information

Article

Applied Hedge Algebra Approach with Multilingual Large Language Models to Extract Hidden Rules in Datasets for Improvement of Generative AI Applications

Hai Van Pham [1,*] and Philip Moore [2]

[1] School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi 10000, Vietnam
[2] School of Information Science and Engineering, Lanzhou University, Lanzhou 730030, China; ptmbcu@gmail.com
* Correspondence: haipv@soict.hust.edu.vn

Abstract: Generative AI applications have played an increasingly significant role in real-time tracking applications in many domains including, for example, healthcare, consultancy, dialog boxes (common types of window in a graphical user interface of operating systems), monitoring systems, and emergency response. This paper considers generative AI and presents an approach which combines hedge algebra and a multilingual large language model to find hidden rules in big data for ChatGPT. We present a novel method for extracting natural language knowledge from large datasets by leveraging fuzzy sets and hedge algebra to extract these rules, presented in meta data for ChatGPT and generative AI applications. The proposed model has been developed to minimize the computational and staff costs for medium-sized enterprises which are typically resource and time limited. The proposed model has been designed to automate question–response interactions for rules extracted from large data in a multiplicity of domains. The experimental results show that the proposed model performs well using datasets associated with specific domains in healthcare to validate the effectiveness of the proposed model. The ChatGPT application in case studies of healthcare is tested using datasets for English and Vietnamese languages. In comparative experimental testing, the proposed model outperformed the state of the art, achieving in the range of 96.70–97.50% performance using a heart dataset.

Keywords: generative AI; language comprehension; multilingual language models; large language models; support systems; technological determinism; chatbot; ChatGPT

check for updates

Citation: Pham, H.V.; Moore, P. Applied Hedge Algebra Approach with Multilingual Large Language Models to Extract Hidden Rules in Datasets for Improvement of Generative AI Applications. *Information* **2024**, *15*, 381. https://doi.org/10.3390/info15070381

Academic Editors: Junwen Duan, Fangfang Li and Tudor Groza

Received: 9 June 2024
Revised: 27 June 2024
Accepted: 27 June 2024
Published: 29 June 2024

1. Introduction

Generative artificial intelligence (hereafter termed GenAI) is a rapidly developing technology which has been employed in the development of ChatGPT by OpenAI (OpenAI: https://openai.com/ (accessed on 10 April 2024)). In a broad and diverse range of applications, GenAI plays a significant role in disruptive innovation (DI), where merging technologies can support smart applications [1]. In addition, GenAI has many societal, ethical, technological, and practical risks, as expressed in Section 2. GenAI models can accommodate multiple domains and the development of GenAI applications can be found in financial systems, computing systems, analysis, technological, and human resources [2–4].

In the realm of AI, while there are multiple GenAI systems (both open source and proprietary systems), a significant focus has been on ChatGPT, a domain stemming from natural language processing (NLP) [5–7]. The development trajectory of ChatGPT was primarily fueled by the objective to engineer an AI language model of high sophistication and versatility. This model is tailored for a spectrum of tasks encompassing text generation, language translation, and analysis of data. At the heart of ChatGPT's foundational technology is the Transformer architecture, a pivotal evolution in AI language processing initially introduced in Ref. [8]. This architecture was designed as a solution to the limitations inherent in previous NLP models, specifically recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Many applications using large language models (LLMs) consider reasoning mechanisms in LLMs combined with ChatGPT for responses [9,10]. An integration of GenAI and LLMs can enable personalized service provision and decision making using engaging technologies in dynamic virtual environments which adapt and respond to users' actions.

A goal of GenAI is to enhance interactions between a chatbot and an LLM(s) in a multiplicity of domains and systems to enable the creation of content including media, images, video, text, and audio. It supports innovative automated interactions in GenAI, NLP, image processing, and computer vision [11]. GenAI provides novel approaches for creating content by filling gaps in the development of the 'metaverse'. Furthermore, LLM(s) and ChatGPT can enhance their responses as they relate to knowledge experience and information generation.

