A Systematic Review on the Generative AI Applications in Human Medical Genetics

Anton Changalidis, Yury Barbitoff, Yulia Nasykhova, Andrey Glotov

Dpt. of Genomic Medicine
D.O. Ott Research Institute of Obstetrics, Gynaecology, and Reproductology
St. Petersburg, Russia
anton@bioinf.me, barbitoff@bioinf.me, anglotov@mail.ru

Abstract

Although traditional statistical techniques and machine learning methods have contributed significantly to genetics and, in particular, inherited disease diagnosis, they often struggle with complex, high-dimensional data, a challenge now addressed by state-of-the-art deep learning models. Large language models (LLMs), based on transformer architectures, have excelled in tasks requiring contextual comprehension of unstructured medical data. This systematic review examines the role of generative AI methods in human medical genomics, focusing on the genetic research and diagnostics of both rare and common diseases. Automated keyword-based search in PubMed, bioRxiv, medRxiv, and arXiv was conducted, targeting studies on LLM applications in diagnostics and education within genetics and removing irrelevant or outdated models. A total of 195 studies were analyzed, highlighting the prospects of their applications in knowledge navigation, analysis of clinical and genetic data, and interaction with patients and medical professionals. Key findings indicate that while transformer-based models perform well across a diverse range of tasks (such as identification of tentative molecular diagnosis from clinical data or genetic variant interpretation), major challenges persist in integrating multimodal data (genomic sequences, imaging, and clinical records) into unified and clinically robust pipelines, facing limitations in generalizability and practical implementation in clinical settings. This review provides a comprehensive classification and assessment of the current capabilities and limitations of LLMs in transforming hereditary disease diagnostics and supporting genetic education, serving as a guide to navigate this rapidly evolving field, while outlining application use cases, implementation guidance, and forward-looking research directions.

***K*eywords** LLM · transformers · genetic diseases · diagnostics

1 Introduction

1.1 Machine Learning, Deep Learning, and Language Models

Machine learning (ML) has become a crucial tool in various fields, from healthcare to research, due to its ability to automate complex tasks and discover patterns in large datasets. Recent reviews highlight the growing impact of ML approaches in biomedical fields, including applications in diagnosing rare diseases and improving clinical outcomes [1, 2].

Traditional machine learning methods, such as decision trees and support vector machines, have been effective in solving well-defined problems where labeled data is abundant. However, these methods often struggle with high-dimensional data, complex relationships, and tasks that require context-dependent understanding, such as natural language processing (NLP) and genomics. One of the major challenges in traditional ML is handling large datasets with long-range dependencies – where information far apart in the data sequence needs to be considered together to make

accurate predictions. Additionally, it often relies on manual feature extraction and struggles with tasks that require a deeper context or understanding of relationships across the data.

With the advent of deep learning (DL), many of these limitations were overcome. Deep learning, particularly with the use of neural networks, enables models to learn directly from raw data by automatically discovering useful patterns and representations. Convolutional Neural Networks (CNNs) excel at processing images [3], while Recurrent Neural Networks (RNNs) were initially used for sequential data like text [4]. However, RNNs also encountered difficulties with tasks that involved understanding relationships across long sequences of text due to their inherent sequential processing. This led to the development of transformer-based architectures, which revolutionized NLP and a range of other fields.

The introduction of transformer models in 2017 marked a significant breakthrough in deep learning [5]. Unlike RNNs, transformers use an attention mechanism that allows the model to focus on different parts of the input data simultaneously, capturing long-range dependencies more effectively. This approach solves the problem of sequential processing and enables the model to understand complex relationships in data, very critical in healthcare and genomics. Transformers are particularly powerful in tasks that require context comprehension, such as text generation, translation, and named entity recognition. Their architecture consists of two main components: the encoder, which processes the input data (e.g., text or any other sequence, such as DNA), and the decoder, which generates the output (e.g., text). These terms refer to different stages of the model's operation: encoding involves breaking down and analyzing input data to form a representation, while decoding reconstructs or predicts the next part of the sequence based on that representation.

BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are two of the most widely known transformer-based models, each tailored for different purposes. BERT is an encoder-only model, designed to understand text in both directions (left to right and right to left), which enables it to capture a more complete context for tasks like text classification and entity recognition. This bidirectional understanding allows the model to make more accurate predictions about the meaning of a word or phrase based on its surrounding context [6]. BERT outputs embeddings for the input, learned numeric vectors that represent a token, span, or the whole sequence; in BERT these embeddings are contextual: the vector for a word depends on its surrounding text, therefore semantically related items lie close in the embedding space and can be compared or fed to downstream classifiers. On the other hand, GPT is a decoder-only model that focuses on generating text, predicting each next word based on the preceding words in a unidirectional fashion. This makes GPT highly effective at tasks, such as text generation, translation, and summarization [7].

The ability of transformers to handle large datasets and maintain coherence over long sequences has led to the development of large language models (LLMs) - models with millions or billions of parameters [8]. These models are capable of performing a variety of tasks by leveraging either full training on large datasets or fine-tuning with smaller, task-specific datasets. Fine-tuning allows the model to adapt to new tasks with minimal additional data, making few-shot or one-shot learning techniques possible: in few-shot learning, the model requires only a few labeled examples to perform well, while in one-shot learning, it can generalize from just a single example. This adaptability enables LLMs to be highly efficient across a range of applications, including research, healthcare, and education, without the need for retraining from scratch [9, 10, 11, 12, 13].

Vision Transformers (ViTs) have further extended this approach beyond text, applying the transformer architecture to image processing tasks [14]. By treating image patches like words in a sentence, ViTs can capture dependencies across different parts of an image, making them highly effective in tasks like image classification and segmentation. The versatility of transformers across multiple domains demonstrates their power and adaptability, making them integral to modern AI applications. The versatility of transformers across multiple domains demonstrates their power and adaptability, making them integral to modern AI applications.

Generative Adversarial Networks (GANs) [15] complement this landscape as specialized models for data generation, enabling the synthesis of highly realistic images, biomedical data, and even artificial genetic sequences through adversarial training.

Meanwhile, foundation models are trained on vast and diverse datasets and subsequently adapted (fine-tuned) to a wide variety of downstream tasks with minimal task-specific data. They shape the backbone of modern AI, providing general-purpose representations that can be adapted to a variety of specialized tasks. These models excel in transferring learned knowledge to new domains, accelerating advances in research, healthcare, and genomics.

Several approaches are commonly used alongside LLMs. The first is retrieval-augmented generation (RAG): before asking the model to answer, we first retrieve relevant passages from a curated corpus/database and pass them in as context. This grounding helps the model stay factual and cite evidence, because it reasons over the provided text rather than trained data [16]. The second approach involves the use of agents: instead of responding immediately, the model plans the steps, calls on tools (e.g., search engines, databases, calculators, code), inspects the results, and only

then produces a response. This enables multi-step, up-to-date answers, but it works best with guardrails (whitelisted tools, sandboxing, logging) [17, 18]. We define these briefly here and analyze design patterns and failure modes in the Discussion.

1.2 Overview of human medical genomics

Medical genomics focuses on the application of genome analysis methods for the prevention, diagnosis, and personalized management of human diseases. The methodology, however, may vary depending on the type of disease in question. Thus, for Mendelian disorders, there are two principal tasks that are inherently interconnected: (i) establishing the correct diagnosis of the disease or syndrome affecting the patient; and (ii) finding the exact genetic cause(s) of the condition (reviewed in [19]). The same two tasks are of paramount importance in cancer, where establishing the mutational profile of the tumor is essential for planning its treatment and prognosis. Another important area is the evaluation of the individual risk of the disease or specific clinical outcomes. Such prediction may be based on both genetic and environmental factors, and is especially relevant in cancer genomics and genomics of complex disease [20]. Importantly, genome analysis is frequently not limited to genome sequencing or array-based genotyping, and may involve a rich set of functional genomics tools (e.g., gene expression analysis or epigenomic profiling), particularly in cancer genomics.

Regardless of the type of disease and methods used, the clinical genomic workflow can be partitioned into three stages, hereafter called pre-analytical, analytical, and post-analytical. This structure is aligned with the ISO 15189:2022 [21], which formalizes the same sequence as pre-examination, examination, and post-examination processes. These stages or "phases of laboratory testing" encompass, respectively, test selection and specimen handling; test execution and interpretation; and report preparation, authorization, and delivery to clinicians [22].

In the context of medical genomics, the pre-analytical stage comprises biospecimen collection, organization and preprocessing of clinical data, determination of tentative diagnosis, and selection of methods that will be used for genetic testing. The next (analytical) stage is the core diagnostic phase, where genomic data are generated, processed, and interpreted. Depending on the data type, their processing and interpretation may involve identification of causal genetic variants, gene expression changes, or other types of molecular biomarkers. As shall be noted later in this review, the collection of genomic data is sometimes omitted, and inference regarding genetic alterations is made on the basis of clinical data or other types of laboratory tests. Finally, the postanalytical phase focuses on communication of the genetic test results to the patient, further patient management and counseling.

While recent reviews have explored the potential of artificial intelligence and, more specifically, transformer models in healthcare and genomics, many have limitations in scope or model specificity. For example, some reviews focus solely on the applications of ChatGPT without a systematic analysis [10, 23, 24], making them outdated or too narrowly focused. Broader reviews, such as those on LLMs in general healthcare applications or in bioinformatics, lack a specific emphasis on genetic diagnostics [25, 26, 27]. In parallel, a recent systematic review and meta-analysis comparing generative AI with physicians provides aggregate diagnostic accuracy estimates but is not focused on genetics as well [28]. Some reviews are limited to a specific disease, such as dementia [29], oncology [30, 31], schizophrenia [32] and often does not have a clear emphasis on transformer-based models[33, 34], moving a the scope of insights away from LLMs for genetic data analysis.

The closest topical review broadly covers AI in clinical genetics: it focuses on conventional DL methods and lacks depth on LLMs and transformers [35]. It is also not systematic or comprehensive, limiting its value as a foundational reference. This systematic review focuses specifically on the application of transformer models and generative AI in the research and diagnosis of hereditary diseases in recent years. To provide a comprehensive perspective, we reviewed models from four key sources: PubMed, bioRxiv, medRxiv, and arXiv, thus including both peer-reviewed studies and the latest preprint models. Since many state-of-the-art models are initially released as open-source in preprint repositories, this approach ensured we did not overlook recent developments. The growing need for efficient data processing and analysis in these domains highlights the potential of LLMs to revolutionize our understanding of genetic data, improve diagnoses, and predict disease outcomes. By exploring the use of LLMs in pre-analytical, analytical, and post-analytical stages, this review aims to provide systematic insights into how these models are transforming diagnostics, automating clinical processes, and supporting personalized medicine. A dedicated section will also assess the performance of models in clinical and research settings, examining both effective and problematic practices and ways to handle them.

Given the rapid release cycle of foundation and clinical LLMs, our goal is not to enumerate every method. Instead, we distill robust task patterns and workflows (e.g., extraction, retrieval-augmented generation, agentic pipelines, ViT-based and multimodal fusion), provide implementation guidance, and highlight near-term opportunities and risks for clinical

deployment. Our search window covers publications up to 31 January 2025; later works are discussed selectively where they materially affect the argument.

2 Methods

To comprehensively analyze the usage patterns of transformer-based models in genetics and hereditary diseases, a systematic review approach was developed according to the latest PRISMA 2020 guidelines for reporting systematic reviews [36], ensuring thorough and transparent coverage of relevant studies. The search strategy was carefully constructed with selected terms relevant to transformer-based models and genetics, and all records were evaluated through a consultative process by two researchers, allowing for in-depth discussions on ambiguous cases, promoting a balanced selection, and reducing potential bias. The full search process is visualized in Figure 1.

2.1 Search strategy

To systematically review the use of LLMs in genetics and hereditary diseases, an initial broad search for relevant articles in English was conducted across multiple major scientific databases. A custom Python script was developed to automate the collection of articles from PubMed, bioRxiv, medRxiv, and arXiv (see *Data Availability* for access to the code repository). The search criteria focused on articles from 2023, 2024, and the beginning of 2025 (January) to ensure the inclusion of the most up-to-date research in this rapidly evolving field (the dataset was downloaded on 31-01-2025). Articles from medRxiv and bioRxiv were accessed through the API available at `https://api.biorxiv.org/` (accessed on 31-01-2025), while arXiv data was retrieved using the Python wrapper `https://github.com/lukasschwab/arxiv.py` for the arXiv API (accessed on 31-01-2025). PubMed articles were accessed via the Biopython package for the PubMed API [37], available at `https://biopython.org/docs/1.76/api/Bio.Entrez.html` (accessed on 31-01-2025). This process yielded an initial dataset of 57,558 articles, forming the basis for further analysis.

The query terms were divided into two groups: one related to genetics and medicine, and the other related to transformer models and LLMs. Relevant articles were required to contain at least one term from each list in their title and/or abstract:

- genomic, genetic, inherited, hereditary, heredity, inheritance, heritability, disease subtype, NGS, next-generation sequencing, next generation sequencing, genome sequencing, phenotype description, variant interpretation, complex trait, medicine, medical, diagnosis, diagnostic, clinical, clinical decision, syndrome.
- LLM, large language model, NLP, natural language processing, GPT, chatGPT, transformer, BERT, Bidirectional Encoder Representation, RAG, retrieval-augmented generation, retrieval augmented generation, generative AI, AI assistant, prompt, chatbot, prompt engineering, attention mechanism, chain-of-thought, chain of thought.

2.2 Inclusion and Exclusion Criteria

After retrieving articles, several steps of filtering and exclusion were conducted. The first step in data processing involved automatically removing duplicate entries and cleaning the data, reducing the dataset to $51,613$ articles. This was done using text processing algorithms to detect similarities in titles and abstracts. Figure 2A illustrates the contribution of each database to the final dataset, with a substantial number of preprints included. Although preprints offer access to the latest research, they lack peer review and may contain unverified results, requiring careful analysis.

A primary semantic analysis was performed to assess the relevance of each article to the research objectives. To identify domain-specific terminology during screening and curation, TF-IDF (Term Frequency-Inverse Document Frequency) scores were calculated for all words and phrases found in article titles and abstracts. This analysis was conducted at multiple levels: for the full corpus, for the selected set of articles, with generic AI/ML phrase filtering, and using a context-preserving fine-tuned approach (full methods, detailed results, and visualizations are in Appendix 8.2.1 and Supplementary Figures 1–4). This helped highlight key terms related to genetics, hereditary diseases, and LLMs. The identified phrases were grouped into three semantic categories:

- LLM-related terms: LLM, large language model, NLP, natural language processing, GPT, chatGPT, transformer, BERT, Bidirectional Encoder Representation, RAG, augmented generation, generative AI, AI assistant, prompt engineering, chatbot, prompt engineering, attention mechanism, chain-of-thought, chain of thought.
- Clinical terms: electronic health record, ehr, clinical, case report, cds, intensive care unit, medical, syndrome, phenotype, complex trait.

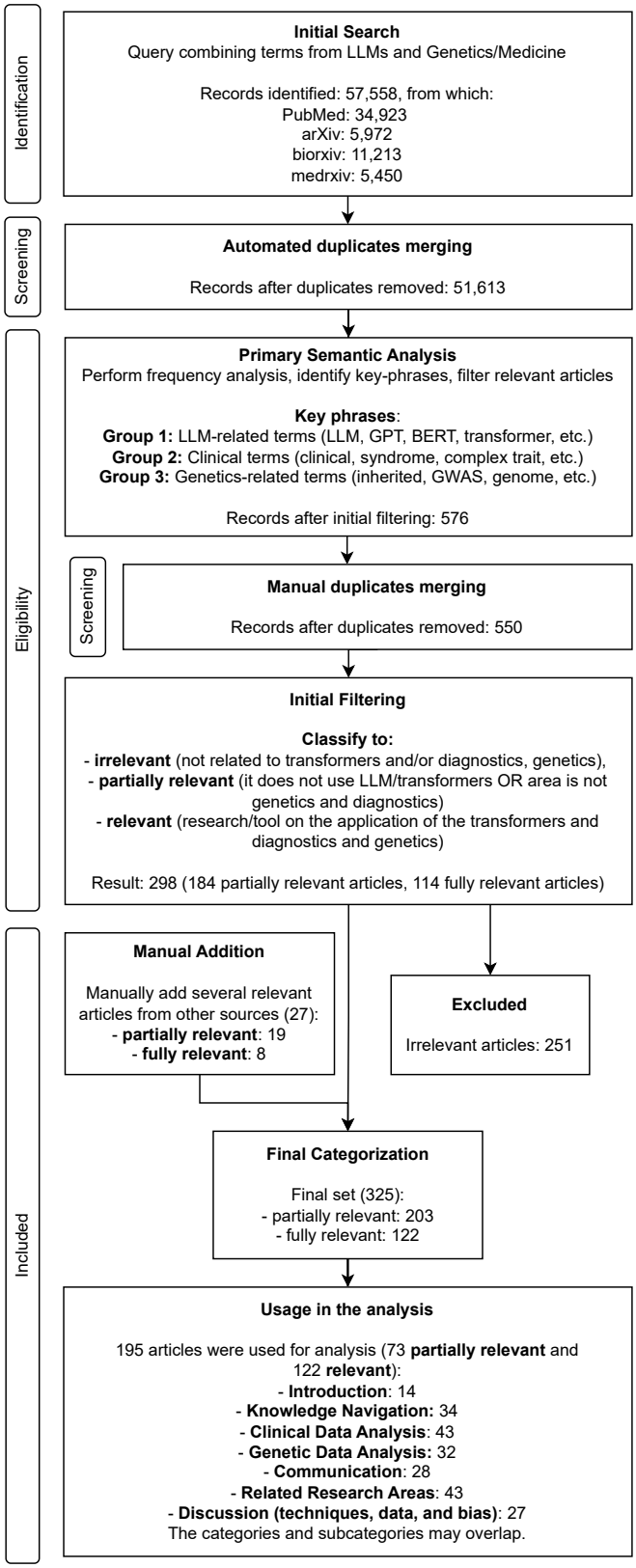


Figure 1: Pipeline of search strategy and filtering of the articles.

