Evidence Aggregator: AI reasoning applied to rare disease diagnostics

Hope Twede*[†], Ashley Mae Conard*, Lynn Pais, Samantha Bryen, Emily O'Heir, Greg Smith, Ron Paulsen, Christina A. Austin-Tse, Alex Bloemendal, Cas Simons, Scott Saponas, Miah Wander, Daniel G. MacArthur, Heidi Rehm

* co-first
[†] corresponding

Keywords: generative AI, rare disease, natural language processing, information retrieval, entity recognition, entity linking, evidence aggregation

Abstract

Retrieving, reviewing, and synthesizing technical information can be time-consuming and challenging, particularly when requiring specialized expertise, as is the case of variant assessment for rare disease diagnostics. To address this challenge, we developed the Evidence Aggregator (EvAgg), a generative AI tool designed for rare disease diagnosis that systematically extracts relevant information from the scientific literature for any human gene. EvAgg provides a thorough and current summary of observed genetic variants and their associated clinical features, enabling rapid synthesis of evidence concerning gene-disease relationships. EvAgg demonstrates strong benchmark performance, achieving 97% recall in identifying relevant papers, 92% recall in detecting instances of genetic variation within those papers, and ~80% accuracy in extracting individual case and variant-level content (e.g. zygosity, inheritance, variant type, and phenotype). Further, EvAgg complemented the process of manual literature review by identifying a substantial number of additional relevant pieces of information. When tested with analysts in rare disease case analysis, EvAgg reduced review time by 34% (p-value < 0.002) and increased the number of papers, variants, and cases evaluated per unit time. These savings have the potential to reduce diagnostic latency and increase solve rates for challenging rare disease cases.

Introduction

Information retrieval (IR) and *synthesis* are fundamental yet challenging tasks, involving the extraction and consolidation of relevant information from vast and often unstructured data sources (1). These tasks are particularly cumbersome when requiring expertise to sift through large literature knowledge bases (2). This challenge is relevant in rare disease (RD), where diagnosis rates are modest and time to diagnosis can be protracted due to the scarcity of well-established evidence for individual genetic variants as well as many gene-disease relationships (GDR) (3), (4).

RDs affect ~300 million people worldwide, creating a substantial collective burden of morbidity and mortality, and often have an underlying genetic cause that is challenging to identify (5) (6) (7). One of the major challenges in the care of individuals affected by RDs is the need to obtain a confident genetic diagnosis in order to provide access to reproductive testing, appropriate treatment and support, and access to clinical trials for new therapies (3) (5) (7) (8). The increasingly widespread use of diagnostic genome sequencing (GS) and exome sequencing (ES) presents the opportunity for variant analysts to have access to vast amounts of genetic information, but also presents substantial scaling challenges given tens of thousands to millions of variants are identified in each individual when typically only one or two variants are thought to be causing their disease (8).

Genomic analysis uses bioinformatic methods for filtering to a tractable set of variants for human review, but interpreting the potential role of the remaining variants in an individual's disease is a complex task that involves aggregating information from a variety of sources (9). Much of this process requires searching for and reviewing academic papers about the disease, genes, and variants in question, which can be time intensive and error prone, making this essential diligence a major bottleneck for RD diagnosis (9).

The objective of this literature review process is to gather multiple lines of evidence to support or rule out a potential causal variant (9) (10). This requires analysts to conduct their own searches and content review using tools such as PubMed (11) or access paid-subscription services, such as the Human Gene Mutation Database (HGMD) and Mastermind, which aggregate literature-mined variant data (12) (13). There are publicly funded or community organized curation efforts that provide freely accessible information on GDRs or variant pathogenicity (e.g., OMIM, ClinVar, ClinGen, GenCC) (14) (15) (16) (17), but none of these resources systematically collect all published literature on genes and variants making them incomplete for use in genomic curation. LitVar2 attempts to aggregate all variation from the published literature (18) though struggles with the lack of standardization of variant descriptions in the primary literature. And none of these resources collate information in ways that make it readily viewable to a genomic analyst performing case review (18).

In response to these challenges, researchers have investigated the possibility of applying natural language processing (NLP) techniques to assist in curation efforts and accelerate diagnostic workflows (12) (19) (20). Recently, large language models (LLM) such as GPT-4 have shown great potential for both IR and complex reasoning tasks (21) (22) (23) (24). Generative AI (GenAI) has already been applied in other biomedical settings (25) (26). AI-based approaches have the potential to make previously untenable ongoing literature curation efforts possible (24). As GenAI models such as GPT-4 are capable of sophisticated contextual reasoning, we hypothesize that LLMs can assist variant analysts in the lengthy task of determining what primary literature is relevant to their case (27) (28).

Here we introduce the Evidence Aggregator (EvAgg), a tool for scientific literature evidence aggregation about rare disease variants in any human gene (**Figure 1**). Starting with a query gene, EvAgg generates a reference resource that consolidates pertinent literature, extracting detailed information about observed variants and their phenotypic impacts. By greatly reducing the amount of time and effort spent manually searching the literature and compiling notes for each potential

GDR relevant to the phenotype of their case, EvAgg enables analysts to more thoroughly analyze cases, reanalyze unsolved cases, and prioritize other work.

We evaluated EvAgg's ability to perform this with a set of standardized benchmarks (see Results section, Benchmark performance) that compared the tool's outputs to a curated dataset of manually curated evidence from the scientific literature. In addition, we directly assessed the utility of EvAgg in our intended use case of rare disease genomic case analysis by conducting a user study of analysts' experience using EvAgg and its impact on their workflow.

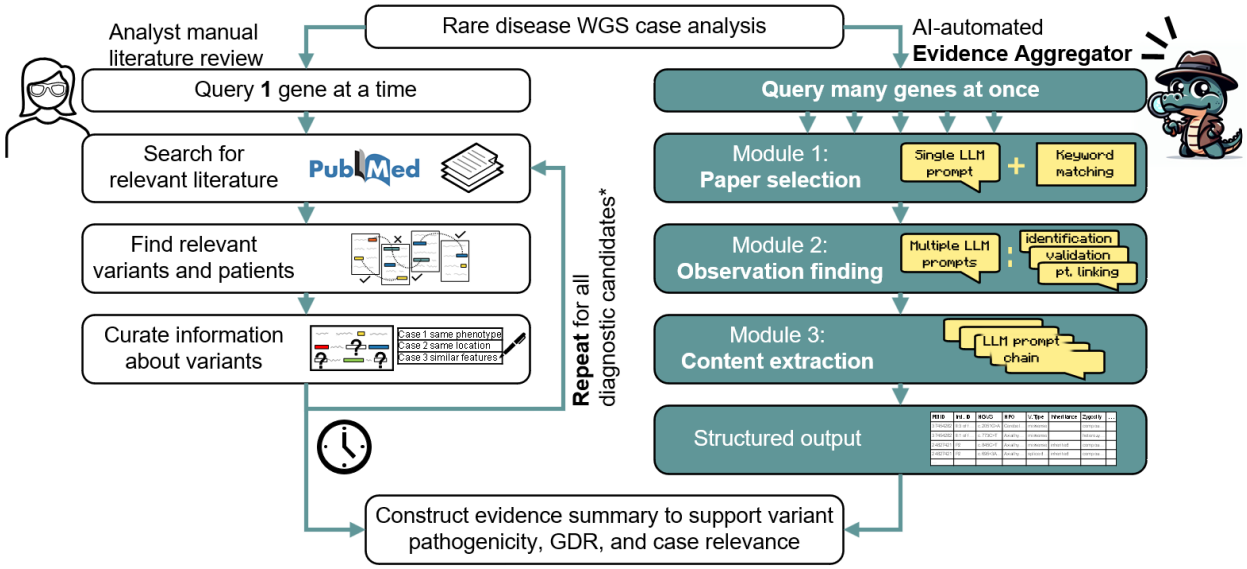


Figure 1: Overview of the Evidence Aggregator. *The Evidence Aggregator leverages generative artificial intelligence (AI) to perform analyst's literature review tasks, which are currently a manual process in their gene/variant analysis workflow. Analysts conduct literature review for one variant and gene at a time, curating gene-level phenotypes, variant locations, and pathogenicity mechanisms to support or refute variant pathogenicity and relevance to the case under review. *Manual analyses of genes and/or variants will range in quantity and rigor of the analysis dependent on preliminary assessment of the literature and degree of associated phenotype match. Evidence Aggregator character icon was AI-generated using DALLE·3. WGS, whole genome sequencing; GDR, gene-disease relationship; LLM, large language model.*

Methods

Pipeline overview and design criteria

The Evidence Aggregator (EvAgg) is an AI-based tool that assists variant analysts in the process of literature review associated with evaluation of pathogenicity of a given genomic variant. The pipeline is implemented as three independent modules: paper selection, observation finding, and content extraction.

