EAGLE-AI: A large language model workflow for automated extraction and scoring of literature evidence linking genes to autism spectrum disorder

Vinicius Furlan[1,2,8], Julian Moran[1,2,8], Nelson B. Salazar[2,3], Olivia Rennie[2,6], Ny Hoang[2], Andrew Wan[1], Marla Mendes de Aquino[1,2], Worrawat Engchuan[1,2], Jacob A.S. Vorstman[2,4,5], Stephen W. Scherer[1,2,3,7]

[1]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada
[2]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada
[3]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada
[4]Department of Psychiatry, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
[5]Department of Psychiatry, Temerty Faculty of Medicine, University of Toronto, Toronto, ON M5T 1R8, Canada
[6]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, M5S 3K3, Canada
[7]McLaughlin Centre, Toronto, ON M5G 0A4, Canada
[8]These authors contributed equally: Vinicius Furlan, Julian Moran

Abstract

We previously developed the Evaluation of Autism Gene Link Evidence (EAGLE) manual curation framework and characterized 223 autism-associated genes now used in the clinical diagnostic setting. However, this effort took years of human work. Here we present proof-of-concept for EAGLE-AI, an automated evidence collection, screening, extraction, and scoring system that incorporates large language models (LLMs). EAGLE-AI deploys multi-agentic workforces for both data extraction and evidence scoring, whereby the final evidence scores indicate the strength of evidence linking a given gene to autism spectrum disorder (ASD, or autism).

On an initial test set of 116 scholarly medical manuscripts screened for ease of machine-readability, EAGLE-AI achieves an F1 score of 91% in reproducing the data extractions of human curators. Its numeric evidence scores differ from those assigned by human scorers by an average margin of 12.6%. However, when input papers were not screened for ease of machine-readability, EAGLE-AI achieved only moderate performance on data extraction and low performance on scoring.

EAGLE-AI's data extraction workforce, which uses GPT-4o and o3 mini, outperforms BiLSTM-CRF and performs comparatively with Llama 3.1-8B fine-tuned on EAGLE manual curations. We also developed a deterministic scoring algorithm that achieves strong performance on screened test papers. In light of this latter finding, EAGLE-AI may benefit from deploying the deterministic algorithm instead of the agentic scoring workforce.

Our findings indicate that EAGLE-AI can successfully automate most of a clinical genomics evidence curation process. We expect that, in allowing new candidate genes and new literature to be rapidly assessed, it will save significant amounts of time. Given our findings, human intervention at the scientific literature retrieval and screening phases is likely to efficiently yield performance gains.

Introduction

Next-generation sequencing has greatly improved our ability to detect genetic variants across the human genome[1]. In leveraging next-generation methods, genetic testing achieves a greater diagnostic yield for many disorders, including autism spectrum disorder (ASD, also called autism)[2]. However, this relies on enough scientific understanding to determine the significance of the genetic variants identified. Here, knowing whether a gene is related to a condition is a crucial first step[3]. In the case of ASD, research has demonstrated a significant genetic component driving the condition[4,5,6,7,8].

Previous studies have shown considerable inconsistency in the genes included in autism gene panels[9,10]. This highlights the need to distinguish between bona-fide confirmed genes and candidate genes, referred to as establishing gene-phenotype clinical validity[11]. To promote consistency among laboratories, researchers, and clinicians, the Clinical Genome Resource (ClinGen) has developed a framework for systematically evaluating the evidence supporting a gene-phenotype association[12]. This framework involves curating publicly available evidence, including genetic and experimental data, and applying a rules-based scoring system. The evidence level is categorized as definitive, strong, moderate, or limited based on the strength of evidence. With this framework, teams of hundreds of ClinGen volunteers from the United States and around the world have evaluated more than 2,420 gene-phenotype relationships and established several gene curation expert panels, including one for Intellectual Disability and Autism[13,14].

Intellectual disability and ASD are often grouped under the umbrella term of neurodevelopmental disorders (NDDs). As an indication for genetic testing and clinical consultation, NDD as a category is commonly accepted by clinical laboratories and genetic clinics, which is reasonable given the high comorbidity among the conditions included in this general term[11,17]. However, for gene curation purposes, it is important to assess the clinical validity for ASD independently from NDDs as a whole, as the implications in clinical management may differ depending on the gene annotation. For instance, ASD often manifests independently of intellectual disability, epilepsy, or other NDDs[18,19]. Moreover, the support needs differ between individuals with ASD and those with both ASD and other comorbid NDDs[20,21]. As genome sequencing is incorporated into clinical care more widely and at earlier stages of development, understanding a gene's relevance to ASD specifically will better empower precision care[22]. Efforts to parse the heterogeneity of autism are ongoing—for example, recent work has identified clinically distinct autism subtypes linked to specific genetic programs and developmental trajectories[23,24,25,26].

To this end, we previously established the Evaluation of Autism Gene Link Evidence (EAGLE) framework for curating ASD clinical genomics evidence[10]. This approach builds on the ClinGen gene-phenotype clinical validity framework by adding a rules-based method for assessing level of confidence in the reported ASD phenotype. EAGLE also adjusts the genetic evidence scoring approach, to account for the variability and complexity of ASD genetics. EAGLE reserves high scores for variants linked to a confirmed ASD diagnosis rather than a case with autism-like features. This helps in distinguishing ASD-associated variants from variants linked to NDDs more broadly.

Previously, ASD gene lists have been defined from large-scale sequencing studies of individuals with ASD. They include lists from the Simons Foundation Powering Autism Research for Knowledge (SPARK)[27], Autism Speaks – MSSNG[28], and the Autism Sequencing Consortium (ASC)[29]. Together, these three projects identify a total of 236 unique genes as being associated with ASD with high confidence. In contrast, EAGLE is used to define an autism gene list based on evidence from the full diversity of the scientific literature[10].

EAGLE scores are currently included in the Simons Foundation Autism Research Initiative (SFARI) database[30] and characterize level of evidence for many prototypical ASD genes—for example, *NRXN1*, *SCN2A*, *MECP2*, *CHD8, DDX3X, SHANK3*, and *PTEN*. However, the EAGLE framework's manual curation workflow is both time-consuming and resource-intensive, as it requires in-depth assessment of ASD clinical case reports, functional studies, and database information.

Since at least 2003, research groups have been using natural language processing models to automate or semi-automate literature curation steps scalably[31]. Transformer architectures ushered in a new wave of interest: for instance, BERT-based models have been used to automate literature data extraction and paper classification tasks[32,33]. During the COVID-19 pandemic, a variety of automation resources emerged for dealing with the onslaught of actionable scientific literature[33,34,35]. With the more recent advent of large language models (LLMs) such as GPTs, other research groups have leveraged these models in automating literature review, such as for paper screening or data extraction[24,25,26]. However, most peer-reviewed work in this area has focused on automating only a single workflow phase[36].

In a similar vein, and to address the costs of manual literature curation in the context of ASD clinical genomics, we developed EAGLE-AI, an evidence collection, screening, extraction, and scoring workflow that incorporates LLMs. This workflow aims to accelerate the accurate identification of ASD-associated genes by operating, with minimal manual intervention, on a large and evolving corpus of literature (Fig. 1).

The data extraction and scoring modules of EAGLE-AI both deploy multi-agent workforces using o3-mini and GPT-4o (Fig. 2). We developed two additional versions of EAGLE-AI: one incorporating a BiLSTM-CRF model, trained in-house for named entity recognition (NER), and another deploying a Llama 3.1-8B model finetuned for EAGLE-compliant data extraction.

EAGLE-AI showcases not only the power of third-party LLMs in extracting data from literature, but also their lesser suitability for evidence scoring when a deterministic rules-based algorithm can be applied instead. We expect that EAGLE-AI's system architecture can generalize, in modular fashion, to automate important aspects of evidence curation for other clinical genomics research questions.

Results

System architecture

In evaluating the association between a genomic variant and ASD, the manual EAGLE curation standard assesses evidence from three categories: genetic evidence, quality of ASD diagnostic evaluation, and experimental evidence (e.g. animal models) (Table 1). Because the EAGLE curation standard is rigorously rules-based, EAGLE-AI implements the same rules in automated fashion for the tasks of data extraction and scoring. As such, the system is comprised of two interconnected modules: an evidence collection service and EAGLE-AI core, which performs the data extraction and scoring steps (Fig. 1).

The evidence collection service maintains a list of genes currently under investigation and submits requests to Papers-Nexus, a purpose-built web API for automatically retrieving and downloading open-access full-text PDFs from PubMed, Elsevier, and Google Scholar using search terms. To keep the EAGLE-AI databases up to date, the evidence collection service requests papers pertaining to a given gene on a scheduled periodic basis.

The evidence collection service subsequently processes the collected papers by converting the PDFs to plain text, extracting table contents, and composing image-to-text descriptions of the paper figures. All text extracted from the papers is converted to embeddings using OpenAI's text-embedding-3-large matrix and stored in a vector database.

The EAGLE-AI workflow handles internal data storage using two distinct databases (Fig. 1). It stores tabular data containing string entries, i.e. paper metadata, extracted cases, and variant information, in a relational database. Embeddings representing each paper's text are indexed using a vector database. Using an API endpoint for o3-mini[47], EAGLE-AI assigns semantic flags to each text chunk, where the flags pertain to genes, phenotype terms, and variants.

EAGLE-AI core deploys eight distinct LLM agents, four of which are dedicated to data extraction and four of which are dedicated to evidence-scoring each gene and candidate ASD case (Fig. 1). The data extraction agents write to the relational database, using a tabular format populated with detailed information on the ASD case subject, gene, etiological genomic variant, and phenotype (Table S1). EAGLE-AI core's scoring agents add a final score, indicating the strength of ASD etiological evidence, to the relational database as well. The relational database is therefore the final repository of EAGLE-AI's data extractions and evidence scores.

