BioMedInformatics

Article

Using Large Language Models for Microbiome Findings Reports in Laboratory Diagnostics

Thomas Krause [1,*], Laura Glau [1], Patrick Newels [2], Thoralf Reis [1], Marco X. Bornschlegl [1], Michael Kramer [3] and Matthias L. Hemmje [1]

[1] Faculty of Mathematics and Computer Science, University of Hagen, 58097 Hagen, Germany; laura.glau@studium.fernuni-hagen.de (L.G.); thoralf.reis@fernuni-hagen.de (T.R); marco-xaver.bornschlegl@fernuni-hagen.de (M.X.B.); matthias.hemmje@fernuni-hagen.de (M.L.H.)
[2] biovis Diagnostik MVZ GmbH, 65552 Limburg, Germany; patrick.newels@biovis.de
[3] ImmBioMed Business Consultants GmbH & Co. KG, 64319 Pfungstadt, Germany; m.kramer@immbiomed.de
* Correspondence: thomas.krause@fernuni-hagen.de

Abstract: Background: Advancements in genomic technologies are rapidly evolving, with the potential to transform laboratory diagnostics by enabling high-throughput analysis of complex biological data, such as microbiome data. Large Language Models (LLMs) have shown significant promise in extracting actionable insights from vast datasets, but their application in generating microbiome findings reports with clinical interpretations and lifestyle recommendations has not been explored yet. Methods: This article introduces an innovative framework that utilizes LLMs to automate the generation of findings reports in the context of microbiome diagnostics. The proposed model integrates LLMs within an event-driven, workflow-based architecture, designed to enhance scalability and adaptability in clinical laboratory environments. Special focus is given to aligning the model with clinical standards and regulatory guidelines such as the In-Vitro Diagnostic Regulation (IVDR) and the guidelines published by the High-Level Expert Group on Artificial Intelligence (HLEG AI). The implementation of this model was demonstrated through a prototype called "MicroFlow". Results: The implementation of MicroFlow indicates the viability of automating findings report generation using LLMs. Initial evaluation by laboratory expert users indicated that the integration of LLMs is promising, with the generated reports being plausible and useful, although further testing on real-world data is necessary to assess the model's accuracy and reliability. Conclusions: This work presents a potential approach for using LLMs to support the generation of findings reports in microbiome diagnostics. While the initial results seem promising, further evaluation and refinement are needed to ensure the model's effectiveness and adherence to clinical standards. Future efforts will focus on improvements based on feedback from laboratory experts and comprehensive testing on real patient data.

Keywords: genomics; diagnostics; AI; LLM; big data; microbiome

check for updates

Citation: Krause, T.; Glau, L.; Newels, P.; Reis, T.; Bornschlegl, M.X.; Kramer, M.; Hemmje, M.L. Using Large Language Models for Microbiome Findings Reports in Laboratory Diagnostics. *BioMedInformatics* **2024**, *4*, 1979–2001. https://doi.org/10.3390/biomedinformatics4030108

Academic Editor: M. Michael Gromiha

Received: 14 June 2024
Revised: 28 August 2024
Accepted: 30 August 2024
Published: 5 September 2024

1. Introduction, Motivation, Problem Statement, Research Question, and Approach

Laboratory diagnostics are one of the foundations of modern healthcare, enabling the identification and management of diseases. With the advent of genomics, the scope of laboratory diagnostics has expanded drastically, shifting towards a data-intensive environment. One example is the analysis of human microbiomes. Microbiomes are complex communities of microorganisms residing in various environments including the human body. Microbiomes in or on the human body are important for health, but can also cause diseases when imbalanced [1]. The gut microbiome, is a promising target for diagnostic tests, with changes in its composition linked to cancer, inflammatory bowel disease, and cardiovascular diseases, among others [1]. The study of microbiomes using data obtained from genetic material is called metagenomics.

The analysis of genomics-based diagnostic data such as microbiome data in medical laboratories is a challenge due to the complexity of the data, the need for reliable, reproducible, and efficient processes, as well as legal requirements [2]. The GenDAI project [3] tries to address these challenges by providing a platform that combines the capabilities of bioinformatics platforms with the requirements of laboratory diagnostics. GenDAI offers automated workflows for various diagnostic tasks such as gene expression analysis [3] and mtDNA dysfunction analysis [4], including result visualization. However, it has not yet addressed the challenge of creating laboratory findings reports, which could further reduce the workload of laboratory staff.

The findings gained from microbiome analysis require expert interpretation to be actionable, as the interactions between different microorganism often necessitate a deeper understanding beyond simple ranges for individual groups of microorganism. The generation of findings reports, i.e., providing natural language clinical interpretations and lifestyle recommendations, is a critical step in the diagnostic process, as it provides clinicians and patients with the information they need to make informed decisions. While some aspects of this creation process can be automated, findings reports have to be carefully revised by clinical pathologist to ensure accuracy and clinical relevance. This is especially true for cases, where the patient's medical history or other factors need special consideration.

Large Language Models (LLMs), trained on vast corpora of texts, including medical literature, have demonstrated remarkable capabilities in numerous natural language processing tasks including medical exams and clinical decision support [5,6]. To our knowledge, however, they have not been applied to the generation of findings reports in the context of laboratory diagnostics. Applying AI methods in the context of laboratory diagnostics raises several ethical and regulatory challenges. Any integration of AI methods into the laboratory workflow must ensure compliance with clinical standards and regulatory requirements, emphasizing data privacy, technical robustness, reliability, and human oversight. Automating the generation of findings reports in laboratory diagnostics while considering these challenges is an unsolved problem.

This paper thus addresses the research question on if and how LLMs can be used to automate the generation of findings reports in the context of microbiome diagnostics while considering the ethical and regulatory challenges. To answer this question, the research framework by Nunamaker et al. [7] was applied. Using this approach, the research question can be broken down into research goals of type "observation", "theory building", "systems development", and "experimentation". The structure of this paper follows this approach as follows: Section 2 provides an overview of the state of the art in laboratory diagnostics, metagenomics, AI methods and regulation, thus addressing the observation goals. Section 3 proposes a model for automating the generation of findings reports in the context of microbiome diagnostics, which is aligned with the theory building goals. Section 4 describes the implementation of this model through a prototype called "MicroFlow", addressing the systems development goals. Then, Section 5 provides an initial evaluation of the prototype by laboratory expert users, focusing on usability and feasibility, using a cognitive walkthrough to address the experimentation goals. Finally, Section 6 briefly discusses the results and planned future work.

2. State of the Art in Science and Technology

In order to find out how LLMs can support clinical diagnostics, it is helpful to first consider the current state of the art in science and technology. In accordance to the research question and the stated goals, it seems reasonable to start such an observation with an overview of metagenomics data analysis in laboratory diagnostics (Observation Goal 1). Then, an overview of existing bioinformatics tools and platforms that support metagenomics analyses will be given, including a discussion of their limitations in relation to the research question (Observation Goal 2). This will be followed by an overview of current AI methods and specifically LLMs to understand the capabilities and limitations of these methods (Observation Goal 3). Finally, the legal and ethical aspects of these findings will be discussed, including considerations like data privacy, algorithmic bias, transparency, explainability, and accountability, to ensure the research complies with current regulations and ethical standards (Observation Goal 4).

2.1. Metagenomic Data Analysis in Laboratory Diagnostics

Metagenomics is the study of nucleotide sequences obtained from the genetic material of microbiome samples. As such, it is a data-intensive process that involves multiple processing and analysis steps. This subsection focuses on the data-centric aspects of metagenomics analysis in laboratory diagnostics.

For cost efficiency reasons, the genetic material of microbiome samples is often not sequenced entirely. Instead, a specific region of the genome is targeted that contains both highly conserved and hypervariable parts, which are affected by evolution and mutations [8,9], such as subregions within the 16S ribosomal RNA (rRNA) gene in bacteria and archaea. By aligning the partially sequenced 16S regions to microbial genomic databases, a precise identification of microorganisms in a sample can be achieved. 16S rRNA sequencing is thus commonly used for detecting predefined microbes in patient samples for diagnostic purposes.

The sequences obtained from the sequencer are commonly stored in FASTQ format. FASTQ is a text file comprising of sequences with a unique ID, an optional description, nucleotide sequence, and quality scores. These text files have a .fq or .fastq extension and are often compressed into .gz files.

The raw data contains many duplicate or nearly duplicate sequences. The Amplicon Sequence Variant (ASV) approach for clustering of 16S rRNA sequences identifies exact sequences and their read counts, taking into account sequencing quality and error probabilities. This method can differentiate between variation due to sequencing error and biological variation [10], resulting in a pool of exact sequences with high statistical confidence. Following the identification of ASVs, the sequences are assigned to known taxonomies using tools like BLAST [11] or the q2-feature-classifier plugin [12] of the QIIME 2 toolkit, which utilize reference databases. The selection of a suitable reference database is crucial, as it can create bias in the identification of taxa. Popular reference databases include SILVA, Greengenes, and RDP [13,14]. After taxonomy assignment, the relative abundance of each microorganism in the sample is calculated based on the number of copies of each sequence.

Within the context of laboratory diagnostics, the metagenomics analysis process is embedded in a larger workflow that includes order entry and sample registration, sample preparation, measurement, data analysis, and findings reporting. To map out this process, several prestudies [15–17] involving interviews with laboratory partners have been conducted at the University of Hagen, with the most important results published in Krause et al. [2,4]. Based on these results, Figure 1 shows a high-level overview of the most important actors and actions in a laboratory diagnostic context. *Physicians* order laboratory tests for their patients and receive a laboratory report with results and interpretation upon completion of the tests. *Lab Scientists* perform tests using a predetermined test protocol and assist in the development of new test protocols. *Measurement Devices* are used by Lab Scientists to perform testing. *Data Analysts* analyze test results and prepare patient findings reports. *Clinical Pathologists* review the patient findings reports, make changes as needed, and approve them. They also approve newly developed test protocols. *Compliance Officers* ensure compliance with relevant regulations, including quality assurance and perform Post-Marketing Surveillance (PMS). A *Laboratory Information Management System (LIMS)* manages specimens, tests, results, and reports.

Focussing on the individual steps of the laboratory workflow, the order entry and sample registration step includes entering all relevant information about the sample and the requested tests order into the LIMS. The measurement step involves the actual measurement of the sample including all necessary preparations. This can involve multiple tests and multiple measurement devices. The data analysis step consists of the analysis of the measurement data including any calculations of derived values and the comparison of the results with reference values. The clinical reporting step comprises of the creation of a report for the requesting physician. This report contains the results of the analysis and any additional information that might be relevant to the physician. The findings reports are often created by laboratory staff, but must be approved and adjusted if necessary by a clinical pathologist. Providing this findings report in the context of an automated laboratory diagnostic workflow is the focus and remaining challenge to be addressed in the remainder of this paper.

