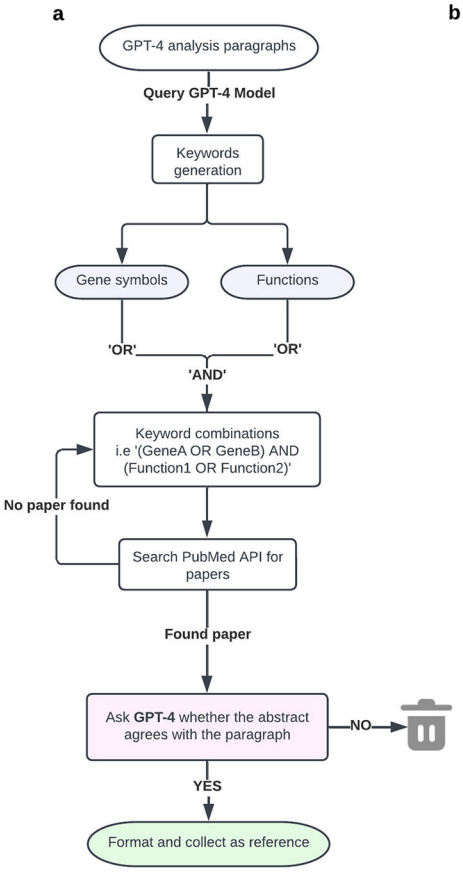
in the corresponding column. A statement was considered ‘unverified’ if no supporting evidence was found within roughly 10 min, using the following method:

- Check simple per-gene statements against information from National Center for Biotechnology Information (NCBI) gene content maintained by the National Library of Medicine (http://www.ncbi.nlm.nih.gov).
    - **a.** For example, ‘Oxytocin (OXT) is a neuropeptide hormone that binds to its receptor, oxytocin receptor (OXTR).’ can be quickly verified by the NCBI Gene entries for the two genes.
    - **b.** If the NCBI entry verifies one or more statements, add the uniform resource locator for the entry to the evidence column, for example, ‘NLM: OXT http://www.ncbi.nlm.nih.gov/gene/5020’.
- For statements not verified by NCBI Gene, search PubMed for publications to provide evidence for the statement. Search strategies include:
    - **a.** Search using gene–keyword pairs, such as ‘TP53 cell cycle’.
    - **b.** For paragraphs that discuss multiple genes, search for review articles with phrases such as ’acute phase response proteins’.
    - **c.** Search for family member proteins together, such as ‘TAS2Rs bitter taste’.

Reviewer evaluation of references

The reviewers evaluated references on the basis of the same criteria with which the LLM was prompted in step 5 of the reference-finding process (above). Reviewers separately recorded whether the title or the abstract successfully provided evidence for a statement of fact, along with the number of irrelevant references for a paragraph.

Extended Data



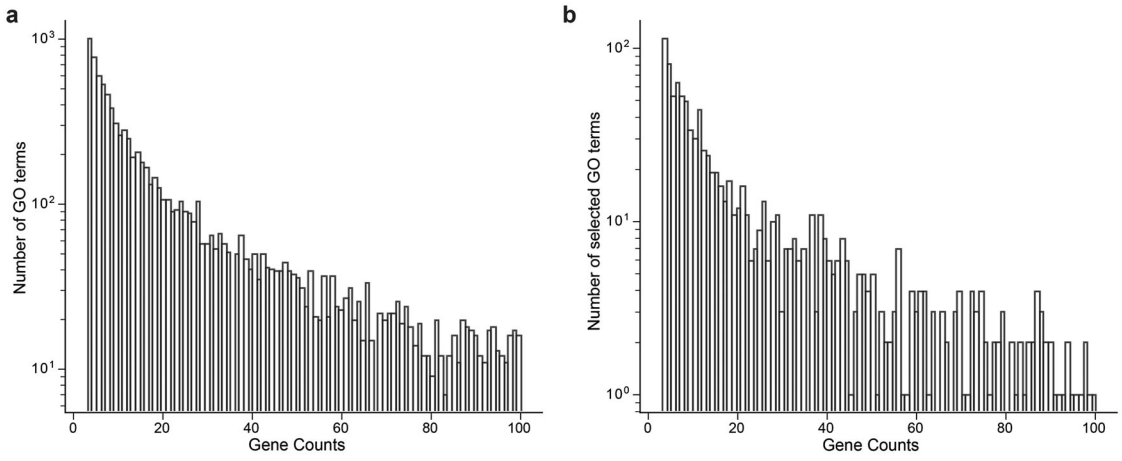
b

System
I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown".

Prompts for keywords extraction	
Gene Symbols	Functions
I have a paragraph *Paragraph:* *{paragraph}* *I would like to search PubMed to find supporting evidence for the statements in this paragraph. Give me a list of gene symbols from the paragraph. Please only include genes. Return the genes as a comma separated list without spacing, if there are no genes in the statements, please return "Unknown"*	*I would like to search PubMed to find supporting evidence for the statements in a paragraph. Give me a maximum of 3 keywords related to the protein functions or biological processes in the statements.* *Example paragraph: Involvement of pattern recognition receptors: TLR1, TLR2, and TLR3 are part of the Toll-like receptor family, which recognize pathogen-associated molecular patterns and initiate innate immune responses. NOD2 and NLRP3 are intracellular sensors that also contribute to immune activation.* *Example response: immune response,receptors,pathogen* *Please don't include gene symbols. Please order keywords by their importance in the paragraph, from high importance to low importance. Return the keywords as a comma separated list without spaces. If there are no keywords matching the criteria, return "Unknown"* *Please find keywords for this paragraph:* *{paragraph}*

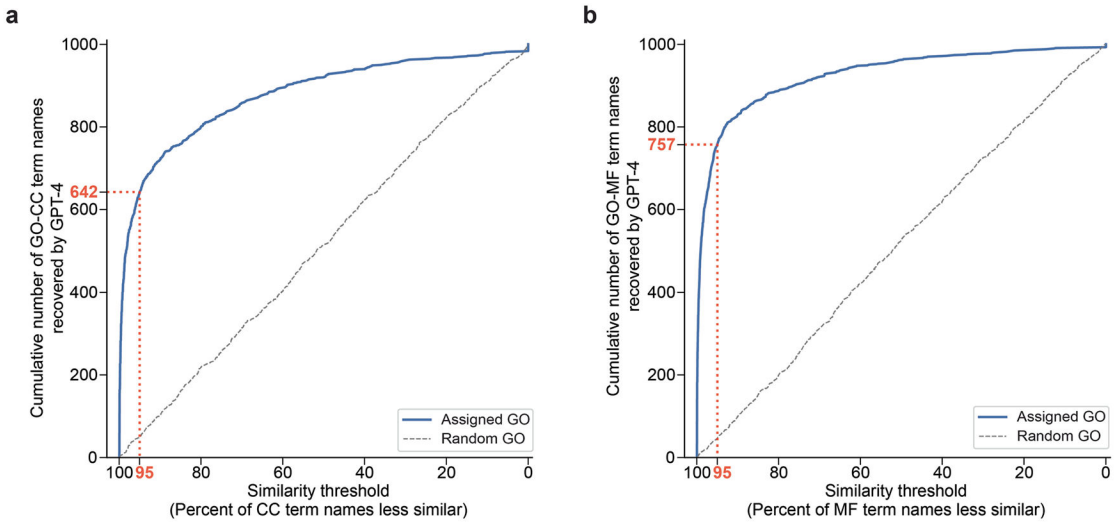
Prompt for abstract support
I have a paragraph *Paragraph:* *{paragraph}* *and an abstract.* *Abstract:* *{abstract}* *Does this abstract support one or more statements in this paragraph? Please tell me yes or no.*

Extended Data Fig. 1 |. Schematic of the citation module.
a, GPT-4 is asked to provide gene symbol keywords and functional keywords separately. Multiple gene keywords and functions are combined and used to search PubMed for relevant paper titles and abstracts in the scientific literature. GPT-4 is queried to evaluate each abstract, saving supporting references. **b**, Prompts used to query the GPT-4 model.