However, a recognized limitation lies in the difficulty in dealing with hidden rules in large datasets and the resulting responses by using a chatbot. In real-world applications, extracting information from large datasets using GenAI systems results in high computational cost and significant hardware and staff resources, as noted above; while large organizations have the resources to implement GenAI, SMEs generally lack the required resources.

In this paper, we present a novel model (hereafter termed *GenAI-Algebra*) which utilizes a combination of hedge algebra approaches and LLM(s) to find hidden rules in large datasets by incorporating the GenAI of ChatGPT. The *GenAI-Algebra:*

- Extracts natural language knowledge from large datasets by leveraging fuzzy rules quantified by hedge algebra.
- Has been designed to extract hidden rules in large datasets with automated question–response interactions in a broad and diverse range of domains and systems.
- Has been developed for resource-limited SME(s).
- In a case study in the medical domain predicated on the human heart (based on the UCI datasets to evaluate the effectiveness of the proposed model), the reported experimental results validate the effectiveness of the proposed model.

Our contributions may be summarized as follows:

- Our *GenAI-Algebra* method can adapt to a multiplicity of domains in both Vietnamese and English. In the case study, *GenAI-Algebra* generates a comprehensive list of potential heart disease diagnoses based on a patient's reported symptoms and medical history by analyzing the patient's information using rules drawn from medical knowledge.
- The customization and fine-tuning of ChatGPT integrated with knowledge bases allows the identification of hidden fuzzy rules quantified by hedge algebra in large datasets.
- Our *GenAI-Algebra* method provides an effective basis upon which the simulation of real-time/real-world interactions [in both English and Vietnamese] can be realised.
- The *GenAI-Algebra* method contributes to symptom analysis, supports differential diagnosis, collects real-time data, and enhances decision-support for clinicians.
- Furthermore, the proposed *GenAI-Algebra* method and ChatGPT can play a valuable role in early detection by extracting relevant historical patient data and prognoses from large datasets; this can ultimately lead to improved patient policy outcomes.
- The *GenAI-Algebra* model is trained by using 'low-rank adaptation' (LoRA) together with 'DeepSpeed' and mass datasets, which results in low computational overhead with reductions in inference time and cost that can lead to enhanced data protection and safety.
- This research aims to address the problem by creating a GenAI model for a chatbot complete with an LLM [12,13] in both the Vietnamese and English languages.

In experimental testing, the proposed *GenAI-Algebra* model achieves a significant performance improvement. In the case study, the proposed model is compared to existing chatbot models, achieving a 92% performance based on the English benchmark.

The remainder of this paper is structured as follows: The state of the art and related research are considered in Section 2 with the proposed *GenAI-Algebra* model introduced in Section 4. The experimental testing is introduced in Section 6. The results with an analysis are set out in Section 7. Section 8 presents a discussion along with open research questions and directions for future research. The paper closes with concluding observations in Section 9.

2. Related Research

In this section, we consider GenAI along with an overview of ChatGPT and LLM.

2.1. Application of GPT Generations

In this section we set out a a brief overview of the applications of GPT through its generations:

- GPT-1: Preliminary text generation; Simple question-answering tasks; language modeling; basic conversational abilities [14,15].
- GPT-2: Enhanced text generation with more coherent and contextually relevant outputs; content creation, such as articles, poetry, and stories; assisting in code writing; advanced conversational abilities; translation and summarization tasks, albeit not its primary design [16,17].
- GPT-3: Advanced and coherent text generation; drafting emails or other pieces of writing; code generation in various programming languages based on prompts; deeper and more contextual question-answering; creation of conversational agents; tutoring in a range of subjects; translation and summarization with improved accuracy; simulating characters for video games; designing and prototyping user interfaces based on textual descriptions [18,19].
- GPT-4: All the capabilities of GPT-3 but with enhanced accuracy, coherence, and depth; potential in more advanced tasks like research assistance; more nuanced conversational abilities; integration into more complex systems; potential applications in specialized fields like healthcare, finance, and other areas requiring expert knowledge [20,21].