First, EvAgg identifies relevant literature containing examples of observed human variants in the gene containing the variant in question (the *query gene*). Once the relevant publications have been identified, EvAgg finds all variants and individual observed cases within those publications, then extracts details about those observed cases that are relevant for the variant analysts' case review (**Figure 1**).

Development of EvAgg pipeline requirements and its evaluation process involved collaborations between researchers, software engineers, and rare disease diagnosticians, informed by an extensive needs-finding exercise (28). EvAgg was heavily optimized to favor recall over precision based on the determination that it was substantially more detrimental to omit relevant content than it was to include extraneous information.

The following sections discuss the implementation of EvAgg's submodules in more detail. See **Figure 2** and **Figure 3** for a graphical representation of the implementation and the **Supplementary Methods** and GitHub repository for additional detail.

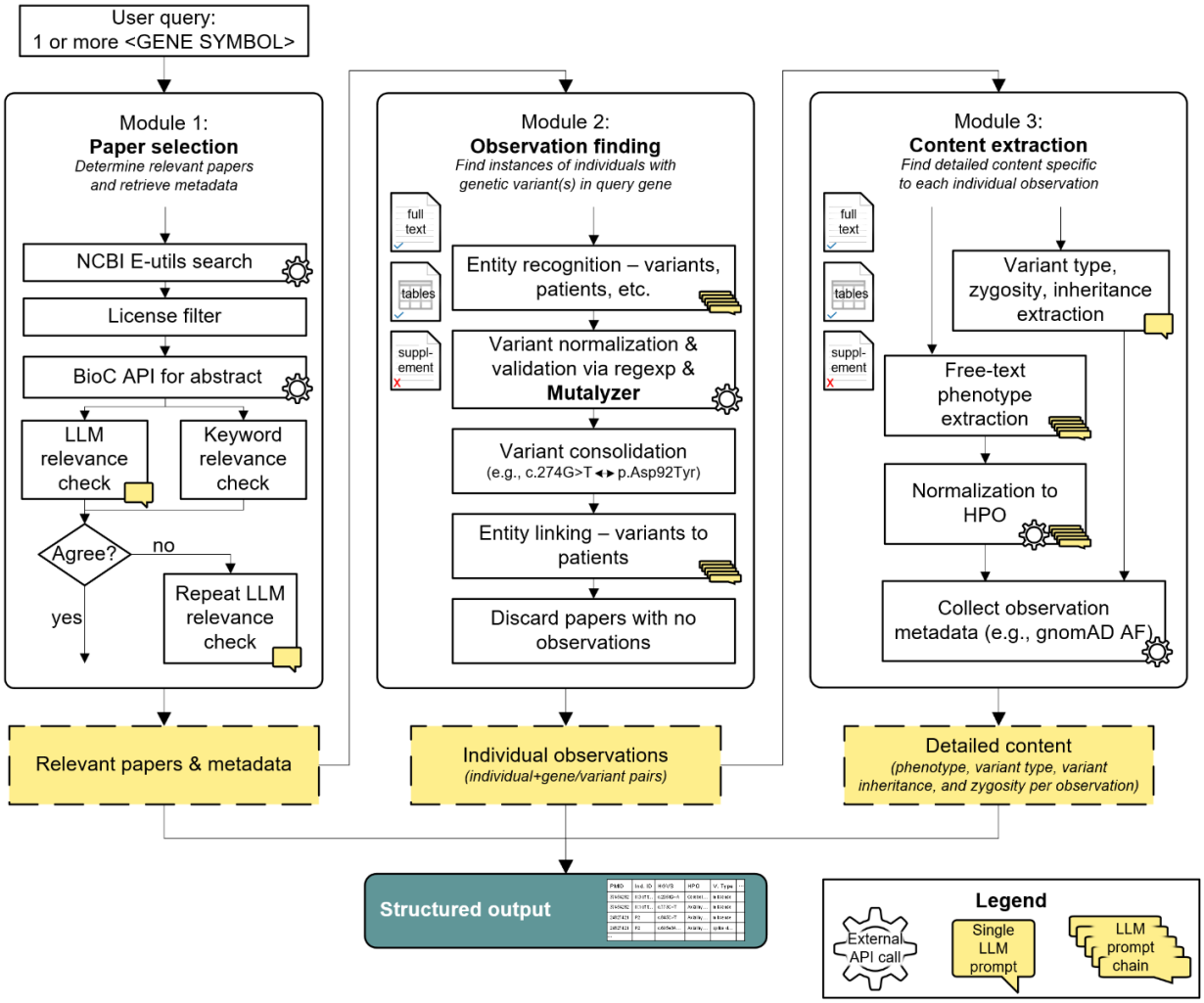
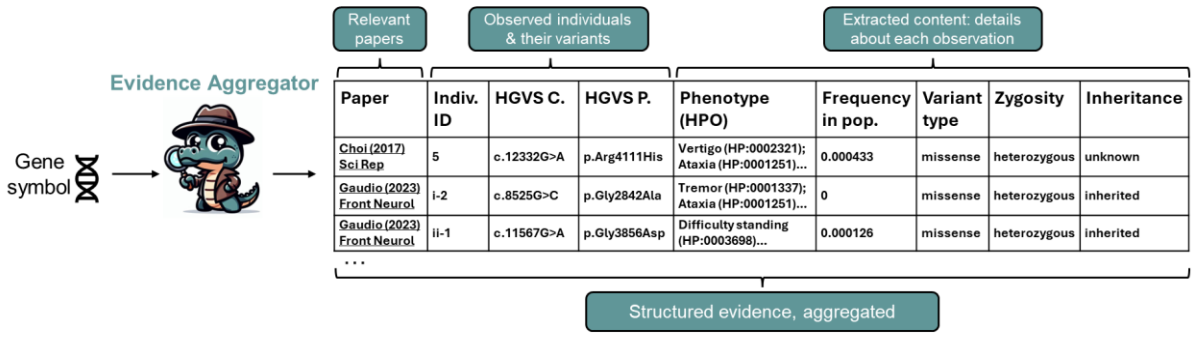


Figure 2: EvAgg schematic. *This figure depicts the processing steps that take place within each module of the EvAgg tool. The first module is responsible for identifying PubMed*

papers that are potentially diagnostically relevant for the gene of interest. The second module has the task of identifying observations of human genetic variation in those papers, where a unique observation is specified by a gene, HGVS variant description, and individual case identifier. The third module then extracts specific details about those individual observations, such as phenotypes and inheritance patterns. EvAgg uses a combination of internal logic, API calls to external web resources, and prompts or prompt chains issued to the LLM. Dashed-outline yellow boxes represent content included in structured output. NCBI, National Center for Biotechnology Information; LLM, large language model; regexp, regular expression (text pattern matching); HPO, Human Phenotype Ontology; AF, allele frequency; API, application programming interface.



