PLOS Computational Biology

RESEARCH ARTICLE

La benchmarking large language models for extracting biobank-derived insights into health and disease

Manuel Corpas[1,2*], **Alfredo Iacoangeli**[3,4,5,6]

1 Life Sciences, University of Westminster, London, United Kingdom, **2** The Alan Turing Institute, London, United Kingdom, **3** Department of Biostatistics and Health Informatics, King's College London, London, United Kingdom, **4** Department of Basic and Clinical Neuroscience, King's College London, London, United Kingdom, **5** Perron Institute for Neurological and Translational Science, Perth, Western Australia, Australia, **6** Biomedical Research Centre, South London and Maudsley NHS Foundation Trust, London, United Kingdom

* m.corpas@westminster.ac.uk

Check for updates

OPEN ACCESS

Citation: Corpas M, Iacoangeli A (2026) La benchmarking large language models for extracting biobank-derived insights into health and disease. PLoS Comput Biol 22(4): e1014224. https://doi.org/10.1371/journal.pcbi.1014224

Editor: Wei Li, University of Maryland School of Medicine, UNITED STATES OF AMERICA

Received: March 26, 2025

Accepted: April 10, 2026

Published: April 20, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014224

Copyright: © 2026 Corpas, Iacoangeli. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium,

Abstract

Biobank-scale datasets such as the UK Biobank have become foundational resources for advancing biomedical discovery. Yet the complexity and heterogeneity of these resources, spanning genomics, imaging, clinical records, and metadata, pose substantial barriers to access and interpretation. Large Language Models (LLMs) offer a promising avenue for making such datasets more navigable through natural language interfaces. However, the extent to which current general-purpose LLMs can retrieve and synthesize biobank-specific insights has not yet been systematically evaluated. In this study, we present a reproducible, multi-metric evaluation framework to benchmark the capabilities of leading LLMs. We evaluated six leading large language models: Gemini 3 Pro, Claude Opus 4.5, Claude Sonnet 4, GPT-5.2, Mistral Large, and DeepSeek V3, on four benchmark tasks designed to assess biobank-related knowledge retrieval. We evaluate model performance across six dimensions (coverage, semantic accuracy, factual correctness, domain knowledge, reasoning quality, and biobank specificity) and assessed output consistency using curated UK Biobank references and a robust random baseline. All models outperformed the baseline by 16× to 25×, with strong statistical separation ($p < 0.001$), confirming meaningful biobank-specific knowledge retrieval. Gemini 3 Pro achieved the highest overall accuracy across tasks such as keyword synthesis, institution recognition, and topic inference, while Claude Sonnet 4 demonstrated the most uniform performance across evaluation dimensions. Our benchmark provides a rigorous framework for evaluating LLMs in biomedical settings. Using the UK Biobank as a real-world testbed, we highlight both the capabilities and limitations of current models, measuring their capacity to recall structured biomedical knowledge consistent with authoritative biobank metadata.

UNCORRECTED PROOF

provided the original author and source are credited.

Data availability statement: All "ground truth" data were retrieved from the UK Biobank Showcase (https://biobank.ndph.ox.ac.uk/showcase/). All scripts and data-processing routines used in this study are publicly accessible under MIT license via GitHub: https://github.com/manuelcorpas/LLM_4_UKB. This repository contains the code for benchmarking LLM, keyword identification, citation analysis, and the evaluation pipeline described in this paper. The results from LLM queries that were subsequently analysed for calculating our coverage metrics are available as a CSV files in the DATA directory of the repository.

Funding: A.I. is funded by South London and Maudsley NHS Foundation Trust, MND Scotland, Motor Neurone Disease Association, National Institute for Health and Care Research, Spastic Paraplegia Foundation, Rosetrees Trust, Darby Rimmer MND Foundation, the Medical Research Council (UKRI), Alzheimer's Research UK and LifeArc. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: MC is associated with Cambridge Precision Medicine Limited. All other authors have declared that no competing interests exist.

Author summary

The UK Biobank is one of the world's most comprehensive biomedical datasets, containing detailed genetic, clinical, and lifestyle information from about 500,000 participants. While this resource has enabled thousands of studies, its complexity makes it difficult to access and interpret - especially for non-specialists or those outside of computational biology. Large Language Models (LLMs), such as ChatGPT, Claude, and Gemini, have emerged as promising tools to bridge this gap by offering intuitive, language-based access to scientific information. However, it remains unclear how effectively they can retrieve and contextualize the nuanced knowledge embedded in biobank-scale datasets. In this study, we benchmarked the performance of six frontier LLMs on tasks derived from UK Biobank metadata and literature. We developed a novel evaluation framework that measures not just keyword coverage, but also semantic accuracy, factual correctness, reasoning quality, and domain-specific understanding. Our results show that while all models performed significantly better than chance, only a few demonstrated consistent, high-fidelity understanding of biobank-specific content. This work provides a rigorous and reproducible method for evaluating LLM knowledge alignment in biomedical settings and highlights both the promise and current limitations of these models in retrieving and synthesizing biobank-relevant information.

Introduction

Biobanks have become central to contemporary biomedical research, offering unprecedented scale and depth for studying population health, disease risk, and genotype-phenotype associations. Among them, the UK Biobank stands out as one of the most comprehensive and widely used resources, encompassing genomic, imaging, clinical, and lifestyle data from over 500,000 participants [1]. Its integration into thousands of scientific studies has made it a cornerstone of data-driven discovery in fields ranging from cardiometabolic diseases to neuropsychiatric traits [2–4].

Large Language Models (LLMs), with their ability to understand and generate human-like language [5], offer a potential means for literature mining, knowledge retrieval, and hypothesis generation [6]. However, while LLMs are increasingly used in biomedical applications, little is known about how well they perform in extracting or synthesizing information specific to biobank settings.

In this study, we address this gap by systematically benchmarking a set of leading LLMs, including GPT-5.2 [7], Claude Opus 4.5 and Claude Sonnet 4 [8], Gemini 3 Pro [9], Mistral Large [10], and DeepSeek V3 [11], on tasks derived from the UK Biobank's publications and application metadata. Our evaluation framework extends beyond keyword retrieval to encompass six dimensions: semantic accuracy, factual correctness, domain knowledge, reasoning quality, response depth, and biobank specificity. By benchmarking models across both structured and open-ended queries,

UNCORRECTED PROOF

and by systematically comparing their performance against a domain-specific random baseline, we capture not only their relative strengths and limitations but also the extent to which their outputs reflect meaningful, non-random knowledge synthesis. Improvement factors of 16× to 25× over random baselines, alongside statistically significant separation ($p < 0.001$), indicate that these LLMs retrieve biobank-relevant information with accuracy substantially exceeding chance, suggesting meaningful encoding of domain-specific knowledge during training.

Methods

Model selection and deployment strategy

To evaluate the capacity of general-purpose large language models (LLMs) to retrieve and synthesize biobank-related biomedical knowledge, we evaluated six leading large language models representing the current state-of-the-art as of January 2026. These models span a range of architectural families, training philosophies, and access paradigms. Included in our evaluation were Gemini 3 Pro (Google's latest multimodal flagship model), Claude Opus 4.5 (Anthropic's most capable reasoning model), Claude Sonnet 4 (Anthropic's balanced performance model), GPT-5.2 (OpenAI's latest generation model), Mistral Large (Mistral AI's flagship model), and DeepSeek V3 (DeepSeek's latest general-purpose model).

All models were accessed via their official APIs in January 2026. Each model was queried with identical prompts, and responses were collected without fine-tuning or custom system prompts to ensure fair comparison of base model capabilities. This allowed us to benchmark each model's general capability without the influence of task-specific optimization, reflecting how biomedical users typically engage with these systems in applied contexts.

To ensure comparability, each model was exposed to the same four queries, presented in identical phrasing and order. Responses were timestamped, archived in raw text and CSV format, and subsequently used for performance evaluation. Across all models, the integrity of the input-output pipeline was maintained to eliminate variability due to interface differences or prompt engineering.

