Society of Medical Education & Research

2026, Volume 6, Issue 1, Page No: 71-92

ISSN: 3108-4826

Journal of Medical Sciences and Interdisciplinary Research

Large Language Model–Driven Automation of Microbiome Diagnostic Reporting in Clinical Laboratories

Mateo Alvarez[1]*, Sofia Herrera[2], Diego Cruz[1], Andres Castro[3]

[1] Department of Clinical Microbiology and LLM Applications, Faculty of Medicine, National University of Colombia, Bogota, Colombia.

[2] Department of Diagnostic Reporting and Natural Language Processing, Faculty of Medicine, University of Chile, Santiago, Chile.

[3] Department of Automation in Laboratory Medicine, Faculty of Medicine, University of Antioquia, Medellin, Colombia.

***E-mail** ✉ mateo.alvarez@unal.edu.co

Abstract

Rapid progress in genomic technologies is reshaping laboratory diagnostics by enabling high-throughput analysis of complex biological information, notably microbiome profiles. Large Language Models (LLMs) have exhibited strong potential for uncovering actionable knowledge from extensive datasets. However, their capacity to produce microbiome findings reports that include clinical interpretations and personalized lifestyle advice has not yet been investigated. This study introduces a novel framework that harnesses LLMs to automate the creation of findings reports for microbiome diagnostics. The framework embeds LLMs in an event-driven, workflow-centric system designed to promote scalability and adaptability in clinical laboratory operations. Emphasis is placed on ensuring conformity with established clinical norms and regulatory frameworks, including the In Vitro Diagnostic Regulation (IVDR) and recommendations from the High-Level Expert Group on Artificial Intelligence (HLEG AI). A prototype implementation, designated “MicroFlow”, was developed to illustrate the approach. Deployment of the MicroFlow prototype confirms the practicality of using LLMs to automatically generate findings reports. Early feedback from laboratory specialists indicates that LLM integration yields promising outcomes, with the resulting reports judged to be credible and clinically relevant. Nonetheless, additional validation employing authentic clinical datasets is required to determine the system’s accuracy and robustness. The present work demonstrates a viable pathway for applying LLMs to support findings report generation in microbiome diagnostics. Although initial findings are encouraging, continued refinement and rigorous evaluation are imperative to validate the model’s performance and compliance with clinical requirements. Planned future activities will incorporate expert input from laboratory professionals and extensive testing on real-world patient samples.

Keywords: Genomics, Diagnostics, AI, LLM, Bbig data, Microbiome

Introduction

Laboratory diagnostics are an essential pillar of modern medicine, supporting disease detection and treatment. The rise of genomics has dramatically extended the capabilities of diagnostic laboratories, transforming them into data-intensive environments. A key example is the investigation of human microbiomes—complex microbial ecosystems that inhabit various sites within and on the human body. These communities are essential for maintaining health but can also contribute to illness when their equilibrium is disturbed [1]. In particular, the gut microbiome has emerged as a valuable target for diagnostic testing, with compositional shifts associated with multiple disorders including cancer, inflammatory bowel disease, and cardiovascular conditions [1]. Genomic analysis of these microbial communities is known as metagenomics.

Interpreting genomics-derived diagnostic data, such as microbiome profiles, in medical laboratories poses

Access this article online https://smerpub.com/

Received: 07 November 2025; **Accepted:** 19 February 2026

How to cite this article: Alvarez M, Herrera S, Cruz D, Castro A. Large Language Model–Driven Automation of Microbiome Diagnostic Reporting in Clinical Laboratories. J Med Sci Interdiscip Res. 2026;6(1):71-92. https://doi.org/10.51847/TTj3AtzOb6

significant challenges because of data complexity, the need for consistent, reproducible, and efficient workflows, and stringent regulatory obligations [2]. The GenDAI project [3] aims to resolve these difficulties by creating an integrated platform that unites advanced bioinformatics tools with the operational standards of clinical diagnostics. The platform currently supports automated pipelines for diverse analyses, such as gene expression profiling [3] and mitochondrial DNA dysfunction assessment [4], including outcome visualization. However, it has yet to incorporate the automated production of comprehensive laboratory findings reports, which could substantially lessen the manual workload of laboratory personnel.

Microbiome-derived insights typically require expert interpretation to translate into clinically useful information. This is because complex interactions among microbial populations often extend beyond straightforward comparisons with reference ranges for individual taxa. Consequently, preparing findings reports—delivering coherent clinical interpretations alongside lifestyle guidance—is a vital stage in the diagnostic pathway. Such reports empower clinicians and patients to reach informed decisions. While partial automation is feasible, clinical pathologists must thoroughly review final reports to safeguard accuracy and contextual relevance, particularly when patient-specific factors or medical history require nuanced consideration.

Large Language Models (LLMs), trained on expansive textual repositories that include medical publications, have achieved notable success across numerous natural language processing applications, such as medical licensing examinations and clinical decision assistance [5, 6]. Nevertheless, their application to automatically generating laboratory findings reports remains unexplored. Integrating AI solutions into laboratory diagnostic processes introduces important ethical and regulatory considerations. Any such integration must uphold clinical standards and legal requirements, prioritizing data protection, technical reliability, system dependability, and sustained human oversight. Thus, automating the production of findings reports in laboratory diagnostics while meeting these constraints remains an open challenge.

This article therefore investigates whether and how LLMs can be effectively utilized to automate the generation of findings reports in microbiome diagnostics, taking full account of associated ethical and regulatory issues. The study adopts the research framework outlined by Nunamaker *et al.* [7], which organizes inquiry into the categories of "observation", "theory building", "systems development", and "experimentation". The paper is structured accordingly: Section 2 surveys the existing literature and practice in laboratory diagnostics, metagenomics, artificial intelligence applications, and relevant regulation, thereby addressing the observation objectives. Section 3 develops a conceptual model for LLM-supported automation of findings reports in microbiome analysis, fulfilling the theory-building aims. Section 4 presents the realization of this model in the "MicroFlow" prototype, corresponding to the systems development goals. Section 5 reports an initial user evaluation conducted with laboratory experts, focusing on usability and practicality through cognitive walkthrough methodology to satisfy the experimentation objectives. Section 6 concludes by discussing the results and outlining future research directions.

Contemporary developments in science and technology

Determining how Large Language Models can assist clinical diagnostics requires a thorough examination of the current scientific and technological context. Consistent with the central research question and associated objectives, this assessment commences with a survey of metagenomics data processing in laboratory diagnostics (Observation Goal 1). Next, it reviews available bioinformatics instruments and platforms used for metagenomics, along with their shortcomings relevant to the current inquiry (Observation Goal 2). It then evaluates modern artificial intelligence approaches, particularly LLMs, to delineate their potential and constraints (Observation Goal 3). The discussion concludes by addressing the regulatory and ethical dimensions, including data privacy, algorithmic bias, transparency, explainability, and accountability, to guarantee that the proposed research meets existing standards and ethical expectations (Observation Goal 4).

Metagenomic data processing in laboratory diagnostics

Metagenomics involves analyzing nucleotide sequences extracted from the collective genetic material of microbial communities in samples. This field demands intensive data handling and comprises several sequential stages of processing and interpretation. This subsection focuses on the data-focused elements of metagenomics in laboratory diagnostic settings.

To optimize costs, complete sequencing of microbial genetic material is seldom performed. Rather, selected genomic segments are targeted; these regions feature both evolutionarily conserved areas and highly variable zones shaped by mutations [8, 9]. A standard example is sequencing variable subregions of the 16S ribosomal RNA (rRNA) gene in bacteria and archaea. Matching these partial 16S sequences against comprehensive microbial reference databases allows for accurate classification of organisms present in the specimen. As a result, 16S rRNA sequencing serves as a routine method for identifying targeted microbes in clinical samples during diagnostic procedures.

Data emerging from sequencing equipment is generally recorded in FASTQ format. This format consists of text files that store a unique identifier, an optional label, the actual nucleotide sequence, and corresponding quality scores. Files typically use .fq or .fastq extensions and are frequently compressed as .gz archives.

The initial data frequently contains redundant or near-identical sequences. The Amplicon Sequence Variant (ASV) strategy resolves this by identifying exact sequences and their abundance counts while accounting for sequencing quality and probable errors. This technique successfully separates technical artifacts from true biological differences [10], producing a collection of precise sequences supported by strong statistical evidence. After ASV identification, sequences are taxonomically classified using alignment tools such as BLAST [11] or the q2-feature-classifier module [12] from the QIIME 2 suite, relying on curated reference databases. Selecting an appropriate database is vital, since it can influence the accuracy of taxonomic assignments and introduce bias. Widely adopted databases include SILVA, Greengenes, and RDP [13, 14]. Once taxonomy is assigned, the proportional representation of each microorganism is determined from the frequency of its corresponding sequences.

In laboratory diagnostics, metagenomic workflows are integrated into a comprehensive operational sequence spanning order placement and sample logging, sample preparation, measurement, data evaluation, and report generation. Multiple preparatory investigations [15-17] conducted at the University of Hagen, which included interviews with laboratory personnel, have helped delineate this process. Key outcomes from these studies appear in Krause *et al*. [2, 4]. Building on this foundation, **Figure 1** illustrates the primary stakeholders and core activities that characterize laboratory diagnostics. Physicians initiate test requests for their patients and obtain a detailed laboratory report with findings and interpretations once testing is complete. Lab Scientists perform assays according to defined protocols and help design new testing methods. Lab Scientists use measurement devices to conduct the analyses. Data Analysts interpret the results and draft initial patient findings reports. Clinical Pathologists evaluate these reports, revise them when appropriate, and provide final authorization. They also endorse newly created test protocols. Compliance Officers monitor conformity with applicable rules, manage quality systems, and carry out Post-Marketing Surveillance (PMS). A Laboratory Information Management System (LIMS) oversees specimen tracking, test management, result storage, and report handling.