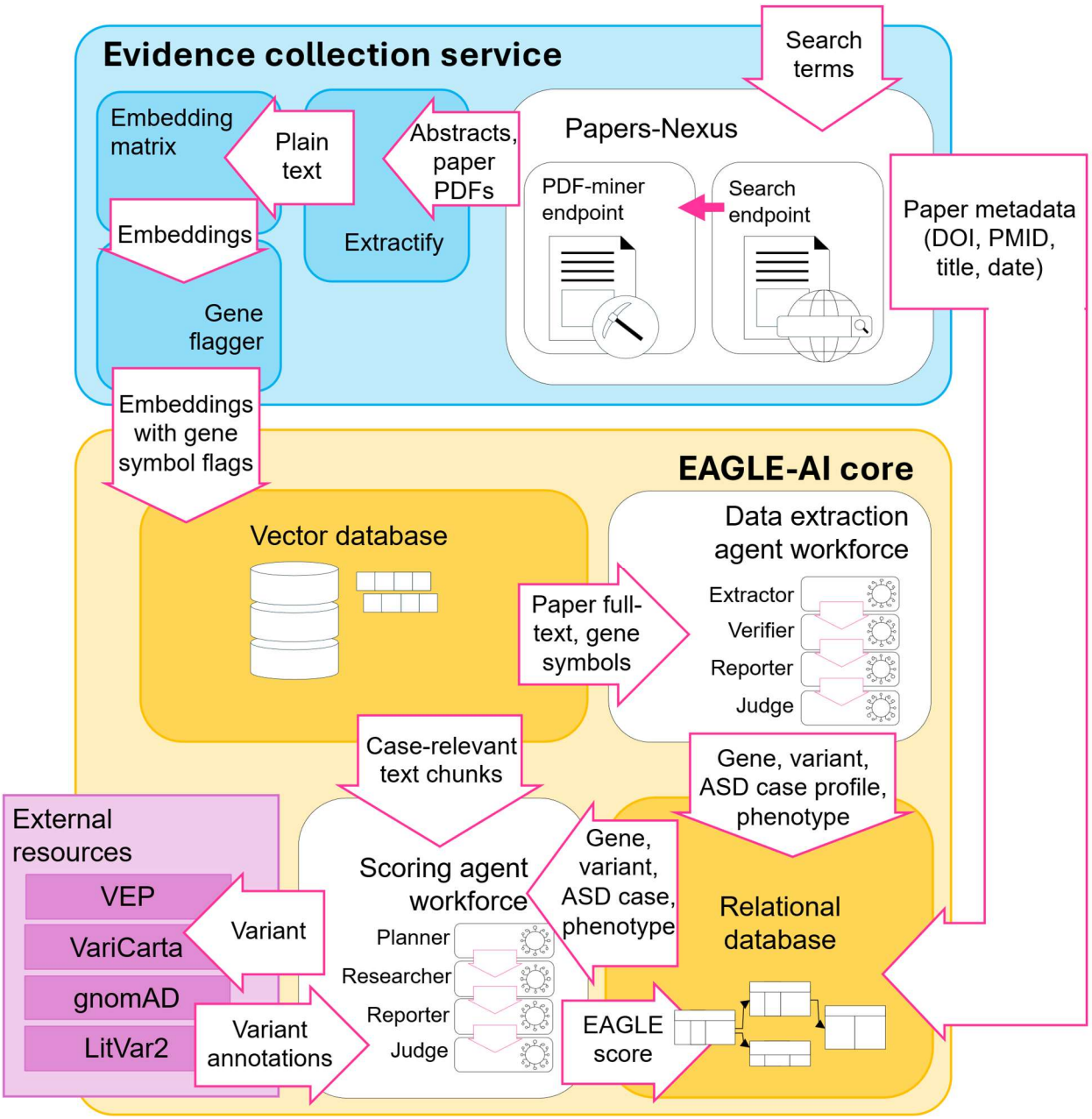

Fig. 1 | **EAGLE-AI system architecture.** Schematic of EAGLE-AI's whole-system architecture, with initial input consisting of natural-language search terms and final output consisting of the populated relational database. Search terms consist of a gene name (e.g. "*DMD*") and the fixed search terms "ASD" and "autism". Database entries are dedicated to single ASD cases and include a case patient description, phenotype description, gene name, genomic variant ID, and evidence score.

Evidence collection service architecture

EAGLE-AI's evidence collection service submits requests to Papers-Nexus, an API developed in-house for performing automated literature searches and bulk full-text retrieval from natural-language search queries. To find relevant paper metadata, EAGLE-AI submits automated requests to the search endpoint of Papers-Nexus, where the request includes a general search term field, a gene search term field, and an exclusion term field. Search criteria consist of a gene name (e.g. "*DMD*") and the fixed search terms "ASD" and "autism".

Using the Bio.Entrez[41] and scholarly[42] Python packages, Papers-Nexus searches for relevant papers in the PubMed, Google Scholar, and Elsevier databases. Its response to the client provides the paper title, author name(s), abstract, PMID, DOI, and citation information for each paper that satisfies the search criteria.

In light of our design goal to have EAGLE-AI extract data from a large corpus of papers with minimal manual input, we developed a PDF-miner endpoint for Papers-Nexus. This endpoint is dedicated to bulk downloading open-access full-text PDFs of papers, where the client's request contains a list of PMIDs or DOIs. Papers-Nexus retrieves full-text PDFs using the PubMed-OA FTP server[43] and the Unpaywall API[44].

Extractify, a module within EAGLE-AI's evidence collection service, uses Adobe's pdf.services[45] and openpyxl[46] to convert retrieved papers and supplementary tables to plain text. For each paper, a gene flagger module deploys an o3-mini model to extract HGNC symbols for all ASD-relevant genes mentioned by the paper, as well as paper metadata. The gene symbols are assigned as flags to the paper's full text. Here, gene symbols are flagged only if the paper presents clinical or experimental evidence linking the gene to ASD.

Evidence collection service retrieval rate

Using a test set of metadata for 172 papers from the PubMed database, we found that Papers-Nexus retrieved full-text PDFs for 114 (66.3%) of the papers. For all remaining papers, Papers-Nexus retrieved the title, author names, and abstract. Papers-Nexus failed to retrieve a paper's full-text if the paper is not open-access or if the paper's host denies transparent requests to automatically download the PDF.

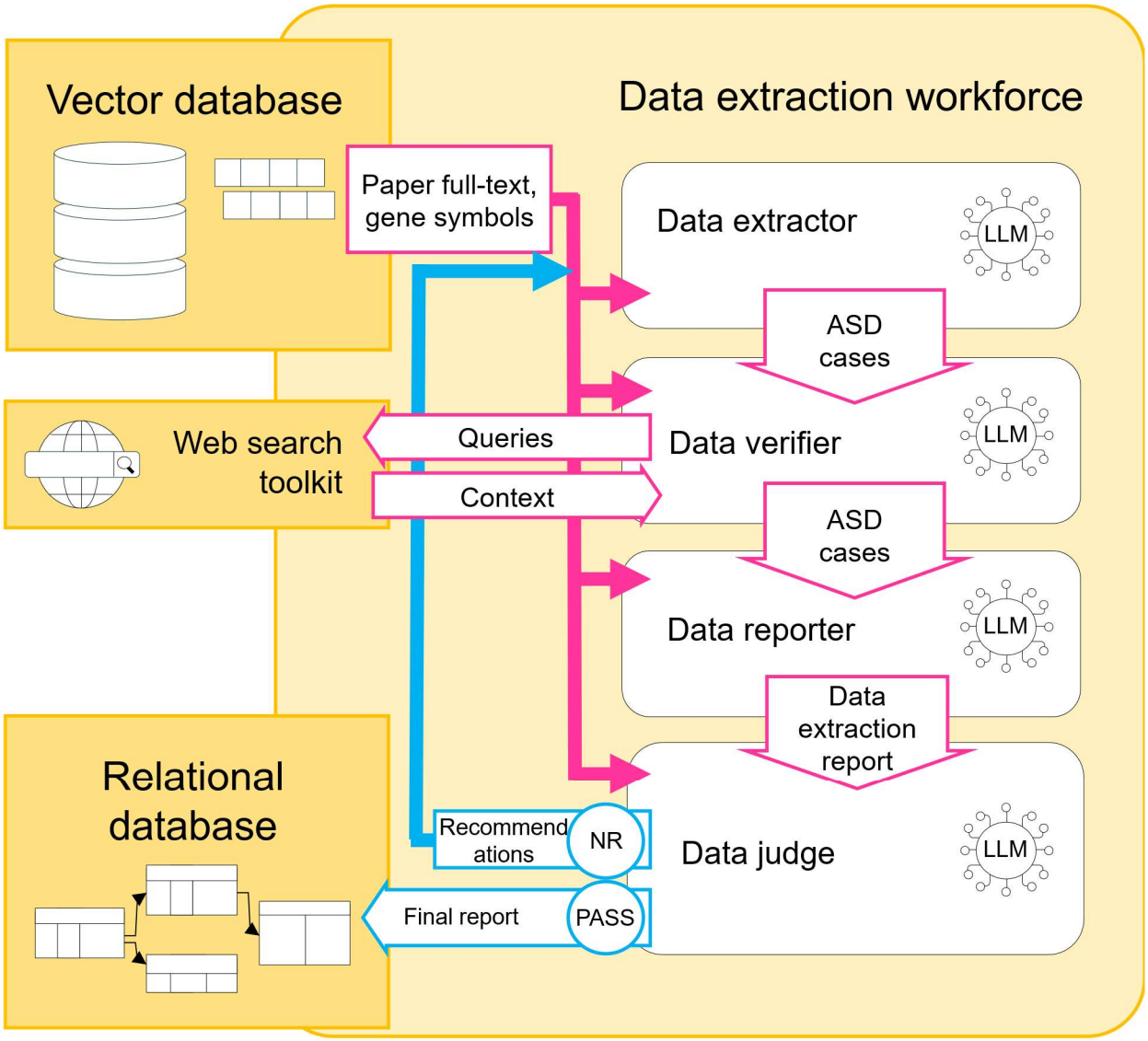

Fig. 2 | **Agentic EAGLE-AI core data extraction workflow.** Schematic of Eagle-AI core's agentic data extraction workforce. The workforce includes an additional task planner and coordinator agent for incorporating recommendations from the judge in the event of a "NEEDS_REVISION" (NR) verdict. Depending on the recommendations, this can yield revision tasks to any of the four workforce agents. The final output report complies with the format presented in Table S1.