Reference corpora and ground truth curation

All evaluation tasks were anchored in data derived from the UK Biobank's openly available metadata schemas [12]. Two key sources were used to construct domain-specific ground truth:

Schema 19 provided a corpus of 8,549 abstracts from peer-reviewed publications that utilized UK Biobank data. These abstracts span a wide range of biomedical domains, including cardiovascular, metabolic, neurological, and psychiatric research. We parsed this collection to extract the most frequent biomedical keywords, methodological terms, and phenotype descriptors. These included concepts such as "genome-wide association study," "cardiovascular disease," "prospective cohort," and "Mendelian randomization." Each concept was manually reviewed and mapped to a set of synonyms and lexical variants, enabling flexible matching to LLM outputs regardless of exact surface form.

Schema 27 contained metadata from 15,046 approved UK Biobank research applications. This dataset included investigator names, project titles, and institutional affiliations. From this corpus, we extracted the top 20 most prolific authors and top 10 most active institutions as reference standards for the authorship and institutional retrieval tasks, respectively.

Prompting and query standardization

To evaluate the models across a range of biomedical retrieval tasks, we formulated four standardized prompts reflecting common types of biobank-related queries. These were selected to span different forms of factual and inferential reasoning:

1. **Keyword Synthesis:** "What is the subject of the most commonly occurring keywords in UK Biobank papers?"
2. **Citation-Aware Topic Retrieval:** "What is the subject of the most cited papers relating to the UK Biobank?"
3. **Author Identification:** "Who are the top 20 most prolific authors publishing on the UK Biobank?"

UNCORRECTED PROOF

4. **Institutional Recognition:** "What are the top 10 leading institutions in terms of number of applications to the UK Biobank?"

The four benchmark questions were selected to represent complementary dimensions of biobank-related information retrieval tasks that researchers commonly perform. Specifically: (1) Keyword Synthesis (Q1) tests the model's ability to aggregate and summarize thematic content across a large corpus, a fundamental task in literature review and trend identification. (2) Citation-Aware Topic Retrieval (Q2) requires integrating bibliometric signals (citation counts) with content understanding, reflecting how researchers prioritize impactful literature. (3) Author Identification (Q3) evaluates entity recognition and ranking capabilities for person names, essential for identifying collaboration opportunities and key contributors to a field. (4) Institutional Recognition (Q4) assesses the model's ability to extract and rank organizational entities from application metadata, which is critical for understanding the research landscape and institutional engagement patterns. These questions span keyword analysis, bibliometric integration, person entity extraction, and organizational entity extraction, four distinct competencies that collectively represent the range of information needs in biobank research discovery. This multi-faceted approach ensures the benchmark evaluates both factual recall and inferential synthesis across different entity types and data sources.

These prompts were designed to be deliberately open-ended, requiring each model not only to retrieve information, but to prioritize relevance and, ideally, demonstrate some capacity for synthesis. In the case of the keyword and citation questions, high-performing models were expected to surface core biomedical themes; for the authorship and institution questions, success required access to correct entities and some factual grounding.

Each prompt was submitted to every model in an isolated context without any conversational history. Responses were automatically extracted and stored for scoring. When responses included bullet points, tables, or summaries, we parsed these structures to preserve semantic content during subsequent evaluation.

Scoring metrics and semantic matching framework

We developed a multi-tiered evaluation strategy that integrates both lexical and semantic similarity. Our primary metrics were Coverage Score, which measures the breadth of relevant concepts retrieved, and Weighted Coverage Score, which accounts for the frequency and salience of those concepts within the biobank literature.

To compute these metrics, we used the *all-MiniLM-L6-v2* SentenceTransformer model to encode both reference terms and model-generated outputs into vector embeddings. Cosine similarity was calculated between all term–response pairs, with a match recorded when similarity exceeded a threshold of 0.20. This threshold was chosen based on empirical tuning to balance precision and recall, and ambiguous matches were manually reviewed to ensure semantic validity.

The Weighted Coverage Score further incorporated frequency weights from Schema 19, assigning greater importance to high-impact terms such as "genome-wide association study," which appears over 1,900 times (Fig 1A). This allowed us to prioritize core biomedical themes over peripheral mentions. These metrics quantify not only how much relevant content each model retrieved, but also how central that content is to UK Biobank research.

Multidimensional evaluation of model capabilities

To assess interpretive competence beyond keyword retrieval, we developed a modular evaluation framework spanning six dimensions of large language model (LLM) performance on biobank-specific tasks. These dimensions were designed to capture semantic fidelity, reasoning depth, and domain-specific expertise relevant to biomedical research. All scores were calculated using custom Python scripts available in the project repository.

- **Semantic Accuracy:** Assessed whether retrieved biomedical concepts were not only semantically relevant but also appropriately contextualized. This score combined embedding-based cosine similarity (using SentenceTransformers) with expert manual review. Only correctly used biomedical concepts contributed to the final score.

UNCORRECTED PROOF

Fig 1. Benchmarking Results from UK Biobank Schema Data. (A) Publications by Year: Annual count of UK Biobank publications from Schema 19 (8,549 abstracts), showing growth from 2013 to 2025. **(B)** Top 15 Keywords: The most frequently cited keywords in UK Biobank papers, with "Humans" (6,547 occurrences) and demographic terms dominating, followed by methodological terms such as GWAS and Mendelian randomization. **(C)** Most Cited Papers: The ten most cited publications associated with the UK Biobank, with citation counts from Schema 19 metadata. Paper titles are truncated for display. **(D)** Top 15 Authors: The most prolific authors by publication count (e.g., George Davey Smith, 122 publications; Naveed Sattar, 119). **(E)** Top 10 Applicant Institutions: Leading institutions by number of approved UK Biobank research applications (Schema 27; 15,046 applications), with the University of Oxford (186 applications) ranking first.

https://doi.org/10.1371/journal.pcbi.1014224.g001

UNCORRECTED PROOF

- **Factual Correctness:** Evaluated entity-level precision in author and institution recognition by calculating the fraction of returned names matching those in the UK Biobank's Schema 27 metadata. Only verifiable matches influenced this score.
- **Reasoning Quality:** Scored based on the presence of interpretive statements, causal relationships, or thematic synthesis, as opposed to flat or unordered lists. Responses exhibiting inference and contextual reasoning received higher ratings.
- **Domain Knowledge:** Measured both the frequency and correct usage of advanced biomedical concepts (e.g., *polygenic score*, *longitudinal analysis*, *Mendelian randomization*), including whether they were explained clearly and used appropriately.
- **Response Depth:** Captured whether responses demonstrated layered insight by integrating multiple aspects of biobank-related evidence (e.g., linking methodological approaches to disease categories). Shallow or unstructured responses were penalized.
- **Biobank Specificity:** Quantified the degree to which outputs were explicitly grounded in UK Biobank content. General biomedical responses that lacked biobank-specific anchoring received lower scores.

Detailed scoring rubrics for each evaluation dimension are provided in S1 Table.

To assess performance stability, we calculated a Consistency Score, defined as a composite of the standard deviation and min-max range across all six metrics for a given model. Lower variability indicated more reliable performance across task types. These distributions are visualized in Fig 3E alongside semantic accuracy scores.

Baseline comparisons

To contextualize LLM performance and assess whether observed results reflect meaningful biobank-specific knowledge retrieval rather than random term association, we implemented a random baseline across all four tasks. For each task, we extracted the full set of relevant terms from UK Biobank metadata (keywords from publication abstracts (Schema 19), author names and paper titles, and institutional affiliations from approved applications (Schema 27)). Simulated baseline responses were then generated by randomly sampling from these term pools, matching the expected output length and formatting to resemble typical LLM completions.

We created 100 independent random samples per task to build empirical distributions of baseline performance. These outputs were scored using the same semantic matching and frequency-weighting pipeline applied to the LLM responses, specifically the Weighted Coverage Score metric (Table 1) This metric was chosen for the baseline comparison because it measures factual retrieval performance (the proportion of ground-truth entities retrieved, weighted by their frequency or citation impact) and can be computed identically for both coherent LLM outputs and randomly assembled term lists.