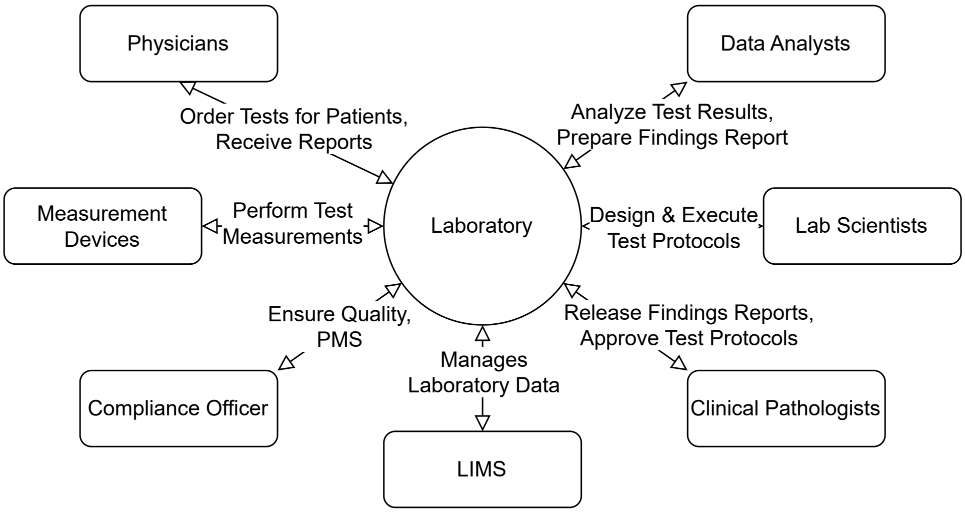

Physicians
Data Analysts
Order Tests for Patients, Receive Reports
Analyze Test Results, Prepare Findings Report
Measurement Devices
Perform Test Measurements
Laboratory
Design & Execute Test Protocols
Lab Scientists
Ensure Quality, PMS
Manages Laboratory Data
Release Findings Reports, Approve Test Protocols
Compliance Officer
Clinical Pathologists
LIMS

Figure 1. High-Level Overview of Laboratory Diagnostic Context.

2.2. Bioinformatics and Diagnostics Platforms

Multiple bioinformatics platforms and tools are available for analyzing microbiome data. QIIME 2 is an open-source, modular software pipeline capable of performing all steps of microbiome analysis, such as identifying ASVs, aligning sequences, or creating visualizations. Its modular architecture allows for adjustments to fit specific data and questions. Other comprehensive toolsets for microbiome analysis include the ASaiM extension of the Galaxy platform. Further tools are discussed in Krause et al. [18]. These bioinformatics platforms provide both flexibility and ease of use for microbiome data analysis. However, they are not optimized for laboratory use cases, as they do not provide features for repeated, automatic processing of many samples and prioritize flexibility over reproducible, fixed, reliable, and clinically validated laboratory workflows. In addition, these tools only cover the data analysis phase of the data and do not provide support for the complete laboratory workflow, including order entry, sample preparation, and clinical result reporting, among others.

To address these limitations and explore the potential of AI in the context of Genomics-Based Diagnostic Data (GBDD) processing, a new diagnostic platform called GenDAI was proposed in Krause et al. [3] and further refined and implemented in Krause et al. [13]. GenDAI is a platform combining the capabilities of bioinformatics platforms with the requirements of laboratory diagnostics. The platform is designed to be highly scalable using container technologies and microservices. With its focus on laboratory diagnostics and AI, GenDAI seems well suited to address the challenges of microbiome data analysis and the generation of findings reports. The model and implementation described in this paper are based on and extend the GenDAI platform. As GenDAI currently does not support the creation of diagnostic findings reports in an automated manner, the model, implementation, and evaluation for such a solution are remaining challenges to be addressed within the remainder of this paper.

2.3. AI Methods

AI is a broad term that encompasses various methods and techniques that enable machines to perform tasks that normally require human intelligence. Large Language Models (LLMs) are a class of pretrained models that are primarily used for natural language processing tasks such as text generation. For text generation tasks, an input text, measured in tokens, is used to predict the next probable token. This method is applied repeatedly to generate longer texts. The initial input text given to the LLM is called the "prompt". This prompt can be used to guide the model by providing context, instructions, background knowledge, examples, or other information that the model should consider when generating text.

LLaMA is a collection of LLMs developed by Meta AI. LLaMA is available in various sizes ranging from seven billion to 65 billion parameters. While the smallest-sized models can be run on consumer-grade hardware, allowing access to powerful LLMs to a broad audience, the larger sizes offer increased performance. LLaMA was followed by Llama 2 and Llama 3, also available in several sizes of up to 70 billion parameters, providing improved performance over the original LLaMA models. All Llama models are free to download and use, making them good candidates for research and development purposes.

While LLMs are powerful, they suffer from several limitations, including so-called "hallucinations". Hallucinations are outputs that are not grounded in facts. These are problematic in the medical domain where they could lead to potentially harmful advice. While hallucinations are a big problem, their probability and impact can be reduced. As an example, injecting domain specific knowledge into the prompts can be used to improve outputs [19,20]. Another technique to reduce hallucinations is the use of "prompt engineering" [20] where the LLM input is carefully designed and evaluated to improve model output performance.Even with these mitigations, hallucinations can not be avoided with complete certaincy. Human oversight and approval of model outputs is hence strongly advisable when used in applications such as medical diagnostics support.

In the context of laboratory findings reports, the interpretation of results is a domain reserved for trained and qualified humans (board-certified clinical pathologists) as it requires a deep understanding of the data and its context, as well as the ability to reason about it. While there are currently no AI methods that can completely replace humans in this task, an approach using LLMs could support humans. A cursory search on scientific publication search engines using the keywords "LLM metagenomics", "LLM laboratory diagnostics", and "LLM diagnostics reporting" did not reveal any publications that use LLMs for the purpose of diagnostic result interpretation in laboratory diagnostics or for the interpretation of qPCR or metagenomics results in general. There are however examples of LLM usage for differential diagnostics [21] or treatment recommendations [22,23], which are focused on other domains, providing promising results. Using LLMs for result interpretation in findings reports is thus another remaining challenge for the remainder of this paper.

2.4. Regulatory Landscape and Responsible AI

The increasing adoption of AI in various sectors has led to growing concerns about its ethical, legal, and social implications. Regulation of AI aims to ensure the development and deployment of safe, ethical, and trustworthy AI systems. This is particularly important in laboratory diagnostics and healthcare in general as the consequences of incorrect decisions can be severe. Regulation not only tries to minimize these risks but also targets broader issues such as data privacy, algorithmic bias, transparency, explainability, and accountability [24,25]. Regulation in the context of laboratory diagnostics is not new. The In-Vitro Diagnostic Regulation (IVDR) is a European Union regulation that aims to ensure the safety and reliability of in-vitro diagnostic devices which includes software used for diagnostic purposes. The IVDR requires, e.g., ensuring adequate risk management, using state-of-the-art technology, and taking into account the complete software lifecycle from conception to deployment and beyond.

Achieving effective regulation in AI is challenging due to multiple reasons. First, the definition of AI is not clear and can vary between experts and over time [24,25]. This lack of a clear definition makes it difficult to determine which systems are subject to regulation [24,25]. Regulation targeting AI but not other technologies with similar risks could favor or disfavor the use of AI which could lead to market distortions [25]. Additional challenges arise from the rapid pace of AI development and the complex interplay between stakeholders like government agencies, industry, and the public. Notwithstanding, several attempts have been made to create guidelines, standards, and legal frameworks for the responsible development and use of AI.

The HLEG AI has published a report on the ethical guidelines for trustworthy AI [26]. The report identifies seven key requirements 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. The report is accompanied by various documents, including a self-assessment list [27] for the development and deployment of AI systems and sector-specific policy recommendations, such as for the healthcare sector [28]. Notably, the sector-specific report recognizes the importance of AI in healthcare and proposes a nuanced approach to regulation and to promote and enable the use of AI [28].

The AI Act is a European Union legislation aiming to regulate the usage and development of AI by differentiating between various risk levels associated with AI applications [29]. The legislation primarily focuses on high-risk applications but also has provisions for unacceptable AI uses and limited-risk applications [29]. Unacceptable AI uses are defined as those that may cause physical or psychological harm such as social scoring systems [29]. High-risk applications entail critical domains such as medical devices and safety components, while limited and low-risk applications involve technologies like chatbots, content generation systems, or spam filters [29]. Low-risk applications are subject to the least stringent regulations. The AI Act encompasses a broad range of AI technologies, including ML, logic and knowledge-based approaches, and statistical or Bayesian methodologies [29]. It applies not only to users within the EU but also to providers offering services in the region, ensuring compliance across borders [29]. In order to ensure the safety and reliability of high-risk AI applications, the AI Act mandates that providers employ a comprehensive Risk Management System (RMS) and a Quality Management System (QMS) [29]. Furthermore, PMS is required for high-risk AI systems to continually evaluate their performance and identify potential issues in real-world scenarios [29]. Violations of the AI Act can result in significant financial penalties, emphasizing the importance of adherence to the stipulated regulations [29].

One aspect of AI that has so far been under-examined is trustworthiness. A KPMG survey found that citizens have low trust in AI systems [30] and that "*trust strongly influences AI acceptance, and hence is critical to the sustained societal adoption and support of AI*" [30]. Motivated by this finding, Bornschlegl [31] recently introduced the TAI-BDM Reference Model that aims to address trustworthiness in AI-supported big data applications. The model describes *Reproducibility*, *Validity*, and *Capability* as the three key dimensions of trustworthiness in human-AI interactions. While these are similar to some of the principles outlined in the beforementioned frameworks, the TAI-BDM Reference Model notes that trust is a process that is built over time and thus emphasizes the trust building process in AI applications [31].

While the AI act is legally binding within the EU, it does not provide direct guidelines on how to develop trustworthy AI systems. The guidelines provided by the HLEG AI together with their self-assessment checklist and the sector-specific guidelines for healthcare on the other hand provide a solid basis for the development and evaluation of models such as GenDAI. The guidelines and requirements of HLEG AI will thus be discussed in more detail in the following paragraphs.

The first criteria for trustworthy AI in the HLEG AI is respecting and enabling *human agency and oversight*. AI systems need to be designed to support human autonomy and decision-making instead of replacing it. They should respect fundamental human rights and should not deceive or manipulate humans, but help them to make informed decisions. Lastly, they should be designed for human oversight.

Technical robustness and safety refers to the requirement that AI systems should be dependable and resilient, even in the face of challenges or adversarial attacks. Fallbacks and safeguards must be implemented to ensure that the system runs within its designed parameters and can be shut down if necessary. AI systems should aim for high accuracy in their predictions. Results should be reproducible, meaning that the system demonstrates the same behavior when executed repeatedly with the same data. Lastly, AI systems should be reliable, meaning that they operate correctly with different inputs and situations. Similar requirements can also be found in the IVDR [32].

Privacy and data governance must be guaranteed during the complete lifecycle of AI systems. This includes the collection, transfer, storage, and processing of data. Access to sensitive data such as patient data must be strictly regulated. These requirements can also be found in specific regulations such as the EU General Data Protection Regulation (GDPR), which is mandatory for the use of patient data in the EU. *Transparency* refers to the requirement that AI systems should be explainable or at least understandable, making the reasoning behind decisions taken by automated algorithms, such as ML methods, more transparent [33]. In the absence of explainability, transparency can be achieved using data provenance and clear communication. Data provenance allows decisions made by an AI system to be linked back to the data and algorithms used, while communication involves informing users and stakeholders about the system's limitations. Understanding and sharing risks is a key aspect of risk management, which is also required by the IVDR [32].

AI systems should be designed to ensure *diversity, non-discrimination, and fairness*. As an example, sex and gender biases have long been a problem in medicine [34]. Historically, women have often been excluded from clinical trials, which means that text books and other medical resources are often biased towards male physiology and symptoms. This exclusion has led to a significant gap in medical knowledge and treatment effectiveness for women, resulting in misdiagnoses, delayed treatment, and overall poorer health outcomes.

