ARTICLE | Article in Press

An AI-powered trisomy 21 Research Assistant

Sutanu Nandi · Zenitha Sundararajan · Marc Subirana-Granés · Joaquin M. Espinosa · Milton Pividori · Kelly D. Sullivan · Matthew D. Galbraith · James C. Costello

Received: 30 June 2026 / Accepted: 5 November 2026
© The Author(s) 2026

Abstract
Down syndrome, caused by trisomy 21, increases diverse co-occurring conditions, and with nearly 35,000 publications indexed in PubMed as of late 2026, keeping pace with this literature is challenging. General-purpose large language models are widely used for information retrieval but rely on broad training data rather than specific evidence. Retrieval-augmented generation (RAG) improves rigor by linking outputs to source texts, but standard implementations rank background text as highly as experimental results. We developed the T21 Research Assistant, a section-aware RAG system that prioritizes Results sections to ground responses in primary experimental evidence. The system draws exclusively from 1,594 open-access original research articles on Down syndrome from PubMed Central, including 158 NIH INCLUDE-funded studies, and uses a multistage pipeline for query validation, retrieval, reranking, synthesis, and citation verification. Built on open-source NVIDIA-Nemotron models, it returns structured, cited responses. Across 30 expert-curated questions, no statistically significant differences were detected between the T21 Research Assistant and leading proprietary and open-source models or two published biomedical RAG baselines on most metrics. The T21 Research Assistant delivered traceable, citation-anchored answers and significantly outperformed its standalone generator model on Recall. The interface is accessible online, and the open-source platform can be deployed locally.

Keywords Down syndrome · Retrieval-augmented generation · Large language models · Biomedical literature mining · Section-aware retrieval

Introduction

Down syndrome (Trisomy 21, T21) affects nearly all systems in the human body and elevates risk for an array of co-occurring conditions (e.g., congenital heart defects, immune dysfunction, thyroid disorders, leukemia, autoimmune conditions, and early-onset Alzheimer disease, among many others)[1,2]. At the same time, individuals with T21 display a lower incidence of many solid tumors[1]. The research community uses both human clinical studies and experimental models, including mouse strains, iPSCs, and other cell lines, to investigate these conditions[1,3]. Studying the link between T21 and co-occurring conditions requires an interdisciplinary integration of a vast body of literature[4–8].

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1038/s41598-026-80099-8

Published online: 10 November 2026

scientific reports

Among the many active research fronts in this rapidly expanding field, significant contributions have advanced our understanding of the molecular, immune, and metabolic changes observed in T21, leading to new therapeutic opportunities[9–15]. Foundational genomic analyses established that T21 consistently activates the interferon transcriptional response across multiple cell types[9–12]. Building on these studies, multi-omics integration has uncovered coordinated immunometabolic dysregulation that stratifies individuals with T21 into clinically distinct subgroups[13], and recent multi-omic analysis of plasma from over 400 people identified perturbed bile acid metabolism and hepatic dysfunction as additional systemic consequences of T21[14]. Translating these molecular insights to clinical practice, a Phase II clinical trial of the JAK1/3 inhibitor tofacitinib demonstrated normalization of autoimmune biomarkers and clinical benefit across dermatological conditions in T21, establishing JAK/STAT signaling as a tractable pharmacological target[15]. Together, these contributions exemplify the breadth of ongoing T21 research and highlight the challenge of tracking advances across this multidisciplinary literature.

With nearly 35,000 PubMed-indexed publications on Down syndrome as of late 2026, it is a challenge to stay updated on research findings across this vast literature. In recent years, large language generative AI models (LLMs) have been widely adopted to answer queries across many domains. Generative AI models are trained on massive general-purpose texts and tend to generate responses biased toward commonly used words and phrases reflective of the broad and generalized model training, rather than drawing on the specific evidence in peer-reviewed Down syndrome research. A further concern is that LLMs are known to generate plausible but non-existent references; a recent audit of 2.5 million biomedical papers identified 4,046 fabricated citations across 2,810 publications, where the rate of fabrication increased 12-fold between 2023 and 2026[16]. Providing the model with relevant context directly from the Down syndrome literature helps to focus answers on what can be experimentally supported, with every claim traceable to the exact text of a public and peer-reviewed publication.

More broadly, agentic AI systems and large language models have been successfully deployed across diverse biomedical domains, including general-purpose multi-task biomedical reasoning[17], autonomous analysis of single-cell RNA sequencing data[18], agentic design of gene-editing experiments[19], self-verified gene-set analysis using curated biological databases[20], LLM-derived embeddings for single-cell biology[21], automated conversion of research papers into interactive AI agents[22], and end-to-end autonomous AI research pipelines[23]. These successes demonstrate that grounding AI systems in domain-specific data and structured agentic workflows substantially improves their reliability and utility for specialized scientific tasks.

Retrieval-augmented generation (RAG) connects a generative large language model to a curated database of domain-specific knowledge, limiting the response to the curated information at the time a query is made[24]. Rather than relying solely on what the model learned during training, the system first retrieves the most relevant context from the database and then asks the model to generate its answer using only those chunks as evidence. Thus, every claim in the response can be traced back to a specific paper and section, making the answer verifiable. RAG systems have been applied across several biomedical domains with measurable improvements in answer accuracy over standard language model baselines[25,26], but their deployment for the specialized condition of Down syndrome, where the literature is multidisciplinary and spans both human and model organism research, remains an emerging area of study and development.

Furthermore, RAG models that rely on simple text retrieval treat all sections of a research publication with the same weight. However, the structure of a manuscript is designed to convey different information, with the Introduction section being used to provide background and set up the research problem, the Methods section describes the experimental design and materials in detail, the Results section reports the observed and measurable findings objectively, and the Discussion section provides for interpretation, limitations, and broader implications. A sentence from a paper's Introduction that provides background context, such as *"Down syndrome is associated with increased Alzheimer disease risk"*, can score just as high in a standard similarity search as a sentence from the Results section reporting the actual measurements behind that association, even though only the second example contains direct experimental evidence. Prioritizing information from the Results, Discussion, and Methods sections will shape the answers generated from RAG models to answer questions with well-supported results, rather

than generalized knowledge statements, while simultaneously facilitating the precise validation and formal attribution of experimental observations.

Here, we present the T21 Research Assistant, a multi-agent RAG that retrieves information from a database of 1,594 open-access Down syndrome original research articles from PubMed Central, including 158 articles funded through the NIH INCLUDE (INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE) Project, ranks them by section priority, and generates structured responses with inline citations through a multilevel agentic pipeline covering retrieval, ranking, writing, citation checking, safety monitoring, LLM-as-Judge quality assessment, and quality review. We evaluated the T21 Research Assistant against a set of 30 positive-control questions (60 questions including negative controls) written by experts in Down syndrome research, each with an expert reference answer and explicit supporting PMIDs, using sentence-level adapted BERTScore (Precision, Recall, and F1) and document-level semantic similarity to measure how well the system's answers match the expert reference answers. The section-priority design and the evaluation framework introduced here are directly applicable to RAG system development in other specialized biomedical domains. The T21 Research Assistant is available at https://bioinformatics.cuanschutz.edu/t21-res-assi/, and the source code is available at https://github.com/snandi-DS/T21_Research_Assistant.

Results

The T21 Research Assistant is a section-aware, multi-agent retrieval-augmented generation (RAG) system that answers Down syndrome research questions using only evidence drawn from peer-reviewed publications (Fig. 1). A user query is embedded and matched against a Milvus[27] vector database of section-level chunks from 1,594 open-access PubMed Central articles (Supplementary Table S1). Articles were screened against predefined exclusion criteria (Supplementary Table S2). Candidate passages are then reranked and weighted by section priority so that Results, Discussion, Methods, and Conclusion text is favored over background framing. A synthesis agent generates a structured, citation-anchored response through an eight-stage agentic pipeline. The relative contribution of dense vector similarity and the cross-encoder rerank score during this step is governed by a single reranking parameter, α (Final Reranking Score $= \alpha \times$ vector_score + $(1-\alpha) \times$ cross_encoder_score). In the sections that follow, we characterize the corpus, determine the reranking configuration, and evaluate the system against expert reference answers, three comparator language models run standalone without a retrieval corpus (Nemotron-3-Super[28], GPT-5.4[29], and Claude Sonnet 4.6[30]), and two published RAG baselines (MedRAG[31] and PaperQA2[32]). The T21 Research Assistant and the two published RAG baselines used the identical document corpus and Nemotron-3-Super as the shared generator, enabling a controlled comparison of their retrieval and response-generation architectures. The three standalone language-model systems were evaluated separately to provide broader performance evaluations. Full architectural, pipeline, and provider-configuration details are provided in the Methods and Supplementary Methods.

A curated corpus of 1,594 Down syndrome original research articles spans the breadth of the field

The T21 Research Assistant corpus contains 1,594 open-access, structured full-text original research articles available through PubMed Central, yielding 65,677 section-level chunks (Table 1). Of these articles, 158 articles were funded through the NIH INCLUDE program and are tagged with INCLUDE provenance in the user interface.