Table 1. Weighted Coverage Scores for Six Frontier LLMs Across Four UK Biobank Benchmark Tasks. Each cell reports the Weighted Coverage Score, which measures the proportion of ground-truth entities retrieved by each model, weighted by entity frequency or citation impact in UK Biobank metadata (Schemas 19 and 27). The Overall score is the mean Weighted Coverage Score across all four tasks.

Model	Keywords	Papers	Authors	Institutions	Overall
Gemini 3 Pro	0.87	0.21	0.50	0.99	0.643
Claude Sonnet 4	0.81	0.16	0.79	0.55	0.577
Claude Opus 4.5	0.81	0.16	0.79	0.55	0.577
Mistral Large	0.86	0.16	0.74	0.51	0.567
DeepSeek V3	0.73	0.00	0.49	0.85	0.517
GPT-5.2	0.73	0.00	0.19	0.90	0.455

https://doi.org/10.1371/journal.pcbi.1014224.t001

UNCORRECTED PROOF

The six multidimensional qualitative measures (Semantic Accuracy, Factual Correctness, Domain Knowledge, Reasoning Quality, Response Depth, and Biobank Specificity) were not applied to the random baseline, as several of these dimensions, particularly Reasoning Quality and Response Depth, require coherent, structured text with logical argumentation and thematic synthesis, properties that randomly sampled term lists lack by construction. The resulting Weighted Coverage Score distributions served as null models for statistical testing.

Results

Ground truth results

First, we present the analysis of the results that we obtained by parsing both Schema 19 (UK Biobank's 8,549 abstracts) and Schema 27 (UK Biobank's 15,046 research applications). Fig 1 (A, B, C, D) shows the results of calculating the answers from our four prompting questions.

Analysis of top keywords in publications

The most prevalent keyword was "Humans", appearing 6,547 times, followed by "Female" (3,469), "Male" (3,277) and "Middle Aged" (2,774) (Fig 1A). These results suggest a strong emphasis on human-related studies, with a particular focus on sex and age demographics.

Geographical representation was also notable, with "United Kingdom" ranking among the top five (2,689 occurrences). Key methodological terms such as "Genome-Wide Association Study" (1,940), "Mendelian Randomization Analysis" (751), and "Prospective Studies" (1,304) indicated a research focus on genetic epidemiology and population-based cohort analyses.

Health-related keywords, including "Cardiovascular Diseases" (711) and "Diabetes Mellitus, Type 2" (544), suggest a strong interest in chronic disease genetics. Other notable terms such as "Risk Factors" (2,264), "Genetic Predisposition to Disease" (1,336) and "Multifactorial Inheritance" (508) further highlight the relevance of polygenic risk and complex trait analyses in these studies.

Most cited articles related to the UK biobank

An analysis of the most cited research articles associated with the UK Biobank (Fig 1B) highlights key areas of scientific interest and impact. The highest-cited publication is on *Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population* (published in the *American Journal of Epidemiology*) [13], which has accumulated 2,548 citations (according the UK Biobank's metadata schema), underscoring the significance of demographic and health-related factors in large-scale biobank studies.

Studies leveraging genome-wide association studies (GWAS) dominate the citation rankings, reflecting the widespread use of the UK Biobank for uncovering genetic risk factors. The second most cited paper, *Genome-wide association analyses identifying 44 risk variants* (*Nature Genetics*) [14], has 2,419 citations, followed closely by research on *Genome-wide polygenic scores for common diseases* (*Nature Genetics*, 2,291 citations) [15].

The impact of body composition on mortality is also a prominent topic, with *Body-mass index and all-cause mortality* (*The Lancet*) accumulating 1,965 citations. Similarly, *Gene discovery and polygenic prediction from a GWAS* (*Nature Genetics*) [16] has received 1,958 citations, reinforcing the role of biobank-scale datasets in predictive genomics.

Mental health research has also gained significant attention, as seen in *Genome-wide meta-analysis of depression* (*Nature Neuroscience*, 1,843 citations) [17], which highlights the genetic basis of psychiatric conditions. Additionally, *Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk* (*Nature Genetics*, 1,777 citations) [18] and *Meta-analysis of GWAS for height* (*Human Molecular Genetics*, 1,710 citations) [19] further illustrate the diversity of genomic research leveraging the UK Biobank.

UNCORRECTED PROOF

Beyond genetics, neuroimaging and brain-related studies are also represented, with *Multimodal population brain imaging in the UK Biobank prospective study* (*Nature Neuroscience*) accumulating 1,577 citations [20]. Finally, *Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies* (*The Lancet Neurology*) has 1,548 citations [21].

Authorship and institutional contributions to UK biobank research

The analysis of publication output reveals the most prolific researchers contributing to studies leveraging the UK Biobank (Fig 1C). George Davey Smith ranks as the most published author, with 122 publications, followed by Naveed Sattar (119), Kari Stefansson (105), and Caroline Hayward (94). Several other notable researchers, including Stephen Burgess (93), Wei Cheng (92), and Carlos Celis-Morales (79), also feature prominently, indicating strong engagement in biobank-based research from diverse fields such as epidemiology, genomics, and cardiometabolic health.

In terms of institutional engagement (Fig 1D), the University of Oxford leads with 186 applications, reflecting its central role in UK Biobank-related studies. The University of Cambridge (74 applications) and Imperial College London (69 applications) follow, showcasing their strong contributions to biobank-driven investigations. Other major UK institutions, such as University College London (69), University of Edinburgh (62), and University of Manchester (61), demonstrate widespread institutional involvement.

Interestingly, UK Biobank Ltd itself appears as a key applicant with 60 applications, likely reflecting in-house research and collaborative projects. Notably, international representation is observed with Sun Yat-Sen University (39 applications), indicating global interest in UK Biobank data.

LLM performance across four biobank-relevant tasks

Fig 2 summarizes the performance of all tested Large Language Models (LLMs) in retrieving information about UK Biobank research from our reference corpora. We evaluated four specific tasks (covering keywords in publications, top cited papers, top authors, and top applicant institutions) using both Coverage Score (breadth of matched concepts) and Weighted Coverage Score (concepts weighted by their relative importance or frequency). In the case of a score of 0.0 this means the model was not able to retrieve any of the "ground truth" results. We aggregated each model's performance across the four tasks to produce an overall benchmark ranking.

Keyword retrieval

Fig 2A compares each model's ability to identify the top 20 most frequent keywords in the UK Biobank literature. Gemini 3 Pro and Mistral Large achieved the highest keyword coverage (0.80), each identifying 16 of the top 20 keywords, with weighted scores of 0.87 and 0.86 respectively. The Claude models (Opus 4.5 and Sonnet 4) followed closely with coverage of 0.75 (weighted 0.81), identifying 15 keywords. GPT-5.2 and DeepSeek V3 both achieved 0.60 coverage (weighted 0.73), identifying 12 keywords each.

Top cited papers

Fig 2B displays results for identifying the most highly cited articles based on our curated list. All models struggled with this challenging task. Gemini 3 Pro performed best with a coverage score of 0.20 (weighted 0.21), identifying 2 of the top 10 most cited papers. Claude Opus 4.5, Claude Sonnet 4, and Mistral Large all achieved identical coverage of 0.20 (weighted 0.16), also identifying 2 papers each. GPT-5.2 and DeepSeek V3 failed to identify any of the top cited papers (coverage 0.00).