AI systems should be designed for the *societal and environmental well-being*. The impact of AI systems on the environment, work force, and society as whole needs to be considered. *Accountability* requires mechanisms to be put in place that ensure that AI systems can be audited and that risks are managed by carefully identifying, assessing, documenting, and minimizing them. This requirement is similar to the requirements for diagnostic devices and software as stipulated in the In-Vitro Diagnostic Regulation [32] and other domain-specific regulations.

Meeting these requirements is another remaining challenge for the solution proposed in the remainder of this paper.

3. Design and Conceptual Modeling

Based on the research by Norman and Draper on User Centered System Design (UCSD) [35] and in line with the TAI-BDM reference model, it is important to consider the user's needs when designing a system. UCSD puts the user and the tasks that the user wants to perform at the center of the design process.

In accordance to our stated research approach, several theory building goals can thus be defined. The results of these goals are used to structure this section. First, the relevant use cases within the GenDAI system need to be identified and described (Theory Bulding Goal 1). Then, the existing GenDAI conceptual model needs to be extended to include the use of AI for the interpretation of microbiome analysis results (Theory Building Goal 2). As a next step, the AI integration needs to be modelled in detail (Theory Building Goal 3). Finally, a conceptual architecture can be created that will be used as the basis for the implementation of the system (Theory Building Goal 4).

3.1. Use Cases

The research question stated in this paper addresses microbiome analyses and the interpretation of their results in findings reports. As such, the use cases for the execution of test protocols, the analysis of test results, and the preparation of findings reports are of particular interest. These use cases will be discussed in more detail.

Test protocol execution as detailed in Figure 2 is performed according to the established test protocols by the lab scientist and is triggered by the order of a physician who fills out an order form and submits patient samples. Test protocol execution consists of preparing samples, performing measurements, and documenting results. The preparation and measurements are associated with quality control procedures. At the end of a test protocol execution, test results are obtained. For microbiome analysis, sample preparation consists of 16S rRNA region extraction and quality control, which includes testing for DNA purity. The measurement is performed on an NGS instrument that allows sequencing of relevant 16S rRNA subregions. Following quality control of the sequence data obtained, the taxonomy of the microorganisms contained in the sample is determined. The results of the taxonomy determination are then used for further analysis.

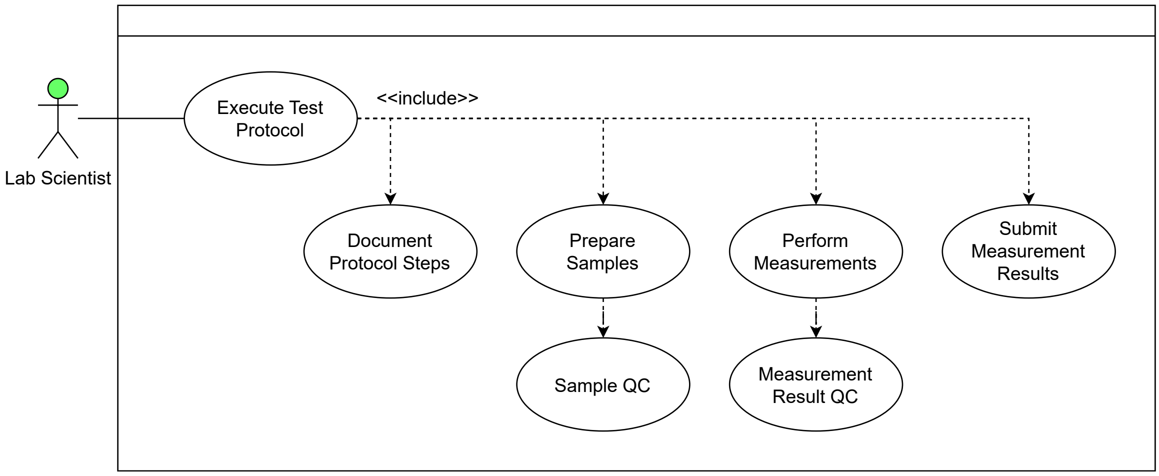


Figure 2. Use Case Diagram for Test Protocol Execution.

The analysis and preliminary interpretation of the test results is performed by the data analyst. The associated use cases are detailed in Figure 3. In the case of microbiome analysis, the analysis consists of determining relevant microbiome parameters, often indicated by the presence or absence of specific groups of microorganisms. The data analyst creates a lab report summarizing the results, their interpretation, and lifestyle recommendations. This lab report must be reviewed and approved by the clinical pathologist. The laboratory report is then sent to the physician who ordered the test.

Figure 4 shows the UML activity diagram of the general diagnostic workflow for patient samples. The sample- and patient-centric workflow starts with the arrival of one or more patient samples at the laboratory. The samples are registered in the LIMS together with the order and patient data. Tests are scheduled for the samples as ordered. After tests have been executed, a findings report is prepared and finally released to the ordering physician. If the test execution was not successful, additional samples might be requested to repeat the test execution.

Figure 5 shows the specific workflow for microbiome analysis. After gathering all samples for the sequencer run of the specific microbiome analysis test protocol, the samples are prepared for sequencing. If the quality for a sample is below expectation, the process is repeated using an additional sample from the same patient. If there are no additional samples, they can be requested to be collected again. The measurement is performed on a sequencing platform. This platform produces a large quantity of data which is then processed by bioinformatics tools to check the sequencing quality and filter bad-quality sequences. If the quality is sufficient, the sequences are mapped to a reference database to determine the taxonomic composition of the samples. Assuming the quality control steps are successful, the measurement results are saved and the analysis is performed by determining relevant microbiome parameters which can be used to create a microbiome profile. These analysis results are checked for plausibility and then saved to complete the test. If any of the quality control steps fails, the test is repeated if possible or additional samples are requested.

Having identified and described the relevant use cases for microbiome analysis and the interpretation of their results in findings reports (Theory Building Goal 1), the next step is to extend the existing GenDAI conceptual model to include the use of AI for the interpretation of microbiome analysis results (Theory Building Goal 2).

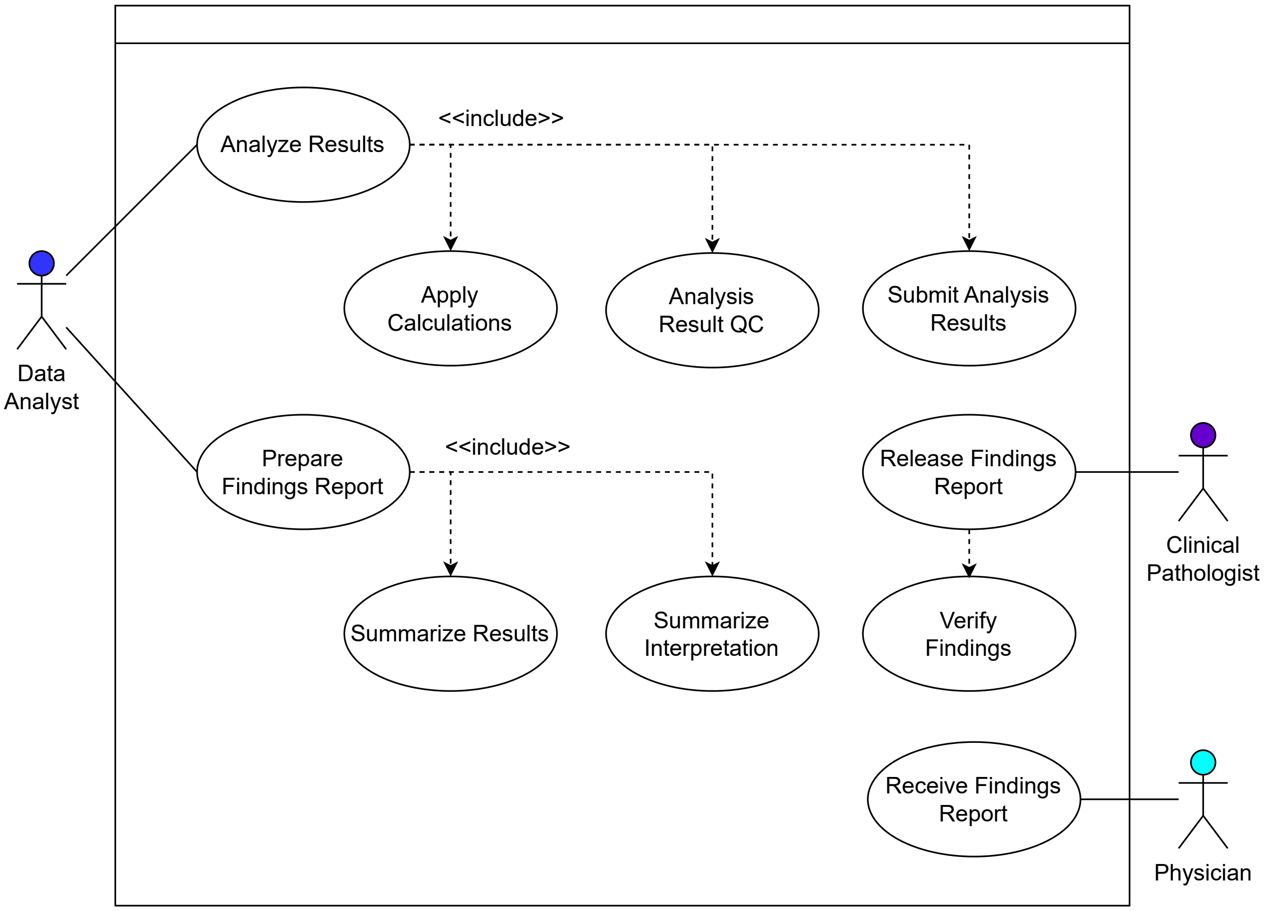


Figure 3. Use Case Diagram for Analysis and Reporting.

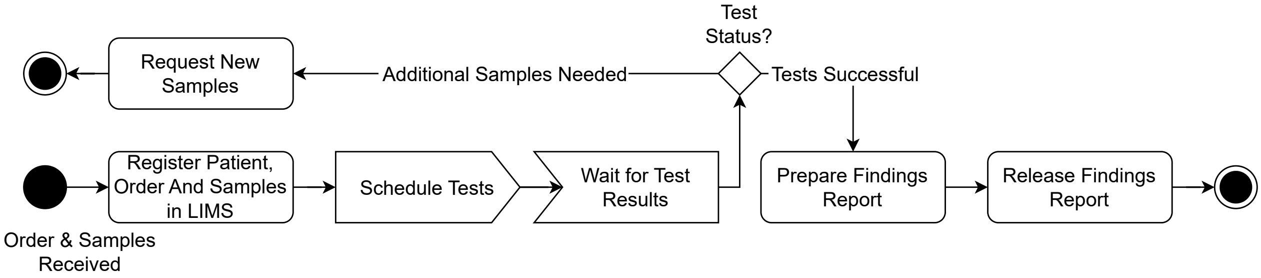


Figure 4. Patient Samples General Diagnostic Workflow Activity Diagram.