Agentic data extraction architecture

EAGLE-AI's data extraction workforce is responsible for extracting three categories of relevant information required for scoring: genomic variant, case ID, and phenotype. It deploys four distinct LLM agents, which perform automated tasks based on their own decision-making from inputs. The agents are assigned to the roles of data extractor, data verifier, data reporter, and data extraction judge. The data extractor, verifier, and judge use OpenAI's o3-mini model[47], whereas the reporter uses GPT-4o[48]. The agentic workforce is managed using the CAMEL-AI multi-agent framework.[50] Input to the data

extraction workforce consists of a single paper's plain full-text and all gene symbols flagged by the gene flagger (Fig. *2*).

For each gene in the paper, the data extractor agent receives a prompt that includes the given gene symbol, instructing the extractor to identify and extract all ASD cases in the paper linked to the given gene. For each case, the case subject must have an ASD diagnosis under specified criteria; the gene must be described with identifying detail; and the paper must describe a specific ASD-driving genomic variant. For all eligible cases, the extractor gives as output the case subject's identifier, the location of the text where the case is described, a brief description of the variant, and an account of pertinent data missing from the paper (Table S2).

A prompt to the verifier agent instructs it to review all sections of the paper full-text and identify any cases missed by the extractor. Using CAMEL's SearchToolkit, the verifier is able seek further information through web search[51]. For each case, its output format resembles that of the extractor (Table S3).

The reporter agent's task is to synthesize all cases from the extractor and verifier. The reporter is instructed to remove any duplicate cases and to reformat the data to match the example presented in Table S1, which in turn matches the schema of EAGLE-AI's relational database.

The judge agent assesses the reporter's output for quality and consistency. In evaluating quality, the judge checks for whether the data types of populated fields comply with the database schema and whether gene and variant details are logically consistent. The judge flags cases with insufficient details about the gene or candidate ASD-driving variant. It also flags any potentially duplicate cases. Its output includes an overall quality score of high, medium, or low (Table S4). Additionally, it issues a verdict of the gene data from the paper: namely, as "PASS" or "NEEDS_REVISION".

As per the CAMEL-AI multi-agent framework[50], an additional task planner agent and coordinator agent handle task assignment to the workforce beyond the initial prompts. In cases where the judge issues a verdict of "NEEDS_REVISION", these agents assign new tasks to the extractor, verifier, reporter, and/or judge based on the judge's account of the issues. The workforce's passed output writes to EAGLE-AI core's relational database.

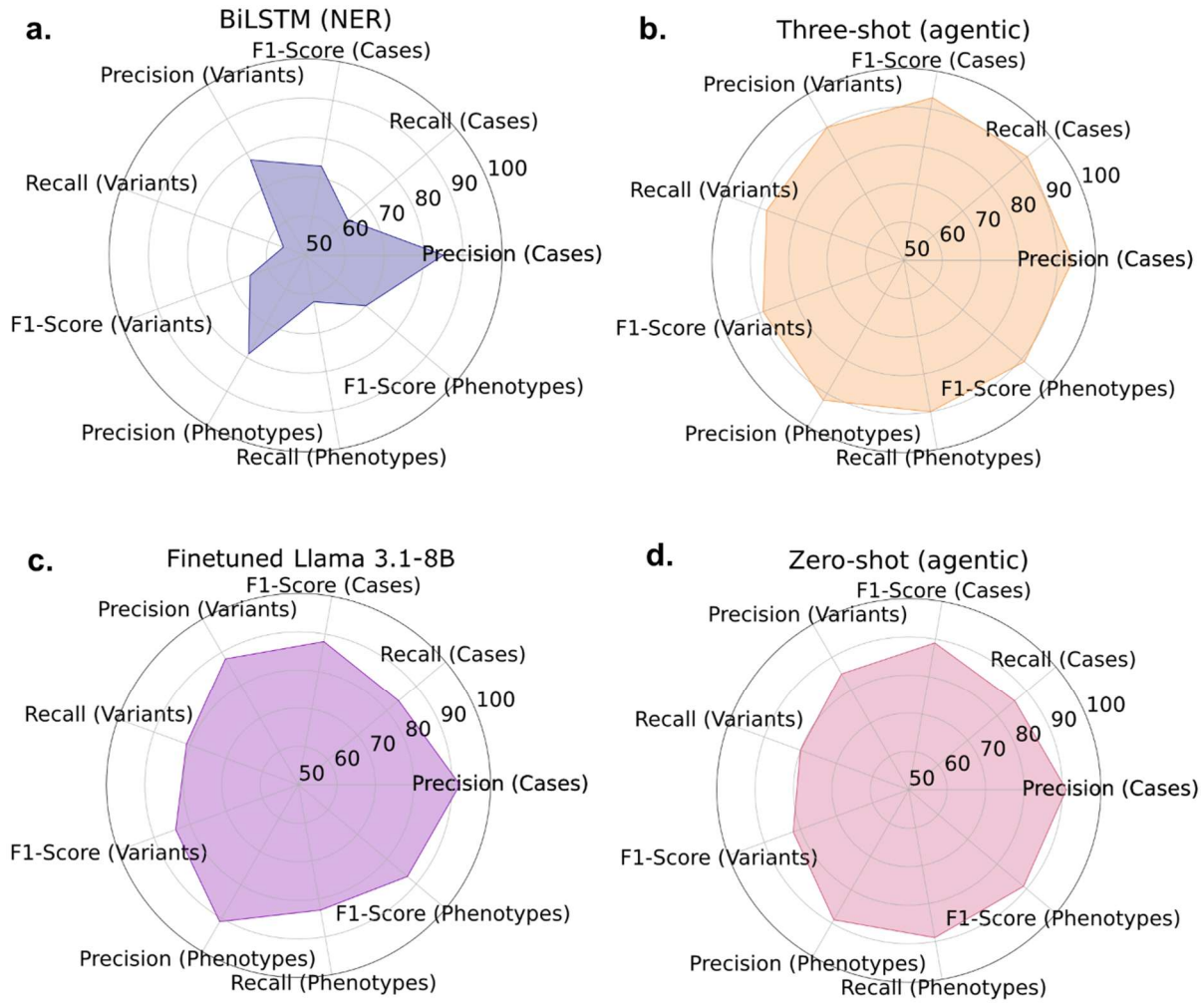

Fig. 3 | **EAGLE-AI data extraction performance.** EAGLE-AI's annotation extraction precision-recall metrics out of 100 using four different approaches, evaluated across manually scored cases from 116 papers; here, the evaluation set was restricted to cases with a final EAGLE score and with a reported variant in *DMD, SHANK1, CACNA1D, DDX3X, DDX53*, or *MECP2*. **a** Extraction performance using the NER model for relevance filtering and text chunk assembly, coupled with a GPT-4o endpoint for data extraction. **b** Extraction performance of the NER model coupled with a finetuned Llama 3.1-8B model for data extraction. **c** Extraction performance of agentic EAGLE-AI core using a zero-shot task prompt to the data extraction workforce. **d** Extraction performance of agentic EAGLE-AI core using a three-shot task prompt to the data extraction workforce.

EAGLE-AI data extraction precision-recall

We evaluated EAGLE-AI core's data extraction precision-recall for four different configurations. Here, we measured how well the agentic workforce reproduced the extractions of the EAGLE manual curators. In the first configuration, the agentic workforce was provided no input-output examples in the extractor prompt ("zero-shot"); in the second, the prompt included three input-output examples ("three-shot").

In the third configuration, we incorporated the BiLSTM-CRF NER model into EAGLE-AI core. Here, we coupled the NER model with regular expression pattern-matching fallbacks and a single GPT-4o endpoint[48] for data extraction.

We also finetuned a single Llama 3.1-8B model[64] for the task of data extraction. In the fourth configuration, this model was coupled with the BiLSTM-CRF NER model: we used the NER model to relevance-filter and assemble paper text chunks, which we then incorporated into the prompt to the finetuned Llama 3.1-8B model for data extraction.

We evaluated all four configurations of EAGLE-AI core on the same set of 116 publications and 96 cases screened for ease of machine-readability. The initial test set was constrained to cases with a final EAGLE score and a reported variant in *DMD, SHANK1*, *CACNA1D*, *DDX3X*, *DDX53*, or *MECP2*. Average F1 score was calculated as the mean of the case, variant, and phenotype F1 scores. Each configuration's extracted cases were matched to the manual curations using a confidence-based case-matching algorithm (see Methods).

The zero-shot agentic workforce achieved an average F1 score of 86% (Fig. 3c), whereas the three-shot agentic workforce achieved average F1 of 91% (Fig. 3d). The NER model achieved an average F1 of 69% (Fig. 3a). Finetuned Llama 3.1-8B achieved average F1 of 86% (Fig. 3b).

Table S5 presents an example whereby, after finetuning, the model was better able to identify the pathogenicity impact of a *DDX3X* variant NM_001356.5:c.1171-2A>G as reported in a paper abstract and case patient description. For a case with a *MECP2* indel variant, Llama 3.1-8B assigned the accurate phenotype confidence level (low) only after finetuning, resulting in an EAGLE score closer to the manual ground truth (Table S6).