Most prolific authors

Fig 2C evaluates retrieval of the top 20 UK Biobank authors by publication count. Mistral Large 2 led on both coverage (0.70) and weighted coverage (0.69), correctly listing many of the highest-output authors such as George Davey Smith,

UNCORRECTED PROOF

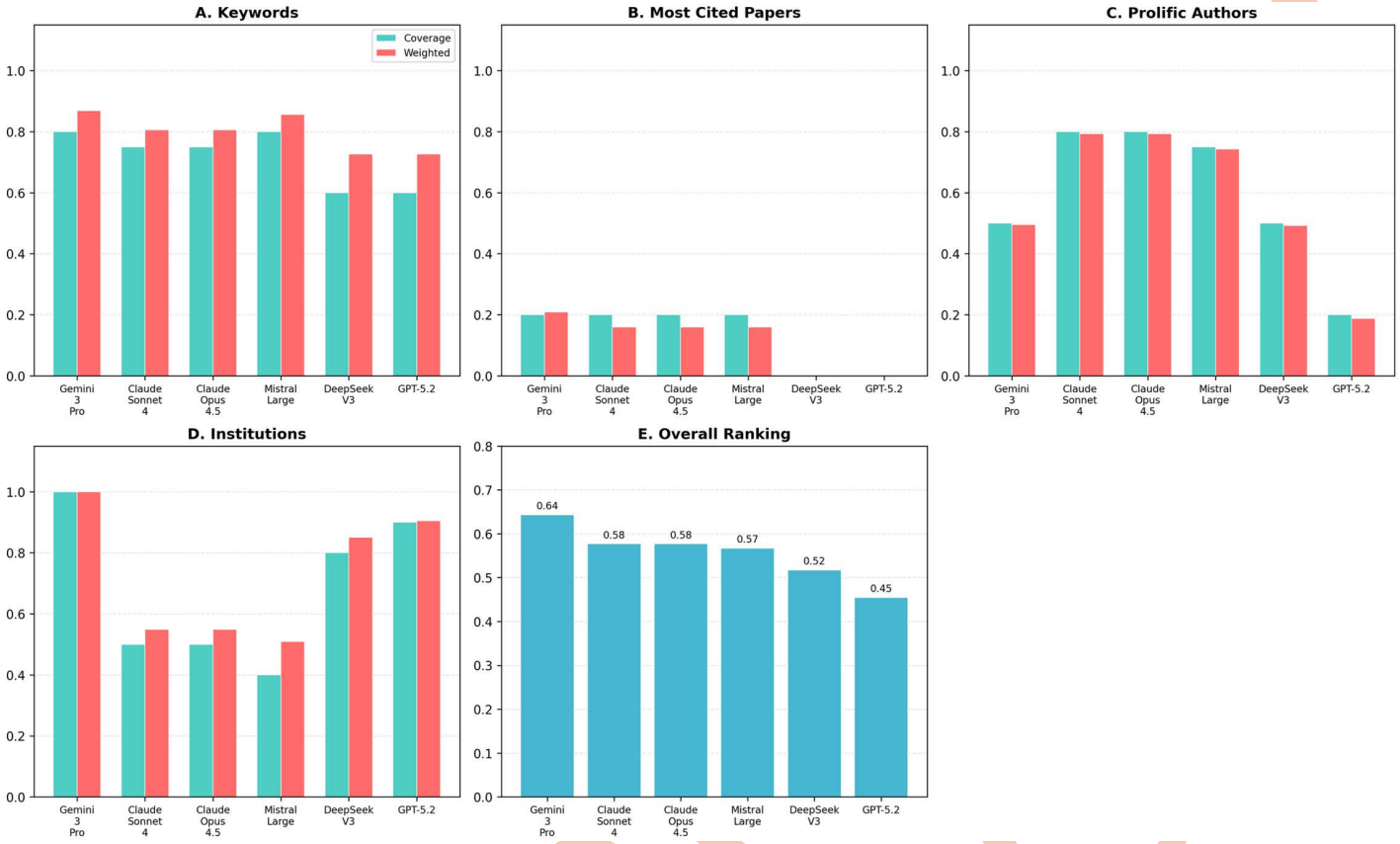


Fig 2. Benchmark Performance of Six Frontier LLMs on UK Biobank Knowledge Retrieval Tasks (January 2026). (A) *Keywords:* Keyword recognition benchmark comparing Coverage Score (proportion of top 20 keywords matched) and Weighted Coverage Score (frequency-weighted) across all six models. Gemini 3 Pro and Mistral Large achieved highest coverage (0.80), followed by Claude models (0.75). GPT-5.2 and DeepSeek V3 scored 0.60. **(B)** *Most Cited Papers:* All models struggled with this challenging task. Gemini 3 Pro scored highest (Coverage 0.20, Weighted 0.21), while GPT-5.2 and DeepSeek V3 failed to identify any matching papers (0.00). **(C)** *Most Prolific Authors:* Each model's output is assessed for matching the top 20 authors by publication count. Coverage Scores capture how many authors each LLM mentions, while Weighted Coverage Scores emphasize authors with more total publications. **(D)** *Applicant Institutions:* Shows performance in detecting the 10 most frequent UK Biobank applicant institutions. Scores are weighted according to the institution's application counts. **(E)** *Overall Ranking:* Overall weighted coverage ranking across all four tasks. Gemini 3 Pro leads (0.643), followed by Claude Sonnet 4 and Claude Opus 4.5 (0.577 each), Mistral Large (0.567), DeepSeek V3 (0.517), and GPT-5.2 (0.455).

https://doi.org/10.1371/journal.pcbi.1014224.g002

Naveed Sattar, and Kari Stefansson. Claude 3.5 Sonnet followed, reaching around 0.60 coverage, while Gemini 3 Pro captured a more moderate 0.45 coverage

Top applicant institutions

Fig 2D focuses on the top 10 institutions with the highest number of UK Biobank applications. Gemini 3 Pro achieved perfect institutional coverage (1.00), correctly identifying all top 10 applicant institutions. GPT-5.2 followed with 0.90 coverage (weighted 0.90), identifying 9 institutions. DeepSeek V3 achieved 0.80 coverage (weighted 0.85), identifying 8 institutions. The Claude models (Opus 4.5 and Sonnet 4) scored 0.50 coverage (weighted 0.55), identifying 5 institutions each. Mistral Large scored lowest on this task with 0.40 coverage (weighted 0.51), identifying 4 institutions.

UNCORRECTED PROOF

Overall accuracy ranking

We computed each model's average Weighted Coverage across all four tasks to generate an overall ranking (Fig 2E, Table 1). Across the four benchmark tasks, Gemini 3 Pro achieved the highest overall weighted coverage score (0.643), followed by Claude Sonnet 4 and Claude Opus 4.5 (both 0.577), Mistral Large (0.567), DeepSeek V3 (0.517), and GPT-5.2 (0.455). The overall coverage scores (unweighted) followed a similar pattern: Gemini 3 Pro (0.625), Claude models (0.563), Mistral Large (0.538), DeepSeek V3 (0.475), and GPT-5.2 (0.425).

Semantic fidelity and interpretive competence across models

To rigorously assess the interpretive capabilities of large language models (LLMs) in the context of biobank-specific information retrieval, we implemented a multidimensional evaluation framework encompassing six orthogonal performance dimensions: semantic accuracy, reasoning quality, domain knowledge, factual correctness, response depth, and biobank specificity. These dimensions were designed to interrogate not only lexical alignment, but also the structural and inferential characteristics of model outputs.

Fig 3A presents a radar chart capturing each model's relative performance across all six dimensions. Gemini 3 Pro demonstrated the most uniformly high scores, suggesting robust semantic generalization, domain fluency, and task-specific grounding. Mistral Large and the Claude models exhibited strong performance in reasoning and methodological coherence but were comparatively less stable across factual and biobank-specific metrics. GPT-5.2 and DeepSeek V3 showed more heterogeneous patterns, achieving moderate scores in some dimensions (e.g., response depth or domain knowledge) but faltering in others.

Fig 3B isolates three core metrics (semantic accuracy, reasoning quality, and domain knowledge) and presents absolute performance values. Each model showed distinct strengths across dimensions. Gemini 3 Pro achieved the highest semantic accuracy (0.79), reflecting strong alignment between model outputs and reference content. Claude 4 led in reasoning quality (0.68), followed by Mistral Large (0.64), suggesting effective thematic synthesis and interpretive structuring. Claude Opus 4.5 demonstrated the strongest domain knowledge (0.74), indicating appropriate contextualization of methodological constructs (e.g., Mendelian randomization, longitudinal designs). DeepSeek V3 and GPT-5.2 showed weaker reasoning quality scores, producing more variable outputs across dimensions.