Topic analysis of the 1,594 articles in the T21 Research Assistant corpus using OpenAlex[33] identified three dominant topics: Down Syndrome and Intellectual Disability Research (652 articles; 40.9%), Prenatal Screening and Diagnostics (269 articles; 16.9%), and Alzheimer's Disease Research and Treatments (41 articles; 2.6%). Additional high-priority clinical themes include Acute Myeloid Leukemia Research (25 articles; 1.6%) and Congenital Heart Disease Studies (24 articles; 1.5%). The remaining topics span genetics, neurodevelopment,

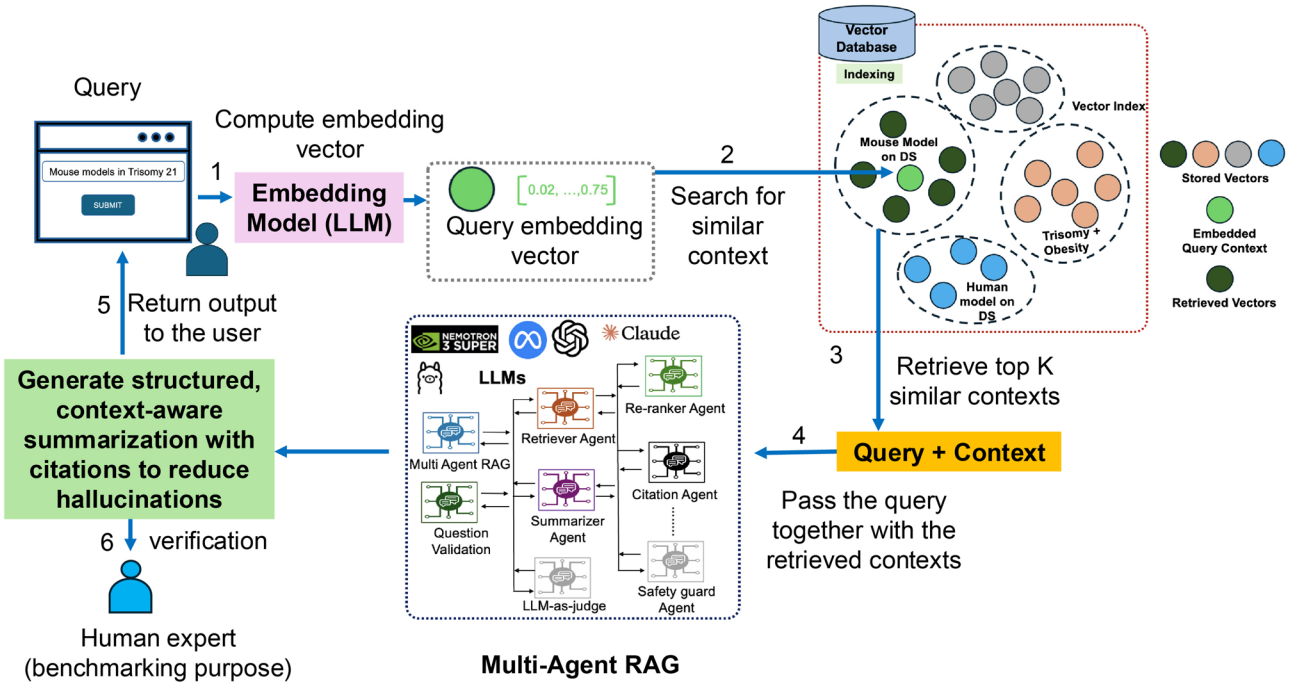


Fig. 1 Overview of the T21 Research Assistant multi-agent RAG workflow. Steps are numbered sequentially. (1) A researcher submits a natural language query through the browser interface. (2) The query is converted to a 2048-dimensional vector by the open-source Nemotron embedding model. (3) The query vector searches the 54,246 evidence-section chunks (Results, Discussion, Methods, and Conclusion) of the Milvus vector database (65,677 stored chunks in total; Abstract and Introduction chunks are stored but excluded from search). (4) The top k most similar chunks are retrieved. (5) The original query and retrieved context pass through the eight-stage multi-agent LangGraph pipeline: Question Validation → Retrieve → Rerank → Synthesize → Raw Evidence Extraction → Safety Guard → Finalize (citation hyperlinking; 150-word cap) → LLM-as-Judge (automated quality scoring and guided rewrite). The pipeline generates a structured, context-aware summary with inline citations [PMCxxxxxxx: Section] and returns the output to the user. (6) Human expert evaluation against reference answers is used for performance evaluation. Provider names within the multi-agent box denote the supported chat providers (NVIDIA Nemotron, Meta Llama, OpenAI GPT, and Anthropic Claude).

Table 1 T21 Research Assistant database separated by paper section (n = 65,677 total indexed vectors). Abstract, Introduction, and Other section chunks are excluded from search in line with the section-priority design.

Section	Chunks (n)	Percentage (%)	Indexed for search
Results	19,886	30.3	Yes
Methods	16,811	25.6	Yes
Discussion	16,268	24.8	Yes
Conclusion	1,281	2.0	Yes
Abstract, Introduction, and Other	11,431	17.4	No
Total	65,677	**100.0**	**N/A**

epigenetics, and immune dysregulation; the top 10 topics by article count are shown in Fig. 2; complete per-article topic assignments are provided in the Supplementary Table S1 corpus manifest.

Response quality is robust to the reranking parameter

To assess the reranking parameter α, we evaluated six values ($\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$) under three section-retrieval conditions: Abstract and Introduction only (C1, background context), the full six-section corpus (C2), and the evidence-only set of Results, Discussion, Methods, and Conclusion sections (C3, the production default),

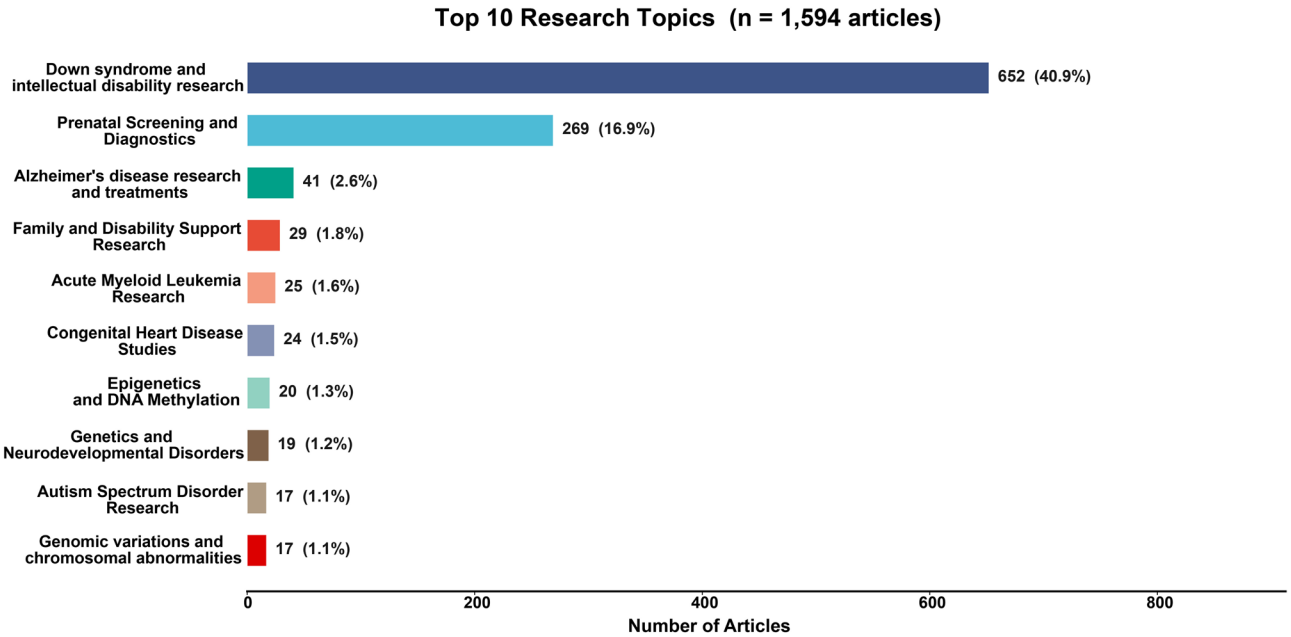


Fig. 2 Top 10 research topics across the T21 Research Assistant corpus ($n = 1{,}594$ articles). Topics were assigned using the OpenAlex topic taxonomy, which classifies each work using a machine-learning classifier trained on bibliographic metadata. Bar lengths indicate the number of articles per topic; percentages are calculated relative to all 1,594 articles.

on all 30 positive-control (Supplementary Table S3) questions (540 pipeline runs). Sentence-level adapted BERTScore F1 (Detailed description, Supplementary Methods) was largely flat across α at every condition, with overlapping standard deviations across the 30 questions. In the production evidence-only condition, performance was insensitive to α across 0.0–0.8 (F1 = 0.556–0.563; paired Wilcoxon between the extremes, $p = 0.95$), and pure vector retrieval without cross-encoder reranking (α = 1.0) produced the numerically lowest mean (F1 = 0.532), although the difference from α = 0.8 was not statistically significant (paired Wilcoxon, $p = 0.63$); α = 0.8, which lies within the insensitive range, was selected empirically from this sweep and used for all further model evaluations. The three retrieval conditions likewise performed within a narrow band, consistent with the interpretation that section-aware retrieval's primary contribution is evidence provenance rather than aggregate answer similarity (Fig. 3).