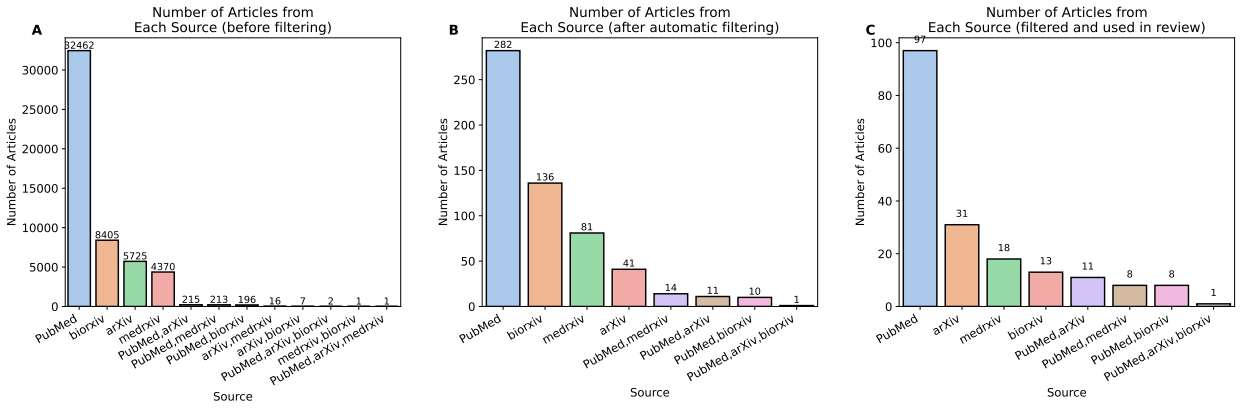


Figure 2: Distribution of articles by source: (A) after automatic deduplication and merging (51,613 records in total); (B) after additional automated filtering for relevance to clinical diagnostics (576 records in total); (C) final set used in this review (articles were manually curated, some of them were merged) (195 records in total).

- Genetics-related terms: inherit, heredit, heritability, gwas, genome-wide, genome wide, association stud, snp, single nucleotide, genetic, variant interpretation, genomic varia, human gen, NGS, generation sequencing.

These key phrases were used to filter the articles based on the presence of at least one term from each group. To ensure coverage of morphologically derived forms (e.g., "inherited", "genomics", "associations"), the terms above were defined using stemmed substrings and matched via regular expressions. Filtering required that each article contain at least one match from each of the three categories.

To avoid false positives caused by accidental substring matches in unrelated words (e.g., "cove**rag**" or "encou**rag**" falsely matching "rag"), an empirically derived exclusion list was applied. This list was constructed by manually reviewing articles irrelevant to the study focus and identifying recurring misleading terms. This list included the following terms or common letter combinations: *tragic, fragment, coverag, encourag, ungs, angs, ongs, ings, eragrostis, smallmouth, fragile, angptl, intragenic, fragment, hallmark, uvrag, leverag, storag, averag, coverag, encourag, forage, liraglutid*. This filtering strategy significantly improved the precision of semantic classification by excluding structurally similar but contextually irrelevant terms.

Additionally, a manual verification step was conducted to identify and remove duplicate entries that were not detected automatically. In several cases, articles had slightly different titles or abstracts but were authored by the same group and described the same study. Based on this content-level similarity and author overlap, duplicates were removed, reducing the dataset from 576 to 550 articles for subsequent analysis. The complete list of included articles is provided in Supplementary Table 1. As previously noted, this step, as well as all subsequent ones, were conducted jointly by two researchers, allowing for careful discussion of ambiguous cases and minimizing potential bias.

After deduplication, articles were manually divided into three classes, based on their relevance to the topic. In order to be considered fully relevant (114 articles), the articles had to meet the following criteria: (i) involve development or evaluation of transformer-based or similar models; and (ii) focus on the extraction, processing, or prediction of genetic information or phenotypic information directly linked to inherited disease (e.g., recognition of rare disease symptoms). 184 articles met only one of these criteria (i.e., described non-transformer models or dealt with adjacent fields of research not directly linked to clinical genetic testing) and therefore were classified as partially relevant. All other articles (252) were considered irrelevant and were excluded.

In addition to automated filtering, 27 manually selected articles of partial (19) and high (8) relevance were included in the final dataset, bringing the total to 325 articles (Supplementary Table 2): 203 partially relevant articles and 122 fully relevant articles. Additional articles were sourced through references from the initially selected studies, as well as through further targeted filtering and searches across the originally extracted dataset.

At this stage, a thorough investigation of the selected articles was conducted. All highly relevant articles, as well as some of the partially relevant ones, were included in the analysis. From the latter category, only those entries were chosen that provided good examples of deep learning methods used in diagnostics, even if not specifically focused on LLMs or transformers. During the review process, each article was assessed based on its relevance to specific sections (see *Results*). Additionally, insights into best and worst practices of transformer usage are outlined in the

discussion section. Since these areas encompass a broad list of tasks, they have been divided into specific applications, and itemized (see the relevant sections and Figure 3).

In total, of the 325 categorized studies, 195 were used for the analysis (122 relevant and 73 partially relevant). Among these, 154 focused on diagnostics and 27 were used as examples of practices discussed in the review (with some articles used in multiple sections). Furthermore, 14 articles were incorporated in the introduction as examples of existing systematic research with a similar topic (see Figures 1,3B).

2.3 Risk of Bias

A large proportion of the selected articles came from preprint databases, such as arXiv, bioRxiv, and medRxiv, meaning they had not yet undergone peer review. This could introduce some bias, as these studies have not been validated by the scientific community. However, given the fast-paced nature of LLM development, many of the most cutting-edge techniques are being developed faster than the peer-review process allows. Consequently, it was deemed essential to include such articles to capture the most current advancements.

Additionally, while this review focuses on the application of LLMs in the specific domain of genetics and hereditary diseases, there may be general-purpose models or methods from broader AI fields that were not included in this focused analysis. These models could still provide valuable insights or advancements applicable to this domain, although they fall outside the scope of this particular review.

2.4 Semantic landscape of the literature (TF-IDF)

TF-IDF profiling showed a consistent progression from generic to domain-specific themes (Supplementary Figure 1). In the full corpus (51,613 articles; SF 1 A), generic phrases such as "language models", "large language", and "artificial intelligence" dominated, confirming broad field coverage prior to curation. The curated set (195 articles; SF 1 B) preserved these anchors and surfaced domain cues (e.g., "precision medicine"): evidence that selection retained the core landscape. After removing generic AI/ML phrases (SF 1 C, SF 3), specific trends emerged, including "precision medicine", "gene expression", "open source", and "genetic testing", alongside disease-focused ("breast cancer", "alzheimer disease"), resource-oriented ("human phenotype ontology"), and technique-oriented ("attention mechanism", "single cell") terms.

A context-preserving fine-tuned analysis was performed to address potential concerns about removing AI/ML terminology (Supplementary Figure 4). This approach first trained the TF-IDF model on the curated dataset with full vocabulary, then applied post-hoc reweighting to down-weight generic terms while preserving semantic relationships. The fine-tuned analysis confirmed that domain-specific trends remain stable across different filtering strategies, validating our findings.
Source comparisons (Supplementary Figures 2, 3, 4) further clarified complementarity. Before filtering, PubMed ($n = 131$) and preprints ($n = 64$) shared 57% of top phrases ($17/30$), indicating strong consensus on core topics. After filtering, overlap dropped to 23% ($7/30$), revealing distinct emphases (Supplementary Figure 3). The fine-tuned analysis showed similar patterns with 23% overlap ($7/30$), confirming the robustness of these findings.

PubMed leaned clinical and translational ("precision medicine", "genetic testing"; established disease terms), whereas preprints highlighted emerging computational motifs ("gene expression", "open source", "attention mechanism") and method-forward phrasing. Many key terms from both technical and biological domains ranked highly in preprints but were absent from PubMed, supporting our dual-source strategy and underscoring that inclusion of preprints offers a more comprehensive view of the field.

Topic modeling further revealed the semantic structure of the literature (Supplementary Figure 5). Eight latent topics were extracted using Latent Dirichlet Allocation (LDA), capturing distinct research themes from clinical variant interpretation to computational method development. The topic overlap visualization demonstrates how different research areas interconnect, with some topics (e.g., clinical diagnostics and precision medicine) closely related while others (e.g., protein structure prediction and variant calling) occupy distinct semantic spaces.

3 Results

Our systematic review identified and used a total of 195 studies that report application of generative AI methods for a wide variety of tasks within the scope of human medical genomics. After careful curation, we have split these studies into four main categories depending on the study design, methods and data types employed: (i) knowledge navigation (34 articles); (ii) clinical data analysis (43 articles); (iii) genetic data analysis (32 articles); and (iv) communication with patients and medical professionals (28 articles). Each category was then subdivided into several subcategories

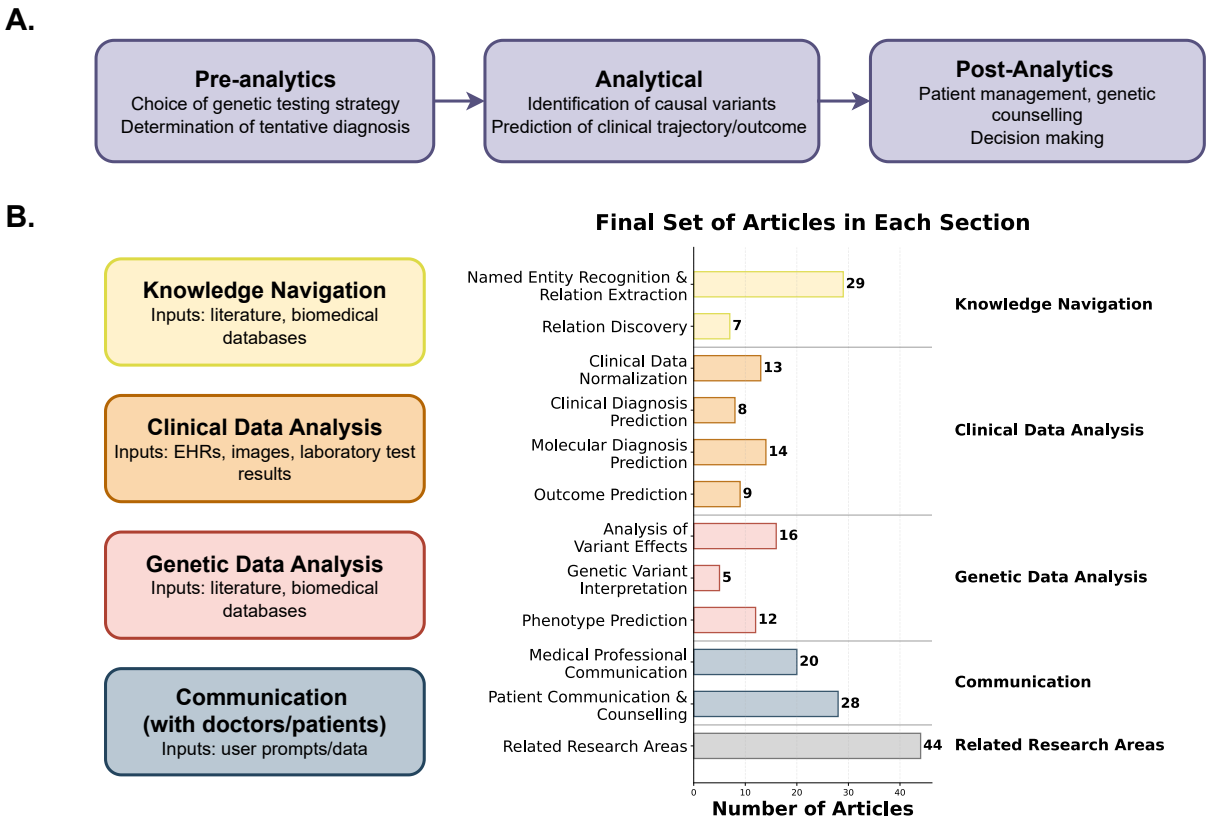


Figure 3: A diagram showing the applications of LLMs in the research and diagnosis of human genetic diseases. A diagram showing the applications of LLMs across stages of the genetic diagnostics workflow. (A) Pre-analytical, analytical, and post-analytical phases of clinical genetics, illustrating how LLMs support test selection, variant identification, and decision-making. (B) To the left: four major functional domains of LLM use: knowledge navigation, clinical data analysis, genetic data analysis, and communication with clinicians or patients. To the right: distribution of the final set of reviewed articles with corresponding subcategories highlighted.

corresponding to major tasks addressed by respective generative AI methods and models (Figure 3B). In the following sections, we will summarize articles from each category, highlighting the most notable studies and discussing prospects for further method development in each area.

3.1 Knowledge Navigation

Most of the studies in the knowledge navigation category dealt with the extraction of structured information from published sources or biomedical databases. In many cases, this goal is achieved through named entity recognition (NER) and relation extraction (RE), and is primarily focused on the extraction of gene-disease or variant-disease relationships from published literature (e.g., [38]). This task is exceptionally important given the vast amount of such information available in the literature, which is, in many cases, not properly reflected in major databases such as Online Mendelian Inheritance in Man (OMIM) [39] or NCBI ClinVar [40]. Data extracted from literature sources are crucial for clinical geneticists and can be used at all stages of the genetic testing workflow. Thus, knowledge about gene-disease associations may aid in the selection of appropriate genetic testing methods and inform interpretation of sequencing results. Complementing these trends, training very small, task-specific encoders is emerging as an accuracy-preserving and controllable alternative to general LLMs, highlighting the promise of small, fine-tunable models for biomedical tasks while reducing hallucination risk [41]. At the same time, there is a rising trend of employing decoder-based LLMs (e.g., GPT-3.5/4, PhenoGPT, GP-GPT) for entity-level tasks, which, despite convenient one/few-shot use (providing one to several examples right in the query) and promising results in some studies, may be suboptimal for extraction tasks due to architectural mismatches [42, 43]. This trend invites further research, as will be described in the Discussion.

Beyond the extraction of simple relations, some studies involved a more sophisticated design. For example, some studies focused on complex multi-entity relationships (e.g., DUVEL (Detection of Unique Variant Ensembles in Literature) [44]). In other studies, the extraction of gene-disease associations was complemented with curated resources and interpretable extraction frameworks (e.g., GPAD (Gene-Phenotype Association Discovery), RelCurator [45, 46]). Several works combined knowledge navigation tasks with question answering, developing specialized tools and models for interactive communication with researchers or clinicians. Examples of such Q&A systems include PubTator 3.0 [47] and BioMedLM [48], and demonstrate improved answer factuality and superior performance compared to general-purpose LLMs. A number of specialized systems, such as ClinVar-BERT, AutoPM3, and VarChat, are optimized specifically for variant interpretation, providing variant impact summaries [49] or extracting pathogenicity evidence for genetic variants from published sources [50, 51].

Aside from the 29 studies involving information extraction, a separate subcategory (comprising 7 studies) focused on the prediction of novel gene-disease relationships. These studies utilized both models specifically trained for solving the task of causal link prediction (e.g., end-to-end disease-gene association prediction model with parallel graph transformer network (DGP-PGTN) [52] or LitGene [53]) as well as interactive large language models (e.g., Med-PaLM 2 [54]). Another notable work described the application of transformer-based models to the identification of causal genes at GWAS loci [55]. While limited in number, these studies illustrate the potential of generative AI methods for hypothesis generation – a goal which, if successfully met, can greatly advance biomedical research in various fields beyond medical genetics.

3.2 Clinical Data Analysis

This category was the largest in our analysis and comprised diverse efforts involving the analysis of electronic health records (EHRs), clinical notes, and results of non-genetic laboratory testing with a goal of phenotypic data organization, providing tentative diagnosis or disease subtypes. Similarly to literature review, these types of analysis are more commonly performed prior to or during genetic testing with a goal of selecting the appropriate testing strategy and enhancing interpretation. However, as shall be described below, there are several attempts to circumvent the need for genetic testing by providing information on actionable genetic markers based solely on other types of data.

The first subcategory of studies focused on extraction and normalization of clinical information from EHRs. Methods employed for this task largely overlap with those used for extraction of information from scientific literature. In purview of clinical data processing, however, the main emphasis is laid onto the extraction and normalization of phenotypic information of the patient, typically by mapping it onto Human Phenotype Ontology (HPO) terms [56] using both encoder- and decoder-based models [57, 58, 42, 59, 60, 61].