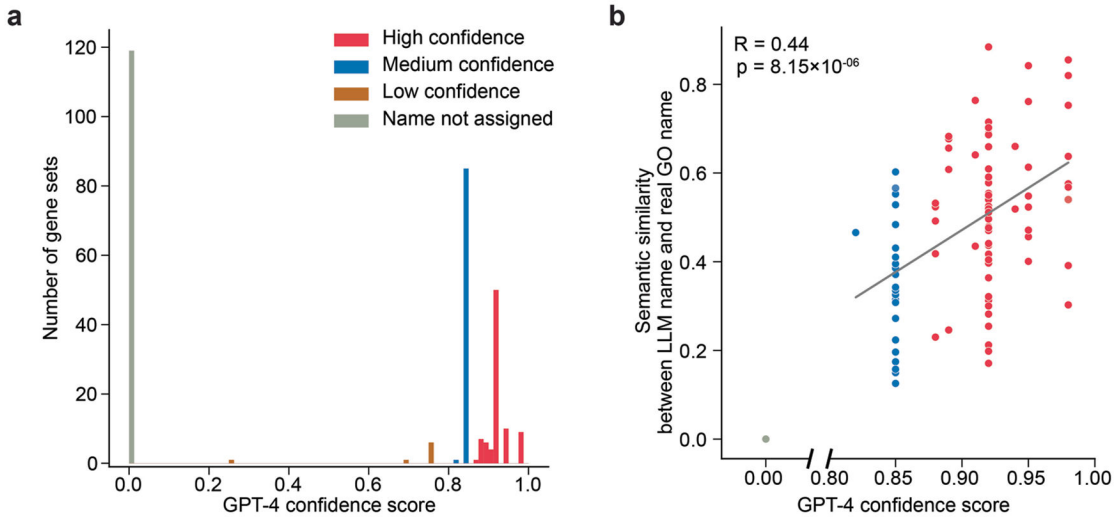
Extended Data Fig. 2 |. Distribution of GO term gene sizes.

a, Distribution of term size (number of genes) for terms in the Biological Process branch (GO-BP). Terms with 3-100 genes shown (n = 8,910). **b**, Distribution of term size for the 1000 GO terms used in Task 1.



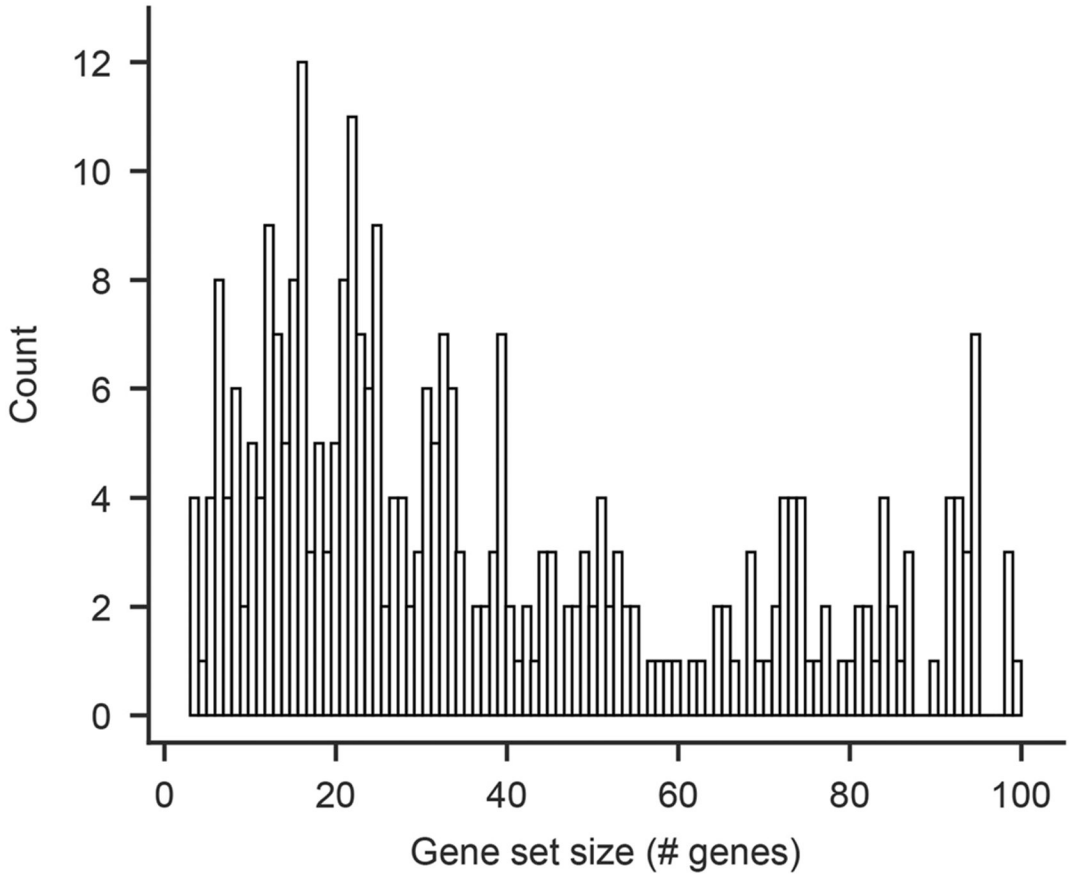
Extended Data Fig. 3 |. Evaluation of GPT-4 in recovery of GO-CC and GO-MF names.

a, Cumulative number of GO-CC term names recovered by GPT-4 (y-axis) at a given similarity percentile (x-axis). 0 = least similar, 100 = most similar. Blue curve: semantic similarities between GPT-4 names and assigned GO-CC term names. Grey dashed curve: semantic similarities between GPT-4 names and random GO-CC term names. The red dotted line marks that 642 of the 1000 sampled GO-CC names are recovered by GPT-4 at a similarity percentile of 95%. **b**, As for panel a, but for GO-MF terms rather than GO-CC. The red dotted line marks that 757 of the 1000 sampled GO-MF names are recovered by GPT-4 at a similarity percentile of 95%.



Extended Data Fig. 4 |. Supplemental analysis of the confidence score.

a, Distribution of confidence scores (n = 300) assigned by GPT-4 with confidence level threshold set based on the distribution pattern. “High confidence” (red): 0.87–1.00; “Medium confidence” (blue): 0.82–0.86; “Low confidence” (dark orange): 0.01–0.81; “Name not assigned” (gray): 0. **b**, Scatter plot of naming accuracy versus GPT-4 self-assessed confidence score for real gene sets drawn from GO (points, n = 100). Accuracy is estimated by the semantic similarity between the GPT-4 proposed name and the real GO term name. The best-fit regression line is shown in dark gray. The correlation coefficient (R) is determined by a two-sided Pearson’s correlation with p-value shown.



Extended Data Fig. 5 |. Distribution of ‘omics gene set sizes.
Distribution shown for all ‘omics gene sets considered in this study (n = 300).