Paper	Indiv. ID	HGVS C.	HGVS P.	Phenotype (HPO)	Frequency in pop.	Variant type	Zygosity	Inheritance
Choi (2017) Sci Rep	5	c.12332G>A	p.Arg4111His	Vertigo (HP:0002321); Ataxia (HP:0001251)...	0.000433	missense	heterozygous	unknown
Gaudio (2023) Front Neurol	i-2	c.8525G>C	p.Gly2842Ala	Tremor (HP:0001337); Ataxia (HP:0001251)...	0	missense	heterozygous	inherited
Gaudio (2023) Front Neurol	ii-1	c.11567G>A	p.Gly3856Asp	Difficulty standing (HP:0003698)...	0.000126	missense	heterozygous	inherited
...								

Figure 3: Evidence Aggregator (EvAgg) structured output example. *EvAgg processes user input gene symbol to produce a structured output table: relevant papers are outputs of Module 1 (Paper Selection), observed individuals & their variants are outputs of Module 2 (Observation Finding), and extracted content (details about each observation) are outputs of Module 3 (Content Extraction). Each table row represents an individual observation: a reported case of an individual with a variant in the query gene, as described in the source publication for that observation. Evidence Aggregator character icon was AI-generated using DALLE·3. Indiv. ID, individual identifier; HGVS C., Human Genome Variation Society variant description in terms of the coding DNA region of the gene; HGVS P., Human Genome Variation Society variant description in terms of the protein-coding amino acids of the gene; HPO, Human Phenotype Ontology; Frequency in pop., population frequency as reported in gnomad v4.*

Paper selection

EvAgg retrieves candidate publications for each query gene from PubMed using REST API calls to the NCBI E-utilities service (29). The query string provided to the API is "`<GENE SYMBOL> pubmed pmc open access[filter]`" which restricts returned publications to only those included in the PubMed Central Open Access (PMC-OA) dataset. Candidate publications are further pre-filtered based on license type, including only licenses that permit data mining and derivative use. All subsequent operations are performed on the full text of candidate publications which EvAgg obtains via the BioC API (30).

EvAgg determines publication relevance by leveraging a consensus approach using two methods: keyword matching and LLM-based classification. Keyword matching involves searching each paper's abstract and title for inclusion and exclusion keywords to retain papers containing variant observations relevant to rare disease (e.g. "mendelian", "monogenic", "rare variant", "variant of uncertain significance") and filtering out papers describing non-monogenic disease and somatic cancer. LLM-based classification leverages a single prompt to predict relevance based on the paper's abstract and title with four few-shot positive and negative examples (see Supplemental Methods). If the two methods yield different results, EvAgg runs the LLM-based classifier a second time to break the tie.

Observation finding

Once EvAgg identifies the candidate publications for each query gene, the next step is to identify all observations of human genetic variation associated with the query gene within those papers.

The observation finding module is implemented via a series of individual prompts, leveraging a task decomposition approach that has been shown to increase LLM performance on complex tasks (31). First, a series of prompts are used to identify candidate entities of interest; including the variants themselves, variant metadata such as reference human genome build (32) and gene transcript identifiers, and individual case identifiers. Second, string representations of variants are normalized to Human Genome Variation Society (HGVS) nomenclature (32), validated, and then consolidated, such that logically linked variant entities (for example, a coding DNA variant description and its associated protein consequence) are treated as single entities. This challenging series of sub-tasks, given the diversity of representations of genetic variants found throughout the literature (33), is addressed using a series of heuristic filters and regular expressions to normalize the nomenclature variability, subsequently leveraging Mutalyzer v3 (34) for validation (HGVS description syntax and biological validity) and consequence prediction. Finally, EvAgg links variants to cases, creating individual observations.

EvAgg discards selected papers with no found observations as presumed false positives from the paper finding module.

Content extraction

We identified the categories of content to extract from each observation from a needs-finding exercise with variant analysts (28). From the source text, EvAgg uses individual prompts or prompt chains (combination of multiple prompts) to extract information about each observation's phenotype (mapped to HPO terms), zygosity, variant type, and inheritance pattern (see **Supplemental table 1** for value lists). In addition to LLM-derived content categories, EvAgg also references gnomAD allele frequencies (v4.1) (35) and publication metadata.

Presentation of pipeline outputs

EvAgg is a standalone tool with a minimal interface. The user provides one or more query genes with optional restrictions and EvAgg produces a structured dataset with relevant observations of genetic variation from source publications.

To facilitate use of EvAgg's outputs by variant analysts, we modified the open-source *seqr* genome analysis platform (36) by adding a minimal user interface that integrates EvAgg's output into a typical *seqr* variant search workflow. The additional UI presents aggregated evidence about a query gene alongside case-specific information.

Prompt implementation detail

When issuing prompts to the LLM, EvAgg provides the entirety of the text of each paper to the model. In addition, when searching for variants, individuals, or phenotypes, EvAgg will also run the associated prompts separately on any tables within a paper, as these are often enriched for the entities of interest. EvAgg retains the union of results from applying these prompts to the full text and the individual tables.

Due to this design decision, we restricted model selection to Azure OpenAI (AOAI) Service models supporting context sizes of 128k+ tokens available at the time of development and evaluation: GPT-4-Turbo, GPT-4o, and GPT-4o-mini.

Benchmarking curated dataset generation

Evaluation of EvAgg performance was complicated, as a structured reference fit for the task did not exist, and benchmarking of paper selection against up-to-date curation services like HGMD prohibit this type of performance comparison in their use terms. In turn, we generated a curated dataset to which pipeline outputs could be compared (**Figure 4**). This task was performed by a subset of the paper authors (AC, HT, LP, SB, EO, RP, MW; "curation team"). The process of generating and verifying the curated dataset was time intensive and proved to be an invaluable resource for the development and validation of EvAgg, so we have made this curated dataset publicly available with the hopes that others in the community will benefit from it and contribute to subsequent versions (see Data availability).

Currently, the primary objective of EvAgg is to facilitate information retrieval from the literature for genes without well-established GDR, so we used the following process to identify the genes for which truth data were generated. First, we identified 1900 genes submitted by ClinGen to GenCC (15) (14). Then we took the subset of those genes where the most recent ClinGen-submitted GDR was "Moderate", "Limited", or "No known GDR" (14) (15). Finally, we randomly selected 15 genes from each of these three categories, leading to a final set of 45 genes.