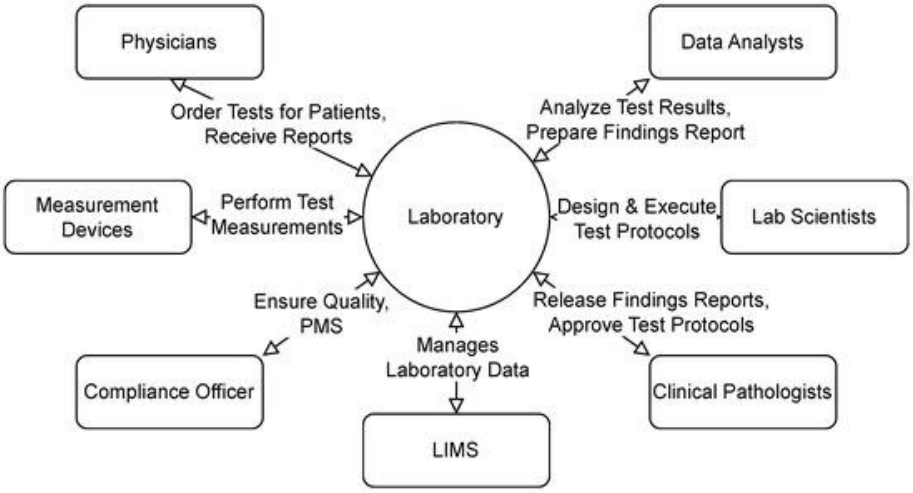


Figure 1. Schematic representation of the laboratory diagnostic setting.

Examining the workflow stages in greater detail, the order entry and sample registration phase entails inputting all necessary details about the specimen and requested examinations into the LIMS. The measurement phase covers physical testing of the sample, including preparatory steps that may involve various assays and instruments. The data analysis phase involves scrutinizing the generated data, performing additional calculations, and benchmarking results against reference standards. The clinical reporting phase focuses on assembling a comprehensive report for the requesting physician. This document presents the analytical outcomes together with any supplementary details relevant to patient care. Although laboratory team members commonly prepare these findings reports, a clinical pathologist must review, modify if required, and formally approve them. Automating the production of such findings reports within a fully integrated laboratory diagnostic process remains the principal unresolved issue explored in the following sections of this article.

Bioinformatics platforms supporting diagnostics

A range of bioinformatics systems and analytical tools have been developed to process microbiome datasets. One prominent example is QIIME 2, an open-source, modular pipeline that handles the full spectrum of microbiome data-processing tasks, from ASV detection and sequence alignment to producing graphical representations. Its component-based structure facilitates adaptation to diverse datasets and specific analytical needs. Complementary comprehensive frameworks include the ASaiM extension built on the Galaxy platform. Krause *et al*. [18] examine additional relevant resources. Collectively, these platforms deliver notable flexibility and accessibility for microbiome studies. However, they fall short in meeting the operational requirements of clinical laboratories. They are not engineered for high-throughput, automated, and repetitive analysis of large sample volumes, and they emphasize adaptability at the expense of the standardized, reproducible, dependable, and clinically validated workflows essential in regulated diagnostic environments. Furthermore, their scope is largely confined to the data analysis stage. It does not extend to the full laboratory diagnostic process, which encompasses test ordering, sample handling, and the delivery of clinical reports.

To address these gaps and harness the possibilities offered by artificial intelligence for Genomics-Based Diagnostic Data (GBDD), the GenDAI platform was originally conceptualized in Krause *et al*. [3] and later further developed in Krause *et al*. [13]. GenDAI merges the advanced analytical features of bioinformatics tools with the practical necessities of laboratory diagnostics. Built using container technologies and microservice architecture, the platform achieves substantial scalability. Its emphasis on laboratory workflows and AI integration positions GenDAI as a strong foundation for addressing both microbiome data challenges and the automated production of findings reports. The framework and prototype detailed in this article build on and expand the GenDAI system. Because GenDAI does not yet support automated generation of diagnostic findings reports, the following sections focus on the design, realization, and assessment of this capability.

Artificial intelligence approaches

The term artificial intelligence covers an extensive set of methodologies that empower machines to execute functions typically associated with human cognition. Large Language Models (LLMs) belong to a group of pretrained systems mainly applied to natural language tasks, particularly text creation. For text generation, the model receives an input expressed in tokens and successively predicts the most probable next token, repeating this process to construct extended passages. The starting input provided to the model is termed the “prompt”. Through carefully formulated prompts, users can supply context, directives, background details, illustrative examples, or other guidance to shape the generated content.

LLaMA constitutes a series of LLMs released by Meta AI, offered in model scales spanning from seven billion to 65 billion parameters. The more compact versions can run on ordinary consumer equipment, democratizing access to sophisticated language models, while the larger variants deliver enhanced capabilities. This was followed by Llama 2 and Llama 3, which extend to up to 70 billion parameters and show meaningful improvements over the original LLaMA family. All Llama models are freely downloadable and deployable, rendering them attractive options for academic research and prototype development.

Although LLMs demonstrate impressive abilities, they are constrained by several drawbacks, most notably the

occurrence of “hallucinations”. These refer to generated statements that are not factually supported. In medical contexts, such outputs carry the risk of producing inaccurate or even dangerous guidance; hallucinations pose a substantial challenge, and strategies exist to reduce their occurrence and severity. One effective approach is to embed domain-specific expertise in prompts to refine results [19, 20]. Another method, known as “prompt engineering” [20], involves deliberately designing and iteratively refining inputs to improve output quality. Despite these interventions, it is impossible to eliminate hallucinations. For this reason, expert human review and endorsement of AI-generated content are strongly recommended in high-stakes areas such as diagnostic assistance.

In laboratory diagnostics, the clinical interpretation of test results is an activity entrusted exclusively to suitably qualified specialists, namely board-certified clinical pathologists. This role necessitates comprehensive knowledge of the underlying data, its broader context, and sophisticated reasoning skills. Although current AI technologies cannot wholly supplant human expertise in this area, LLMs can serve as valuable aids to professionals. An informal literature search across academic databases employing terms such as “LLM metagenomics”, “LLM laboratory diagnostics”, and “LLM diagnostics reporting” located no prior studies applying LLMs to the interpretation of laboratory findings, qPCR data, or metagenomic results. However, promising applications of LLMs have been reported in related fields, including differential diagnosis [21] and therapeutic guidance [22, 23]. Therefore, leveraging LLMs specifically for result interpretation and drafting findings reports is a significant challenge that this paper addresses in the sections that follow.

Regulatory environment and responsible artificial intelligence

The expanding integration of artificial intelligence across multiple domains has raised substantial concerns regarding its ethical, legal, and societal consequences. Regulatory efforts seek to promote the creation and implementation of secure, morally sound, and dependable AI technologies. This imperative is especially critical in laboratory diagnostics and healthcare more broadly, where erroneous outcomes can have grave repercussions. Beyond mitigating potential hazards, regulation addresses broader matters, including data confidentiality, algorithmic bias, transparency, interpretability, and accountability [24, 25]. Oversight of laboratory diagnostics is not new. The In-Vitro Diagnostic Regulation (IVDR), a new European Union regulation, seeks to guarantee the safety and performance of in-vitro diagnostic products, including software used in diagnostic contexts. Among other stipulations, the IVDR mandates robust risk management practices, adoption of cutting-edge methods, and consideration of the full software lifecycle from initial design through deployment and ongoing maintenance.

Establishing robust AI regulation presents several difficulties. Primarily, artificial intelligence lacks a universally accepted definition and evolves with expert perspectives and technological progress [24, 25]. This ambiguity complicates efforts to identify systems that fall under regulatory scope [24, 25]. Rules that single out AI while overlooking comparable risks in other technologies risk creating imbalances that distort market dynamics [25]. Further obstacles stem from the swift evolution of AI capabilities and the intricate relationships among governments, industry participants, and society. Despite these hurdles, numerous initiatives have produced ethical guidelines, technical standards, and legislative structures to foster responsible AI advancement and application.

The High-Level Expert Group on Artificial Intelligence (HLEG AI) has released ethical guidelines for achieving trustworthy AI [26]. This document outlines seven essential principles for trustworthy AI: (i) Human Agency & Oversight, (ii) Technical Robustness & Safety, (iii) Privacy & Data Governance, (iv) Transparency, (v) Diversity, Non-discrimination & Fairness, (vi) Societal & Environmental Well-being, and (vii) Accountability. Supporting materials include a self-assessment questionnaire [27] for AI system development and deployment, as well as domain-specific policy advice, notably for healthcare [28]. The healthcare-focused report acknowledges AI’s significant potential in medicine and advocates a balanced regulatory strategy that safeguards patients while facilitating beneficial innovation [28].

The AI Act represents European Union legislation designed to govern the creation and application of AI systems through a risk-based classification [29]. The framework concentrates chiefly on high-risk uses while also addressing prohibited practices and lower-risk applications [29]. Prohibited applications include those likely to cause physical or psychological damage, such as certain social scoring mechanisms [29]. High-risk categories cover critical areas like medical devices and

safety-related components. Limited- or low-risk applications include tools such as chatbots, content generators, and spam detection systems [29], which face minimal regulatory burden. The AI Act covers diverse AI techniques, including machine learning, knowledge-based systems, and statistical or Bayesian methods [29]. Its reach extends to providers supplying services to the EU market, regardless of location, thereby promoting cross-border compliance [29]. For high-risk AI systems, the legislation requires establishing a comprehensive Risk Management System (RMS) and a Quality Management System (QMS) [29]. Continuous Post-Marketing Surveillance (PMS) is also obligatory to monitor real-world performance and detect emerging concerns [29]. Non-compliance may incur substantial financial sanctions, underscoring the necessity of strict adherence [29].

Trustworthiness in AI remains a comparatively underexplored dimension. A survey by KPMG revealed limited public confidence in AI technologies [30], noting that "trust strongly influences AI acceptance, and hence is critical to the sustained societal adoption and support of AI" [30]. In response, Bornschlegl [31] proposed the TAI-BDM Reference Model to enhance trustworthiness in AI-driven big data solutions. The model identifies Reproducibility, Validity, and Capability as central pillars of trustworthiness in human-AI collaboration. Although overlapping with elements of earlier frameworks, the TAI-BDM Reference Model stresses that trust develops gradually and therefore highlights processes for cultivating confidence in AI applications [31].