Extended Data Table 1 |

Engineered prompt for gene set analysis

System content	You are an efficient and insightful assistant to a molecular biologist
Task instruction	Write a critical analysis of the biological processes performed by this system of interacting proteins. Base your analysis on prior knowledge available in your training data. After completing your analysis, propose a short descriptive name for the most prominent biological process performed by the system.

Confidence score assignment instruction	After completing your analysis, please also assign a confidence score to the process name you selected. This score should follow the name in parentheses and range from 0.00 to 1.00. A score of 0.00 indicates the lowest confidence, while 1.00 reflects the highest confidence. This score helps gauge how accurately the chosen name represents the functions and activities within the system of interacting proteins. When determining your score, consider the proportion of genes in the protein system that participate in the identified biological process. For instance, if you select "Ribosome biogenesis" as the process name but only a few genes in the system contribute to this process, the score should be lower compared to a scenario where a majority of the genes are involved in "Ribosome biogenesis".
Format instruction	Put your chosen name at the top of the analysis as 'Process: <name>'.
Analytical approach	Be concise, do not use unnecessary words. Be factual, do not editorialize. Be specific, avoid overly general statements such as 'the proteins are involved in various cellular processes'. Avoid listing facts about individual proteins. Instead, try to group proteins with similar functions and discuss their interplay, synergistic or antagonistic effects, and functional integration within the system. Also, avoid choosing generic process names such as 'Cellular Signaling and Regulation'. If you cannot identify a prominent biological process for the proteins in the system, I want you to communicate this in your analysis and name the process: "System of unrelated proteins". Provide a score of 0.00 for a "System of unrelated proteins".
One-shot example	To help you in your work, I am providing an example system of interacting proteins and the corresponding example analysis output. The example system of interacting proteins is: PDX1, SLC2A2, NKX6-1, GLP1, GCG. The example analysis output is: Process: Pancreatic development and glucose homeostasis (0.96) 1. PDX1 is a homeodomain transcription factor involved in the specification of the early pancreatic epithelium and its subsequent differentiation. It activates the transcription of several genes including insulin, somatostatin, glucokinase and glucose transporter type 2. It is essential for maintenance of the normal hormone-producing phenotype in the pancreatic beta-cell. In pancreatic acinar cells, it forms a complex with PBX1 b and MEIS2b and mediates the activation of the ELA1 enhancer. 2. NKX6-1 is also a transcription factor involved in the development of pancreatic beta-cells during the secondary transition. Together with NKX2-2 and IRX3, controls the generation of motor neurons in the neural tube and belongs to the neural progenitor factors induced by Sonic Hedgehog (SHH) signals. 3. GCG and GLP1, respectively glucagon and glucagon-like peptide 1, are involved in glucose metabolism and homeostasis. GCG raises blood glucose levels by promoting gluconeogenesis and is the counter regulatory hormone of Insulin. GLP1 is a potent stimulator of Glucose-Induced Insulin Secretion (GSIS). Plays roles in gastric motility and suppresses blood glucagon levels. Promotes growth of the intestinal epithelium and pancreatic islet mass both by islet neogenesis and islet cell proliferation. 4. SLC2A2, also known as GLUT2, is a facilitative hexose transporter. In hepatocytes, it mediates bi-directional transport of glucose across the plasma membranes, while in the pancreatic beta-cell, it is the main transporter responsible for glucose uptake and part of the cell's glucose-sensing mechanism. It is involved in glucose transport in the small intestine and kidney too. To summarize, the genes in this set are involved in the specification, differentiation, growth and functionality of the pancreas, with a particular emphasis on the pancreatic beta-cell. Particularly, the architecture of the pancreatic islet ensures proper glucose sensing and homeostasis via a number of different hormones and receptors that can elicit both synergistic and antagonistic effects in the pancreas itself and other peripheral tissues.
User input	Here are the interacting proteins: Proteins: {protein list}

The full prompt used to query the LLMs, separated into sections matching Fig. 1a.

Extended Data Table 2 |

Overview of five language models

Models	Version Release	Params	Context Length (tokens)	Company	Estimated Time Usage (second/gene set)	Estimated Cost ($/gene set)
GPT-4 Turbo	Nov 2023	~1.7T	128k	OpenAI	36.5 ‡	4.8×10^{-2}
Gemini Pro	Dec 2023	Unspecified	32k	Google	7.9	0.0
GPT-3.5 Turbo	Nov 2023	~175B	16k	OpenAI	9.6	2.8×10^{-3}
Mixtral Instruct	Dec 2023	13B (active), 47B (total)	32k	MistralAI	46.4	0.0 †
Llama2	July 2023	70B	4k	Meta	61.8	0.0 †

† Does not consider the cost to host an open-source model.

‡ GPT-4 compute time was significantly shorter (1.1s) when asking for a gene set name but not further analysis.

List of facts for five large language models used in this study.

Extended Data Table 3 |

Confidence assessment by GPT-4 versus human

		GPT-4 Self-assessed Confidence Score		Total
		High	Medium	
Human Reviewer's Proposed Confidence *	High	10	6	17
	Medium	1	8	10
Total		11	14	25

† Fisher's exact test p-value (two-sided) = 0.033

* Reviewer provided with GPT-4 gene set name and analysis essay but blinded to its self-reported confidence score.

We asked a human reviewer to read GPT-4's proposed name and supporting analysis text for 25 gene sets and independently assign high or medium confidence (the reviewer was blinded to GPT-4's own confidence assessment). The agreement between human and GPT-4 confidence assessment is presented in this table. Significance is determined using a two-sided Fisher's exact test.

Extended Data Table 4 |

Clusters named by **LLM** (GPT-4) versus enrichment (g:Profiler)

Minimal Gene Cluster Specificity *			GPT-4 LLM		Total
			Named †	Unnamed	
0%	g:Profiler Enrichment	Named ‡	124	105	229
		Unnamed	11	60	71
		Total	135	165	**300**
5%			Named †	Unnamed	
	g:Profiler Enrichment	Named ‡	33	51	84
		Unnamed	29	187	216
		Total	62	238	**300**
10%					

Minimal Gene Cluster Specificity*			GPT-4 LLM		Total
			Named†	Unnamed	
			Named†	Unnamed	
	g:Profiler Enrichment	Named‡	11	22	33
		Unnamed	31	236	267
		Total	42	258	**300**
20%			Named†	Unnamed	
	g:Profiler Enrichment	Named‡	0	4	4
		Unnamed	21	275	296
		Total	21	279	**300**

†Omics clusters named by GPT-4 LLM analysis with default settings (self-confidence > 0).

‡Omics clusters named by g:Profiler enrichment analysis with default settings (BH adjusted significance $p \leq 0.05$).

*Addl. requirement of a minimum % genes in cluster supporting the name (normalized by Jaccard index, Methods).

The number of omics gene clusters named by GPT-4 or by GO enrichment analysis using g:Profiler at the required specificity threshold measured by Jaccard Index (left most column).

Extended Data Table 5 |

Engineered prompt for identifying genes supporting a proposed name

Analyze the provided text, which describes a gene set's common functions and suggests a name reflecting its predominant function(s). Your task is to identify genes that support this name based solely on the information given in the text.

Context: Gene sets are groups of genes that share common biological functions, pathways, or other characteristics. Naming these sets based on their predominant functions helps researchers quickly understand their significance.