Its innovative approach has been instrumental in the development of impactful language models, including the GPT series by OpenAI, such as GPT-2 and GPT-3, which are integral to the genesis of ChatGPT. The ChatGPT model is built on the GPT-3.5 architecture, a streamlined adaptation of OpenAI's 2020 GPT-3 model. This iteration, GPT-3.5, is a more compact version, containing 6.7 billion parameters in contrast to the 175 billion parameters of the original GPT-3 [22,23]. Despite its reduced parameter count, GPT-3.5 demonstrates impressive capabilities in various NLP tasks, including understanding language, generating text, and translating languages. ChatGPT, specifically trained on an extensive textual dataset, is finely tuned to craft conversational replies, adept at providing responses that closely resemble human interaction [24,25].

2.2. Generative Artificial Intelligence and Chatbots

GenAI has recently provided advanced methods capable of generating text, images, or other media, using generative models along with the development of many GenAI applications. However, GenAI models present issues and risks [26]. The swift progress in artificial intelligence (AI) and NLP has given rise to language models that are both sophisticated and adaptable [27,28]. GenAI encompasses AI models capable of producing new data by learning patterns and structures from pre-existing data. These models can generate diverse content, including text, images, music, and more, utilizing deep learning methods and neural networks [29,30]. Notably, ChatGPT (a creation of OpenAI) stands out as a versatile tool with a wide range of uses [31–33].

A chatbot called ChatGPT, which is a software application, typically utilizes GenAI and an LLM [34]. ChatGPT is a Transformer-based deep neural network integrated with LLM prompts as input in a smart system [35]. Applications of the chatbot use GenAI and LLMs for human–chatbot interactions [36]. While the aim of a chatbot is to mimic a human conversation, GenAI-driven chatbots' have demonstrated the capability to provide responses in applications for interactions in a variety of domains [26,37–39].

GenAI models can respond to either positive or negative aspects of GenAI and chatbots with a focus on ChatGPT. Investigations into GenAI have identified its disruptive nature, with open research questions identifying the need for ongoing research to fully understand the socio-technical impact of GenAI and an understanding of hidden data in mass datasets in order to respond to questions and provide answers in real time. Moreover, GenAI-driven chatbots can be designed with instructions, guidelines, and considerations [26,37] to:

- Consider sensitive information or information inappropriate to chatbots.
- Consider the safety and privacy of conversations of users.
- Create chatbots with GenAI adoption.

Chatbots have been considered in a range of applications and systems where future research into information systems design forms an important topic. It is an observation and the argument made in [40] to determine the achievements made in chatbots. While technologies in present-day AI are capable of applying GenAI in 'real-world applications' [41], studies have not focused on exploring data in large datasets together with LLM(s). In considering LLM(s), the generation techniques currently used to provide a response are predicated on human preference(s) employed by the LLMs. Human-preference datasets can be collected from rules or utilize public datasets. For fine-tuning of LLMs, these models can provide safer responses to better meet user requirements.

3. Preliminaries

In this section, we introduce pre-trained language models (Section 3.1), multimodal models (Section 3.2), hedge algebras for extracting rules in large datasets (Section 3.3), fuzzy sets (Section 3.4), the frame of cognition (Section 3.5), and linguistic variables (Section 3.6). The proposed approach, *GenAI-Algebra* model, is introduced in Section 4.

3.1. Pre-Trained Language Models

The Transformer architecture is a cornerstone in the development of cutting-edge models such as GPT-3 [42] and DALL-E-2 [43]. The Transformer architecture is designed to address the shortcomings of earlier models such as RNN models, particularly the handling of variable-length sequences and contextual understanding.