Although the AI Act carries legal force across the EU, it offers limited practical direction on constructing trustworthy systems. In contrast, the HLEG AI guidelines, along with their self-assessment tools and healthcare-specific recommendations, provide a robust foundation for designing and evaluating platforms such as GenDAI. These principles will therefore receive more detailed consideration in the following discussion.

The initial HLEG AI requirement concerns human agency and oversight. AI solutions should be developed to enhance rather than supplant human autonomy and judgment. They must uphold fundamental rights, avoid manipulation or deception, and empower users to reach well-informed choices. Finally, provisions for appropriate human supervision must be incorporated.

Technical robustness and safety require AI systems to remain dependable and resilient under varying conditions, including potential adversarial interference. Appropriate fallback mechanisms and protective measures should ensure operation within intended boundaries and allow for safe deactivation when needed. Systems should achieve high predictive accuracy and reproducible behavior when presented with identical inputs. Reliability also requires consistent, correct functioning across diverse inputs and operational contexts. The IVDR [32] also stipulates comparable standards.

Privacy and data governance must be upheld throughout the entire AI system lifecycle, covering data acquisition, transmission, retention, and utilization. Strict controls are essential for sensitive information such as patient records. These obligations align with regulations including the EU General Data Protection Regulation (GDPR), which governs the handling of personal health data within the Union.

Transparency requires that AI decision-making processes be explainable, or at least comprehensible, to stakeholders. Where full explainability of complex algorithms such as machine learning models is unattainable, transparency can be advanced through meticulous data provenance tracking and clear stakeholder communication regarding system capabilities and limitations [33]. Linking decisions back to source data and algorithms, alongside open disclosure of risks, forms a vital part of risk management as required by the IVDR [32].

AI development should promote diversity, non-discrimination, and fairness. For example, longstanding sex and gender biases in medical research and practice [34] have resulted in the under-representation of women in clinical studies. This has produced medical literature and resources disproportionately oriented toward male physiology, contributing to diagnostic inaccuracies, postponed interventions, and inferior health outcomes for female patients.

Systems should also be crafted with societal and environmental well-being in mind. Their broader effects on ecosystems, employment, and communities warrant careful evaluation. Accountability entails establishing mechanisms to audit AI operations and systematically identify, evaluate, record, and mitigate risks. These expectations mirror the stipulations for diagnostic devices and associated software under the In Vitro Diagnostic Regulation [32] and related sector-specific rules.

Meeting these multifaceted requirements is another important challenge for the approach developed in the subsequent sections of this paper.

System design and conceptual framework

Informed by the foundational work on User-Centered System Design (UCSD) by Norman and Draper [35] and consistent with the TAI-BDM reference model, effective system development must strongly emphasize user requirements. UCSD positions the user and their specific tasks as the primary focus throughout the design process. Following the research strategy outlined earlier, several theory-building objectives can be established. These objectives organize the content of this section. The process begins with the identification and detailed description of relevant use cases within the GenDAI platform (Theory Building Goal 1). The existing GenDAI conceptual model is then expanded to incorporate AI-based interpretation of microbiome analysis outcomes (Theory Building Goal 2). This is followed by a comprehensive modeling of the AI integration component (Theory Building Goal 3). Finally, an overarching conceptual architecture is developed to guide subsequent system implementation (Theory Building Goal 4).

Use cases

The primary research question examined in this study focuses on microbiome testing and the clinical interpretation of results within diagnostic findings reports. Consequently, use cases concerning test protocol performance, result evaluation, and report generation are central. These are described in detail below.

Test protocol execution, depicted in **Figure 2**, is conducted by laboratory scientists following standardized protocols. The procedure begins with a physician's test request, which includes completing an order form and submitting patient specimens. Key activities include sample preparation, measurement, and outcome recording, all accompanied by appropriate quality assurance steps. Upon completion, the test produces results. In microbiome analysis, preparation involves extracting the 16S rRNA region and verifying DNA quality. Measurements utilize next-generation sequencing instruments to target relevant 16S rRNA segments. After an initial quality assessment of the sequencing data, microorganisms are taxonomically identified. These classifications then support further analysis.

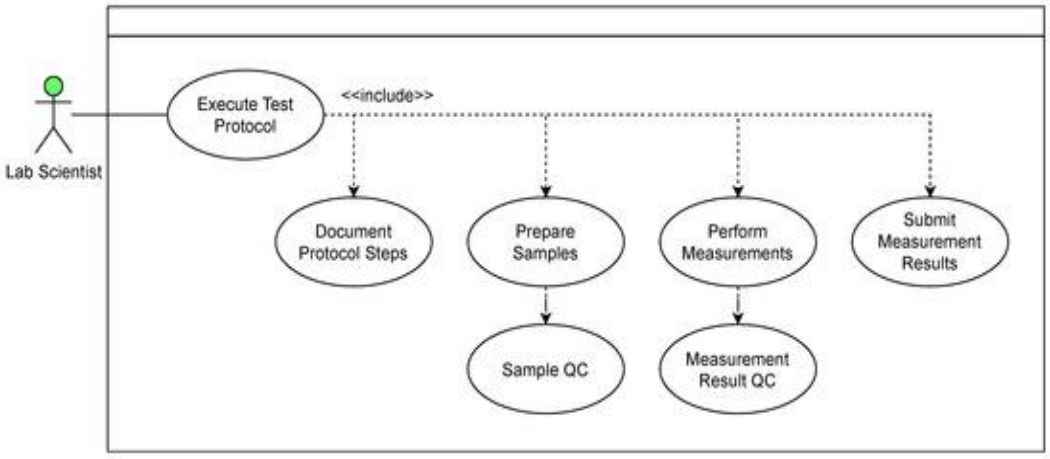


Figure 2. Use case diagram for test protocol execution.

Data analysts evaluate and preliminarily interpret test outcomes. The associated use cases appear in **Figure 3**. For microbiome investigations, this stage entails assessing key microbial indicators, often based on the detection or absence of particular organism groups. The data analyst compiles a laboratory report that presents the results, provides clinical context, and offers lifestyle guidance. This document requires careful review and formal approval by a clinical pathologist before transmission to the requesting physician.

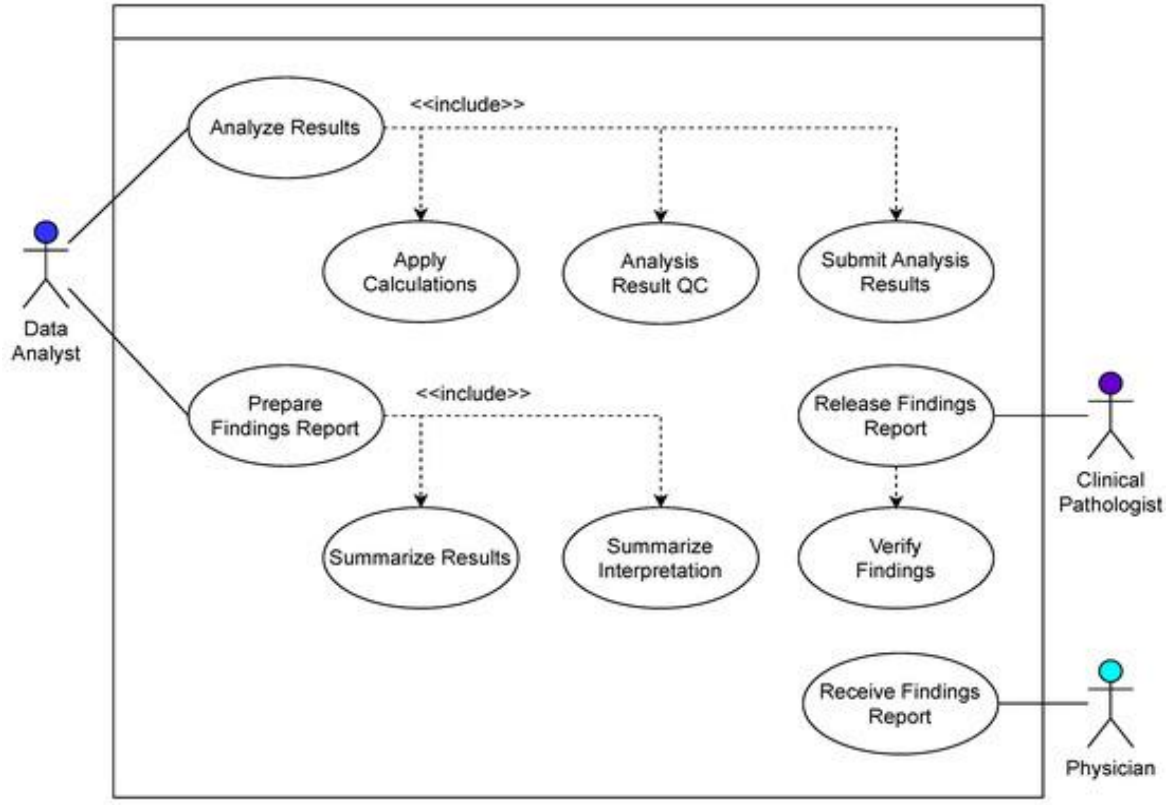


Figure 3. Use case diagram for analysis and reporting.

Figure 4 shows the overall diagnostic workflow for patient specimens in a UML activity diagram. This patient-focused sequence begins when samples arrive at the laboratory. Specimens, along with accompanying order and patient details, are entered into the Laboratory Information Management System (LIMS). The system then schedules and executes ordered tests. Once testing is complete, the system creates a findings report and delivers it to the ordering physician. If tests are unsuccessful, repeat testing with additional samples may be necessary.

