electronics

MDPI

Article

Augmenting Large Language Models with Rules for Enhanced Domain-Specific Interactions: The Case of Medical Diagnosis

Dimitrios P. Panagoulias *, Maria Virvou and George A. Tsihrintzis

Department of Informatics, University of Piraeus, 80 Karaoli ke Dimitriou ST, 18534 Piraeus, Greece; mvirvou@unipi.gr (M.V.); geoatsi@unipi.gr (G.A.T.)
* Correspondence: panagoulias_d@unipi.gr

Abstract: In this paper, we present a novel Artificial Intelligence (AI) -empowered system that enhances large language models and other machine learning tools with rules to provide primary care diagnostic advice to patients. Specifically, we introduce a novel methodology, represented through a process diagram, which allows the definition of generative AI processes and functions with a focus on the rule-augmented approach. Our methodology separates various components of the generative AI process as blocks that can be used to generate an implementation data flow diagram. Building upon this framework, we utilize the concept of a dialogue process as a theoretical foundation. This is specifically applied to the interactions between a user and an AI-empowered software program, which is called "Med | Primary AI assistant" (Alpha Version at the time of writing), and provides symptom analysis and medical advice in the form of suggested diagnostics. By leveraging current advancements in natural language processing, a novel approach is proposed to define a blueprint of domain-specific knowledge and a context for instantiated advice generation. Our approach not only encompasses the interaction domain, but it also delves into specific content that is relevant to the user, offering a tailored and effective AI–user interaction experience within a medical context. Lastly, using an evaluation process based on rules, defined by context and dialogue theory, we outline an algorithmic approach to measure content and responses.

Keywords: AI-empowered software engineering; generative AI; dialogue theory; large language models; natural language processing; rule-augmented systems; medical diagnosis; evaluation

Citation: Panagoulias, D.P.; Virvou, M.; Tsihrintzis, G.A. Augmenting Large Language Models with Rules for Enhanced Domain-Specific Interactions: The Case of Medical Diagnosis. *Electronics* **2024**, *13*, 320. https://doi.org/10.3390/electronics13020320

Academic Editors: Fabio Grandi and Liang Song

Received: 23 November 2023
Revised: 31 December 2023
Accepted: 9 January 2024
Published: 11 January 2024

1. Introduction

Healthcare experiences for patients are multifaceted, encompassing dynamic doctor–patient interactions, diverse diagnosis and treatment methods, adherence to recommended lifestyle or suggested behavioral changes, and ongoing preventive health measures. A patient's healthcare journey is clearly non-linear, forming a comprehensive and interwoven sequence of events and encounters [1–4]. For example, the diagnostic procedure in medicine often combines different approaches, which are influenced by the context, patient symptoms, clinician expertise, and available diagnostic tools [5]. Indeed, as illustrated and summarized in Figure 1, the diagnostic process begins with gathering patient data, including a medical history and possibly a physical examination. This information is then analyzed for patterns to assist in decision making. The process further refines and confirms initial hypotheses about the condition using the collected data, which leads to the creation of a treatment strategy, the monitoring of patient progress, and the tracking of disease progression.

Recent advancements and ongoing research have allowed significant progress in digitizing a great portion of the healthcare process, with the aim to alleviate the burdens and costs of primary care, while improving patients' experiences. Some methodologies utilize natural language processing (NLP), big data analysis, and machine learning (ML) technologies [6–8] to digitize, compress, and accelerate healthcare processes. Indeed, these

technologies are promising to revolutionize patient care and disease management by automating tasks, streamlining workflows, reducing manual labor, and simplifying daily activities for all stakeholders [9,10].

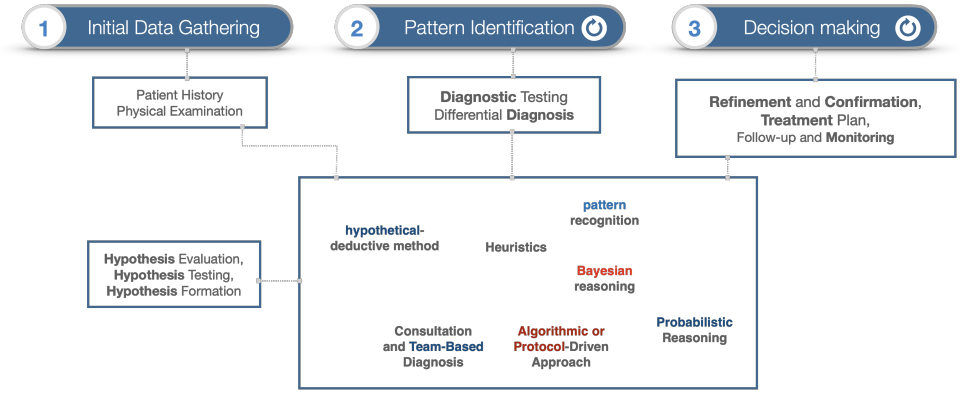

Figure 1. Medical diagnosis pathways.

One such emergent technology showing great promise to revolutionize healthcare is the technology of large language models (LLMs). Indeed, LLMs demonstrate a remarkable capability of understanding medical texts and identifying (diagnosing) a range of symptoms and health conditions. An exemplary LLM is GPT by OpenAI, powering ChatGPT, which generates accurate, human-like text responses [11]. Other notable LLMs include Google's BERT (Bidirectional Encoder Representations from Transformers) [12], Meta's Llama (Large Language Model Meta AI) [13], and Stanford's Alpaca (fine-tuned from the Llama model) [14]. While each LLM and NLP approach has its limitations, selectively integrating elements from various technologies can offer both efficacy and cost-effectiveness.

In recent studies, a novel general three-step methodology was proposed to evaluate the potential of LLMs and, more specifically, ChatGPT in medical diagnosis and treatment [15]. The evaluation of ChatGPT's performance, as per its communication capability in radiology [16] and oncology [17], has also been conducted. It was found that, under different circumstances, ChatGPT performed at an average to optimum level. Moreover, it was found that ChatGPT and other NLPs/LLMs could potentially perform better under the supervision and assistance of a medical expert, who could evaluate the ChatGPT answers better than a patient.

Based on these previous findings, in our current work, we introduce a novel rule-augmented AI-empowered system in which a rule-based decision mechanism is integrated with an LLM engine and various external machine learning and analytical APIs.

Our system includes the following novelties and key contributions.

- The domain space of AI–user interaction is associated with rules of dialogue to be followed, as detailed later in the paper in Table 1 in Section 4.1. This provides a theoretical basis for the evaluation of the performance and the assessment of an LLM's ability to remain within these constraints, which aim to simulate real-time/real-world interactions. The process is systemized and generalized to reach a measurable conclusion on the LLM answers, within a domain-specific context and a dialogue-defined space.
- Using NLP algorithms, we define the blueprint of domain-specific knowledge and the domain-specific content that is relevant to the user. This enhances the AI–user interaction experience within a medical context.
- A methodology is introduced, represented through a process diagram, aimed at defining generative AI processes and functions with a rule-augmented approach for the prototyping of AI-empowered systems.
- The system, which utilizes the GPT-4 engine, has undergone extensive evaluation through multiple-choice questions that focus on symptomatology in the field of general pathology.

The previous functionalities have been fully implemented in our rule-augmented AI-empowered system and are presented in the remaining sections of the paper. Overall, our system is characterized by the incorporation of ML tools that simulate several of the common tools used by a general practitioner in an initial physical examination. Enhanced with these functionalities, the current version of our system provides medical assistance, closely replicating the behavior, objectives, tasks, and tools of a general practitioner when offering diagnostic recommendations to primary care patients.

More specifically, the paper is organized as follows. Section 2 is devoted to highlighting background theories and context with regard to both LLMs and the state of primary care worldwide. Section 3 presents an overview of the developed rule-augmented AI-empowered system. Section 4 details the system architecture from a micro and a macro level. Section 5 includes a system evaluation and Section 6 summarizes the paper, articulates and presents its key findings, and offers insights on future related research endeavours.

2. Background Theories and Context

In this section, our focus is to establish a comprehensive background pertinent to the methodologies used. We delve into various aspects of NLP and explore its diverse applications within the medical sphere. Notably, the integration of NLP and LLMs in healthcare has been significant [18]. These technologies are increasingly employed for a range of purposes, including the extraction of vital data from electronic health records (EHR), supporting decision making in clinical settings, and analyzing patient sentiments through their reviews and feedback [19,20].

2.1. Natural Language Processing

NLP, a pivotal AI sub-field, focuses on the interaction and interpretation of human language by computers. It facilitates various tasks, including translation, sentiment analysis, and conversational interfaces. The evolution of NLP spans from rule-based approaches to sophisticated ML techniques, giving rise to advanced models such as GPT and BERT [12,21,22]. Essential concepts in NLP encompass tokenization, part-of-speech tagging, named entity recognition, and parsing. The overarching aim is the effective comprehension of human language, facilitating the extraction of meaning and simulation of reasoning to accomplish specific tasks. NLP employs an array of techniques and models, ranging from rule-based systems to advanced ML algorithms. Prominent models in NLP include the following.

- NLP using pattern matching and substitution: These initial NLP systems depend on manually crafted rules and lexicons. An iconic example is the ELIZA chatbot [23], created in 1964. ELIZA was one of the first programs capable of attempting the Turing test.
- ML models: This category encompasses traditional models like naive Bayes, support vector machines (SVM), and decision trees, commonly applied in text classification and sentiment analysis.
- Neural networks: Inspired by the human brain, these models include recurrent neural networks (RNNs) and convolutional neural networks (CNNs), suitable for tasks needing an understanding of a language's sequential nature.
- Embedding models: These models produce dense vector representations of words or larger text units, capturing semantic meanings. Notable examples include Word2Vec, GloVe, and FastText [24].
- Sequence-to-sequence models: Capable of transforming input sequences into output sequences, these models are integral to machine translation and text summarization, often based on an encoder–decoder architecture with attention mechanisms [25].
- Large language models (LLMs): LLMs are designed to perform a wide range of NLP tasks, from translation to question answering and to text generation, without needing task-specific training data. LLMs are further discussed in the following section.

2.2. Large Language Models (LLMs)

The GPT series, including GPT-3 and GPT-4, comprises autoregressive models known for generating contextually coherent text. GPT decodes the received input, using language pattern understanding, to produce a relevant and coherent output. GPT models are especially powerful for applications like content creation, dialogue generation, and tasks that require the production of new text based on given prompts.