Beyond normalization of phenotype descriptors, a large number of models are built for suggesting genetic diagnosis on the basis of the patient's phenotypic features using textual (EHRs) or visual information (e.g., portrait photos or data from other imaging methods). In the former category, generative LLMs such as GPT-3.5, GPT-4, and Gemini (which are obsolete at the moment) have been applied to suggest candidate diagnoses in autoinflammatory and neurogenetic disorders, or predict cancer predisposition genes from textual EHR summaries [62, 63, 64]. Models with visual data inputs are also designed to predict both tentative diagnoses and causal genetic alterations. For example, an older CNN-based model called DeepGestalt has proven its efficacy in syndromic features identification [65], with its newer version, GestaltMML (multimodal Transformers over facial photos, clinical notes, and metadata), having improved accuracy due to its multimodal design [66]. In oncology, a large number of models have been built to predict the mutational profile of the tumor based on histopathology data (whole slide images, WSIs). Examples of such efforts include prediction of gene mutation status [67, 68, 69, 70, 71, 72] or aggregate genomic features such as tumor mutational burden [73]. A peculiar feature of these approaches is that they are designed as a substitute for, rather than being complementary to, costly genetic testing.

Finally, a series of studies focused on the prediction of various clinical outcomes in patients using a mixture of genetic and non-genetic information. Notable examples of such studies include stratification of survival risk in breast cancer patients [74] or genetics-informed subtyping of Alzheimer's disease patients [75].

3.3 Genetic Data Analysis

While bioinformatic analysis of genomic data is commonly considered to be the most complicated step of a medical genomics workflow, only a minority of studies identified by our review directly employed generative AI for genetic variation analysis. The respective methods were focused on three major tasks: (i) phenotype-agnostic prediction of functional impact of genetic variants; (ii) prioritization of genetic variants in the context of NGS results interpretation; and (iii) aggregation of genetic variation data for prediction of the patient's phenotype (typically, in connection with complex disease).

In the first subcategory, much of the promise of generative AI is connected with the development of domain-specific models (foundation models) to understand the language of biological molecules (e.g., DNA or proteins). While biological sequences differ from natural language due to a lack of easily identifiable "words", AI methods have already demonstrated their extraordinary capabilities in solving fundamental tasks. Nobel prize-winning AlphaFold [76] is the most notable example of such models that demonstrated groundbreaking performance in protein folding. A number of well-established methods have been developed on top of the protein language model employed by AlphaFold, including AlphaMissense, a tool that has become a de facto gold standard in the evaluation of pathogenicity of amino acid substitutions [77]. Beyond prediction of impact for amino acid substitutions in proteins, a range of models for working with DNA sequence have been proposed, with some showing promising results in tasks related to genetic variation analysis (e.g., prediction of splice sites, epigenetic marks, enhancer sequence, promoter sequence, enhancer activity, chromatin profile, and others) [78, 79, 80, 81]. These models are already trained to understand the context of a sequence, and their representations can be fine-tuned for a diverse range of downstream tasks. In addition, models trained for specific tasks also exist – their advantage is that they can be much smaller and therefore require fewer computational resources. In a notable study, transformers have advanced splice site prediction for identifying disease-relevant splice variants [82]. Transformer-based variant annotation is not limited to point mutations - for instance, a tool called PhenoSV applies attention-based modeling to structural variants (SVs) to capture how both non-coding and coding structural variants affect gene function [83].

Another important and particularly challenging area of bioinformatic analysis of genome sequencing data is the identification of causal genetic variants among millions present in each individual genome [19]. In this field, a variety of generative AI methods have also shown their exceptional performance. For example, authors of the Mendelian Approach to Variant Effect pRedICtion (MAVERICK) tool report ranking the causal variant among the top five variants in over 95% of the cases [84]. Other tools, such as Genetic Transformer (GeneT), also report high performance in variant prioritization [85], and benchmarking studies confirm substantial improvement of clinical variant classification from using state-of-the-art models and other techniques, such as fine-tuning and RAG (for details see discussion), including LLMs [86, 87].

Finally, a set of generative AI-based methods has been developed to enhance polygenic risk prediction in complex diseases. A recurrent strategy employed in several studies is the application of LLMs and other models for the construction of informative predictive features (such as epigenetic markers) based on the individual genotypes (e.g., Epi-PRS [88] or epiBrainLLM [89]). Other studies attempt to use transformer architectures for modeling epistatic interactions between genes [90] or for simple classification of patients into subtypes based on their genotype, as exemplified by a study in Parkinson's disease [91].

3.4 Interaction with Patients and Medical Professionals

The last category of generative AI applications in medical genetics leverages the unique capacity of LLMs to communicate with the user in natural language. Such communication typically involves medical question answering, and can assist both medical professionals and patients. As mentioned in previous sections, interactive chatbots have been developed and used for various tasks mentioned earlier in this review, including knowledge navigation, clinical data analysis, and genetic variant interpretation [92, 93, 49, 64, 63, 94, 95, 96]. However, the use of generative AI for interaction with researchers, doctors, and patients is not limited to Q&A tasks. In this subsection, we will briefly describe other notable works involving communication with patients or medical professionals.

In the realm of interaction with medical professionals, one study has reported the use of generative AI to address privacy challenges of using real patient images in genetics education. A study on Kabuki and Noonan syndromes found that AI-generated facial images, created using StyleGAN [97] methods, were nearly as effective as real photos in training pediatric residents to recognize phenotypic features [98]. While real images were rated slightly more helpful, synthetic ones notably increased diagnostic confidence and reduced uncertainty.

Some works are focused on the development of interactive assistants for the interpretation of genetic test results. One notable example is the study by Yang et al. [99] who have constructed an LLM module for textual summaries of submodules of a knowledge graph. Another notable case is the Just-DNA-seq platform that integrates a custom GPT model called GeneticsGenie to facilitate the interpretation of genetic test results by users with no genetics background [100]. In another effort, an AI assistant was developed for the interpretation of pharmacogenomic test results [101]. Besides interpretation of genetic testing results, a range of studies have explored the application of LLMs in genetics question answering [102, 103], counseling [104, 105], and education [98, 106]. It has to be noted, however, that studies reveal variability in accuracy, especially in nuanced topics, such as inheritance patterns or ethical subtleties of genetic risk communication [103, 106]. Besides, models still risk hallucination and outdated references, highlighting the need for oversight and continual retraining [106].

Taken together, all of the aforementioned applications are well aligned with general trends in the field of generative AI methods, which are increasingly being used as personal assistants in various fields. However, a range of technical and ethical concerns still raise doubts regarding the implementation of LLMs in clinical genetics in the near future (see Discussion for a more in-depth analysis of the outstanding issues).

3.5 Related Research Areas

Although this review focuses on the applications of generative AI models in human genetics and diagnostics, several adjacent research areas, while not directly related to human genome analysis, offer valuable insights and transferable lessons. These applications were excluded from the main focus due to limited direct relevance; however, they highlight the variety of possible usage across biological and medical domains and may inspire future applications in human genetic research.

Studies applying LLMs to microbial genomes have demonstrated the potential of language models to encode meaningful representations of whole genomes. For example, models trained on bacterial or fungal species can predict traits such as antibiotic resistance or habitat specificity [107, 108, 109]. While distinct from human genetics, these works show how transformer-based models can capture population structure and gene interactions in complex biological systems.

Transformer models have also been applied to protein sequences for predicting gene ontology terms and functional annotations [110]. These studies operate in the space of proteomics, yet demonstrate modeling principles that could be extended to human gene function prediction or variant interpretation.

A substantial body of work with ViTs focuses on cancer imaging, particularly for tasks such as tumor segmentation, subtype classification, and spatial analysis from whole-slide images [111, 112, 113, 114, 115, 116]. While not always grounded in genomic data, these tasks intersect with genetic diagnostics when molecular subtypes play a role in treatment stratification.

Epigenetic regulation and cross-species prediction of gene expression using sequential and imaging data represent another promising direction [109, 117, 112]. These studies explore how attention-based models can generalize across evolutionary distances, enabling predictions in under-characterized organisms and informing functional annotation pipelines.

LLMs are being increasingly integrated into gene editing workflows: from automating guide RNA design and protocol generation (e.g., CRISPR-GPT) to predicting cellular responses to perturbations at single-cell resolution (e.g., scLAMBDA) [118, 119]. Together, these applications illustrate how transformer models support both the interpretation and manipulation of gene function in human genetics.

Transformer models have also been used to investigate the role of genetic support in the success or failure of clinical trials [120]. While not directly diagnostic, such applications emphasize the growing role of human genetic evidence in pharmaceutical development and clinical decision-making.

Taken together, these diverse research directions extend the scope of transformer-based models beyond traditional genetics. By leveraging techniques and datasets from related fields, such as microbial biology, cancer diagnostics, and synthetic biology, future work in human genetics may benefit from models and insights developed in adjacent domains.

4 Discussion

As transformer models, more specifically, encoder-only (e.g., BERT family), decoder-only (e.g., GPT family), and encoder-decoder (e.g., T5 [121] – a flexible text-to-text transformer architecture not originally trained for biomedical tasks, but widely adapted for biomedical NLP [122, 123]), are increasingly integrated into specialized fields such as genetics, it is essential to evaluate their strengths and limitations. This research systematically reviewed the application of these models in genetic diagnostics, incorporating resources from PubMed, bioRxiv, medRxiv, and arXiv to ensure both peer-reviewed depth and inclusion of the latest model developments, focusing on hereditary disease diagnostics while retaining adjacent work for context.

In this section, we aim to provide a guideline for selecting the models and strategies for researchers willing to integrate generative AI tools in their workflows. We list and discuss the main benefits and limitations of different models and techniques, and highlight key developments needed to ensure the reliability and trustworthiness of LLM applications. We also touch upon emerging trends and techniques that may enhance their effectiveness.

4.1 Model selection guide

In this subsection, we summarizes major families of generative AI models across a range of data types they process (modalities), including text, sequence, images, and their combinations (Table 1). We try to link the models to respective tasks and examples reviewed in the previous section.

Table 1: Architectures and capabilities map for genetics and adjacent tasks.

Family	Modality (examples)	Proper application	Limitations / cautions	Examples
Encoder-only	Text (clinical notes, biomedical literature)	NER/RE, mapping of terms to standardized vocabularies (ontologies)	Long documents may truncate, weak for free-form generation	ClinVar-BERT [50], PathoBERT [124], Big Bird [125]
Decoder-only	Text (clinical narratives, reports, instructions)	Drafting reports, Q&A, guideline summarization, next event prediction	Hallucinations without retrieval, version drift	GPT [7], ChatGPT [8], GeneGPT [126], Comet [127]
Encoder-decoder	Text (summaries, structured templates)	Summarization, controlled generation (using templates)	All limitations of encoder- and decoder-only	T5 [121], TSCA-Net [128]
Foundation models	Biological sequences (DNA, RNA, protein)	Tasks involving sequence analysis (e.g., variant and regulatory effect prediction, epigenomic signal transfer)	Miss long-range effects, less reliable for rare or cross-species data; require validation	GENA-LM [129], Nucleotide Transformer [117], MethylGPT [78], AlphaMissense [77], MIPPI [130], CellFM [131], Enformer [117]
ViT, Hybrid CNN-Transformer	Images (MRI, WSI, facial phenotypes)	Predictions based on imaging data (e.g., disease subtyping, prediction of genetic alterations)	Sensitive to data quality and bias, require expert annotation, limited by ethical constraints	CroMAM [72], BPGT [68], PromptBio [132], ChromTR [133], Tokensome [134], Gestalt-MML [66]
Multimodal models	Multimodal (imaging, genomics, text, tabular data)	Integration of diverse data types	Modality imbalance; missing-modality handling	MGI [135], BioFusionNet [74], Genetic InfoMax [136]

Models based on transformer architecture (including encoder-only, decoder-only, and encoder–decoder variants) can be used to address a wide range of tasks involving biomedical text processing. Thus, encoder-style models (e.g., BioBERT, ClinVar-BERT, PathoBERT, Big Bird [125]) remain the most reliable for structured extraction and annotation (NER/RE, HPO normalization), and for structuring clinical data or texts. Decoder LLMs (including both general-purpose, such as ChatGPT, or specialized, such as GeneGPT) add value for generative tasks, such as clinician-facing Q&A, report drafting, and exploratory hypothesis generation, but typically require retrieval and tool calling for robust, auditable performance; mixed systems (RAG, agents – described in the next section) reduce hallucinations when grounded in curated resources and databases. Additionally, decoder-only models can be used for next-event prediction (e.g., Comet [127]). Lastly, full encoder-decoder (seq2seq) architectures are useful for controlled, template-constrained generation of text (e.g., highly structured summaries or other tasks). Some studies use decoder-only models (e.g., ChatGPT) where a transformer encoder would likely perform better for extraction-focused workloads [137, 55, 138]. It is important to note that choosing the right architecture is critical. Although early comparisons often focused on GPT-3.5 vs. GPT-4, newer models and configurations combining reasoning and tool-use have been released. While they can improve results, studies consistently report hallucinations, outdated knowledge, and stylistic artifacts distinguishable from expert writing [105, 139, 103, 138].

Foundation models for DNA/RNA/protein (e.g., GENA-LM, Nucleotide Transformer, MethylGPT, CellFM [131], Enformer [140]) provide reusable representations for splice/regulatory effect prediction, epigenomic signal transfer, variant effect scoring, and downstream tasks including drug response and trait/PRS modeling. Such pre-trained models can be fine-tuned for any specific task, provided with data. Their main cautions concern tokenization granularity, long-range dependencies, domain/species shift, and calibration on rare regions or reliance on predicted structures.

If the goal is to process other types of data beyond text, specialized model types have to be used. For instance, vision backbones and hybrid CNN-Transformer systems address a range of image processing tasks, including working with MRI, microscopic images, and facial phenotypes for tasks such as mutation status prediction (e.g., CroMAM, BPGT, PromptBio), karyotyping (e.g., ChromTR, Tokensome), and syndrome suggestion (e.g., GestaltMML). When the foal is to align or fuse imaging, sequences, or text, multimodal models (e.g., MGI, BioFusionNet, Genetic InfoMax) can be developed (see next subsection for discussion of related techniques). As mentioned in earlier sections, these models allow for more accurate predictions compared to single-modality models (e.g., for tasks such as predicting cancer patient survival [74]). However, gains from using image data or multimodal fusion depend more on data acquisition methods (e.g., imaging instruments or staining techniques), and may be particularly vulnerable to class balance or other issues characteristic of traditional machine learning frameworks. These problems may have a comparable or even more dramatic impact than model size, and special attention has to be dedicated to data preparation. Techniques, such as normalization, site-balanced splits, and external validation mitigate common risks.

Finally, complex specialized architectures (Epi-PRS, Prophet, TransBTS) combine multiple mechanisms (convolution, attention, classical ML) to model relations across sequences, images, and text. Such pipelines can substantially improve machine understanding of clinical-genetic signals but require deeper domain knowledge in model training and testing.

Overall, model selection should be driven by task, data, and safety requirements: encoders for extraction and normalization, decoder LLMs (with retrieval/tools) for controllable generation, encoder-decoder models for structured seq2seq outputs, biological foundation models when specific pattern understanding is needed, and multimodal/vision architectures where phenotype-genotype links are image-mediated. Using the latest versions, domain adaptation, and careful prompting improves performance, but rigorous evaluation remains essential [141, 95, 63, 142]. In the next section, we will consider techniques that can improve the results of model usage.

4.2 Model Strategies

We now discuss how to achieve effective use of transformer models, since outcomes depend on how systems are composed. Table 2 aggregates prompting, retrieval, tool-use, long-context modeling, multimodal fusion, privacy-preserving training, and evaluation patterns, indicating when to use them, expected benefits, typical limitations, and concrete mitigations.

Prior to model training, data quality control (QC) and preprocessing are important. Artifacts, missing data points, or inconsistent phenotype capture degrade model inputs [152, 153, 143, 144]. Preprocessing (e.g., segmentation, standardization) improves data quality by reducing artifacts, but increases can also encode hidden bias or lead to reproducibility issues [143, 144]. Fully documenting steps and QC, pinning versions, and testing on external datasets are essential for robustness and reproducibility.

General LMs (e.g., BioBERT, GENA-LM, GPT-4) often miss disease- or site-specific patterns, labeled cohorts are small (risk of overfitting/forgetting), access to protected data is constrained, and guidelines keep changing, models must be adapted for the target task while remaining flexible [154, 155, 156]. One of the possible solutions is using fine-tuning and domain adaptation. Full fine-tuning (i.e., task-specific re-training) can maximize alignment when labels and compute suffice, but risks overfitting and forgetting on small cohorts. Parameter-efficient methods (PEFT – small parameter updates [157]; e.g., LoRA[145]/adapters[146]/QLoRA[147]) keep the backbone of the model frozen, reduce compute and exposure of protected information, and enable rapid iteration across disease/task variants [147]. Mixed and continual training (keeping learning gradually) helps retain broad knowledge useful for retrieval and classification while still specializing: by exposing the model to diverse data/tasks at once, it builds more general representations, reduces overfitting, and stays flexible. In contrast, the classic pretrain-then-finetune pipeline deepens task-specific skills but is more prone to overfitting on small cohorts. Evidence suggests mixed training yields models that remain adaptable for downstream tasks, important in genetics, where knowledge and guidelines evolve, balancing specialization and generalization, e.g. for genetic counseling or variant interpretation [158]. Recent findings also show that domain-specific pretraining alone does not guarantee superiority: randomly initialized models can match or exceed genomic foundation models in downstream tasks [159]. Moreover, very small, task-focused LMs with selective incremental learning can be competitive for pathway inference while reducing hallucinations [41].