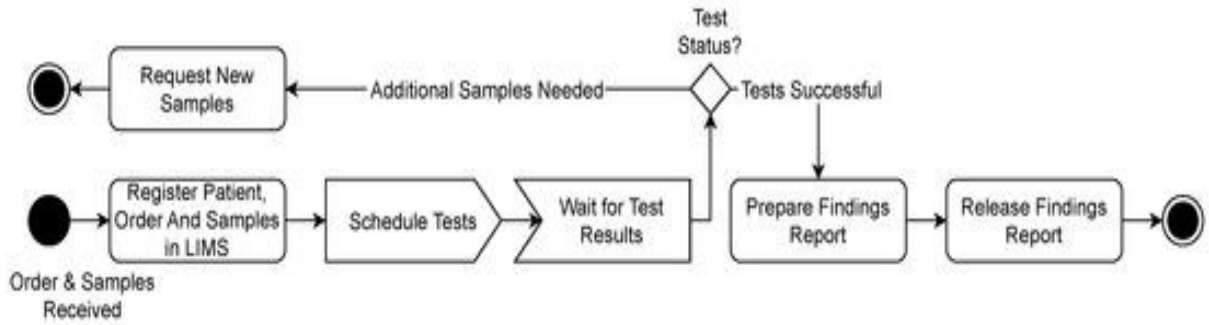


Figure 4. Patient samples general diagnostic workflow activity diagram.

Figure 5 presents the detailed workflow specific to microbiome analysis. After collecting all required samples for a sequencing run under the designated protocol, specimens are readied for analysis. Samples failing quality criteria are reprocessed using reserve material from the same patient. Should no reserve samples be available, new collections are requested. Sequencing takes place on an appropriate platform, yielding substantial data volumes that undergo bioinformatics processing to evaluate quality and eliminate substandard reads. When quality standards are met, sequences are compared to reference databases to

establish taxonomic profiles. Successful completion of quality controls allows storage of measurement results and subsequent determination of key microbiome parameters to generate an overall profile. These findings are assessed for consistency before final recording. Failure at any quality checkpoint triggers test repetition where possible or requests for fresh samples.

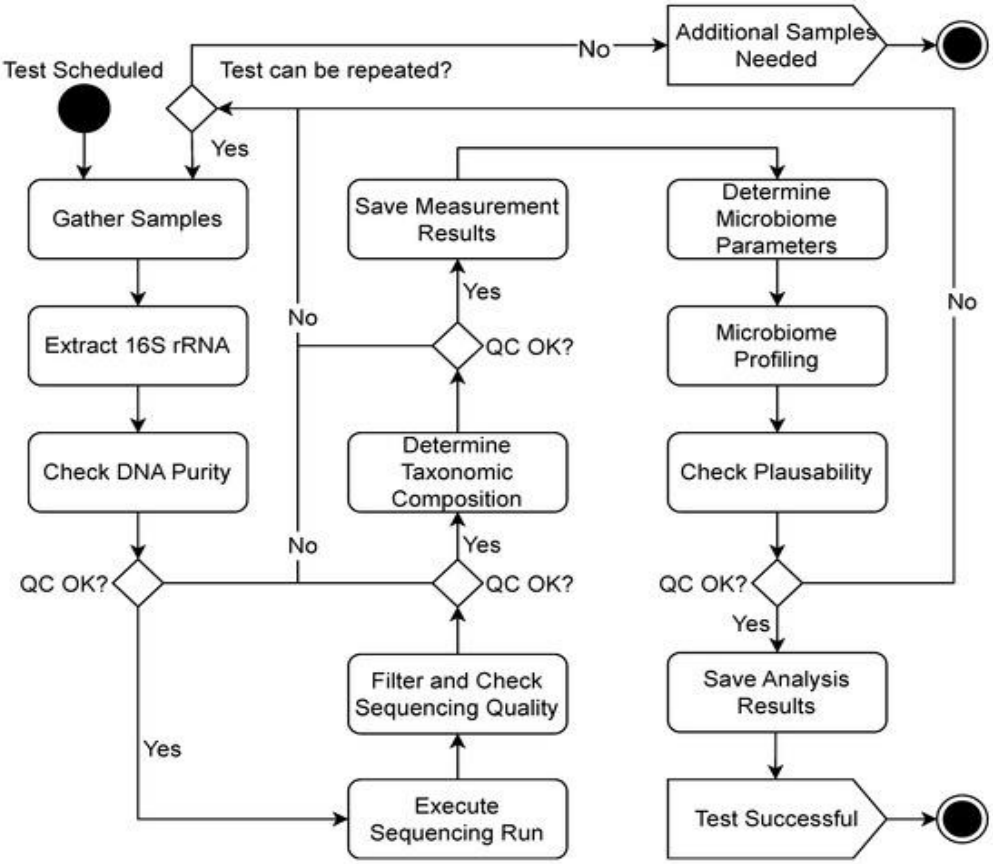


Figure 5. Microbiome analysis workflow activity diagram.

With the relevant use cases for microbiome analysis and interpretation of results in findings reports now identified and described (Theory Building Goal 1), attention turns to extending the current GenDAI conceptual model to integrate artificial intelligence for interpreting microbiome analysis results (Theory Building Goal 2).

Conceptual model

Once the relevant use cases have been established and described, User Centered System Design (UCSD) develops a conceptual reference model. This model articulates user objectives alongside the system's interaction mechanisms through a unified terminology. **Figure 6** presents a diagrammatic overview of the revised conceptual reference model. Because GenDAI must accommodate a diverse array of clinical diagnostic procedures, the model is intentionally formulated at an abstract level. While Krause *et al.* [3] presented a previous conceptual model for GenDAI, the version introduced here reflects considerable modifications to incorporate the contributions of the current work.

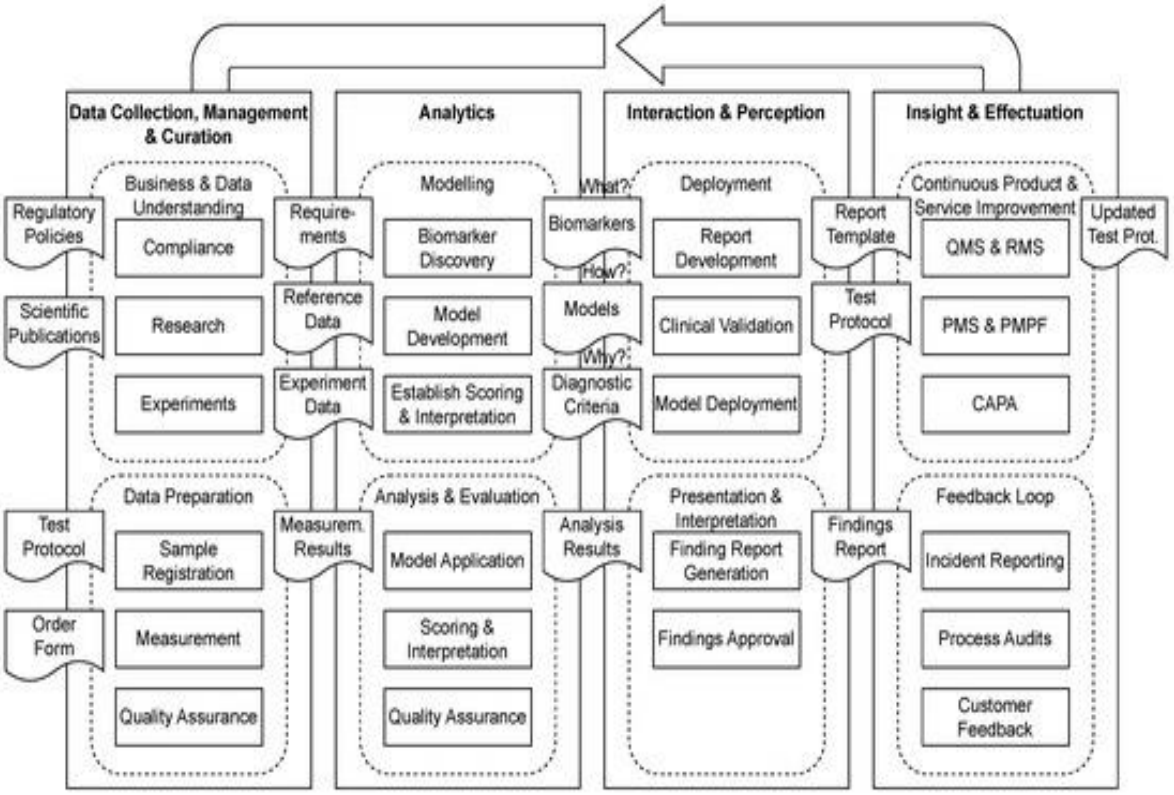


Figure 6. GenDAI conceptual reference model.

The diagram in **Figure 6** separates the test protocol development process (upper section) from the test execution process (lower section). Both processes progress through the same sequence of phases from left to right: Data Collection, Management & Curation; Analytics; Interaction & Perception; and Insight & Effectuation.

In the test protocol development pathway, the Data Collection, Management & Curation phase commences with Business & Data Understanding activities. These activities assemble the critical information needed to design or enhance diagnostic tests, drawing on multiple sources. Laboratory researchers systematically review scientific publications to uncover new insights and methodologies suitable for diagnostic translation. Such findings undergo experimental validation and adaptation for laboratory application. Regulatory compliance requirements play a central role in shaping business understanding and determining the quality controls and validation protocols necessary for new test development.

The Analytics phase in test protocol development focuses on detecting or validating potential biomarkers through detailed examination of experimental results and reference datasets compiled earlier. Researchers then integrate identified biomarkers with supplementary knowledge to build mathematical or statistical models. These models deliver a condensed yet comprehensive representation of relevant biological processes associated with particular diseases or physiological states. Scoring and interpretation mechanisms are subsequently created to convert model outputs into specific clinical conclusions. Development of these scoring systems typically involves establishing reference ranges derived from populations of both healthy and diseased individuals, as stipulated by regulatory standards. Collectively, biomarker discovery, model formulation, and the creation of scoring and interpretation tools constitute the Modeling task.

During the Interaction & Perception phase of test protocol development, attention shifts to incorporating the newly developed models and scoring systems into routine laboratory operations. This stage includes designing effective report templates and visual elements to convey analytical results clearly and accessibly. The Insight & Effectuation phase addresses sustained product and service optimization, encompassing risk management, quality oversight, and post-marketing surveillance. Observations and data generated in this

phase feed back into the initial stages to strengthen business understanding and support ongoing refinement of test protocols.