Input:
1. A list of gene symbols in the gene set, provided in comma-separated format:
<gene set>{genes_in_text}</gene set>

2. The analysis text:
<text>{text}</text>

3. The suggested name for the gene set:
<name>{name}</name>

Instructions:
1. Evaluate each gene from the provided list that is mentioned in the text.
2. Determine if the text makes a definite assertion about the gene that supports the given name.
- A definite assertion clearly states a gene's function or role without using speculative language.
- Example of a definite assertion: "XRCC1 is involved in the DNA damage response"
- Example of a non-definite assertion: "E2F1 may be involved in homologous recombination"
3. If a gene is mentioned multiple times, consider the strongest assertion made about it.
4. In case of contradictory statements about a gene, favor the most recent or specific assertion.
5. For each gene you determine supports the name:
- Briefly explain your reasoning (max 50 words per gene)
- Assign a confidence level (High, Medium, Low) based on the strength of the assertion
6. Handle acronyms or alternative gene names as equivalent to official gene symbols.
7. If no genes seem to support the name or if all genes support it, state this observation.

Output your analysis in the following format:

-- Summary --
[Provide a brief summary (max 100 words) of why the selected genes support the given name]

-- Explanation --
[Gene Symbol]: [Confidence Level]
[Explanation of reasoning (max 50 words)]

[Repeat for each supporting gene]

-- genes supporting the name: [List of gene symbols of genes supporting the name]

Do not critique the analysis or the name. Base your evaluation solely on the information provided in the text.

The full prompt used to query GPT-4 to identify genes supporting the proposed name.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by National Institutes of Health grants U24 CA269436 (R.T.P., D.F., K.S., T.I. and D.P.), OT2 OD032742 (M.H., T.I. and D.P.), U24 HG012107 (D.F., T.I. and D.P.) and U01 MH115747 (S.A. and T.I.). Additional support was received from Schmidt Futures (M.H. and T.I.). We thank X. Zhao and A. Singhal for insightful comments, M. R. Kelly for providing the NeST data raw files and C. Churas and J. Lenkiewicz for helping to improve the GitHub repository.

Data availability

All data used in this paper are publicly available. The full GO (2023-11-15 release) is downloaded from http://release.geneontology.org/2023-11-15/ontology/index.html. The selected NeST gene set is available to download from https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation/blob/main/data/Omics_data/NeST_IAS_clixo_hidef_Nov17.edges. The L1000 data used in this study are available at https://maayanlab.cloud/static/hdfs/harmonizome/data/lincscmapchemical/gene_attribute_edges.txt.gz. The viral infection data are available at https://maayanlab.cloud/static/hdfs/harmonizome/data/geovirus/gene_attribute_matrix.txt.gz. Detailed information on data download and parsing procedures, along with all datasets used in this paper, are available in our GitHub repository at https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation.

Code availability

The code to run the LLM gene set analysis pipeline and to reproduce results for the evaluation tasks is available via GitHub at https://github.com/idekerlab/llm_evaluation_for_gene_set_interpretation or Code Ocean[77] (https://doi.org/10.24433/CO.7045777.v1) with the MIT License. Note that LLM outputs are inherently stochastic and the precise names and analysis text produced by the models are not guaranteed to be the same from run to run. We minimized the variability of the outputs as described in 'Controlling the variability of LLM responses' section in Methods.

References

1. Zeeberg BR et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4, R28 (2003). [PubMed: 12702209]
2. Breitling R, Amtmann A & Herzyk P Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinf. 5, 34 (2004).
3. Beissbarth T & Speed TP GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20, 1464–1465 (2004). [PubMed: 14962934]

4. Subramanian A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005). [PubMed: 16199517]
5. Al-Shahrour F. et al. From genes to functional classes in the study of biological systems. BMC Bioinf. 8, 114 (2007).
6. Backes C. et al. GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res. 35, W186–W192 (2007). [PubMed: 17526521]
7. Huang DW, Sherman BT & Lempicki RA Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc 4, 44–57 (2009). [PubMed: 19131956]
8. Chen EY et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf. 14, 128 (2013).
9. Pomaznoy M, Ha B & Peters B GOnet: a tool for interactive Gene Ontology analysis. BMC Bioinf. 19, 470 (2018).
10. Cerami EG et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 39, D685–D690 (2011). [PubMed: 21071392]
11. Fabregat A. et al. The Reactome pathway Knowledgebase. Nucleic Acids Res. 44, D481–D487 (2015). [PubMed: 26656494]
12. Pico AR et al. WikiPathways: pathway editing for the people. PLoS Biol. 6, e184 (2008). [PubMed: 18651794]
13. Kanehisa M, Goto S, Sato Y, Furumichi M & Tanabe M KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–D114 (2012). [PubMed: 22080510]
14. Pillich RT et al. NDEx IQuery: a multi-method network gene set analysis leveraging the Network Data Exchange. Bioinformatics 39, btad118 (2023). [PubMed: 36882166]
15. Wang S. et al. Typing tumors using pathways selected by somatic evolution. Nat. Commun 9, 4159 (2018). [PubMed: 30297789]
16. Ashburner M. et al. Gene Ontology: tool for the unification of biology. Nat. Genet 25, 25–29 (2000). [PubMed: 10802651]
17. Gene Ontology Consortiumet al. The Gene Ontology knowledgebase in 2023. Genetics 224, iyad031 (2023). [PubMed: 36866529]
18. Kanehisa M & Goto S KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000). [PubMed: 10592173]
19. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28, 1947–1951 (2019). [PubMed: 31441146]
20. Kanehisa M, Furumichi M, Sato Y, Kawashima M & Ishiguro-Watanabe M KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 51, D587–D592 (2023). [PubMed: 36300620]
21. Croft D. Reactome: a database of biological pathways. Nat. Preced 10.1038/npre.2010.5025.1 (2010).
22. Jassal B. et al. The reactome pathway knowledgebase. Nucleic Acids Res. 48, D498–D503 (2020). [PubMed: 31691815]
23. Sollis E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 51, D977–D985 (2023). [PubMed: 36350656]
24. Blake JA et al. The Mouse Genome Database genotypes::phenotypes. Nucleic Acids Res. 37, D712–D719 (2009). [PubMed: 18981050]
25. Weng M-P & Liao B-Y MamPhEA: a web tool for mammalian phenotype enrichment analysis. Bioinformatics 26, 2212–2213 (2010). [PubMed: 20605928]
26. Keenan AB et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 47, W212–W224 (2019). [PubMed: 31114921]
27. Rubin JD et al. Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment. Commun. Biol 4, 661 (2021). [PubMed: 34079046]
28. Franzén O, Gan L-M & Björkegren JLM PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019, baz046 (2019).