T21 Research Assistant performs within a narrow band of leading proprietary and open-source models

To evaluate the T21 Research Assistant's performance, we first compared against three standalone language models: Nemotron-3-Super[28](open-source, via Ollama), GPT-5.4[29] (proprietary), and Claude Sonnet 4.6[30] (proprietary). We evaluated all four systems on 30 positive control questions using sentence-level adapted BERTScore Precision, Recall, and F1, and Semantic Similarity (document-level, on full responses in Supplementary Table S4), all computed with the nvidia/llama-nemotron-embed-vl-1b-v2 (2,048-dimensional) embedding model[34] to ensure metric differences reflect response quality rather than embedding artefacts (Fig. 4). Aggregated across all 30 questions, the four systems performed within a narrow band on every metric, with no single system leading on all four. Together with the two published RAG baselines described below (MedRAG[31] and PaperQA2[32]), six systems in total are compared; Fig. 4 presents all six, while the statistics reported in the following paragraph concern the four language-model systems.

Among the four language-model systems, full-length, sentence-level adapted BERTScore Recall was highest for the T21 Research Assistant (0.574 ± 0.119), above GPT-5.4 (0.554 ± 0.123), Claude Sonnet 4.6 (0.547 ± 0.115), and Nemotron-3-Super standalone (0.527 ± 0.119). The advantage over Nemotron-3-Super standalone (the

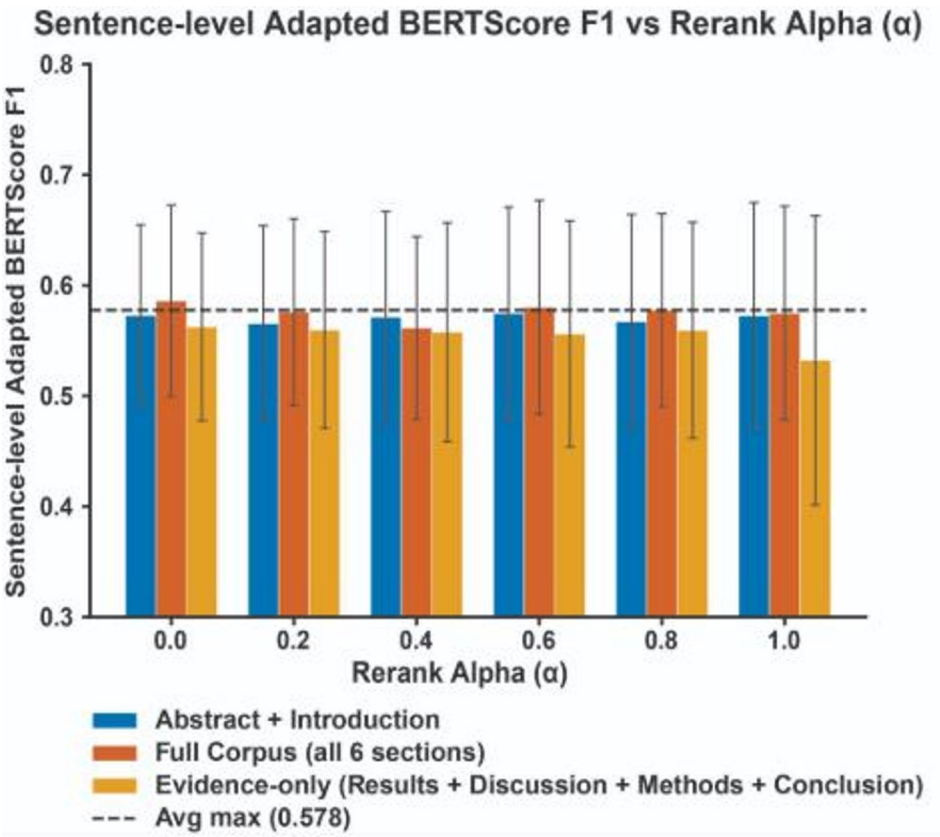


Fig. 3 Reranking parameter robustness. Sentence-level adapted BERTScore F1 across the rerank weight parameter α (α ∈ {0.0, 0.2, 0.4, 0.6, 0.8, 1.0}) for the T21 Research Assistant under three section-retrieval conditions: Abstract and Introduction only (C1, background context), the full six-section corpus (C2), and the evidence-only set of Results, Discussion, Methods, and Conclusion sections (C3, the production default). Grouped bars show mean sentence-level adapted BERTScore F1 for each retrieval condition at each alpha value; error bars indicate ± 1 SD across the 30 positive-control questions.

identical generator model run without the retrieval pipeline) was the only comparison significant after Bonferroni correction (two-sided Wilcoxon signed-rank, corrected $p = 0.008$), directly quantifying the contribution of corpus-grounded retrieval. Full-length, sentence-level adapted BERTScore F1 was highest for the T21 Research Assistant (0.558 ± 0.100), followed closely by GPT-5.4 (0.557 ± 0.104), Claude Sonnet 4.6 (0.556 ± 0.098), and Nemotron-3-Super standalone (0.533 ± 0.098). Sentence-level adapted BERTScore Precision was highest for Claude Sonnet 4.6 (0.573 ± 0.091), followed by GPT-5.4 (0.567 ± 0.091), Nemotron-3-Super standalone (0.550 ± 0.090), and the T21 Research Assistant (0.547 ± 0.097). Semantic Similarity was highest for GPT-5.4 (0.672 ± 0.078), followed by Claude Sonnet 4.6 (0.664 ± 0.081), the T21 Research Assistant (0.649 ± 0.090), and Nemotron-3-Super standalone (0.640 ± 0.086). No Precision, F1, or Semantic Similarity comparison remained significant after correction. The T21 Research Assistant led on Recall, the metric most directly tied to covering the key experimental findings in the expert reference, and was numerically best on F1, while trailing the proprietary models modestly on Precision and Semantic Similarity; across all four metrics, the absolute spread between the best and worst system was small (no more than 0.05 units), indicating that the open-source T21 Research Assistant performed in the same range as leading proprietary models on this performance evaluation. We note that the absence of statistical significance at $n = 30$ means no difference was detected, not that the systems are equivalent. The lower Precision and Semantic Similarity may reflect differences in wording and level of detail, and not necessarily in the findings conveyed: the expert reference answers state the main findings briefly, whereas the T21 Research Assistant, which retrieves evidence from the prioritized Results, Discussion, Methods, and Conclusion sections, tends to include more technical terminology and experimental detail from the source papers.

We additionally tested 30 negative-control questions that should not be answered because they are based on false premises, invented gene names (e.g., IFNAR9), general errors, and clinical trials or approvals that do not exist. The T21 Research Assistant returned an explicit refusal for 22 of the 30 questions (including one declined request for personal treatment advice) and explained each refusal, which follows from question validation and