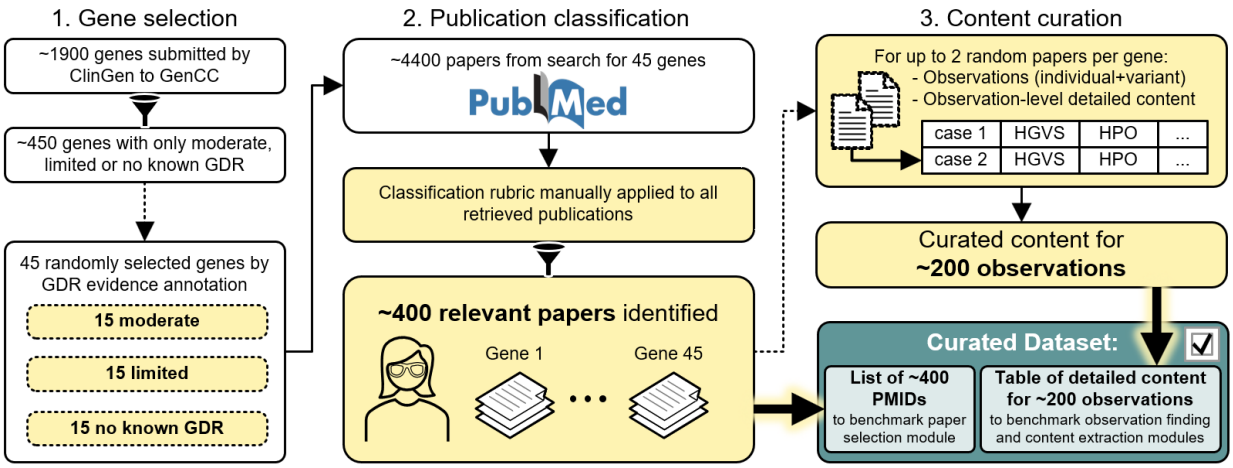


Figure 4: Benchmark curated dataset generation process. *1) Forty-five genes were randomly selected from the subset of ClinGen submissions to GenCC with only moderate, limited, or no known gene-disease relationship (GDR). 2) Approximately 4400 papers were returned by a PubMed search for these genes and manually assessed for diagnostic relevance using a standardized rubric, with ~400 papers deemed relevant (Supplementary methods). 3) From these ~400 relevant papers, we selected up to 2 papers for each of the 45 genes and curated detailed content for observations of individuals with genetic variants in the gene of interest. The curated dataset, consisting of the complete list of ~400 relevant papers and curated content for ~200 observations, was used for subsequent benchmarking. Yellow boxes represent curation team actions. Funnel-shaped connectors represent filtration steps, dotted line arrows represent random sampling steps, dotted outline shapes represent random samples, bold arrows with yellow highlight represent inclusion of information in curated dataset. GDR, gene-disease relationship; HGVS, human genome variation society; HPO, human phenotype ontology.*

For each of the 45 genes, our curation team performed a deep review of all papers returned by PubMed search for each gene, focusing on publications that appear after the GDR curation was approved by a ClinGen expert panel. Curators selected relevant papers based on a rubric they designed to reflect variant analysts' approaches to identifying publications containing observations of human genetic variation that support or refute variant pathogenicity while analyzing a rare disease case (**Supplementary methods, Curated dataset generation rubric**).

We randomly selected at most two relevant papers for each gene from the PMC-OA subset with licenses that permitted derivative use. For each of the selected papers, curators extracted individual observation content: publication information, individual case identifiers, variant cDNA and protein HGVS annotation, variant type, variant genome build (e.g. hg19), gene transcript, zygosity, inheritance, phenotype text description, phenotypes encoded as Human Phenotype Ontology (HPO) terms (37), study type, and functional study information when available.

Once curators constructed the curated dataset, we randomly split the curated dataset into a development (dev) and evaluation (eval) set on a per gene basis, with 70% of the genes in the dev

set and the remainder in the eval set. The dev set was used for iterative development of EvAgg while the eval set was held in reserve for benchmarking after development was completed.

Error analysis

To assess pipeline performance, identify consistent error modes, and determine if EvAgg was able to identify additional relevant information in the literature, we conducted an in-depth review of putative errors. This also provided an opportunity to refine the curated dataset, as manual curation could introduce human error due to fatigue effects. Our approach to performing this review was to focus on consistent disagreements (i.e., those observed in most pipeline executions) between the curated dataset and EvAgg, obtain multiple independent assessments of each error, and minimize bias based on whether EvAgg or the curated dataset curation team was the source of any given piece of information.

After executing the five replicates of benchmark pipeline runs, we compiled consistent disagreements between pipeline outputs and the curated dataset. We formatted these putative errors into true-false or multiple-choice questions. Each question was randomly assigned to be reviewed by two expert variant analysts (manuscript authors LP, SB, and EO). Reviewers were blinded as to which of the candidate answers to each question was provided by EvAgg and which was provided by the curated dataset. Reviewers independently answered all questions assigned to them, adhering to the curated dataset generation rubric to ensure consistency. When reviewer responses to questions differed, the larger curation team reviewed all questions until all errors were resolved.

Following error analysis completion, warranted updates were made to a revised version of the curated dataset.

Pipeline benchmarking

We benchmarked three aspects of the EvAgg pipeline: selecting relevant papers, finding observations of human genetic variation within those papers, and extracting specific details about those observations (e.g. individual's phenotype).

For paper selection, to determine if EvAgg could find all the papers we were looking for, we calculated *recall* (the proportion of papers within the curated dataset that EvAgg successfully retrieved) and to determine EvAgg's ability to filter out irrelevant papers we calculated *precision* (the proportion of papers returned by EvAgg that were included in the curated dataset).

Like for paper selection, observation finding recall and precision were calculated in comparison to curated dataset observations. Each unique observation was identified by an individual+variant pair. To better understand whether missing or extraneous observations could be attributed to entity recognition (i.e., finding the correct variants within a paper) or entity linking (correctly associating those variants with individuals), we separately assessed precision and recall of variant finding using the same approach while treating all individual case identifiers as equivalent.

Finally, for all content categories except phenotype (inheritance, variant type, zygosity), we assessed accuracy by comparing the proportion of curated dataset content that exactly matched the values produced by EvAgg. Direct comparison of pipeline outputs to the curated dataset for

observed phenotypes was challenging as curated dataset values were manual transformations of free text phenotypes into specific HPO terms, a subjective task that itself has been the subject of much study (38) (39). Recognizing that the primary use of phenotypes in a rare disease diagnostic context is to assess affected organ/systems, we benchmarked phenotypes based on agreement of broader phenotypic categories.

All benchmark runs were conducted five times each and reported performance metrics include the mean and standard deviation across those five runs. To better understand the impact of LLM variability, we evaluated the between-run consistency for paper finding.

Benchmarking was performed repeatedly on the dev set during implementation to assess common error modes in the pipeline and provide the opportunity to address them. Benchmarking on the eval set was performed only once after completing development.

User study design and analysis

We designed and conducted a user study to assess utility and user experience of incorporating EvAgg into rare disease case review. Study participants were individuals from collaborating institutions who have been previously trained to perform rare disease case analysis within the *seqr* platform.

To control for time-of-day and fatigue effects, we split the protocol into two sessions with the same start time on different days. Study sessions were recorded and conducted virtually with screen-sharing. During each study session, the study administrator (HT) observed a single participant performing rare disease case analysis for one hour, taking notes including timestamped participant actions. The first study session served as a baseline for user experience without the use of EvAgg. EvAgg outputs were available during the second study session.

To minimize the impact of differences between participants and the difficulty of individual cases, we created fixed-order, approximately difficulty-matched sets of four unsolved cases selected from the Broad Institute's Rare Genomes Project (raregenomes.org). Participants were randomized to one of the case sets on the first day, analyzing the reciprocal set on the second day (**Figure 5**).

In Likert-scale surveys following use of EvAgg, we solicited comparison of their experience during case analysis with/without the EvAgg outputs as well as their sentiment and trust towards EvAgg. Semi-structured interview discussions included feedback about specific features contributing to sentiment and trust, importance of accuracy, characterization of how EvAgg may impact their work, and features which would improve their user experience. Survey and semi-structured interview feedback was analyzed jointly to identify key themes.

We observed workflow changes from time-stamped user action recording as well as timestamped artifacts generated by *seqr*. We used a mixed-effects model to analyze the impact of EvAgg on individual behavioral metrics, accounting for both fixed effects (study session and case set) and random effects (individual participants).