29. Zhang X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. 47, D721–D728 (2019). [PubMed: 30289549]
30. Hu C. et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 51, D870–D876 (2023). [PubMed: 36300619]
31. Devlin J, Chang M-W, Lee K & Toutanova K BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (eds Burstein J et al..) 4171–4186 (Association for Computational Linguistics, 2019).
32. Brown TB et al. Language models are few-shot learners. In Advances in Neural Information Processing Systems (eds Larochelle H, et al.) 1877–190 (NeurIPS, 2020).
33. Vaswani A et al. Attention is all you need. Neural Inf. Process Syst 30, 5998–6008 (2017).
34. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
35. Touvron H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
36. Jiang AQ et al. Mixtral of experts. Preprint at https://arxiv.org/abs/2401.04088 (2024).
37. Gemini Team et al. Gemini: a family of highly capable multimodal models. Preprint at https://arxiv.org/abs/2312.11805 (2023).
38. Joachimiak MP, Harry Caufield J, Harris NL, Kim H & Mungall CJ Gene set summarization using large language models. Preprint at https://arxiv.org/abs/2305.13338 (2023).
39. Moghaddam SR & Honey CJ Boosting theory-of-mind performance in large language models via prompting. Preprint at https://arxiv.org/abs/2304.11490 (2023).
40. Hebenstreit K, Praas R, Kiesewetter LP & Samwald M An automatically discovered chain-of-thought prompt generalizes to novel models and datasets. Preprint at https://arxiv.org/abs/2305.02897 (2023).
41. Wei J. et al. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (eds Koyejo S et al.) 24824–24837 (NeurIPS, 2022).
42. Caufield JH et al. Structured prompt interrogation and recursive extraction of semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. Bioinformatics 40, btae104 (2024). [PubMed: 38383067]
43. Miller GA & Charles WG Contextual correlates of semantic similarity. Lang. Cogn. Process 6, 1–28 (1991).
44. Xiong M. et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In The Twelfth International Conference on Learning Representations (ICLR, 20234).
45. Fu J, Ng S-K, Jiang Z & Liu P GPTScore: evaluate as you desire. In Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers (eds Duh K et al.) 6556–6576 (Association for Computational Linguistics, 2024).
46. Kolberg L. et al. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 51, W207–W212 (2023). [PubMed: 37144459]
47. Duan Q. et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 42, W449–W460 (2014). [PubMed: 24906883]
48. Subramanian A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000. Profiles Cell 171, 1437–1452.e17 (2017). [PubMed: 29195078]
49. Barrett T. et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 41, D991–D995 (2013). [PubMed: 23193258]
50. Zheng F. et al. Interpretation of cancer mutations using a multiscale map of protein systems. Science 374, eabf3067 (2021). [PubMed: 34591613]

51. Pinkas DM et al. Structural complexity in the KCTD family of Cullin3-dependent E3 ubiquitin ligases. Biochem. J 474, 3747–3761 (2017). [PubMed: 28963344]
52. Dhanoa BS, Cogliati T, Satish AG, Bruford EA & Friedman JS Update on the Kelch-like (KLHL) gene family. Hum. Genomics 7, 13 (2013). [PubMed: 23676014]
53. Pleiner T. et al. WNK1 is an assembly factor for the human ER membrane protein complex. Mol. Cell 81, 2693–2704.e12 (2021). [PubMed: 33964204]
54. Berthold J. et al. Characterization of RhoBTB-dependent Cul3 ubiquitin ligase complexes—evidence for an autoregulatory mechanism. Exp. Cell. Res 314, 3453–3465 (2008). [PubMed: 18835386]
55. McCormick JA et al. Hyperkalemic hypertension-associated cullin 3 promotes WNK signaling by degrading KLHL3. J. Clin. Invest 124, 4723–4736 (2014). [PubMed: 25250572]
56. Sohara E & Uchida S Kelch-like 3/Cullin 3 ubiquitin ligase complex and WNK signaling in salt-sensitive hypertension and electrolyte disorder. Nephrol. Dial. Transpl 31, 1417–1424 (2016).
57. Tang H, Finn RD & Thomas PD TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. Bioinformatics 35, 518–520 (2019). [PubMed: 30032202]
58. Groh BS et al. The antiobesity factor WDTC1 suppresses adipogenesis via the CRL4WDTC1 E3 ligase. EMBO Rep. 17, 638–647 (2016). [PubMed: 27113764]
59. Ji W & Rivero F Atypical rho GTPases of the RhoBTB subfamily: roles in vesicle trafficking and tumorigenesis. Cells 5, 28 (2016). [PubMed: 27314390]
60. Nori H, King N, McKinney SM, Carignan D & Horvitz E Capabilities of GPT-4 on medical challenge problems. Preprint at https://arxiv.org/abs/2303.13375 (2023).
61. López Espejel J, Ettifouri EH, Yahaya Alassan MS, Chouham EM & Dahhane W GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot learning and performance boosting through prompts. Nat. Lang. Process. J 5, 100032 (2023).
62. Yu H. et al. Evaluation of retrieval-augmented generation: a survey. Preprint at https://arxiv.org/abs/2405.07437 (2024).
63. Yao S. et al. ReAct: synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations (ICLR, 2022).
64. Nair V, Schumacher E, Tso G & Kannan A DERA: enhancing large language model completions with dialog-enabled resolving agents. In Proc. 6th Clinical Natural Language Processing Workshop (eds Naumann T et al.) 122–161 (2023).
65. Shinn N et al. Reflexion: language agents with verbal reinforcement learning. In Advances in Neural Information Processing Systems (eds Oh A et al.) 8634–8652 (NeurIPS, 2023).
66. Li G, Al Kader Hammoud HA, Itani H, Khizbullin D & Ghanem B CAMEL: Communicative Agents for 'Mind' Exploration of Large Scale Language Model Society. In Advances in Neural Information Processing Systems (eds Oh A et al.) 36, 51991–52008 (NeurIPS, 2023).
67. Schick T. et al. Toolformer: language models can teach themselves to use tools. In Advances in Neural Information Processing Systems (eds Oh A et al.) 68539–68551 (NeurIPS, 2023).
68. Shen Y et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face. In Advances in Neural Information Processing Systems (eds Oh A et al.) 36, 38154–38180 (NeurIPS, 2023).
69. Keskar NS, McCann B, Varshney LR, Xiong C & Socher R CTRL: a conditional transformer language model for controllable generation. Preprint at https://arxiv.org/abs/1909.05858 (2019).
70. Holtzman A, Buys J, Du L, Forbes M & Choi Y The curious case of neural text degeneration. In The Eighth International Conference on Learning Representations (ICLR, 2020).
71. Smith B. et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol 25, 1251–1255 (2007). [PubMed: 17989687]
72. Tirmizi SH et al. Mapping between the OBO and OWL ontology languages. J. Biomed. Semant 2, S3 (2011).
73. Liu F, Shareghi E, Meng Z, Basaldella M & Collier N Self-alignment pretraining for biomedical entity representations. In Proc. 2021 Conference of the North American Chapter of the Association

for Computational Linguistics: Human Language Technologies (eds Toutanova K et al.) 4228–4238 (Association for Computational Linguistics, 2021).

74. Wolf T. et al. Transformers: state-of-the-art natural language processing. In Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (eds Liu Q & Schlangen D) 38–45 (Association for Computational Linguistics, 2020).
75. Rouillard AD et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database 2016, baw100 (2016).
76. Seal RL et al. Genenames.org: the HGNC resources in 2023. Nucleic Acids Res. 51, D1003–D1009 (2023). [PubMed: 36243972]
77. Hu M. et al. Evaluation of Large Language Models for Discovery of Gene Set Function (Code Ocean, 2024); 10.24433/CO.7045777.V1