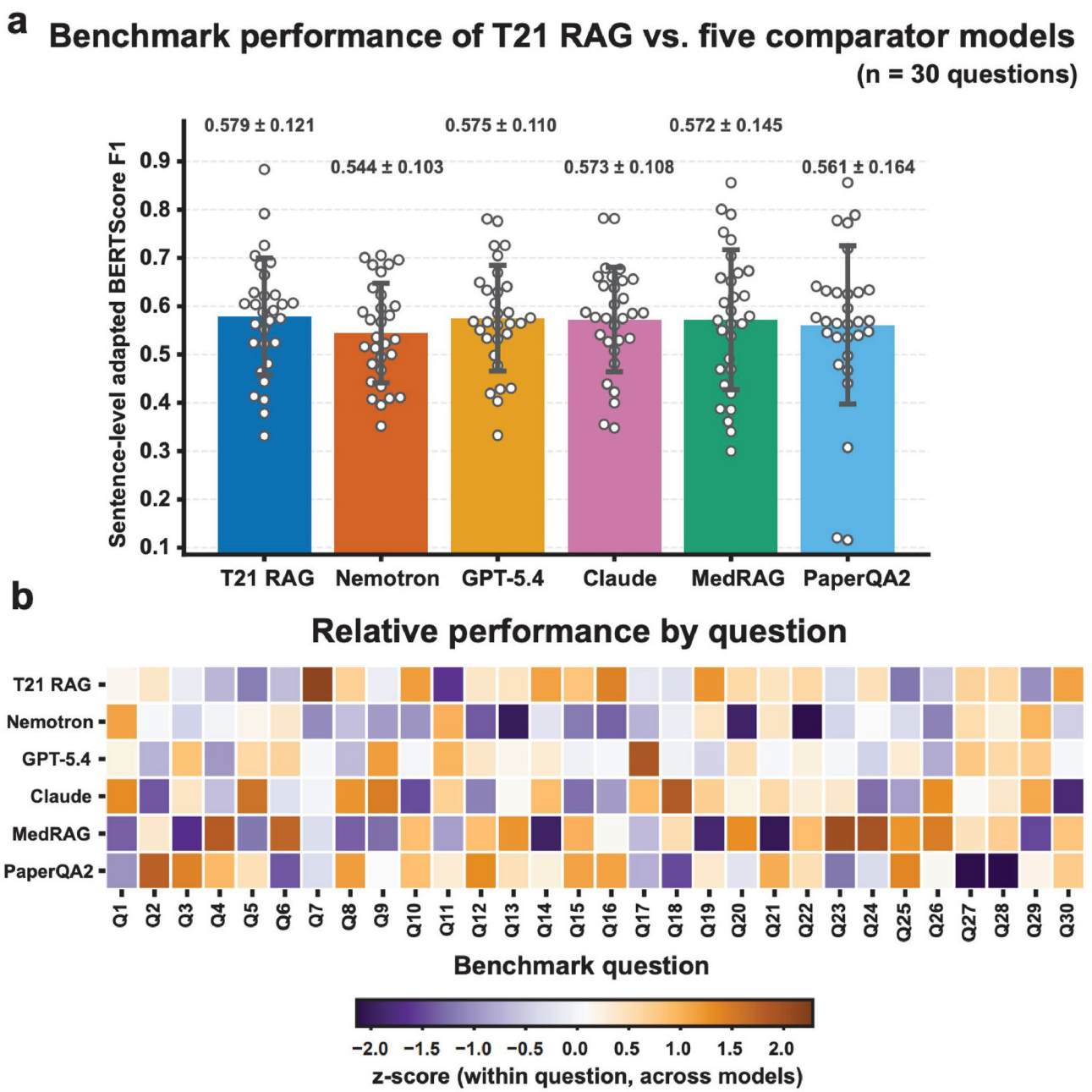


Fig. 4 Performance of the T21 Research Assistant against five comparator methods on 30 positive control questions. (**a**) Size-matched sentence-level adapted BERTScore F1 (mean ± SD; open circles show individual questions) for the T21 Research Assistant (0.579 ± 0.121), Nemotron-3-Super standalone (0.544 ± 0.103), GPT-5.4 (0.575 ± 0.110), Claude Sonnet 4.6 (0.573 ± 0.108), MedRAG (0.572 ± 0.145), and PaperQA2 (0.561 ± 0.164), all computed with the open-source nvidia/llama-nemotron-embed-vl-1b-v2 (2,048-dimensional) embedding model[34]. The T21 Research Assistant achieves the numerically highest mean size-matched F1 of the six methods. The only comparison significant after Bonferroni correction is full-length Recall for the T21 Research Assistant versus Nemotron-3-Super standalone (two-sided Wilcoxon signed-rank, corrected $p = 0.008$). (**b**) Relative model performance per question. For each of the 30 questions, size-matched F1 scores are z-score normalized by column (per question).

the requirement that every claim be grounded in retrievable evidence. Of the eight questions where a substantive response was returned, four explicitly noted that the queried entity, trial, or measurement was absent from the retrieved evidence, and four made spurious connections with genes or across model systems (i.e., mouse, human, canine). PaperQA2 refused 24 of 30 questions, returning a generic "insufficient information" message, while MedRAG refused to answer only 11 of 30 questions. The standalone language models rarely refused. GPT-5.4, Claude Sonnet 4.6, and Nemotron-3-Super answered nearly all 30, usually correcting the false premise, but sometimes accepting it. Nemotron-3-Super performed worst, inventing quantitative results on eight questions, including expression values for the fictional gene IFNAR9. Per-question responses are provided in Supplementary Table S5.

Published biomedical RAG baselines perform within the same narrow band without section-level provenance

To compare the T21 Research Assistant against published retrieval-augmented systems rather than base language models alone, we ran two controlled baselines over the identical document corpus and the identical generator model (nemotron-3-super), so that the arms differ only in their retrieval and answering machinery. MedRAG[31], a general biomedical RAG toolkit, chunks flat, unlabeled text at approximately 1,000 characters and retrieves with MedCPT (Contrastive Pre-trained Transformers[35], a biomedical dense retrieval model that serves as MedRAG's default retriever, run in its default configuration (the optional cross-encoder reranking and retriever fusion were not enabled), with no section labels, no section prioritization, and no multi-agent verification. PaperQA2[32], a RAG system built for full-text scientific papers, was additionally configured with the same embedding model used by the T21 Research Assistant, providing a controlled architectural comparison with a shared corpus, embedder, and generator. However, differences in document chunking, retrieval strategies, prompt templates, and response-generation workflows mean that this comparison evaluates the combined effects of these architectural choices rather than isolating the contribution of any single component. On the 30 positive-control questions, MedRAG achieved a sentence-level adapted BERTScore F1 = 0.554 (precision = 0.538, recall = 0.585) at a mean answer length of 280 words, and PaperQA2 achieved F1 = 0.525 (0.561 after size matching) at 133 words, against the T21 Research Assistant's F1 = 0.558 (0.579 after size matching) at 95 words; paired differences were not significant (two-sided Wilcoxon, uncorrected, $p = 0.46$ and $p = 0.24$; every model answered all 30 questions, $n = 30$). Full-length Semantic Similarity favored the baselines (MedRAG 0.679 and PaperQA2 0.659, versus 0.649 for the T21 Research Assistant; uncorrected $p = 0.045$ and $p = 0.036$), consistent with their longer answers; neither difference persisted after size matching ($p = 0.11$ and $p = 0.16$). After reference-guided size matching, which reduced differences in effective response length but did not eliminate potential selection bias, the T21 Research Assistant ranked numerically first among all six evaluated methods (Fig. 4a), with the per-question comparison shown in Fig. 4b. The published RAG baselines produced longer responses on average (280 words for MedRAG and 133 words for PaperQA2, compared with 95 words for the T21 Research Assistant), with differences in sentence-level adapted BERTScore Precision and Recall. Although greater response length may influence semantic coverage, the present evaluation does not establish a causal relationship between verbosity and these performance differences. Both baselines provide citations at the paper level rather than the section-level provenance offered by the T21 Research Assistant. Full baseline configurations, including every deviation from the deployed defaults, are described in the Supplementary Methods (Sect. 9).

Paper chunk embeddings cluster by section type and theme

To illustrate the complete retrieval and synthesis path on a single question, we selected question Q10 ("Are individuals with Down syndrome more prone to high blood pressure?") as a representative case. This question is asked in plain language, yet its evidence-based answer runs counter to the intuitive expectation. Individuals with Down syndrome typically exhibit lower, not higher, blood pressure. A question of this type therefore tests

whether a system reports what the literature shows rather than what general prior expectation would suggest. The expert ground truth answer and the T21 Research Assistant response share the core vocabulary of the question domain, with shared terms including "individuals", "Down syndrome", "blood pressure", and "lower" appearing in both (Fig. 5a).

To visualize the embedding space of the 1,594 corpus articles, along with the associated queries and retrieved chunks, a UMAP projection of all 65,677 section-level chunk embeddings was generated. The projection yielded a two-dimensional representation in which thematically related chunks clustered together (Fig. 5b). Chunks colored by OpenAlex topic form thematically coherent groupings, with the most frequent topics (e.g., prenatal screening and diagnostics, Alzheimer disease research, and acute myeloid leukemia research) occupying distinct regions of the space. Coloring by paper section shows that Results chunks ($n = 19,886$) and Discussion chunks ($n = 16,268$) are broadly distributed across the space, reflecting the diverse topics covered in the Down syndrome corpus. Methods chunks ($n = 16,811$) form denser, locally coherent clusters consistent with shared methodological language across studies. Conclusion chunks ($n = 1,281$) are sparse, occupying peripheral positions near the Results and Discussion cluster boundaries (Fig. 5c).

