Phenomics Assistant: An Interface for LLM-based Biomedical Knowledge Graph Exploration

Shawn T O'Neil[1*], Kevin Schaper[1], Glass Elsarboukh[1], Justin T Reese[2], Sierra A T Moxon[2], Nomi L Harris[2], Monica C Munoz-Torres[1], Peter N Robinson[3], Melissa A Haendel[1], Christopher J Mungall[2]

[1]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, 80045, USA; [2]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA; [3]The Jackson Laboratory for Genomic Medicine, Farmington CT, 06032, USA

* Corresponding author: shawn@tislab.org

Abstract

We introduce Phenomics Assistant, a prototype chat-based interface for querying the Monarch knowledge graph (KG), a comprehensive biomedical database. While unaided Large Large Language models (LLMs) are prone to mistakes in factual recall, their strong abilities in summarization and tool use suggest new opportunities to help non-expert users query and interact with complex data, while drawing on the KG to improve reliability of the answers. Leveraging the ability of LLMs to interpret queries in natural language, Phenomics Assistant enables a wide range of users to interactively discover relationships between diseases, genes, and phenotypes.

To assess the reliability of our approach and compare the accuracy of different LLMs, we evaluated Phenomics Assistant answers on benchmark tasks for gene-disease association and gene alias queries. While comparisons across tested LLMs revealed differences in their ability to interpret KG-provided information, we found that even basic KG access markedly boosts the reliability of standalone LLMs. By enabling users to pose queries in natural language and summarizing results in familiar terms, Phenomics Assistant represents a new approach for navigating the Monarch KG.

Introduction

Large language models (LLMs) represent a new paradigm in human-computer interaction, allowing users to work with systems in their native language. LLMs excel in summarizing,

paraphrasing, and explaining in-context information [1,2], spurring the growth of knowledge-backed AI agents, capable of using tools to search for and contextualize externally-sourced information [3]. The information they generate, however, is not always accurate, particularly for information that is not well represented in training data [4].

Knowledge graphs (KGs) are a powerful approach for integrating heterogeneous data and enabling the data to be queried to discover new insights; they are widely used in biomedicine and beyond. In translational research, KGs are frequently used to represent known relationships between biomedical entities such as diseases, genes, and phenotypes, where insights into these relationships can lead to improved treatments [5]. The Monarch Initiative KG includes millions of known, curated associations across hundreds of thousands of entities for dozens of species [6]. However, the sheer volume and complexity of genetic data pose significant challenges in terms of accessibility and interpretation. While there are a number of interfaces for querying such data (Figure 1), including specialized query languages [7], graphical interfaces [8], and information-rich websites and APIs [6,9], using these effectively often requires domain-specific vocabulary and knowledge, limiting their utility for a broader range of users, including clinicians and researchers without extensive bioinformatics training. Integrating the translational information stored in biomedical KGs with the user-friendly features of LLMs thus presents a promising direction for enhancing the accessibility of the Monarch KG.
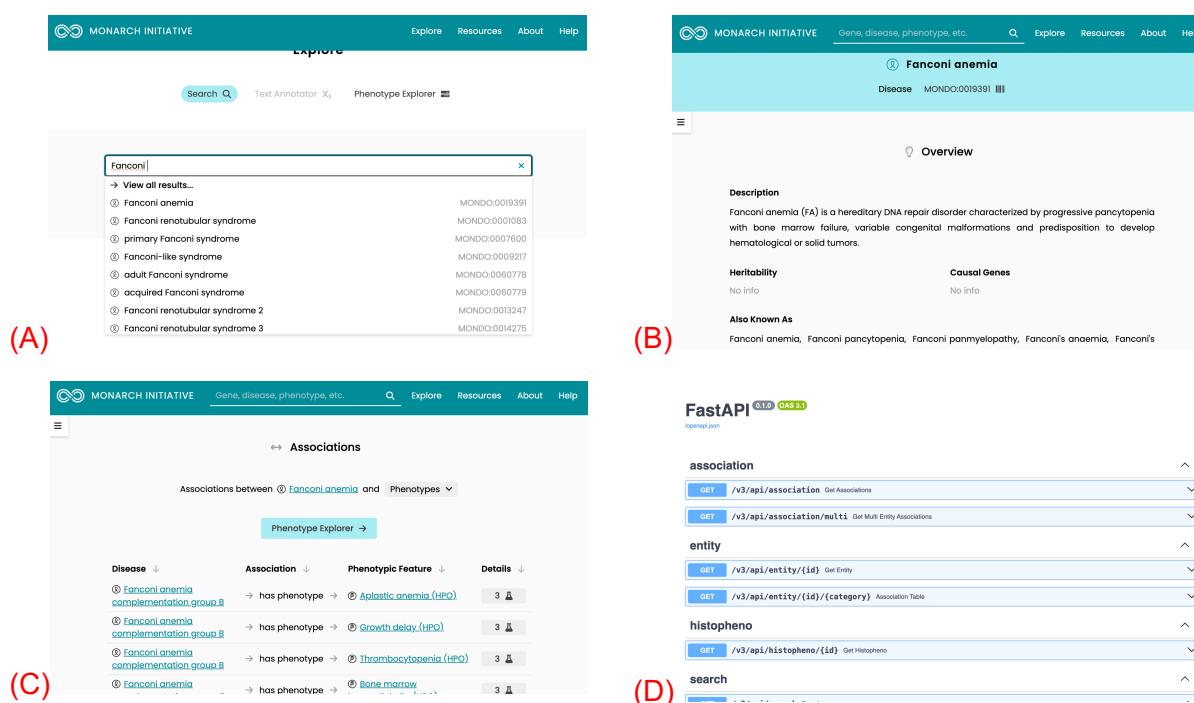

Figure 1: Monarch Initiative interfaces. While biomedical databases such as Monarch host vast amounts of information, their interfaces are generally designed for domain experts to search (A), quickly review related information (B) and known associations (C), or provide access via API (D).

In this paper we introduce Phenomics Assistant, an LLM-based interface for searching, retrieving, and summarizing information in the Monarch Initiative KG. We describe its prototype

implementation as a user-friendly web application, and analyze its performance on natural language benchmark tasks for gene name identification and gene-disease associations, comparing several LLM models with and without KG access. These results highlight the importance of providing curated KG information to LLMs, and reveal differences in the ability of different LLMs to use the provided information. On average, LLMs with KG access produced between 1.9X and 5.1X more correct answers than those without KG access in our tests.

The integration of LLMs with domain-specific, curated knowledge bases like the Monarch Knowledge Graph presents a new avenue for scientific question-answering in the field of genomics. While LLMs are adept at generating coherent and contextually relevant responses, their reliance on training data can lead to inaccuracies or omissions, particularly in specialized domains [4]. Phenomics Assistant addresses this by grounding LLM responses in the verified data of the KG, enhancing the information provided with links to sources and other information. By allowing users to pose queries in natural language and summarizing data in familiar terms, Phenomics Assistant democratizes access to complex genomic information, making it more readily available to a diverse audience.