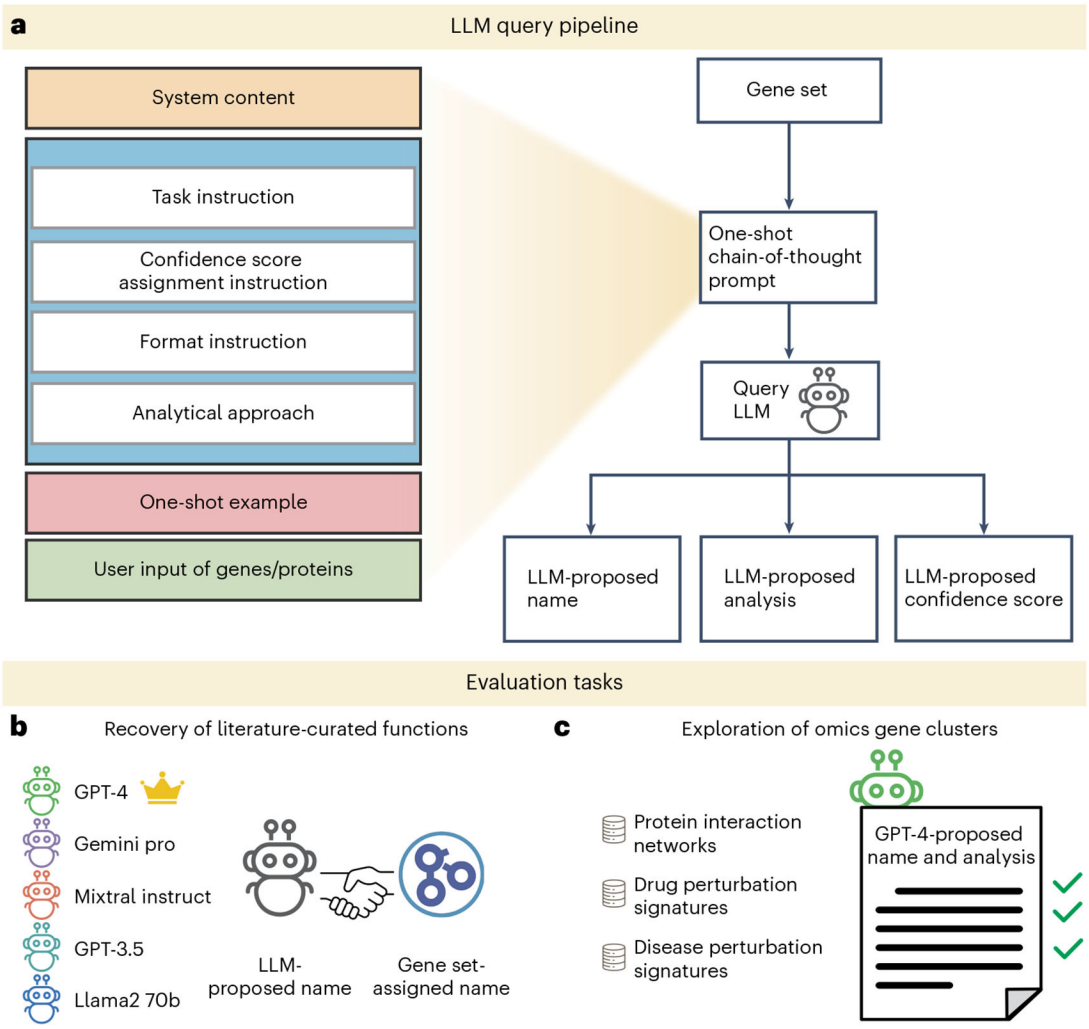


Fig. 1 | Use and evaluation of LLMs for functional analysis of gene sets.

a, The LLM prompt (left boxes) includes system content, detailed chain of thought instructions, and an example gene set query with desired response (full prompt given in Extended Data Table 1). The specific list of genes is inserted into the 'user input of genes/proteins' field at the end of the prompt template, resulting in generation of a proposed name, a supporting analysis essay and a confidence score (right flowchart). **b**, Benchmarking LLM names against names assigned by GO (evaluation task 1). The proposed name from each of five LLMs (left robot icons) is compared with the name assigned by the GO curators (handshake icon). GPT-4 (crowned) was the winning model for this task. **c**, Exploration of gene sets discovered in omics data (evaluation task 2). The GPT-4 name and analysis are scored for novelty and accuracy (right green check marks). Gene sets derived from three different data types (left database icons).

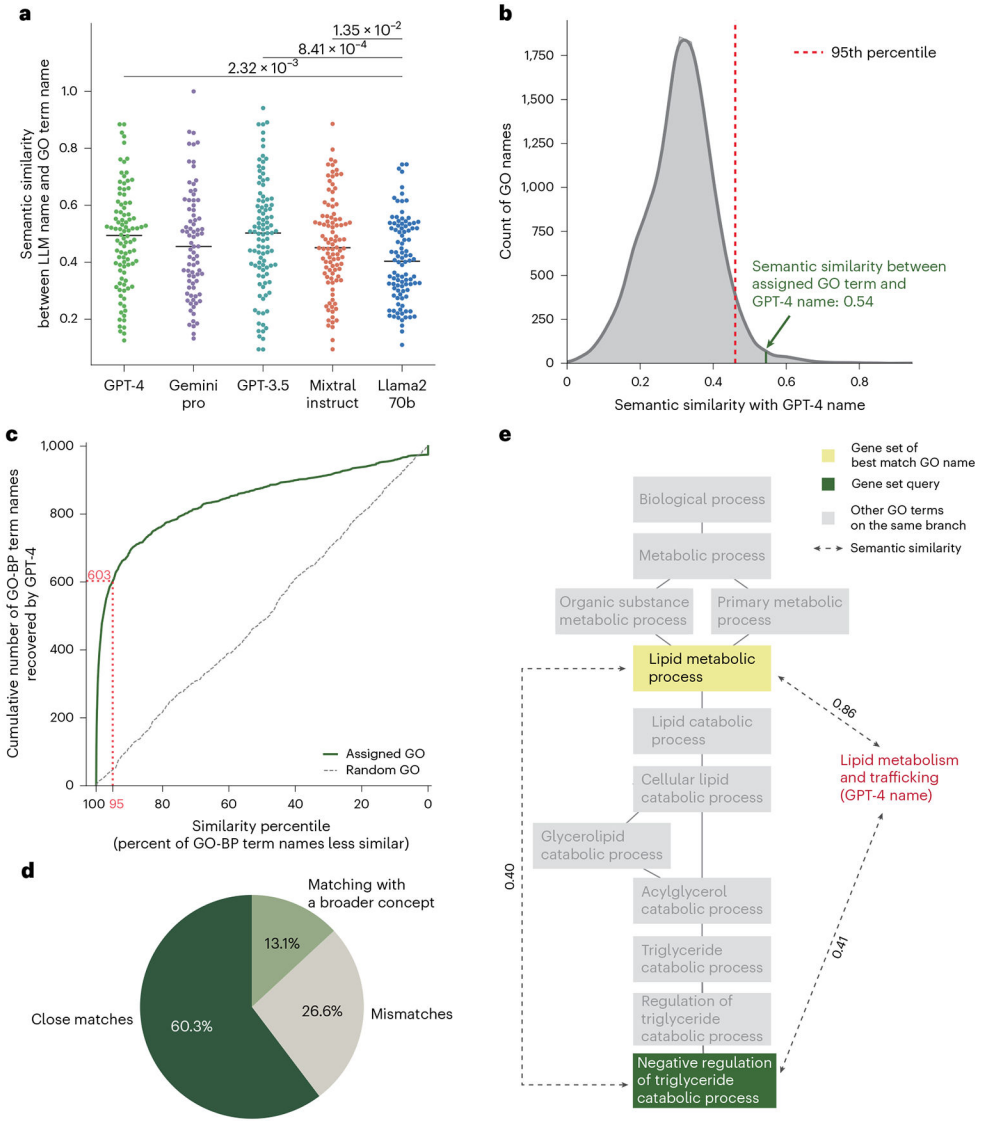


Fig. 2 | Evaluation of LLMs in recovering GO gene set names.