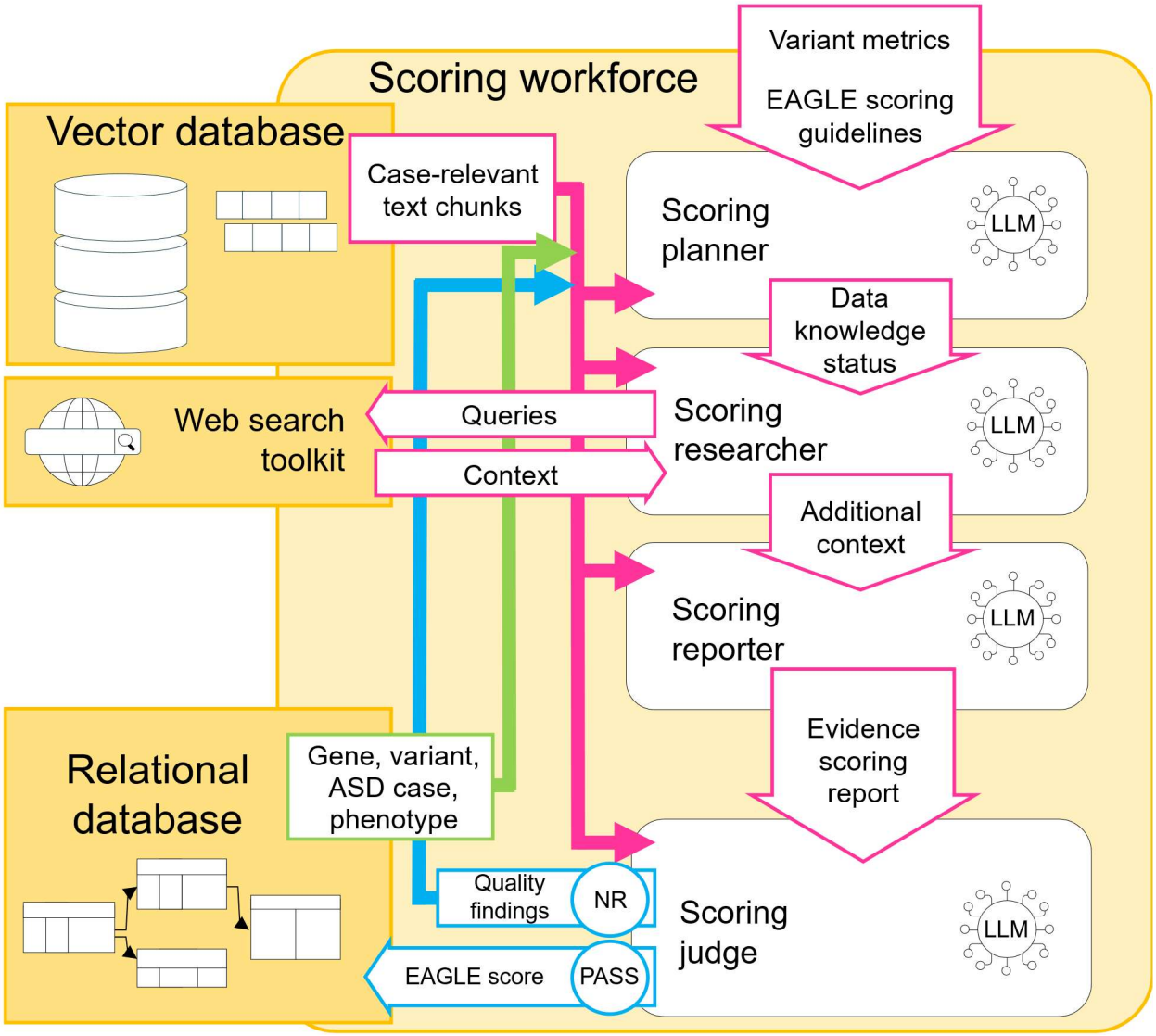

Fig. 4 | **Agentic EAGLE-AI core scoring workflow.** Schematic of the scoring workforce within agentic EAGLE-AI core. Initial inputs consist of case-relevant paper text chunks from the vector database; extracted gene, variant, ASD case, and phenotype data from the relational database; variant metrics from external API calls; and a prompt containing the EAGLE scoring guidelines. The workforce includes an additional task planner and coordinator agent for incorporating recommendations from the judge in the event of a "NEEDS_REVISION" (NR) verdict.

Agentic evidence scoring architecture

Fig. 4 presents a schematic of EAGLE-AI core's multi-agent workforce for scoring possible ASD cases and their reported variants. Much like the data extraction workforce, the scoring workforce deploys four distinct LLM agents, with similar task breakdowns: a planner, researcher, reporter, and judge. As of the current version of EAGLE-AI core, the scoring planner and judge use o3-mini[47], while the scoring researcher and reporter used GPT-4o[48].

Since the EAGLE scoring guidelines (Table 1) require variant information such as null or missense status, functional data, de-novo status, population frequency, and splice-site-modifying status, an upstream variant annotator agent searches for these metrics using third-party APIs (Fig. 1). The annotator sends requests to the Ensemble variant effect predictor's (VEP)[52] API for transcript-level, protein-level, and clinical impact; VariCarta's[53] for known ASD association and pathogenicity scores (CADD 1.3[54], SIFT[55], and polyphen[56]); gnomAD's[57] for population frequency; and LitVar2's[61] for HGVS protein variant notation and pathogenicity status.

The planner agent receives the query case ID, the case's candidate ASD-driving variant annotated with the third-party metrics, all paper text chunks relevant to the case as stored in the vector database, and all associated data from the data extraction module as stored in the relational database (Fig. 1). It receives also the EAGLE scoring guidelines (Table 1) and a checklist of data required for applying the guidelines (Table S7). Any missing data are tagged by the planner agent as requiring additional search. For all data labelled as missing, the researcher agent uses the suggested search terms and the CAMEL-AI web search toolkit to attempt to retrieve the data using a suite of search engines[51].

A downstream reporter agent leverages all input data from the relational database, vector database, annotator, and scoring researcher in applying the EAGLE scoring guidelines as per Table 1. It returns a report for the case, summarizing the data and assigning an EAGLE score to the case's genetic evidence, ASD phenotype evaluation, and experimental evidence (Table S8). It also summarises any evidence against the case's ASD diagnosis or against the variant's driving of ASD risk. It yields a final EAGLE score for the case, with a maximum of 2 points for genetic evidence and a maximum of 6 points for experimental evidence.

The judge agent appraises the quality of the report and issues a final verdict of "PASS" or "NEEDS_REVISION". In the event of a "PASS" verdict, the workforce writes the case's EAGLE score to the relational database. In the event of a "NEEDS_REVISION" verdict, the scoring workforce incorporates the judge's feedback and assigns follow-up tasks to the agents as appropriate (Fig. 4).

EAGLE-AI scoring error

We tested performance of the agentic scoring workforce when provided with three correct scoring examples in the initial task prompt (“three-shot").

Additionally, we measured performance of a deterministic if-else scoring algorithm written in Python. We coupled the algorithm with the BiLSTM-CRF NER model, which was used for relevance-screening and assembling text chunks. To generate the data extractions, we used post-NER regular expression matching to augment entity labelling for the purposes of relevance filtering. We then submitted a single dynamic prompt to a GPT-4o-mini endpoint[49], which instructed the model to extract the data (Table S10). We used the resulting extractions as input to the algorithm.

The algorithm applies the rules-based EAGLE scoring guidelines for genetic, phenotypic, and experimental evidence as described in Table 1. However, because if-else logic cannot program subjective decision-making, the algorithm applies only the suggested scores instead of the discretionary score ranges specified in brackets.

Across 55 cases, the scoring workforce achieved a case-wise symmetric mean absolute percentage error (SMAPE) of 12.6% as per equation [1] (see Methods). As this is an error term, this indicates 87.4% correspondence with manual scores. Across 128 cases, the deterministic scoring algorithm achieved a case-wise SMAPE of 10.1%, thereby outperforming the agentic scoring workforce. Distributions of scores for both the agentic workforce and the scoring algorithm closely match that of the manually curated scores (Fig. 5a, c). For both the workforce and the algorithm, Bland-Altman plots do not indicate error bias with respect to score magnitude (Fig. 5b, d). For the workforce, 80.7% of its case scores have zero difference from their corresponding manual case scores; for the algorithm, 87.5% of its case scores have zero difference.

Whole-system performance on eight genes without prior screening

We computed performance metrics for the whole EAGLE-AI system (Fig. 1) using papers pertaining to eight genes that were excluded from all prior training, test, and validation sets. The eight genes evaluated were *CACNA1A*, *CASK*, *CDKL5*, *CHAMP1*, *CHD3*, *FMR1*, *GIGYF2*, and *GRIN1*. We evaluated all 36 papers related to these genes retrieved by Papers-Nexus; the papers contained a total of 85 candidate ASD cases. Here, we evaluated performance of EAGLE-AI while using the agentic workforces for the data extraction and scoring steps, where the workforces received three-shot task prompts. The 36 papers were not screened for ease of machine-readability or access. Table S15 includes the manual data extractions, EAGLE-AI's data extractions, and the raw performance data.

EAGLE-AI achieved moderate performance on data extraction, with average accuracy of 72% (Table 3). EAGLE-AI's scoring accuracy was low, with gene-wise SMAPE of 45.0% (Table 3).