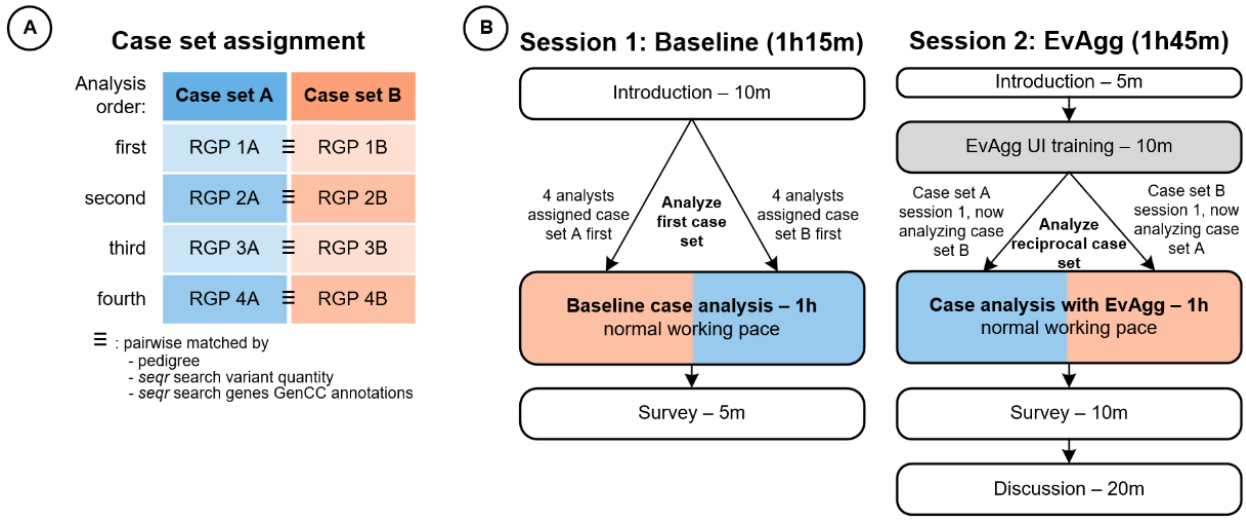


Figure 5: User study design. *(A) User study participants analyzed unsolved cases from the rare genomes project (RGP) over the course of two days. Cases in each set were presented in a fixed order, with individual cases being pairwise difficulty-matched (i.e. RGP case 1A was of similar difficulty to RGP case 1B). Participants were randomly assigned one of the two case sets for session 1 and analyzed the reciprocal case set for session 2. (B) The session 1 procedure involved 1 hour of standard procedures for case analysis and concluded with a short survey of the experience. The procedure for session 2 began with training on the use of EvAgg followed by 1 hour of case analysis with access to the tool. The day concluded with a survey and participant interview. During each hour of case analysis, analysts were asked to work at their normal pace and were not expected to complete all cases within the hour, as the intention was to measure effects of EvAgg on their workflow, not to evaluate individual analyst's "performance". RGP, Rare Genomes Project; ≡, represents pairwise-matched RGP cases approximately matched by affected status of family members in pedigree, quantity of variants returned in seqr variant search, and ClinGen/GenCC annotation of genes returned in seqr search.*

Responsible AI practices and institutional review

EvAgg engineering and open-source release underwent Microsoft subject matter expert review for Responsible AI considerations. Transparency documentation was developed to highlight the key components for users: permitted uses, user responsibility, limitations, and data attributions. This documentation can be found in the EvAgg README on GitHub (see Data Availability).

Performance was carefully designed to align closely to end-use, with the application-specific, expert-curated curated dataset evaluated for performance. User study design underwent subject matter expert review and IRB review (OHRP-IRB00009672) at Microsoft.

Data and code availability

The Evidence Aggregator (EvAgg) code is fully available and documented in an open source repository on GitHub at https://github.com/microsoft/healthfutures-evagg. All source data,

including the manually generated ground truth, the quantitative data from the user study, and the results of the error analysis are available in the same repository. All scripts necessary to regenerate the quantitative findings in this manuscript are available in the same repository.

Results

We evaluated the performance of Evidence Aggregator (EvAgg) against the benchmark of human curators performing manual information retrieval and in the context of rare disease case analysis, the primary intended use case.

Curated dataset characteristics

The manually generated curated dataset was sourced from PubMed search results containing 4401 papers of potential interest. The depth of literature concerning any given gene varied widely; the average number of papers per gene requiring review was 112.85 ± 92.65 (mean ± std throughout). Of these, 10.05 ± 15.11 per gene were deemed relevant given our rubric (392 total). This suggests that during manual review of literature, only approximately 1 in 11 papers read is likely to contain variant-informed evidence relevant to the task of diagnosis. Of the 392 relevant papers, 133 were annotated as "in-scope" for benchmark analyses based on their presence in the PubMed Open Access subset and meeting licensing requirements.

From the set of 133 in-scope papers, we identified observations in 41 papers, representing 24 of the 45 genes of interest. This process resulted in manual curation of relevant information about 224 unique observations representing 167 different genetic variants. There were on average 9.33 ± 12.91 observations (6.96 ± 10.09 variants) per gene and 5.60 ± 9.53 observations (4.28 ± 8.13) per paper.

Error analysis

While the curation team was rigorous in development of the curated dataset, this task was specialized for comparison with EvAgg outputs and not an exact match with the typical literature review workflow during case analysis. This complex task required intense, often lengthy, periods of undivided attention and strict adherence to the rubric, potentially introducing fatigue and natural human error. Evaluation of putative errors was conducted to both gauge true performance of EvAgg, as well as refine the curated dataset.

Across the three subtasks (paper selection, observation finding, and content extraction), despite overall excellent performance, there were a substantial number of consistent disagreements between the initial version of the curated dataset and pipeline outputs. There were 470 points on which the curated dataset and pipeline outputs disagreed half the time or more, with 74 (15.7%) attributable to paper selection, 81 (17.23%) to observation finding, and 226 (48.1%) to phenotype extraction from 52 distinct observations.

Two curators independently provided the same disagreement resolution in 188 (61%) of cases. Manual review of 470 disagreements showed that for 235 of them (50%) reviewers ultimately deemed the EvAgg outputs as more relevant/correct than the original curated dataset. These findings provide an interesting picture of the complexity of the task of manual targeted information retrieval from the literature, and the potential for AI-based IR to greatly assist in this process.

See **Figure 6** and **Supplemental table 5** for additional detail on the error analysis results.

Figure 6: Breakdown of error analysis shows EvAgg errors and identifies additional information for inclusion in the curated dataset. *Each subplot shows the proportion of each putative error type that was adjudicated as either a verified error by EvAgg (orange, EvAgg errors) or required an update to the curated dataset (blue, EvAgg additions to curated dataset). (A) Breakdown of the paper selection task putative errors with majority of both extra papers (papers selected by EvAgg that were not part of the original curated dataset) and missing papers (papers from the curated dataset that EvAgg didn't select as relevant) resulting in curated dataset updates after review. (B) Depicts the same for observation finding putative errors and shows that while EvAgg ultimately correctly identified a large number of observations that were overlooked in the curated dataset, the majority of observations that EvAgg omitted were indeed errors. (C) Shows the error analysis for content extraction, with the vast majority of content extraction errors originating in the assessment of phenotype with a fairly even distribution between errors that let to curated dataset updates and those that did not.*

Benchmark performance

The following results discuss performance using the evaluation (eval) subset after curated dataset revision. Full benchmark analyses (dev and eval, pre- and post-error analysis with curated dataset revision) are available in the **Supplementary results**. Additionally, a performance comparison of GPT-4-Turbo and other candidate LLMs is included in the **Supplementary results**.