Predicated on the self-attention mechanism, the Transformer architecture empowers the model to process various segments of an input sequence in parallel. The Transformer comprises two main components: an encoder, that processes the input sequence into a set of representations; and a decoder, that translates these representations into an output sequence. Each layer within the encoder and decoder is composed of a multi-head attention mechanism alongside a feed-forward neural network. The multi-head attention, a pivotal element of the Transformer, assigns varying degrees of importance to different tokens, enhancing the model's capability to manage long-range dependencies, and thereby, bolstering its performance across numerous NLP tasks. The architecture's inherent parallelizability and its capacity to prioritize data-driven learning over inductive biases make it especially apt for large-scale pre-training, thus allowing Transformer-based models to excel in a multitude of downstream tasks [44].

The advent of the Transformer architecture has solidified its status as a preeminent framework in NLP, owing to its parallel processing and potent learning proficiencies. Transformer-based pre-trained language models are generally bifurcated into two categories depending on their training paradigms: autoregressive language modeling and masked language modeling [45]. Masked language modeling, exemplified by BERT [46] and its enhanced counterpart RoBERTa [47], entails predicting the likelihood of a hidden token given the surrounding context. BERT, a flagship model for this approach, undertakes masked language modeling and next-sentence prediction as its core tasks. RoBERTa builds on BERT's foundation, augmenting its performance by expanding the training dataset and introducing more rigorous pre-training challenges. XL-Net [48] extends the BERT premise, employing permutation strategies during training to diversify the order of token prediction, thereby enriching the model's contextual awareness. Autoregressive language models like GPT-3 [43] and OPT [48], in contrast, predict the subsequent token based on the sequence of preceding tokens which aligns them more with generative tasks.

The core concept driving pre-trained language models is the emulation of a "well-read" entity capable of comprehending language to perform any designated task within that linguistic framework (illustrated in Figure 1). Initially, the language model ingests a vast expanse of non-annotated data, such as the entirety of *Wikipedia*, to acquire a fundamental grasp of word usage and general language patterns. Subsequently, the model is specialized for a specific NLP task by fine-tuning it with a smaller, task-oriented dataset, culminating in a final model adept at executing the target task.

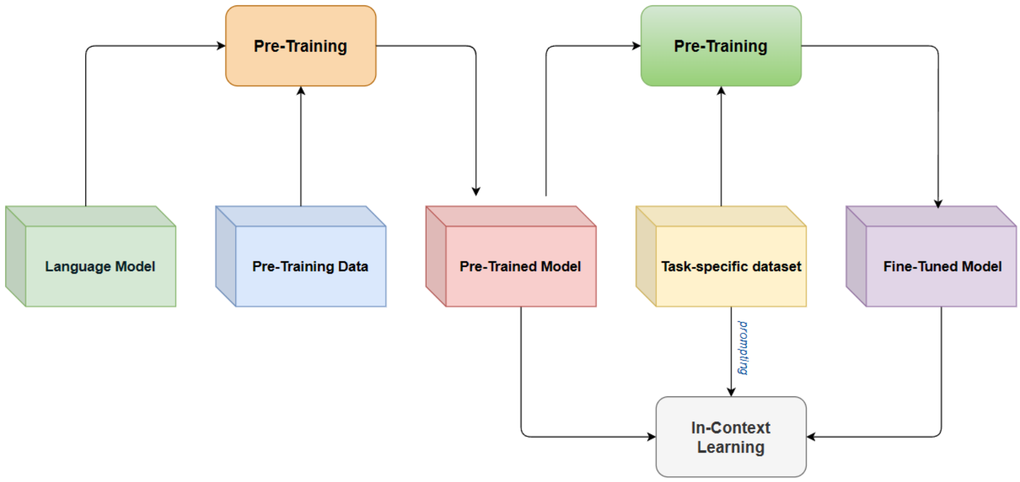


Figure 1. The taxonomy of pre-trained language models.