Phenomics Assistant is still in active development. A demonstration deployment can be accessed from the GitHub repository at https://github.com/monarch-initiative/phenomics-assistant.

Related Work

Augmenting LLMs with external tools or curated data is a common approach to improving their factual accuracy and reasoning abilities [3]. Many LLM systems incorporate free-text document databases, for example to support question-and-answer tasks over scientific literature [10]. LLMs have demonstrated proficiency in accurately summarizing structured data, and they can be used with relational databases to assist non-specialists in data exploration and querying [11]. LLMs may also be configured to access APIs; this technique powers ChatGPT "plugins," allowing the AI to access information or take action on users' behalf [12], and has been used to access data via scientific APIs such as PubMed's E-utils [13].

A variety of applications have integrated LLMs and knowledge graphs, including assisting in the development or curation of KGs by extracting entities and relationships from free text [14–18]. Conversely, a number of researchers have explored providing KG data to LLMs to improve the reliability of the answers that are generated [19]. In biomedical applications, some approaches utilize semantic similarity search via embedding vectors. Recent approaches include embedding search followed by neighborhood retrieval, filtering the associations by query similarity for LLM summarization [20], neighborhood filtering based on a query classification [21], additionally considering document collections [22], rewriting neighborhood descriptions in text [23], decomposing queries into logical constructs [24], fine-tuning model weights with additional KG-sourced training data [25], and developing specialized graph neural networks for improved reasoning [26].

As we discuss below, Phenomics Assistant accesses KG information via the Monarch API rather than directly, placing it closer in spirit to ChatGPT plugins and other API-accessing LLM utilities, and thus more readily deployable over existing infrastructure. As our evaluations show, this API-backed approach makes effective use of KG structure, including information about classes of diseases and phenotypes, resulting in significant improvements over unaided LLMs.

Methods

Architecture

Phenomics Assistant consists of several components illustrated in Figure 2. Users interact with a prototype chat-based user interface that connects to an LLM framework, mediating access to a subset of functionality of the Monarch Initiative API.
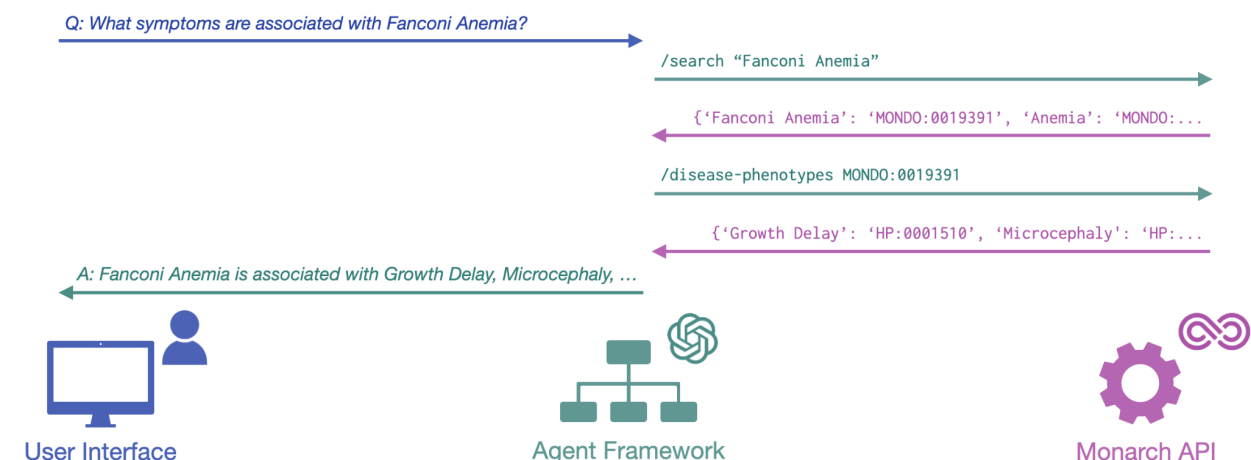

Figure 2: Phenomics Assistant architecture. Users pose questions via the UI in natural language (blue), and these are translated to Monarch API calls by the LLM and agent framework (green). Responses from the API (pink) are evaluated by the LLM, and may trigger followup calls and responses, until a final answer is returned to the user in natural language.

API: The standard Monarch Initiative API provides a wide variety of functions, including keyword-based entity search, flexible entity-association retrieval, and semantic similarity search. While LLMs can be adapted to call such functions as "tools," minimizing the number and complexity of available tools reduces potential for errors caused by inappropriate tool use [27]. For Phenomics Assistant we thus utilize a small set of LLM-focused functions, including search and individual association-type lookups, described in Table 1. Importantly, the association endpoints consider subclass closures for queried entities, matching the behavior of the Monarch website interface. Gene associations for Ehler-Danlos Syndrome (MONDO:0017314), for example, will also include associations for the autosomal dominant and recessive subtypes (MONDO:0007524, MONDO:0002014). Gene associations always include both causal and

correlated relationships, distinguishing between the two. Returned lists of associations include URLs to publications or evidence information when available from the KG.

Table 1: Available API functions and parameters. Function names, parameters, and their descriptions are supplied as part of the prompt to the LLM. All functions also include optional *limit* and *offset* parameters (described as part of the LLM prompt as "The maximum number of search results to return" and "Offset for pagination of results" respectively, with defaults 0 and 10).

Function (Description)	Parameters (Description)
/search ("Search for entities in the Monarch knowledge graph")	term ("The ontology term to search for.") category ("A single category to search within as a string. Valid categories are: biolink:Disease, biolink:PhenotypicQuality, and biolink:Gene". Default: "biolink:Disease".)
/disease-genes ("Get a list of genes associated with a disease")	disease_id ("The ontology identifier of the disease.")
/disease-phenotypes ("Get a list of phenotypes associated with a disease")	disease_id ("The ontology identifier of the disease.")
/gene-diseases ("Get a list of diseases associated with a gene")	gene_id ("The ontology identifier of the gene.")
/gene-phenotypes ("Get a list of phenotypes associated with a gene")	gene_id ("The ontology identifier of the gene.")
/phenotype-diseases ("Get a list of diseases associated with a phenotype")	phenotype_id ("The ontology identifier of the phenotype.")
/phenotype-genes ("Get a list of genes associated with a phenotype")	phenotype_id ("The ontology identifier of the phenotype.")