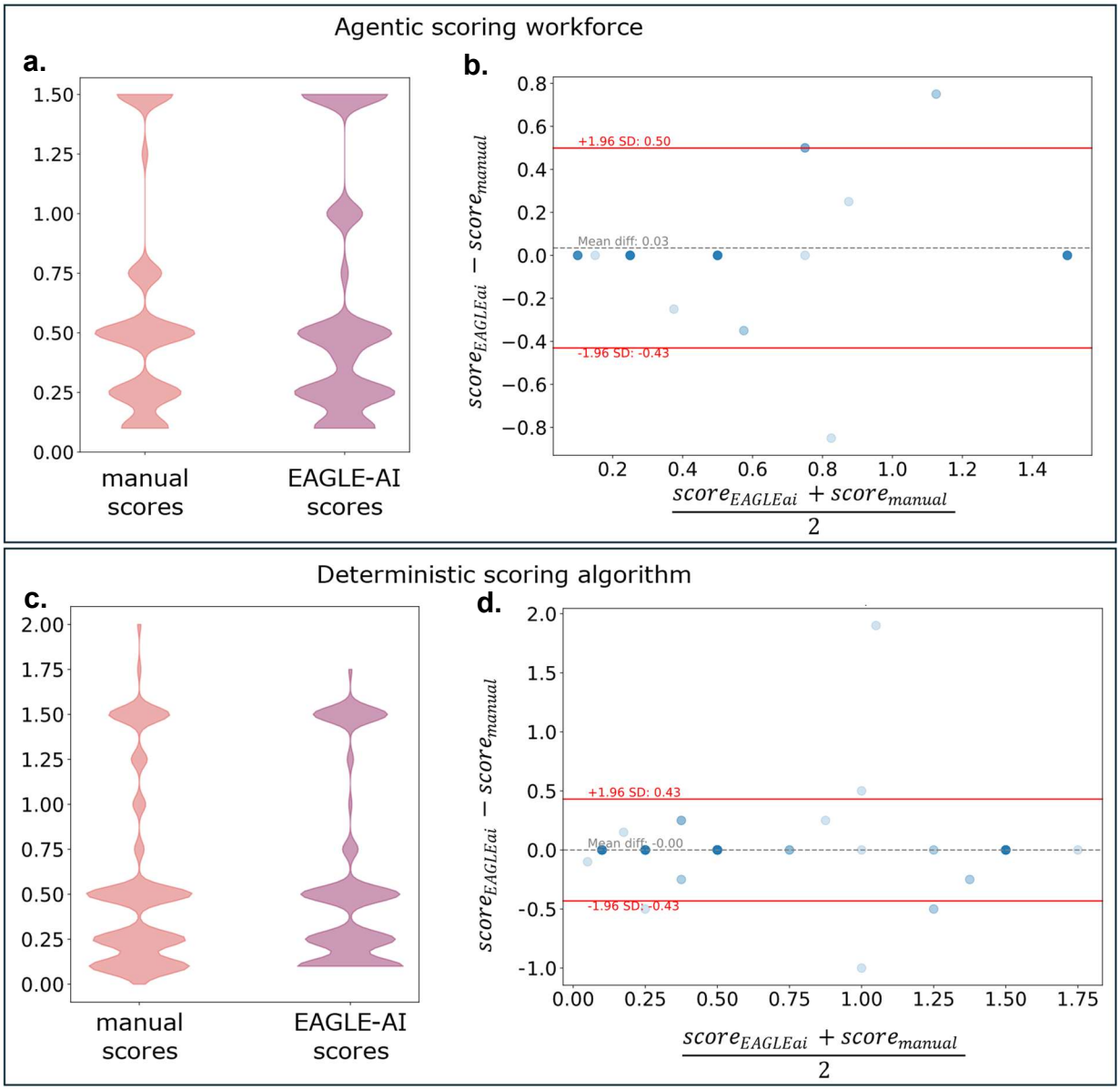

Fig. 5 | **EAGLE-AI scoring performance.** EAGLE-AI's scoring performance under different conditions, where greater score means greater strength of evidence linking a case's variant to ASD. **a, b** Agentic scoring workforce performance across 116 papers and 55 ASD cases. **a** Violin plots of manually curated case scores (orange) v. agentic (purple) EAGLE-AI case scores. **b** Bland-Altman plot for agentic EAGLE-AI's scores benchmarked against manual scores, where darker points indicate overlap; here, each point represents an ASD case. **c, d** Deterministic scoring performance across 90 papers and 128 ASD cases. **c** Violin plots of manually curated case scores (orange) v. deterministic algorithm's (purple) scores. **d** Bland-Altman plot for deterministic algorithm's scores benchmarked against manual scores.

Discussion

On the initial set of 116 papers pertaining to six ASD-associated genes, EAGLE-AI achieved near-human performance on both data extraction (Fig. 3) and scoring tasks (Fig. 4). The multi-agent extraction workforce, incorporating GPT-4o and o3-mini, outperformed alternative versions using BiLSTM-CRF or finetuned Llama-3.1-8B. Our findings serve as proof-of-concept for applying third-party LLMs, organized in multi-agent workforces, to clinical genomics evidence curation. We expect this approach will save significant amounts of time for humans.

We interpret these results with an important caveat: namely, these 116 papers were screened for favourable text extraction conditions. Converting the entirety of a paper to plain text is nontrivial, so we first screened these papers for ease of machine-readability (see Methods). We conducted this assessment in order to isolate performance of the data extraction and scoring modules. However, this means the input dataset may not fully represent the diversity of relevant scientific literature available for testing.

During this initial round of performance assessment, we treated human curation as the gold standard. For this reason, instances of super-human performance, whereby EAGLE-AI avoids errors made by human curators, were not possible to detect. The reported performance metrics of F1 93% and scoring error 12.6% may therefore understate EAGLE-AI's true performance on these tests.

On follow-up, whereby we evaluated the whole-system performance on eight genes without prior screening of the papers, the assessment was manual (see Methods). Here, we noticed instances whereby, on review, EAGLE-AI included more correct details about ASD case phenotypes than did the manual curators (Table S15). In all such instances, the details pertained to highly specific co-occurring conditions and test results (e.g. anatomical abnormalities, brain imaging findings). Under the phenotype accuracy rating system used for performance assessment, these contributed only 10 out of 100 points to the phenotype accuracy rating (see Methods). We therefore updated the manually curated descriptions with EAGLE-AI's findings and continued to treat the curations as the gold standard. Beyond these details in the phenotype descriptions, we did not observe instances of super-human performance from EAGLE-AI.

The follow-up assessment without paper screening exposed EAGLE-AI's limitations. On this test, data extraction performance was moderate at 72% accuracy and scoring performance was low with error of 45%. Here, EAGLE-AI struggled most with variant recall at 41% (Table S14), meaning that more than half the time it failed to extract the correct ASD-associated variant as reported in the literature. Most often, this was because the variants were reported in supplementary tables that EAGLE-AI could not access through Papers-Nexus. A "missing supplementary" flagger, which prompts human curators to retrieve and upload supplementary materials to EAGLE-AI in the event they are missing, may help address this. Given Papers-Nexus's 66.3% PDF retrieval rate, a similar "missing paper" flagger would improve EAGLE-AI's coverage of the literature. This is one example whereby automated tools for evidence curation may require ongoing human intervention.

The tables in the papers themselves also contributed to low variant recall. Many papers present cases and their ASD-driving variants in a large table within the full-text. Since these tables are coded as images within PDFs, EAGLE-AI relies on image-to-text conversion to read the table contents, which is more error-prone when the table is in landscape orientation and reads up-to-down. We plan to apply OpenCV's optical character recognition and image rotation tools[70] to standardize all paper tables before extraction.

Regarding EAGLE-AI's lower performance on scoring, context overload may have contributed. As the 36 follow-up papers were not screened, EAGLE-AI more often encountered critical missing information and depended on background knowledge and the web-search toolkit more. The quality of information from web search was variable and had the potential to confuse the scoring workforce. For instance, when scoring *FMR1*—a gene canonically associated with Fragile X Syndrome, which is accompanied by autism in about 40% of male and 15% of female carriers[71]—EAGLE-AI treated Fragile X Syndrome diagnoses as equivalent to ASD. This in turn yielded erroneous scores for many of the *FMR1* cases.

Additionally, the follow-up paper set included review papers or purely functional studies that did not reproduce ASD-specific phenotypes, which EAGLE-AI sometimes failed to exclude. This latter phenomenon also accounts for why data extraction performance was generally lower. To refine context inputs to EAGLE-AI, we are considering deprecating the web search toolkit, grounding the model to information in the paper texts, and placing a stronger emphasis on inclusion and exclusion criteria in the data extraction workforce prompts. Furthermore, the difference between our initial and follow-up results suggested that human intervention to screen out reviews and off-topic papers may improve EAGLE-AI's performance considerably.

On a test set of 90 papers, our deterministic scoring algorithm performed comparably to the agentic scoring workforce (Fig. 4) and achieved a lower scoring error of 10.1%. EAGLE's scoring framework is rigorously rules-based[10] such that we could viably program it with if-else algorithmic logic. While LLMs are well suited for extracting data of a standardized format from scientific literature, this finding supports using traditional rules-based algorithms for evidence scoring instead of LLM agents. We intend to deprecate EAGLE-AI's agentic scoring workforce and replace it with the deterministic algorithm, with the API calls to external resources instead integrated into the data extraction workforce. In our current work, we use EAGLE-AI to regularly assess genes for their role in ASD—this includes genes emerging from our own discovery research and those added to the SFARI database.

Setting aside the current performance limitations, we propose that EAGLE-AI's architecture should be able to generalize to clinical genomics evidence curation as a whole. For instance, EAGLE-AI's workflow may be able to replace or augment the manual curation approach used by ClinGen's gene-disease validity working groups.

Notably, the human genome project initiated early efforts to digitize data at large scale and to standardize nomenclature, genome references, and databases[72,73,74,75]. For this reason, within the realm of biomedical annotation, artificial intelligence may perform better when applied to genetics data.