3.2. Multimodal Models

Multimodal generation has become a crucial aspect of modern AI-generated content models (AIGCs). The essence of multimodal generation lies in constructing models capable of generating raw modalities, such as images or sounds, by learning complex connections and interactions across different data types [21]. Multimodal interactions can be intricate, posing challenges to learning a shared representational space. However, the development of robust modality-specific foundational architectures has spawned methods to meet these challenges. We will explore state-of-the-art multimodal models in various domains including vision–language, text–audio, text–graph, and text–code generation, primarily focusing on their application in downstream tasks.

A multimodal architecture [exemplified by GPT-4] comprises an encoder for converting image and text inputs into vector representations, a decoder for generating text from these vectors, and an attention mechanism that enables both components to focus on pertinent elements of the inputs and outputs. The generation methods may be summarized as follows:

a Vision–language generation: Here, the encoder–decoder framework is extensively applied for uni-modal generation challenges in both computer vision and natural language processing. In vision–language multimodal generation, this architecture serves as a foundational structure. The encoder is tasked with learning a contextualized representation of the input, while the decoder is responsible for generating raw modalities that encapsulate cross-modal interactions and coherence.

b Text–audio generation: Here, text–audio multimodal processing has experienced significant advancements. Prevailing models typically concentrate on synthesis tasks like speech synthesis or recognition tasks such as automatic speech recognition, which involve translating written text to spoken speech or transcribing spoken words into machine-readable text, respectively. Text–audio generation is a distinct endeavor that entails crafting new audio or text utilizing multimodal approaches, differing from synthesis and recognition tasks in both objectives and methodologies.

c Text–graph generation: This mode holds substantial promise in enhancing NLP systems. Text, often laden with redundant information and lacking in logical structure, can be challenging for machines. Knowledge graphs (KG) offer a structured, organized representation of content, outlining semantic relationships within language processing systems. An increasing number of studies focus on deriving KGs from text to support text generation that encompasses complex concepts across multiple sentences. Semantic parsing is another facet of text–graph generation, aiming to convert text into logical forms like abstract meaning representations (AMRs) [49], which differ from KG by providing machine-interpretable representations. KG-to-text generation, conversely, generates coherent text based on pre-constructed KGs. Beyond NLP, text–graph generation is pushing the boundaries of computer-aided drug design, linking molecule graphs with descriptive language to aid molecular comprehension and discovery.

d Text–code generation: This mode seeks to automate the creation of valid programming code from natural language descriptions, providing coding assistance. LLMs have shown remarkable potential in generating programming language (PL) code from natural language (NL) descriptions. While early models treated text–code generation as a pure language task, the intrinsic modal differences between NL and PL necessitate strategies for capturing their mutual dependencies during semantic alignment. Text–code models must also handle PL(s) structural complexity and syntax, presenting additional challenges in semantic comprehension. These models also aim for multilingual support, enhancing their generalization capabilities.

Visual demonstrations (see [50]) illustrate the model processing images, responding to questions about them, extracting and interpreting text, captioning images, and engaging in visual IQ tests achieving accuracy in the range of 22–26%. Training requires each modality to be transmuted into a common embedding space representation, entailing sequences of vectors of uniform length derived from both text and images. Text processing is relatively straightforward due to its discrete nature, with each token obtaining an embedding during training that brings semantically similar words closer in the embedding space. For images, the MetaLM approach is employed, leveraging a pre-trained image encoder that feeds into a connector layer, aligning the image-derived embeddings with the text embedding dimension.

Overall, ChatGPT employs the Transformer architecture, which is key for state-of-the-art models like GPT-3. It uses a self-attention mechanism for better handling of long-term dependencies in NLP tasks. Two main types of pre-trained language models are used: autoregressive language modeling (like GPT-3) and masked language modeling (like BERT). ChatGPT also incorporates in-context learning and reinforcement learning from human feedback for improved performance.

3.3. Hedge Algebras for Extracting Rules in Large Datasets

Linguistic information involved in multi-criteria decision problems with a logic-based approximate reasoning method has been developed by Chen et al. [51] to provide decision-support based on information provided.