Retrieval and response placement for question Q10

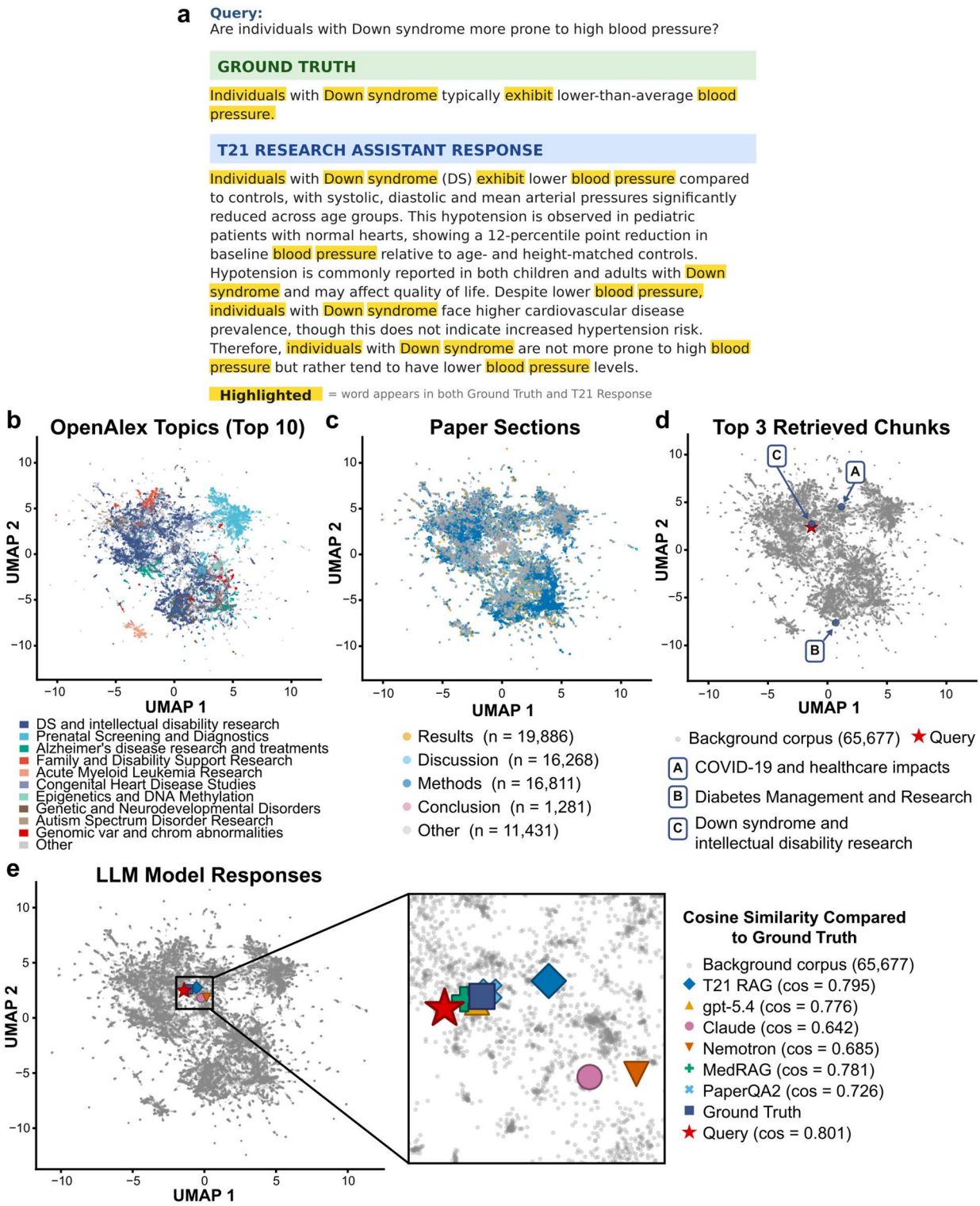
For question Q10, the top-3 retrieved chunks were projected onto the manifold and colored by Nemotron cross-encoder rerank score (Fig. 5d). The highest-scoring chunk (C) originated from the Results section of a study of blood pressure and body mass index in adults with Down syndrome[36]. Chunk B originated from the Results section of a study identifying carbonyl reductase 1 as a regulator of blood pressure in Down syndrome[37], and chunk A from the Results section of a pediatric COVID-19 cohort study reporting cardiovascular comorbidity in Down syndrome[38]. The query embedding lies within the retrieved chunk distribution, confirming that the open-source Nemotron embedding and reranking pipeline targeted the relevant region of the literature space.

The placement of all six method-generated responses relative to the expert ground truth in the same space is shown in Fig. 5e. The T21 Research Assistant response was the closest to the ground truth (cosine similarity 0.795), followed by MedRAG (0.781), GPT-5.4 (0.776), PaperQA2 (0.726), Nemotron-3-Super standalone (0.685), and Claude Sonnet 4.6 (0.642); the query itself achieved a cosine similarity of 0.801 to the ground truth. At the sentence level, the size-matched sentence-level adapted BERTScore F1 likewise ranked the T21 Research Assistant first among the six methods on this question (0.883 against 0.856 for the best comparator). Document-level proximity and sentence-level coverage therefore capture complementary aspects of answer quality, and the system's answer correctly reported the lower blood pressure documented in the retrieved evidence.

A structured, user-friendly interface links each response to its source text

The user interface through which the T21 Research Assistant delivers responses is itself an expression of the system's transparency principles (Fig. 6). Rather than returning a single text block, the T21 Research Assistant structures output across six expandable panels that show, step by step, which papers were retrieved, which passages were used as evidence, and how the answer was assembled from them. The Summary panel provides a 150-word narrative that integrates findings across retrieved papers with inline [PMCxxxxxxx: Results]-style hyperlinks anchoring each claim to its source document and section. The Key Evidence panel lists discrete evidence items labelled by section type (Results, Discussion, Methods), allowing users to verify the evidential basis of individual claims without consulting the full source papers. The Reasoning panel documents which evidence was selected and why, articulates the inferential chain connecting retrieved findings to the synthesized response, quantifies confidence relative to available evidence, and flags gaps and caveats. The Deep Dive panel provides an extended response without word limits for users who would like a more detailed response. The Highlighted Topic, MeSH Terms and Keywords from Top Retrieved Papers panel displays the OpenAlex[33] topic assignment, MeSH (Medical Subject Headings) controlled vocabulary terms, and author-assigned keywords from the top retrieved papers,

https://doi.org/10.1038/s41598-026-80099-8



supporting rapid appraisal of source relevance and thematic scope. The Export & Copy panel enables seamless export of the structured response into downstream workflows. Together, these panels enable researchers to critically assess AI-generated responses without requiring expertise in the underlying computational pipeline. Every citation link identifies and highlights the text that is referenced within the full-text article using browser text fragments. In addition, copy buttons are provided on each output text box.

Fig. 5 UMAP visualization of corpus structure, retrieval behavior, and response placement for positive control question Q10 ("Are individuals with Down syndrome more prone to high blood pressure?"). (**a**) The expert ground truth answer and the T21 Research Assistant-synthesized response for Q10; words appearing in both are highlighted. (**b**) Down syndrome PubMed Central corpus manifold colored by OpenAlex topic (top 10 of 219 topics shown); $n = 65{,}677$ section-level vectors. (**c**) Corpus space colored by paper section. (**d**) Background corpus (grey) overlaid with the top-3 retrieved chunks colored by open-source Nemotron cross-encoder rerank score and the query embedding (star). (**e**) Background corpus (grey) overlaid with the query (star), expert ground truth (square), and the six method-response embeddings. The boxed region is to resolve the overlapping response markers for visualization purposes; query cosine similarity to ground truth = 0.801.

Discussion

The T21 Research Assistant demonstrates that a section-aware, corpus-grounded RAG system can deliver traceable, citation-anchored answers to Down syndrome research questions. In this exploratory performance evaluation, the clearest quantitative finding is that corpus-grounded retrieval significantly improved sentence-level adapted BERTScore Recall over the identical generator model run without the retrieval pipeline; the contribution of section-aware retrieval itself lies in evidence provenance, determining which passages ground each answer and the section label carried by every citation, rather than in aggregate answer scores. A second observation is that the entirely open-source platform, incorporating NVIDIA Nemotron models for embedding, reranking, and synthesis, achieved performance within the range of the evaluated proprietary commercial models. The system also declined to answer the majority of negative-control questions that lacked literature support. Equally central to the system's utility is the user experience: structured output panels, section-level citations, and text-fragment deep links allow users to verify each claim against the exact supporting passage in the source article.

Beyond the primary findings, the four-system comparison illustrates another important point. In comparing performance metrics, a failure to reject the null using standard statistical comparisons is not evidence of equivalence, and scores that place two responses on a par do not establish that they carry the same evidential weight, which cannot be captured with a single performance metric. A response traceable to a specific result in the peer-reviewed literature can be checked at its source, while one that cannot be traced offers no such recourse. Similarity measures score surface agreement, not provenance. In this context, we interpret the lower Precision and Semantic Similarity as potentially reflecting a vocabulary mismatch between the technically dense Results-section text that the system retrieves and the concise, summary-style expert reference answers, which Down syndrome research experts wrote as brief summaries of the main findings of the supporting publications (each answer accompanied by explicit PMIDs; Supplementary Table S3), although differences in response structure and factual coverage may also contribute. The T21 Research Assistant therefore meets its primary design objective, which is to prioritize evidential coverage of primary experimental findings (BERTScore Recall) while showing no statistically detectable difference from leading proprietary models on Precision, F1, or document-level Semantic Similarity. That recall advantage cannot be attributed to the language model alone; it reflects the combined pipeline of section-priority retrieval, Nemotron cross-encoder reranking, LLM-as-Judge quality refinement, and structured Context-Content-Conclusion (C-C-C)[39] synthesis, although the present design does not distinguish between the contributions of each individual component.