In contrast, BERT operates by analyzing both preceding and succeeding words in a sentence, thereby enriching its understanding of the sentence context. Both models are built upon the Transformer architecture, first introduced in [22], which employs an "attention" mechanism to assign varying significance to different words. Central to these models is an encoder, which converts sequences of words into contextually enriched vector representations. The novel self-attention mechanism in these models allows them to consider the inter-dependencies of words over longer ranges, significantly improving their predictive accuracy. Notably, BERT employs a bidirectional training approach, enabling word prediction based on both the preceding and subsequent context. This is in contrast to GPT's unidirectional methodology.

Prior to the widespread adoption of neural networks and Transformer models in NLP, statistical models were the mainstay. Key among these were the following.

- Markov Models: Based on the principle named after mathematician Andrey Markov, these probabilistic models assume that the probability of each subsequent state depends only on the current state. Their application is particularly notable in sequential tasks like language modeling.
- Hidden Markov Models (HMMs): An extension of Markov models, HMMs include hidden states and observable outputs. They find applications in NLP tasks, notably in part-of-speech tagging and named entity recognition.
- Conditional Random Fields (CRFs): These are statistical frameworks used in NLP to model the probability of outputs given specific inputs. Unlike HMMs, CRFs take into account the entire sequence of words, thereby yielding more accurate results.
- n-gram Models: These models predict the next item in a sequence by considering the previous $(n - 1)$ items. Predicated on the assumption that a word's probability is dependent solely on its preceding words, n-gram models are prevalent in areas like speech recognition and machine translation.
- Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups. In NLP, these groups or topics help us to understand why data parts are similar, positing each document as a topic mixture with each word attributed to a document's topic.

2.3. Problems with NLP and Evaluation Pipeline

Below, we list some important problems and concerns associated with NLP, especially when employed in the medical domain. Some of these are currently being addressed by the companies that commercially provide the state-of-the-art models.

- Hallucinations: Generation of outputs that seem plausible but are entirely fabricated or inaccurate [26,27].
- Bias: LLMs learn and reproduce the biases that exist in their training datasets [28].
- Lack of explainability: Generative AI systems typically do not provide explicit explanations for the conclusions that they reach or the answers that they provide [29,30]. Explainable AI (XAI) ensures that users comprehend the characteristics of the utilized models and provide a transparent representation of the used algorithms that generate a response, a classification, or a recommendation. Considering user ability and adding personalization in XAI is also an important factor that can increase transparency and lead to the greater adoption of AI-empowered systems [31]. Current GAI systems lack explainability, particularly in terms of personalized explanations.

- Real-time validation: The responses are not derived from real-time information. Instead, they are based on the dataset that was used to train the model that typically contains information from a period up to the date of the training of the tool [27].
- Limitations in mathematical operations: This limitation is partially addressed using Python modules for calculations and by providing updated models more frequently.
- Content—token size limitation: This limitation is partially addressed by increasing the token size limits and charging higher usage costs.

In previous works [15,32], we proposed a methodology to evaluate the domain-specific proficiency of ChatGPT or other LLMs, focusing on reliability and precision. These metrics are based on the context of the answers, their accuracy, and the quality of the references used. Our approach utilizes a three-tiered scoring scale (1–3) to assess various aspects, categorizing the context, references, and value added to the system as follows:

- correct (3), generic (2), or incorrect (1);
- actionable (3), generic (2), or non-actionable (1);
- precise (3), generic (2), or misleading (1);
- under-extended (2), exactly aligned (1), or over-extended (1).

The evaluation specifically focuses on (A) the validity and accuracy of answers as per the context and references returned in the LLM response, (B) the specificity and usefulness of the LLM-generated response to physicians and patients alike, and (C) the economic value (potentially) added to the system. The entire assessment process is overseen by a medical professional and can be seen in Figure 2.

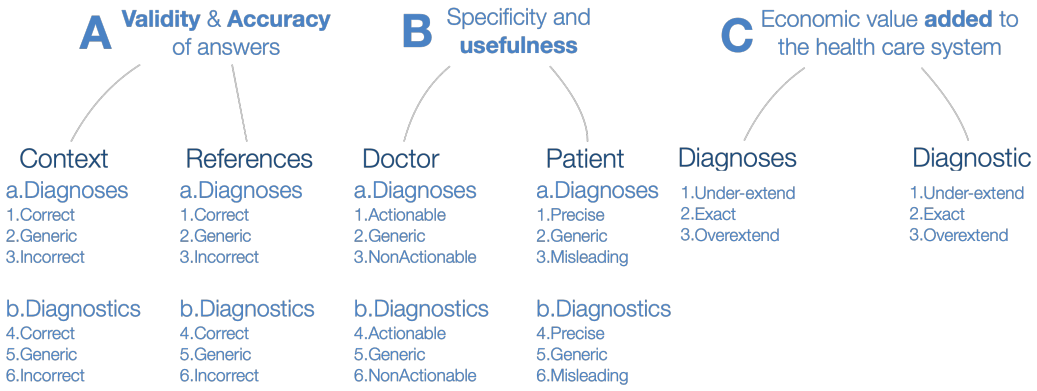

Figure 2. Methodology for evaluation of the domain-specific proficiency of ChatGPT.

2.4. Transformers and Attention Mechanism

The Transformer model has been very influential in the field of NLP and constitutes the engine of the state-of-the-art LLMs, still powering the latest ChatGPT engine as of the latest update of November 2023. In Figure 3, the architecture of a Transformer is presented, along with a description of each step and a brief explanation of the related mathematical formulae. This section contextualizes our study within the broader scope of NLP progress, but also provides a necessary technical foundation for the analysis and development of further innovations in the generative artificial intelligence (GAI) space.

The Transformer model is based on a mechanism, referred to as self-attention, that directly models the relationships between words in a sentence, regardless of their respective positions in the sentence.

- Encoder: The left part of the diagram represents the encoder, which processes the input data. The input sequence is processed through multiple layers of multi-head attention and feed-forward networks, with each of these layers followed by the residual connection and linear normalization steps.
  - Embeddings: The numerical representations of words, phrases, or other types of data. In the case of LLMs, they represent words or tokens. Each word or

token is mapped to a vector of real numbers that captures semantic and syntactic information about the word. The words with similar meanings or used in similar contexts will have similar vector representations.

    * Input Sequence Embedding: Input tokens' conversion into vectors of a fixed dimension.

- Positional Encoding: Adds information about the positional order of the respective words of the sequence.
- Multi-Head Attention: Applies self-attention multiple times in parallel to capture different aspects of the data. This allows joint attention to information from different representation subspaces, referred to as heads, at different positions. Using multiple heads, the model captures different types of dependencies from different representational spaces. For example, one head might learn to pay attention to syntactic dependencies, while another might learn semantic dependencies. The mathematical representation of this is as follows. Let us we denote the linear transformations that produce the queries, keys, and values for head i with W_i^Q, W_i^K, W_i^V, respectively, and the output linear transformation with W^O. Then, the multi-head attention operation MultiHead can be defined as

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O,$$

where each head head_i is computed as

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

and Attention is the scaled dot-product attention function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Here, d_k is the dimensionality of the key vectors, while the division by $\sqrt{d_k}$ is the scaling factor.

1. Q, K, V: The input to the multi-head attention layer is first linearly transformed into three different sets of vectors: queries (Q), keys (K), and values (V). This is done for each attention head using different, learned linear projections.
2. Scaled dot-product attention: For each head, the scaled dot-product attention is independently calculated. The dot product is computed between each query and all keys, which results in a score that represents how much focus to place on other parts of the input for each word. These scores are scaled down by the dimensionality of the keys (typically the square root of the key dimension) to stabilize the gradients during training. A softmax function is applied to the scaled scores to obtain the weights on the values.
3. Attention output: The softmax weights are then used to create a weighted sum of the value vectors. This results in an output vector for each head that is a combination of the input values, weighted by their relevance to each query.
4. Concatenation: The output vectors from all heads are concatenated. Since each head may learn to attend to different features, concatenating them combines the different learned representation subspaces.
5. Linear transformation: The concatenated output undergoes a linear transformation to produce the final product of the multi-head attention layer.
6. Feed-forward network: A fully connected feed-forward network is applied to each position separately and identically.

7. Residual connection and linear normalization: Applies residual connections and layer normalization.

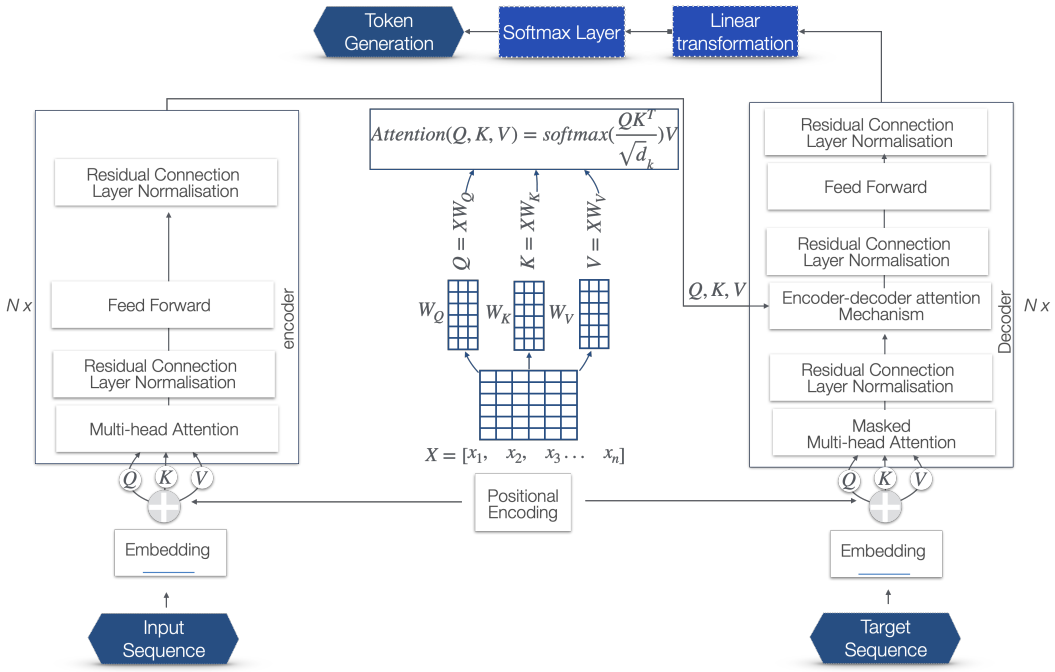

Figure 3. Transformer—data flow chart.