For human beings, language serves as a fundamental basis for cognition in the decision-making process; this process can be viewed as a consecutive series of decisions resulting in a final Boolean decision. The nature of decision making is to identify and select the optimal decision from a range of appropriate alternative options. As a consequence, in natural languages, human reasoning should incorporate linguistic [semantic] elements [words, phrases, adjectives, etc.] to describe alternatives based on a comparison between their properties [52].

In the algebraic approach, every linguistic domain can be interpreted as an algebra. For example, $AX = (X; G; H \leq)$ where $(X; \leq)$ is a poset, (G) is a set of the primary generators, and (H) is a set of unary operations representing linguistic hedges [52]. Values of the linguistic variable ***Truth*** may range, for example, from *True* through *VeryTrue*, *ProbablyFalse*, and *VeryProbablyFalse*, and so on. The values can be obtained from a set of generators (primary terms) such as $G = \{False, True\}$ using hedges from a set $H = \{Very, More, Probably, \ldots\}$ as unary operations.

3.4. Fuzzy Sets

This section provides a brief definition and characteristics of fuzzy sets; for a detailed exposition of set theory and 'real-world' practical examples, see [53,54]. Fuzzy set theory was proposed by Zadeh in 1965 in [55] with the notion of providing computerized systems with the capability to understand and process knowledge expressed in natural language. The membership function of an ordinary set can only take values in the range $[0, 1]$. Let (A) be the set of all points (objects) in a certain value domain or field, the fuzzy set (X) on the reference domain (A) is the set of all pairs $(a, E(a))$, where $(a \in A)$ and (E) are mappings, as in Equation (2):

$$[E : \rightarrow [0.1]] \tag{1}$$

The mapping (E) is called the membership function of the fuzzy set (X). The set (A) is called the base set of the fuzzy set (X). The value (E) represents the degree of membership of element a in the fuzzy set. The closer it is to (1), the higher the degree membership in (X).

When building fuzzy sets, the membership function value varies in the range $[0, 1]$. The degree of membership for common fuzzy sets is always highest in the middle and gradually reduces on both sides, which comes from the notion that this relationship represents phenomena in reality. There are always one or a few values with the highest membership in the fuzzy set (X). When these values increase or decrease past those thresholds, their membership in (X) will also decrease.

3.5. The Frame of Cognition

The frame of cognition (FoC) (F) of a linguistic variable (L) is a finite set of ordered fuzzy sets on the reference domain of the variable (L). These fuzzy sets are assigned a significance value of L. This value is called a term; the chosen term must be able to be used in expressing the meaning of (L). Therefore, the process of labeling necessitates a specific comprehension of the linguistic variable under consideration.

The terms used in (L) in the FoC must to be ordered based on their inherent semantics, for example, "young", "middle-aged", "old". Below are two graphical models (Figures 2 and 3) for two FOC linguistic variables, "heart rate" and "quantifier", where the attribute "heart rate" has an FoC consisting of five fuzzy sets corresponding to five terms: "very low", "low", "medium", "high", and "very high" (Figure 2a). Considering the As for the "quantifier" (Q value in LS), it has five fuzzy sets corresponding to five terms "non", "few", "a half", "many", and "almost" (Figure 2b).

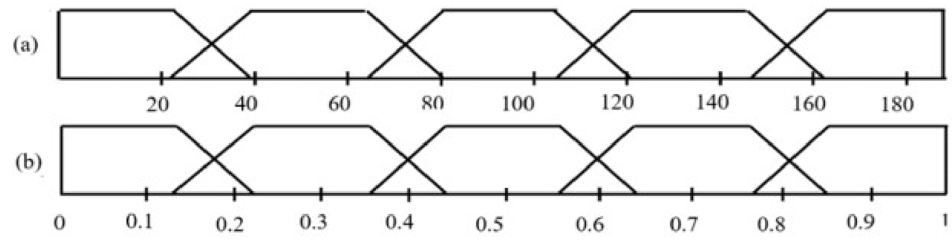


Figure 2. An example of fuzzy set mapping of sub figures (**a**) and (**b**) for a numerical reference domain.