Because the T21 Research Assistant retrieves from a defined, curated corpus of Down syndrome literature, every generated response is traceable to identifiable, peer-reviewed source documents, ensuring that credit is attributed to original research publications and that the boundaries of the system's knowledge are explicit and auditable. A citation-verification step confirms that every cited claim maps to a peer-reviewed source paper and section, and the supporting evidence is surfaced directly in the Key Evidence panel alongside each claim, so users can verify any reported claim against the original passage. This contrasts with proprietary models trained on opaque web-scale text sources, where the origin of encoded knowledge cannot be determined or cited.

Proprietary LLMs are subject to data-use agreements, financial costs, latency variability, and model update cycles that may be incompatible with institutional data governance requirements or the resource constraints common in academic research settings. By delivering strong evidential coverage with no statistically significant

performance differences detected relative to leading proprietary models on an evaluation of 30 questions, the T21 Research Assistant provides a replicable template for building open-source, cost-effective, and auditable biomedical literature tools. The open-source architecture enables methodological transparency through publicly available embedding, reranking, and synthesis models, including the embedding model (nvidia/llama-nemotron-embed-vl-1b-v2[34]), the cross-encoder (nvidia/llama-nemotron-rerank-vl-1b-v2[40]), and the synthesis model (Nemotron-3-Super[28]), together with version-controlled source code and documented configurations. These resources support independent implementation and evaluation of the system. However, exact reproduction of individual generated responses may be affected by inference-provider configurations, model-serving environments, and residual nondeterminism in hosted inference services, even when nominally deterministic generation settings are used.

The system has several limitations. The T21 Research Assistant is constrained by the papers it can access and by the capabilities of the synthesis model. Publications not available as open full text on PubMed Central, and publications outside the 1,594-publication corpus, cannot yet be retrieved regardless of their relevance. Answer quality varies with the chat model, and the performance results reported here were produced under standardized conditions with specific model versions (Nemotron-3-Super via Ollama, GPT-5.4, and Claude Sonnet 4.6, all accessed in September 2026). Quantitative similarity scores measure semantic agreement with expert reference answers, not independent factual correctness or citation fidelity. Future blinded expert preference rankings and direct human assessments could provide a more complete evaluation of answer quality.

In addition, the reranking parameter α was selected using the same 30 positive-control questions that constitute the final performance evaluation; no held-out question set was used. Although measured performance was insensitive to α across 0.0-0.8, the range containing the deployed value, and none of the comparator systems was tuned on these questions, this overlap may nonetheless favor the T21 Research Assistant, and the comparisons should be interpreted as exploratory rather than confirmatory.

Pipeline latency is a further practical consideration for the eight-stage architecture, and several optimizations are already implemented. Suppression of the reasoning model's hidden chain-of-thought at generation time reduced single synthesis calls from minutes to seconds, addressing the dominant latency factor for reasoning-class models; in-memory response caching keyed on the query makes repeated questions near-instant; and per-call timeouts cap worst-case latency. Future directions include parallelizing independent pipeline stages, caching intermediate retrieval and reranking results across related queries, and evaluating smaller distilled open-source models for the lower-stakes early stages (such as the question judge) while retaining larger models for final synthesis.

Keeping the corpus current is equally important given the pace of Down syndrome research. The corpus is maintained by an incremental ingestion pipeline that takes newly published PMIDs, applies the same eligibility gate as the original corpus build (publication-type exclusions, PMC full-text availability, and a dual retraction screen combining the MEDLINE publication-type flag with a Crossref check incorporating Retraction Watch), fetches and enriches metadata through the same OpenAlex, NCBI, Crossref, and PMC chain, and embeds and indexes new articles into the existing vector collection without a full rebuild. Each corpus refresh is written to a date-stamped, checkpointed build directory with a per-record screening audit, so successive corpus versions are directly comparable and citable. We intend to run this refresh every two months.

Taken together, these results show that a domain-specific, multi-agent RAG system can provide investigators with rapid, evidence-based access to peer-reviewed literature. This architecture addresses a critical challenge for the Down syndrome research community by making explicit, for every response, which papers and passages were retrieved and how the answer was assembled from them, all auditable through inline citations. With the ongoing growth of the underlying database, it provides a verifiable and scalable framework to support subsequent developments in autonomous retrieval and systematic quality assessment.

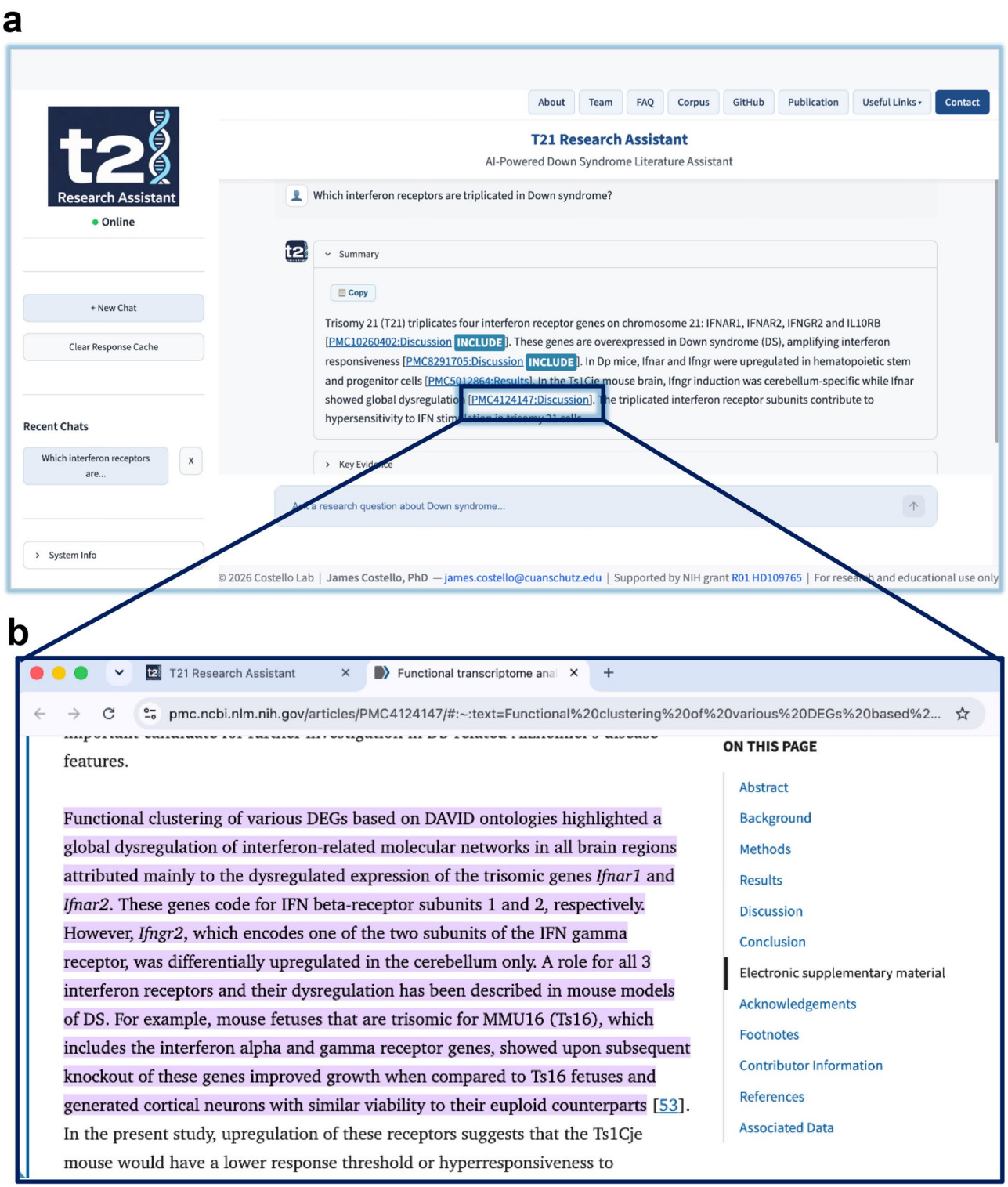


Fig. 6 Screenshot of the T21 Research Assistant web interface (Version 1.0; © 2026 Costello Lab, University of Colorado Anschutz) responding to the query "Which interferon receptors are triplicated in Down syndrome?" (**a**) The structured output is organized into six expandable panels: (i) Summary: a 150-word narrative with inline PubMed Central hyperlinks anchoring each claim to its source article and section; (ii) Key Evidence: discrete evidence items labelled by section type; (iii) Reasoning: inferential chain, confidence assessment, and caveats; (iv) Deep Dive: an uncapped extended response; (v) Highlighted Topic, MeSH Terms and Keywords from Top Retrieved Papers; and (vi) Export & Copy: structured response export. (**b**) The hyperlink brings the user to a new window with the text supporting the claim in the response highlighted in the referenced article.