- Decoder: The right part of the diagram represents the decoder that generates the output.
  - Target sequence embedding: Converts target tokens into vectors and shifts them to the right.
  - Masked multi-head attention: Prevents positions from attending to subsequent positions during training.
  - Encoder–decoder attention mechanism: Attends to the encoder's output and the decoder's input. The keys (K) and values (V) come from the output of the encoder. The similarity between the queries and keys is calculated. This involves taking the dot product of the queries with the keys, scaling it (usually by dividing by the square root of the dimension of the key vectors), and then applying a softmax function to obtain the weights for the values.
  - Feed-forward network: Following the attention mechanisms, there is a feed-forward network. It consists of two linear transformations with a ReLU activation in between.
  - Residual connection and linear normalization: Applies residual connections and layer normalization.
  - Linear transformation before softmax: In the final layer of the decoder, the Transformer model applies a linear transformation to the output of the previous layer. This linear transformation, typically a fully connected neural network layer (often referred to as a dense layer), projects the decoder's output to a space whose dimensionality is equal to the size of the vocabulary.
  - Softmax function: After this linear transformation, a softmax function is applied to these projected values, which creates a probability distribution over the vocabulary based on the positional attributes.
  - Token selection: The probability distribution for each potential token is analyzed considering the context of the sequence. This analysis determines which tokens are most likely to be the appropriate next elements in the sequence. The token

selection can be done using various strategies like greedy decoding, sampling, or beam search.
- Token generation: Based on this probability distribution, tokens are generated as the output for each position in the sequence.
- Sequence construction: The selected tokens are combined to form the output text sequence. This can involve converting sub-word tokens back into words and dealing with special tokens such as those that represent the start and end of a sentence.
- Post-processing: Post-processing is performed, based on syntactical, grammatical, and language rules.

This architecture is highly parallelizable and reduces the need for recurrent or convolutional layers. Further details of this architecture are analyzed in [22]. By providing a previous basic overview of how a Transformer works, we can show how the methodology that we use provides an efficient shortcut to creating our domain-specific system's engine.

2.5. Telehealth

Telehealth uses digital technologies to remotely deliver selected healthcare services. Telehealth's importance is prominent where resources are limited and the goal of healthcare cost reduction is important. The continuous monitoring of treatment is another important domain of application of telehealth. It also aims to educate patients and healthcare providers, support consultations between primary care providers and specialists for quicker diagnoses and treatment, more effectively manage hospital patient loads, and more actively engage patients in their own care [33,34].

Telehealth has been widely adopted across various medical specialties due to its versatility. Primary care can handle routine check-ups and minor health issues remotely, while psychiatry and psychology benefit from teletherapy and telepsychiatry. Similarly, radiology allows for the remote sharing and analysis of medical images, and cardiology and neurology utilize remote monitoring for conditions like heart rhythm abnormalities and epilepsy. Dermatology and endocrinology practices can remotely diagnose and manage skin conditions and diseases like diabetes. Geriatrics and pediatrics are also benefiting from telehealth, especially for patients with mobility issues or for the management of minor concerns and follow-ups. Chronic disease management, including hypertension, COPD, and asthma, is another area where telehealth plays a crucial role.

As technology advances, more medical specialties are incorporating telehealth into their practices. While physical examinations remain a limitation, many conditions can be diagnosed and treated effectively using a combination of the patient history, visual examination, and remotely collected data. However, the above areas do not constitute an exhaustive list and the potential for telehealth continues to grow.

2.6. The State of Primary Care

In many healthcare systems worldwide, including those in Europe and the United States, costs and delays [35] are significant issues, although their intensity and nature vary by region. In the European Union, most countries have universal healthcare systems funded through taxation or mandatory health insurance, with some variations in patient costs and the option of private insurance. While the primary care quality is generally high, there are differences between and within countries, with concerns often centered around waiting times for specialty care.

In contrast, the USA operates mainly on an insurance-based system, with many facing co-payments, deductibles, and other out-of-pocket costs. This can deter some from seeking primary care, and the quality of care varies by factors like location and socioeconomic status. Debates about the benefits and drawbacks of moving towards a universal healthcare system are still ongoing [36].

Both the European Union and the USA face challenges in primary care accessibility and preventable mortality. Over 100,000 deaths annually in the USA may be due to preventable

medical errors, including access failures. Similarly, thousands may be dying annually in the European Union due to preventable causes. Treatable mortality [37,38] rates are published yearly both in Europe and the USA. In 2020, the treatable mortality rates accounted for 39 deaths per 100,000 population in Switzerland (lowest) and for 225 deaths per 100,000 population in Hungary (highest) [39]. These challenges are intertwined with issues like quality of care, equity of access, and patient satisfaction, and each region addresses them based on its specific healthcare models and cultural values.

Lastly, many developing countries have limited budgets for healthcare, which can lead to inadequate infrastructures, low salaries for healthcare workers, and insufficient medical supplies.

2.7. NuhealhtSoft: An AI-Empowered Software Platform for Medical Exam Classification and Health Recommendations

NuhealthSoft is a recently developed software platform to facilitate users in understanding their blood exams using various presentation and analytical techniques, including semantic grouping. The system also provides its users with services that identify patterns and, thus, health states in their blood exams and dietary intake. In more detail, NuhealthSoft [40] employs advanced ML techniques combined with comprehensive nutritional and biochemical data. This approach enables the platform to effectively categorize individuals based on various health metrics. These metrics include blood pressure, weight, and indicators of metabolic syndrome, as supported by several studies and references [10]. Additionally, NuhealthSoft moves beyond mere classification by offering personalized nutritional advice. This guidance is specifically tailored to meet the unique dietary needs and blood work results of each user.

The development and refinement of NuhealthSoft necessitate close collaboration with medical professionals and regulatory bodies. These stakeholders play a crucial role in validating the system's effectiveness and compliance with health standards. This collaboration has been a pivotal aspect of our research. We have focused on understanding and integrating the specific requirements of doctors to facilitate a seamless validation process. Simultaneously, we aim at maintaining a transparent and user-friendly framework for the end-users of NuhealthSoft. This dual focus ensures that the system is not only medically sound and compliant, but also accessible and beneficial to those whom it serves.

In this paper, Med | Primary AI assistant is presented, which has been developed as an an add-on for the NuhealthSoft suite. Specifically, Med | Primary AI assistant has been developed to analyze symptoms and provide health advice from a general practitioner's perspective.

3. System Overview of Med | Primary AI Assistant

In Figure 4, the objectives, tasks, and available tools are outlined. The purpose is to build an AI-empowered system that can perform these objectives and complete the tasks with use of the available tools. In essence, Figure 4 provides the blueprint of the domain-specific knowledge of the system. To ensure that the shortcomings of LLMs are addressed, the system also encompasses rules, i.e., a rule-augmented application is developed. The rules are used to

- engineer prompts based on domain specification;
- extract semantically important words and associated with external services and classifiers and external sensors; and
- create an evaluation basis, to ensure alignment with domain specifications and requirements based on the dialogue's theoretical context.

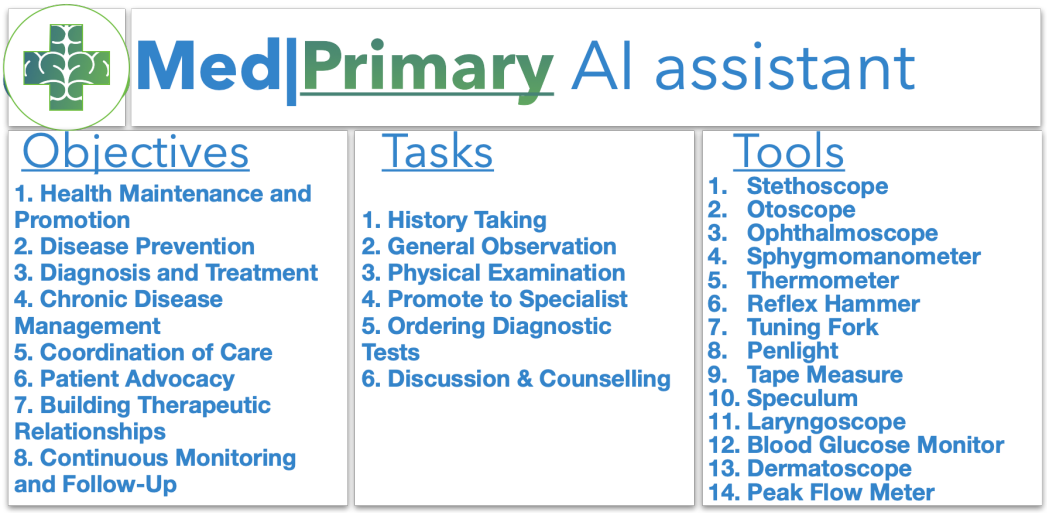

Figure 4. General practitioner's objectives, tasks and tools.

3.1. System Description

Med | Primary AI assistant, included in the NuhealthSoft suite, utilizes LLMs; for the purpose of this study, we have used and tested the GPT-4 model.

GPT-4 is an advanced multimodal model, currently processing text inputs and producing text outputs and chat completion tasks. It outperforms previous models with its extensive general knowledge and enhanced reasoning skills. While it shares similarities with GPT-3.5-turbo in being optimized for chat interactions, it is also proficient in executing traditional completion tasks.

Our system also encompasses analytical services and ML models to extract useful information from health data, while only providing the necessary information. The limiting of token usage maintains a manageable input and also ensures the computational and mathematical validity of the provided information, thus augmenting the quality of the response.

3.2. Use Cases

In Med | Primary AI assistant, a user can interact with the system in two main ways.