Clinical notes and omics sequences contain distal dependencies that short-context models miss; tokenization can also contradict biology (sequence chunking vs. function). Long-context sequence models (e.g., GENA-LM, Nucleotide Transformer) deal with this by targeting both long- and short-distal dependencies in long notes and genomes. Hybrid windows (local+global), task-specific heads, and comparisons to short-context baselines help maintain accuracy. What is more, decisions and solutions should often rely on data with multiple modalities, such as images, genomics, and text (otherwise, insights from complementary signals will be lost). There are two common ways for multimodal fusion, which means combining such data. The first approach is contrastive learning to place all modalities in the same shared

Table 2: From clinical data problems to LLM-based solutions: techniques/patterns with benefits, limitations, and mitigations.

Problem or task	Technique / Pattern	Benefits	Limitations	Mitigations	Example Models
Noisy or unstandardized data	Modality-specific preprocessing & QC	Cleaner input, higher signal-to-noise ratio	Sensitive to small changes, reproducibility risk	Standardized workflows and QC protocols, external validation	[143, 144]
Local or data-specific patterns	Fine-tuning or domain adaptation	Higher accuracy on small, focused datasets	Overfitting, loss of general knowledge	Lightweight fine-tuning, frozen backbone, external cohort testing	LoRA[145]/ adapters[146]/ QLoRA[147]
Long-range dependencies	Long-context transformers	Distal genomic or textual relations captured	Tokenization-biology mismatch, distal trade-offs	Combine local and global contexts, add task-specific layers, benchmark against short-range models	GENA-LM[129], Nucleotide Transformer[79]
Multimodal or missing inputs	Contrastive learning or cross-attention fusion	Usage of complementary signals, greater robustness	Modality imbalance, missing input at inference	Hyperparameters tuning, validate on diverse, multi-site data	CroMAM[72], BioFusionNet[74]
Need for structured reasoning	Prompting, Chain-of-Thoughts, one/few-shot	Consistent reasoning, reusable templates	Prompt leakage, verbosity, unstable zero-shot behavior	Few-shot verified prompts, separate reasoning/final output, regular review	BioGPT[126], Med-PaLM 2 [54]
Need for reference-grounded answers	Retrieval Augmented Generation (RAG)	Reduced hallucinations, improved factual grounding	Weak retrieval, outdated sources	Curated indices, freshness policies, inline citations	GeneGPT and others [87, 126]
Need for tool or API execution	Agentic AI	Automated tasks decomposition and workflow execution	Tool errors, latency, unsafe calls	Restriction to verified tools, safety checks, human oversight	BioAgents [148], BioChatter [149]
Data privacy protection	Federated Learning	Collaboration without raw data sharing	Complex setup, coordination overhead	Site-specific evaluation plans, standardized protocols	SF-GWAS [150]
Fair testing and leakage prevention	Evaluation- or leakage-aware benchmarks	Transparent comparison, contamination control	Hidden leakage, overfitting to test data	External test sets, preregister benchmarks, no train/test overlap	CARDBiomedBench [151]

space and train a model to understand the relationship between data points by learning to differentiate between similar and dissimilar pairs [160, 161]. The second one is using cross-attention between modalities or late fusion to let one modality use information from another [162]. Typical problems are modality imbalance, a missing modality at test time, and domain shift from site/scanner/stain differences. Practical fixes include curriculum learning and hyperparameters (e.g. temperature) tuning (for contrastive), missing-modality heads, or modality dropout. Additionally, specialists often need structured, step-wise outputs without retraining; naive prompts can leak context, get verbose, and behave unstably in a zero-shot setting (when output examples are not included in the query prompt). Prompting, including adding specific instructions for reasoning (Chain-of-Thought) [163, 164], providing one to several examples of the desired structure and result (this technique is called one or few-shot learning), helps to create structured and desired outputs. However, it is still vulnerable to leakage and verbosity, therefore periodically validating/updating prompt libraries remains important. Furthermore, clinical answers must be fact-checked and reference-backed, since relying on internal model memory can produce hallucinations (confident mistakes). RAG searches databases and websites before generating answers, therefore reducing hallucinations and providing a controllable, citable trace. Key practices include using curated indices (e.g., ClinVar/OMIM/HPO), enforcing freshness policies, applying document-grounded scoring, requiring inline citations, and using deterministic decoding with version pinning to ensure actuality, auditability, and

stability [165, 101, 104, 92, 95, 87]. These approaches help mitigate the risk of relying solely on a model's internal memory.

Biomedical workflows are often complex and require calculations, ontology/database queries, and code execution. Otherwise analysis steps are not reproducible and reliable. Tool-use and agents decompose complex workflows into callable steps (calculations, ontology/database queries, code execution) and preserve traces for reproducibility. They help with HPO mapping, risk scores, hypothesis generation, and experiment planning. Open frameworks and systems, such as BioChatter [149] or BioAgents [148] demonstrate constrained, locally deployable, retrieval-enhanced pipelines for biomedical tasks. Advanced agentic systems for genetics include BioDiscoveryAgent for perturbation-experiment design [166] and a chatbot agent to facilitate family communication of hereditary risk in familial hypercholesterolemia [167].

Finally, multi-site collaboration is often required, while raw data cannot be shared. Federated learning enables privacy-preserving training and cross-site collaboration without raw data exchange, aligning with regulatory expectations [150, 168, 169, 170]. However, these techniques usually require additional technical expertise.

Taken together, these practices underscore that effectiveness depends not only on model scale but also on task alignment, prompt design, real-time access to knowledge, and auditable reasoning tools, which are key ingredients for trustworthy clinical deployment. In the next section, we will consider evaluation and benchmarks, which are necessary for credible claims about the model's quality.

4.3 Data and Benchmarks

The growing use of generative AI is closely tied to the quality of available datasets and benchmarks. Reliable evaluation and generalization depend not only on model design but also on data diversity, integrity, and task-relevant benchmarking protocols [152, 153, 143, 144].

LLM applications in genetic diagnostics require reliability; therefore, robust benchmarks are vital for comparing models and ensuring trust. CARDBiomedBench [151] exemplifies this shift, offering a multi-domain Q&A benchmark in biomedicine. Its design is based on curated expert knowledge and data augmentation, which exposes real gaps in model reasoning and safety, even among state-of-the-art systems. The number of domain benchmarks, reported scores, and proposed tracking methods continues to grow [137, 171, 102, 96, 51, 42, 172], helping move beyond general-purpose NLP benchmarks toward the nuanced reasoning required in biomedical decision-making.

Recent work in other technical domains has highlighted the threat of benchmark leakage, where models inadvertently see test data during pretraining [173, 174]. Complementing these findings, another study shows a chronological "task contamination" effect: LLMs score markedly higher on datasets released before their training data cutoff than on post-cutoff sets, with supporting evidence from training-data inspection and membership-inference attacks, underscoring how pretraining overlap can inflate zero/few-shot results [175]. These studies demonstrate how leakage can inflate performance and undermine credibility, motivating leakage-aware protocols and transparent documentation of training data, especially sensitive domains, such as biomedicine.

As noted throughout this review, modern clinical models must integrate diverse data types: text, images, genomics, structured records, which requires both scalable architectures and consistent input quality. Recent methods improve efficiency in multimodal fusion (e.g., contrastive learning, cross-attention) [176, 75, 66, 74, 135, 177, 81], while preprocessing helps standardize specific modalities (e.g., segmentation [178, 179, 128], or facial axes standardization [180]).

Independent of architecture, version pinning (model, tokenizer, prompts, decoding parameters), leakage-aware evaluation, and traceability (logged sources, tool traces, decision checkpoints) improve safety and reproducibility [181, 182]. For transparent assessment and regulatory preparedness, healthcare reporting checklists such as MI-CLEAR-LLM are recommended [183]. Together, these developments underscore that the value of LLMs in genetics is not solely defined by model architecture. Equally important are the integrity of training and evaluation datasets, the representativeness of benchmarks, and the methods used to integrate and align multimodal inputs.

4.4 Biases

Despite their impressive capabilities, LLMs often reflect biases present in their training data, which can affect clinical utility. Several studies have revealed racial and demographic biases in generated medical reports and other outputs [184, 185], while others show variations in performance across age-specific manifestations of genetic disorders [186] or reviewer experience levels [187]. Language also remains a critical source of disparity: most biomedical models are English-centric, limiting accessibility and accuracy in other languages. Resources such as MedLexSp for Spanish [188],

Chinese medical conversational Q&A corpora [189], and domain adaptation efforts for Japanese genetic counseling [104] demonstrate how localized models and lexicons can help reduce these gaps.

In genomics and precision medicine, the lack of diversity in training data has long limited the generalizability of AI insights for underrepresented groups. Over 80% of genome-wide association studies to date have been conducted on individuals of European ancestry[190], leading to predictive tools that underperform in other populations. For example, polygenic risk scores trained predominantly on Eurocentric cohorts show substantially lower accuracy when applied to individuals of African, Hispanic, or other ancestries, reflecting poor out-of-distribution generalization and exacerbating health disparities[191, 192]. These gaps highlight that without deliberate interventions to include diverse data, AI systems, including LLMs, may fail to equitably serve marginalized communities.
Overall, recent findings underscore the need for targeted fairness efforts and rigorous ethical evaluation in deploying AI. In practice, we recommend reporting results stratified by site and language (separate performance metrics per hospital/registry and clinical language); notably, even FDA-cleared AI tools rarely report performance by patient demographics, underscoring the importance of transparent subgroup evaluation. Using ancestry-aware sampling during model development (i.e. balancing or weighting cohorts to better reflect the target population) is another key step, alongside technical bias-mitigation measures[193]. For instance, integrating data augmentation and algorithmic debiasing techniques can help ensure models maintain consistent performance across subgroups[194]. We also advice scheduling fairness checks in production, periodic bias audits that monitor performance gaps across demographic subgroups, to catch and remediate any emerging disparities. Such responsible AI practices, combined with language- or population-specific model adaptations, are essential to mitigate bias and promote more equitable clinical AI systems[193].

5 Conclusions

As detailed in this review, transformer-based models have made significant progress in various critical tasks within the research and diagnosis of human genetic diseases. Generative AI methods have proven their efficiency in diverse tasks related to knowledge navigation, analysis of clinical and genetic data, and interaction with researchers, medical specialists, and patients. Owing to the peculiar architecture of generative models, they have found their application beyond standard classification tasks, and are now widely used for complex tasks such as genetic variant interpretation, generation of novel biological hypotheses, or prediction of complex epigenomic features for polygenic risk assessment.

Generative AI tools, including LLMs, hold clear potential for supporting various professional roles involved in genetic medicine. For clinical geneticists, LLM-powered systems (described in this article as well as newly developed) can assist in providing definitive diagnosis, prediction of individual risks, and interactions with patients. For researchers and bioinformaticians, such models offer solutions for complicated tasks involving processes of vast amounts of genomic or other high-throughput data. As LLMs mature, we anticipate their deployment in software environments designed to assist these distinct expert groups, enhancing the quality and speed of inherited disease diagnostics.

Naturally, this review cannot cover every tool and model in a field that evolves so rapidly. Rather, it provides a structured overview that can serve as a classifier and guide, helping researchers and practitioners navigate the fast-growing landscape of LLM applications in human medical genomics.

6 Data Availability

All data and code pertinent to the results presented in this work are available at `https://github.com/TohaRhymes/llm_in_diagnostics`.

7 Acknowledgments

This research was supported by the Ministry of Science and Higher Education of the Russian Federation (project "Multicenter research bioresource collection "Reproductive Health of the Family" contract No. 075-15-2025-478 from 29 May 2025).

References

[1] Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3), 2022.

[2] P Roman-Naranjo, A M Parra-Perez, and J A Lopez-Escamez. A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases. *J Biomed Inform*, 143:104429, June 2023.

[3] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, Mar 2021.

[4] Safwan Mahmood Al-Selwi, Mohd Fadzil Hassan, Said Jadid Abdulkadir, Amgad Muneer, Ebrahim Hamid Sumiea, Alawi Alqushaibi, and Mohammed Gamal Ragab. Rnn-lstm: From applications to modeling techniques and beyond—systematic review. *Journal of King Saud University - Computer and Information Sciences*, 36(5):102068, 2024.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[9] Duo Du, Fan Zhong, and Lei Liu. Enhancing recognition and interpretation of functional phenotypic sequences through fine-tuning pre-trained genomic models. *Journal of Translational Medicine*, 22(1):756, Aug 2024.

[10] Stefania Zampatti, Cristina Peconi, Domenica Megalizzi, Giulia Calvino, Giulia Trastulli, Raffaella Cascella, Claudia Strafella, Carlo Caltagirone, and Emiliano Giardina. Innovations in medicine: Exploring chatgpt's impact on rare disorder management. *Genes*, 15(4), 2024.

[11] Samuel J. Aronson, Kalotina Machini, Jiyeon Shin, Pranav Sriraman, Sean Hamill, Emma R. Henricks, Charlotte Mailly, Angie J. Nottage, Sami S. Amr, Michael Oates, and Matthew S. Lebo. Preparing to integrate generative pretrained transformer series 4 models into genetic variant assessment workflows: Assessing performance, drift, and nondeterminism characteristics relative to classifying functional evidence in literature, 2024.

[12] Daiju Ueda, Shannon L. Walston, Toshimasa Matsumoto, Ryo Deguchi, Hiroyuki Tatekawa, and Yukio Miki. Evaluating gpt-4-based chatgpt's clinical potential on the nejm quiz. *BMC Digital Health*, 2(1):4, Jan 2024.

[13] Matthew J Laye and Michael B Wells. Rapid creation of Knowledge-Balanced student groups using ChatGPT4. *Med Sci Educ*, 34(3):523–525, April 2024.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[17] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.

[18] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023.

[19] Yury A Barbitoff, Mikhail O Ushakov, Tatyana E Lazareva, Yulia A Nasykhova, Andrey S Glotov, and Alexander V Predeus. Bioinformatics of germline variant discovery for rare disease diagnostics: current approaches and remaining challenges. *Briefings in Bioinformatics*, 25(2):bbad508, 01 2024.

[20] Hannah Wand, Samuel A Lambert, Cecelia Tamburro, Michael A Iacocca, Jack W O'Sullivan, Catherine Sillari, Iftikhar J Kullo, Robb Rowley, Jacqueline S Dron, Deanna Brockman, Eric Venner, Mark I McCarthy, Antonis C Antoniou, Douglas F Easton, Robert A Hegele, Amit V Khera, Nilanjan Chatterjee, Charles Kooperberg, Karen Edwards, Katherine Vlessis, Kim Kinnear, John N Danesh, Helen Parkinson, Erin M Ramos, Megan C Roberts, Kelly E Ormond, Muin J Khoury, A Cecile J W Janssens, Katrina A B Goddard, Peter Kraft, Jaqueline A L MacArthur, Michael Inouye, and Genevieve L Wojcik. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, 591(7849):211–219, March 2021.

[21] Medical laboratories — requirements for quality and competence. Includes POCT requirements; supersedes ISO 15189:2012 and ISO 22870:2016.

[22] Kenneth A. Fleming, Mahendra Naidoo, Michael Wilson, John Flanigan, Susan Horton, Modupe Kuti, Lai Meng Looi, Christopher Price, Kun Ru, Abdul Ghafur, Jianxiang Wang, and Nestor Lago. High-quality diagnosis: An essential pathology package. In Dean T. Jamison, Hellen Gelband, Susan Horton, Prabhat Jha, Ramanan Laxminarayan, Charles N. Mock, and Rachel Nugent, editors, *Disease Control Priorities: Improving Health and Reducing Poverty*, chapter 11. The International Bank for Reconstruction and Development / The World Bank, Washington, DC, 3 edition, November 2017. See Box 11.1: Three Phases of Laboratory Testing.

[23] Jinge Wang, Zien Cheng, Qiuming Yao, Li Liu, Dong Xu, and Gangqing Hu. Bioinformatics and biomedical informatics with chatgpt: Year one review, 2024.

[24] Madhan Jeyaraman, Swaminathan Ramasubramanian, Sangeetha Balaji, Naveen Jeyaraman, Arulkumar Nallakumarasamy, and Shilpa Sharma. ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. *World J Methodol*, 13(4):170–178, September 2023.

[25] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R Chaurasia, Nirav R Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A Pfeffer, and Nigam H Shah. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328, January 2025.

[26] Liang Cheng. Attention mechanism models for precision medicine. *Briefings in Bioinformatics*, 25(4):bbae156, 05 2024.

[27] Anqi Lin, Junpu Ye, Chang Qi, Lingxuan Zhu, Weiming Mou, Wenyi Gan, Dongqiang Zeng, Bufu Tang, Mingjia Xiao, Guangdi Chu, Shengkun Peng, Hank Z H Wong, Lin Zhang, Hengguo Zhang, Xinpei Deng, Kailai Li, Jian Zhang, Aimin Jiang, Zhengrui Li, and Peng Luo. Bridging artificial intelligence and biological sciences: a comprehensive review of large language models in bioinformatics. *Briefings in Bioinformatics*, 26(4):bbaf357, 07 2025.