Agent Framework: Some LLMs are trained to "call" external functions when provided with callable function metadata as part of the conversation context, by responding to queries with the names and parameters of functions to call. These specially-formatted responses are then parsed and executed locally before including the results in followup responses to the model. Utilizing OpenAI models that support function-calling [28], we developed an agent-based framework that extracts function metadata (functions, parameters, and descriptions from Table 1) and provides it to the LLM, executes LLM-specified calls, and returns results in JSON format. The framework also manages LLM choice (e.g. GPT 3.5 or GPT 4), conversation history, toxicity checks using OpenAI's moderation API [29], and the system prompt, which is used by many LLMs to guide model behavior. Phenomics Assistant's system prompt instructs the model to use lay language descriptions and include links to external pages when possible (Suppl. Table 1).

User Interface: A web application provides a chat interface to Phenomics Assistant. The current interface is developed with the Streamlit web framework, enabling rapid feature prototyping. Users enter questions in natural language, and the interface provides real-time feedback on functions being called, optionally including call parameters and results in the conversation stream (Figure 3). Responses include formatted links to external pages or publications when available (Figure 4). Such transparency features can increase confidence in results by allowing users to check answers against returned data and follow up with external resources.
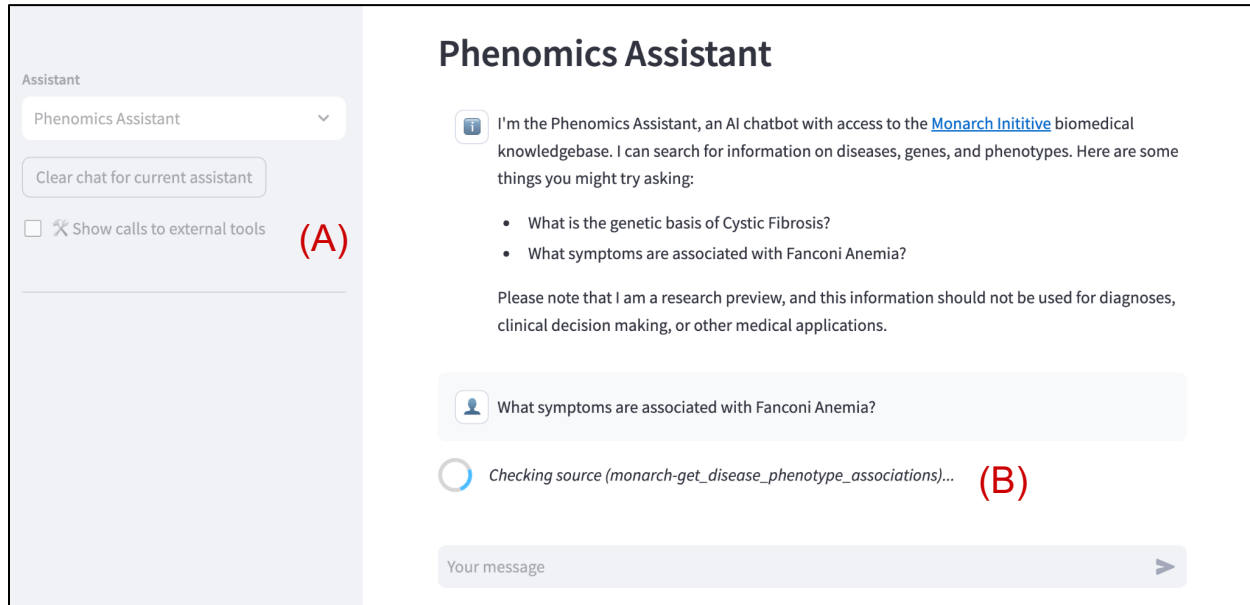

Figure 3: Phenomics Assistant prototype user interface. Users can select from different Assistant variations and optionally include calls and responses in the conversation stream (A). Real-time feedback alerts the user to specific endpoints or functions being called (B).

Fanconi anemia is associated with a variety of symptoms, which in medical terms are referred to as phenotypes. Here are some of the symptoms associated with Fanconi anemia:

1. **Delayed CNS myelination**: This refers to a slower than normal development of the myelin sheath, the protective covering that surrounds nerve fibers in the central nervous system. Read more. Publication

2. **Unilateral renal agenesis**: This is a condition where one kidney fails to develop. Read more. Publication

3. **Depressed nasal tip**: The tip of the nose is lower than usual. Read more. Publication

4. **Micropenis**: An abnormally small penis. Read more. Publication

5. **Cleft palate**: A split or opening in the roof of the mouth. Read more. Publication

6. **Renal hypoplasia**: Abnormally small kidneys. Read more. OMIM Record

7. **Bone marrow hypocellularity**: A lower than normal number of cells in the bone marrow. Read more. OMIM Record

8. **Anemia**: A condition in which there is a deficiency of red cells or of hemoglobin in the blood. Read more. OMIM Record

9. **Anal atresia**: Absence or abnormal narrowing of an opening to the anus. Read more. Publication

10. **Cafe-au-lait spot**: Light brown skin patches. Read more. Publication

Please note that the presence and severity of these symptoms can vary among individuals with Fanconi anemia.

Figure 4: Example query response. Links are included to relevant pages and data sources when available. This response required two API calls (not shown), first to `/search` for "Fanconi Anemia" to retrieve the relevant entity identifier in the KG, and second to `/disease-phenotypes` to retrieve the relevant associations.

Evaluation

We compared the accuracy of three different LLMs (OpenAI's `gpt-3.5-turbo-0613`, `gpt-4-0613`, and `gpt-4-1106-preview` models) with and without KG access on the *gene alias* and *gene-disease association* tasks of the GeneTuring benchmark dataset [30]. Each task consists of 50 question and gold-standard answer pairs. Gene alias questions ask for an official gene symbol for a non-standard gene name, for example "What is the official gene symbol of LMP10?" with the gold-standard answer being "PSMB10". For this task, the GeneTuring authors specify a Jaccard similarity score, comparing the set of symbols mentioned in a given answer to the size-one set containing the gold standard, thereby penalizing additional mentioned names. Gene-disease association asks for the set of gene names associated with a disease, for example "What are genes related to Distal renal tubular acidosis?" with the gold-standard answer being "SLC4A1, APT6V0A4". This task prescribes a recall metric, computed as the percentage of gold-standard genes mentioned in an answer (thus not penalizing additional mentioned names). The GeneTuring authors sourced data for the gene alias task from NCBI, and for gene-disease associations from OMIM [31], which is also a source for Monarch data[6].

All models use `temperature = 0`, minimizing (but not eliminating) stochastic variation in model responses. Models with KG access are configured similarly to those available in the Phenomics Assistant web interface at the time of testing, including the system prompt (Suppl. Table 1). Models without KG access were given the system prompt "You are a helpful assistant."