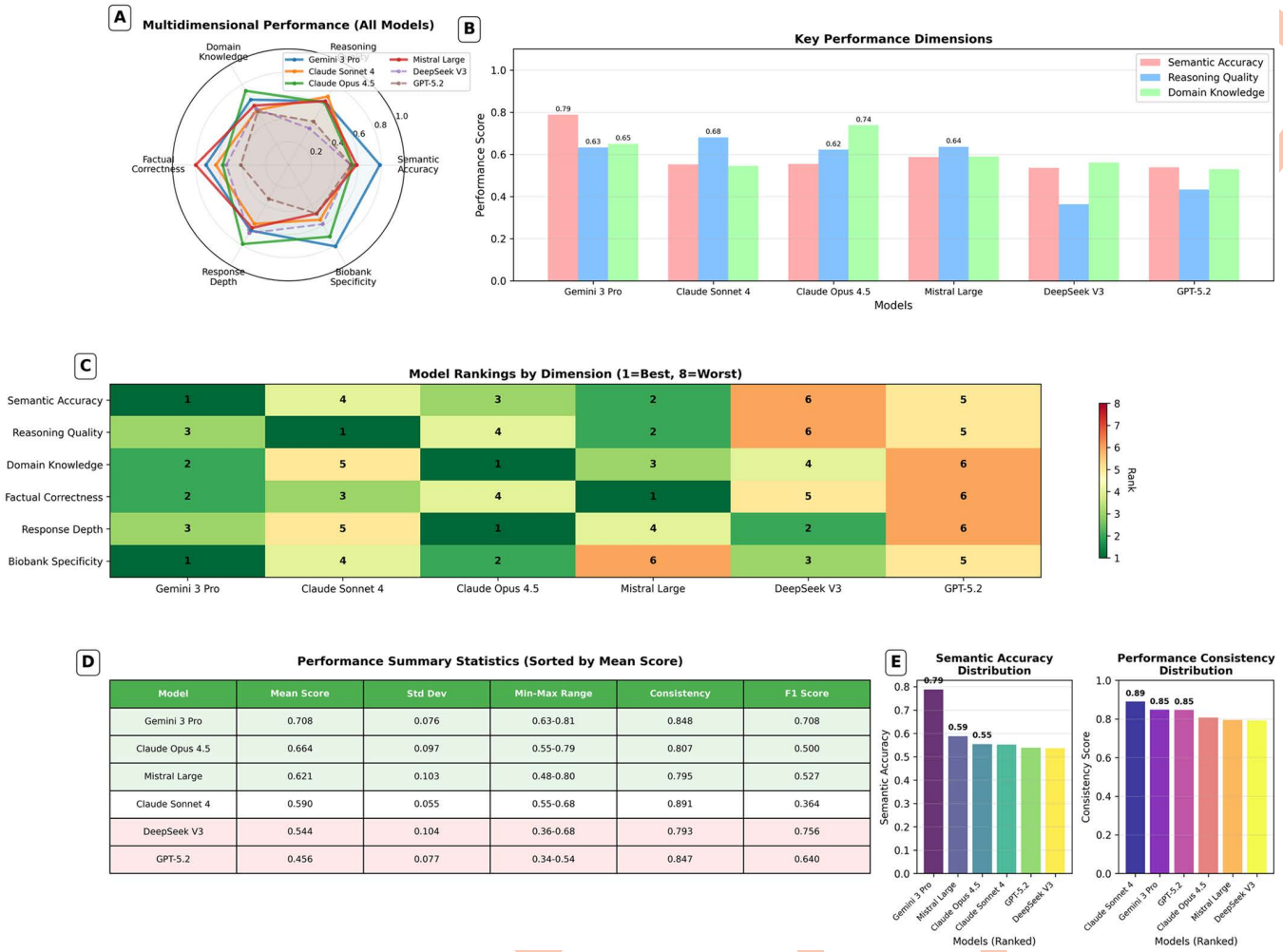
Fig 3C ranks each model across the six metrics. The rank heatmap reveals that Gemini 3 Pro was consistently positioned in the top two across all categories. Notably, Claude Opus 4.5 ranked first in domain knowledge, suggesting fluency in methodologically dense language without adequate referential grounding. Mistral Large showed strength in domain knowledge but more variable factual correctness.

Aggregate performance statistics are reported in Fig 3D, with models sorted by mean composite score. Gemini 3 Pro again led with a mean score of 0.708 and a consistency index of 0.848, denoting not only strong average performance but low intra-model variance. Mistral Large and DeepSeek V3 displayed broader ranges, reflecting performance volatility across task types. Claude Opus 4.5 yielded a relatively high F1 value (0.500), suggesting that in specific domains or tasks, their outputs achieved high precision, albeit inconsistently.

Finally, Fig 3E visualizes the distribution of semantic accuracy and consistency scores. Gemini 3 Pro achieved the highest semantic accuracy (0.79), followed by Mistral Large (0.59), with remaining models clustering around 0.55. For consistency, Claude 4 led (0.89), followed by Gemini 3 Pro and GPT-5.2 (both 0.85). Notably, GPT-5.2 exhibited high consistency despite moderate semantic accuracy, suggesting uniformly moderate performance rather than uniformly high accuracy across dimensions.

While most models surpassed baseline lexical recall, only a minority (chiefly Gemini 3 Pro and, to a lesser extent, the Claude models and Mistral Large) demonstrated semantically robust, domain-aware, and contextually coherent representations of biobank-derived knowledge.

UNCORRECTED PROOF



Model	Mean Score	Std Dev	Min-Max Range	Consistency	F1 Score
Gemini 3 Pro	0.708	0.076	0.63-0.81	0.848	0.708
Claude Opus 4.5	0.664	0.097	0.55-0.79	0.807	0.500
Mistral Large	0.621	0.103	0.48-0.80	0.795	0.527
Claude Sonnet 4	0.590	0.055	0.55-0.68	0.891	0.364
DeepSeek V3	0.544	0.104	0.36-0.68	0.793	0.756
GPT-5.2	0.456	0.077	0.34-0.54	0.847	0.640

Fig 3. Multidimensional Performance Analysis of Six Frontier Large Language Models on UK Biobank Benchmark (January 2026). (A) Radar plot showing normalised scores (0–1 scale) across six evaluation dimensions for all six models. Each axis represents one dimension: Semantic Accuracy, Factual Correctness, Domain Knowledge, Reasoning Quality, Response Depth, and Biobank Specificity. Gemini 3 Pro (blue) shows the highest overall coverage, while Claude models (orange, green) demonstrate strength in reasoning quality. **(B)** Grouped bar chart comparing three key evaluation dimensions (Semantic Accuracy, Reasoning Quality, Domain Knowledge) across all six models. Values labelled directly on bars. **(C)** Heatmap of model rankings (1 = best, 6 = worst) for each of the six evaluation dimensions. Colour gradient: green indicates top ranks, red indicates bottom ranks. **(D)** Summary statistics table displaying Mean Score, Standard Deviation, Score Range, Consistency Score, and F1 Score for each model, sorted by Mean Score in descending order. **(E)** Distribution plots showing model-wise performance for Semantic Accuracy (left panel) and Performance Consistency (right panel). Models ranked by respective scores. Claude 4 shows highest consistency (0.89) across dimensions.

https://doi.org/10.1371/journal.pcbi.1014224.g003

Comparison to random baseline performance

To validate that LLM performance reflects genuine biobank knowledge rather than coincidental term overlap, we compared each model's Weighted Coverage Score (Table 1)against a random baseline distribution (Fig 4). All six models substantially outperformed the random baseline. The mean Weighted Coverage Score across LLMs (0.597) exceeded the random baseline mean by a factor of 16x to 25x (Fig 4A, B), with all comparisons reaching statistical significance at $p < 0.001$ (Mann-Whitney U test; Fig 4C). These results confirm that the models encode meaningful biobank-specific knowledge acquired during pre-training.