a, The performance of each LLM (colors) scored by semantic similarity between its proposed name for a gene set and the name assigned by GO curators. Results for 100 GO terms are shown (dots; the horizontal black lines show median semantic similarities). Significant difference in distributions is determined using a two-sided Mann–Whitney *U* test. **b**, The percentile calibration of semantic similarity between the GO and GPT-4 names for a gene set, shown for the GO term 'response to X-ray' and the corresponding GPT-4 name 'DNA damage response and repair'. The plot shows the semantic similarity between these two names (vertical dark-green line, 0.54) versus the complete distribution of semantic similarity scores between the GPT-4 name and each name in the GO biological process database (GO-BP, gray). The GPT-4 name score is converted to a percentile, that is, the percentage of all names in GO with lower similarity (here, 99%). The dashed red line denotes the 95th percentile threshold. **c**, The cumulative number of GO term names recovered by GPT-4 (**y** axis) at a given similarity percentile (x axis). 0, least similar; 100, most similar. The dark-

green curve shows the semantic similarities between GPT-4 names and assigned GO term names. The dashed gray curve shows the semantic similarities between GPT-4 names and random GO term names. The dotted red line marks the number of GO names recovered by GPT-4 at the 95th similarity percentile. **d**, A pie chart summarizing the results of the GPT-4 name/GO name similarity comparison. **e**, A hierarchical view of the GO term ‘negative regulation of triglyceride catabolic process’ and its ancestors. Blue box: gene set query; yellow box: gene set of best match GO name (most similar GO name to GPT-4 name); dashed lines with arrows: semantic similarities between names; red text: GPT-4 proposed name.

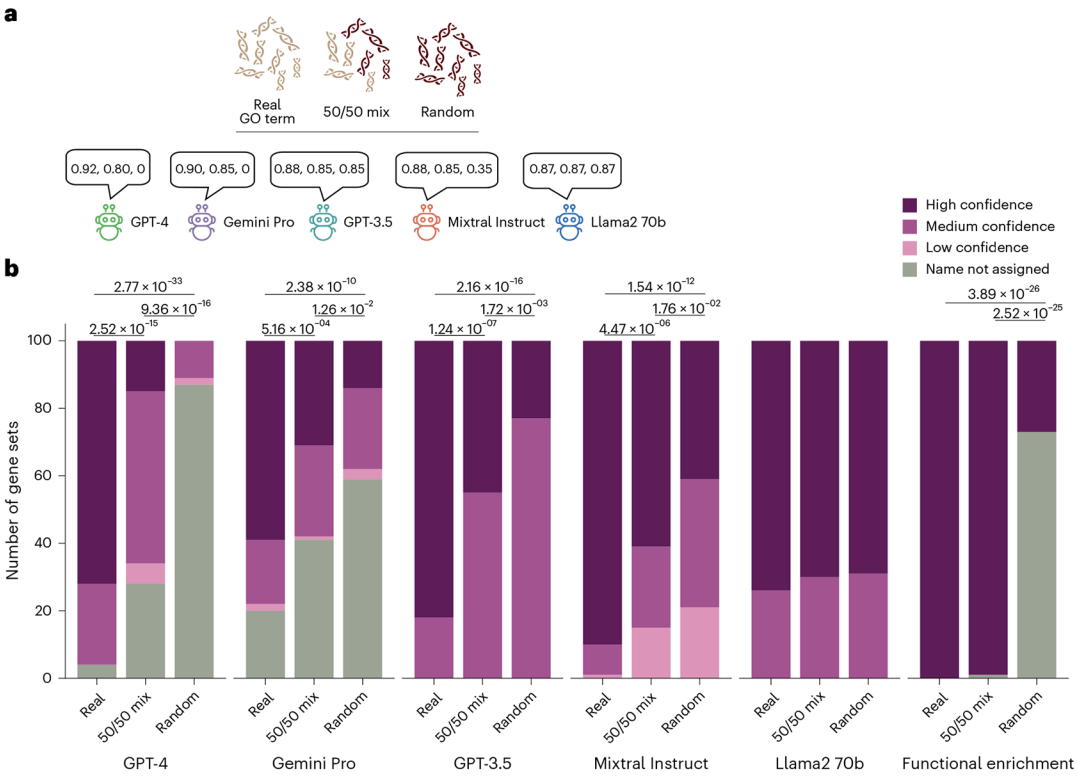


Fig. 3 |. Evaluation of LLM self-confidence.

a, Investigation of model-assigned confidence scores (chat bubbles) for the ability to distinguish real GO terms from 50/50 mix and random gene sets (light DNA strands from the same GO term, dark DNA strands randomly selected from outside the GO term). **b**, Bar graphs showing the confidence rating assigned by each model for real, contaminated or random gene sets. Increasing shades of purple indicate low to high score bins. ‘High confidence’ (dark purple): 0.87–1.00; ‘medium confidence’ (medium purple): 0.82–0.86; ‘low confidence’ (light purple): 0.01–0.81; ‘name not assigned’ (gray): 0. For comparison with functional enrichment (rightmost group of bars), ‘high confidence’ for a gene set is defined as BH-adjusted $P \leq 0.05$ (dark purple, g:Profiler[46] with Benjamini–Hochberg correction), otherwise ‘name not assigned’ (gray) is used. A significant difference in confidence distributions between real, 50/50 mix and random is determined using a two-sided chi-squared test.