[28] Hirotaka Takita, Daijiro Kabata, Shannon L. Walston, Hiroyuki Tatekawa, Kenichi Saito, Yasushi Tsujimoto, Yukio Miki, and Daiju Ueda. A systematic review and meta-analysis of diagnostic performance comparison between generative ai and physicians. *npj Digital Medicine*, 8(1):175, Mar 2025.

[29] Juan A. Berrios Moya. Addressing the gaps in early dementia detection: A path towards enhanced diagnostic models through machine learning, 2024.

[30] Emily M Webster, Muhammad Danyal Ahsan, Luiza Perez, Sarah R Levi, Charlene Thomas, Paul Christos, Andy Hickner, Jada G Hamilton, Kemi Babagbemi, Evelyn Cantillo, Kevin Holcomb, Eloise Chapman-Davis, Ravi N Sharaf, and Melissa K Frey. Chatbot artificial intelligence for genetic cancer risk assessment and counseling: A systematic review and Meta-Analysis. *JCO Clin Cancer Inform*, 7:e2300123, September 2023.

[31] Aya Mudrik, Abraham Tsur, Girish N Nadkarni, Orly Efros, Benjamin S Glicksberg, Shelly Soffer, and Eyal Klang. Leveraging large language models in gynecologic oncology: A systematic review of current applications and challenges. *medRxiv*, 2024.

[32] Antoine Deneault, Alexandre Dumais, Marie Désilets, and Alexandre Hudon. Natural language processing and schizophrenia: A scoping review of uses and challenges. *Journal of Personalized Medicine*, 14(7), 2024.

[33] Priyanka Venkatapathappa, Ayesha Sultana, Vidhya K S, Romy Mansour, Venkateshappa Chikkanarayanappa, and Harish Rangareddy. Ocular pathology and genetics: Transformative role of artificial intelligence (AI) in anterior segment diseases. *Cureus*, 16(2):e55216, February 2024.

[34] Xiang Dai, Sarvnaz Karimi, and Nathan O'Callaghan. Identifying health risks from family history: A survey of natural language processing techniques, 2024.

[35] Dat Duong and Benjamin D Solomon. Artificial intelligence in clinical genetics. *Eur J Hum Genet*, 33(3):281–288, January 2025.

[36] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[37] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[38] Dao-Ling Huang, Quanlei Zeng, Yun Xiong, Shuixia Liu, Chaoqun Pang, Menglei Xia, Ting Fang, Yanli Ma, Cuicui Qiang, Yi Zhang, Yu Zhang, Hong Li, and Yuying Yuan. A combined manual annotation and Deep-Learning natural language processing study on accurate entity extraction in hereditary disease related biomedical literature. *Interdisciplinary Sciences: Computational Life Sciences*, 16(2):333–344, June 2024.

[39] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7, January 2005.

[40] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 11 2017.

[41] Pranta Saha, Joyce Reimer, Brook Byrns, Connor Burbridge, Neeraj Dhar, Jeffrey Chen, Steven Rayan, and Gordon Broderick. Reconstructing biological pathways by applying selective incremental learning to (very) small language models, 2025.

[42] Kitty B Murphy, Brian M Schilder, and Nathan G Skene. Harnessing generative ai to annotate the severity of all phenotypic abnormalities within the human phenotype ontology. *medRxiv*, 2024.

[43] Yanjun Lyu, Zihao Wu, Lu Zhang, Jing Zhang, Yiwei Li, Wei Ruan, Zhengliang Liu, Xiaowei Yu, Chao Cao, Tong Chen, Minheng Chen, Yan Zhuang, Xiang Li, Rongjie Liu, Chao Huang, Wentao Li, Tianming Liu, and Dajiang Zhu. Gp-gpt: Large language model for gene-phenotype mapping, 2024.

[44] Charlotte Nachtegael, Jacopo De Stefani, Anthony Cnudde, and Tom Lenaerts. DUVEL: an active-learning annotated biomedical corpus for the recognition of oligogenic combinations. *Database*, 2024:baae039, 05 2024.

[45] K. M. Tahsin Hassan Rahit, Vladimir Avramovic, Jessica X. Chong, and Maja Tarailo-Graovac. Gpad: a natural language processing-based application to extract the gene-disease association discovery information from omim. *BMC Bioinformatics*, 25(1):84, Feb 2024.

[46] Heonwoo Lee, Junbeom Jeon, Dawoon Jung, Jung-Im Won, Kiyong Kim, Yun Joong Kim, and Jeehee Yoon. RelCurator: a text mining-based curation system for extracting gene-phenotype relationships specific to neurodegenerative disorders. *Genes Genomics*, 45(8):1025–1036, June 2023.

[47] Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546, 04 2024.

[48] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024.

[49] Federica De Paoli, Silvia Berardelli, Ivan Limongelli, Ettore Rizzo, and Susanna Zucca. VarChat: the generative AI assistant for the interpretation of human genomic variations. *Bioinformatics*, 40(4), March 2024.

[50] Weijiang Li, Xiaomin Li, Ethan Lavallee, Alice Saparov, Marinka Zitnik, and Christopher Cassa. From text to translation: Using language models to prioritize variants for clinical review. December 2024.

[51] Shumin Li, Yiding Wang, Chi-Man Liu, Yuanhua Huang, Tak-Wah Lam, and Ruibang Luo. Autopm3: Enhancing variant interpretation via llm-driven pm3 evidence extraction from scientific literature. *bioRxiv*, 2024.

[52] Yang Li, Zihou Guo, Keqi Wang, Xin Gao, and Guohua Wang. End-to-end interpretable disease–gene association prediction. *Briefings in Bioinformatics*, 24(3):bbad118, 03 2023.

[53] Ala Jararweh, Oladimeji Macaulay, David Arredondo, Olufunmilola M Oyebamiji, Yue Hu, Luis Tafoya, Yanfu Zhang, Kushal Virupakshappa, and Avinash Sahu. Litgene: a transformer-based model that uses contrastive learning to integrate textual information into gene representations. *bioRxiv*, 2024.

[54] Tao Tu, Zhouqing Fang, Zhuanfen Cheng, Svetolik Spasic, Anil Palepu, Konstantina M. Stankovic, Vivek Natarajan, and Gary Peltz. Genetic discovery enabled by a large language model. *bioRxiv*, 2023.

[55] Suyash S. Shringarpure, Wei Wang, Sotiris Karagounis, Xin Wang, Anna C. Reisetter, Adam Auton, and Aly A. Khan. Large language models identify causal genes in complex trait gwas. *medRxiv*, 2024.

[56] Michael A Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B Addo-Lartey, Anna V Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M Bagley, Eduard Bakštein, James P Balhoff, Gareth Baynam, Susan M Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F Bodenstein, Pablo Botas, Kaan Boztug, Jolana Čady, Tiffany J Callahan, Rhiannon Cameron, Seth J Carbon, Francisco Castellanos, J Harry Caufield, Lauren E Chan, Christopher G Chute, Jaime Cruz-Rojo, Noémi Dahan-Oliel, Jon R Davids, Maud de Dieuleveult, Vinicius de Souza, Bert B A de Vries, Esther de Vries, J Raymond DePaulo, Beata Derfalvi, Ferdinand Dhombres, Claudia Diaz-Byrd, Alexander J M Dingemans, Bruno Donadille, Michael Duyzend, Reem Elfeky, Shahim Essaid, Carolina Fabrizzi, Giovanna Fico, Helen V Firth, Yun Freudenberg-Hua, Janice M Fullerton, Davera L Gabriel, Kimberly Gilmour, Jessica Giordano, Fernando S Goes, Rachel Gore Moses, Ian Green, Matthias Griese, Tudor Groza, Weihong Gu, Julia Guthrie, Benjamin Gyori, Ada Hamosh, Marc Hanauer, Kateřina Hanušová, Yongqun Oliver He, Harshad Hegde, Ingo Helbig, Kateřina Holasová, Charles Tapley Hoyt, Shangzhi Huang, Eric Hurwitz, Julius O B Jacobsen, Xiaofeng Jiang, Lisa Joseph, Kamyar Keramatian, Bryan King, Katrin Knoflach, David A Koolen, Megan L Kraus, Carlo Kroll, Maaike Kusters, Markus S Ladewig, David Lagorce, Meng-Chuan Lai, Pablo Lapunzina, Bryan Laraway, David Lewis-Smith, Xiarong Li, Caterina Lucano, Marzieh Majd, Mary L Marazita, Victor Martinez-Glez, Toby H McHenry, Melvin G McInnis, Julie A McMurry, Michaela Mihulová, Caitlin E Millett, Philip B Mitchell, Veronika Moslerová, Kenji Narutomi, Shahrzad Nematollahi, Julian Nevado, Andrew A Nierenberg, Nikola Novák Čajbiková, John I Nurnberger, Jr, Soichi Ogishima, Daniel Olson, Abigail Ortiz, Harry Pachajoa, Guiomar Perez de Nanclares, Amy Peters, Tim Putman, Christina K Rapp, Ana Rath, Justin Reese, Lauren Rekerle, Angharad M Roberts, Suzy Roy, Stephan J Sanders, Catharina Schuetz, Eva C Schulte, Thomas G Schulze, Martin Schwarz, Katie Scott, Dominik Seelow, Berthold Seitz, Yiping Shen, Morgan N Similuk, Eric S Simon, Balwinder Singh, Damian Smedley, Cynthia L Smith, Jake T Smolinsky, Sarah Sperry, Elizabeth Stafford, Ray Stefancsik, Robin Steinhaus, Rebecca Strawbridge, Jagadish Chandrabose Sundaramurthi, Polina Talapova, Jair A Tenorio Castano, Pavel Tesner, Rhys H Thomas, Audrey Thurm, Marek Turnovec, Marielle E van Gijn, Nicole A Vasilevsky, Markéta Vlčková, Anita Walden, Kai Wang, Ron Wapner, James S Ware, Addo A Wiafe, Samuel A Wiafe, Lisa D Wiggins, Andrew E Williams, Chen Wu, Margot J Wyrwoll, Hui Xiong, Nefize Yalin, Yasunori Yamamoto, Lakshmi N Yatham, Anastasia K Yocum, Allan H Young, Zafer Yüksel, Peter P Zandi, Andreas Zankl, Ignacio Zarante, Miroslav Zvolský, Sabrina Toro, Leigh C Carmody, Nomi L Harris, Monica C Munoz-Torres, Daniel Danis, Christopher J Mungall, Sebastian Köhler, Melissa A Haendel, and Peter N Robinson. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Res*, 52(D1):D1333–D1346, January 2024.

[57] Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. *Patterns*, 5(1):100887, 2024.

[58] Abdulkadir Albayrak, Yao Xiao, Piyush Mukherjee, Sarah S. Barnett, Cherisse A. Marcou, and Steven N. Hart. Enhancing human phenotype ontology term extraction through synthetic case reports and embedding-based retrieval: A novel approach for improved biomedical data annotation. *Journal of Pathology Informatics*, 16:100409, 2025.

[59] Ekin Soysal and Kirk Roberts. Phenormgpt: a framework for extraction and normalization of key medical findings. *Database*, 2024:baae103, 10 2024.

[60] Daniel B. Hier, Thanh Son Do, and Tayo Obafemi-Ajayi. A simplified retriever to improve accuracy of phenotype normalizations by large language models. *Frontiers in Digital Health*, 7, 2025.

[61] Davy Weissenbacher, Siddharth Rawal, Xinwei Zhao, Jessica R C Priestley, Katherine M Szigety, Sarah F Schmidt, Mary J Higgins, Arjun Magge, Karen O'Connor, Graciela Gonzalez-Hernandez, and Ian M Campbell. PhenoID, a language model normalizer of physical examinations from genetics clinical notes. January 2024.

[62] Joshua Pillai and Kathryn Pillai. Accuracy of generative artificial intelligence models in differential diagnoses of familial mediterranean fever and deficiency of interleukin-1 receptor antagonist. *Journal of Translational Autoimmunity*, 7:100213, 2023.

[63] Stefania Zampatti, Juliette Farro, Cristina Peconi, Raffaella Cascella, Claudia Strafella, Giulia Calvino, Domenica Megalizzi, Giulia Trastulli, Carlo Caltagirone, and Emiliano Giardina. Ai-powered neurogenetics: Supporting patient's evaluation with chatbot. *Genes*, 16(1), 2025.

[64] Iyad Sultan, Haneen Al-Abdallat, Zaina Alnajjar, Layan Ismail, Razan Abukhashabeh, Layla Bitar, and Mayada Abu Shanap. Using ChatGPT to predict cancer predisposition genes: A promising tool for pediatric oncologists. *Cureus*, 15(10):e47594, October 2023.

[65] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1):60–64, Jan 2019.

[66] Da Wu, Jingye Yang, Cong Liu, Tzung-Chien Hsieh, Elaine Marchi, Justin Blair, Peter Krawitz, Chunhua Weng, Wendy Chung, Gholson J. Lyon, Ian D. Krantz, Jennifer M. Kalish, and Kai Wang. Gestaltmml: Enhancing rare genetic disease diagnosis through multimodal machine learning combining facial images and clinical texts, 2024.

[67] Bangwei Guo, Xingyu Li, Miaomiao Yang, Hong Zhang, and Xu Steven Xu. A robust and lightweight deep attention multiple instance learning algorithm for predicting genetic alterations. *Computerized Medical Imaging and Graphics*, 105:102189, 2023.

[68] Gexin Huang, Chenfei Wu, Mingjie Li, Xiaojun Chang, Ling Chen, Ying Sun, Shen Zhao, Xiaodan Liang, and Liang Lin. Predicting genetic mutation from whole slide images via biomedical-linguistic knowledge enhanced multi-label classification, 2024.

[69] Kai Sun, Yuanjie Zheng, Xinbo Yang, and Weikuan Jia. A novel transformer-based aggregation model for predicting gene mutations in lung adenocarcinoma. *Medical & Biological Engineering & Computing*, 62(5):1427–1440, May 2024.

[70] Vivek Kumar Singh, Yasmine Makhlouf, Md Mostafa Kamal Sarker, Stephanie Craig, Juvenal Baena, Christine Greene, Lee Mason, Jacqueline A James, Manuel Salto-Tellez, Paul O'Reilly, and Perry Maxwell. KRASFormer: a fully vision transformer-based framework for predictingKRASgene mutations in histopathological images of colorectal cancer. *Biomed Phys Eng Express*, 10(5), July 2024.

[71] Farhan Akram, Daniël P de Bruyn, Quincy C C van den Bosch, Teodora E Trandafir, Thierry P P van den Bosch, Rob M Verdijk, Annelies de Klein, Emine Kiliç, Andrew P Stubbs, Erwin Brosens, and Jan H von der Thüsen. Prediction of molecular subclasses of uveal melanoma by deep learning using routine haematoxylin-eosin-stained tissue slides. *Histopathology*, 85(6):909–919, July 2024.

[72] Jisen Guo, Peng Xu, Yuankui Wu, Yunyun Tao, Chu Han, Jiatai Lin, Ke Zhao, Zaiyi Liu, Wenbin Liu, and Cheng Lu. Cromam: A cross-magnification attention feature fusion model for predicting genetic status and survival of gliomas using histological images. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7345–7356, 2024.

[73] Ching-Wei Wang, Tzu-Chien Liu, Po-Jen Lai, Hikam Muzakky, Yu-Chi Wang, Mu-Hsien Yu, Chia-Hua Wu, and Tai-Kuang Chao. Ensemble transformer-based multiple instance learning to predict pathological subtypes and tumor mutational burden from histopathological whole slide images of endometrial and colorectal cancer. *Medical Image Analysis*, 99:103372, 2025.

[74] Raktim Kumar Mondol, Ewan K. A. Millar, Arcot Sowmya, and Erik Meijering. Biofusionnet: Deep learning-based survival risk stratification in er+ breast cancer through multifeature and multimodal data fusion. *IEEE Journal of Biomedical and Health Informatics*, 28(9):5290–5302, September 2024.

[75] Diego Machado Reyes, Hanqing Chao, Juergen Hahn, Li Shen, Pingkun Yan, and for the Alzheimer's Disease Neuroimaging Initiative. Identifying progression-specific alzheimer's subtypes using multimodal transformer. *Journal of Personalized Medicine*, 14(4), 2024.

[76] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[77] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

[78] Kejun Ying, Jinyeop Song, Haotian Cui, Yikun Zhang, Siyuan Li, Xingyu Chen, Hanna Liu, Alec Eames, Daniel L McCartney, Riccardo E Marioni, Jesse R Poganik, Mahdi Moqri, Bo Wang, and Vadim N Gladyshev. MethylGPT: a foundation model for the DNA methylome. November 2024.

[79] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, Feb 2025.

[80] Hongyang Li, Sanjoy Dey, Bum Chul Kwon, Michael Danziger, Michal Rosen-Tzvi, Jianying Hu, James Kozloski, Ching-Huei Tsou, Bharath Dandala, and Pablo Meyer. Bmfm-dna: A snp-aware dna foundation model to capture variant effects, 2025.

[81] Diego Machado Reyes, Myson Burch, Laxmi Parida, and Aritra Bose. A multimodal foundation model for discovering genetic associations with brain imaging phenotypes. *medRxiv*, 2024.