UNCORRECTED PROOF

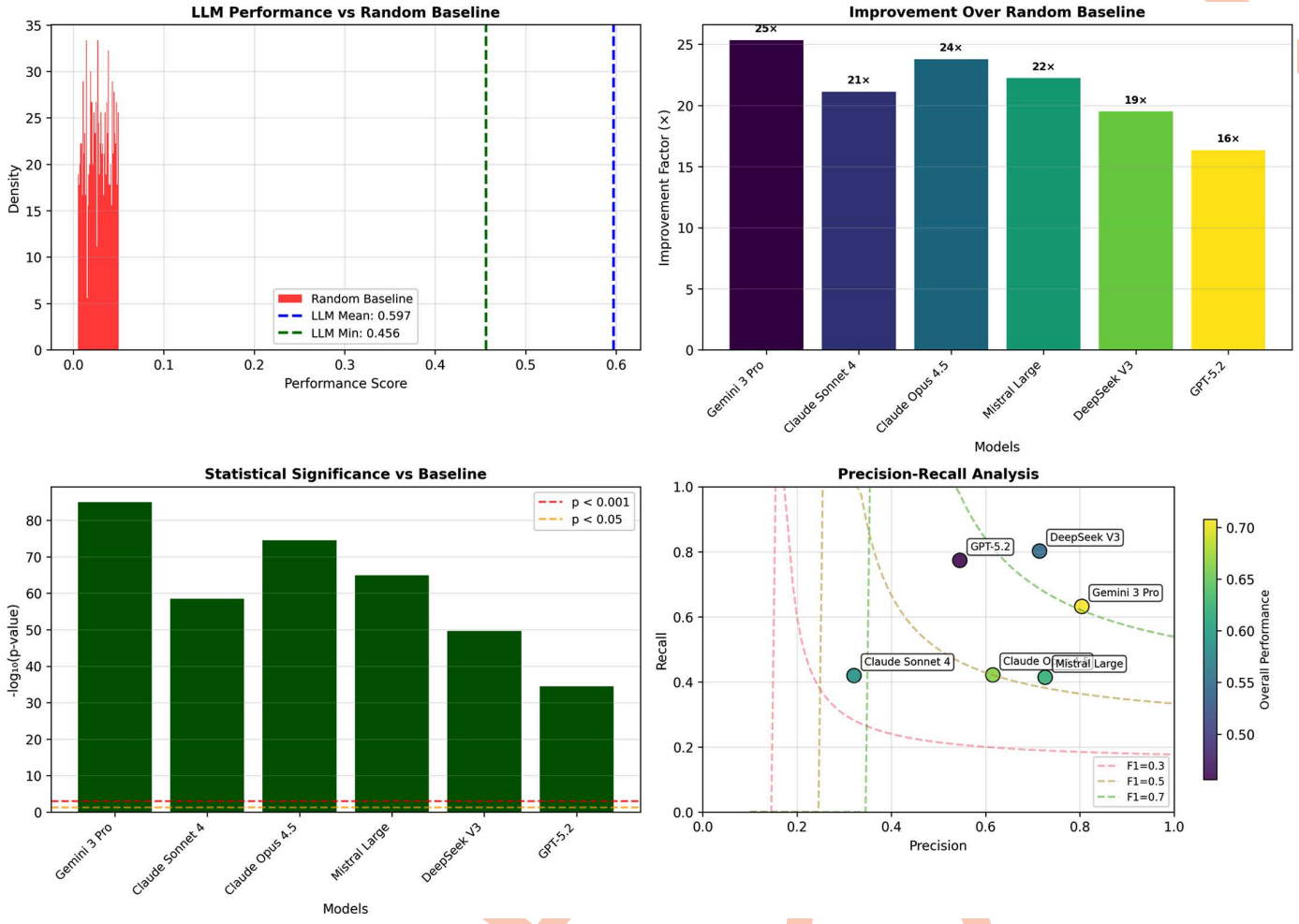


Fig 4. LLM Weighted Coverage Performance Compared to Random Baseline. All panels use the Weighted Coverage Score (mean across all four benchmark tasks: Keywords, Papers, Authors, and Institutions) as the evaluation metric. This is the same "Overall" score reported in Table 1. The Weighted Coverage Score measures the proportion of ground-truth entities retrieved by each model, weighted by entity frequency or citation impact in UK Biobank metadata. Qualitative evaluation dimensions (Reasoning Quality, Response Depth, etc.) are not applied to random outputs, as they require coherent text. **(A)** Density histogram showing the distribution of Weighted Coverage Scores for the random baseline (n = 1,000 iterations of random term sampling from UK Biobank vocabulary pools). Vertical dashed lines indicate the mean Weighted Coverage Score across all six LLMs (blue, 0.597) and the minimum LLM score (green, 0.456). All six LLMs fall far outside the random baseline distribution. **(B)** Improvement factors for each LLM over the random baseline, calculated as (Model Weighted Coverage Score)/ (Random Baseline Mean Weighted Coverage Score). All models achieve 16x to 25x improvement over chance, indicating genuine encoding of biobank-specific knowledge rather than coincidental term overlap. **(C)** Statistical significance of each LLM versus the random baseline, tested using the Mann-Whitney U test. The y-axis displays -log10(p-value); horizontal lines indicate significance thresholds at $p < 0.001$ (red) and $p < 0.05$ (orange). All models achieve $p < 0.001$. **(D)** Precision-Recall analysis for entity extraction (authors and institutions) across all models. Point color indicates Overall Weighted Coverage Score (colorbar). Dashed curves represent F1 score iso-lines at 0.3, 0.5, and 0.7. Model names are annotated adjacent to each data point.

https://doi.org/10.1371/journal.pcbi.1014224.g004

Fig 4 (top-left panel) displays the kernel density distribution of random baseline scores (red), overlaid with the minimum (0.456) and mean (0.597) performance scores achieved by LLMs. The non-overlapping distributions confirm that even the weakest model (GPT-5.2) significantly outperformed the best-case baseline. Top-right, we show relative improvement factors over baseline performance. Gemini 3 Pro led with a 25× improvement, followed by Claude Opus 4.5 (24×), Mistral Large (22×), and Claude Sonnet 4 (21×). DeepSeek V3 achieved 19× improvement, while GPT-5.2, the weakest performer, still exceeded the baseline by 16×.

UNCORRECTED

The bottom-left panel displays −log10(p-value) comparisons from one-sided Z-tests between LLMs and the random sample distribution. Gemini 3 Pro showed the strongest statistical separation, followed by Claude Opus 4.5 and Mistral Large. GPT-5.2, despite being the weakest performer, still achieved highly significant separation from baseline ($p < 0.001$).

Finally, Fig 4 (bottom-right) introduces a precision-recall analysis contextualized by interpolated F1 isocurves. Gemini 3 Pro attained the best balance of high recall and precision, while DeepSeek V3 and GPT-5.2 exhibited high recall. The Claude models showed balanced precision-recall profiles.

Discussion

This study benchmarks the ability of general-purpose large language models (LLMs) to recall biobank-relevant information, using curated UK Biobank metadata and literature as ground truth. The analysis draws from 8,549 publication abstracts and over 15,000 application records to evaluate LLM performance across four core tasks: keyword synthesis, citation retrieval, author recognition, and institutional mapping. Our results provide a structured comparison of six frontier models (as of January 2026) using coverage, semantic, and inferential metrics.

Core themes in UK biobank research

Analysis of the UK Biobank literature highlights two dominant thematic clusters: demographic characterization and genetic epidemiology. The frequent appearance of terms such as “Female,” “Male,” “Middle Aged,” and “United Kingdom” reflects the resource’s focus on population-level stratification across sex, age, and region. These demographic markers are complemented by methodological terms like “Genome-Wide Association Study” and “Mendelian Randomization Analysis,” underscoring the program’s central role in uncovering the genetic architecture of complex diseases. The most cited studies further reflect this pattern, with a strong emphasis on polygenic scoring, causal inference, and risk stratification across cardiometabolic and neurological domains.