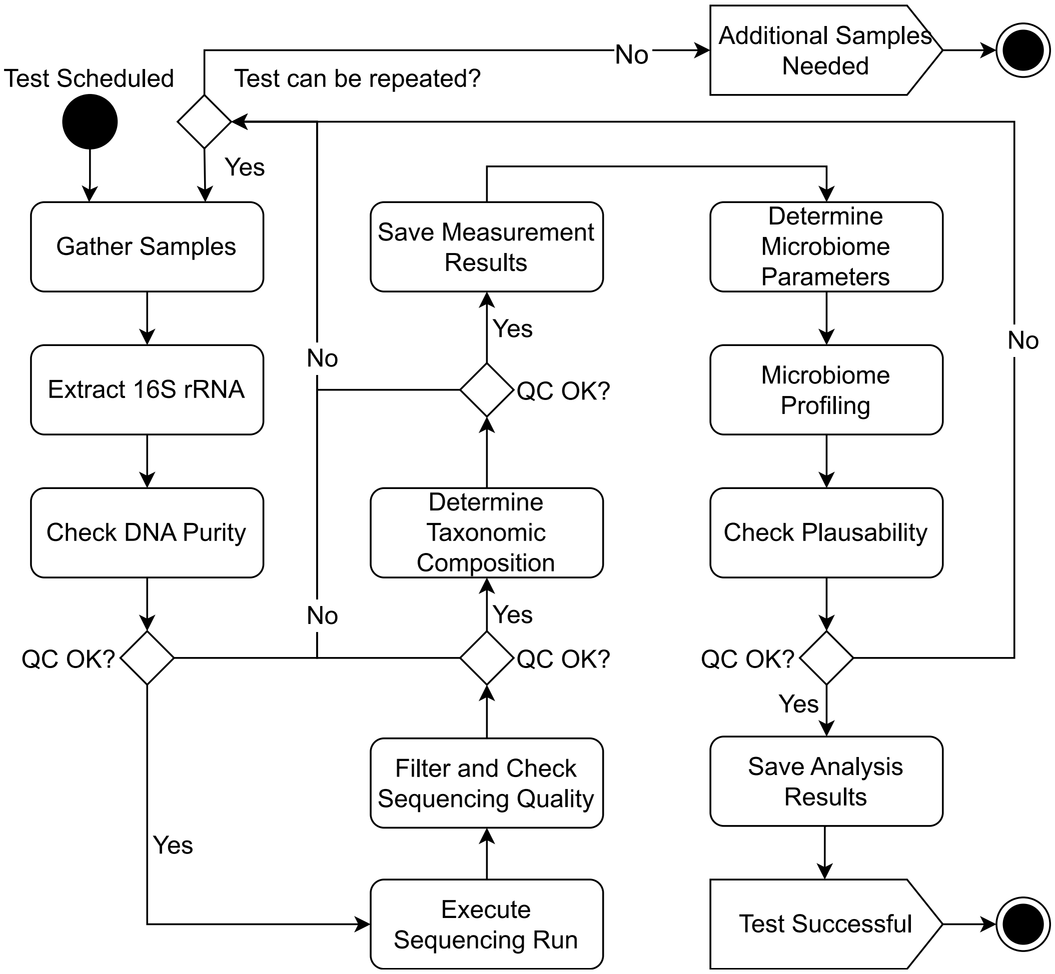


Figure 5. Microbiome Analysis Workflow Activity Diagram.

3.2. Conceptual Model

After introducing and describing the use cases, the next step in UCSD is to derive a conceptual reference model that describes the user's goals and the system's interface mechanisms in a common language. A graphical representation of the resulting conceptual reference model is shown in Figure 6. Due to the need for GenDAI to support a wide range of clinical diagnostic tests, the model is kept abstract. While a conceptual model for GenDAI was previously published in Krause et al. [3], the model presented here has been significantly updated to incorporate the developments described in this paper.

Figure 6 shows the test protocol development process on the top and the test execution process at the bottom. From left to right, the model uses the phases *Data Collection, Management & Curation*, *Analytics*, *Interaction & Perception*, and *Insight & Effectuation*.

The *Data Collection, Management & Curation* phase of the test protocol development process starts with *Business & Data Understanding* tasks, which are used to gather the necessary information to develop new or improved tests. The required information and knowledge is gathered from various sources. As part of the research process, the laboratoy screens scientific publications for new findings and methods that can be utilized in a diagnostic setting. These findings have to be confirmed and adapted to laboratory diagnostics by performing experiments. The regulatory requirements necessary for compliance form an important part of the business understanding. These regulatory requirements are used to define the necessary quality controls and validation steps for the development of test protocols.

Continuing with the test protocol development, the *Analytics* phase is used to identify or confirm possible biomarkers by analyzing the experiment and reference data gathered in the previous phase. These biomarkers can be combined and, together with additional knowledge, integrated into models. These mathematical or statistical models provide a compressed but comprehensive view of biological processes that are relevant to a particular disease or health condition. Finally, scoring and interpretation systems are developed that link the model results to concrete diagnostic interpretations. These scoring systems usually require the establishment of reference values based on a population of healthy and diseased individuals. Regulatory policies define requirements on how to establish these reference values. Biomarker discovery, model development, and the establishment of scoring systems and interpretations can be grouped in the *Modeling* task.

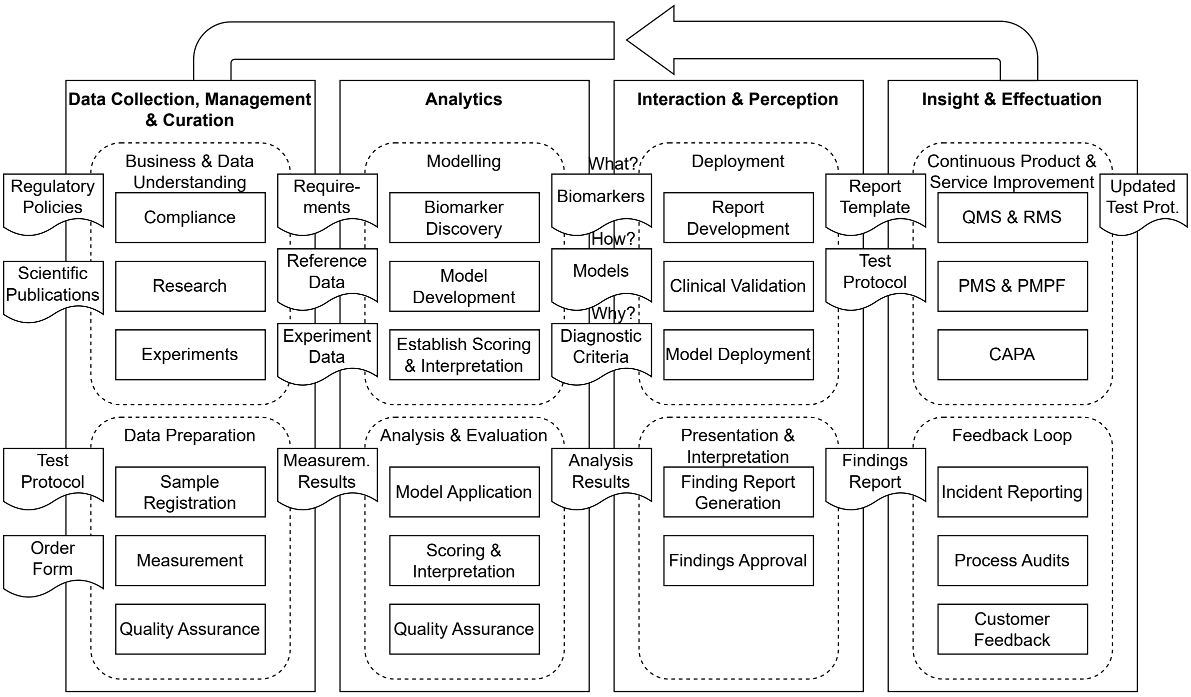


Figure 6. GenDAI Conceptual Reference Model.

The *Interaction & Perception* phase during test protocol development is concerned with the deployment of the developed models and scoring systems into the productive process. This involves the development of suitable report templates including visualizations that can be used to present the results of the analysis in an understandable way. Finally, the *Insight & Effectuation* phase of test protocol development contains tasks related to continuous product and service improvements such as risk and quality management, and post-marketing surveillance. New information gathered in this phase is fed back to the beginning of the process to raise business understanding and to develop new and improved test protocols.

Continuing with the test execution process, the *Data Collection, Management & Curation* phase is used to execute measurements and retrieve the associated data. The phase also includes any necessary data preprocessing and quality assurance. The *Analytics* phase is used to perform the actual analysis of the measurement data. This includes the application of models developed in the test protocol development process, the application of scoring systems, and the interpretation of the results. It also includes additional quality assurance steps. In the *Interaction & Perception* phase, the results of the analysis are rendered into a findings report which is approved by the clinical pathologist. Finally, in the *Insight & Effectuation* phase the findings reports are released to the ordering physician. Customer feedback or other insights from the test execution are fed back into the continuous improvement task of the test protocol development process.

This conceptual model addressed Theory Building Goal 2 by extending the existing GenDAI conceptual model to include the use of AI for the interpretation of microbiome analysis results. The next step is to model the AI integration in detail (Theory Building Goal 3).

3.3. AI Integration

In order to create findings reports, it is necessary to interpret the analysis results and derive recommendations for action. Ultimately, these tasks are the responsibility of the trained medical expert, the clinical pathologist. As certain interpretations and recommendations for action can be repeated depending on the test, text modules are sometimes used in practice, which can be individually adapted if necessary. In addition, the creation of draft findings can be delegated as long as the final review, editing, and approval is carried out by a doctor. In principle, the creation of draft findings is thus a task that is suitable for automation. However, if text modules are not to be used exclusively, the automation requires a high degree of domain competence, as the interpretation of the results and the derivation of recommendations for action require a high level of medical expertise.

An LLM will be used for the automated creation of findings, which receives patient data, the results of the analyses, as well as further background/expert knowledge and instructions as input and delivers a draft findings report as output. While the instructions can be used to adjust the expected format of the output, background knowledge can be used to add specific domain knowledge that the laboratory posesses and that was not part of the model training or that should receive special attention.

The process of creating a draft report is shown in Figure 7. The Electronic Health Record (EHR) containing patient data is first pseudonymized. The patient data relevant to the findings, such as age, gender and additional medical information, is then extracted and converted into a text format together with the analysis results of the tests performed. The additional medical background knowledge, as well as text templates and instructions, are also converted into text format. Together, these texts are concatenated as input for the LLM. The LLM is usually called via its own API, which is mapped in Figure 7 using request and response messages. The response of the LLM is evaluated and returned as a draft finding. This draft report must then be checked and approved by the clinical pathologist.

As previously discussed, the legal and ethical aspects for integrating AI into a clinical diagnostic workflow must be carefully considered. While a legal opinion is beyond the scope of this paper, the guidelines of the HLEG AI provide a good basis for the development of GenDAI. GenDAI should thus be designed to respect user privacy, be technically robust and safe, allow transparency, ensure diversity, non-discrimination and fairness, allow accountability, enable human agency and oversight, and be beneficial to society and the environment. The integration of AI into GenDAI must fulfil these guidelines. For the purpose of this paper, it will be assumed that providing improved and efficient diagnostic services to patients is beneficial to society and the environment. The following sections will discuss how GenDAI can be designed to meet the other guidelines.