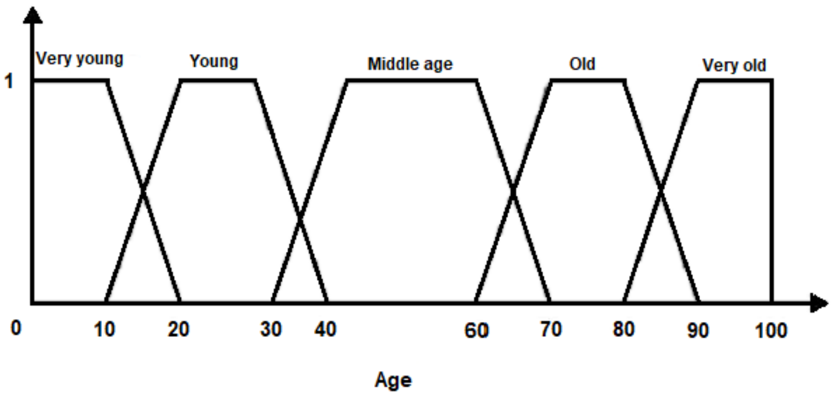


Figure 3. An example of five fuzzy sets semantically representing the linguistic values of the variable age in the reference domain [0, 100] (unit: age).

Based on these five fuzzy sets, we can see that every age value in the range [0, 10] belongs to the groups of very low, low, average, high, and very high to some extent. Suppose we have an age value of 2, we will have a membership value corresponding to each fuzzy set as $E_{veryLow} = 0.5$, $E_{low} = 0.5$, $E_{medium} = 0$, $E_{high} = 0$, and $E_{veryHigh} = 0$.

3.6. Linguistic Variables

Linguistic variables are variables whose values are words or sentences in natural or artificial languages. For example, when considering a person's age, we can consider this a linguistic variable called AGE and receive linguistic values such as 'very young', 'young', 'middle-aged', 'tall' , 'very high'. For each of these linguistic values, assign it a corresponding membership function that defines a fuzzy set on the domain of numeric values [0, 100] (age units) of the AGE attribute.

4. The Proposed *GenAI-Algebra* Model

In this section, we introduce our *GenAI-Algebra* model, consisting of proprietary data and user questions as inputs, outputs as answers, the vector database, and the submodel. The proposed model aims to create a multilingual chatbot with its GenAI for instant responses. An overview of the proposed system architecture is shown in the conceptual model in Figure 4.

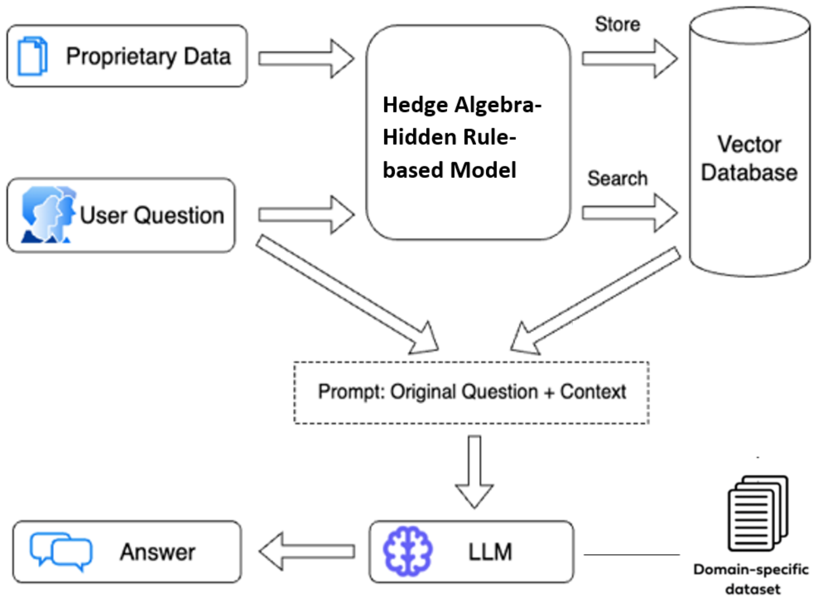


Figure 4. System architecture overview with data processing pipeline, model architecture, training process, and deployment.