Methods

System architecture

The T21 Research Assistant is built using LangChain (v0.3+)[41] for language model integration and LangGraph for deterministic pipeline orchestration (Fig. 1). The system comprises three integrated layers: (i) a data ingestion pipeline that retrieves full-text publications from PubMed Central and enriches them with metadata from the OpenAlex, NCBI efetch, and CrossRef APIs, segments full texts at section boundaries, and encodes them as dense vector representations; (ii) a RESTful web API that receives queries, executes the eight-stage, multi-agent RAG pipeline, and returns structured answers with inline citations; and (iii) a browser-based Streamlit interface for submitting natural language questions and reading annotated responses. The core computational stack uses open-source NVIDIA Nemotron models for embedding, reranking, and synthesis, and the open-weight gpt-oss:120b-cloud model for LLM-as-Judge scorin[28,34,40,42]. Queries are embedded with nvidia/llama-nemotron-embed-vl-1b-v2 (2,048-dimensional; the 4,096-dimensional nvidia/llama-embed-nemotron-8b is supported only as a legacy local alternative)[34], reranking uses an open-source Nemotron cross-encoder, and the default chat provider is Nemotron-3-Super. The same embedding model is used for all evaluation metrics and for UMAP visualization, so the entire retrieval, synthesis, and evaluation pipeline operates within a single unified open-source embedding space, with no dependency on proprietary commercial API subscriptions. The pipeline provides a generic provider interface, so proprietary models from OpenAI (GPT-5.4) and Anthropic (Claude Sonnet 4.6) can be configured as alternative synthesis providers; the performance evaluation reported here used only Nemotron-3-Super, GPT-5.4, and Claude Sonnet 4.6. Full implementation details of the architecture and provider configuration are given in the Supplementary Methods.

Document database and processing

The corpus comprises 1,594 open-access peer-reviewed original research articles on Down syndrome retrieved from PubMed Central (PMC) using a MeSH-based full-text search restricted to the PMC Open Access Subset. From 2,075 records identified by the query (snapshot of 2 September 2026), articles were retained when they passed the publication-type gate (case reports, comments, editorials, letters, and other non-research types excluded), carried a PMCID, were not retracted (screened against both the MEDLINE publication-type flag and the Crossref retraction registry, which incorporates Retraction Watch), had usable PMC full text, and yielded at least four of the six canonical sections, specifically Abstract, Introduction, Results, Methods, Discussion, Conclusions. In total, 1,891 manuscripts were eligible after the metadata gate, 1,837 with usable full text, and 1,669 with at least four sections. A final publication-type verification against the enriched OpenAlex and PMC metadata then removed 75 records that the initial MEDLINE-flag gate had passed but that were not original research articles (64 reviews, 8 preprints, 2 errata, and 1 other record), yielding the final corpus of 1,594 original research articles. For each retained article, structured metadata were collected from the OpenAlex, NCBI efetch, and PMC OAI-PMH APIs, and full-text XML was parsed into the previously listed six canonical sections. Among these articles, 158 were funded through the NIH INCLUDE program and are tagged with an INCLUDE label in the user interface. The corpus spans publication years 1970–September 2026. Articles were first segmented at section boundaries, and section texts were further split into overlapping sub-chunks using RecursiveCharacterTextSplitter from langchain-text-splitters (chunk size 1,200 characters, 200-character overlap, splitting on paragraph, then sentence, then word boundaries), so that every stored sub-chunk carries an unambiguous section label (mean = 41.2 sub-chunks per article). All chunks and associated metadata (PMCID, section name, MeSH terms, and bibliographic details) were stored in a Milvus vector database (IVF_FLAT index; nprobe = 64), yielding 65,677 indexed section-level chunks. The complete corpus manifest (one row per article) and the screening audit (one row per screened record, with the inclusion or exclusion criterion) are provided as Supplementary Table S1 (corpus manifest) and Supplementary Table S2 (screening) and are downloadable from the deployed interface.

Multi-level agentic pipeline

The T21 Research Assistant implements an eight-stage LangGraph state machine that routes each query sequentially through: Question Validation → Retrieve → Rerank → Synthesize → Raw Evidence Extraction → Safety Guard → Finalize → LLM-as-Judge (Fig. 1). Each incoming query first passes through an LLM classifier that labels it VALID or INVALID; this first-stage prompt instructs the model to determine whether the query is a Down syndrome or trisomy 21 research question, labelling in-scope queries VALID and out-of-scope queries INVALID; only valid Down syndrome queries proceed. Valid queries are embedded and searched against the Milvus index with per-section retrieval (12 candidate chunks fetched independently from each of the four active sections: Results, Discussion, Methods, and Conclusion), with a per-paper cap of two chunks per PMCID. The open-source Nemotron cross-encoder scores each candidate on a 0–100 scale, integrated with normalized vector similarity via Final Reranking Score = $\alpha \times$ vector_score + $(1-\alpha) \times$ cross_encoder_score, followed by a section-priority multiplier (1.0 for Results, Discussion, Methods, Conclusion; 0.0 for Abstract and Introduction). Top-ranked chunks are synthesized under a Context–Content–Conclusion structured prompt[39] (temperature = 0.0) into a 150-word response with inline [PMCxxxxxxx: Section] citations using Nemotron-3-Super[28]. Subsequent stages perform raw evidence extraction, safety screening, citation hyperlinking with a strict 150-word ceiling, and automated quality assessment by a separate LLM judge (gpt-oss:120b-cloud)[42] that triggers a guided rewrite when any criterion scores below 4 on a 1–5 scale. Full stage-by-stage specifications, prompts, and parameters are provided in the Supplementary Methods. Decoding uses temperature 0.0 with a fixed seed, which minimizes sampling variability but does not guarantee identical outputs across hosted API calls; exact repeatability within a deployment is provided by response caching.

Section-priority retrieval design

A key design decision was the exclusion of Abstract and Introduction chunks from the searchable index, since these sections restate established knowledge in a declarative style that closely mirrors query phrasing. Results, Discussion, Methods, and Conclusion sections receive a section priority weight of 1.0 and are eligible for retrieval; Abstract and Introduction sections receive a weight of 0.0 and are blocked at the database level (Table 1). The reranking parameter α, which controls the relative contribution of dense vector similarity and the cross-encoder score, was selected empirically by evaluating $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ under three section-retrieval conditions (Abstract and Introduction only, the full six-section corpus, and the evidence-only set of Results, Discussion, Methods, and Conclusion sections) across all 30 positive-control evaluation questions (Supplementary Table S3), with BERTScore F1 computed against matched expert reference answers. The same 30 positive-control questions were used for both this parameter sweep and the final evaluation; no separate held-out question set was used. The full α-sweep procedure and provider configuration details are described in the Supplementary Methods.

Evaluation metrics and statistical testing

System answers were evaluated against expert reference answers using an embedding-based framework. We implemented a sentence-level adapted BERTScore[43], computing Precision, Recall, and F1 from greedy cosine matching between candidate and reference sentences embedded with nvidia/llama-nemotron-embed-vl-1b-v2 (2,048-dimensional). Lexical overlap metrics such as ROUGE[44] and BLEU[45] were not used, as they penalise valid paraphrases in knowledge-intensive synthesis tasks. A document-level Semantic Similarity (cosine similarity between full response and full reference embeddings) was computed with the same model. To control for response length, a symmetric best-sentence extraction procedure normalized all evaluated systems to the ground truth word and sentence count before BERTScore computation. UMAP[46] (n_neighbors = 15, min_dist = 0.1, metric = cosine) was applied to all 65,677 chunk embeddings to visualize retrieval focus and answer placement. All metrics were computed on the 30 positive control questions (Supplementary Table S4); 30 negative control

questions (Supplementary Table S5) probing false premises and out-of-scope topics were evaluated for all six systems by classifying each response as an explicit refusal, a premise-correcting answer, or a premise acceptance. Each expert reference answer is accompanied by explicit supporting PMIDs (Supplementary Table S3). Pairwise comparisons used the two-sided Wilcoxon signed-rank test ($n = 30$; $\alpha = 0.05$), with results reported as mean ± SD. P-values for the three language-model comparisons are Bonferroni-corrected, because this is the comparison family from which significance is claimed. Comparisons against the two retrieval baselines are descriptive and are reported with uncorrected p-values, labeled as such, so that differences favoring the baselines are not masked by the correction (Supplementary Methods). Full definitions of the BERTScore formulas, the size-matching algorithm, embedding parameters, and the complete evaluation question set are provided in Supplementary Methods and Supplementary Table S3.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-80099-8.

Acknowledgements We thank members of the Costello Lab, as well as Lucas A. Gillenwater, Ph.D., and Casey S. Greene, Ph.D., for their valuable feedback on this work, and acknowledge the NIH INCLUDE Project and the Global Down Syndrome Foundation. During the preparation of this work, the authors used Anthropic Claude Sonnet 4.6 to improve the grammar and flow of the text; all content was subsequently reviewed and edited by the authors, who take full responsibility for the published work.