- The first is by freely (without constraints and rules) providing symptoms and descriptions of their health state and obtaining a series of diagnoses, proposed diagnostic exams, or a referral to a medical specialist. While, in this case, the patient has no constraints, using specific knowledge input, the LLM will provide assistance if the user's input is not useful for the LLM to complete its predefined tasks and objectives.
- The second is by using a more constrained and step-by-step approach, for the LLM to obtain a more comprehensive background on the user's symptomatology and age. In both cases, data can be retrieved by health sensors and analyzed by the included analytical and machine learning services [41].

In Figure 5, two main actors are presented. The first actor is the patient, who will provide the symptoms directly to Med | Primary AI assistant or via form inputs. Health sensors can provide more context and data. Finally, the patient can review the process and output. The doctor, as the second actor, can evaluate and validate the primary care AI interactions (inputs–outputs) and the patient's review of the the primary care AI. This process is essential for the system to improve and for more services to address primary care AI shortcomings.

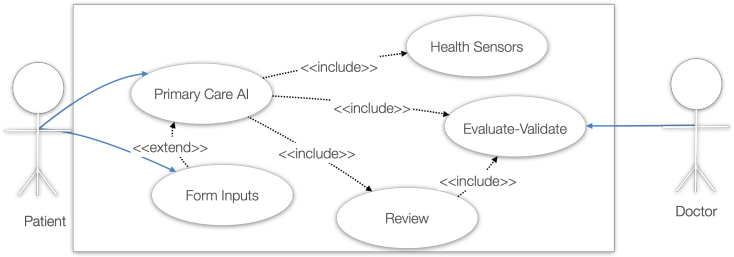

Figure 5. Med | Primary AI assistant use case.

4. System Architecture Analysis

In this section, the structural elements, flow of data, and organization of Med | Primary AI assistant are outlined and described. As shown in Figure 4, the specific objectives, tasks, and tools are considered as building blocks. For example, if a service is provided to facilitate the process of diagnostic analysis, it will only belong in the domain of a general practitioner's competencies and, thus, constructs the blueprint of domain-specific knowledge of the system.

4.1. Modeling the Domain Space

The main sources of the system pertain to the management of inputs and outputs during a conversation between a patient and a general practitioner, specifically addressing questions and answers [42–44].

For context, there are six (6) types of theoretical questions and answers that can be applied in any domain.

1. Questions
   - (a) Informational: Query for specific information.
   - (b) Instructional: Query related to specific task, i.e., a command to do something.
   - (c) Reflective: To confirm or clarify previous statements.
   - (d) Rhetorical: Are not meant to be answered and are rather used for emphasis.
   - (e) Open-ended: Are meant to encourage a detailed response or discussion.
   - (f) Closed-ended: Can be answered with a a yes or no.
2. Answers
   - (a) Direct: Provide a straightforward response.
   - (b) Elaborated: Provide additional context and information beyond what was requested.
   - (c) Clarifying: Aim at requiring clarity, where a query is ambiguous.
   - (d) Reflected: Ensure that the question is answered in a way that mimics the question's sentiment.
   - (e) Deferred: When an answer cannot be provided and the one that provides it offers guidance on where or how to find it.
   - (f) Non-Answers: When the choice is to not answer.

4.2. Domain Settings

Questions and answers are the building blocks of the conversations and the input and output of the flow chart pictured in Figure 6. A conversation in the medical domain usually includes informational and open-ended questions, followed by direct or elaborated answers. For the case of this system (Table 1), a deferred answer is also an option when the question cannot be answered, either due to the fact that it requires a more specialized evaluation (i.e., a referral to a specialist) or when it is outside the scope of the domain (dialogue reset). Dialogue rules are set to

- provide a basis for evaluating the performance;
- assess the model's ability to remain within constraints that aim to simulate real-time communication protocols.

Table 1. Dialogue rule augmentation.

Question (q)	Domain (d)	Medical (m)	Answer (a)	Answer Content (c)	Use Case Examples
Open Ended	Yes	Yes	Direct or Elaborated	Define Ability (Ab), Reflect (Re)	Figure 7
Informational	Yes	Yes	Direct or Elaborated	Define Ability (Ab), Reflect (Re)	Figure 8
Any	No	Yes	(deferred) Refer To Specialist	Define Ability (Ab), Reflect (Re)	Figure 9
Any	No	No	(deferred) Dialogue Reset	Define Ability (Ab), Reflect (Re)	Figure 10

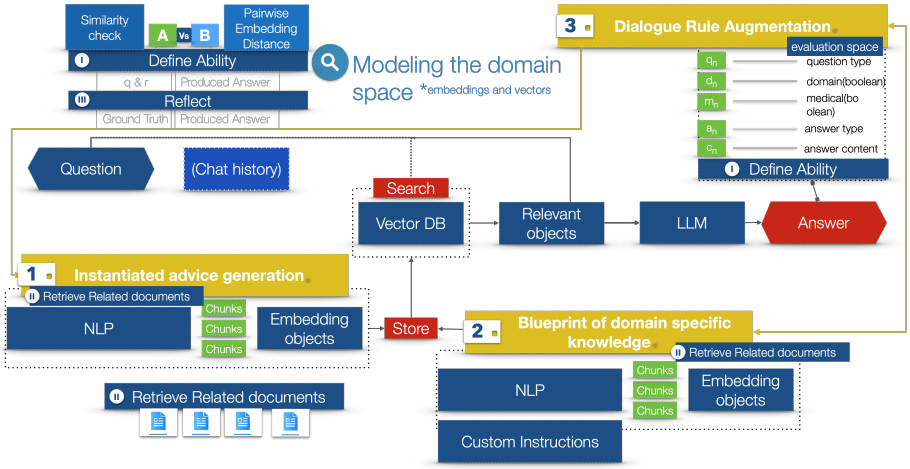

Figure 6. Domain settings, flow chart.

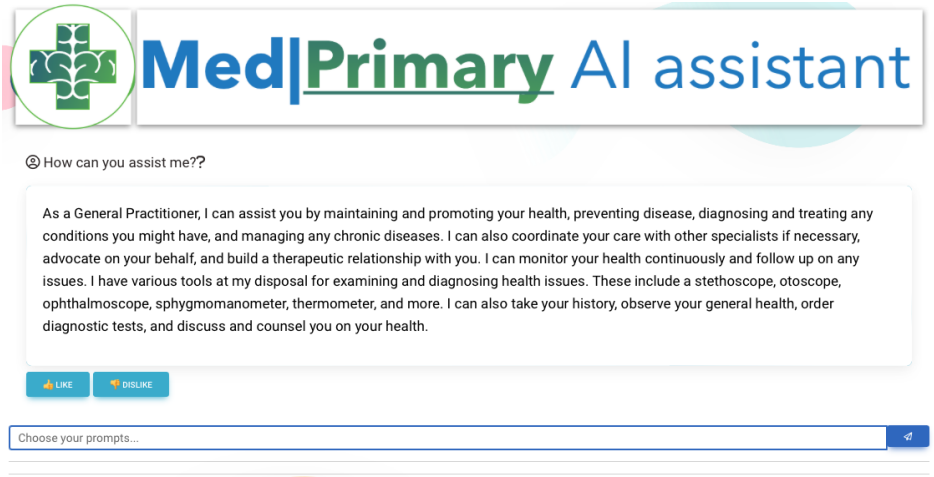

Figure 7. Open-ended question—direct answer, system objectives, and tasks.

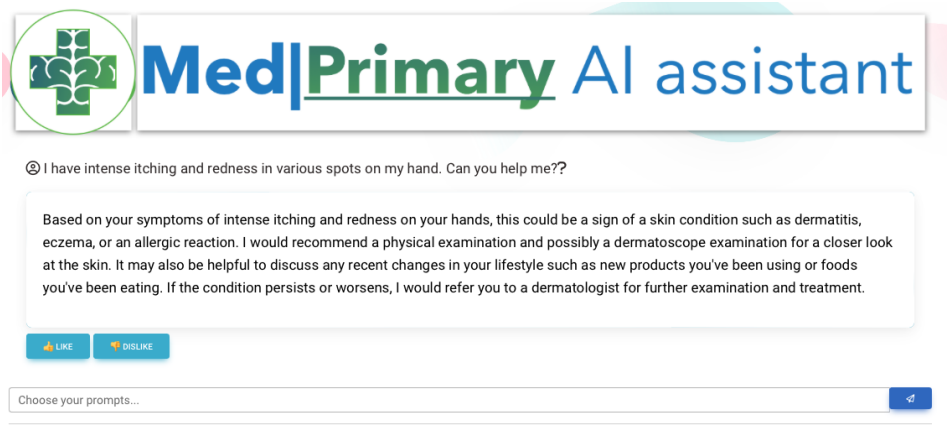

Figure 8. Informational question—elaborated answer, skin problem with referral.

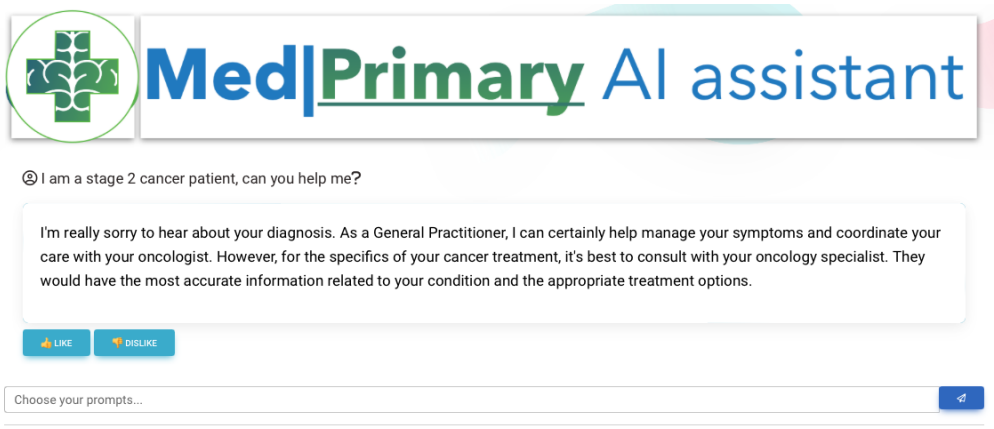

Figure 9. Open-ended question—deferred answer, cancer patient.