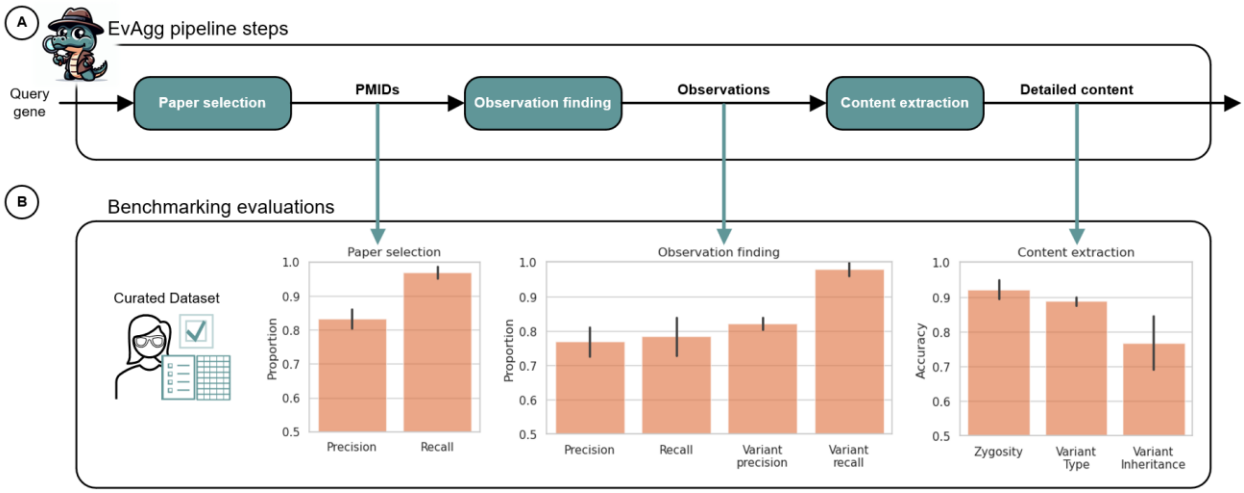


Figure 7: Benchmarking overview and results. *(A) Individual steps of the EvAgg pipeline were separately evaluated against the curated dataset to assess tool performance in different aspects of the task. (B) Task performance when benchmarked against the evaluation subset from the curated dataset: precision and recall for the paper selection and observation finding tasks, accuracy (i.e., the frequency of matching curated dataset content) for the content extraction task. Evidence Aggregator character icon was AI-generated using DALLE·3.*

Selecting relevant papers

EvAgg repeatedly selected similar sets of papers with a given paper appearing 89.3% of the time, indicating stable performance despite expected noisiness of the underlying LLM generations.

EvAgg excelled at retrieving all the papers flagged as relevant in the curated dataset, with an average recall of 0.97 ± 0.017 (**Figure 7b**). This constitutes an average of 1.6 false negatives (of 53 total papers) per run.

While optimizing recall was the primary focus of the development process, precision remained suitable for the intended application, with an average of 0.832 ± 0.28, corresponding to an average of 10.4 false positives (of 53 average papers retrieved) per run.

Finding observations of genetic variation

For each of the papers in the curated dataset with associated observations, we evaluated EvAgg's ability to specifically identify those observations. The average recall for this task was 0.783 ± 0.055. In comparison to the recall of 0.950 ± 0.009 for the slightly different task of identifying relevant genetic variants (irrespective of the associated individual), this degraded performance on observation finding is due to failures in either correctly identifying specific individuals or correctly linking those individuals with the genetic variant they possess.

As was the case in paper selection, precision on the observation finding task was slightly lower than recall, averaging 0.767 ± 0.042. Precision on variant finding was unsurprisingly higher at 0.821 ± 0.018.

Extracting detail about observations

The average accuracy in determining zygosity was 0.922 ± 0.027, for variant type it was 0.888 ± 0.012, and for variant inheritance it was 0.768 ± 0.076 (**Figure 7b)**.

The proportion of observations with exact concordance of generalized phenotypes between the curated dataset and pipeline was 0.862 ± 0.047. We observe that on average, the number of phenotypic terms reported by EvAgg exceeded the number of terms provided in the curated dataset (**Figure 8b**): 2.68 ± 3.13 in truth vs 3.91 ± 4.02 in EvAgg output, which corresponds to the relatively high recall for phenotype extraction (0.961 ± 0.025) when compared to precision (0.892 ± 0.033). See **Figure 8c** for detail. Overall, this suggests that EvAgg is extremely capable of extracting all the phenotypes described in the curated dataset, and errors primarily involve inclusion of extraneous terms.

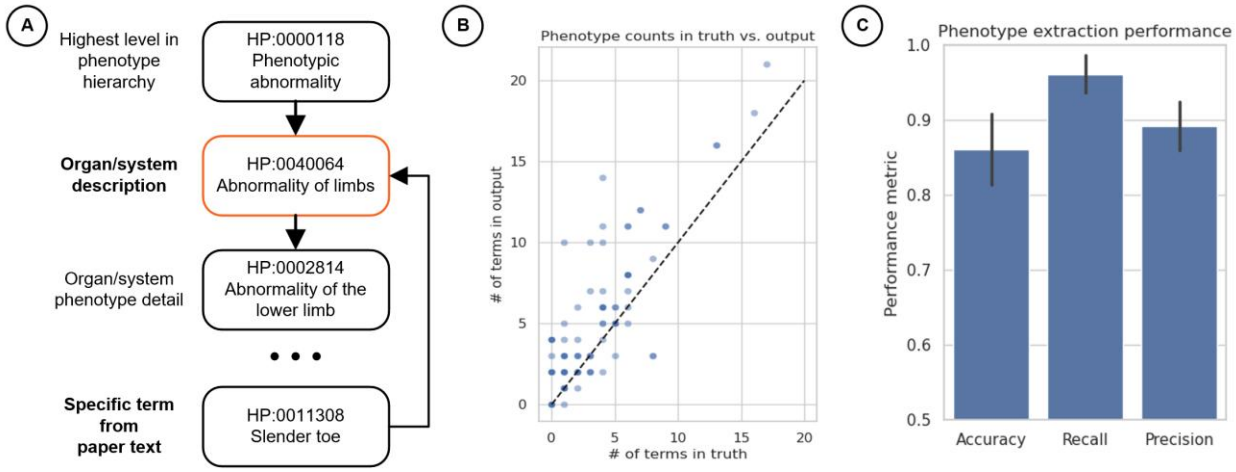


Figure 8: Phenotype benchmarking methodology and results. *(A) Depiction of the phenotype generalization approach. Specific HPO terms derived from paper text were generalized to corresponding organ/system description terms in the HPO (generally a child of parent term "HP:0000118 – Phenotypic abnormality") before making comparisons. (B) This scatter plot shows that EvAgg consistently identifies more phenotypic terms per observation than are found in the curated dataset. (C) Bar plot separately showing the accuracy, recall, and precision of phenotype extraction. Accuracy is defined as the proportion of observations with a perfect match of generalized organ/system description terms, recall defined as the proportion of observations where all curated dataset organ/system description terms were included in EvAgg's output, and precision defined as the proportion of observations where there were no extraneous terms included in EvAgg's output. Relatively high recall as compared to precision is concordant with the consistent overidentification of HPO terms shown in subplot B.*

User study in rare disease case analysis

The study included eight highly educated participants with 1-10 years of experience in rare disease genome case analysis. Two study participants were authors (LP, SB) involved in curated dataset curation and error analysis but were not involved in engineering, benchmarking, or user study planning and administration.

Workflow changes

We collected detailed notes and compared objective measures of the workflow: number of EvAgg interactions, number of cases/variants reviewed, number of publications read (opened full-text and spent time to review), and time spent per case/variant. Participants were able to review cases 34% faster when they had EvAgg available (27.2 ± 11.3 vs 17.9 ± 4.8 min/case, p<0.002) (**Figure 9**). Analysts reviewed substantially more papers with EvAgg (2.5 ± 2.2 vs 6.0 ± 4.8, p<0.02) despite no significant change in the number of manual publication searches conducted (p=0.39).