In the test execution pathway, the Data Collection, Management & Curation phase includes performing measurements, acquiring associated data, and completing essential preprocessing and quality assurance procedures. The Analytics phase carries out the primary evaluation of measurement outputs. This involves applying models and scoring systems established during protocol development, interpreting results, and conducting additional quality checks. In the Interaction & Perception phase, processed results are formatted into a formal findings report for the clinical pathologist's review and approval. Finally, the Insight & Effectuation phase releases the approved report to the ordering physician. Feedback received from users or other insights emerging from test execution are redirected into the continuous improvement cycle of the test protocol development process.

Through this conceptual model, Theory Building Goal 2 is achieved by augmenting the original GenDAI framework with provisions for artificial intelligence in interpreting microbiome analysis results. The following step involves a more detailed specification of the AI integration (Theory Building Goal 3).

AI integration

Generating findings reports requires both interpreting analytical outcomes and formulating appropriate clinical recommendations. These responsibilities ultimately rest with qualified medical professionals, specifically clinical pathologists. In practice, when certain interpretations and recommendations recur across similar cases, laboratories often employ reusable text modules that can be customized as needed. Laboratories can also delegate drafting initial report versions, provided a physician retains final review, modification, and approval. Consequently, producing draft reports lends itself well to automation. However, when fixed text modules are insufficient, effective automation requires substantial domain expertise, since accurate interpretation and recommendation generation require advanced medical knowledge.

To facilitate automated report generation, an LLM is employed. It receives inputs comprising patient information, analysis results, supplementary background or expert knowledge, and explicit instructions, and produces a draft findings report. Instructions allow control over the desired structure and style of the report. At the same time, background knowledge enables the inclusion of laboratory-specific expertise that may not have been part of the model's original training or that warrants particular emphasis.

Figure 7 illustrates the workflow for producing a draft report. Patient data from the Electronic Health Record (EHR) is first pseudonymized. Relevant details—such as age, gender, and additional clinical information—are then extracted and combined with test results in textual form. Domain-specific medical knowledge, report templates, and guiding instructions are similarly transformed into text. These elements are merged to form the complete input for the LLM. The model is typically accessed through its dedicated application programming interface, represented in **Figure 7** by request and response exchanges. The LLM's output is assessed and delivered as a draft report, which subsequently undergoes thorough examination and approval by the clinical pathologist.

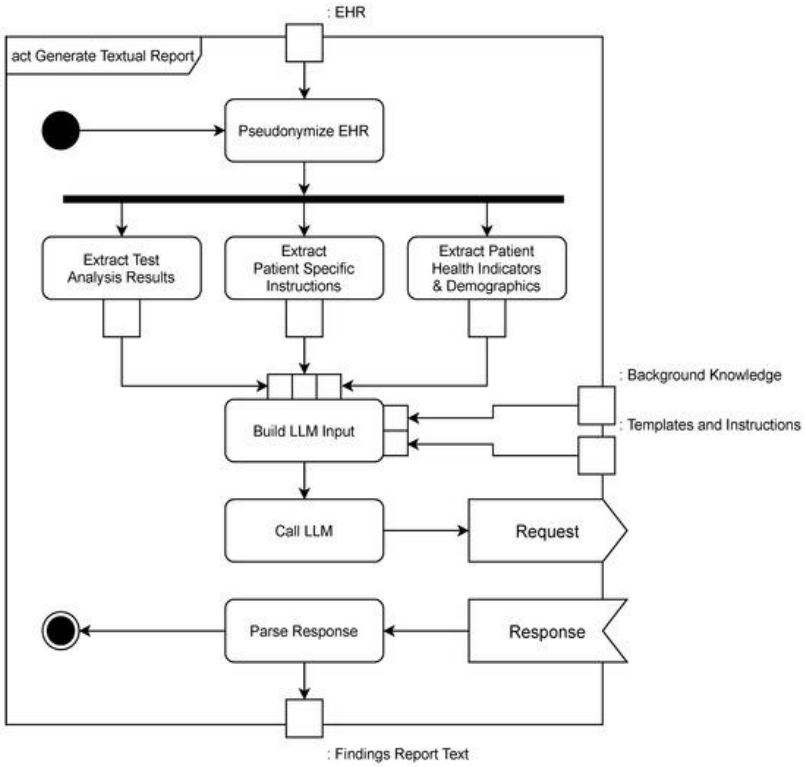


Figure 7. LLM for clinical interpretation and recommendations activity diagram.

As noted earlier, incorporating AI into clinical diagnostic workflows demands careful attention to legal and ethical considerations. Although a formal legal analysis lies outside the scope of this study, the HLEG AI guidelines offer a valuable foundation for GenDAI's design. Accordingly, the platform is developed to uphold user privacy, ensure technical robustness and safety, promote transparency, support diversity, non-discrimination and fairness, facilitate accountability, maintain human agency and oversight, and contribute positively to society and the environment. The AI integration within GenDAI is engineered to satisfy these principles. For this paper, providing more efficient and enhanced diagnostic services to patients is regarded as beneficial to society and the environment. Subsequent sections address how GenDAI can fulfill the remaining guidelines.

Protecting patient privacy and ensuring data security are paramount when deploying LLMs for clinical interpretation and recommendation. This is especially critical when models operate outside the laboratory's local infrastructure. Interpretation of results often relies on sensitive patient attributes such as age, gender, or medical history, which could potentially identify individuals and therefore require stringent safeguards. To align with relevant guidelines and the principle of data minimization, GenDAI prioritizes anonymized or pseudonymized data. The platform also accommodates both local and cloud-based LLM deployments, with local implementations serving as the default for operational environments. External providers are utilized solely for testing and evaluation activities.

Technical robustness and safety in LLM applications involve multiple considerations. Although adversarial attacks are not a primary risk in the intended use cases, protective mechanisms and fallback procedures remain essential to prevent erroneous or misleading outputs. In the GenDAI context, every LLM-generated report—including clinical interpretations and recommendations—undergoes mandatory human review by a clinical pathologist, who may revise or entirely rewrite the document as required. Should the system become unavailable, laboratory personnel can prepare reports manually, subject to standard pathologist approval, consistent with existing workflows. Temporary outages or unsatisfactory outputs can be addressed by reinitiating the report generation process. As previously described, the review, editing, and approval stages are integral to current laboratory procedures and remain unchanged with the introduction of LLMs. Technical robustness further requires high standards of quality and

accuracy in LLM outputs. This is vital for GenDAI, as unreliable results would undermine the system's utility and limit achievable automation levels. To address this, GenDAI provides interfaces to multiple LLMs, encompassing both proprietary commercial models and open-source alternatives. Integration with the Hugging Face platform enables flexible configuration of any model available in its extensive model repository.

A notable advantage of Large Language Models is their ability to adapt to diverse contexts without explicit retraining. This characteristic strengthens reliability as defined by the HLEG AI framework. However, reproducibility remains a limitation. Although certain methods can compel LLMs to generate identical outputs for identical inputs, these techniques carry drawbacks and function reliably only for specific model versions operating in controlled environments.

Rather than insisting on verbatim replication of outputs, it is more practical to reinterpret the reproducibility criterion as requiring outputs that are internally consistent for the same input — that is, free from contradictions. This interpretation corresponds to the standards expected of human specialists. It aligns with the trustworthiness definition in the TAI-BDM reference model, which emphasizes the capacity to produce consistent results given identical inputs. Although this represents a less stringent standard, it remains challenging to satisfy with contemporary LLMs, which are susceptible to hallucinations that occasionally generate incorrect information. To minimize such occurrences, GenDAI employs meticulously crafted prompts that constrain outputs to contextually appropriate content. For instance, prompts may specify lists of plausible diagnoses and interventions or outline explicit criteria that must be satisfied for any diagnosis or recommendation to be considered valid. This domain expertise, incorporated via the prompt, can be readily configured and regularly refreshed according to emerging scientific evidence and input from clinical pathologists.

Explainability constitutes a crucial requirement for AI applications in medical diagnostics. As outlined in Section 2.4, this element features prominently in regulatory initiatives concerning AI. Explainability exists on a continuum rather than as a binary attribute, generally diminishing as model complexity increases. Similarly, explainability expectations are context-dependent, varying according to the intended application and the potential risks associated with errors. The regulatory approaches discussed earlier do not impose absolute mandates for explainability but instead regard it as a component of the broader principle of transparency, which is fundamental to trustworthy AI.

The LLMs selected for GenDAI fall into the category of black-box models, which offer limited explainability. Nevertheless, current and forthcoming regulations do not prohibit their use in high-risk settings, provided there is adequate transparency regarding their operation and limitations, coupled with appropriate risk management measures. To reduce potential risks in GenDAI, LLM-generated content is restricted to producing draft reports that invariably undergo verification and approval by a clinical pathologist. Furthermore, models are prompted to supply verifiable justifications for proposed diagnoses and recommendations. Because LLMs can produce convincing yet inaccurate information—including fabricated references—clinical pathologists must be educated about this phenomenon and provided appropriate training. The likelihood of hallucinations can be substantially lowered by embedding structured frameworks within prompts, including predefined sets of diagnoses, validation criteria, and suitable interventions.

The application of LLMs in clinical interpretation and recommendation is likely to reflect existing societal and scientific biases. For example, although sex represents a significant biological factor, women have historically been underrepresented in clinical research [34]. Diagnostic approaches derived from such biased data may favor male physiology, and LLMs trained on these sources tend to perpetuate the same imbalances. In this case, the model mirrors gaps present in the underlying scientific literature. Addressing such issues fundamentally requires improvements in the source knowledge base.

LLM training corpora frequently contain substantial volumes of non-scientific material from social media and the broader internet, including discriminatory language and stereotypes that can affect generated outputs. Considerable research has focused on detecting and reducing these biases, with partial success [35, 36].

GenDAI adopts several complementary strategies to mitigate bias-related risks. First, data minimization principles limit the information supplied to the LLM. While factors such as age and sex may be clinically relevant for microbiome interpretation, non-essential details like names or addresses—which could heighten bias risks—are excluded from prompts. Second, expert knowledge and carefully formulated instructions

embedded in the prompt guide the model's responses and help correct potential biases. Given the specialized context of microbiome diagnostics and the use of refined prompting techniques, the probability of the LLM producing overtly discriminatory content beyond existing scientific limitations is considered low. Nevertheless, the clinical pathologist meticulously reviews all LLM outputs to confirm the absence of biased or discriminatory language.