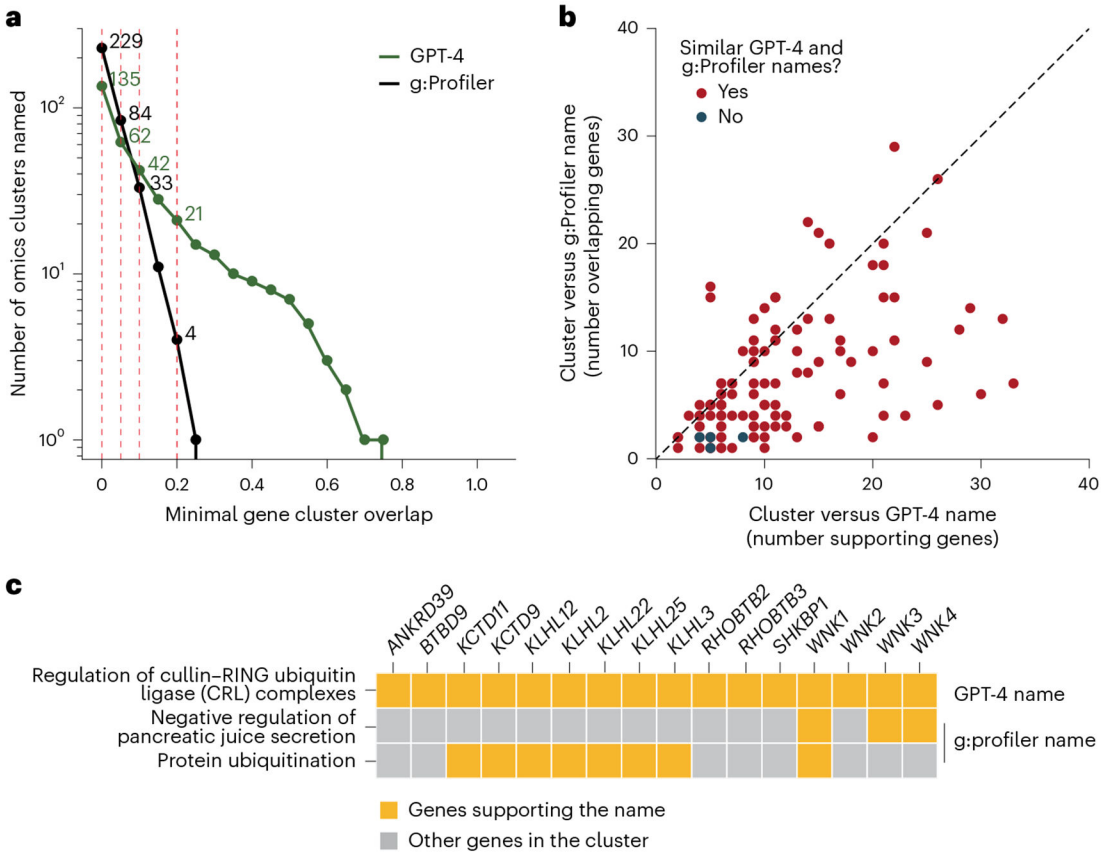


Fig. 4 |. Evaluation of GPT-4 in naming 'omics gene clusters.

a, The number of omics gene clusters (*y* axis, $\log_{10}$ scale) named by GPT-4 (dark green) or by GO enrichment analysis using g:Profiler (black; BH-adjusted $P \leq 0.05$) versus the gene cluster specificity threshold measured by the Jaccard index (*x* axis; Methods). The vertical dashed red lines mark the same specificity thresholds shown in Extended Data Table 4. **b**, The number of cluster genes overlapping the genes associated with g:Profiler enriched GO term (*y* axis) is plotted against the number of genes in support of the GPT-4 name (*x* axis). The red points represent GPT-4 names highly similar to a significant g:Profiler name (semantic similarity ≥0.5); otherwise, navy color is used. The dotted black diagonal denotes equal specificity for the GPT-4 and g:Profiler names. **c**, Alternate names for cluster NeST:2-105 are shown (rows), with yellow boxes indicating which names support each of the cluster genes (columns). The GPT-4 name is shown first in bold (top), while the remaining rows highlight two of the significant g:Profiler results: the GO term with the best *P* value of enrichment (middle) and the term most conceptually similar to the GPT-4 name (bottom).

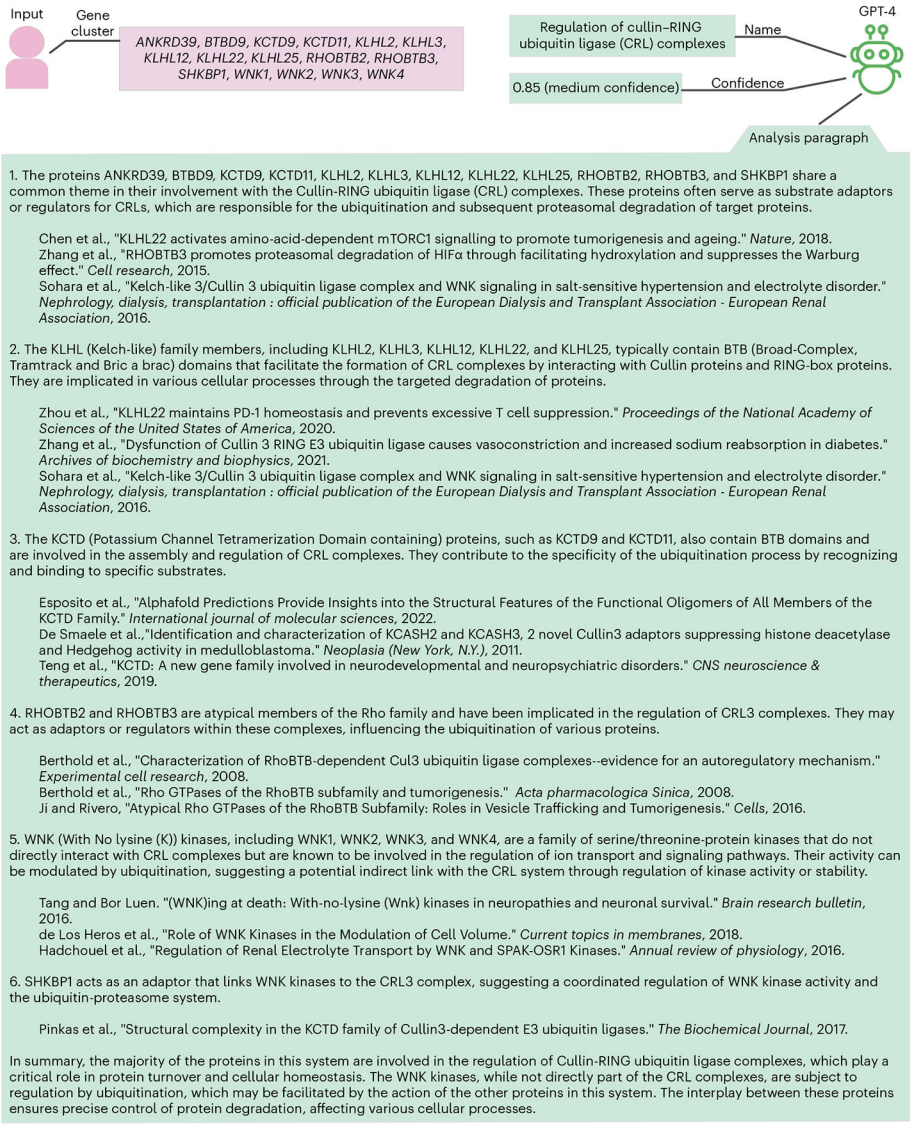


Fig. 5 l. Representative analysis for protein interaction clusters (NeST:2-105).

Input gene set, 16 genes (top left pink box); GPT-4 generated cluster name (top right green box); GPT-4 confidence score (middle right green box); GPT-4 analysis text (bottom green box). Each generated paragraph is followed by the associated citations found by the citation module (Extended Data Fig. 1 and Methods).

Table 1 |

Best and worst **LLM** names for GO terms by semantic similarity

GO name (GO term ID)	LLM **name**	**Semantic similarity**	LLM
Synaptic vesicle exocytosis (GO:0016079)	Synaptic vesicle exocytosis	1.00	Gemini Pro
Synaptic vesicle exocytosis (GO:0016079)	Synaptic vesicle exocytosis and neurotransmitter release	0.94	GPT-3.5
Pentose-phosphate shunt (GO:0006098)	Pentose phosphate pathway	0.89	GPT-3.5
Glucose-6-phosphate transport (GO:0015760)	Glucose-6-phosphate metabolism and transport	0.89	Mixtral Instruct
Protein quality control for misfolded or incompletely synthesized proteins (GO:0006515)	Protein quality control and degradation	0.88	GPT-4
Negative regulation of fat cell differentiation (GO:0045599)	Regulation of Wnt signaling and cellular stress response	0.13	GPT-4
Negative regulation of CD8-positive, alpha–beta T cell differentiation (GO:0043377)	Regulation of iron homeostasis	0.11	Llama2 70b
Negative regulation of peptide secretion (GO:0002792)	Glucose homeostasis and energy metabolism	0.09	GPT-3.5
Negative regulation of peptide secretion (GO:0002792)	Glucose homeostasis and energy metabolism	0.09	Mixtral Instruct
Negative regulation of CD8-positive, alpha–beta T cell differentiation (GO:0043377)	Regulation of ion transport and cellular homeostasis	0.09	GPT-3.5