Patient privacy and data protection must be considered when integrating LLMs for clinical interpretation and recommendation. This applies in particular if the models are not run locally within the laboratory. The assessment and interpretation of test results may depend on patient data such as age, gender or medical history, which may allow individual patients to be identified and are therefore particularly worthy of protection. In order to meet the requirements of these guidelines and the principle of data minimization, GenDAI uses anonymized or pseudonymized data as far as possible. GenDAI also supports both local and cloud-based implementations of LLMs for clinical interpretation and recommendation, whereby only local implementations are used by default in productive use. Thid party providers are available for testing and evaluation purposes.

Technical Robustness and Safetey in the context of LLMs entails various aspects. While adversarial attacks are not a concern for the planned use cases, including safeguards and fallbacks are nonetheless essential to ensure that the models do not produce wrong or misleading results. In the GenDAI use cases, all LLM outputs, namely the clinical reports including recommendations are subject to human review by the clinical pathologist, who can decide to adjust or rewrite the report if necessary, thereby providing a safeguard. If the system is unavailable, laboratory staff can prepare the findings report, which the clinical pathologist must approve, as done in the existing laboratory workflow. If the system is only temporarily unavailable or failed to produce sensical output, the report generation task can be restarted. As discussed in the previous section, the review, editing, and approval process is already part of the laboratory workflow and is not changed by the introduction of LLMs. Technical robustness also includes achieving high quality and accuracy of the results produced by the LLM. This is a key requirement for the planned GenDAI use cases, as inaccurate results would diminish the usefulness of the system and thus the level of automation. GenDAI thus implements interfaces to multiple LLMs, including both commercial and open-source solutions. The interface to hugging face allows the configuration of any model available in the hugging face model hub.

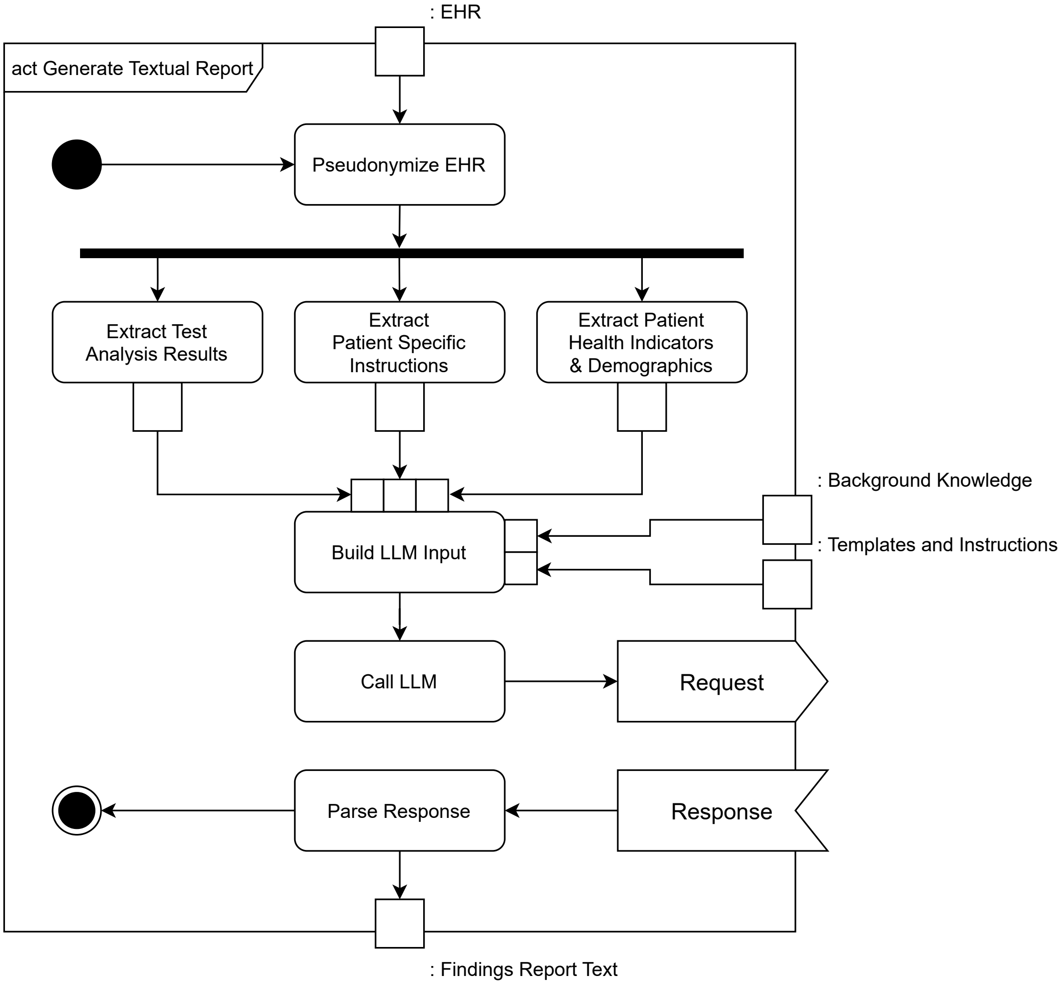

: EHR
act Generate Textual Report
Pseudonymize EHR
Extract Test Analysis Results
Extract Patient Specific Instructions
Extract Patient Health Indicators & Demographics
: Background Knowledge
: Templates and Instructions
Build LLM Input
Call LLM
Request
Response
Parse Response
: Findings Report Text

Figure 7. LLM for Clinical Interpretation and Recommendations Activity Diagram.

One of the strengths of LLMs is to adapt to a variety of contexts without needing explicit (re-)training. This improves the reliability as defined by HLEG AI. On the downside, the reproducibility achieved by LLM is limited. While there are some techniques to force LLM to produce the exact same output for the same input, they have disadvantages and only work for the exact version of the LLM running in a specific environment.

However, rather than comparing the exact output of LLMs when using the same input, it seems appropriate to reinterpret the reproducibility requirement to mean that the same input should result in outputs that are consistent with each other, i.e., they should not contradictory. This definition aligns with the expectations of human experts. It is also compatible with the definition of trustworthiness in the TAI-BDM reference model which defines reproducibility as the ability to obtain consistent results when the same input is provided to the system. While this is a weaker requirement, it is still hard to achieve with current LLMs which are prone to hallucinations, meaning that sometimes they produce false outputs. To reduce hallucinations, GenDAI uses carefully designed prompts to limit the possible outputs to those that are consistent with the context. For example, the prompt can include a list of possible diagnoses and interventions. The prompt can also include a list of criteria that must be met for a diagnosis or intervention to be valid. This expert knowledge that is added as part of the prompt can be configured in the system and continuously updated as needed based on the latest scientific findings and feedback from the clinical pathologist.

The explainability of AI models is an important requirement for use in medical diagnostics. As described in Section 2.4, this aspect is also part of the regulatory efforts for AI. Explainability is a continuous spectrum and not a binary value. In general, explainability decreases as the complexity of the models increases. The requirements for explainability are generally not binary either. Instead, they depend on what the models are used for and what risk exists in the event of misconduct. The regulatory efforts discussed in Section 2.4 also do not place any hard requirements on explainability, but rather view explainability as a sub-area of transparency, which is elementary for trustworthy AI.

The LLMs planned for use in GenDAI belong to the so-called black box models where explainability is very low. However, such models are not prohibited in current and upcoming regulation, even for high-risk applications, as long as there is transparency about the functioning and limits of the models and appropriate risk management is in place. In order to minimize the risk in GenDAI, the LLM models should only be used for the creation of draft findings, which must always be checked and approved by a clinical pathologist. In addition, the models should be designed in such a way that they provide their own verifiable justifications for the diagnoses and recommendations for action. Since current LLMs are prone to so-called hallucinations in which false information can be generated, but which is convincingly reproduced and even provided with supposed sources, it is important to sensitize the clinical pathologists to the problem and to train them accordingly. The risk of hallucinations can also be significantly reduced if a fixed framework for possible diagnoses, suitable criteria and interventions is already specified in LLM prompt.

When using LLMs for clinical interpretation and recommendation, it is to be expected that the outputs reflect existing biases in society and science. For example, sex is an important biologic variable, but women have historically been underrepresented in studies [34]. Clinical diagnostics based on these biased studies can thus also be biased to male patients, and LLMs trained on this data will likely reproduce the same bias. In this example, the LLM reflects a gap in knowledge in the scientific literature. There is no easy way to improve the outputs except by improving the underlying knowledge.

LLM training data often also contains a large amount of non-scientific training data including text from social media and the internet in general. This data includes text with discriminatory language and stereotypes, which can influence LLM output. Significant effort has been spent on identifying and mitigating these problems in LLM with some success [36].

GenDAI tries to mitigate this risk using multiple strategies. First, data minimization is used to reduce the amount of data that is fed into the LLM. While age and sex can be important for the interpretation of microbiome analysis results, other data such as name or addresses could increase the risk of biases and are thus omitted from the prompt. Second, the LLM is guided in its output by expert knowledge and instructions that is added to the prompt and can be used to correct biases if necessary. Given the narrow context of microbiome diagnostics and the use of carefully designed prompts, the risk of the LLM to output discriminatory language or biases beyond the underlying scientific knowledge is expected to be low. Nonetheless, the outputs of the LLM will also be carefully reviewed by the clinical pathologist to ensure that they are free of discriminatory language and biases.

To ensure accountability, the implementation must ensure that all input data and the resulting findings can be audited. For this purpose, the input data and the results of the analysis are persisted in such a way that a later review is possible. As many laboratories do not yet have fully digital archive systems, GenDAI also makes all relevant input data, interim results and results available as a printable report. In addition to auditability, the management of risks is also an important component of accountability. Risks must be identified, assessed, documented and minimized. Such an RMS is already required by ISO 14971 [37] and ISO 22367 [38] for all medical laboratories and therefore does not need to be newly developed for GenDAI.

The responsible use of AI requires human agency and oversight, especially in critical areas such as medicine. Human agency and oversight are already ensured by the requirement for the clinical pathologist to approve all reports. It is only necessary to ensure that the reports contain the necessary information for decision support, which is already ensured as part of accountability. Due to the still insufficiently researched risk of hallucinations in LLMs in the field of diagnostics, it is however necessary to sensitize the clinical pathologists to the problem and to train them accordingly in order to ensure effective agency and oversight.

After this description of the integration of AI into the GenDAI conceptual model (Theory Building Goal 3), the next step is to provide a conceptual architecture that describes the principal components of the system.

3.4. Conceptual Architecture

As an intermediate step before the implementation, the conceptual model can be extended by a conceptual architecture. The conceptual architecture describes the principal components of a system in a logical grouping based on their functionality. The conceptual architecture for GenDAI is shown in Figure 8. As GenDAI supports several use cases that have previously been modelled and implemented, the architecture encompasses both existing components and new components. The new components are related to the microbiome analysis and the integration of LLMs for the creation of draft findings reports.