Accountability requires that all input data and resulting findings be auditable. GenDAI therefore stores input information and analytical outputs in a manner that supports subsequent review. Recognizing that many laboratories lack comprehensive digital archiving systems, the platform also generates printable reports containing all pertinent input data, intermediate results, and outcomes. In addition to auditability, effective risk management forms a core component of accountability. Risks must be systematically identified, evaluated, documented, and mitigated. Such a Risk Management System (RMS) is already mandated by standards including ISO 14971 [37] and ISO 22367 [38] for medical laboratories and thus requires no separate development for GenDAI.

Responsible AI deployment, particularly in sensitive domains such as medicine, necessitates human agency and oversight. This is already embedded in the requirement that clinical pathologists review and approve every report. The accountability measures described above address ensuring that reports contain sufficient information for informed decision-making. However, given the currently incomplete understanding of hallucination risks in diagnostic applications of LLMs, targeted education and training for clinical pathologists are essential to maintain effective human oversight and agency.

This discussion of AI integration within the GenDAI conceptual model completes Theory Building Goal 3. The next step is to present a conceptual architecture that outlines the system's primary components.

Conceptual architecture

Before proceeding with implementation, the conceptual model can be further developed by introducing a conceptual architecture. This architecture provides a high-level representation of the system's main components, logically grouped by function. The conceptual architecture for GenDAI is presented in **Figure 8**. Because GenDAI accommodates various use cases previously modeled and realized, the architecture integrates both established and newly developed components. The additional components focus on microbiome analysis and incorporating large language models (LLMs) to support the generation of preliminary findings reports.

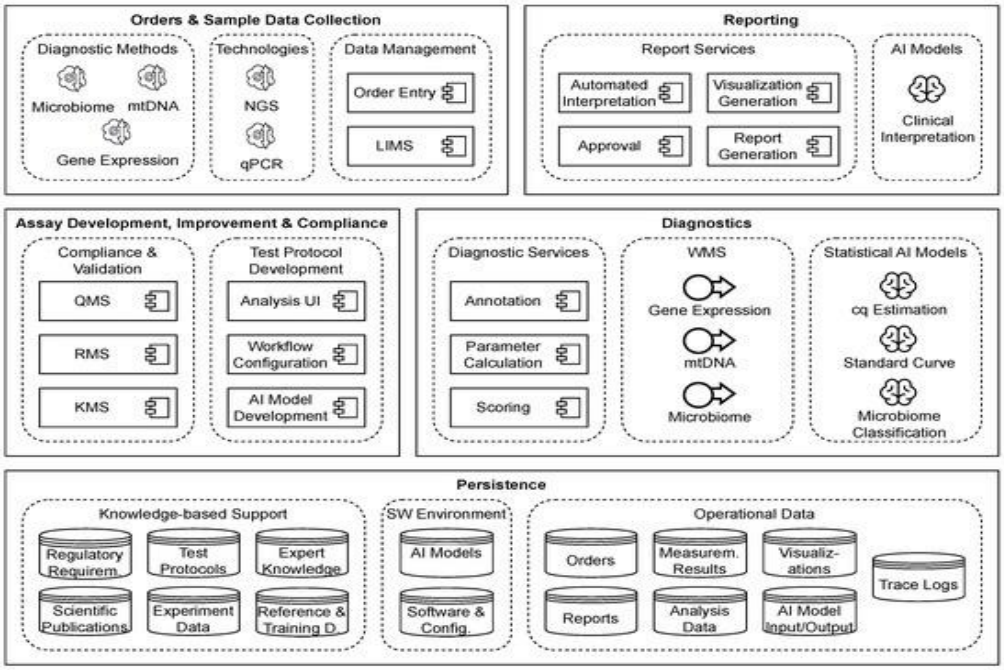


Figure 8. Conceptual architecture of GenDAI.

The overall structure is organized into five primary areas distributed across three distinct layers. The base layer is the Persistence area, which identifies the core data repositories the system needs. These repositories originate directly from the artifacts defined in the GenDAI conceptual model. The uppermost layer captures the main inputs and outputs associated with GenDAI and the laboratory environment as a whole. Positioned on the left is the Order & Sample Data Collection area, which includes the elements required for order processing, sample acquisition, and initial preparation. It also highlights the primary input artifacts linked to the measurement technologies and diagnostic approaches currently supported by GenDAI. On the right, the Reporting area encompasses the elements essential for producing findings reports, including the LLM for interpreting results, the review and approval mechanism, and other supporting elements needed to compile a complete findings report.

The intermediate layer is divided into two distinct areas. On the left, the Assay Development, Improvement, and Compliance area outlines the components involved in creating and refining assays while ensuring conformity with pertinent regulatory requirements. This includes training and fine-tuning AI models and deploying them for operational use. It further addresses the design and implementation of new diagnostic workflows or workflow configurations, together with associated reporting templates. Additionally, it incorporates vital mechanisms for regulatory compliance, including the RMS and QMS.

Within the same middle layer, the Diagnostics area specifies the components required for the routine processing and analysis of measurement data. Central to this is the Workflow Management System (WMS), which governs the execution of diagnostic workflows, along with all integrated tools and modules employed during these processes. These include bioinformatics applications, encapsulated computational procedures, and the utilization of AI models for GenDAI functions other than microbiome analysis.

In light of this conceptual architecture, which satisfies Theory Building Goal 4, the next stage entails implementing the conceptual model and architecture by developing a prototype.

Implementation

Building upon the foundations established earlier, the conceptual model and architecture introduce targeted extensions to the original GenDAI framework. These additions address microbiome analysis capabilities and incorporate large language models (LLMs) to facilitate the production of initial findings reports. The prior model was realized as a prototype under the designation “GenomicInsights” [13]. The microbiome-focused diagnostic workflow, together with its report-generation functionality, was developed as an extension to this existing prototype and named “MicroFlow”. This section first summarises the essential characteristics of the GenomicInsights prototype before exploring the MicroFlow extension, which forms the primary subject of the current work. The described implementation realizes the System Development Goal of the overall research strategy. Source code for both the GenomicInsights prototype and its MicroFlow extension is openly available via GitHub [39].

GenomicInsights operates as a web application designed to test protocols as automated workflows grounded in an event-driven architecture (EDA). This paradigm, which supports distributed computing, is well suited to environments managing extensive datasets that demand exceptional scalability and operational resilience. By using asynchronous events for inter-component interaction, EDA achieves significant decoupling, which improves fault tolerance and expandability. The manner in which EDA is applied within GenomicInsights is illustrated in **Figure 9**.

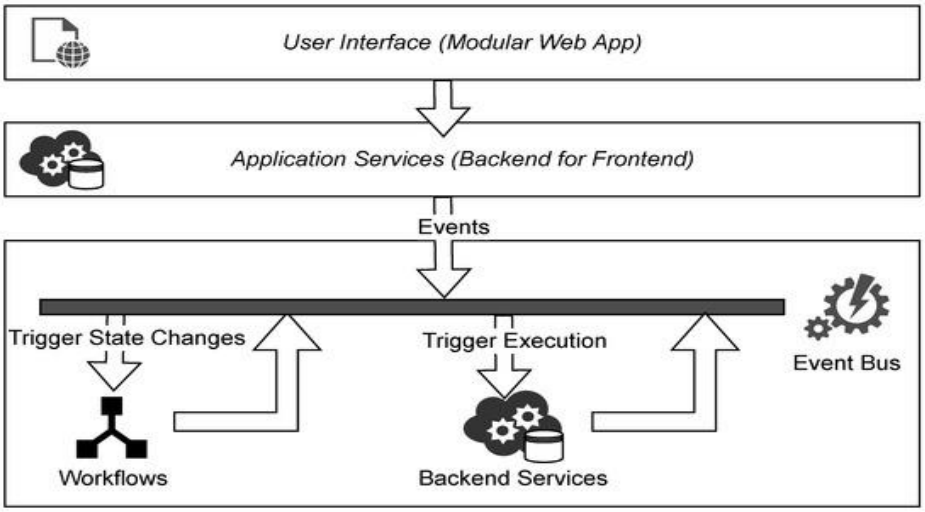


Figure 9. EDA and microservice-oriented technical architecture.

Central to event management in GenomicInsights is the Event Hub, which ingests and routes system-generated events. Various components subscribe to events relevant to their functions and can publish new events to the hub as circumstances arise, such as task completion or the emergence of updated data. To enhance decoupling and support dynamic workflow coordination, GenomicInsights integrates the event-driven model with a Workflow Management System (WMS) that constructs workflows capable of reacting to events and initiating further ones.

Application services function as a backend-for-frontend layer supporting the user interface [40], while dedicated backend services manage asynchronous, resource-intensive operations including analytical processing and report creation. The WMS relies on Apache Airflow (v2.9), a widely used open-source workflow orchestration platform. Airflow expresses workflows as directed acyclic graphs (DAGs) and executes them across distributed resources. Apache Kafka (v3.7) underpins the event hub, ensuring dependable messaging among all system elements.

The MicroFlow extension utilizes this infrastructure to implement a specialized microbiome analysis workflow. Workflow initiation prompts the Apache Airflow engine to dispatch an event via the Kafka hub announcing incoming data. The microbiome analysis microservice then processes this event using the QIIME 2 toolkit to evaluate 16S rRNA sequencing results, thereby producing a taxonomic microbiome profile in accordance with the modeling presented earlier in **Figure 5**. Upon completion, the microservice issues a follow-up event to the hub, advancing the workflow. The next stage triggers draft report generation through another event sent to the hub and received by the reporting microservice.

This reporting microservice engages the designated LLM, as schematized in **Figure 7**. The interaction incorporates a carefully structured prompt encompassing patient details, analytical outcomes, supplementary expert insights, and precise directives. These directives instruct the LLM to function as a clinical pathologist responsible for drafting a microbiome analysis report. The output must be a formal letter that interprets results and provides actionable recommendations. Listing 1 displays the prompt template utilized, translated into English, with placeholders retained for analysis results, patient data, and expert knowledge.