[82] Benedikt A. Jónsson, Gísli H. Halldórsson, Steinþór Árdal, Sölvi Rögnvaldsson, Eyþór Einarsson, Patrick Sulem, Daníel F. Guðbjartsson, Páll Melsted, Kári Stefánsson, and Magnús Ö. Úlfarsson. Transformers significantly improve splice site prediction. *Communications Biology*, 7(1):1616, Dec 2024.

[83] Zhuoran Xu, Quan Li, Luigi Marchionni, and Kai Wang. PhenoSV: interpretable phenotype-aware model for the prioritization of genes affected by structural variants. *Nat Commun*, 14(1):7805, November 2023.

[84] Matt C. Danzi, Maike F. Dohrn, Sarah Fazal, Danique Beijer, Adriana P. Rebelo, Vivian Cintra, and Stephan Züchner. Deep structured learning for variant prioritization in mendelian diseases. *Nature Communications*, 14(1):4167, Jul 2023.

[85] Lungang Liang, Yulan Chen, Taifu Wang, Dan Jiang, Jishuo Jin, Yanmeng Pang, Qin Na, Qiang Liu, Xiaosen Jiang, Wentao Dai, Meifang Tang, Yutao Du, Dirong Peng, Xin Jin, and Lijian Zhao. Genetic transformer: An innovative large language model driven approach for rapid and accurate identification of causative variants in rare genetic diseases. *medRxiv*, 2024.

[86] Youssef Boulaimen, Gabriele Fossi, Leila Outemzabet, Nathalie Jeanray, Oleksandr Levenets, Stephane Gerart, Sebastien Vachenc, Salvatore Raieli, and Joanna Giemza. Integrating large language models for genetic variant classification, 2024.

[87] Shuangjia Lu and Erdal Cosgun. Boosting gpt models for genomics analysis: generating trusted genetic variant annotations and interpretations through rag and fine-tuning. *Bioinformatics Advances*, 5(1):vbaf019, 02 2025.

[88] Wanwen Zeng, Hanmin Guo, Qiao Liu, and Wing Hung Wong. How to improve polygenic prediction from whole-genome sequencing data by leveraging predicted epigenomic features? *medRxiv*, 2024.

[89] Qiao Liu, Wanwen Zeng, Hongtu Zhu, Lexin Li, Wing Hung Wong, and Alzheimer's Disease Neuroimaging Initiative. Leveraging genomic large language models to enhance causal genotype-brain-clinical pathways in alzheimer's disease. *medRxiv*, 2024.

[90] Ingoo Lee, Zach Wallace, Sungjoon Park, Hojung Nam, Amit R. Majithia, and Trey Ideker. Mechanistic genotype-phenotype translation using hierarchical transformers. *bioRxiv*, 2024.

[91] Diego Machado Reyes, Mansu Kim, Hanqing Chaoh, Juergen Hahn, Li Shen, and Pingkun Yan. Genomics transformer for diagnosing parkinson's disease. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04, 2022.

[92] Mujahid Ali Quidwai and Alessandro Lagana. A rag chatbot for precision medicine of multiple myeloma. *medRxiv*, 2024.

[93] Anindita Nath, Savannah Mwesigwa, Yulin Dai, Xiaoqian Jiang, and Zhong-ming Zhao. GENEVIC: GENetic data exploration and visualization via intelli- gent interactive console. *Bioinformatics*, page btae500, 08 2024.

[94] Stefan Lukac, Davut Dayan, Visnja Fink, Elena Leinert, Andreas Hartkopf, Kristina Veselinovic, Wolfgang Janni, Brigitte Rack, Kerstin Pfister, Benedikt Heitmeir, and Florian Ebner. Evaluating chatgpt as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Archives of Gynecology and Obstetrics*, 308(6):1831–1844, Dec 2023.

[95] Katherine J Hewitt, Isabella C Wiest, Zunamys I Carrero, Laura Bejan, Thomas O Millner, Sebastian Brandner, and Jakob Nikolas Kather. Large language models as a diagnostic support tool in neuropathology. *The Journal of Pathology: Clinical Research*, 10(6):e70009, 2024.

[96] Zacharie Hamilton, Aseem Aseem, Zhengjia Chen, Noor Naffakh, Natalie M Reizine, Frank Weinberg, Shikha Jain, Larry G Kessler, Vijayakrishna K Gadi, Christopher Bun, and Ryan H Nguyen. Comparative analysis of generative Pre-Trained transformer models in Oncogene-Driven Non-Small cell lung cancer: Introducing the generative artificial intelligence performance score. *JCO Clin Cancer Inform*, 8:e2400123, December 2024.

[97] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[98] Rebekah L Waikel, Amna A Othman, Tanviben Patel, Suzanna Ledgister Hanchard, Ping Hu, Cedrik Tekendo-Ngongang, Dat Duong, and Benjamin D Solomon. Recognition of genetic conditions after learning with images created using generative artificial intelligence. *JAMA Netw. Open*, 7(3):e242609, March 2024.

[99] Jiaxin Yang, Xinhao Zhuang, Zhenqi Li, Gang Xiong, Ping Xu, Yunchao Ling, and Guoqing Zhang. Cpmkg: a condition-based knowledge graph for precision medicine. *Database*, 2024:baae102, 09 2024.

[100] Kulaga Anton, Borysova Olga, Karmazin Alexey, Koval Maria, Usanov Nikolay, Fedorova Alina, Evfratov Sergey, Pushkareva Malvina, Ryangguk Kim, and Tacutu Robi. Just-dna-seq, open-source personal genomics platform: longevity science for everyone, 2024.

[101] Mullai Murugan, Bo Yuan, Eric Venner, Christie M Ballantyne, Katherine M Robinson, James C Coons, Liwen Wang, Philip E Empey, and Richard A Gibbs. Empowering personalized pharmacogenomics with generative AI solutions. *J Am Med Inform Assoc*, 31(6):1356–1366, May 2024.

[102] K Keat, R Venkatesh, Y Huang, R Kumar, S Tuteja, K Sangkuhl, B Li, L Gong, M Whirl-Carrillo, T E Klein, M D Ritchie, and D Kim. PGxQA: A resource for evaluating LLM performance for pharmacogenomic QA tasks. *Pac Symp Biocomput*, 30:229–246, 2025.

[103] Scott P McGrath, Beth A Kozel, Sara Gracefo, Nykole Sutherland, Christopher J Danford, and Nephi Walton. A comparative evaluation of ChatGPT 3.5 and ChatGPT 4 in responses to selected genetics questions. *J Am Med Inform Assoc*, 31(10):2271–2283, October 2024.

[104] Takuya Fukushima, Masae Manabe, Shuntaro Yada, Shoko Wakamiya, Akiko Yoshida, Yusaku Urakawa, Akiko Maeda, Shigeyuki Kan, Masayo Takahashi, and Eiji Aramaki. Evaluating and enhancing japanese large language models for genetic counseling support: Comparative study of domain adaptation and the development of an expert-evaluated dataset. *JMIR Med. Inform.*, 13:e65047, January 2025.

[105] Jharna M. Patel, Catherine E. Hermann, Whitfield B. Growdon, Emeline Aviki, and Marina Stasenko. Chatgpt accurately performs genetic counseling for gynecologic cancers. *Gynecologic Oncology*, 183:115–119, 2024.

[106] Nephi Walton, Sara Gracefo, Nykole Sutherland, Beth A. Kozel, Christopher J. Danford, and Scott P. McGrath. Evaluating chatgpt as an agent for providing genetic education. *bioRxiv*, 2023.

[107] Bryan Naidenov and Charles Chen. Gene-language models are whole genome representation learners. *bioRxiv*, 2024.

[108] Zhufeng Li, Sandeep S Cranganore, Nicholas Youngblut, and Niki Kilbertus. Whole genome transformer for gene interaction effects in microbiome habitat specificity, 2025.

[109] Laura Weinstock, Jenna Schambach, Anna Fisher, Cameron Kunstadt, Ethan Lee, Elizabeth Koning, William Morrell, Wittney Mays, Warren Davis, and Raga Krishnakumar. A hybrid machine learning model for predicting gene expression from epigenetics across fungal species. *bioRxiv*, 2024.

[110] Azwad Tamir and Jiann-Shiun Yuan. Protgo: A transformer based fusion model for accurately predicting gene ontology (go) terms from full scale protein sequences, 2024.

[111] Yin Li, Kaiyi Zheng, Shuang Li, Yongju Yi, Min Li, Yufan Ren, Congyue Guo, Liming Zhong, Wei Yang, Xinming Li, and Lin Yao. A transformer-based multi-task deep learning model for simultaneous infiltrated brain area identification and segmentation of gliomas. *Cancer Imaging*, 23(1):105, Oct 2023.

[112] Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez, Humaira Noor, Wei Yao, Christian Wohlfart, Antoaneta Vladimirova, Kathleen Marchal, and Olivier Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15(1):9886, Nov 2024.

[113] Ziwei Hu, Jianchao Wang, Qinquan Gao, Zhida Wu, Hanchuan Xu, Zhechen Guo, Jiawei Quan, Lihua Zhong, Ming Du, Tong Tong, and Gang Chen. Weakly supervised classification for nasopharyngeal carcinoma with transformer in whole slide images. *IEEE J Biomed Health Inform*, PP, July 2024.

[114] Karim Gasmi, Najib Ben Aoun, Khalaf Alsalem, Ibtihel Ben Ltaifa, Ibrahim Alrashdi, Lassaad Ben Ammar, Manel Mrabet, and Abdulaziz Shehab. Enhanced brain tumor diagnosis using combined deep learning models and weight selection technique. *Frontiers in Neuroinformatics*, Volume 18 - 2024, 2024.

[115] Ethan Hillis, Kriti Bhattarai, and Zachary Abrams. Evaluating generative ai's ability to identify cancer subtypes in publicly available structured genetic datasets. *Journal of Personalized Medicine*, 14(10), 2024.

[116] Ping Yang, Wengxiang Chen, and Hang Qiu. Mmgcn: Multi-modal multi-view graph convolutional networks for cancer prognosis prediction. *Computer Methods and Programs in Biomedicine*, 257:108400, 2024.

[117] Pratik Ramprasad, Nidhi Pai, and Wei Pan. Enhancing personalized gene expression prediction from dna sequences using genomic foundation models. *Human Genetics and Genomics Advances*, 5(4):100347, 2024.

[118] Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A. Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments, 2024.

[119] Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with sclambda. *bioRxiv*, 2024.

[120] Olesya Razuvayevskaya, Irene Lopez, Ian Dunham, and David Ochoa. Genetic factors associated with reasons for clinical trial stoppage. *Nature Genetics*, 56(9):1862–1867, Sep 2024.

[121] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[122] Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. Large language models for biomedical text simplification: Promising but not there yet, 2024.

[123] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models?, 2023.

[124] Priyadarshini Rai, Atishay Jain, Shivani Kumar, Divya Sharma, Neha Jha, Smriti Chawla, Abhijit Raj, Apoorva Gupta, Sarita Poonia, Angshul Majumdar, Tanmoy Chakraborty, Gaurav Ahuja, and Debarka Sengupta. Literature mining discerns latent disease–gene relationships. *Bioinformatics*, 40(4):btae185, 04 2024.

[125] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33, 2020.

[126] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 02 2024.

[127] Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale, 2025.

[128] Yinghua Fu, Junfeng Liu, and Jun Shi. Tsca-net: Transformer based spatial-channel attention segmentation network for medical images. *Computers in Biology and Medicine*, 170:107938, 2024.

[129] Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: A family of open-source foundational models for long dna sequences. *bioRxiv*, 2023.

[130] Zhe Liu, Wei Qian, Wenxiang Cai, Weichen Song, Weidi Wang, Dhruba Tara Maharjan, Wenhong Cheng, Jue Chen, Han Wang, Dong Xu, and Guan Ning Lin. Inferring the effects of protein variants on Protein-Protein interactions with interpretable transformer representations. *Research (Wash D C)*, 6:0219, September 2023.

[131] Yuansong Zeng, Jiancong Xie, Ningyuan Shangguan, Zhuoyi Wei, Wenbing Li, Yun Su, Shuangyu Yang, Chengyang Zhang, Jinbo Zhang, Nan Fang, Hongyu Zhang, Yutong Lu, Huiying Zhao, Jue Fan, Weijiang Yu, and Yuedong Yang. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16(1):4679, May 2025.

[132] Ling Zhang, Boxiang Yun, Xingran Xie, Qingli Li, Xinxing Li, and Yan Wang. Prompting whole slide image based genetic biomarker prediction, 2024.

[133] Chao Xia, Jiyue Wang, Xin You, Yaling Fan, Bing Chen, Saijuan Chen, and Jie Yang. Chromtr: chromosome detection in raw metaphase cell images via deformable transformers. *Frontiers of Medicine*, 18(6):1100–1114, Dec 2024.

[134] Haoxi Zhang, Xinxu Zhang, Yuanxin Lin, Maiqi Wang, Yi Lai, Yu Wang, Linfeng Yu, Yufeng Xu, Ran Cheng, and Edward Szczerbicki. Tokensome: Towards a genetic vision-language gpt for explainable and cognitive karyotyping, 2024.

[135] Jiaying Zhou, Mingzhou Jiang, Junde Wu, Jiayuan Zhu, Ziyue Wang, and Yueming Jin. Mgi: Multimodal contrastive pre-training of genomic and medical imaging, 2024.

[136] Yaochen Xie, Ziqian Xie, Sheikh Muhammad Saiful Islam, Degui Zhi, and Shuiwang Ji. Genetic infomax: Exploring mutual information maximization in high-dimensional imaging genetics studies, 2023.

[137] Thomas Labbe, Pierre Castel, Jean-Michel Sanner, and Majd Saleh. ChatGPT for phenotypes extraction: one model to rule them all? *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2023:1–4, July 2023.

[138] Daniel B. Hier, S. Ilyas Munzir, Anne Stahlfeld, Tayo Obafemi-Ajayi, and Michael D. Carrithers. High-throughput phenotyping of clinical text using large language models, 2024.

[139] Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew E Fenech, Kasper Norman, Henrik Støvring, and Troels Krarup Hansen. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: A turing test-inspired survey among employees of a danish diabetes center. *PLoS One*, 18(8):e0290773, August 2023.

[140] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, Oct 2021.

[141] Gianluca Mondillo, Alessandra Perrotta, Simone Colosimo, and Vittoria Frattolillo. Chatgpt as a bioinformatic partner. *medRxiv*, 2024.

[142] Mohamad-Hani Temsah, Amr Jamal, Khalid Alhasan, Abdulkarim A Temsah, and Khalid H Malki. OpenAI o1-preview vs. ChatGPT in healthcare: A new frontier in medical AI reasoning. *Cureus*, 16(10):e70640, October 2024.

[143] Ismail Chahid, Aissa Kerkour Elmiad, and Mohammed Badaoui. Data preprocessing for machine learning applications in healthcare: A review. In *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–6, 2023.

[144] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, Jul 2020.

[145] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[146] Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. Adapters: A unified library for parameter-efficient and modular transfer learning, 2023.

[147] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[148] Nikita Mehandru, Amanda K. Hall, Olesya Melnichenko, Yulia Dubinina, Daniel Tsirulnikov, David Bamman, Ahmed Alaa, Scott Saponas, and Venkat S. Malladi. Bioagents: Democratizing bioinformatics analysis with multi-agent systems, 2025.

[149] Sebastian Lobentanzer, Shaohong Feng, Noah Bruderer, Andreas Maier, BioChatter Consortium, Cankun Wang, Jan Baumbach, Jorge Abreu-Vicente, Nils Krehl, Qin Ma, Thomas Lemberger, and Julio Saez-Rodriguez. A platform for the biomedical application of large language models. *Nat Biotechnol*, 43(2):166–169, February 2025.

[150] Hyunghoon Cho, David Froelicher, Jeffrey Chen, Manaswitha Edupalli, Apostolos Pyrgelis, Juan R. Troncoso-Pastoriza, Jean-Pierre Hubaux, and Bonnie Berger. Secure and federated genome-wide association studies for biobank-scale datasets. *Nature Genetics*, 57(4):809–814, Apr 2025.

[151] Owen Bianchi, Maya Willey, Chelsea X. Alvarado, Benjamin Danek, Marzieh Khani, Nicole Kuznetsov, Anant Dadu, Syed Shah, Mathew J. Koretsky, Mary B. Makarious, Cory Weller, Kristin S. Levine, Sungwon Kim, Paige Jarreau, Dan Vitale, Elise Marsan, Hirotaka Iwaki, Hampton Leonard, Sara Bandres-Ciga, Andrew B Singleton, Mike A Nalls, Shekoufeh Mokhtari, Daniel Khashabi, and Faraz Faghri. Cardbiomedbench: A benchmark for evaluating large language model performance in biomedical research. *bioRxiv*, 2025.

[152] SWJ Nijman, AM Leeuwenberg, I Beekers, I Verkouter, JJL Jacobs, ML Bots, FW Asselbergs, KGM Moons, and TPA Debray. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*, 142:218–229, 2022.

[153] Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, and Nan Liu. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine*, 142:102587, 2023.