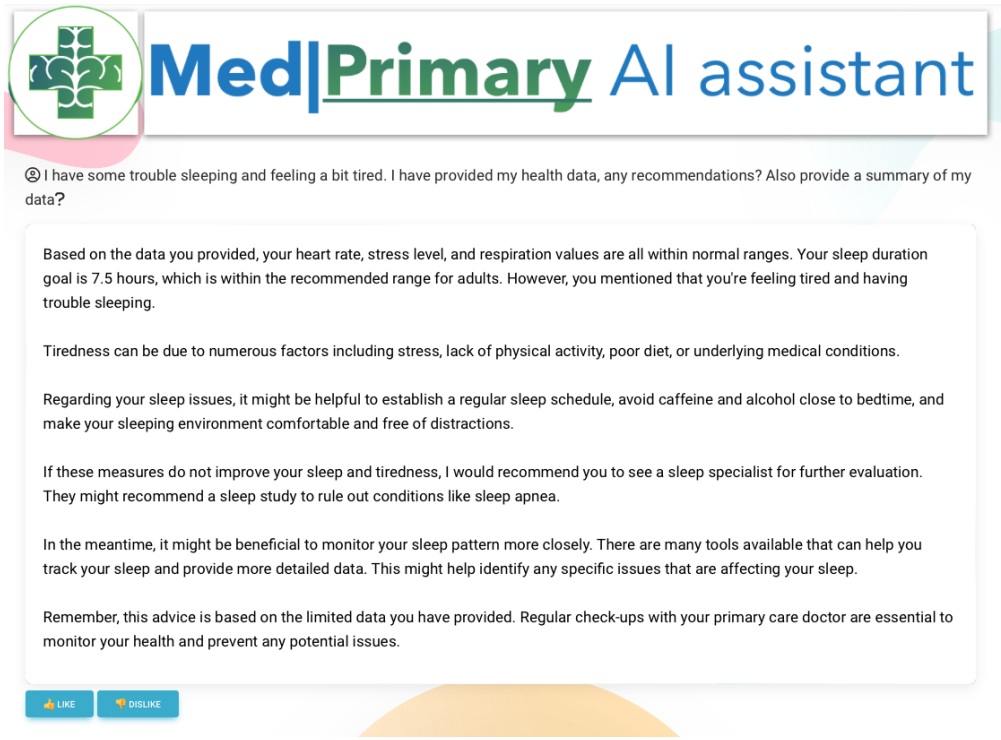

Figure 10. Informational question—elaborated answer, symptoms with health metrics.

The instantiated advice generation and blueprint of domain-specific knowledge encourage the structuring of informational and open-ended questions. Rules are also defined to lead to a certain type of answer that ensures a diagnosis, the promotion of necessary diagnostics, or the proposal of a medical specialist, mostly associated with the described symptoms.

To ensure the safety of users, the Dialogue Rule Augmentation stage (as shown in Table 1) establishes the framework for evaluating the extracted answers, as depicted in Figure 6. This evaluation is conducted through the 'Define Ability Process (I)' and the 'Reflect Process (III)', which are incorporated as functions in Algorithm 1. Notably, in the medical domain, professors evaluate students across a spectrum of themes and real-time scenarios. These evaluations include the process of patient interaction, as well as the assessment of symptoms and the subsequent course of action, as referenced in [45,46]. Here, we systematize and generalize the process to reach to a measurable conclusion of LLMs answers, within a domain-specific context and a dialogue-defined space. Our contributions

in the generalization of the domain settings using embeddings can be seen in Figure 6 within the yellow cards.

Each component that defines the answer is represented by the appropriate letter in Table 1 and Figure 6. In a range of (n), a final score would finalize a decision as per the quality of the answer, which in essence represents the ability (A) of the LLM to comprehend the answer and also the effectiveness of the domain setup (Figure 6). In more detail, the items are as follows.

- Question q.
- Domain d.
- Medical m.
- Answer a.
- Answer content c: Derived from the 'instantiated advice generation' (2) and the 'blueprint of domain-specific knowledge' (1).
- Retrieve documents r: Sourced from (1) and (2) to show basis of produced answer.
- Process I (ability assessment): To define ability, we compare the answer produced by the LLM (a) to the ground truth (i.e., the correct answer) using similarity checks and pairwise embedding distance algorithms. This involves retrieving the documents (r) on which the answer was based.
- Process III (reflective capacity): The reflective capacity is calculated by applying similarity checks and pairwise embedding distance algorithms between a composite of the question (q) and answer type (a) and the answer content (c) produced by the LLM.

Algorithm 1 Evaluation process based on rules.

Require: Question (q), Domain (d), Medical (m)
Ensure: Answer (a), Grade of Answer Content (c)

Begin
if (q == "Informational" OR q == "Open-Ended") AND (d == "Yes") AND (m == "Yes") **then**
    $a \leftarrow$ "Direct or Elaborated"
else if (q == "Any") AND (d == "No") **then**
    if (m == "No") **then**
        $a \leftarrow$ "(deferred) Dialogue Reset"
    else if (m == "Yes") **then**
        $a \leftarrow$ "(deferred) Refer To Specialist"
    end if
end if
$c \leftarrow$ AnswerContent
$r \leftarrow$ RetrieveDocuments
$grade(Ab) \leftarrow$ Define Ability(c)
$grade(Re) \leftarrow$ Reflect(c)
return $a, grade$
End
function DEFINEABILITY(AnswerContent)
    // A grading logic for Answer content(1)
    // Return grade(Ab)
end function
function REFLECT(AnswerContent,documents)
    // A reflection(2) logic for answer content and document retrieval(3)
    // Return grade(Re)
end function

For Process I (Define Ability),

$$\text{Ability} = \text{Similarity}(\text{LLM Answer}, \text{Ground Truth})$$

where

$$\text{LLM Answer} = \text{Function}(a, c, r)$$

$$\text{Ground Truth} = \text{Known Correct Answer}$$

For Process II (Reflect -reflective capacity),

$$\text{Reflective Capacity} = \text{Similarity}((q + a), c)$$

Here, "Similarity" represents the similarity check and pairwise embedding distance algorithms, and "Function" is the method by which the LLM produces its answer based on the answer type (a), answer content (c), and retrieved documents (r).

The main assumption made for the creation of embeddings in GAI, instead of a bottom-up approach when fine tuning or recreating a model, is based on the fact that the trained LLMs used in systems like Bard (google) and ChatGPT (openAI) are tested and evaluated in numerous tasks. Moreover, the magnitude of their training datasets is such that recreating a similar model would incur additional costs and evaluation procedures. While the fine tuning is a more straightforward strategy than recreating one, again, there are significant costs and similar evaluation requirements.

Retrieval-augmented generation (RAG) is particularly effective for general tasks as it combines the benefits of a large language model with external data sources, enhancing the breadth and specificity of its responses. This approach is suitable for a wide range of applications where expert involvement is not critical and the focus is on augmenting the generative capabilities with a diverse set of information sources.

On the other hand, GAI, particularly in sensitive areas, requires a more nuanced approach. In scenarios such as healthcare, legal advice, or personalized recommendations, the requirements are higher. Therefore, employing GAI in these domains demands rigorous testing, diligent constraint implementation, and continuous monitoring to ensure safety, accuracy, and ethical compliance. The involvement of domain experts becomes crucial for the validation of the outputs and provision of guidance on the model's usage boundaries. This ensures that the generated responses and decisions are not merely based on data and algorithms but are also aligned with human expertise and ethical standards.

The creation of embeddings offers a rapid method of incorporating essential context into pre-trained LLMs, which becomes particularly effective when these models are employed in specific applications. Additionally, we ensure that all supplementary data utilized by the model are provided by medical experts and align with the guidelines set forth by the relevant medical boards. The need for model fine tuning is determined based on the outcomes observed. If necessary, this fine tuning can occur later in the release and production pipeline, after a thorough evaluation tailored to the specific domain requirements and specifications. When the embeddings are input into a trained Transformer model, particularly for conversational purposes, the model utilizes the weights (as shown in Figure 3) that were acquired during its training phase. This process enables the model to more swiftly adapt to a predetermined conversational context [47].

The process of creating and training embeddings typically involves several key steps, outlined below, to ensure reproducibility.

1. Creation of embeddings
    (a) Training: Embeddings are usually created through supervised or unsupervised learning on large text corpora. At this stage, the model's trained weights are used to create vectors.
    (b) Dimensionality: The vectors usually have hundreds of dimensions. Dimensionality reduction techniques (like PCA or t-SNE) can be applied for visualization.
    (c) Contextualization: Traditional embeddings (Word2Vec, GloVe) do not consider the context, meaning that they represent a word with the same vector regardless of its usage. Modern embeddings (BERT, GPT) are contextual, adjusting the representation based on the word's usage in a sentence.

(d) Transfer Learning: Pre-trained embeddings can be fine-tuned on a smaller dataset for specific tasks, leveraging the general language understanding learned during pre-training while adapting to the nuances of the task at hand.
(e) Evaluation: The quality of embeddings is usually evaluated based on their performance in downstream NLP tasks like text classification, sentiment analysis, or named entity recognition.

Using the created embeddings, the following processes are the splitting, chunking, and storage of the information in vector databases, which would either define the instantiated advice generation or the blueprint of domain-specific knowledge.

2. Vector stores are databases that specialize in storing, indexing, and querying high-dimensional vectors. These vectors can represent various types of data, such as images, text, or other complex data types, transformed into numerical representations [48]. They are extremely useful for a similarity search, which, in Figure 6, is the red rectangle named search, pointing to the vector database.
3. A search is the process of retrieving documents stored in the vector store, either from the instantiated advice generation or the blueprint of domain-specific knowledge, based on a similarity threshold, manually defined. The higher the similarity threshold, the more restrictive the rules; thus, less documents are returned for processing.

In the Q&A chain, the aforementioned process is outlined as a generic algorithm in Algorithm 2.

Algorithm 2 Q&A Chain

$DB, Embeddings \leftarrow$ Vector.db() (Vector Database Settings)
$Retriever, LLM \leftarrow$ chainer$(DB, Embeddings)$ (LLM engine properties)
procedure RunQAChain$(Query : QueryModel, CurrentUser)$ (*Function)
    $Question \leftarrow Query.question$
    $QAChain \leftarrow$ prompter$(Retriever, LLM)$ (Retriever== Similarity parameters and search depth, LLM== llm engine parameters)
    $Result \leftarrow$ responder$(QAChain, Question)$
    return {"response" : $Result$}
end procedure

Step 1 of the algorithm involves initializing or loading a vector database. The database contains embeddings, which are high-dimensional vectors of keys and values, representing the documents and instructions provided by the user. Step 2 involves the binding of components to process queries. This involves setting up an embedding filter, i.e., similarity search parameters; a retriever, where the properties to be retrieved from the DB are set; and an LLM. The exact roles of these components depend on the specific implementation and use case. In Step 3, the procedure is initialized, based on a received question that is posed by a specific user. Lastly, a response is returned from the model as a result, usually in a json or xml format.