Agent answers are provided as free text in markdown format (Figure 4). Computing metric scores thus requires extracting identifiers mentioned in answers for comparison to gold-standard answers. For example, given the question "What are genes related to Congenital disorder of deglycosylation?" with expected gold-standard answer of "MAN2C1, NGLY1" and LLM-generated answer of "Genes associated with Congenital disorder of deglycosylation include MAN2C1, PMM2, and ALG6", we must compute the recall of the set (MAN2C1, PMM2, ALG6) against (MAN2C1, NGLY1). We accomplish this in an automated fashion by instantiating `gpt-4-0613`-based "evaluator" agents provided with functions that compute recall or Jaccard scores as appropriate from generated answer text. These LLM-callable functions take entity lists as parameters and return computed scores, allowing the evaluators to use function-calling for simultaneous named entity extraction and computational evaluation. While powerful, such model-based evaluations themselves require validation [32], particularly when applied to model-supplied outputs [33]. To validate this approach, we compute scores manually for 20 questions selected at random for each task (40 total) for comparison to evaluator-produced scores. Finally, we use single-sided, paired Wilcoxon signed-rank tests (`wilcox.test` in R version 4.2.2) to assess improvement in scores for models with KG access compared to those without.

Results

The gene alias and disease-association tasks prescribe Jaccard similarity and recall metrics respectively, with both having a [0,1] range and 1.0 perfect score. Across models and tasks our results were largely bimodal, with most answers scoring 0.0 or 1.0. Manual evaluation for 20 random questions (out of 300 across tested models) from each task identified between 0 and 6 gene names per answer for scoring; LLM-evaluator extractions and scores agreed in 100% of these questions using case-sensitive exact-match criteria.

Figure 5 illustrates score counts across agents and tasks, revealing strong performance increases for models that are able to query the knowledge graph. For gene-disease associations, addition of knowledge graph resources increased the number of fully correct (score 1.0) answers by an average of 1.9X across models, and decreased wholly incorrect (0.0) answers by an average of 5.3X. Correct answer counts for gene alias increased by 5.1X with KG access and incorrect answer counts decreased by 2.7X. Table 2 analyzes these gains statistically; with the exception of GPT 3.5 for the gene-diseases association task, the estimated median score improvement with KG access is 1.0.
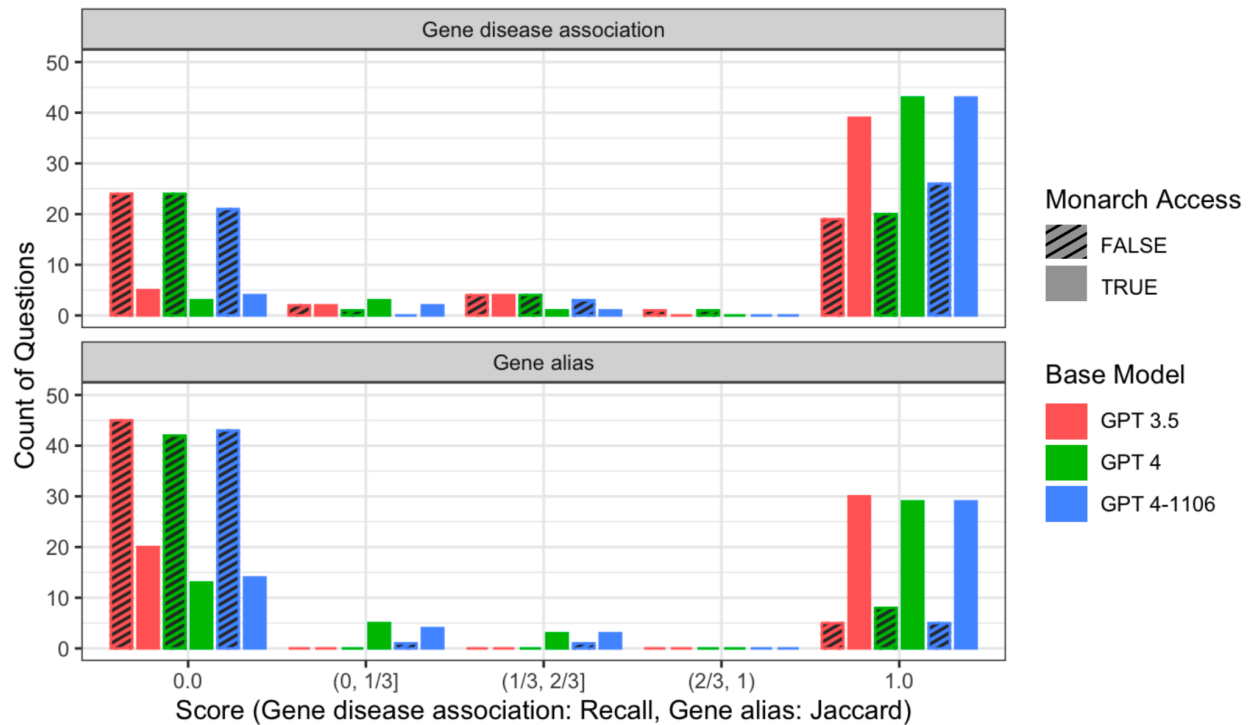
Figure 5: Evaluation results. Results are shown for three models (`gpt-3.5-turbo-0613`, `gpt-4-0613`, and `gpt-4-1106-preview`) with and without access to the Monarch knowledge graph API, on two GeneTuring evaluation tasks, gene-disease association and gene alias. Scores of 0.0 and 1.0 are counted separately; other scores are binned according to ranges shown in parentheses below the X axis.

Table 2: Statistical comparisons of answer scores with and without KG access, per base model and task. Scores and score differences are non-normal, so we use single-sided paired Wilcoxon signed rank tests to estimate median score improvements for models with KG access compared to those with no access. *p* values are Bonferroni corrected across the six tests.

Task	Base model	Estimated median KG-access score improvement	*p* value (adjusted)
Gene alias	GPT 3.5	1.0	0.00000482
Gene alias	GPT 4	1.0	0.0000113
Gene alias	GPT 4-1106	1.0	0.00000158
Gene disease association	GPT 3.5	0.75	0.000167
Gene disease association	GPT 4	1.0	0.0000158
Gene disease association	GPT 4-1106	1.0	0.000193

Score differences across models with KG backing were small in comparison, suggesting that KG access is the major determinant of performance. For comparison, non-KG-backed GPT-3.5

scored 1.0 on 39% (19/50) of gene-disease answers, in line with results reported by the GeneTuring authors. The latest model, `gpt-4-1106-preview`, improves on this metric to 52% (26/50) without KG access.