[154] Sarthak Pati, Sourav Kumar, Amokh Varma, Brandon Edwards, Charles Lu, Liangqiong Qu, Justin J Wang, Anantharaman Lakshminarayanan, Shih-Han Wang, Micah J Sheller, Ken Chang, Praveer Singh, Daniel L Rubin, Jayashree Kalpathy-Cramer, and Spyridon Bakas. Privacy preservation for federated learning in health care. *Patterns (N Y)*, 5(7):100974, July 2024.

[155] Patrick Rockenschaub, Ela Marie Akay, Benjamin Gregory Carlisle, Adam Hilbert, Joshua Wendland, Falk Meyer-Eschenbach, Anatol-Fiete Näher, Dietmar Frey, and Vince Istvan Madai. External validation of ai-based scoring systems in the icu: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 25(1):5, Jan 2025.

[156] Miranda Durkie, Emma-Jane Cassidy, Ian Berry, Martina Owens, Clare Turnbull, Richard H Scott, Robert W Taylor, Zandra C Deans, Sian Ellard, Emma L Baple, and Dominic J McMullan. Acgs best practice guidelines for variant classification in rare disease 2024 (v1.2). Best practice guidelines, Association for Clinical Genomic Science (ACGS), United Kingdom, February 2024. Ratified by ACGS Quality Subcommittee on 20 Feb 2024.

[157] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023.

[158] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation, 2024.

[159] Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveen K Kanithi, Marco AF Pimentel, Ronnie Rajan, and Shadab Khan. Genomic foundationless models: Pretraining does not promise performance. *bioRxiv*, 2024.

[160] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[161] Arya Hadizadeh Moghaddam, Mohsen Nayebi Kerdabadi, Mei Liu, and Zijun Yao. Contrastive learning on medical intents for sequential prescription recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 748–757. ACM, October 2024.

[162] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey, 2025.

[163] Joseph Lee, Shu Yang, Jae Young Baik, Xiaoxi Liu, Zhen Tan, Dawei Li, Zixuan Wen, Bojian Hou, Duy Duong-Tran, Tianlong Chen, and Li Shen. Knowledge-driven feature selection and engineering for genotype data with large language models, 2025.

[164] J. van Uhm, M.M. van Haelst, and P.R. Jansen. Ai-powered test question generation in medical education: The dailymed approach. *medRxiv*, 2024.

[165] Emma Coen, Guilherme Del Fiol, Kimberly A Kaphingst, Emerson Borsato, Jackie Shannon, Hadley Stevens Smith, Aaron Masino, and Caitlin G Allen. Chatbot for the return of positive genetic screening results for hereditary cancer syndromes: a prompt engineering study. August 2024.

[166] Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments, 2025.

[167] Nicole L. Walters, Zoe T. Lindsey-Mills, Andrew Brangan, Sarah K. Savage, Tara J. Schmidlen, Kelly M. Morgan, Eric P. Tricou, Megan M. Betts, Laney K. Jones, Amy C. Sturm, and Gemme Campbell-Salome. Facilitating family communication of familial hypercholesterolemia genetic risk: Assessing engagement with innovative chatbot technology from the impact-fh study. *PEC Innovation*, 2:100134, 2023.

[168] Moein Amin, Eloy Martínez-Heras, Daniel Ontaneda, and Ferran Prados Carrasco. Artificial intelligence and multiple sclerosis. *Curr Neurol Neurosci Rep*, 24(8):233–243, June 2024.

[169] Giulia Calvino, Cristina Peconi, Claudia Strafella, Giulia Trastulli, Domenica Megalizzi, Sarah Andreucci, Raffaella Cascella, Carlo Caltagirone, Stefania Zampatti, and Emiliano Giardina. Federated learning: Breaking down barriers in global genomic research. *Genes*, 15(12), 2024.

[170] Dmitry Kolobkov, Satyarth Mishra Sharma, Aleksandr Medvedev, Mikhail Lebedev, Egor Kosaretskiy, and Ruslan Vakhitov. Efficacy of federated learning on genomic data: a study on the uk biobank and the 1000 genomes project. *Frontiers in Big Data*, Volume 7 - 2024, 2024.

[171] Constantine Tarabanis, Sohail Zahid, Marios Mamalis, Kevin Zhang, Evangelos Kalampokis, and Lior Jankelson. Performance of publicly available large language models on internal medicine board-style questions. *PLOS Digit. Health*, 3(9):e0000604, September 2024.

[172] Jack W. O'Sullivan, Anil Palepu, Khaled Saab, Wei-Hung Weng, Yong Cheng, Emily Chu, Yaanik Desai, Aly Elezaby, Daniel Seung Kim, Roy Lan, Wilson Tang, Natalie Tapaskar, Victoria Parikh, Sneha S. Jain, Kavita Kulkarni, Philip Mansfield, Dale Webster, Juraj Gottweis, Joelle Barral, Mike Schaekermann, Ryutaro Tanno, S. Sara Mahdavi, Vivek Natarajan, Alan Karthikesalingam, Euan Ashley, and Tao Tu. Towards democratization of subspeciality medical expertise, 2024.

[173] Xin Zhou, Martin Weyssow, Ratnadira Widyasari, Ting Zhang, Junda He, Yunbo Lyu, Jianming Chang, Beiqi Zhang, Dan Huang, and David Lo. Lessleak-bench: A first investigation of data leakage in llms across 83 software engineering benchmarks, 2025.

[174] Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need, 2025.

[175] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore, 2023.

[176] Michal Golovanevsky, Eva Schiller, Akira Nair, Eric Han, Ritambhara Singh, and Carsten Eickhoff. One-versus-others attention: Scalable multimodal integration for biomedical data, 2024.

[177] Reza Shirkavand, Liang Zhan, Heng Huang, Li Shen, and Paul M. Thompson. Incomplete multimodal learning for complex brain disorders prediction, 2023.

[178] Lei Yuan, Jianhua Song, and Yazhuo Fan. MCNMF-Unet: a mixture Conv-MLP network with multi-scale features fusion unet for medical image segmentation. *PeerJ Comput Sci*, 10:e1798, January 2024.

[179] Pengcheng Shi, Xutao Guo, Yanwu Yang, Chenfei Ye, and Ting Ma. Nextou: Efficient topology-aware u-net for medical image segmentation, 2023.

[180] Antonia Alomar, Ricardo Rubio, Laura Salort, Gerard Albaiges, Antoni Payà, Gemma Piella, and Federico Sukno. Automatic facial axes standardization of 3d fetal ultrasound images, 2024.

[181] Edward Raff, Michel Benaroch, Sagar Samtani, and Andrew L. Farris. What do machine learning researchers mean by "reproducible"? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12812–12821. AAAI Press, 2025. Booz Allen Hamilton; University of Maryland, Baltimore County; Syracuse University; Indiana University.

[182] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Mag.*, 46(2), April 2025.

[183] Seong Ho Park, Chong Hyun Suh, Jeong Hyun Lee, Charles E Kahn, and Linda Moy. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*, 25(10):865–868, October 2024.

[184] Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(1):176, September 2024.

[185] Shu Lin, Saket Pandit, Tara Tritsch, Arkene Levy, and Mohammadali M Shoja. What goes in, must come out: Generative artificial intelligence does not present algorithmic bias across race and gender in medical residency specialties. *Cureus*, 16(2):e54448, February 2024.

[186] Amna A. Othman, Kendall A. Flaharty, Suzanna E. Ledgister Hanchard, Ping Hu, Dat Duong, Rebekah L. Waikel, and Benjamin D. Solomon. Assessing large language model performance related to aging in genetic conditions. *medRxiv*, 2025.

[187] Gabriel Levin, Rene Pareja, David Viveros-Carreño, Emmanuel Sanchez Diaz, Elise Mann Yates, Behrouz Zand, and Pedro T Ramirez. Association of reviewer experience with discriminating human-written versus chatgpt-written abstracts. *International Journal of Gynecological Cancer*, 34(5):669–674, 2024.

[188] Leonardo Campillos-Llanos. Medlexsp – a medical lexicon for spanish medical natural language processing. *Journal of Biomedical Semantics*, 14(1):2, Feb 2023.

[189] Yixuan Weng, Bin Li, Fei Xia, Minjun Zhu, Bin Sun, Shizhu He, Kang Liu, and Jun Zhao. Large language models need holistically thought in medical conversational qa, 2023.

[190] Artur Francisco Schumacher-Schuh, Andrei Bieger, Olaitan Okunoye, Kin Ying Mok, Shen-Yang Lim, Soraya Bardien, Azlina Ahmad-Annuar, Bruno Lopes Santos-Lobato, Matheus Zschornack Strelow, Mohamed Salama, Shilpa C. Rao, Yared Zenebe Zewde, Saiesha Dindayal, Jihan Azar, Lingappa Kukkle Prashanth, Roopa Rajan, Alastair J. Noyce, Njideka Okubadejo, Mie Rizig, Suzanne Lesage, Ignacio Fernandez Mata, and Global Parkinson's Genetics Program (GP2). Underrepresented populations in parkinson's genetics research: Current landscape and future directions. *Movement Disorders*, 37(8):1593–1604, 2022.

[191] Benson R. Kidenya and Gerald Mboowa. Inclusiveness of the all of us research program improves polygenic risk scores and fosters genomic medicine for all. *Communications Medicine*, 4(1):227, Nov 2024.

[192] Kathryn Step, Carene Anne Alene Ndong Sima, Ignacio Mata, and Soraya Bardien. Exploring the role of underrepresented populations in polygenic risk scores for neurodegenerative disease risk prediction. *Frontiers in Neuroscience*, Volume 18 - 2024, 2024.

[193] Ellison B. Weiner, Irene Dankwa-Mullan, William A. Nelson, and Saeed Hassanpour. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. *PLOS Digital Health*, 4(4):1–12, 04 2025.

[194] Debesh Jha, Gorkem Durak, Abhijit Das, Jasmer Sanjotra, Onkar Susladkar, Suramyaa Sarkar, Ashish Rauniyar, Nikhil Kumar Tomar, Linkai Peng, Sirui Li, Koushik Biswas, Ertugrul Aktas, Elif Keles, Matthew Antalek, Zheyuan Zhang, Bin Wang, Xin Zhu, Hongyi Pan, Deniz Seyithanoglu, Alpay Medetalibeyoglu, Vanshali Sharma, Vedat Cicek, Amir A. Rahsepar, Rutger Hendrix, A. Enis Cetin, Bulent Aydogan, Mohamed Abazeed, Frank H. Miller, Rajesh N. Keswani, Hatice Savas, Sachin Jambawalikar, Daniela P. Ladner, Amir A. Borhani, Concetto Spampinato, Michael B. Wallace, and Ulas Bagci. Ethical framework for responsible foundational models in medical imaging. *Frontiers in Medicine*, Volume 12 - 2025, 2025.

8 Appendix

8.1 Appendix A: Supplementary Tables - Annotated Article Dataset

Two supplementary tables were compiled to support the analysis presented in this study. Supplementary Table 1 contains the complete list of articles included after initial collection, deduplication, and manual verification. Supplementary Table 2 provides an extended, manually annotated version of the dataset with additional semantic tags and classification columns.

Supplementary Table 1 (ST1) presents the cleaned dataset after the removal of duplicates and initial triage. Duplicate entries were identified not only through automatic preprocessing but also through joint manual assessment by two researchers, ensuring a consistent and conservative approach to inclusion. ST1 includes metadata such as the article title, abstract, source, review status, and initial relevance tag.

Supplementary Table 2 (ST2) expands upon this initial dataset by including additional annotations used in the systematic analysis. These include fine-grained labels for specific tasks inside these stages (`final_category`, `subcategory`), and three binary relevance flags (`not_relevant`, `partly_relevant`, `relevant`). 27 manually selected articles were also added at this stage (eight highly relevant and 19 partially relevant), resulting in a total of 325 articles in ST2. These additions were motivated by expert review and targeted searches within the originally collected corpus and cited references.

Detailed descriptions of column meanings and classification codes are available in the project GitHub repository[1].

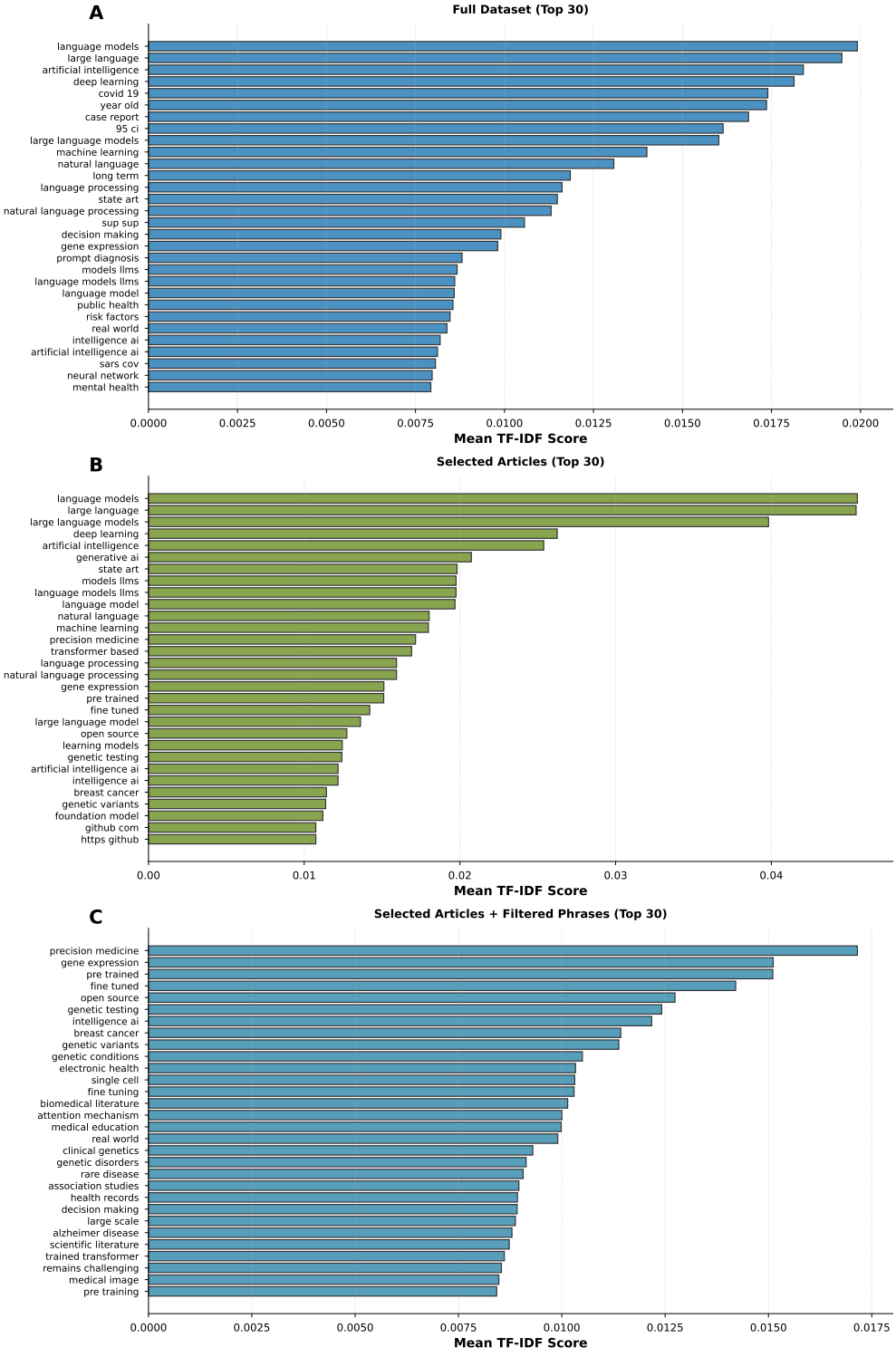
8.2 Appendix B: F-IDF and Filtering Methods

To characterize the semantic landscape of LLM applications in medical genomics, we employed two complementary text mining approaches: Term Frequency-Inverse Document Frequency (TF-IDF) analysis for identifying domain-specific terminology, and Latent Dirichlet Allocation (LDA) for discovering latent thematic structure. Both methods were applied to article titles and abstracts from the curated dataset.