The LLM model selection is fully flexible. Development efforts employed the meta-llama/Llama-2-70b-chat-hf model sourced from the Hugging Face repository. An API connection to the cloud-based HuggingChat service [41] was implemented to streamline prototype testing and future assessments of alternative LLMs. After generating the report, the microservice publishes a completion event to the hub. The WMS receives this signal, finalizes the workflow, and enables users to retrieve the draft report (**Figure 10**).

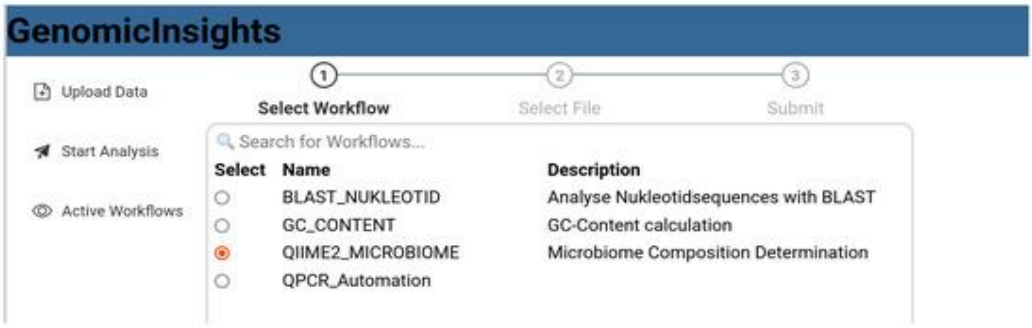


Figure 10. Start Screen of the MicroFlow Extension.

Listing 1. MicroFlow Prompt Template (Translated).

- Acting as a physician in a medical diagnostic laboratory, you are responsible for preparing a thorough diagnostic report for an individual patient. A stool sample was received from the patient, and 16S rRNA sequencing was performed to characterize the intestinal microbiome composition: {microbiome percentages}.
- Incorporate the microbiome composition into the letter as a table. Integrate the additional patient details provided below into the report: {patient information}.
- Present the patient details in a table positioned at the beginning of the document, as part of the letterhead.
- The following expert knowledge is also supplied for reference: {expert knowledge}.
- Refer to this expert knowledge exclusively when the mentioned bacteria appear within the provided microbiome composition data.
- Omission of a bacterium from the composition data does not indicate its total absence.
- Do not conclude that bacteria are absent from the data.
- Refrain from including the expert knowledge as a distinct section within the letter; it is furnished solely as supporting background for your analysis.
- Compose one paragraph summarising the analytical findings and comparing them to profiles observed in healthy donors. Include a separate paragraph containing practical recommendations for correcting any microbiome imbalances identified.
- Address the patient personally while maintaining a professional and formal tone throughout. Prepare the entire letter in German. Thank you.

Once the appropriate workflow and input data files have been selected, processing begins, and the user is taken to the workflow monitoring interface shown in **Figure 11**. This view provides continuous oversight of workflow execution status.

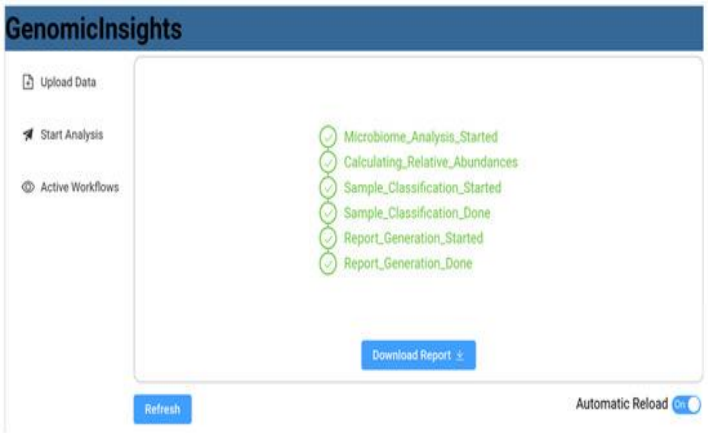


Figure 11. MicroFlow workflow progress interface.

Additional diagnostic and debugging capabilities are available through the Apache Airflow user interface, which displays workflow runs and associated details as shown in **Figure 12**. Detailed inspection of specific workflow instances is also possible within this interface.

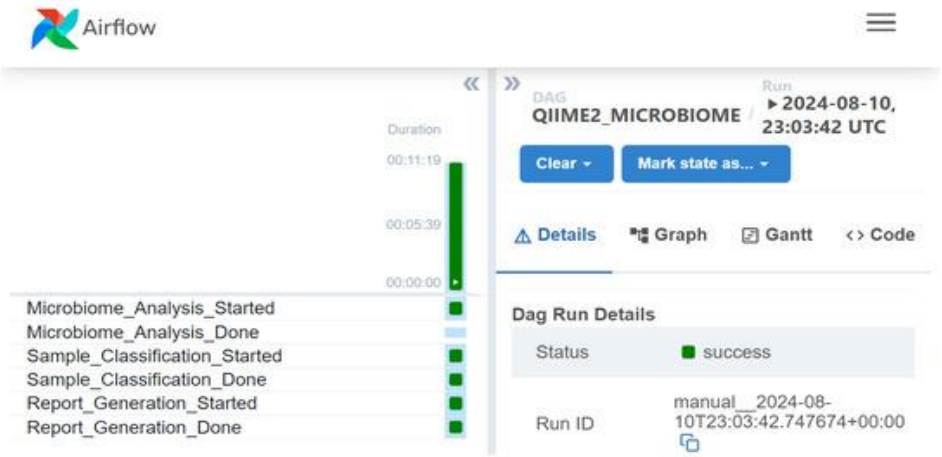


Figure 12. Apache Airflow user interface.

After a successful workflow run, the completed draft findings report can be downloaded directly from the MicroFlow prototype. The following section analyses a representative example of such a report, which presents the initial evaluation of the MicroFlow prototype.

Evaluation

A primary difficulty in assessing the usability and practicality of MicroFlow involved grounding the evaluation in the genuine requirements of end users. Consistent with the UCSD and TAI-BDM methodologies, engaging actual users in the assessment process is essential to enable iterative refinement driven by user feedback. A cognitive walkthrough [42] is an appropriate technique for this purpose, in which one or more evaluators systematically examine a sequence of tasks and pinpoint potential problems. The MicroFlow evaluation used a cognitive walkthrough conducted with two domain experts. This subsection outlines the outcomes of the qualitative assessment and presents implications along with directions for subsequent research. In line with the adopted research framework, this section fulfills the Experimentation Goal and completes the research objectives of the present study.

The two expert participants in the cognitive walkthrough were Prof. Michael Kramer (M.D.) and Patrick Newels (M.Sc.). Michael Kramer is a clinical pathologist, directs a clinical pathology laboratory, and leads a laboratory consulting organization, ImmBioMed Business Consultants GmbH & Co. KG, located in Pfungstadt, Germany, specializing in the translational advancement of laboratory diagnostic methods. Patrick Newels is head of the medical science department at biovis Diagnostik MVZ GmbH in Limburg-Eschhofen, Germany.

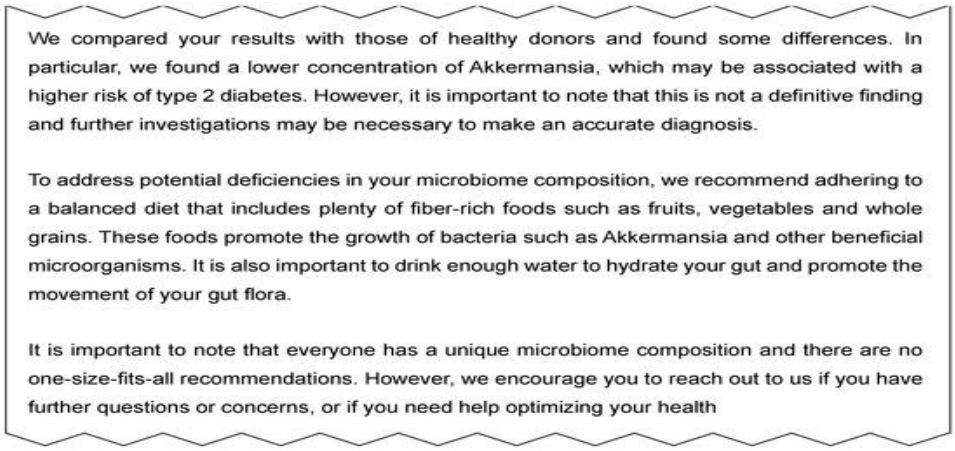
The objective of the cognitive walkthrough was to identify usability challenges in MicroFlow and assess the overall viability of the proposed LLM integration. The exercise adopted a formative perspective, emphasizing qualitative observations over numerical measurements. For the assessment, the biovis laboratory supplied sequencing data from a control sample containing known microbial concentrations for processing by MicroFlow. The data were provided in paired-end FASTQ format. Participants were introduced to the MicroFlow user interface and instructed on manually launching a microbiome analysis workflow using the sample data. They were also shown how to track workflow progress within the system. Finally, MicroFlow presented a completed draft report. Participants were then invited to offer their observations on both the prototype and the produced report.

The cognitive walkthrough elicited favorable responses regarding MicroFlow, reinforcing the potential value of incorporating LLMs into diagnostic workflows. Users described the prototype as straightforward and user-friendly. Among the limited recommendations were proposals to simplify adding workflows with varying configurations to accommodate diverse testing protocols. The LLM integration was positively received, as the model produced coherent and practically relevant drafts.

However, the walkthrough format precluded an in-depth examination of the LLM outputs to confirm the absence of hallucinations. Participants recommended that future LLM-generated content include citations to supporting literature, enabling clinical pathologists to validate the information more readily.