Model Comparisons

Although performance of GPT 3.5 and GPT 4-1106 was largely similar across the 100 posed questions, Table 3 lists 10 cases where the latter outperformed the former, and 4 cases of the opposite. Instances where GPT 4-1106 outperformed 3.5 appear to be influenced by suboptimal ordering of initial results from the `/search` function.. For example, a search for CXorf40B in response to "What is the official symbol of CXorf40B?" returns two entries: first CXorf40B, the name for chromosome X open reading frame 40B in the species *Gallus gallus*, and second EOLA2, the name for the homologous gene in *Homo sapiens.* Arguably, in this instance GPT 4-1106 better interpreted the question intent. Similarly, answering gene-disease questions requires a two step process: first, a search by disease name for the appropriate identifier, which may return multiple hits, followed by fetching associations for one of the results. In response to "What are genes related to Trichoepithelioma?" a search for "trichoepithelioma" returns multiple results, including Vulvar Trichoepithelioma (MONDO:0002201) first, followed by Familial Multiple Trichoepithelioma (MONDO:0011114). In this example GPT 3.5 followed up on the former, which is not associated with the prescribed gold-standard answer of CYLD, while 4-1106 followed up on the latter, which is. Similarly, a search for "proteasome-associated autoinflammatory syndrome" (PRAAS) matches multiple entries, and in the results PRAAS Type 5 (PRAAS5) is listed first; again, GPT 3.5 followed up with the first entry while 4-1106 chose the more appropriate generic condition (PRAAS) to fetch associations. As a final example, spinal muscular atrophy with congenital bone fractures (SMABF) has two types caused by different genes: SMABF1 caused by TRIP4, and SMABF2 caused by ASCC1. The search result listed both diseases and their descriptions, which include the causal gene names. Both 3.5 and 4-1106 followed up by fetching associations for the first, SMABF1. However, while 3.5 listed only the result of the latter query (TRIP4), 4-1106 included both gene names from the earlier search-provided information.

Table 3 lists four questions where GPT 3.5 outperformed 4-1106. Three of these are for gene alias questions, where GPT 4 included all results returned by the initial search, regardless of the species, while 3.5 listed only the first. The use of Jaccard similarity thus penalizes 4-1106's comprehensiveness in comparison to 3.5. The question about diabetes is malformed - "What are genes related to Type diabetes mellitus?". In this case the initial search listed Type I diabetes first and Type II diabetes second. GPT 3.5 followed up with associations for Type I, while GPT 4-1106 followed up with Type II.

Table 3: Questions for which GPT 4-1106 + KG performed differently than GPT 3.5 + KG. Better performance by GPT 4-1106 is largely due to improved interpretation of search results independent of result order. The best answers and scores in each row are bolded.

Task	Question	Gold Standard	Score, 4-1106 + KG	Answer Genes, 4-1106 + KG	Score, 3.5 + KG	Answer Genes, 3.5 + KG
Gene alias	What is the official gene symbol of MPS4B?	GLB1	1	**GLB1**	0	
Gene alias	What is the official gene symbol of CXorf40B?	EOLA2	1	**EOLA2**	0	CXorf40B
Gene disease association	What are genes related to Trichoepithelioma?	CYLD	1	**CYLD**	0	
Gene disease association	What are genes related to Proteasome-associated autoinflammatory syndrome?	PSMB9, PSMG2, POMP, PSMB10	1	**PSMB10, PSMG2, PSMB8, PSMB4, PSMB9, POMP**	0.25	PSMB10
Gene alias	What is the official gene symbol of DCSCRIPT?	ZNF366	0.5	**ZNF366, ZNF366.L**	0	znf366
Gene alias	What is the official gene symbol of 15-LOX?	ALOX15	0.5	**ALOX15, ALOX15B**	0	ALOX15B
Gene disease association	What are genes related to Spinal muscular atrophy with congenital bone fractures?	TRIP4, ASCC1	1	**TRIP4, ASCC1**	0.5	TRIP4
Gene disease association	What are genes related to Gastrointestinal defects and immunodeficiency syndrome?	PI4KA, TTC7A	1	**PI4KA, TTC7A**	0.5	PI4KA
Gene alias	What is the official gene symbol of PTH1?	PTH	0.3333333 333	**PTH1R, PTH, PTRH1**	0	PTH1

Gene alias	What is the official gene symbol of GCS1?	MOGS	0.25	**GCS1, ADAP1, MOGS, Mogs**	0	GCS1
Gene disease association	What are genes related to Type diabetes mellitus? [*Malformed question– see main text*]	HNF1B, IL6, GPD2, HMGA1, IRS1, NEUROD1, IL6	0.1428571 429	HNF1A, IL6, ITPR3, PTPN22	0.428571 4286	**HNF1B, AKT2, GCK, ABCC8, PAX4, PPP1R3A, SLC2A2, HNF1A, TCF7L2, WFS1, SLC30A8, RETN, IGF2BP2, ENPP1, GPD2, HMGA1**
Gene alias	What is the official gene symbol of CT116?	LYPD6B	0.5	LYPD6B, SPANXN1	1	**LYPD6B**
Gene alias	What is the official gene symbol of PTP-SL?	PTPRR	0.2	PTPRR, Ptprr, ptprr, ptprr.L, ptprr.S	1	**PTPRR**
Gene alias	What is the official gene symbol of hCAP?	RNGTT	0.1	RNGTT, SCLT1, SMC3, DCD, NCAPD3, NCAPG, NCAPG2, NCAPD2, GEMIN4, NCAPH	1	**RNGTT**

Even for the most advanced GPT 4-1106 model, KG access resulted in increased scores on 30 gene alias and 19 gene-disease association questions (Suppl. Table 2). This finding is unsurprising, because LLMs are trained on large corpora of texts which may include inaccuracies, or may state factual truths in ways that are hard for a machine to understand. Of these 49 questions, GPT 4-1106 without KG access listed gene identifiers for 29, only three of which were partially correct (score > 0). GPT 4-1106 scored higher without KG access than with for only three questions; two of these were gene alias questions (seeking official gene names for DCSCRIPT and 15-LOX, see also Table 3) where KG-provided information provided additional context penalized by the Jaccard metric. The third sought genes related to

Hyperphenylalaninemia. Here base GPT 4-1106 identified the six-gene gold standard set exactly, but when provided with KG access followed up on the first search result, Mild hyperphenylalaninemia (MONDO:0019335), which is only associated with one of these six.

Table 4 shows mean scores across questions per task for GPT 4-1106 with KG access, compared to similarly computed averages from other published methods on the same tasks. In the original GeneTuring work, Hou et al. consider multiple models, including GPT 3, ChatGPT (GPT 3.5), and New Bing (a version of GPT 4 with web search capabilities) [30]. Jin et al. developed GeneGPT, an approach similar to Phenomics Assistant with access to NCBI APIs and backed by the now discontinued OpenAI Codex model [13]. GPT 4-1106 with KG access performs best on gene-disease association tasks, slightly ahead of the web-enabled New Bing. GeneGPT performs best on gene alias tasks.