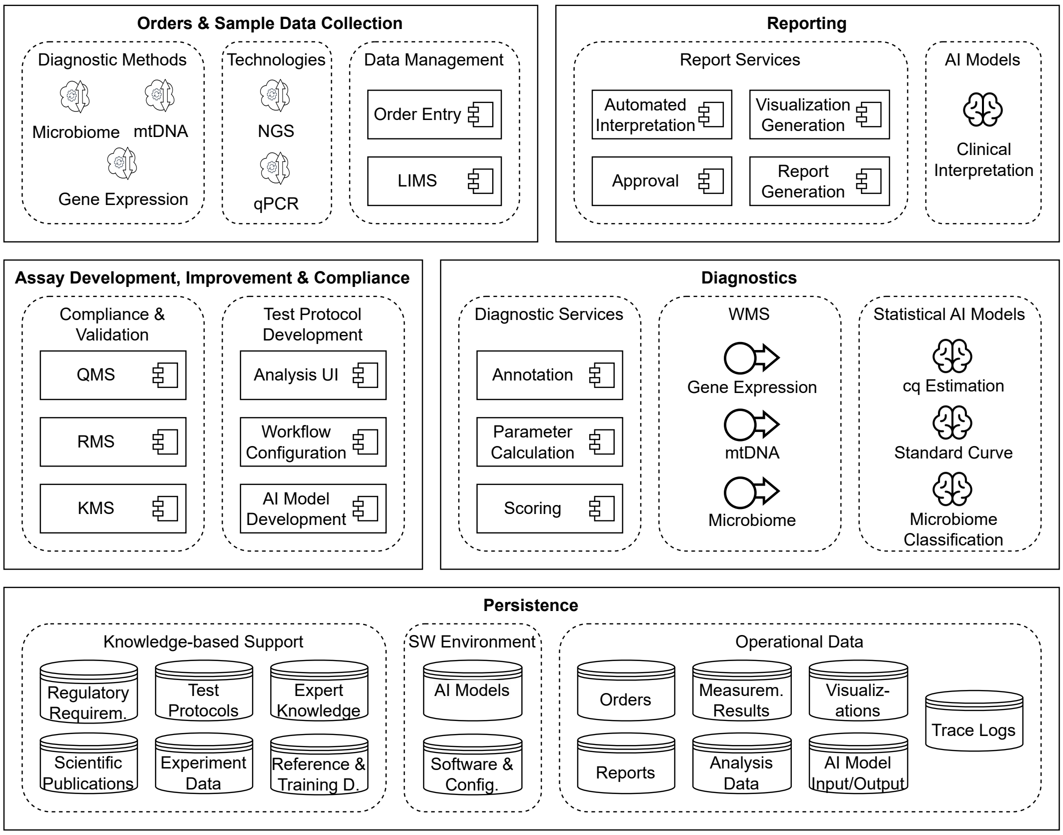


Figure 8. GenDAI Conceptual Architecture.

The conceptual architecture is divided into five principal areas arranged in three layers. The bottom layer contains the *Persistence* area, which shows the most important data stores required for the solution. These data stores are derived from the artifacts identified in the GenDAI conceptual model. The top layer represents the principal input and output of GenDAI and the laboratory as a whole. On the left side, the *Order & Sample Data Collection* area, consists of the components necessary for order entry, sample data collection, and preparation. It also shows the main input artifacts based on the measurement technologies and diagnostic methods currently supported by GenDAI. On the right side, the *Reporting* area, shows the components necessary for the creation of findings reports. This includes the LLM for result interpretation, the approval process, and all other required components for the creation of a findings report.

The middle layer is split into two areas. The *Assay Development, Improvement, and Compliance* area on the left side shows the components necessary for the development and improvement of the assays and the compliance with the relevant regulations. This includes the training or fine-tuning of the AI models and their deployment. It also includes the development and deployment of new diagnostic workflows or workflow configurations and coresponding reporting templates. Finally it includes essential components for the compliance with relevant regulations, such as the RMS or QMS.

The *Diagnostics* area within the middle layer shows the components necessary for the productive analysis of the measurement data. This includes the Workflow Management System (WMS) responsible for the execution of diagnostic workflows and all components that are used within the workflow. These components include bioinformatics tools, encapsulated calculations, and the application of AI models for other GenDAI applications beyond microbiome analysis.

Based on this conceptual architecture, which addresses Theory Building Goal 4, the next step is to implement the conceptual model and architecture in a prototype.

4. Implementation

The conceptual model and architecture described in the previous section extend the existing GenDAI conceptual model with additions for microbiome analysis and the integration of LLMs for the creation of draft findings reports. The existing model was previously implemented as a prototype with the name "GenomicInsights" [13]. The microbiome diagnostic workflow with the finding report generation was implemented as an extension of this prototype under the name "MicroFlow". This section will reiterate the key aspects of the GenomicInsights prototype and subsequently discuss the MicroFlow extension that is the focus of this paper. This implementation addresses the System Development Goal within our research approach. The code for the GenomicInsights prototype and the MicroFlow extension is available on GitHub [39].

GenomicInsights is a web-based application that allows the execution of test protocols as automated workflows using an event-driven architecture (EDA). EDA is a distributed computing paradigm that is particularly suitable for systems that process large amounts of data and require high scalability and reliability. By using asynchronous events as the primary communication mechanism, EDAs decouple components, thus increasing fault tolerance and scalability. The application of EDA to GenomicInsights is shown in Figure 9.

The core of the event communication in GenomicInsights is the *Event Hub* which is responsible for receiving and distributing events that occur within the system. Individual components can subscribe and react to events relevant to them. Components can also publish to the event hub when relevant events occur. This can be, for example, the completion of a task or the availability of new data. To further decouple components and to allow a flexible orchestration of components, GenomicInsights combines the event-driven architecture with a WMS that allows the creation of workflows that react to events and trigger new events.

Application-services are used as a backend-for-frontend to support the UI [40], while backend services operate asynchronously and support long-running operations like the analysis tasks and the report generation. The WMS is provided by Apache Airflow (v2.9), an open-source workflow automation tool. Airflow allows the definition of workflows as directed acyclic graphs (DAGs) and the execution of them in a distributed manner. Apache Kafka (v3.7) is used for the event hub to provide communication between all components of the system.

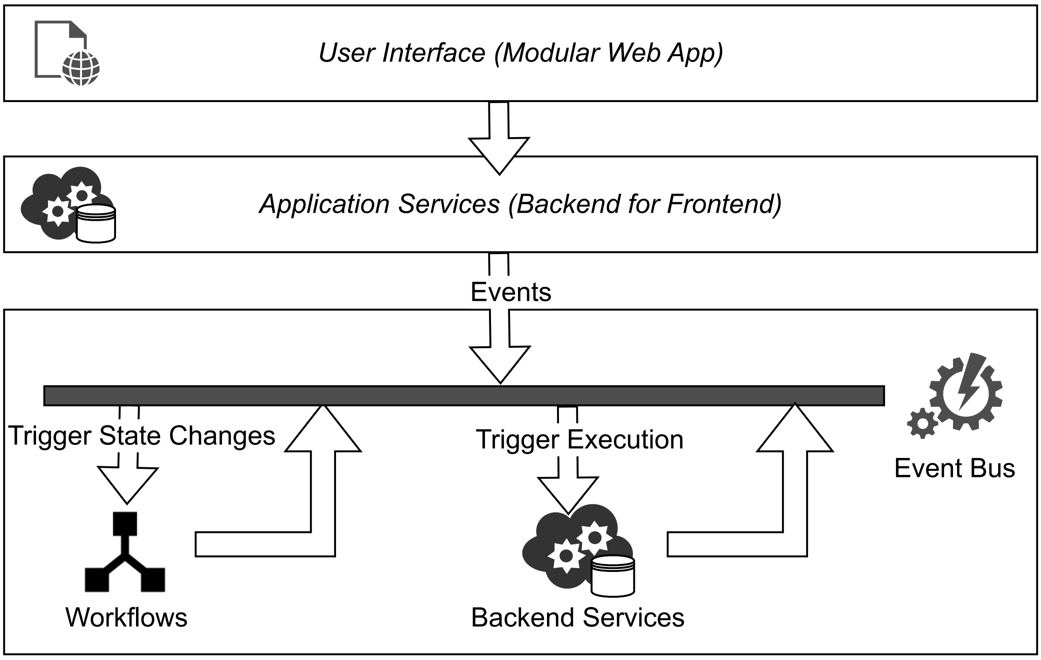


Figure 9. EDA & Microservice-Based Technical Architecture Diagram.

Using these components, the MicroFlow extension includes a workflow for microbiome analysis. As soon as the workflow is started, the Apache Airflow workflow engine sends an event to the Kafka event hub, indicating the arrival of new data. The event is received by the microbiome analysis microservice which utilizes the QIIME 2 toolset to analyze 16S rRNA sequencing data, yielding a taxonomic profile of the microbiome as previously modelled in Figure 5. After successful analysis, the microservice sends an event to the event hub, signaling the workflow to progress to the next step. As a next step the workflow triggers the creation of a draft findings report by sending a request event to the event hub, which is received by the reporting microservice.

The reporting microservice calls the underlying LLM as visualized in Figure 7 in the previous section. This call includes a prompt that contains the patient data, the analysis results, additional expert knowledge, and instructions. The instructions tell the LLM to act as a clinical pathologist and to create a draft findings report for microbiome analysis results. They also tell the LLM to format these findings in the form of a letter and to include both an interpretation of the results and recommendations for action. Listing 1 shows the used prompt template with the placeholders for the analysis results, patient data, and expert knowledge, translated into English.

The LLM to be used can be freely configured. During development, the meta-llama/Llama-2-70b-chat-hf model in the Hugging Face repository was used. To simplify the prototype implementation and evaluation of other LLMs in the future, an API for the cloud-based HuggingChat was used [41]. After the creation of the draft findings report, the microservice sends an event to the event hub, which is received by the WMS and triggers the finalization of the workflow, allowing the user to download the draft findings report.

Figure 10 shows the start screen of the GenomicInsights prototype with the MicroFlow extension. The user can manually select a workflow to execute from the list of available workflows. It should be noted that the manual triggering of workflows is only intended for testing and evaluation purposes. In productive use, the workflows would be triggered automatically by the system based on the arrival of new data files in a drop folder.

Listing 1. MicroFlow Prompt Template (Translated).