4.3. System Architecture

The system architecture is outlined in Figures 11 and 12. In the micro-level process diagram, the internal design and communication paths of the different modules are analyzed. In the macro-level diagram, the application's overall structure is detailed.

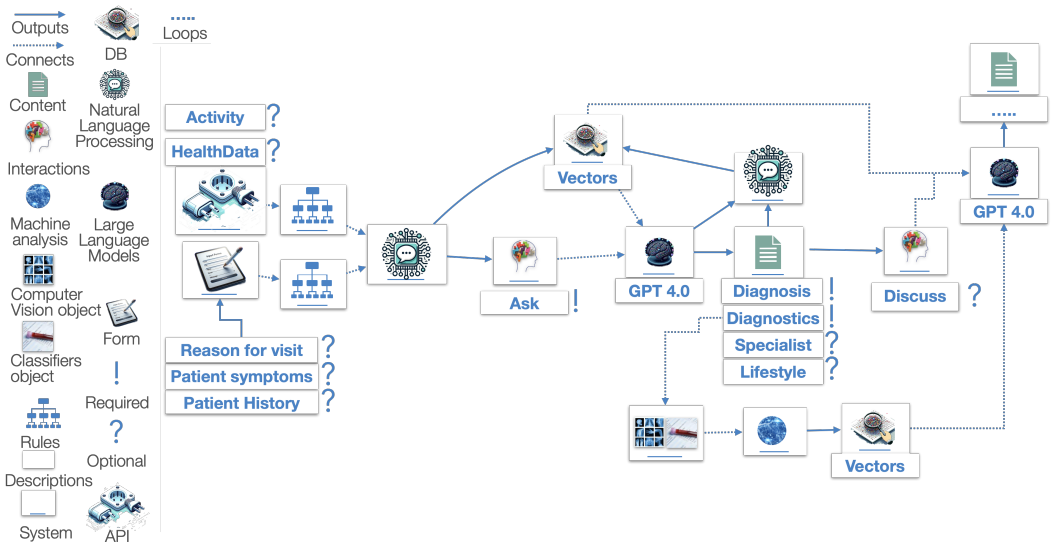

Figure 11. Process diagram—micro level.

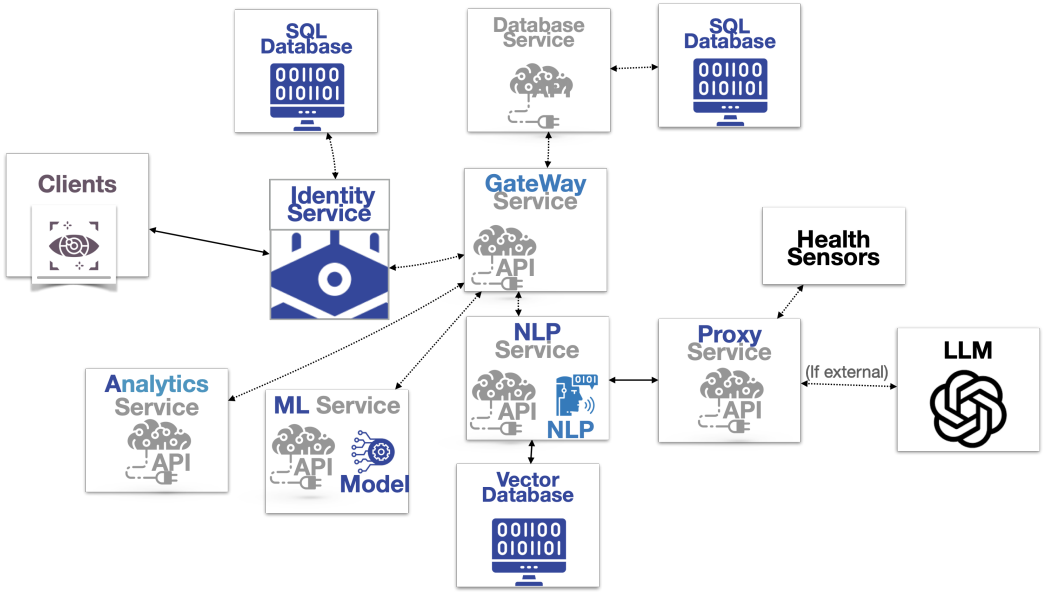

Figure 12. System architecture—macro level.

4.3.1. Micro Level

In the process diagram, we introduce a novel methodology to define GAI processes and functions, in the scope of a rule-augmented approach. On the left side, the different components are noted, and, on the right side, these components are implemented and we describe the information flow, the constraints, and the expected outcomes. In detail, the items are as follows.

- Outputs: Information exchange.
- Connects: Implies dependency.
- DB: Type of database procedure.
- Content: Generation of data.
- Interactions: User interaction, which leads to a generation.
- Natural Language Processing: Any NLP process.
- Large Language Model: LLM processing or LLM API call.
- Computer Vision Object: ML process related to computer vision.
- Classifier Object: ML process related to classification.
- Machine Analysis: ML service output in textual or numeric format.

- Rules: Rules that are used to augment the system and limit malfunctions.
- White Box: Description space for processes.
- White Box with Blue Line: Defines a function or system.
- Form: A type of user input form, in a predefined context, i.e., using questionnaires or pre-selected inputs.
- !: Required described process.
- ?: Optional described process.
- API: External communication process.
- Three Dots (. . .): Indicates a loop or a repetitive–iterative process.

The process starts with the input of the optional definition of reasons for the (doctor) visit or/and patient symptoms and an optional patient history. In parallel, the user can optionally upload or connect activity and health data via an API. An example of the data format extracted via the API can be seen in the Garmin Health snapshot schema. Based on specific rules, the data are then fed into the NLP system, which outputs an engineered prompt and an embedding to be saved in the Vectors DB system. The engineered prompt is the Ask Required interaction, which, alongside the Vectors DB, provides the necessary information for the LLM system (powered here by GPT-4), to provide a required diagnosis, the required diagnostics, and an optional referral to a specialist or an optional lifestyle intervention strategy. For the diagnostics, the user is also provided with computer vision systems and classifiers, for examination analysis and the transposition of data into usable objects. These objects are again saved into embedding objects in the Vectors DB system. The Discuss output is optional and can lead to a Q&A, as previously described in the Theoretical Dialogue section.

4.3.2. Macro Level

In Figure 12, the microservice architecture that can optimally support the rule-augmented AI application is detailed. This approach encompasses the overall structure of the entire software system, including how the different modules and components interact. In this specific application, which enables a range of analytical services and external APIs, a microservice architecture allows each service to be deployed using the most appropriate infrastructure. At the same time, the independent testing of each AI service or module facilitates an easier-to-manage workflow.

Especially when dealing with Python AI libraries, one common issue is the varying dependencies and requirements that they may have. These libraries often rely on specific versions of other libraries, which can lead to compatibility issues when multiple AI services are bundled together in a monolithic application.

The different components of the system architecture are as follows.

- Client: The user interface that provides the main space for the patients to interact.
- Identity Service: The authorization and authentication infrastructure that validates the client based on user profiles stored in the first SQL database.
- Gateway Service: Acts as an intermediary that processes and routes requests from clients to various services within a system.
  - Database service.
  - Analytics service: the component that is dedicated to analyzing data and generating insights, related to health metrics and diagnostic examinations.
  - ML service: a packaged component that provides machine learning capabilities to the system. This module can be integrated into existing systems to add features like prediction, classification, and anomaly detection based on extreme value theory.
  - NLP Service: the software module that encompasses the required processing features.
    - Vector database: type of database designed specifically to handle vectors, and, more specifically, in this case, the transposition to incoming embedding data.

- Proxy service: intermediary for requests from other APIs seeking resources from other servers. In this case, it handles communication with external health sensors and LLM engines.
  - Health sensors: APIs provided by wearable manufacturers like Fitbit or Apple HealthKit or medical devices used in clinical settings.
  - LLMs: APIs of powerful natural language processing engines for question and answering.

4.4. Services Simulating a General Practitioner

In this section, a brief description is provided for some key services that can facilitate the user when providing him/her with detailed diagnostic outcomes. These outcomes are to be processed using NLP techniques and provided back into the system, for the diagnosis pipeline to complete via one of the possible outcomes (diagnosis, diagnostics, specialist referral, lifestyle suggestion). It should also be stated that the services aim to simulate the available tools and competencies of a general practitioner. In this study, the focus is the simulation and automation of a physical examination using available technologies and incorporating them into an intuitive, fast, and simple-to-use rule-augmented system.

A blood examination is considered, alongside a view of a user's routine and a daily snapshot of the user's basic bio-metrics.

4.4.1. Blood Exam Analyzer

The blood exam analyzer consists of tools that extract information from the relevant examinations provided by the user for the construction of the second prompt. As part of this use case, the blood test analyzer and the blood exam classifier are utilized. This particular technology consists of a specific conditional logic that extracts those blood variables that are outside the normal ranges. Our system also consists of machine learning algorithms [4,49] that can identify similarities with specific weight groups, based on blood exams, and thus recognize other possible health states related to an unbalanced biochemical profile. Metabolic syndrome is also identified through a similar process [4,10].

4.4.2. External Sensors

Connecting with external sensors, the doctor, or, in this case, the AI system, an approximation to a physical examination can be achieved. Various health data can be extracted by health sensors with great accuracy, such as the following.