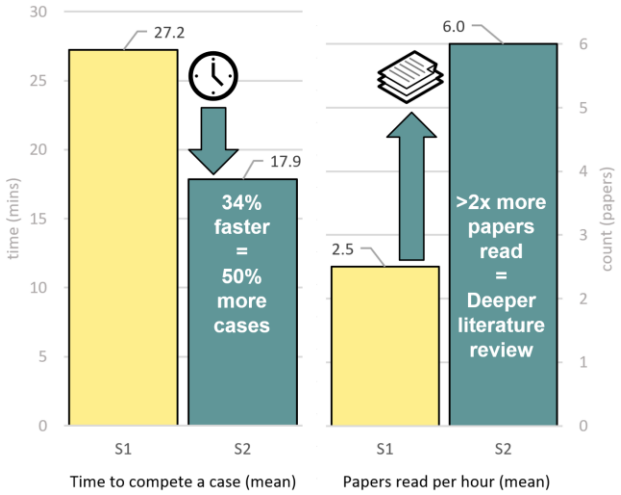


Figure 9: Workflow changes observed with EvAgg use. *Changes measured from user study in terms of time per completed case analysis and total number of publications read during 1 hour of case analysis. See* ***Supplemental figure 3*** *for workflow changes stratified by case set and participant for additional detail. S1, measure from user study session 1 (baseline without EvAgg, yellow bar); S2, measure from user study session 2 (case analysis with EvAgg available, green bar).*

User experience

All participants expressed positive sentiment and desire to use EvAgg frequently, thought it was easy to use, and believed they could learn to use EvAgg very quickly (**Figure 10**, **Supplemental figure 2**). When asked for the impact of using EvAgg "tomorrow", one participant stated *"even if tomorrow is Saturday I would use it Saturday as well if it would save me time".* Key likes described by most participants were: time savings, workflow alterations such as phenotype-based screening, and table features like links to publications. The top feature request was for EvAgg availability for all/more genes.

The majority of participants felt EvAgg made cases less difficult. All participants perceived time savings, with the caveat that "rabbit holing" and inexperience with the tool could lessen actual time savings. Most participants imagined they would use the saved time to dig in deeper into cases while half described altering their workflows to complete other tasks. EvAgg was predominantly viewed as an early resource for gathering information.

Trust

Participants indicated mixed trust, with more than 50% rating their trust positively (**Figure 10**). Participants expressed that their trust would be increased once they had a better understanding of how EvAgg works and its limitations. While accuracy was generally considered important due to the nature of the task ("*I think [accuracy is] very important because it's patients that we're assessing"*), accuracy of phenotype and variant location/change were considered the most important aspects of accuracy. Every participant indicated tolerance for inaccuracy, with most reasoning that they would double check all information and expect some amount of error ("*I'm not concerned if there is something not accurate because I'm not taking the information in the table blindly*"). Notably, trust did not seem to be correlated with sentiment, as all participants expressed positive sentiment and desire to immediately start using EvAgg.

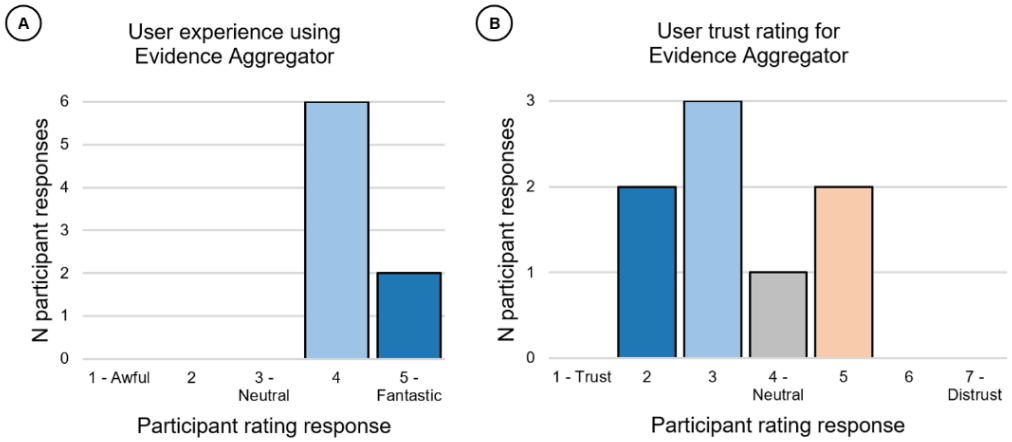


Figure 10: User sentiment and trust toward EvAgg as reported by participants. *A) User sentiment from survey data reported on a 5-point Likert scale: scores ranged from 1 (strongly dislike: "Awful") to 5 (strongly like: "Fantastic"). B) User trust from survey data reported on a 7-point Likert scale: scores ranged from 1 ("completely trust") to 7 ("completely distrust").*

Discussion

Our study demonstrates the potential of generative AI (GenAI) to significantly improve information retrieval and reasoning in rare disease diagnostics, reducing the time and effort required for analysts. The Evidence Aggregator (EvAgg) is effective in retrieving and reasoning over relevant literature, as indicated with high sensitivity, specificity, and accuracy on benchmark evaluations.

Notably, our results show over 95% recall when selecting relevant rare disease papers and variants within those papers, with high performance linking variants and individual cases as well as extracting relevant content from those papers. Evaluation of putative errors revealed that EvAgg was able to identify additional relevant papers, observations, and content, demonstrating the applicability of GenAI as an assistant in complex domain-specific curation efforts.

Our user study highlights EvAgg's utility for experienced analysts to more efficiently analyze cases. Beyond our objective findings, users expressed positive sentiment and wanted a tool like EvAgg to be integrated into their workflows "tomorrow", perceiving case analysis as easier with the number one request being more genes with EvAgg-derived evidence. We observed that analysts were able to complete more cases in the allotted time, review more variants, and read more papers within each case analysis.

User study subjects indicated that understanding how EvAgg works, including performance characteristics and error modes, is important for increasing trust. The rigorous task-specific curated dataset generation and error analysis ensure performance metrics of EvAgg are accurate and relevant, which contributes to user trust. We have made our benchmarking curated dataset available in our GitHub repository and welcome feedback and contributions from the community.

GenAI has the potential to democratize access to summarized information, making it more accessible to a wider audience. This could significantly impact the field by providing timely and relevant information to researchers and clinicians. Application of GenAI tools like EvAgg is advantageous in contrast to resource-intensive manual curation and bespoke natural language processing modeling techniques which require vast amounts of training data and compute resources.

Limitations

EvAgg's performance is limited by the underlying model. Future models may outperform GPT-4-Turbo, which we used in most benchmark analyses. The cost of using advanced AI models may be prohibitive in resource-limited settings. Strategies like text chunking and few-shot examples could enhance future pipeline versions. Currently, EvAgg processes inputs and outputs only in English.

The curated dataset development was rigorous, but the data set is small. This underscores the need for shared curated datasets and benchmarks for robust evaluation and comparison using more analysts and more cases.

While we were able to observe workflow changes in our user study, a larger longitudinal study is needed to quantify the effect on user experience during case analysis with EvAgg assistance.

The information returned by EvAgg is not exhaustive. Cost and performance for strong/definitive GDR genes with extensive literature was not evaluated. We did not explore processing abstracts or user input files/credentials for closed-access papers. Increasing open access and suitable license terms is crucial to facilitate improvements in the capability of GenAI and other computational approaches to assist in synthesizing scientific knowledge.

Summary

In the context of rare disease diagnosis where the burden to continuously review academic literature is extremely high and the time availability to do so is low, automated tools to facilitate this process have the potential to significantly impact clinical outcomes. We have demonstrated that GenAI-based approaches like EvAgg have the potential to alleviate some of this time burden and also identify overlooked relevant information. Additionally, when employed in a representative diagnostic setting, access to EvAgg appears to significantly accelerate case analysis. We have made EvAgg and the curated curated dataset data used for benchmarking publicly available with the hopes that the community will carry out their own evaluations, improve the tool's capabilities, and ideally leverage it in their own workflows.

Author contributions

All authors were involved in the development of the research proposal. AMC, HT, LP, SB, EO, RP, and MW created the curated dataset and performed error analysis comparing the curated dataset to pipeline outputs. AMC, MW, and GS developed the EvAgg software package. HT designed and executed the user study. AMC, HT, and MW performed data analysis. HT, RP, AMC, and MW conducted responsible AI review. All authors were involved in the authoring of the manuscript.