Table 4: Mean task scores for 4-1106 + Monarch compared to results reported by Hou et al. and Jin et al. [13,30]**.**

Task	GPT 3	ChatGPT	New Bing	GeneGPT (best)	4-1106 with KG
Gene disease association	0.34	0.31	0.84	0.76	**0.87**
Gene alias	0.09	0.07	0.66	**0.84**	0.62

Discussion

Phenomics Assistant is designed to help users to explore the primary entities in the Monarch KG – genes, diseases, and phenotypes – and known relationships between them. While the Monarch API provides keyword search and relationship information, the Assistant handles organization and summarization of the results for the user, along with links to available evidence and the Monarch website for more information. While not shown by default, the raw calls and retrieved data may be inspected by the user for transparency. In use, we have observed that the Assistant also successfully interprets incorrect or loosely-specified requests, for example searching for "Cystic Fibrosis" in response to "What genes are involved with CF?" and correcting misspelled disease names, such as searching for the correct "Ehlers-Danlos Syndrome" when asked about the incorrect "Ehlen-Danlos Syndrome." On the other hand, while the natural-language descriptions of complex diseases and phenotypes are more accessible to a lay audience, they are in many cases generated by the LLM. While plausible, in a separate study we have observed that nuanced issues can infiltrate AI-generated definitions [15], and further work is needed to inform responses with KG-provided definitions.

Overall, we found that evaluation scores were significantly higher for models with access to Monarch data. For gene-disease associations the weakest model with KG access outperformed the strongest base model with 1.5X more correct answers; for gene aliases, over 3X more.

However, unaided models performed reasonably for the gene-disease association task, with the latest GPT 4-1106 improving over GPT 3.5 with an additional 7 correct answers.

While access to Monarch data was the largest determinant of performance, model quality is also an important factor in how that data is utilized. This was particularly true when a list of retrieved information was given in suboptimal order. In several cases, GPT 3.5 used only the first entry returned, even when other choices would be more appropriate from the question context. Such order effects are well known [34], and can even contribute to hallucination (inaccurate results) in the presence of biased few-shot learning examples [35]. GPT 4-1106 was improved in this regard, and in one example (SMABF) incorporated information from both an initial search and subsequent association.

Finally, these results highlight the importance of both data representation and evaluation criteria when assessing performance of retrieval-based systems. All variants of Phenomics Assistant performed better on finding gene-disease associations than on identifying gene aliases, which may be expected given that source data for the former (OMIM) was present in the Monarch KG at the time of testing, while the source for the latter (NCBI gene aliases) was not. This distinction is reflected in opposite trends for GeneGPT, which utilizes NCBI APIs but not OMIM. This is confounded, however, by the different scoring criteria prescribed by the GeneTuring evaluation tasks. Gene-disease associations are scored according to the percentage of gold-standard genes identified, accommodating extra information provided in answers, whereas the Jaccard-based scoring for gene alias tasks does not. We prompted models to be comprehensive in answers, and since the Monarch KG contains information on many species, gene alias answers were frequently penalized by the Jaccard metric for providing aliases for multiple species, whereas GeneTuring references only human genes and diseases. Still, the same trend was seen for unaided base models, agreeing with the original GeneTuring test results and potentially impacted by GPT 4's relative verbosity [36].

While Phenomics Assistant demonstrates good performance generally, open questions and future considerations remain. Many LLM-backed applications perform semantically-aware search via text embeddings [37], whereas the search functionality provided by the Monarch API is keyword-based. Given the observed sensitivity of some models to search result order, re-ranking results by embedding similarity, or implementing embedding-based semantic search directly, may further improve results. Similarly, we've implemented association retrieval between genes, diseases, and phenotypes, but the Monarch KG contains many other entity types and associations, and validating generic association-retrieval functionality for LLM use is a high priority.

Although we've focused the prototype Phenomics Assistant on the Monarch KG, additional access to other knowledge graphs could readily expand its capabilities to new domains. Furthermore, interface elements employed by many KG-driven applications such as plots, tables, and other widgets may complement the chat-only UI. Finally, we have thus far only tested OpenAI function-calling LLMs. Other closed-source models such as Claude (Anthropic)

and Gemini (Google) have also started supporting function calling; some open source models have recently begun incorporating this feature [38].

Regardless of the base model, these results highlight the utility of LLM-based user interfaces in interacting with curated knowledge, dramatically improving the completeness and accuracy of generative AI models. Simultaneously, LLMs show promise in democratizing access to large, complex knowledge bases, effectively searching, summarizing, and contextualizing the information for end users.

Data and Code Availability

Phenomics Assistant components are all open source and available on GitHub.

Main code repository: https://github.com/monarch-initiative/phenomics-assistant

User interface, with link to demonstration deployment: https://github.com/monarch-initiative/phenomics-assistant (see link in README)

API: https://github.com/monarch-initiative/oai-monarch-plugin

Agent framework: https://github.com/monarch-initiative/agent-smith-ai

Evaluation and results: https://github.com/monarch-initiative/oai-plugin-evals

Bibliography

1. Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking Large Language Models for News Summarization. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2301.13848

2. Leinonen J, Denny P, MacNeil S, Sarsa S, Bernstein S, Kim J, et al. Comparing Code Explanations Created by Students and Large Language Models. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2304.03938

3. Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, et al. Augmented Language Models: a Survey. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.07842

4. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large Language Models Struggle to Learn Long-Tail Knowledge. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning. PMLR; 23--29 Jul 2023. pp. 15696–15707.

5. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6: 252ra123.

6. Putman TE, Schaper K, Matentzoglu N, Rubinetti VP, Alquaddoomi FS, Cox C, et al. The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. Nucleic Acids Res. 2023. doi:10.1093/nar/gkad1082

7. Cox S, Ahalt SC, Balhoff J, Bizon C, Fecho K, Kebede Y, et al. Visualization Environment for Federated Knowledge Graphs: Development of an Interactive Biomedical Query Language and Web Application Interface. JMIR Med Inform. 2020;8: e17964.

8. Morton K, Wang P, Bizon C, Cox S, Balhoff J, Kebede Y, et al. ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. Bioinformatics. 2019;35: 5382–5384.

9. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43: D789–98.

10. Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, et al. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. NPJ Digit Med. 2021;4: 68.

11. Zhang W, Wang Y, Song Y, Wei VJ, Tian Y, Qi Y, et al. Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2310.17894

12. Iqbal U, Kohno T, Roesner F. LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins. arXiv [cs.CR]. 2023. Available: http://arxiv.org/abs/2309.10254

13. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: Augmenting Large Language Models with Domain Tools for Improved Access to Biomedical Information. ArXiv. 2023. Available: https://www.ncbi.nlm.nih.gov/pubmed/37131884

14. Carta S, Giuliani A, Piano L, Podda AS, Pompianu L, Tiddia SG. Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2307.01128

15. Toro S, Anagnostopoulos AV, Bello S, Blumberg K, Cameron R, Carmody L, et al. Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2312.10904

16. Matentzoglu N, Harry Caufield J, Hegde HB, Reese JT, Moxon S, Kim H, et al. MapperGPT: Large Language Models for Linking and Mapping Entities. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2310.03666

17. Baek J, Aji AF, Saffari A. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2306.04136

18. Mihindukulasooriya N, Tiwari S, Enguix CF, Lata K. Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text. The Semantic Web – ISWC 2023. Springer Nature Switzerland; 2023. pp. 247–265.