- Heart Rate: Continuous heart rate monitoring, including resting heart rate and abnormal heart rate alerts.
- Sleep: Tracks sleep patterns, including sleep stages (light, deep, REM) and sleep quality.
- Stress: Measures stress levels throughout the day.
- Steps and Floors Climbed: Tracks daily step count and floors climbed using an altimeter.
- Calories Burned: Estimates calories burned through various activities.
- Intensity Minutes: Tracks vigorous activity minutes as per health recommendations.
- Body Battery: Monitors body energy levels to suggest the best times for activity and rest.
- Pulse Oximetry: Measures blood oxygen saturation, which can be essential at high altitudes or for tracking sleep issues.
- Respiration Rate: Monitors breathing rate throughout the day and night.
- Women's Health: Tracks menstrual cycle or pregnancy.
- VO2 Max: Estimates the maximum volume of oxygen that can be utilized during intense exercise.
- GPS Tracking: Offers detailed tracking for outdoor activities, including pace, distance, and routes.
- Activity Profiles: Multiple sports profiles for tracking different activities like running, swimming, cycling, golfing, and more.

- Incident Detection: Some models offer incident detection during certain activities, which can send one's location to emergency contacts if a fall is detected.
- Mobility Metrics: Monitors how fast one walks, the timing of each step, and how often one stands up.

In the Garmin Health snapshot (see Listing 1), a json object, as an example of a useful retrieved health metric, is provided. In the external sensor algorithm, a summary of the processes of extraction and conversion of these data is provided, for replication purposes.

Listing 1. Garmin Health snapshot. Extracted and transformed into json file.

```
{'calendarDate': '2023-10-23',
    'minHeartRate': 50,
        'maxHeartRate': 131,
            'includesActivityData': True,
                    'restingHeartRate': 61,
                'averageStressLevel': 42,
            'bodyBatteryMostRecentValue': 18,
        'highestRespirationValue': 17.0,
    'lowestRespirationValue': 12.0,
'latestRespirationValue': 14.0}
```

Algorithm External Sensors

get_weekly_data: Collects data for the past 7 days using a provided data retrieval function.
daily_snapshot: Collects last data using a data retrieval function.

Pseudocode

```
function get_weekly_data(SensorData):
    start_date <- today - 7 days
    weekly_data <- empty~list

    for i in 0 to 6:
        current_Date <- start_date - i days
        data <- SensorData(current_Date)
        append data to~weekly_data

    return~weekly_data

function daily_snapshot(healthSensor, anarray):
    extracted_data <- empty~dictionary

    for each key in anarray:
        extracted_data[key] <- healthSensor.get(key, None)
    return extracted_data
```

4.5. Prototype—Use Cases

In this section, we present a series of screenshots, where different conversational use cases are considered. The general Med | Primary AI assistant is constructed following the methodology discussed in the previous section, where a blueprint of domain-specific knowledge has been constructed using embeddings. Moreover, instantiated advice generation is provided, in the form of embeddings, where the user can upload specific examination data using the provided services (Figure 12).

4.5.1. Use Case 1

In Figure 7, a user requests, in an open-ended question, the ways in which the assistant can provide help. The answer provided outlines the blueprint of domain-specific knowledge, which is analyzed in Figure 4. The tools, objectives, and tasks are described and returned as a direct answer.

4.5.2. Use Case 2

In the second use case (Figure 8), a user requests a consultation based on the described symptoms—an informational question. The system provides an elaborated answer, where some initial consultation is provided. A recommendation for a physical examination is also suggested and a provision for potential lifestyle changes and the use of products. This is an elaborated answer where the discussion can continue for a more definite diagnosis to be acquired.

4.5.3. Use Case 3

In this use case, shown in Figure 9, we show an example of a deferred answer, where the user is referred to a specialist (Table 1). Here, an open-ended question is provided that is within the medical domain but outside the blueprint of domain-specific knowledge. Thus, the system suggests a medical specialist, an oncologist, to better assess the related query.

4.5.4. Use Case 4

In this final use case, shown in Figure 10, we present the ways in which the data retrieved from external sensors are utilized. As already discussed in the previous sections, the data are transposed into embeddings and then analyzed, if necessary, by the system. The user provides some symptoms (information question –> elaborated answer) and states that data have been uploaded. Moreover, a summary of the data is requested. This descriptive prompt is designed in such a way as to best outline the system's capabilities. In a real-world scenario, since the blueprint of domain-specific knowledge is already created, the process would be more intuitive and only the symptoms would be required. The AI assistant would assess these symptoms and analyze the health data if necessary to provide a response.

5. System Evaluation

To effectively assess our system, we have utilized a selection of multiple-choice quiz questions sourced from 'The Internet Pathology Laboratory for Medical Education', an esteemed resource hosted by the University of Utah's Eccles Health Sciences Library [50]. These quizzes are meticulously designed to cater to students and professionals in healthcare sciences, with a particular focus on pathology. This selection is aligned with the specific educational needs and curricular requirements of medical students of pathology and practitioners.

More specifically, our system, which leverages the advanced capabilities of the GPT-4 model, has been tested across three thematic pillars of general pathology. These pillars encompass a comprehensive range of topics critical to the field.

Atherosclerosis and Thrombosis: We explored 50 questions in this category, delving into the complexities of atherosclerotic diseases and thrombotic processes. This section aimed to evaluate the system's understanding of cardiovascular pathologies, their etiologies, and the intricate mechanisms underlying these conditions. Overall, 48 out of the total of 50 questions were correctly answered.

Cellular Injury: A set of 55 questions tested the system's grasp of cellular injury mechanisms. This included queries on cellular responses to stress, pathophysiological changes in cell injury, and the various stages and outcomes of such injuries, mirroring real-world scenarios encountered in medical practice. Overall, 50 out of the total of 55 questions were correctly answered.

Embryology: In this segment, 52 questions were presented, focusing on the developmental stages and anomalies of embryology. The system's performance in this area was crucial to ascertain its ability to handle complex developmental biology concepts and their implications in pathological states. Overall, 45 out of the total of 52 questions were correctly answered.

Nutrition: Lastly, a set of 40 questions pertaining to nutrition was used. These questions were designed to assess the system's understanding of nutritional science, its role in health and disease, and its integration into pathological conditions. Overall, 37 out of the total of 40 questions were correctly answered.

In the evaluation process, each question was carefully crafted to present a realistic medical scenario, encompassing a range of symptoms and conditions that were specific to different gender and age groups. We provided the totality of questions and choices and requested the correct choice as a response. This approach was intended to simulate real-world clinical challenges, thereby testing the system's ability to apply its knowledge in a practical, context-sensitive manner.

For example, consider the following question from the Cellular Injury quiz.

A 50-year-old woman with a history of unstable angina suffers an acute myocardial infarction. Thrombolytic therapy with a tissue plasminogen activator (tPA) is administered to restore the coronary blood flow. Despite this therapy, the extent of myocardial fiber injury may increase due to which of the following cellular abnormalities?

- *[A.] Cytoskeletal intermediate filament loss*
- *[B.] Decreased intracellular pH from anaerobic glycolysis*
- *[C.] Increased free radical formation*
- *[D.] Mitochondrial swelling*
- *[E.] Nuclear chromatin clumping*
- *[F.] Reduced protein synthesis*

This question exemplifies the complexity and depth of the quizzes. It not only tests the system's grasp of specific medical knowledge but also its ability to analyze and apply this knowledge in diagnosing and understanding the progression of a disease. The inclusion of multiple answer choices, ranging from four to five options per question, further enhances the challenge, requiring the system to discern the most appropriate response from several plausible alternatives.

Such questions are integral to evaluating the system's proficiency in medical reasoning, particularly in pathology and related healthcare fields. They are designed not only to test the recall of factual information but also to assess the system's understanding of intricate physiological processes and its ability to make informed clinical decisions. This holistic approach ensures a thorough assessment of the system's capabilities in handling complex medical scenarios.

This comprehensive testing approach not only gauges the system's proficiency in handling specific medical knowledge but also its ability to integrate and apply this knowledge in a way that is coherent and contextually relevant to the field of pathology. The GPT-4 model had total precision of 91.37%, answering correctly 180 out of 197 questions. Although the system demonstrates a high success rate, further evaluation by medical experts and extensive testing across diverse scenarios are essential. Generally, systems empowered by LLMs like GPT-4, as used in this research, should be considered and treated as assistants and not replacements for human experts, particularly in sensitive domains such as the medical field, considering the critical impact of decision making in such disciplines.

6. Discussion of Results and Future Research

In this paper, the application of AI and particularly LLMs and NLP in healthcare is explored. A novel AI-empowered system is introduced, which is enhanced with rule-based algorithms and incorporates GPT models and other ML tools, to provide diagnostic advice. This system is tailored to address the complexities of healthcare experiences, specifically from a general practitioner's perspective. The research is organized into various

sections, covering theoretical foundations, system design and implementation, and practical use cases.

A key contribution of this work is the creation of a blueprint of domain-specific knowledge, serving as a contextual foundation for an AI system augmented with LLMs and rule-based logic. By generalizing the process, a measurable conclusion can be reached on the quality of the LLM's answers within a domain-specific context and a dialogue-defined space. These rules are formulated from a dialogue theory perspective, ensuring meaningful and relevant interactions. The system design is innovatively constructed and presented in a cost-effective manner, emphasizing reproducibility and scalability. The proposed AI-empowered, rule-augmented healthcare application integrates rules, external APIs, and modern methodologies to utilize current LLMs efficiently. This forms the basis for innovative approaches in the medical domain. Finally, the GPT-4-empowered system has undergone comprehensive evaluation in the field of general pathology, achieving a 91.37% accuracy rate in a set of 197 multiple-choice questions.

For future research and development, two critical areas are highlighted.

- Cost Analysis: Understanding the financial implications of deploying and using this AI system in healthcare is vital. This involves assessing the initial setup costs, ongoing operational expenses, and the potential financial benefits or savings that it might bring to healthcare providers and patients. This analysis will help to determine the economic feasibility and scalability of the system.
- Value-Based Care: This aspect focuses on comparing the costs and outcomes of care provided by different healthcare providers, considering both automated systems like the one proposed and traditional care methods. Key elements include the following.
  - Evaluating Effectiveness of Interventions: This involves measuring the impact of healthcare interventions on patient outcomes such as mortality rates, morbidity rates, and improvements in health-related quality of life. The AI system's role in facilitating timely interventions and improving these outcomes needs to be examined.
  - Patient Perspectives on Effectiveness: Assessing the value of care from the patient's point of view is crucial. This involves gathering and analyzing patient feedback to understand their experiences and satisfaction with the care provided, both through traditional means and the AI system.