Beyond keywords and citations, metadata analyses reveal a robust institutional footprint centered on UK-based institutions, with Oxford, Cambridge, and Imperial College London leading in application volume. Prominent individual researchers (including George Davey Smith, Naveed Sattar, and Kari Stefansson) emerged consistently in both citation and authorship metrics, confirming the consistency and quality of the ground truth corpus.

Model performance and semantic competence

While all models substantially outperformed a null baseline across tasks, their relative strengths varied sharply depending on the evaluation dimension.

Gemini 3 Pro achieved the highest overall performance and led in semantic accuracy (0.79), reflecting strong alignment between model outputs and ground truth concepts. Claude Sonnet 4 demonstrated the greatest consistency across tasks (0.89), indicating stable performance across evaluation dimensions. Claude Opus 4.5 led in domain knowledge (0.74), while both Claude models showed strong author identification capabilities (0.80 coverage). Mistral Large demonstrated competence in reasoning quality (0.64), ranking second on this dimension.

GPT-5.2 showed strong performance on institutional recognition (0.90) but struggled with author identification (0.20). DeepSeek V3 achieved moderate scores across dimensions, with institutional mapping (0.80) being its strongest task. Performance variation across task types highlights the importance of multi-dimensional evaluation.

Importantly, the precision-recall analysis indicates divergent calibration profiles. Gemini 3 Pro balanced sensitivity and specificity effectively. DeepSeek V3 achieved high recall (0.80) with strong precision, while GPT-5.2, despite high recall, showed diminished precision, indicating a tendency toward overgeneration of partially relevant content. The Claude models and Mistral Large exhibited more conservative recall with moderate to high precision. These findings point to

UNCORRECTED PROOF

fundamental differences in how LLMs prioritize inference, lexical matching, and factual grounding under open-ended biomedical prompting.

Random baseline comparison and statistical validation

To rigorously test whether LLM outputs reflect structured knowledge retrieval rather than stochastic lexical overlap, we benchmarked all models against a random baseline across all task categories. Random outputs were constructed by sampling from the UK Biobank term distributions (Schemas 19 and 27), with structural formatting and token count matched to actual model completions. These baselines yielded uniformly poor performance, with a mean semantic score near zero and no overlap with the empirical distribution of LLM-generated scores.

In contrast, all LLMs achieved substantial performance gains relative to this null distribution. Gemini 3 Pro achieved the largest improvement (exceeding the baseline by a factor of 25×) followed by Claude Opus 4.5 (24×) and Mistral Large (22×). Even the weakest model, GPT-5.2, surpassed the best-case baseline output by 16×. These results suggest that model outputs cannot be attributed to frequency priors or random recombination of biomedical terms.

Statistical validation using one-sided Z-tests confirmed significant separation between LLM and baseline responses for all models, with p-values < 0.001. This robust separation reinforces that LLM outputs reflect non-random retrieval of biobank-relevant information rather than chance lexical associations.

Interpretation and practical implications

The results of this benchmarking study highlight the emerging utility of general-purpose LLMs as viable tools for semantic querying and interpretive synthesis within large-scale biomedical corpora such as the UK Biobank. These capabilities position top-tier LLMs as promising candidates for tasks such as metadata harmonization, cohort stratification, and hypothesis generation within translational research pipelines.

At the same time, the analysis reveals considerable variance across models and tasks. Certain systems exhibited strengths in reasoning or domain-specific vocabulary yet lacked precision or factual grounding. Others demonstrated high recall but poor calibration, generating outputs that were partially relevant but prone to overgeneralization. These discrepancies emphasize the limitations of traditional evaluation metrics and the need for more rigorous, task-specific assessment frameworks in biomedical NLP.

Crucially, the observed variation in performance consistency underscores the importance of reproducible and interpretable benchmarking practices. Without fine-grained evaluation protocols, differences in LLM outputs may be masked or misinterpreted, particularly in high-stakes settings where domain specificity and inferential accuracy are critical.

Benchmark value and reusability

A natural question concerns the lasting value of a benchmark in the rapidly evolving LLM landscape, where model rankings may shift with each new release. We argue that our contribution extends beyond the specific rankings reported here in three important ways. First, we provide a reusable methodology and open-source framework for evaluating LLMs on biomedical research tasks. The code, ground truth datasets, and evaluation pipeline (available at https://github.com/manuelcorpas/LLM_4_UKB) can be readily applied to benchmark future models, enabling longitudinal tracking of progress in this domain. Second, we establish baseline performance expectations for this task category. Our random baseline analysis demonstrates that current LLMs genuinely encode biobank knowledge (achieving 16-25× performance above chance), providing a calibration point against which future models and methodologies can be compared. Third, the multidimensional evaluation framework itself represents a methodological contribution. The six dimensions we define (semantic accuracy, factual correctness, domain knowledge, reasoning quality, response depth, and biobank specificity) can be adapted to evaluate LLM performance on other specialized biomedical knowledge retrieval tasks beyond the UK Biobank

UNCORRECTED PROOF

context. Finally, the specific findings about model characteristics provide actionable insights for researchers selecting tools for biobank-related tasks today. For example, Gemini 3 Pro's strength in institutional knowledge and keyword recognition, the Claude models' superior author identification capabilities, and the general challenge of biobank specificity observed across all models inform practical tool selection. The framework ensures these insights can be systematically updated as new models emerge.

Limitations

While our evaluation framework captures a broad spectrum of interpretive capabilities, several limitations should be acknowledged. First, the study focuses on retrospective task performance using structured prompts and predefined ground-truth corpora; it does not evaluate models in generative or forward-looking tasks such as hypothesis development, data harmonization, or longitudinal pattern inference. These use cases are critical to translational research and warrant dedicated methodological assessment.

Second, although our semantic scoring pipeline incorporates embedding-based similarity and expert review, it remains constrained by the representation biases of the underlying language model used for vectorization. Semantically equivalent responses that diverge lexically may be underweighted if they fall below the cosine similarity threshold. Future approaches could incorporate multi-embedding ensembles or dynamic thresholding to improve sensitivity.

Third, our study benchmarks general-purpose models accessed via public prompts. We did not include domain-specialized models fine-tuned on biomedical literature (e.g., PubMedBERT [22] or Galactica [23]) or retrieval-augmented generation (RAG) architectures, which may offer enhanced factual grounding and context integration. Including such systems in future comparisons could provide valuable insight into the tradeoffs between generalizability and specialization.

Finally, while our random baseline offers a stringent null model for statistical benchmarking, comparisons against established symbolic or rule-based information retrieval (IR) systems were not performed. Including classical baselines such as BM25 [24], concept graph traversal [25], or MeSH-based indexing could enrich future evaluations and help disentangle the contributions of semantic modeling versus domain priors [26].

Future directions

As LLMs become increasingly embedded within biomedical research and clinical informatics pipelines, future evaluation paradigms must evolve to capture their full operational scope. Beyond entity retrieval and semantic matching, there is a pressing need to benchmark models on complex reasoning tasks, including multi-hop inference, causal explanation, and counterfactual generation [27]. Incorporating interpretability audits (such as attribution mapping and calibration diagnostics [28]) will also be critical for building trust and ensuring model transparency in high-stakes settings.

Further, expanding evaluation to clinically actionable endpoints (such as treatment prioritization, risk score generation, or eligibility screening) would better align benchmarking efforts with translational objectives [29]. These tasks require not only semantic alignment but robust integration of structured medical logic, patient heterogeneity, and evolving clinical guidelines [30].

Finally, the incorporation of temporally-aware prompts, multimodal inputs (e.g., tabular biomarkers, imaging reports), and structured biomedical ontologies (e.g., SNOMED CT [31], UMLS, or MeSH) represents a promising frontier for enhancing factual grounding and context sensitivity [32]. Benchmarking models in these contexts will be essential for determining their readiness to support real-world biomedical decision-making, from cohort design to individualized care planning.