Author contributions S.N. conceived and designed the study, developed the software, performed the formal analysis and visualization, curated data, and wrote the original draft. Z.S. curated data and contributed to validation. M.S.G. contributed to validation and manuscript review and editing. J.M.E. contributed to manuscript review and editing and acquired funding. M.P. contributed to validation, manuscript review and editing, and acquired funding. K.D.S. curated data and contributed to validation and manuscript review and editing. M.D.G. curated data and contributed to validation, manuscript review and editing, and acquired funding. J.C.C. conceived the study, supervised the work, wrote the original draft, reviewed and edited the manuscript, and acquired funding. All authors reviewed the manuscript.

Funding This work was supported by the National Institutes of Health (NIH) under grant nos. HD109765 (J.C.C.), U2CHL156291 (M.D.G., J.M.E.), and HG011898 (M.P.).

Data availability The T21 Research Assistant is publicly available at https://bioinformatics.cuanschutz.edu/t21-res-assi/. The source code is available at https://github.com/snandi-DS/T21_Research_Assistant. An archival snapshot of the exact code version described in this manuscript is deposited in the Zenodo repository at: https://doi.org/10.5281/zenodo.23322961. All publications used in the corpus are open-access articles retrieved from PubMed Central. The corpus manifest, screening audit, and benchmark question set generated in this study are provided in the Supplementary Tables S1–S5. A preprint of this work is available on bioRxiv at https://doi.org/10.64898/2026.06.08.730893.

Code availability The complete source code for the T21 Research Assistant, including the data ingestion pipeline, the multi-agent RAG pipeline, the web interface, and the evaluation notebooks, is openly available at: https://github.com/snandi-DS/T21_Research_Assistant. An archival snapshot of the exact code version described in this manuscript is deposited in the Zenodo repository at: https://doi.org/10.5281/zenodo.23322961 (https://doi.org/10.5281/zenodo.23322961).

Declarations

Competing interests J.C.C. is a cofounder and Chief Science Officer of OncoRx Insights. The other authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you

will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Antonarakis, S. E. et al. Down syndrome. *Nat. Rev. Dis. Primers*. **6**, 9 (2020).
2. Santoro, J. D. et al. Neurologic complications of Down syndrome: a systematic review. *J. Neurol.* **268**, 4495–4509 (2021).
3. Russo, M. L., Sousa, A. M. M. & Bhattacharyya, A. Consequences of trisomy 21 for brain development in Down syndrome. *Nat. Rev. Neurosci.* **25**, 740–755 (2024).
4. Bull, M. J. et al. Health supervision for children and adolescents with down syndrome. *Pediatrics* **149**, e2022057010 (2022).
5. Weijerman, M. E. & de Winter, J. P. Clinical practice. The care of children with Down syndrome. *Eur. J. Pediatr.* **169**, 1445–1452 (2010).
6. Lott, I. T. & Head, E. Dementia in Down syndrome: unique insights for Alzheimer disease research. *Nat. Rev. Neurol.* **15**, 135–147 (2019).
7. Dierssen, M. Down syndrome: the brain in trisomic mode. *Nat. Rev. Neurosci.* **13**, 844–858 (2012).
8. Mégarbané, A. et al. The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of Down syndrome. *Genet. Med.* **11**, 611–616 (2009).
9. Sullivan, K. D. et al. Trisomy 21 consistently activates the interferon response. *Elife* **5**, (2016).
10. Waugh, K. A. et al. Triplication of the interferon receptor locus contributes to hallmarks of Down syndrome in a mouse model. *Nat. Genet.* **55**, 1034–1047 (2023).
11. Galbraith, M. D. et al. Multidimensional definition of the interferonopathy of Down syndrome and its response to JAK inhibition. *Sci. Adv.* **9**, eadg6218 (2023).
12. Powers, R. K. et al. Trisomy 21 activates the kynurenine pathway via increased dosage of interferon receptors. *Nat. Commun.* **10**, 4766 (2019).
13. Gillenwater, L. A. et al. Integrated analysis of immunometabolic interactions in Down syndrome. *Sci. Adv.* **10**, eadq3073 (2024).
14. Dunn, L. N. et al. Altered hepatic metabolism in Down syndrome. *Cell. Rep.* **45**, 116835 (2026).
15. Rachubinski, A. L. et al. JAK inhibition decreases the autoimmune burden in Down syndrome. *Elife* **13**, (2024).
16. Topaz, M., Roguin, N., Gupta, P., Zhang, Z. & Peltonen, L. M. Fabricated citations: an audit across 2·5 million biomedical papers. *Lancet* **407**, 1779–1781 (2026).
17. Huang, K. et al. Autonomous biomedical research with an artificial intelligence agent. *Science* **393**, eadz4351 (2026).
18. Alber, S. et al. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. *Nat. Methods*. https://doi.org/10.1038/s41592-026-03029-6 (2026).
19. Qu, Y. et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-025-01463-z (2025).
20. Wang, Z. et al. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nat. Methods*. **22**, 1677–1685 (2025).
21. Chen, Y. & Zou, J. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nat. Biomed. Eng.* 1–11 (2024).
22. Miao, J., Davis, J. R., Zhang, Y., Pritchard, J. K. & Zou, J. Paper2Agent: Reimagining research papers as interactive and reliable AI agents. https://doi.org/10.48550/arXiv.2509.06917 (2025).
23. Lu, C. et al. Towards end-to-end automation of AI research. *Nature* **651**, 914–919 (2026).
24. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. https://doi.org/10.48550/arXiv.2005.11401 (2020).
25. Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S. & Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digit. Health*. **4**, e0000877 (2025).
26. Liu, S., McCoy, A. B. & Wright, A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J. Am. Med. Inf. Assoc.* **32**, 605–615 (2025).
27. Retrieval-Augmented Generation (RAG) with Milvus and LangChain. https://milvus.io/docs/integrate_with_langchain.md.
28. NVIDIA Corporation. nemotron-3-super. https://ollama.com/library/nemotron-3-super (2026).
29. Introducing & GPT-5.4. *OpenAI* https://openai.com/index/introducing-gpt-5-4/.
30. Claude Sonnet. https://www.anthropic.com/claude/sonnet.

31. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. https://doi.org/10.48550/arXiv.2402.13178 (2024).
32. Skarlinski, M. D. et al. Language agents achieve superhuman synthesis of scientific knowledge. (2024).
33. Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. https://doi.org/10.48550/arXiv.2205.01833 (2022).
34. NVIDIA. nvidia/llama-nemotron-embed-vl-1b-v2 · Hugging Face. https://huggingface.co/nvidia/llama-nemotron-embed-vl-1b-v2 (2025).
35. Jin, Q. et al. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* **39**, btad651 (2023).
36. Pucci, F. et al. Blood pressure levels and body mass index in Brazilian adults with Down syndrome. *Sao Paulo Med. J.* **134**, 330–334 (2016).
37. Malbon, A. J. et al. Carbonyl reductase 1: a novel regulator of blood pressure in Down syndrome. *Biosci. Rep.* **45**, 157–170 (2025).
38. Leung, C. et al. Risk for severe illness and death among pediatric patients with Down syndrome hospitalized for COVID-19, Brazil. *Emerg. Infect. Dis.* **29**, 26–35 (2023).
39. Mensh, B. & Kording, K. Ten simple rules for structuring papers. *PLoS Comput. Biol.* **13**, e1005619 (2017).
40. NVIDIA. nvidia/llama-nemotron-rerank-vl-1b-v2 · Hugging Face. https://huggingface.co/nvidia/llama-nemotron-rerank-vl-1b-v2 (2025).
41. LangChain. Observe, evaluate, and deploy reliable AI agents. https://www.langchain.com/ (2022).
42. OpenAI. gpt-oss:120b-cloud. Ollama Library. https://ollama.com/library/gpt-oss:120b-cloud (2025).
43. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR 2020)* (2020).
44. Lin, C. Y. & ROUGE A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81 (2004).
45. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. (Association for Computational Linguistics, 2002).
46. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/arXiv.1802.03426 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sutanu Nandi[1] · Zenitha Sundararajan[2] · Marc Subirana-Granés[3] · Joaquin M. Espinosa[1,2] · Milton Pividori[3] · Kelly D. Sullivan[2,4] · Matthew D. Galbraith[1,2] · James C. Costello[1,2,3]

✉ James C. Costello
james.costello@cuanschutz.edu

[1] Department of Pharmacology, University of Colorado Anschutz, Mail Stop 8303, #L18-6114, 12801 East 17th Ave, Aurora, CO 80045, USA

[2] Linda Crnic Institute for Down Syndrome, University of Colorado Anschutz, Aurora, CO, USA

[3] Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO, USA

[4] Department of Pediatrics, Section of Developmental Biology, University of Colorado Anschutz, Aurora, CO, USA