Some features of EAGLE-AI are tailored to the EAGLE curation task. However, we designed EAGLE-AI with modularity in mind. To apply the workflow to a different curation task, for example to curating for a different disease or disorder, much of the architecture is reusable. For a workflow that performs only data extraction, one would need to change only the search terms, prompts to the data extraction workforce, and schema of the relational database. For each new curation task, however, the prompts and schema would require design consultation with both manual curators and domain subject-matter experts.

For additional evidence scoring based on ClinGen's semi-quantitative ("Definitive", "Strong", "Moderate", "Limited") framework[13], a tailored scoring algorithm would likely need to be developed for each genetic disease or disorder/condition. For curation tasks whereby scoring requires considerable subjective decision-making, an agentic LLM workforce may prove more suitable. However, one might re-apply a suite of general principles (e.g. nonsense de-novo variants or deletions present stronger evidence for a monogenic role than inherited or missense variants), such that the development time for disease-specific scoring modules may be short.

With the advent of LLMs, we expect that automated workflows for evidence curation will become more common. As a result, the human labour of curating evidence will likely transition away from manual data extraction or evidence scoring and toward designing database schemas, LLM prompts, and evidence scoring algorithms. Moreover, researchers and publishers may come to reconsider how scientific material is published, for example by more often supporting open-access and machine-readable formats.

EAGLE-AI, in its current state, still requires human intervention while retrieving papers and supplementary materials, as well as when screening papers for suitability. However, further development should ameliorate these limitations, and the workflow succeeds in automating most of the EAGLE evidence curation framework with near-human performance. Our findings serve as proof-of-concept for applying third-party LLMs, organized in multi-agent workforces, to the task of clinical genomics evidence curation.

Methods

Named entity recognition (NER) model architecture

We developed a version of EAGLE-AI core that does not use LLM agentic workforces but rather depends on a bilateral long short-term memory (BiLSTM) neural network with a conditional random-field (CRF) layer. Using PyTorch[62], we trained the BiLSTM-CRF model to perform named entity recognition (NER) of genes, variants, and medical condition terms from the extracted literature text. The entity mentions are used to relevance-filter and assemble text chunks before loading them into the EAGLE-AI data extraction module. Here, the data extraction module queries a single GPT-4o API endpoint[48] for data extraction or, alternatively, a finetuned Llama 3.1-8B model[64].

The NER model consists of an embedding layer, an attention layer, a BiLSTM layer, and a CRF layer. We used pretrained BioWordVec embeddings[63] for the embedding layer, with a vocabulary size of 10,475 and where each embedding vector is 200-long. Each embedding vector is concatenated to a corresponding 150-long output from a character

encoder, consisting sequentially of a linear layer and ReLU activation function. Both the embedding layer and the character encoder were further trained alongside the rest of the model.

The BiLSTM layer consists of two feed-forward LSTM networks with 192-long hidden states. Its output is a tensor of emission scores, which feed into an attention layer consisting sequentially of a linear layer, Tanh activation, multiplication by a context vector, and SoftMax activation. Output from the attention layer feeds into the CRF layer; the CRF layer also receives a target tensor, which contains the true classification scores as per the labelled train/test data.

The NER model labels each token as one of 31 named entities pertaining to case, variant, and phenotype information. On the test dataset, it achieved an F1 score of 56% across 201 "PHENOTYPE" entities, 87% across 26 "ASD" entities, 65% across 71 "DISEASE" entities, and 77% across 154 "GENE" entities.

NER model training

We trained the NER model in PyTorch[62] using the BioRED dataset, which consists of 600 clinical PubMed abstracts[65]. Tokens were generated by first splitting the text into sentences by punctuation boundaries, then separating words by whitespace and isolating leading or trailing punctuation marks as individual tokens. Here, the labelled entities of interest pertained to genes, genomic variants, diseases/disorders, patient case information, signs/symptoms, and experimental validation. We converted the labelling scheme from beginning, inside, outside (BIO) to a binary entity labelling scheme whereby each token is labelled as either inside an entity or outside any entity. We treated output of the CRF layer as the loss function. The model was trained for 36 epochs at a learning rate of 2e-2 and with dropout of 0.5.

Llama 3.1-8B finetuning

We finetuned the Llama 3.1-8B model[64] using a JSON of manually extracted data as per the EAGLE curation guidelines described in Table 1. The JSON contained query and target response pairs. 20 query-response pairs gave probing questions and target answers about the EAGLE extraction and scoring guidelines; 390 subsequent query-response pairs consisted of assembled paper text chunks and the human-extracted data and scores.

Text was tokenized using LlamaTokenizerFast. Using a low-rank (LoRA) matrix finetuning approach, we targeted parameters in the query, key, value, and output projection tensors of the model's attention mechanism, as well as the gate, up, and down projection tensors of the multilayer perceptron block. Finetuning was conducted over a single epoch using the SFTTrainer from the HuggingFace transformer reinforcement learning library[66], with a cross-entropy loss function and a learning rate of 2e-4. For parameter optimization, we used paged AdamW 32-bit.

Data extraction precision-recall evaluation

In testing performance of EAGLE-AI core's data extraction precision-recall, we drew from evidence extracted by EAGLE's manual curators. We constrained the cases to only those that had a final EAGLE score and had a reported variant in *DMD, SHANK1, CACNA1D, DDX3X, DDX53*, or *MECP2*. We also screened papers for ease of machine-readability: we checked that machine-extracted text contained minimal misidentified special characters, missing words, incorrect table structures, or missing supplementary data. In total, we included 116 papers.

EAGLE's manual curation workflow is already established and described elsewhere[10]. We collected and stored the source literature from which the manual curators extracted patient/subject information, phenotype descriptions, and variant information. We stored embeddings for the documents in EAGLE-AI's vector database and used them as input to EAGLE-AI core.

In evaluating EAGLE-AI's data extraction performance, we treated presence of an extraction as positive and absence of an extraction as negative. We evaluated the case ID extractions, variant ID extractions, and phenotype descriptions as separate categories. For each data extraction category, we computed the true positive, false positive, and false negative rate using the definitions in Table 2. This gave us the corresponding precision, recall, and F1-score of EAGLE-AI core's data extractions.

We matched cases extracted by EAGLE-AI to their corresponding manually curated cases by first matching by normalized publication name and author. We achieved further case-specific matching by adding confidence scores in the event of a matching gene name, case ID, or inheritance pattern.

We tested EAGLE-AI's multi-agent workforce when using task prompts with zero input-output examples ("zero-shot") (Fig. 3c) and three input-output examples ("three-shot") (Fig. 3d), respectively. We also tested data extraction precision-recall of the NER model when used for relevance filtering and text chunk assembly for a given case: during testing, we coupled the NER model with either a single GPT-4o API endpoint[48] for data extraction (Fig. 3a) or the finetuned Llama 3.1-8B model for data extraction (Fig. 3b).

Evidence scoring error evaluation

In evaluating EAGLE-AI core's scoring error, we computed the symmetric mean absolute percentage error (SMAPE) as per equation [1]. Here, n denotes the number of genes when computing gene-wise SMAPE or the number of candidate ASD cases when computing case-wise SMAPE.

[1] $$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|scor\ \ _{EAGLEAI} - scor\ \ _{manual}|}{(|scor\ \ _{EAGLEAI}| + |score_{manual}|)/2}$$

We matched cases scored by EAGLE-AI to their corresponding manually curated cases by first matching by normalized publication name and author. We achieved further case-specific matching by adding confidence scores in the event of a matching gene name, case ID, or inheritance pattern. For each case extracted by EAGLE-AI, we treated its

highest-confidence match as the corresponding manually curated case when applying equation [1].

We evaluated the scoring error of EAGLE-AI's multi-agent workflow when given zero-shot and three-shot task prompts.

We also evaluated scoring performance of EAGLE-AI when using the NER model augmented by regular expression pattern matching and coupled to a single GPT-4o-mini endpoint[49] for data extraction using the prompt in Table S10. Here, scoring was performed using a deterministic algorithm as per the rules in Table 1.

Whole-system evaluation on eight genes without prior screening

We evaluated performance of the whole EAGLE-AI system using the agentic LLM workforces for the data extraction and scoring steps. Here, the workforces received three-shot prompts. The genes evaluated were *CACNA1A*, *CASK*, *CDKL5*, *CHAMP1*, *CHD3*, *FMR1*, *GIGYF2*, and *GRIN1*, which had never before been scored using the EAGLE framework. These genes were selected because some have strong evidence for association with ASD as per the SFARI database[30], while others have relatively little evidence in the literature. We included 36 papers related to these genes; papers were included only if they were successfully downloaded by Papers-Nexus and assigned to these genes by the gene flagger module.

We evaluated EAGLE-AI's performance across 85 candidate ASD cases extracted by manual curators from the 36 papers. The manual curators extracted the gene name, paper name, case ID, phenotype description, phenotype confidence level, sex of the case, age of the case, variant ID, zygosity of the variant, variant inheritance (i.e. de-novo or inherited), and variant impact (e.g. missense). The manual curators also assigned evidence scores to each of the cases using the EAGLE standard operating procedures (Table 1).

We measured EAGLE-AI's performance by manually matching EAGLE-AI's output to the output of the curators. We matched cases by case ID; in the absence of a matching case ID, we matched cases by phenotype description, age of the case, sex of the case, and/or paper name. We matched variants by rsID, ClinVar ID, HGVS NM number and coding position, HGVS NC or NG number and genome position, HGVS NP number and protein position, or chromosome:position:ref>alt notation relative to a reference genome. We designated matches as true positive (TP), false positive (FP), false negative (FN), or true negative (TN).

Sometimes, a case and variant extracted by the manual curators matched to a case extracted by EAGLE-AI, but EAGLE-AI was confidently wrong about the variant. Under these conditions, we counted the variant as both FP and FN. From the count of TPs, FPs, FNs, and TNs, we computed the precision, recall, and F1 score out of 100 percent for both the cases and variants (Table 3).