These areas emphasize the need to balance technological advancement with practical, patient-centered care. Future research should also focus on ethical considerations, data privacy, and the integration of AI systems with existing healthcare infrastructures. The ultimate goal is to enhance healthcare delivery while ensuring that it is accessible, affordable, and aligned with patient needs and values.

Author Contributions: Conceptualization, D.P.P.; software, D.P.P.; validation, M.V. and G.A.T.; writing—original draft, D.P.P., M.V. and G.A.T.; writing—review and editing, M.V. and G.A.T.; visualization, D.P.P.; supervision M.V. and G.A.T. 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: Data are contained within the article.

Acknowledgments: This work has been partly supported by the University of Piraeus Research Center. Theoretical/medical support and technical/medical advice as per the validity of our hypothesis was provided on 30 October 2023 by the medical doctors of Dermacen S.A. https://www.dermatologikokentro.gr (accessed: 30 October 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
GAI	Generative Artificial Intelligence
LLM	Large Language Model
NLP	Natural Language Processing
EHR	Electronic Health Records
HMM	Hidden Markov Models
CRF	Conditional Random Fields
LDA	Latent Dirichlet Allocation
RAG	Retrieval-Augmented Generation

References

1. Trebble, T.M.; Hansi, N.; Hydes, T.; Smith, M.A.; Baker, M. Process mapping the patient journey: An introduction. *BMJ* **2010**, *341*, c4078. [CrossRef]
2. Gualandi, R.; Masella, C.; Viglione, D.; Tartaglini, D. Exploring the hospital patient journey: What does the patient experience? *PLoS ONE* **2019**, *14*, e0224899. [CrossRef] [PubMed]
3. McCarthy, S.; O'Raghallaigh, P.; Woodworth, S.; Lim, Y.L.; Kenny, L.C.; Adam, F. An integrated patient journey mapping tool for embedding quality in healthcare service reform. *J. Decis. Syst.* **2016**, *25*, 354–368. [CrossRef]
4. Panagoulias, D.P.; Virvou, M.; Tsihrintzis, G.A. Nuhealthsoft: A Nutritional and Health Data Processing Software Tool from a patient's perspective. In Proceedings of the 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Dijon, France, 19–21 October 2022; pp. 386–393.
5. Balogh, E.P.; Miller, B.T.; Ball, J.R. *Improving Diagnosis in Health Care*; The National Academies Press: Washington, DC, USA, 2015.
6. Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* **2017**, *69*, 218–229. [CrossRef]
7. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef] [PubMed]
8. Xiao, C.; Choi, E.; Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1419–1428. [CrossRef] [PubMed]
9. Davenport, T.; Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **2019**, *6*, 94. [CrossRef] [PubMed]
10. Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. SVM-Based Blood Exam Classification for Predicting Defining Factors in Metabolic Syndrome Diagnosis. *Electronics* **2022**, *11*, 857. [CrossRef]
11. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774 .
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
14. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Stanford Alpaca: An Instruction-Following LLaMA Model. 2023. Available online: https://github.com/tatsu-lab/stanford_alpaca (accessed on 1 January 2024).
15. Panagoulias, D.; Palamidas, F.; Virvou, M.; Tsihrintzis, G.A. Evaluating the potential of LLMs and ChatGPT on medical diagnosis and treatment. In Proceedings of the 14th IEEE International Conference on Information, Intelligence, Systems, and Applications (IISA2023), Volos, Greece, 10–12 July 2023.
16. Gordon, E.B.; Towbin, A.J.; Wingrove, P.; Shafique, U.; Haas, B.; Kitts, A.B.; Feldman, J.; Furlan, A. Enhancing patient communication with Chat-GPT in radiology: Evaluating the efficacy and readability of answers to common imaging-related questions. *J. Am. Coll. Radiol.* **2023**. [CrossRef] [PubMed]
17. Floyd, W.; Kleber, T.; Pasli, M.; Qazi, J.; Huang, C.; Leng, J.; Ackerson, B.; Carpenter, D.; Salama, J.; Boyer, M. Evaluating the Reliability of Chat-GPT Model Responses for Radiation Oncology Patient Inquiries. *Int. J. Radiat. Oncol. Biol. Phys.* **2023**, *117*, e383. [CrossRef]
18. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D.; et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef]
19. Locke, S.; Bashall, A.; Al-Adely, S.; Moore, J.; Wilson, A.; Kitchen, G.B. Natural language processing in medicine: A review. *Trends Anaesth. Crit. Care* **2021**, *38*, 4–9. [CrossRef]

20. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [CrossRef]
21. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December, 2017; Volume *30*, pp. 5998–6008
23. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
24. Wang, B.; Wang, A.; Chen, F.; Wang, Y.; Kuo, C.C.J. Evaluating word embedding models: Methods and experimental results. *Apsipa Trans. Signal Inf. Process.* **2019**, *8*, e19. [CrossRef]
25. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-art speech recognition with sequence-to-sequence models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
26. OpenAI. *Better Language Models and Their Implications*; OpenAI: San Francisco, CA, USA, 2019.
27. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
28. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4349–4357.
29. Gunning, D.; Stefik, M; Choi, J; Miller, T; Stumpf, S; Yang, G; XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *37*, eaay7120. [CrossRef] [PubMed]
30. Holzinger, A.; Goebel, R.; Fong, R.; Moon, T.; Müller, K.R.; Samek, W. xxAI-beyond explainable artificial intelligence. In Proceedings of the xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, Vienna, Austria, 18 July 2020; Revised and Extended Papers; Springer: Berlin/Heidelberg, Germany, 2022; pp. 3–10.
31. Panagoulias, D.P.; Sarmas, E.; Marinakis, V.; Virvou, M.; Tsihrintzis, G.A.; Doukas, H. Intelligent Decision Support for Energy Management: A Methodology for Tailored Explainability of Artificial Intelligence Analytics. *Electronics* **2023**, *12*, 4430. [CrossRef]
32. Panagoulias, D.; Palamidas, F.; Virvou, M.; Tsihrintzis, G.A. Evaluation of ChatGPT-supported diagnosis, staging and treatment planning for the case of lung cancer. In Proceedings of the 20th ACS/IEEE International Conference on Computer Systems and Applications, AICSSA 2023, Giza, Egypt, 4–7 December 2023.
33. Blandford, A.; Wesson, J.; Amalberti, R.; AlHazme, R.; Allwihan, R. Opportunities and challenges for telehealth within, and beyond, a pandemic. *Lancet Glob. Health* **2020**, *8*, e1364–e1365. [CrossRef] [PubMed]
34. Snoswell, C.L.; Chelberg, G.; De Guzman, K.R.; Haydon, H.H.; Thomas, E.E.; Caffery, L.J.; Smith, A.C. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. *J. Telemed. Telecare* **2023**, *29*, 669–684. [CrossRef] [PubMed]
35. Kraft, A.D.; Quimbo, S.A.; Solon, O.; Shimkhada, R.; Florentino, J.; Peabody, J.W. The health and cost impact of care delay and the experimental impact of insurance on reducing delays. *J. Pediatr.* **2009**, *155*, 281–285. [CrossRef] [PubMed]
36. Martin, D.; Miller, A.P.; Quesnel-Vallée, A.; Caron, N.R.; Vissandjée, B.; Marchildon, G.P. Canada's universal health-care system: Achieving its potential. *Lancet* **2018**, *391*, 1718–1735. [CrossRef] [PubMed]
37. Goodair, B.; Reeves, A. Outsourcing health-care services to the private sector and treatable mortality rates in England, 2013–20: An observational study of NHS privatisation. *Lancet Public Health* **2022**, *7*, e638–e646. [CrossRef] [PubMed]
38. Yang, H.; Kim, S.; Park, J. Exploring avoidable, preventable, treatable mortality trends and effect factors by income level. *Eur. J. Public Health* **2023**, *33*, ckad160–1115. [CrossRef]
39. Treatable Mortality in Europe: Time Series. Available online: https://www.statista.com/statistics/1421315/treatable-mortality-in-europe-time-series (accessed on 18 December 2023).
40. NuhealtSoft Suite. Available online: https://www.diskinside.com/nuhealthsoft/ (accessed on 8 January 2024)
41. Panagoulias, D.P.; Virvou, M.; Tsihrintzis, G.A. Rule-Augmented Artificial Intelligence-empowered Systems for Medical Diagnosis using Large Language Models. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA, 6–8 November 2023.
42. Gorsky, P.; Caspi, A.; Chajut, E. The "theory of instructional dialogue": Toward a unified theory of instructional design. In *Understanding Online Instructional Modeling: Theories and Practices*; IGI Global: Hershey, PA, USA, 2008; pp. 47–69.
43. Wilson, D.C. Chapter Three: A Framework for Clarifying. In *A Guide to Good Reasoning: Cultivating Intellectual Virtues*; McGraw-Hill College: New York, NY, USA, 2020.
44. García-Carrión, R.; López de Aguileta, G.; Padrós, M.; Ramis-Salas, M. Implications for social impact of dialogic teaching and learning. *Front. Psychol.* **2020**, *11*, 140. [CrossRef]
45. Mitchell, M.L.; Henderson, A.; Groves, M.; Dalton, M.; Nulty, D. The objective structured clinical examination (OSCE): optimising its value in the undergraduate nursing curriculum. *Nurse Educ. Today* **2009**, *29*, 398–404. [CrossRef] [PubMed]
46. Majumder, M.A.A.; Kumar, A.; Krishnamurthy, K.; Ojeh, N.; Adams, O.P.; Sa, B. An evaluative study of objective structured clinical examination (OSCE): students and examiners perspectives. *Adv. Med Educ. Pract.* **2019**, *10*, 387–397. [CrossRef] [PubMed]

47. Customizing Conversational Memory. Available online: https://python.langchain.com/docs/modules/memory/conversational_customization (accessed on 29 September 2023).
48. Vector Stores-LlamaIndex. Available online: https://gpt-index.readthedocs.io/en/v0.7.8/core_modules/data_modules/storage/vector_stores.html (accessed on 20 November 2023).
49. Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. An Extreme Value Analysis-Based Systemic Approach in Healthcare Information Systems: The Case of Dietary Intake. *Electronics* **2023**, *12*, 204. [CrossRef]
50. The Internet Pathology Laboratory for Medical Education. Available online: https://webpath.med.utah.edu/webpath.html (accessed on 15 December 2023).