The LLM landscape continues to evolve rapidly. The models evaluated here: Gemini 3 Pro, Claude Opus 4.5, Claude Sonnet 4, GPT-5.2, Mistral Large, and DeepSeek V3, represent the frontier as of January 2026. This evaluation updates our initial experiments conducted in late 2024 [33], which included earlier model generations (Gemini 2.0 Flash, Claude 3.5 Sonnet, GPT-4o, among others). The transition from our 2024 to 2026 evaluation demonstrates both the reusability of

UNCORRECTED PROOF

our benchmark framework and the continued advancement in model capabilities. We encourage future studies to apply our open-source framework to evaluate subsequent model releases and contribute to longitudinal tracking of progress in biobank-related knowledge retrieval.

Conclusion

This study introduces a multidimensional benchmarking framework for evaluating large language models (LLMs) in the context of the UK Biobank. By leveraging structured prompts and curated outputs, we demonstrate that general-purpose LLMs vary widely in their ability to retrieve, contextualize, and synthesize domain-specific knowledge.

Our results reveal that top-performing models (particularly Gemini 3 Pro) consistently exhibit high semantic accuracy and stable performance across diverse query types, whereas others display fragmented or error-prone behavior. Crucially, we show that coverage-based metrics alone are insufficient to capture these differences, underscoring the importance of deeper, task-sensitive evaluation criteria for biomedical applications.

Beyond benchmarking, our analysis also surfaces key thematic insights. Research involving UK Biobank data is dominated by demographic stratification (e.g., sex and age), genetic epidemiology (notably GWAS and Mendelian randomization), and chronic disease phenotypes such as cardiovascular disease and type 2 diabetes. We also identify leading contributors, including prolific authors and high-application institutions, that shape the research landscape. These findings reinforce the centrality of the UK Biobank in population-scale genomics and highlight emerging areas of opportunity for more diverse or underexplored domains.

As LLMs become increasingly embedded in biomedical research workflows, the ability to systematically characterize both their interpretive capabilities and the structures of the datasets they interrogate will be essential. The framework and findings presented here offer a rigorous foundation for future model development and evaluation, and also a lens into one of the world's most important biomedical data resources.

Supporting information

S1 Text. Reproducibility of Coverage Score and Weighted Coverage Score. Detailed explanation of the methodology for calculating Coverage Score and Weighted Coverage Score metrics, including semantic similarity matching using sentence embeddings, synonym normalization, and frequency weighting approaches.
(DOCX)

S1 Table. Scoring Rubrics for Multidimensional Evaluation Framework. Detailed scoring rubrics for each of the six evaluation dimensions (Semantic Accuracy, Factual Correctness, Domain Knowledge, Reasoning Quality, Response Depth, and Biobank Specificity) used in the multidimensional evaluation framework, including score ranges and specific criteria for each dimension.
(DOCX)

Acknowledgments

We acknowledge the UK Biobank for providing the extensive datasets and metadata that made this benchmarking study possible. Their commitment to open and collaborative research continues to advance our understanding of health and disease on a global scale.

Author contributions

Conceptualization: Manuel Corpas.

Data curation: Manuel Corpas.

Formal analysis: Manuel Corpas.

UNCORRECTED PROOF

Investigation: Manuel Corpas, Alfredo Iacoangeli.

Methodology: Manuel Corpas.

Resources: Manuel Corpas.

Software: Manuel Corpas.

Validation: Manuel Corpas, Alfredo Iacoangeli.

Visualization: Manuel Corpas.

Writing – original draft: Manuel Corpas.

Writing – review & editing: Manuel Corpas, Alfredo Iacoangeli.

References

1. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z PMID: 30305743
2. Garg M, Karpinski M, Matelska D, Middleton L, Burren OS, Hu F, et al. Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in the UK Biobank. Nat Genet. 2024;56(9):1821–31. https://doi.org/10.1038/s41588-024-01898-1 PMID: 39261665
3. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat Genet. 2021;53(2):185–94. https://doi.org/10.1038/s41588-020-00757-z PMID: 33462484
4. Gadd DA, Hillary RF, Kuncheva Z, Mangelis T, Cheng Y, Dissanayake M, et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. Nat Aging. 2024;4(7):939–48. https://doi.org/10.1038/s43587-024-00655-7 PMID: 38987645
5. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev. 2024;57:260. https://doi.org/10.1007/s10462-024-10888-y
6. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. 2025;27:e59069. https://doi.org/10.2196/59069
7. ChatGPT. [cited 13 Feb 2025]. Available: https://chatgpt.com
8. Claude. [cited 13 Feb 2025]. Available: https://claude.ai/new
9. Gemini – chat to supercharge your ideas. In: Gemini [Internet]. [cited 16 Feb 2025]. Available: https://gemini.google.com
10. Le Chat. In: Mistral AI [Internet]. [cited 13 Feb 2025]. Available: https://chat.mistral.ai
11. DeepSeek. [cited 13 Feb 2025]. Available: https://chat.deepseek.com
12. Schema. [cited 20 June 2025]. Available: https://biobank.ndph.ox.ac.uk/showcase/schema.cgi
13. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017;186(9):1026–34. https://doi.org/10.1093/aje/kwx246 PMID: 28641372
14. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50(5):668–81. https://doi.org/10.1038/s41588-018-0090-3 PMID: 29700475
15. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–24. https://doi.org/10.1038/s41588-018-0183-z PMID: 30104762
16. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018;50(8):1112–21. https://doi.org/10.1038/s41588-018-0147-3 PMID: 30038396
17. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22(3):343–52. https://doi.org/10.1038/s41593-018-0326-7 PMID: 30718901
18. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51(3):404–13. https://doi.org/10.1038/s41588-018-0311-9 PMID: 30617256
19. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641–9. https://doi.org/10.1093/hmg/ddy271 PMID: 30124842
20. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci. 2016;19(11):1523–36. https://doi.org/10.1038/nn.4393 PMID: 27643430

UNCORRECTED PROOF

21. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019;18(12):1091–102. https://doi.org/10.1016/S1474-4422(19)30320-5 PMID: 31701892
22. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare. 2021;3(1):1–23. https://doi.org/10.1145/3458754
23. Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: A Large Language Model for Science. arXiv. 2022. https://doi.org/10.48550/arXiv.2211.09085
24. Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Found Trends Inf Retr. 2009;4(1–2):333–89. https://doi.org/10.1561/1500000019
25. Rodriguez MA, Neubauer P. The Graph Traversal Pattern. In: arXiv.org [Internet]. 2010 [cited 20 June 2025]. Available: https://arxiv.org/abs/1004.1001v1
26. Jimeno-Yepes AJ, Plaza L, Mork JG, Aronson AR, Díaz A. MeSH indexing based on automatically generated summaries. BMC Bioinformatics. 2013;14:208. https://doi.org/10.1186/1471-2105-14-208 PMID: 23802936
27. Ho X, Duong Nguyen A-K, Sugawara S, Aizawa A. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics; 2020. p. 6609–25.
28. Wang AQ, Karaman BK, Kim H, Rosenthal J, Saluja R, Young SI, et al. A Framework for Interpretability in Machine Learning for Medical Imaging. IEEE Access. 2024;12:53277–92. https://doi.org/10.1109/access.2024.3387702 PMID: 39421804
29. Krewski D, Saunders-Hastings P, Baan RA, Barton-Maclaren TS, Browne P, Chiu WA, et al. Development of an Evidence-Based Risk Assessment Framework. ALTEX. 2022;39(4):667–93. https://doi.org/10.14573/altex.2004041 PMID: 36098377
30. Javed H, El-Sappagh S, Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artif Intell Rev. 2024;58:12. https://doi.org/10.1007/s10462-024-11005-9
31. El-Sappagh S, Franda F, Ali F, Kwak K-S. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak. 2018;18(1):76. https://doi.org/10.1186/s12911-018-0651-5 PMID: 30170591
32. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y). 2025;6(6):101260. https://doi.org/10.1016/j.patter.2025.101260 PMID: 40575123
33. Corpas M, Iacoangeli A. Large language models for mining biobank-derived insights into health and disease. Res Sq. 2025. https://doi.org/10.21203/rs.3.rs-6098960/v1

UNCORRECTED PROOF