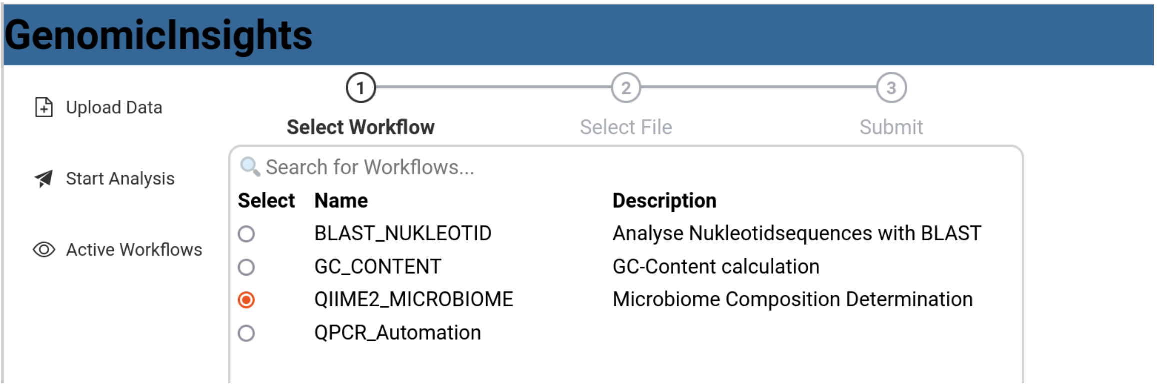
```
You are a doctor in a medical diagnostic laboratory and are
creating a detailed medical diagnostic report for a patient. The
patient has submitted a stool sample and the composition of the
intestinal microbiome was examined using 16S rRNA sequencing. The
microbiome composition is as follows: {microbiome percentages}.

The microbiome composition should be presented as a table in the
letter. Please also consider the following additional information
about the patient in the report: {patient information}.

The information about the patient should also be presented in
tabular form above the text, i.e., in the letterhead.
In addition, I provide you with the following expert knowledge:
{expert knowledge}.

Please use this expert knowledge only if the corresponding
bacteria are mentioned in the microbiome composition.
If a bacterium is not listed in the microbiome composition, this
does not mean that it is not present.
If it is not listed, no statement can be made about this bacterium.
Please do not list the expert knowledge separately in the letter.
This is only intended for you as background knowledge.

And please write a paragraph about the analysis and the comparison
to healthy donors and a paragraph with recommendations for action
to address any deficiencies in the microbiome composition.
The letter should also address the patient directly and sound
professional. Please write the letter in German. Thank you very
much!
```



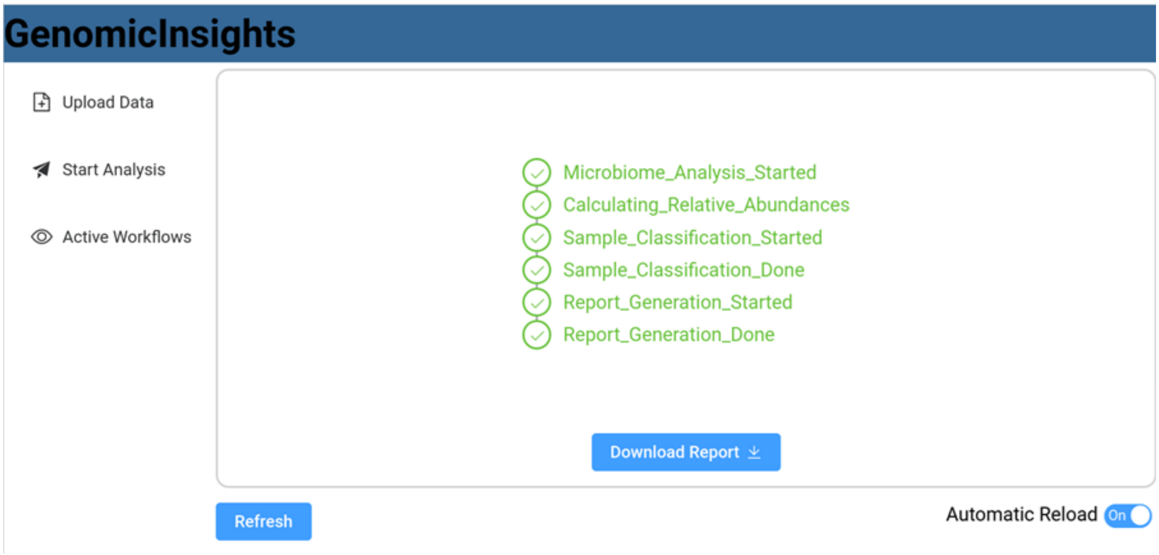
Select	Name	Description
○	BLAST_NUKLEOTID	Analyse Nukleotidsequences with BLAST
○	GC_CONTENT	GC-Content calculation
●	QIIME2_MICROBIOME	Microbiome Composition Determination
○	QPCR_Automation	

Figure 10. MicroFlow Start Screen.

After selecting the workflow and corresponding data files, the execution of the workflow starts and the user is redirected to the workflow progress screen shown in Figure 11. This screen allows the user to monitor the progress of the workflow execution.

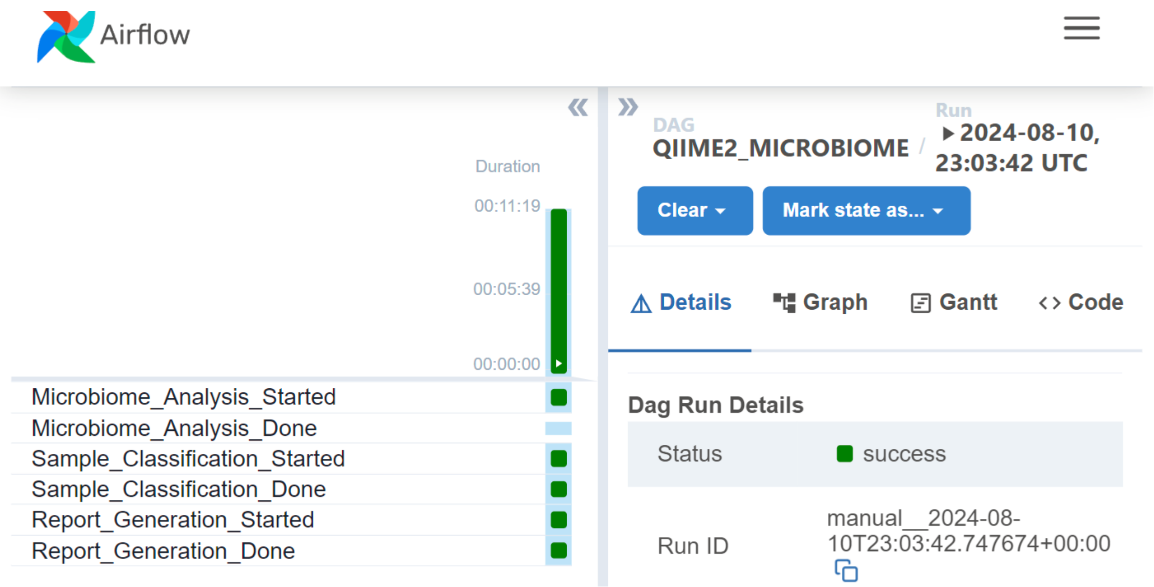
For more detailed information and for debugging purposes, the user can also use the Apache Airflow UI to view the list of workflow runs other details as shown in Figure 12. Individual instances of the workflow can also be monitored in the Airflow UI.

After successful execution, the draft findings report can be downloaded from the MicroFlow prototype. An example draft report will be discussed in the next section, which describes the initial evaluation of the MicroFlow prototype.



✓ Microbiome_Analysis_Started
✓ Calculating_Relative_Abundances
✓ Sample_Classification_Started
✓ Sample_Classification_Done
✓ Report_Generation_Started
✓ Report_Generation_Done

Figure 11. MicroFlow Workflow Progress Screen.



Task	
Microbiome_Analysis_Started	🟩
Microbiome_Analysis_Done	▢
Sample_Classification_Started	🟩
Sample_Classification_Done	🟩
Report_Generation_Started	🟩
Report_Generation_Done	🟩

DAG QIIME2_MICROBIOME / Run ▸ 2024-08-10, 23:03:42 UTC

Dag Run Details

Status	🟩 success
Run ID	manual__2024-08-10T23:03:42.747674+00:00

Figure 12. Apache Airflow UI.

5. Evaluation

One of the main challenges in evaluating the usability and feasibility of MicroFlow was to ensure that the evaluation is based on the actual needs of the users. Following the UCSD and TAI-BDM approaches, it is thus important to involve actual users in the evaluation, so that the system can be iteratively improved based on user needs. A suitable method for this is the cognitive walkthrough [42] in which one or more evaluators work through a series of tasks while identifying potential issues. To evaluate MicroFlow, a cognitive walkthrough with two expert users has been performed. This subsection describes the results of this qualitative evaluation and provides conclusions and future work. Within the selected research approach, this section addresses the Experimentation Goal, concluding the research goals of this paper.

The two expert users that participated in the cognitive walkthrough were Prof. Michael Kramer (M.D.) and Patrick Newels (M.Sc.). Michael Kramer is a clinical pathologist, head of a clinical pathology lab, and head of a laboratory consulting firm, the ImmBioMed Business Consultants GmbH & Co. KG, Pfungstadt, Germany focusing on the translational development of laboratory diagnostic procedures. Patrick Newels is the head of the medical science department at biovis Diagnostik MVZ GmbH, Limburg-Eschhofen, Germany.

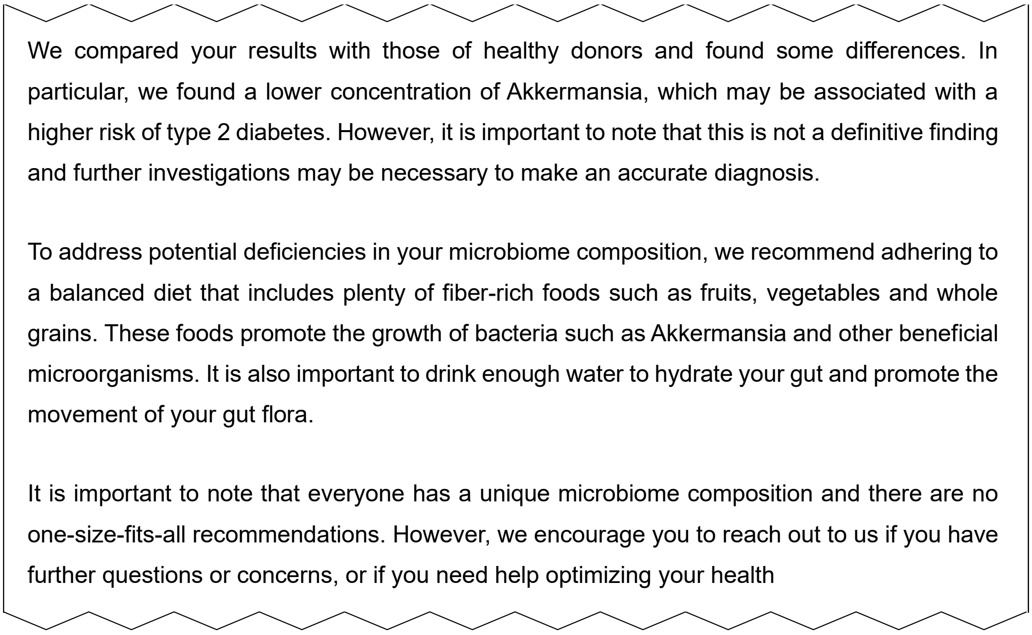
The goal of the cognitive walkthrough was to identify usability issues in MicroFlow and to evaluate the general feasibility of the proposed LLM integration. The evaluation was formative in nature, focusing on qualitative insights rather than quantitative metrics. For testing, the biovis laboratory provided sequencing data of a control sample with known concentrations of microorganisms to be processed by MicroFlow. The sequencing data was in paired-end FASTQ format. The participants were shown the user interface of MicroFlow and how to manually execute a microbiome analysis workflow for the sample data. The participants were shown how the execution can be monitored within MicroFlow. Finally, a finished draft report produced by MicroFlow was presented to the participants. The participants were asked to give feedback on the MicroFlow prototype and the generated report.

MicroFlow received positive feedback during the cognitive walkthrough, supporting the idea that the integration of LLMs into the diagnostic workflow is worth pursuing. The prototype was found to be easy to use and intuitive. Minor suggestions included the possibility to easily add workflows with different configurations to support different test protocols. The integration of the LLM was also well received, with the LLM providing plausible and useful drafts. The format of the cognitive walkthrough did not allow for a detailed evaluation of the LLM output however to determine if the LLM output was free of hallucinations. The participants suggested that the LLM output should include literature references that support its findings, so that the clinical pathologist can verify the information more easily.