In response, subsequent efforts will prioritize refined prompt engineering techniques to guide the LLM toward the required output format. Although the current MicroFlow prototype relied on the Llama 2 model, future investigations should examine the recently introduced Llama 3 model alongside other LLMs. Moreover, LLM output evaluation should be expanded to encompass thorough scrutiny for hallucinations and additional inaccuracies. To investigate the model's capacity to leverage supplementary expert knowledge supplied by the laboratory, we tested draft report generation with an illustrative example of such information. Specifically, the LLM was informed that reduced levels of Akkermansia are linked to type 2 diabetes.

The generated draft reports appear in German, consistent with the laboratory's location and the participating experts. For inclusion in this paper, we translated the draft report into English using DeepL [43]. The document opens with a standard introduction covering the patient and test outcomes, presented as a table summarizing identified microorganisms and their relative abundances (**Figure 13**). It then proceeds to an interpretive analysis of the microbial composition. As intended, the report highlights the absence of Akkermansia in the sample and suggests this absence may be a risk factor for type 2 diabetes. The subsequent paragraph proposes a fiber-rich diet to promote beneficial gut bacteria, including Akkermansia. The report ends with a disclaimer noting the individuality of each person's gut microbiome and the consequent unsuitability of universal recommendations.



We compared your results with those of healthy donors and found some differences. In particular, we found a lower concentration of Akkermansia, which may be associated with a higher risk of type 2 diabetes. However, it is important to note that this is not a definitive finding and further investigations may be necessary to make an accurate diagnosis.

To address potential deficiencies in your microbiome composition, we recommend adhering to a balanced diet that includes plenty of fiber-rich foods such as fruits, vegetables and whole grains. These foods promote the growth of bacteria such as Akkermansia and other beneficial microorganisms. It is also important to drink enough water to hydrate your gut and promote the movement of your gut flora.

It is important to note that everyone has a unique microbiome composition and there are no one-size-fits-all recommendations. However, we encourage you to reach out to us if you have further questions or concerns, or if you need help optimizing your health

Figure 13. Excerpt from a MicroFlow draft report.

In summary, the findings indicate that embedding LLMs within diagnostic workflows is achievable and can help automate findings report generation while retaining essential human supervision. Incorporating expert knowledge directly into LLM prompts was shown to augment model capabilities without requiring specialized training. Nevertheless, more extensive evaluation remains required to establish the reliability and precision of LLMs when applied to microbiome diagnostics.

Conclusion

This study introduced a conceptual model and its practical implementation for embedding large language models (LLMs) within laboratory diagnostics, with particular emphasis on generating draft findings reports for microbiome analysis. Employing a systematic research strategy, the work defined and successfully pursued multiple research goals spanning observation, theory construction, system development, and experimentation. The conceptual model and architecture extend the pre-existing GenDAI framework, while the MicroFlow prototype implementation confirms the fundamental viability of the proposed LLM integration. Prototype evaluation demonstrated that LLM incorporation is feasible and holds promise for streamlining the creation of findings reports. A notable constraint of this study is the absence of comprehensive

testing of LLM outputs against a broad, authentic patient dataset, which is necessary to fully gauge accuracy, dependability, and the potential for hallucinations. Planned future research will involve assessing additional LLMs, conducting rigorous evaluations of AI model performance in real clinical environments, improving the verifiability of LLM outputs, and applying advanced prompt engineering to optimize results. As an immediate continuation, and supported by a recently granted EU MCSA research award, these enhancements, along with further developments, will be incorporated into the ongoing evolution of the GenDAI model and MicroFlow prototype. These will then be evaluated using real patient data to improve diagnostic processes and therapeutic management of Inflammatory Bowel Disease (IBD).

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Ogunrinola GA, Oyewale JO, Oshamika OO, Olasehinde GI. The Human Microbiome and Its Impacts on Health. Int J Microbiol. 2020;2020:8045646.
2. Krause T, Jolkver E, Mc Kevitt P, Kramer M, Hemmje M. A Systematic Approach to Diagnostic Laboratory Software Requirements Analysis. Bioengineering (Basel). 2022;9(4):144.
3. Krause T, Jolkver E, Bruchhaus S, Kramer M, Hemmje M. GenDAI—AI-Assisted Laboratory Diagnostics for Genomic Applications. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2021; Houston, TX, USA. IEEE; 2021.
4. Krause T, Glau L, Jolkver E, Leonardi-Essmann F, Mc Kevitt P, Kramer M, et al. Design and Development of a qPCR-based Mitochondrial Analysis Workflow for Medical Laboratories. Bio Med Informatics. 2022;2(4):643-53.
5. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv [Preprint]. 2023. arXiv:2303.13375.
6. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. medRxiv [Preprint]. 2023. 2023.02.21.23286254.
7. Nunamaker JF, Chen M, Purdin TD. Systems Development in Information Systems Research. J Manag Inf Syst. 1990;7(1):89-106.
8. Méndez-García C, Bargiela R, Martínez-Martínez M, Ferrer M. Metagenomic Protocols and Strategies. In: Nagarajan M, editor. Metagenomics. Cambridge (MA): Academic Press; 2018. p. 15-54.
9. Field KG, Olsen GJ, Lane DJ, Giovannoni SJ, Ghiselin MT, Raff EC, et al. Molecular phylogeny of the animal kingdom. Science. 1988;239(4841):748-53.
10. Chiarello M, McCauley M, Villéger S, Jackson CR. Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. PLoS One. 2022;17(2):e0264443.
11. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:421.
12. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90.
13. Krause T, Zickfeld M, Bruchhaus S, Reis T, Bornschlegl MX, Buono P, et al. An Event-Driven Architecture for Genomics-Based Diagnostic Data Processing. Appl Biosci. 2023;2(2):292-307.
14. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT – How do these taxonomies compare? BMC Genomics. 2017;18(Suppl 2):114.
15. Jolkver E. Verarbeitung von RT-qPCR Daten in der Labordiagnostik [Bachelor's Thesis]. Hagen (Germany): FernUniversität Hagen; 2022.
16. Glau L. Validation of qPCR Data in the Field of Medical Diagnostics [University Project]. Hagen (Germany): FernUniversität Hagen; 2022.
17. Glau L. Development of a System for Automated Microbiome Analysis and Subsequent LLM-Supported Report Generation in the Field of Medical

Diagnostics [Master's Thesis]. Hagen (Germany): FernUniversität Hagen; 2024.
18. Krause T, Andrade BGN, Afli H, Wang H, Zheng H, Hemmje M. Understanding the Role of (Advanced) Machine Learning in Metagenomic Workflows. In: Reis T, Bornschlegl MX, Angelini M, Hemmje M, editors. Advanced Visual Interfaces (AVI 2020); 2020; Ischia, Italy. Berlin/Heidelberg: Springer Nature; 2021. p. 56-82.
19. Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv [Preprint]. 2023. arXiv:2302.12813v3.
20. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [Preprint]. 2023. arXiv:2303.12712v5.
21. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards Accurate Differential Diagnosis with Large Language Models. arXiv [Preprint]. 2023. arXiv:2312.00164v1.
22. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep. 2023;13:20159.
23. Williams CY, Miao BY, Butte AJ. Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the Emergency Department. medRxiv [Preprint]. 2023. 2023.10.19.23297276.
24. Buiten MC. Towards Intelligent Regulation of Artificial Intelligence. Eur J Risk Regul. 2019;10(1):41-59.
25. Smuha NA. From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. Law Innov Technol. 2021;13(1):57-84.
26. High-Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. Luxembourg: Publications Office of the European Union; 2019.
27. High-Level Expert Group on AI. Assessment List for Trustworthy Artificial Intelligence (ALTAI). Luxembourg: Publications Office of the European Union; 2020.
28. High-Level Expert Group on AI. Sectoral Considerations on Policy and Investment Recommendations for Trustworthy AI. Luxembourg: Publications Office of the European Union; 2020.
29. Edwards L. The EU AI Act: A Summary of Its Significance and Scope. London: Ada Lovelace Institute; 2022 [cited 2024 Aug 20]. Available from: https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/
30. Gillespie N, Lockey S, Curtis C, Pool J, Akbari A. Trust in Artificial Intelligence: A Global Study. Brisbane (Australia): The University of Queensland; Sydney (Australia): KPMG; 2023.
31. Bornschlegl MX. Towards Trustworthiness in AI-Based Big Data Analysis. Hagen (Germany): FernUniversität Hagen; 2024.
32. The European Parliament and the Council of the European Union. In Vitro Diagnostic Regulation (EU) 2017/746. Brussels: Official Journal of the European Union; 2017.
33. Krause T, Wassan JT, Mc Kevitt P, Wang H, Zheng H, Hemmje M. Analyzing Large Microbiome Datasets Using Machine Learning and Big Data. Bio Med Informatics. 2021;1(2):138-65.
34. Plevkova J, Brozmanova M, Harsanyiova J, Sterusky M, Honetschlager J, Buday T. Various aspects of sex and gender bias in biomedical research. Physiol Res. 2020;69(Suppl 3):S367-78.
35. Norman DA, Draper SW, editors. User Centered System Design. Mahwah (NJ): Erlbaum; 1986.
36. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. arXiv [Preprint]. 2023. arXiv:2303.08774.
37. ISO 14971:2019. Medical Devices — Application of Risk Management to Medical Devices. Geneva: International Organization for Standardization; 2019.
38. ISO 22367:2020. Medical Laboratories — Application of Risk Management to Medical Laboratories. Geneva: International Organization for Standardization; 2020.
39. Krause T, Zickfeld M, Müller K, Glau L. GenomicInsights [GitHub repository]. 2024 [cited 2024 Aug 20]. Available from: https://github.com/aKzenT/GenomicInsights
40. Krause T, Jolkver E, Kramer M, Mc Kevitt P, Hemmje M. A Scalable Architecture for Smart Genomic Data Analysis in Medical Laboratories. In: Blum L, editor. Applied Data Science. Berlin/Heidelberg: Springer; 2023.
41. Soulter. HuggingChat Python API [GitHub repository]. 2024 [cited 2024 Aug 20]. Available from: https://github.com/Soulter/hugging-chat-api

42. Mahatody T, Sagar M, Kolski C. State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions. Int J Hum-Comput Interact. 2010;26(8):741-85.
43. Kutylowski J. DeepL [Online]. 2024 [cited 2024 Aug 20]. Available from: https://www.deepl.com/