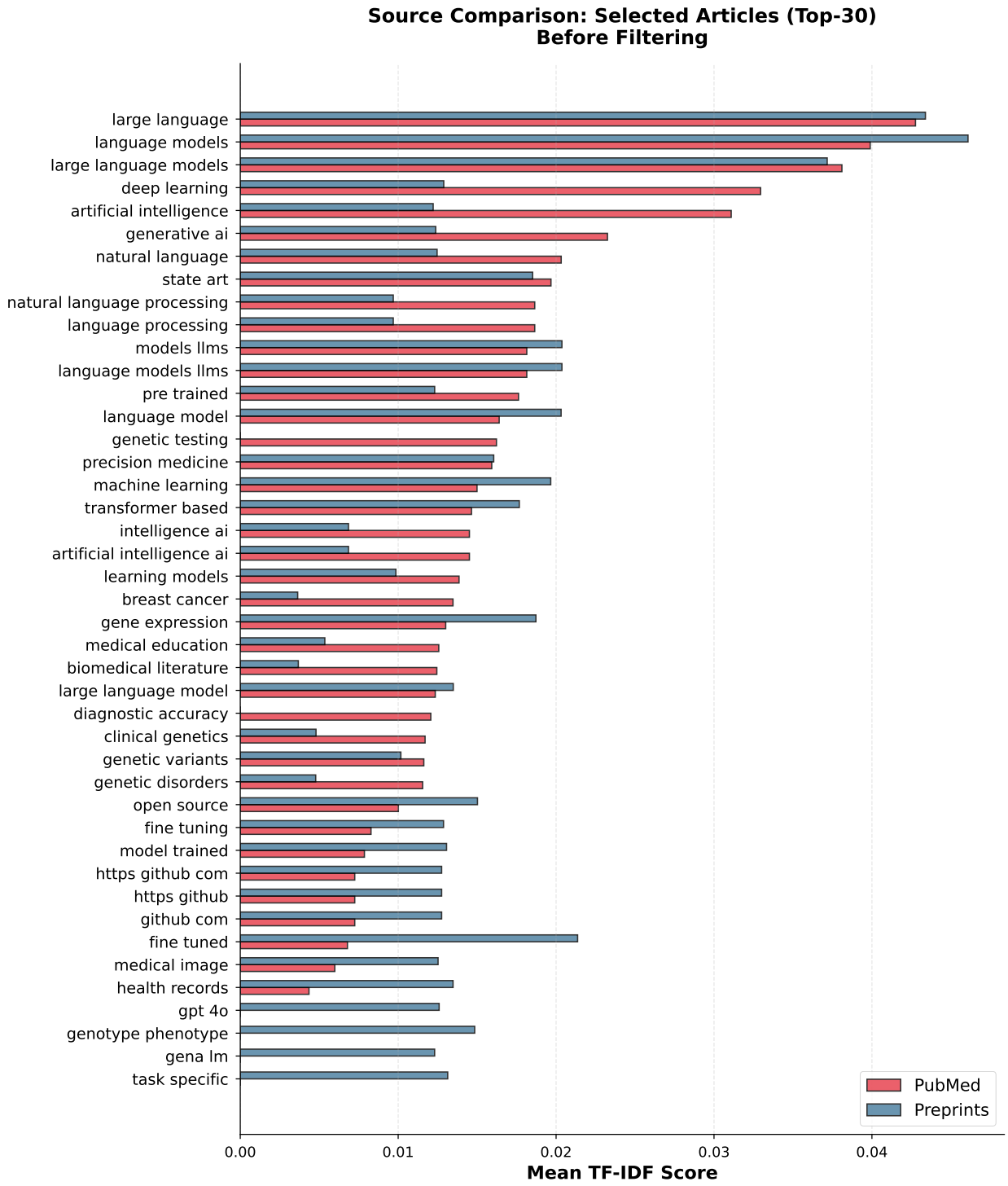
8.2.1 Term Frequency-Inverse Document Frequency (TF-IDF) Analysis

As mentioned earlier, TF-IDF helped to identify areas of the research in applications of LLMs. It was applied at multiple stages: the full corpus (51,613 records after deduplication), the curated review set (195 articles), and filtered variants where generic AI/ML phrases were removed (to move beyond obvious LLM keywords such as "language model", "deep learning" – full pattern list in the source code).
Bigram-trigram TF-IDF scores were computed (scikit-learn's `TfidfVectorizer` with `ngram_range=(2,3)` and `max_features=1000`), lower-casing and removing English stop-words plus custom artifacts (e.g., "et al").
Additionally, a context-preserving fine-tuning approach was implemented. This method first trains the TF-IDF model on the curated dataset (195 articles) with full terminology context, then applies post-hoc reweighting by zeroing out generic AI/ML anchor terms and renormalizing the document vectors. This preserves the semantic context during initial feature extraction while down-weighting generic phrases in the final ranking. The fine-tuned analysis confirms that the domain-specific trends reported in the main text (e.g., precision medicine, gene expression, genetic testing) remain stable once obvious anchors are de-emphasized, demonstrating that our findings are robust to different filtering strategies.
Sources within the curated set were compared by stratifying PubMed ($n = 131$) versus preprints (bioRxiv/medRxiv/arXiv; $n = 64$). For each comparison, the union of the two top-30 lists was applied. This helped to capture the shift from generic to domain-specific terminology and highlight complementary emphases between peer-reviewed and preprint venues.
Additional representations of TF-IDF analysis are shown in four supplementary figures: Supplementary Figure 1 (three-stage progression of TF-IDF after selecting articles and filtering words), Supplementary Figure 2 (comparison of sources before filtering), Supplementary Figure 3 (comparison of sources after filtering), and Supplementary Figure 4 (comparison of sources using fine-tuned analysis).
In addition to standard English stop-words, generic AI/ML phrases and artifacts were excluded to surface domain-specific terminology. The final list included the following common terms: large language, language model, llm, llms, generative ai, foundation model, foundation models, deep learning, deep neural, neural network, neural networks, machine learning, artificial intelligence, artificial neural, natural language, language processing, nlp, transformer model, transformer models, reinforcement learning, supervised learning, unsupervised learning, state art, based, using, https, github, model, models, learning, data.

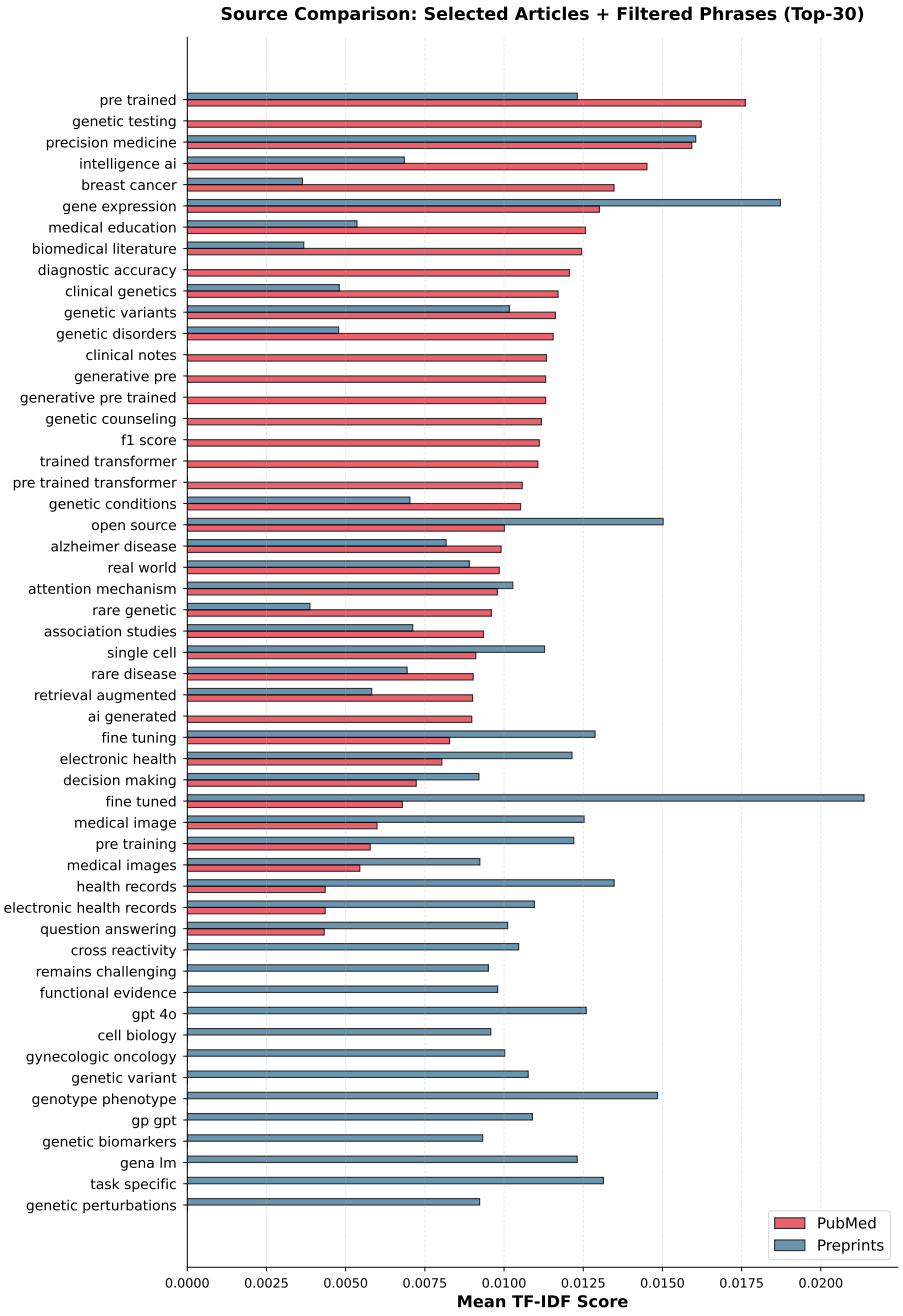
[1] `https://github.com/TohaRhymes/llm_in_diagnostics`



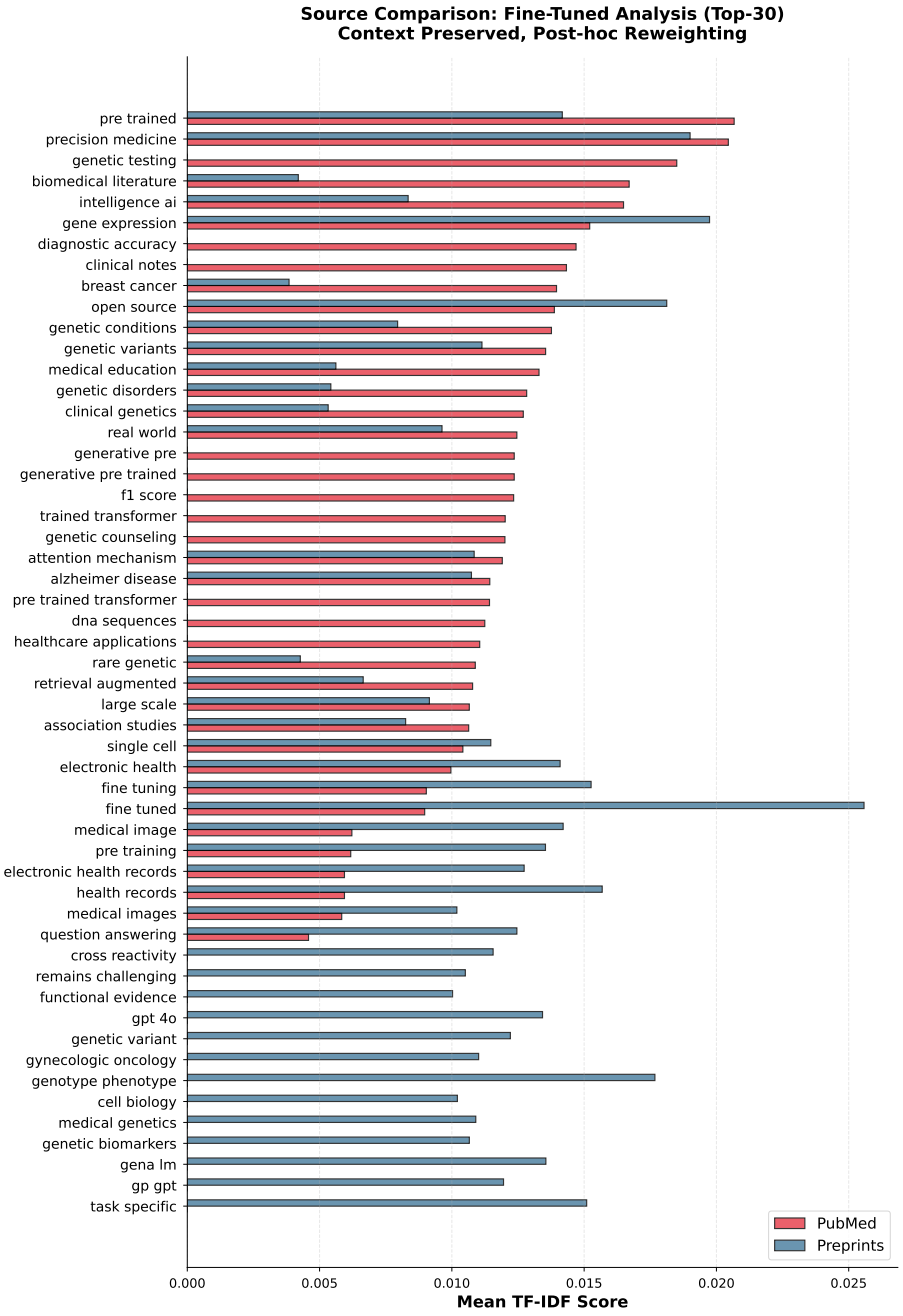
Supplementary Figure 1: Progression of TF-IDF analysis from full corpus to filtered insights. (A) Full dataset (51,613 articles): generic anchors dominate; (B) Selected articles (195): core themes retained; (C) Selected articles + Filtered words: domain-specific trends (e.g., precision medicine, gene expression, human phenotype ontology, single cell) become prominent. Horizontal bars show the top-30 phrases per stage.



Supplementary Figure 2: Source comparison before filtering generic phrases. Grouped bars show TF-IDF scores on the same scale for the union of top-30 phrases across PubMed ($n = 128$) and preprints ($n = 64$). Strong overlap (57%) indicates consensus on core topics before filtering.



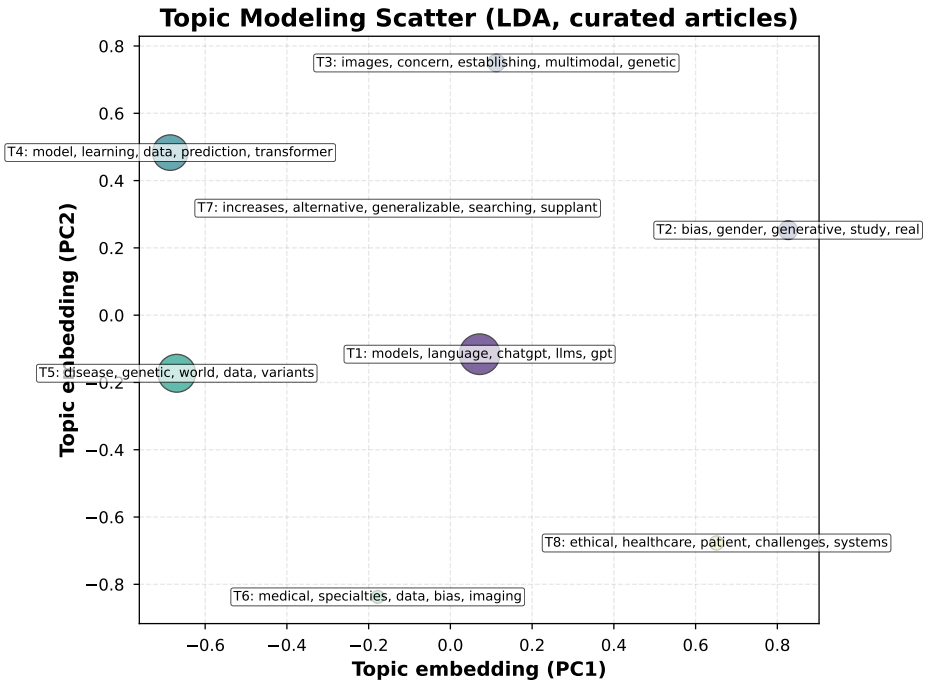
Supplementary Figure 3: Source comparison of research trends after article selection and filtering of generic AI/ML phrases. Grouped bars show actual TF-IDF scores for the union of top-30 phrases from PubMed ($n = 128$) and preprints ($n = 64$) on the same scale. Overlap drops to 23%, revealing distinct emphases: PubMed emphasizes clinical/translational terms while preprints highlight computational methods.



Supplementary Figure 4: Source comparison using fine-tuned analysis (context preserved, post-hoc reweighting). This analysis includes training TF-IDF with full context, then down-weighting generic AI/ML terms. Results confirm that domain-specific trends remain stable (27% overlap between sources), validating the filtered analysis approach. The fine-tuned method preserves semantic relationships while surfacing specific research emphases.

8.2.2 Latent Dirichlet Allocation (LDA) Topic Modeling

To address the visual representation of topic overlap, we performed Latent Dirichlet Allocation (LDA) topic modeling on the curated dataset. Eight topics were extracted using gensim with automatic hyperparameter optimization. Topic similarity was quantified using Jensen-Shannon divergence between topic-word distributions, and topics were arranged in 2D space using a custom force-directed layout algorithm that positions similar topics closer together. The resulting scatter plot (Supplementary Figure 5) displays topics as bubbles sized by prevalence, with labels showing the top five terms per topic. This visualization provides an intuitive view of how topics relate to each other and their relative importance in the literature.



Supplementary Figure 5: Topic modeling visualization showing overlap and relationships between themes. Eight LDA topics fitted on the curated dataset ($n = 195$) are displayed in 2D space using Jensen-Shannon divergence-based layout. Bubble size reflects topic prevalence. Labels show top five terms per topic. Topics closer together share more vocabulary, illustrating the semantic landscape of LLM applications in medical genomics.

Supplementary Table 3 (ST3) presents topic modeling metadata from LDA analysis, including topic identifiers, visualization coordinates, prevalence values, and the top five representative terms for each of eight identified topics. Supplementary Table 4 (ST4) provides the document-topic probability distribution matrix, where each row represents one article from ST2 and columns represent the probability of assignment to each of the eight topics, plus source categorization.