To address this feedback in future work, the use of enhanced prompt engineering to instruct the LLM on the exact desired output format seems most promising. While this initial MicroFlow prototype used the Llama 2 model, future work should also evaluate the use of the recently released Llama 3 model and other LLMs. The evaluation of the LLM output should also be extended to include a more detailed analysis of the output for hallucinations and other errors. To test the ability of the LLM using additional expert background knowledge provided by the laboratory, the draft report generation was tested with an example of such background knowledge. Specifically, the LLM was told that lower concentrations of Akkermansia are associated with type 2 diabetes.

Figure 13 show the relevant excerpts of the generated draft report. The draft reports are in german as the laboratory and the participants are both located in Germany. For this paper, we translated the draft report to english using DeepL [43]. The report starts with a generic introduction to the patient and the test results as a list of the detected microorganisms and their abundance (not included in Figure 13). The report then continues with an interpretation of this composition. Specifically, as desired, the report notes the absence of Akkermansia in the sample and suggests that this could be a risk factor for type 2 diabetes. The next paragraph in the report recommends a diet rich in fiber to increase the abundance of healthy gut bacteria such as Akkermansia. The report concludes with a disclaimer that each gut microbiome is unique and that as such no recommendation will be suitable for all patients.

In conclusion we find that the integration of LLMs into the diagnostic workflow is possible and can support the automatization of the findings report creation without sacrificing human oversight. The use of expert knowledge as part of the LLM prompt was validated to extend the capabilities of the LLM without requiring custom training. Further evaluation is however necessary to determine the effectiveness and accuracy of LLMs in the context of microbiome diagnostics.



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. MicroFlow Draft Report Excerpt

6. Conclusions and Future Work

This paper presented a conceptual model and implementation for the integration of LLMs in laboratory diagnostics, enabling the creation of draft findings reports in microbiome diagnostics. Based on a structured research approach, several research goals in the area of observation, theory building, system development, and experimentation were outlined and subsequently addressed. The conceptual model and architecture extend the existing GenDAI conceptual model, while the implementation of the MicroFlow prototype demonstrates the general feasibility of the proposed LLM integration. The evaluation of the prototype showed that the integration of LLMs is possible and could support the automatization of the findings report creation. A limitation of this paper is the lack of a detailed evaluation of the LLM output on a larger, real-world patient dataset to determine the accuracy and reliability of the LLMs and to determine the risk of hallucinations. Future work will focus on evaluating additional LLMs, systematically assessing the accuracy of the AI models in a real-world clinical setting, improving the verifiability of LLM outputs, and employing enhanced prompt engineering to refine the LLM's output. As a next step, and as part of a recently awarded EU MCSA research grant, these and other improvements will be integrated into the further development of the GenDAI model and MicroFlow prototype, which will be evaluated on real patient data to enhance diagnostics and management of Inflammatory Bowel Disease (IBD).

Author Contributions: Conceptualization, T.K. and L.G.; methodology, M.L.H.; software, L.G.; validation, T.K., L.G., P.N. and M.K.; investigation, T.K. and L.G.; writing—original draft preparation, T.K.; writing—review and editing, L.G., P.N., T.R., M.X.B. and M.K.; visualization, T.K. and L.G.; supervision, M.L.H. and M.X.B.; project administration, T.K. and L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Patrick Newels is employee at biovis DiagnostikMVZ GmbH, Michael Kramer is the owner and managing director of ImmBioMed Business Consultants GmbH & Co. KG. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ogunrinola, G.A.; Oyewale, J.O.; Oshamika, O.O.; Olasehinde, G.I. The Human Microbiome and Its Impacts on Health. *Int. J. Microbiol.* **2020**, *2020*, 8045646. [CrossRef]
2. Krause, T.; Jolkver, E.; Mc Kevitt, P.; Kramer, M.; Hemmje, M. A Systematic Approach to Diagnostic Laboratory Software Requirements Analysis. *Bioengineering* **2022**, *9*, 144. [CrossRef] [PubMed]
3. Krause, T.; Jolkver, E.; Bruchhaus, S.; Kramer, M.; Hemmje, M. GenDAI—AI-Assisted Laboratory Diagnostics for Genomic Applications. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021. [CrossRef]
4. Krause, T.; Glau, L.; Jolkver, E.; Leonardi-Essmann, F.; Mc Kevitt, P.; Kramer, M.; Hemmje, M. Design and Development of a qPCR-based Mitochondrial Analysis Workflow for Medical Laboratories. *BioMedInformatics* **2022**, *2*, 643–653. [CrossRef]
5. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv* **2023**, arXiv:2303.13375.
6. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. *medRxiv* **2023**, 2023.02.21.23286254. [CrossRef]
7. Nunamaker, J.F.; Chen, M.; Purdin, T.D. Systems Development in Information Systems Research. *J. Manag. Inf. Syst.* **1990**, *7*, 89–106. [CrossRef]
8. Méndez-García, C.; Bargiela, R.; Martínez-Martínez, M.; Ferrer, M. Metagenomic Protocols and Strategies. In *Metagenomics*; Nagarajan, M., Ed.; Academic Press: Cambridge, MA, USA, 2018; pp. 15–54. [CrossRef]
9. Field, K.G.; Olsen, G.J.; Lane, D.J.; Giovannoni, S.J.; Ghiselin, M.T.; Raff, E.C.; Pace, N.R.; Raff, R.A. Molecular phylogeny of the animal kingdom. *Science* **1988**, *239*, 748–753. [CrossRef]
10. Chiarello, M.; McCauley, M.; Villéger, S.; Jackson, C.R. 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*, e0264443, [CrossRef]
11. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421, [CrossRef]
12. Bokulich, N.A.; Kaehler, B.D.; Rideout, J.R.; Dillon, M.; Bolyen, E.; Knight, R.; Huttley, G.A.; Gregory Caporaso, J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **2018**, *6*, 90, [CrossRef]
13. Krause, T.; Zickfeld, M.; Bruchhaus, S.; Reis, T.; Bornschlegl, M.X.; Buono, P.; Kramer, M.; Mc Kevitt, P.; Hemmje, M. An Event-Driven Architecture for Genomics-Based Diagnostic Data Processing. *Appl. Biosci.* **2023**, *2*, 292–307. [CrossRef]
14. Balvočiūtė, M.; Huson, D.H. SILVA, RDP, Greengenes, NCBI and OTT—How do these taxonomies compare? *BMC Genom.* **2017**, *18*, 114, [CrossRef]
15. Jolkver, E. Verarbeitung von RT-qPCR Daten in der Labordiagnostik. Bachelor's Thesis, FernUniversität Hagen: Hagen, Germany, 2022.
16. Glau, L. *Validation of qPCR Data in the Field of Medical Diagnostics*; University Project; FernUniversität Hagen: Hagen, Germany, 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, FernUniversität Hagen, Hagen, Germany, 2024.
18. Krause, T.; Andrade, B.G.N.; Afli, H.; Wang, H.; Zheng, H.; Hemmje, M. Understanding the Role of (Advanced) Machine Learning in Metagenomic Workflows. In *Proceedings of the Advanced Visual Interfaces, Ischia, Italy, 9 June and 29 September 2020*; Reis, T., Bornschlegl, M.X., Angelini, M., Hemmje, M., Eds.; Springer Nature: Berlin/Heidelberg, Germany, 2021; pp. 56–82.
19. Peng, B.; Galley, M.; He, P.; Cheng, H.; Xie, Y.; Hu, Y.; Huang, Q.; Liden, L.; Yu, Z.; Chen, W.; et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. *arXiv* **2023**, arXiv:2302.12813v3.
20. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712v5.
21. McDuff, D.; Schaekermann, M.; Tu, T.; Palepu, A.; Wang, A.; Garrison, J.; Singhal, K.; Sharma, Y.; Azizi, S.; Kulkarni, K.; et al. Towards Accurate Differential Diagnosis with Large Language Models. *arXiv* **2023**, arXiv:2312.00164v1.
22. Truhn, D.; Weber, C.D.; Braun, B.J.; Bressem, K.; Kather, J.N.; Kuhl, C.; Nebelung, S. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci. Rep.* **2023**, *13*, 20159. [CrossRef]
23. Williams, C.Y.; Miao, B.Y.; Butte, A.J. Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the Emergency Department. *medRxiv* **2023**, 2023.10.19.23297276. [CrossRef]
24. Buiten, M.C. Towards Intelligent Regulation of Artificial Intelligence. *Eur. J. Risk Regul.* **2019**, *10*, 41–59. [CrossRef]
25. Smuha, N.A. From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law Innov. Technol.* **2021**, *13*, 57–84. [CrossRef]
26. High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI*; Publications Office of the European Union: Luxembourg, 2019. [CrossRef]

27. High-Level Expert Group on AI. *Assessment List for Trustworthy Artificial Intelligence (ALTAI)*; Publications Office of the European Union: Luxembourg, 2020. [CrossRef]
28. High-Level Expert Group on AI. *Sectoral Considerations on Policy and Investment Recommendations for Trustworthy AI*; Publications Office of the European Union: Luxembourg, 2020. [CrossRef]
29. Edwards, L. The EU AI Act: A Summary of Its Significance and Scope; Ada Lovelace Institute: London, UK, 2022. Available online: https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/ (accessed on 20 August 2024).
30. Gillespie, N.; Lockey, S.; Curtis, C.; Pool, J.; Akbari, A. *Trust in Artificial Intelligence: A Global Study*; The University of Queensland: Brisbane, Australia; KPMG: Sydney, Australia, 2023. [CrossRef]
31. Bornschlegl, M.X. *Towards Trustworthiness in AI-Based Big Data Analysis*; FernUniversität Hagen: Hagen, Germany, 2024. [CrossRef]
32. The European Parliament and the Council of the European Union. *In Vitro Diagnostic Regulation*; Official Journal of the European Union: Brussels, Belgium, 2017.
33. Krause, T.; Wassan, J.T.; Mc Kevitt, P.; Wang, H.; Zheng, H.; Hemmje, M. Analyzing Large Microbiome Datasets Using Machine Learning and Big Data. *BioMedInformatics* **2021**, *1*, 138–165. [CrossRef]
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*, S367–S378, [CrossRef]
35. Norman, D.A.; Draper, S.W. (Eds.) *User Centered System Design*; Erlbaum: Mahwah, NJ, USA, 1986.
36. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
37. *ISO 14971:2019*; Medical Devices—Application of Risk Management to Medical Devices. ISO International Organization for Standardization: Geneva, Switzerland, 2019.
38. *ISO 22367:2020*; Medical Laboratories—Application of Risk management to Medical Laboratories. ISO International Organization for Standardization: Geneva, Switzerland, 2020.
39. Krause, T.; Zickfeld, M.; Müller, K.; Glau, L. GenomicInsights GitHub Repository. Available online: https://github.com/aKzenT/GenomicInsights (accessed on 20 August 2024).
40. Krause, T.; Jolkver, E.; Kramer, M.; Mc Kevitt, P.; Hemmje, M. A Scalable Architecture for Smart Genomic Data Analysis in Medical Laboratories. In *Applied Data Science*; Blum, L., Ed.; Springer: Berlin/Heidelberg, Germany, 2023.
41. Soulter. HuggingChat Python API GitHub Repository. Available online: https://github.com/Soulter/hugging-chat-api (accessed on 20 August 2024).
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*, 741–785. [CrossRef]
43. Kutylowski, J. DeepL. Available online: https://www.deepl.com/ (accessed on 20 August 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.