19. Yu W, Zhu C, Li Z, Hu Z, Wang Q, Ji H, et al. A Survey of Knowledge-enhanced Text Generation. ACM Comput Surv. 2022;54: 1–38.

20. Soman K, Rose PW, Morris JH, Akbas RE, Smith B, Peetoom B, et al. Biomedical

knowledge graph-enhanced prompt generation for large language models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2311.17330

21. Suneera, Prakash J, Singh PK. Question answering over knowledge graphs using BERT based relation mapping. Expert Syst. 2023;40. doi:10.1111/exsy.13456

22. Guo Q, Cao S, Yi Z. A medical question answering system using large language models and knowledge graphs. Int J Intell Syst. 2022;37: 8548–8564.

23. Wu Y, Hu N, Bi S, Qi G, Ren J, Xie A, et al. Retrieve-rewrite-answer: A KG-to-Text enhanced LLMs framework for knowledge graph question answering. arXiv [cs.CL]. 2023. Available: https://github.com/wuyike2000/Retrieve-Rewrite-Answer

24. Choudhary N, Reddy CK. Complex Logical Reasoning over Knowledge Graphs using Large Language Models. arXiv [cs.LO]. 2023. Available: http://arxiv.org/abs/2305.01157

25. Varshney D, Zafar A, Behera NK, Ekbal A. Knowledge grounded medical dialogue generation using augmented graphs. Sci Rep. 2023;13: 3310.

26. Yasunaga M, Ren H, Bosselut A, Liang P, Leskovec J. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. 2021; 535–546.

27. Ruan J, Chen Y, Zhang B, Xu Z, Bao T, Du G, et al. TPTU: Large Language Model-based AI Agents for Task Planning and Tool Usage. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2308.03427

28. OpenAI Platform. [cited 8 Nov 2023]. Available: https://platform.openai.com/docs/guides/function-calling

29. O'Neil ST, Mungall C. monarch-initiative/agent-smith-ai: v1.0.0. doi:10.5281/zenodo.8361491

30. Hou W, Ji Z. GeneTuring tests GPT models in genomics. bioRxiv. 2023. p. 2023.03.11.532238. doi:10.1101/2023.03.11.532238

31. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33: D514–7.

32. Kamalloo E, Dziri N, Clarke CLA, Rafiei D. Evaluating open-domain question answering in the era of large language models. Annual Meeting of the Association for Computational Linguistics. 2023 [cited 8 Nov 2023]. doi:10.48550/ARXIV.2305.06984

33. Dai S, Zhou Y, Pang L, Liu W, Hu X, Liu Y, et al. LLMs may Dominate Information Access: Neural Retrievers are Biased Towards LLM-Generated Texts. arXiv [cs.IR]. 2023. Available: http://arxiv.org/abs/2310.20501

34. Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. Muresan S, Nakov P, Villavicencio A, editors. 2022; 8086–8098.

35. Turpin M, Michael J, Perez E, Bowman SR. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv [cs.CL]. 2023.

Available: http://arxiv.org/abs/2305.04388

36. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2309.07430

37. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2306.08302

38. Srinivasan VK, Dong Z, Zhu B, Yu B, Mao H, Mosk-Aoyama D, et al. NexusRaven: a commercially-permissive Language Model for function calling. 2023 Workshop on …. 2023. Available: https://openreview.net/pdf?id=Md6RUrGz67

Supplementary Materials

Suppl. Table 1

Suppl. Table 1: System prompts used by models for evaluation. Information regarding the 'function call setting' pertains to functionality disabled during evaluation.

Models	System Prompt
3.5 + Monarch, 4 + Monarch, 4-1106 + Monarch	You are the Monarch Assistant, an AI-powered chatbot that can answer questions about data from the Monarch Initiative knowledge graph. You can search for entities such as genes, diseases, and phenotypes by name to get the associated ontology identifier. You can retrieve associations between entities via their identifiers. Users may use synonyms such as 'illness' or 'symptom'. Do not assume the user is familiar with biomedical terminology. Always add additional information such as lay descriptions of phenotypes. If the user changes the show function call setting, do not make any further function calls immediately. IMPORTANT: Include markdown-formatted links to the Monarch Initiative for all results using the templates provided by function call responses.
3.5 Base, 4 Base, 4-1106 Base	You are a helpful assistant.

Suppl. Table 2

Suppl. Table 2: Questions where 4-1106 + Monarch was scored differently than 4-1106 Base, ordered by score difference. Answer Genes columns list gene names extracted from free-text agent answers by evaluator agents. 4-1106 Base without Monarch access outperformed 4-1106 + Monarch in three instances due to scoring criteria and model performance (see Results).

Task	Question	Gold Standard	Score, 4-1106 + Monarch	Answer Genes, 4-1106 + Monarch	Score, 4-1106 Base	Answer Genes, 4-1106 Base
Gene alias	What is the official gene symbol of HsT1192?	SYT4	1	SYT4	0	
Gene alias	What is the official gene symbol of PFDN3?	VBP1	1	VBP1	0	PFDN3, Prefoldin Subunit 3
Gene alias	What is the official gene symbol of LEP9?	LCE2A	1	LCE2A	0	
Gene alias	What is the official gene symbol of LMP10?	PSMB10	1	PSMB10	0	PSMB8
Gene alias	What is the official gene symbol of CTTNBP1?	SHANK2	1	SHANK2	0	CTTNBP2NL
Gene alias	What is the official gene symbol of SGEF?	ARHGEF26	1	ARHGEF26	0	SGEF
Gene alias	What is the official gene symbol of ZNF482?	ZBTB6	1	ZBTB6	0	ZNF482
Gene alias	What is the official gene symbol of M12.219?	ADAMDEC1	1	ADAMDEC1	0	
Gene alias	What is the official gene symbol of CXorf40B?	EOLA2	1	EOLA2	0	ENSG00000284747, CXorf40B
Gene alias	What is the official gene symbol of PLK-	PLK5	1	PLK5	0	PLK1, PLK2, PLK3, PLK4