For each of the 85 cases, we measured the phenotype accuracy rating out of 100 points. If EAGLE-AI was correct about ASD status, it received +50 points. If EAGLE-AI was correct about ASD confidence level and reason, it received +20 points. If EAGLE-AI was correct about major co-occurring conditions, i.e. epilepsy, seizures, intellectual disability,

and developmental delay, it received +20 points. If EAGLE-AI was correct about detailed co-occurring conditions and testing results, e.g. other psychiatric disorders, electroencephalogram results, brain imaging results, micro or macrocephaly, or anatomical abnormalities, it received +10 points. Table 3 reports the mean phenotype accuracy rating as a percent.

For each of the eight genes, we also measured the gene-wise scoring error as per equation [1], where both $score_{EAGLEAI}$ and $score_{manual}$ are summed across all cases for a given gene. Table 3 reports the mean scoring error across the eight genes.

Data availability

Much of the data generated and analyzed during the current study are included in the published article, its supplementary information files, and the EAGLE-AI proof-of-concept data repository at https://github.com/juliandeanmoran/EAGLE-AI_poc_publicData. Additional EAGLE manual curation data used in training the models of the current study, beyond those data included in the supplementary information files, are too large to make publicly available; however they are available from the corresponding author upon reasonable request.

Code availability

Code for performance assessment is available at the public repository at https://github.com/juliandeanmoran/EAGLE-AI_poc_publicData. The underlying code and some of the training/test/validation datasets for the current study are not publicly available for proprietary reasons.

Acknowledgements

We thank all research authors whose data published under open access enabled development of EAGLE-AI. We also acknowledge support from the Northbridge Chair in Paediatric Research, the SickKids Psychiatry Associates Chair in Developmental Psychopathology, the University of Toronto McLaughlin Centre, the Canada Foundation for Innovation Major Sciences Initiative, and the SickKids Foundation.

Author contributions

V.F. developed Extractify, the gene flagger module, the search endpoint of Papers-Nexus, and all versions of EAGLE-AI core. He conducted bulk automated performance assessments of all EAGLE-AI core data extraction and scoring module versions. J.M. developed the PDF-miner endpoint of Papers-Nexus, conducted manual curation and follow-up performance assessment for agentic EAGLE-AI on eight novel genes, tested and debugged EAGLE-AI core and Extractify, assisted in automated performance assessment of the scoring modules, and was the principal writer of the manuscript. N.B.S. and O.R. manually curated EAGLE-AI's ground truth datasets and tested and gave feedback on EAGLE-AI. N.B.S. also assisted in the manual performance assessment of EAGLE-AI on eight genes. V.F., O.R., and N.H. contributed to writing the manuscript. A.W. contributed to developing Papers-Nexus. V.F., M.M.A., and W.E. initiated the project and conceptualized the framework. J.A.S.V. and S.W.S. provided supervision, funding acquisition, oversight, and guidance throughout the project. All authors read, gave feedback on, and approved the manuscript.

Competing interests

J.A.S.V. serves as a consultant for NoBias Therapeutics Inc. and has received speaker fees for Henry Steward Talks Ltd.. S.W.S. is on the Scientific Advisory Committees of Population Bio and Deep Genomics, and intellectual property from aspects of his research held at the Hospital for Sick Children are licensed to Athena Diagnostics and Population Bio. These relationships did not influence data interpretation or presentation during this study, but are still being disclosed for potential future considerations.

Table 1A: EAGLE scoring guidelines, genetic evidence

Non-splice site	Default score (range)	Splice site	Default score (range)
Variant is de novo	2 (0 – 2.5)	Variant is de novo, affects canonical splice site	2 (0 – 2.5)
Variant is de novo missense	0.5 (0 – 1)	Variant is de novo, affects non-canonical splice site	1 (0 – 1.5)
Variant is predicted or proven null (not de novo)	1.5 (0 – 2)	Variant affects inherited canonical splice site	1.5 (0 – 2)
Other variant type not predicted or proven null and not de novo	0.1 (0 – 0.1)	Variant affects inherited non-canonical splice site	0.5 (0 – 1)

EAGLE guidelines for scoring genetic evidence as submitted to agentic EAGLE-AI core's scoring workforce. Each scored variant can receive a maximum genetic score of 5.5.

Table 1B: EAGLE scoring guidelines, ASD phenotype evaluation

Confidence level	Indicators
High confidence	Expert clinician diagnosis, validated assessment methods, or explicit DSM or ICD criteria mentioned
Medium confidence	Description of symptoms in social / communicative and repetitive domains indicative of ASD
Low confidence	Mentions only "ASD", "autism", "autistic features", "autistic traits", etc.

If low confidence, apply deduction …		If medium confidence, apply deduction …	
genetic evidence score	deduction	genetic evidence score	deduction
2	-1	2	-0.5
1.5	-0.5	1.5	-0.25
1	-0.5	1	-0.25
0.5	-0.25	0.5	-0.1

EAGLE guidelines for scoring ASD phenotype evaluation as submitted to agentic EAGLE-AI core's scoring workforce. Each case receives a confidence level of high, medium, or low. If the confidence level is medium or low, a downstream algorithm performs a deduction to the genetic evidence score, depending on the initial genetic evidence score. If the case cannot meet the minimum confidence level (low), the case is excluded.

Table 1C: EAGLE scoring guidelines, experimental evidence

Evidence category	Evidence type	Suggested points default	Suggested points range	Max score
Function	Biochemical function	0.5	0 – 2	2
	Protein interaction	0.5	0 – 2	
	Expression	0.5	0 – 2	
Functional alteration	Patient cells	1	0 – 1	4
	Non-patient cells	0.5	0 – 4	
Models	Non-human model organism	2	0 – 4	4
	Cell culture model	1	0 – 2	
Rescue	Rescue in human	2	0 – 4	4
	Rescue in non-human model organism	2	0 – 4	
	Rescue in cell culture model	1	0 – 2	
	Rescue in patient cells	1	0 – 2	
Total allowable points for experimental evidence				6

EAGLE guidelines for scoring experimental evidence as submitted to agentic EAGLE-AI core's scoring workforce.

Table 2: EAGLE-AI core data extraction performance testing definitions

Case ID and Variant ID extractions	Exact match after normalization	True Positive
	Less than exact match	False Positive
	Absent variant ID	False Negative

Phenotype description extractions	Embedding cosine similarity ≥ 0.7	True Positive
	Embedding cosine similarity < 0.7	False Positive
	Absent description	False Negative

Definitions of true positive, false positive, and false negative used in computing precision, recall, and F1-score for case ID, variant ID, and phenotype description data extractions. We matched EAGLE-AI's data extractions to their corresponding manual ground truth using the patient/subject case ID. Variant IDs and case IDs were normalized for whitespace and punctuation. Phenotype descriptions extracted by both EAGLE-AI and the manual curators were converted to embeddings using the OpenAI text-embedding-3-large matrix. Matches were assessed in reference to ground truths produced by manual EAGLE curators.

Table 3: EAGLE-AI performance on 8 genes excluded from prior training, test, and validation sets

Total number of candidate ASD cases	85
Accuracy (cases)	78%
Accuracy (variants)	67%
Mean phenotype accuracy rating	71%
Average data extraction accuracy	72%
Gene-wise scoring error (SMAPE)	45%

Genes evaluated were CACNA1A, CASK, CDKL5, CHAMP1, CHD3, FMR1, GIGYF2, and GRIN1. 36 papers about these genes were evaluated. Mean phenotype accuracy rating was computed out of 100 points.

References

1. Satam, H. et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology* **12**, 997; https://doi.org/10.3390/biology12070997 (2023).
2. Stefanski, A., Calle-Lopez, Y., Leu, C., Perez-Palma, E., Pestana-Knight, E. Clinical sequencing yield in epilepsy, autism spectrum disorder, and intellectual disability: a systematic review and meta-analysis. *Epilepsia*. **62**, 143–151; 10.1111/epi.16755 (2020).
3. Selvanayagam, T. et al. Clinical utility of genome sequencing in autism: illustrative examples from a genomic research study. *J. Med. Genet.* **62**, 413–421; https://doi.org/10.1136/jmg-2024-110463 (2025).
4. Quick, V. S., Wang, B. & State, M. W.. Leveraging large genomic datasets to illuminate the pathobiology of autism spectrum disorders. *Neuropsychopharmacol.* **46**, 55–69; 10.1038/s41386-020-0768-y (2021).
5. Trost B. et al. Genomic architecture of autism from comprehensive whole-genome sequence annotation. *Cell* **185**, 4409–4427; 10.1016/j.cell.2022.10.009 (2022).
6. Tammimies, K. et al. Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. *JAMA* **314**, 895–903; doi:10.1001/jama.2015.10078 (2015).
7. Feliciano, P. et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genom. Med.* **4**, 19; 10.1038/s41525-019-0093-8 (2019).
8. Fu, J. M. et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331; https://doi.org/10.1038/s41588-022-01148-2 (2022).

9. Hoang, N., Buchanan, J. A. & Scherer, S. W. Heterogeneity in clinical sequencing tests marketed for autism spectrum disorders. *NPJ Genom. Med.* **3**, 27; 10.1038/s41525-018-0066-3 (2018).

10. Schaaf, C. P. et al. A framework for an evidence-based gene list relevant to autism spectrum disorder. *Nat. Rev. Genet.* **21**, 367–376; 10.1038/s41576-020-0231-2 (2020).

11. Srivastava, S. et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet. Med.* **21**, 2413–2421; 10.1038/s41436-019-0554-6 (2019).

12. Strande, N. T. et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the Clinical Genome Resource. *Am. J. Hum. Genet.* **100**, 895–906; 10.1016/j.ajhg.2017.04.015 (2017).