Competing interests

The authors declare the following competing interests: HT, AMC, GS, RP, SS, and MW are all full-time employees of Microsoft. HR, CAA-T, LP, EO, AB, CS, SB, and DM receive research funding from Microsoft.

References

1. *Structured information extraction from scientific text with large language models.* **John Dagdelen, Alexander Dunn, Sanghoon Lee, et al.** 2024, Nat Commun, Vol. 15, p. 1418.

2. *Text-mining and information-retrieval services for molecular biology.* **Martin Krallinger & Alfonso Valencia.** 6, 2005, Genome Biol, Vol. 224.

3. *Time to make rare disease diagnosis accessible to all.* **Heidi L Rehm.** 2, s.l. : Nature Publishing Group US New York, 2022, Nature Medicine, Vol. 28, pp. 241-242.

4. *Solving the unsolved rare diseases in Europe.* **Holm Graessner, Birte Zurek, Alexander Hoischen & Sergi Beltran.** 9, s.l. : Springer International Publishing Cham, 2021, European Journal of Human Genetics, Vol. 29, pp. 1319-1320.

5. *Rare-disease genetics in the era of next-generation sequencing: discovery to translation.* **Kym M Boycott, Megan R Vanstone, Dennis E Bulman & Alex E MacKenzie.** 2013, Nature Reviews Genetics, Vol. 14, pp. pages681–691.

6. *The landscape for rare diseases in 2024.* **The Lancet Global Health.** 2024, The Lancet Global Health, p. e341.

7. *The impact of clinical genome sequencing in a global population with suspected rare genetic disease.* **Erin Thorpe, Taylor Williams, Chad Shaw, et al.** 2024, Am J Hum Genet., pp. 1271-1281.

8. *Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis.* **Lisa J Ewans, Andre E Minoche, Deborah Schofield, et al.** 2022, Eur J Hum Genet., pp. 1121-1131.

9. *Best practices for the interpretation and reporting of clinical whole genome sequencing.* **Christina A Austin-Tse, Vaidehi Jobanputra, Denise L Perry, et al.** 1, s.l. : Nature Publishing Group UK London, 2022, NPJ genomic medicine, Vol. 7, p. 27.

10. *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* **Sue Richards, Nazneen Aziz, Sherri Bale, et al.** 5, 2015, Genet Med., Vol. 17, pp. 405-424.

11. *PubMed: the bibliographic database.* **Kathi Canese & Sarah Weis.** 1, s.l. : {National Center for Biotechnology Information (US) Bethesda, 2013, The NCBI handbook, Vol. 2.

12. *The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies.* **Peter D Stenson, Matthew Mort, Edward V Ball, et al.** 6, s.l. : Springer, 2017, Hum Genet., Vol. 136, pp. 665-677.

13. *Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation.* **Lauren M Chunn, Diane C Nefcy, Rachel W Scouten, et al.** 2020, Front. Genet.

14. *The gene curation coalition: a global effort to harmonize gene--disease evidence resources.* **Marina T DiStefano, Scott Goehringer, Lawrence Babb, et al.** 8, s.l. : Elsevier, 2022, Genetics in Medicine, Vol. 24, pp. 1732-1742.

15. *ClinGen—the clinical genome resource.* **Heidi L Rehm, Jonathan S Berg, Lisa D Brooks, et al.** 23, s.l. : Mass Medical Soc, 2015, New England Journal of Medicine, Vol. 372, pp. 2235-2242.

16. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* **Ada Hamosh, Alan F Scott, Joanna S Amberger, et al.** suppl_1, s.l. : Oxford University Press, 2005, Nucleic acids research, Vol. 33, pp. D514-D517.

17. *ClinVar: public archive of relationships among sequence variation and human phenotype.* **Melissa J Landrum, Jennifer M Lee, George R Riley, et al.** 2014, Nucleic acids research, pp. D980–D985.

18. *Tracking genetic variants in the biomedical literature using LitVar 2.0.* **Alexis Allot, Chih-Hsuan Wei, Lon Phan, et al.** 2023, Nat Genet, Vol. 55, pp. 901–903.

19. *AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature.* **Johannes Birgmeier, Maximilian Haeussler, Cole A Deisseroth, et al.** 2020, Science translational medicine, p. 544.

20. *PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge.* **Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, et al.** W1, 2024, Nucleic Acids Research, Vol. 52, pp. W540-W546.

21. *Exploring ChatGPT for next-generation information retrieval: Opportunities and challenges.* **Yizheng Huang & Jimmy X Huang.** 2024, Web Intelligence.

22. *Large Language Models are Zero-Shot Reasoners.* **Takeshi Kojima, Shixiang Shane Gu, Machel Reid, et al.** 2022. 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

23. *Artificial intelligence (AI)—it's the end of the tox as we know it (and I feel fine)*.* **Nicole Kleinstreuer & Thomas Hartung.** 2024, Archives of Toxicology, Vol. 98, pp. 735-754.

24. *Large language models for generative information extraction: A survey.* **Derong Xu, Wei Chen, Wenjun Peng, et al.** 2024, Frontiers of Computer Science.

25. *Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial.* **Ethan Goh, Robert Gallo, Jason Hom, et al.** 2024, JAMA Netw Open.

26. *A review of evaluation approaches for explainable AI with applications in cardiology.* **Ahmed M Salih, Ilaria Boscolo Galazzo, Polyxeni Gkontra, et al.** 2024, Artif Intell Rev.

27. *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* **Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, et al.** 2023, arXiv.

28. *AI-Enhanced Sensemaking: Exploring the Design of a Generative AI-Based Assistant to Support Genetic Professionals.* **Angela Mastrianni, Hope Twede, Aleksandra Sarcevic, et al.** 2024, arXiv (preprint).

29. *Database resources of the national center for biotechnology information.* **Eric W Sayers, Jeffrey Beck, Evan E Bolton, et al.** 2021, Nucleic acids research, pp. D20–D26.

30. *PMC text mining subset in BioC: about three million full-text articles and growing.* **Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan & Zhiyong Lu.** 2019, Bioinformatics, pp. 3533–3535.

31. *Break It Down: A Question Understanding Benchmark.* **Tomer Wolfson, Mor Geva, Ankit Gupta, et al.** 2020, Transactions of the Association for Computational Linguistics.

32. *HGVS Recommendations for the Description of Sequence Variants: 2016 Update.* **Johan T den Dunnen, Raymond Dalgleish, Donna R Maglott, et al.** 2016, Human mutation, pp. 564–569.

33. *Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature.* **Kyubum Lee, Chih-Hsuan Wei & Zhiyong Lu.** 3, 2021, Briefings in Bioinformatics, Vol. 22.

34. *Mutalyzer 2: next generation HGVS nomenclature checker.* **Mihai Lefter, Jonathan K Vis, Martijn Vermaat, et al.** 2021, Bioinformatics, pp. 2811–2817.

35. *Analysis of protein-coding genetic variation in 60,706 humans.* **Monkol Lek, Konrad J Karczewski, Eric V Minikel, et al.** 2016, Nature, pp. 285–291.

36. *seqr: A web-based analysis and collaboration tool for rare disease genomics.* **Lynn S Pais, Hana Snow, Ben Weisburd, et al.** 2022, Hum Mutat., pp. 698-707.

37. *The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.* **Peter N Robinson, Sebastian Köhler, Sebastian Bauer, et al.** 2008, The American Journal of Human Genetics, pp. 610-615.

38. *PhenoBERT: A Combined Deep Learning Method for Automated Recognition of Human Phenotype Ontology.* **Yuhao Feng, Lei Qi & Weidong Tian.** 2023, IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1269-1277.

39. *PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology.* **Ling Luo, Shankai Yan, Po-Ting Lai, et al.** 2021, Bioinformatics, pp. 1884–1890.

40. **Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, et al.** *Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models.* s.l. : Journal of Healthcare Informatics Research, 2023.