	5?					
Gene alias	What is the official gene symbol of C20orf86?	ANKRD60	1	ANKRD60	0	MFSD3
Gene alias	What is the official gene symbol of PKCL?	PRKCH	1	PRKCH	0	PRKCL1
Gene alias	What is the official gene symbol of CKBBP1?	RNF7	1	RNF7	0	Ckb
Gene alias	What is the official gene symbol of SEP3?	SEPTIN3	1	SEPTIN3	0	SEPALLATA3
Gene alias	What is the official gene symbol of AGTIL?	ASIP	1	ASIP	0	AGT
Gene alias	What is the official gene symbol of ASV?	SRC	1	SRC	0	
Gene alias	What is the official gene symbol of PCPB?	CPB2	1	CPB2	0	PYCR1
Gene alias	What is the official gene symbol of C11orf27?	UBTFL1	1	UBTFL1	0	HMBS
Gene alias	What is the official gene symbol of CTGLF6?	AGAP9	1	AGAP9	0	
Gene alias	What is the official gene symbol of NPAP60L?	NUP50	1	NUP50	0	NUP210L
Gene alias	What is the official gene symbol of TSH2B?	H2BC1	1	H2BC1	0	TSHB
Gene alias	What is the official gene symbol of BMSC-MCP?	SLC25A33	1	SLC25A33	0	
Gene alias	What is the official gene symbol of C20orf195?	FNDC11	1	FNDC11	0	MFSD3
Gene alias	What is the official gene symbol of QSCN6L1?	QSOX2	1	QSOX2	0	NXF1

Gene alias	What is the official gene symbol of C6orf186?	METTL24	1	METTL24	0	MROH8
Gene disease association	What are genes related to Sensorineural deafness with mild renal dysfunction?	BSND	1	BSND	0	MYO7A, COL4A3, COL4A4, COL4A5, SLC26A4, GJB2, GJB6, OTOF, WFS1, MITF
Gene disease association	What are genes related to Neurodevelopmental disorder with nonspecific brain abnormalities and with or without seizures?	DLL1	1	DLL1	0	SCN1A, MECP2, CDKL5, STXBP1, TSC1, TSC2, FOXG1, ARX, SLC6A1, CHD2, SYNGAP1
Gene disease association	What are genes related to Immunodeficiency due to defect in MAPBP-interacting protein?	LAMTOR2	1	LAMTOR2	0	UFL1
Gene disease association	What are genes related to Chronic atrial and intestinal dysrhythmia?	SGO1	1	SGO1	0	SCN5A, HLA genes, KCNQ1, ANK2
Gene disease association	What are genes related to Split-foot malformation with mesoaxial polydactyly?	MAP3K20	1	MAP3K20	0	TP63, WNT10B, FBXW4, DLX5, DLX6, BHLHA9
Gene disease association	What are genes related to Congenital disorder of deglycosylation?	NGLY1, MAN2C1	1	MAN2C1, NGLY1	0	PMM2, ALG6, MPI, ALG3, ALG12, PMM1
Gene disease association	What are genes related to Spinal muscular atrophy with congenital bone fractures?	TRIP4, ASCC1	1	TRIP4, ASCC1	0	SMN1, SMN2
Gene disease association	What are genes related to EDICT syndrome?	MIR184	1	MIR184	0	CUL4B
Gene disease	What are genes related to	SCP2	1	SCP2	0	POLR3A

association	Leukoencephalopathy with dystonia and motor neuropathy?				
Gene disease association	What are genes related to Siddiqi syndrome?	FITM2	1	FITM2	0	USB1
Gene disease association	What are genes related to Immunodeficiency with inflammatory disease and congenital thrombocytopenia?	ARPC1B	1	ARPC1B	0	WASP, CYBB, DOCK8, GATA1, ANKRD26, MYH9, RAB27A, STXBP2, ITGB3, GPIBA
Gene disease association	What are genes related to Corneal fleck dystrophy?	PIKFYVE	1	PIKFYVE	0	
Gene disease association	What are genes related to Intracranial hemorrhage in brain cerebrovascular malformations?	IL6	1	KRAS, IL6	0	CCM1/KRIT1, CCM2/MGC4607, CCM3/PDCD10, ENG, ACVRL1, RASA1, SMAD4, ADAMTS13, COL3A1, NOTCH3, TGFBR1, TGFBR2, FOXC1, PITX2, SOX17
Gene disease association	What are genes related to Neurodevelopmental disorder with gait disturbance?	TCEAL1	1	TCEAL1	0	SCN1A, MECP2, FMR1, DMD, GNAO1, ADGRG1, KMT2A, TSC1, TSC2, ANKRD11, SLC6A1
Gene disease association	What are genes related to Lichtenstein-Knorr syndrome?	SLC9A1	1	SLC9A1	0	
Gene disease association	What are genes related to Gastrointestinal defects and immunodeficiency syndrome?	PI4KA, TTC7A	1	PI4KA, TTC7A	0	FOXP3, ITGB2, XIAP, BIRC4, IL10RA, IL10RB, SKIV2L, TTC37, RAG1, RAG2, ADA, CYBB
Gene alias	What is the official gene symbol of AF10?	MLLT10	1	MLLT10	0.5	KMT2A, MLLT10

Gene alias	What is the official gene symbol of CT116?	LYPD6B	0.5	LYPD6B, SPANXN1	0	
Gene disease association	What are genes related to Proteasome-associated autoinflammatory syndrome?	PSMB9, PSMG2, POMP, PSMB10	1	PSMB10, PSMG2, PSMB8, PSMB4, PSMB9, POMP	0.5	PSMB8, PSMB4, PSMB9, PSMA3, POMP, PSMD12
Gene disease association	What are genes related to Pigmented nodular adrenocortical disease?	PRKAR1A, PDE11A, PDE8B	1	PDE11A, PRKAR1A, PDE8B, PRKACA	0.66666 66667	PRKAR1A, PDE11A, PRKACB
Gene alias	What is the official gene symbol of GCS1?	MOGS	0.25	GCS1, ADAP1, MOGS, Mogs	0	GANAB
Gene alias	What is the official gene symbol of PTP-SL?	PTPRR	0.2	PTPRR, Ptprr, ptprr, ptprr.L, ptprr.S	0	PTPRS
Gene disease association	What are genes related to Type diabetes mellitus?	HNF1B, IL6, GPD2, HMGA1, IRS1, NEUROD1, IL6	0.142857 1429	HNF1A, IL6, ITPR3, PTPN22	0	INS, PTPN22, CTLA4, IL2RA, IL2RB, ERBB3
Gene alias	What is the official gene symbol of hCAP?	RNGTT	0.1	RNGTT, SCLT1, SMC3, DCD, NCAPD3, NCAPG, NCAPG2, NCAPD2, GEMIN4, NCAPH	0	
Gene alias	What is the official gene symbol of DCSCRIPT?	ZNF366	0.5	ZNF366, ZNF366.L	1	ZNF366
Gene alias	What is the official gene symbol of 15-LOX?	ALOX15	0.5	ALOX15, ALOX15B	1	ALOX15
Gene disease association	What are genes related to Hyperphenylalaninemia?	PTS, GCH1, QDPR, PCBD1,	0.166666 6667	PAH	1	PTS, GCH1, QDPR, PCBD1, DNAJC12, PAH

		DNAJC12, PAH			