13. ClinGen Consortium. The Clinical Genome Resource (ClinGen): advancing genomic knowledge through global curation. *Genet. Med.* **27**, 101228; 10.1016/j.gim.2024.101228 (2025).

14. Riggs, E. R. *et al.* Clinical validity assessment of genes frequently tested on intellectual disability/autism sequencing panels. *Genet. Med.* **24**, 1899–1908; 10.1016/j.gim.2022.05.001 (2022).

15. Balzotti, M. et al. Clinical validity of expanded carrier screening: Evaluating the gene-disease relationship in more than 200 conditions. *Hum. Mutat.* **41**, 1365–1371; 10.1002/humu.24033 (2020).

16. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. (American Psychiatric Publishing, 2013).

17. Carter, M. T. & Scherer, S. W. Autism spectrum disorder in the genetics clinic: a review. *Clin. Genet.* **83**, 399–407; 10.1111/cge.12101 (2013).

18. Assuah, F. B. et al. A literature review of similarities between and among patients with autism spectrum disorder and epilepsy. *Cureus* **15**, e33946; 10.7759/cureus.33946 (2023).

19. Etyemez, S. et al. The role of intellectual disability with autism spectrum disorder and the documented co-occurring conditions: a population-based study. *Autism Res.* **15**, 2399–2408; 10.1002/aur.2831 (2022).

20. Vorstman, J. A. S. & Scherer, S. W. Contemplating syndromic autism. Genet. Med. 25, 100919; https://doi.org/10.1016/j.gim.2023.100919 (2023).

21. Vorstman, J. A. S. et al. Autism genetics: opportunities and challenges for clinical translation. *Nat. Rev. Genet.* **18**, 362–376; https://doi.org/10.1038/nrg.2017.4 (2017).

22. Fernandez, B. A., Marshall, C. R., Vorstman, J. A. S. & Scherer, S. W. Molecular testing in autism spectrum disorder. In *Diagnostic Molecular Pathology* (2nd edn), 291–301; https://doi.org/10.1016/B978-0-12-822824-1.00016-X (2024).

23. Kim, S.-W. et al. Evaluation of familial phenotype deviation to measure the impact of de novo mutations in autism. *Genome Med*. **17**, 93; 10.1186/s13073-025-01532-7 (2025).

24. Leblond, C. S. et al. Meta-analysis of *SHANK* mutations in autism spectrum disorders: a gradient of severity in cognitive impairments. *PLoS Genet*. **10**, e1004580; 10.1371/journal.pgen.1004580 (2014).

25. Sanders, S. J. et al. Progress in understanding and treating SCN2A-mediated disorders. *Trends Neurosci.* **41**, 442–456; 10.1016/j.tins.2018.03.011 (2018).

26. Litman, A. et al. Decomposition of phenotypic heterogeneity in autism reveals underlying genetic programs. *Nat. Genet.* **57**, 1611–1619; https://doi.org/10.1038/s41588-025-02224-z (2025).

27. SPARK Consortium. SPARK: a US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493; 10.1016/j.neuron.2018.01.015 (2018).

28. Yuen, R. K. et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).

29. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215; 10.1038/nature13772 (2014).

30. Abrahams, B. S. et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 36; https://doi.org/10.1186/2040-2392-4-36 (2013).

31. Shi, M. *et al.* A machine learning approach for the curation of biomedical literature. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings 25* 597–604 (Springer Berlin Heidelberg, 2003).

32. Varela-Vega, A., Posada-Reyes, A. & Méndez-Cruz, C. Automatic extraction of transcriptional regulatory interactions of bacteria from biomedical literature using a BERT-based approach. *Database* **2024**, baae094; 10.1093/database/baae094 (2024).

33. Chen, Q., Du, J., Allot, A. & Lu, Z. LitMC-BERT: Transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 2584–2595; 10.1109/TCBB.2022.3173562 (2022).

34. Wang, L. L. & Lo, K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Brief. Bioinform.* **22**, 781–799; 10.1093/bib/bbaa296 (2021).

35. Zhang, Y. *et al.* Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. *Syst. Rev.* **11**, 11; 10.1186/s13643-021-01881-5 (2022).

36. Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A. & Lenert, L. A. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J. Am. Med. Inform. Assoc.* **32**, 1071–1086; 10.1093/jamia/ocaf063 (2025).

37. Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M. & Naugler, C. Automated paper screening for clinical reviews using large language models: data analysis study. *J. Med. Internet Res.* **26**, e48996; 10.2196/48996 (2024).

38. Schmidt, L., Hair, K., Graziosi, S., Campbell, F., Kapp, C., Khanteymoori, A., Craig, D., Engelbert, M. & Thomas, J. Exploring the use of a large language model for data extraction in systematic reviews: a rapid feasibility study. 3rd Workshop on Augmented Intelligence for Technology-Assisted Review Systems (2024).

39. Hart, R. K. *et al.* HGVS Nomenclature 2024: improvements to community engagement, usability, and computability. *Genome Med.* **16**, 149; 10.1186/s13073-024-01421-5 (2024).

40. Wang, M. *et al.* hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 update. *Hum. Mutat.* **39**, 1803–1813; 10.1002/humu.23615 (2018).

41. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423; 10.1093/bioinformatics/btp163 (2009).

42. Cholewiak, S. A., Ipeirotis, P., Silva, V. & Kannawadi, A. SCHOLARLY: simple access to Google Scholar authors and citation using Python [software]. Version 1.5.1. https://github.com/scholarly-python-package/scholarly (2021).

43. National Library of Medicine. *PMC Open Access Subset*. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/ (2024).

44. Piwowar, H., Orr, R. & Priem, J. *Unpaywall*. https://unpaywall.org/products/api (n.d.).

45. Adobe Inc. *Adobe PDF Services API* [Software]. Adobe Inc., San Jose, CA. (2025).

46. openpyxl contributors. *openpyxl* [Software]. https://pypi.org/project/openpyxl/ (2025).

47. OpenAI. *o3-mini API* [Software]. OpenAI, San Francisco, CA. (2024).

48. OpenAI. *GPT-4o API* [Software]. OpenAI, San Francisco, CA. (2024).

49. OpenAI. *GPT-4o-mini API* [Software]. OpenAI, San Francisco, CA. (2024).

50. Li, G., et al. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *arXiv* 2303.17760 (2023).

51. CAMEL-AI contributors. *CAMEL-AI 0.2.36: Key Modules: Tools*. https://docs.camel-ai.org/key_modules/tools.html (2024).

52. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122; 10.1186/s13059-016-0974-4 (2016).

53. Belmadani, M. *et al.* VariCarta: A comprehensive database of harmonized genomic variants found in autism spectrum disorder sequencing studies. *Autism Res.* **12**, 1728–1736; 10.1002/aur.2200. (2019).

54. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894; 10.1093/nar/gky1016 (2019).

55. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814; 10.1093/nar/gkg509 (2003).

56. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **7**, Unit 7.20; 10.1002/0471142905.hg0720s76 (2013).

57. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443; 10.1038/s41586-020-2308-7 (2020).

58. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74; 10.1038/nature15393 (2015).

59. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985; 10.1093/nar/gkt1113 (2014).

60. Rehm, H. L. *et al.* ClinGen — the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242; 10.1056/NEJMsr1406261 (2015).

61. Allot, A. *et al.* Tracking genetic variants in the biomedical literature using LitVar 2.0. *Nat. Genet.* **55**, 901–903; 10.1038/s41588-023-01492-4. (2023).

62. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *arXiv* 1912.01703 (2019).

63. Zhang, Y., Chen, Q., Yang, Z., Lin, H. & Lu, Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data* **6**, 52.; 10.1038/s41597-019-0055-0 (2019).

64. Meta AI. *Llama 3.1–8B* [large language model]. https://huggingface.co/meta-llama/Llama-3.1-8B (n.d.).

65. Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N. & Lu, Z. BioRED: a rich biomedical relation extraction dataset. *Brief Bioinform.* **23**, bbac282; 10.1093/bib/bbac282. (2022).

66. HuggingFace. Transformer reinforcement learning (TRL) library [software]. https://huggingface.co/docs/trl/v0.21.0/en/index (n.d.).

67. Chaves, L. D., *et al*. Skewed X-chromosome inactivation in women with idiopathic intellectual disability is indicative of pathogenic variants. *Mol. Neurobiol.* **60**, 3758–3769; 10.1007/s12035-023-03311-0 (2023).

68. Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. On faithfulness and factuality in abstractive summarization. *arXiv* 2005.00661 (2020).

69. Lin, S., Hilton, J. & Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv* 2109.07958 (2022).

70. Bradski, G. The OpenCV Library. *Dr. Dobb's J. Software Tools* **25**, 120–125 (2000).

71. Richter, J.D., & Zhao, X. The molecular biology of FMRP: new insights into Fragile x Syndrome. *Nat. Rev. Neurosci.* **22**, 209-22; 10.1038/s41583-021-00432-0.

72. Bruford, E. A., Braschi, B., Denny, P., Jones, T. E. M., Seal, R. L. & Tweedie, S. Guidelines for human gene nomenclature. *Nat. Genet.* **52**, 754–758; 10.1038/s41588-020-0669-3 (2020).

73. Green, E. D., Watson, J. D. & Collins, F. S. Human Genome Project: twenty-five years of big biology. *Nature* **526**, 29–31; 10.1038/526029a (2015).

74. Trost, B., Loureiro, L. O. & Scherer, S. W. Discovery of genomic variation across a generation. *Hum. Mol. Genet.* **30**, R174–R186; 10.1093/hmg/ddab209 (2021).

75. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798; 10.1093/nar/gku1205 (2015).

76. Amadori, E. et al. Targeted re-sequencing for early diagnosis of genetic causes of childhood epilepsy: the Italian experience from the ‘beyond epilepsy’ project. *Ital. J. Pediatr.* **46**, 92; 10.1186/s13052-020-00860-1 (2020).