The proposed *GenAI-Algebra* model can be applied to advance the diagnosis of heart disease and extract datasets by analyzing patient data to support doctors, leveraging its LLMs for responses in real time.

- Proprietary data: Datasets are preprocessed and parameters are adjusted to process these data based on rules in the submodel hedge algebra hidden rule-based model, including heart datasets in the mass datasets.
- User questions: Users can give questions and make requests from the proposed system, as well as interactive prompts, contexts, and original questions.
- Hedge algebra hidden rule-based model: The submodel is to execute hidden rules considered from fuzzy rules with hedge algebra into the vector database. These rules are also updated to the vector database, which responds to LLMs.
- Vector database: Prompts from questions and contexts of a domain can be requested from the database, which responds to LLMs.
- LLMs: Stanford University has provided an approach which utilizes a publicly accessible backbone called *LLaMA* [56] and fine-tunes it using BLOOM on their public website. The adaptability of BLOOM [57] to both English and Vietnamese allows the development of a multilingual chatbot that is capable of generating contextually relevant responses in both the English and Vietnamese languages.

To optimize hardware resources for model training, reducing the training time and costs, the proposed method allows organizations (including SMEs) to implement a chatbot adapted for both English and Vietnamese; the aim is the development of a multilingual chatbot capable of generating contextually relevant responses in both languages. The approach uses BLOOM [57] with optimization for the training process and efficiently utilizes GPU memory; the LoRA [58] with the DeepSpeed ZeRO-Offload [59] method are used to optimize parameters to enable hardware performance.

In the proposed model, the input to the model consists of instruction prompts which can be in the form of inputs for the chatbot to respond to, as given by Equation (2):

$$\mathcal{C} = (t_1,\ p_1)\ (t_2, p_2)\ \ldots\ (t_N,\ p_N) \tag{2}$$

where dataset C contains N samples, for example, i, and N is the number of *instruction–output* pairs. t_n is the nth instruction, and p_n is the output for the nth instruction.

To input texts of length L, the attention scores for the ith query $g_i \in R^{1xd}$, $(1 \le i \le L)$ in each head, given the first i keys $K \in R^{ixd}$, where d presents a head dimension, are given by Equation (3):

$$Sofmax(g_i K^T) \tag{3}$$

4.1. The Proposed LSmd Algorithm with Hedge Algebra

In this section, we set out the proposed LSmd algorithm with hedge algebra to identify the hidden rules. The proposed LSmd algorithm is to extract fuzzy rules in large datasets for heart disease. For a case study of a dialog of a doctor in healthcare, for heart disease patient database D, let $\{f1, f2, f3, ...fn\}$ be fields of database D, $d_i = \{f1_i, f2_i, ...fn_i\}$ be the ith record of D, and $fk_i(k \in [1, n])$ be the f_k field value of the record d_i.

Inputs: Attribute field f_k satisfies $(\forall d_i \in D, fk_i \in \mathbb{R})$, filter condition F: "fj = fil".

Outputs: LS sentences of the form "Q F y is/have S", truth value T of each LS sentence.

LSmd algorithm steps:

- Step 1: Choose the parameters for the hedge algebra architecture corresponding to the f_k attribute;
- Step 2: Generate a frame of cognition for the attribute f_k and the quantifier Q;
- Step 3: Calculate the average value of f_k corresponding to each label in the frame of cognition;
- Step 4: Calculate the truth value of the conclusion corresponding to each quantifier.

The proposed system architecture is described in detail in Figure 4; it aims to extract fuzzy rules in large datasets from heart disease, as shown in Figure 5.

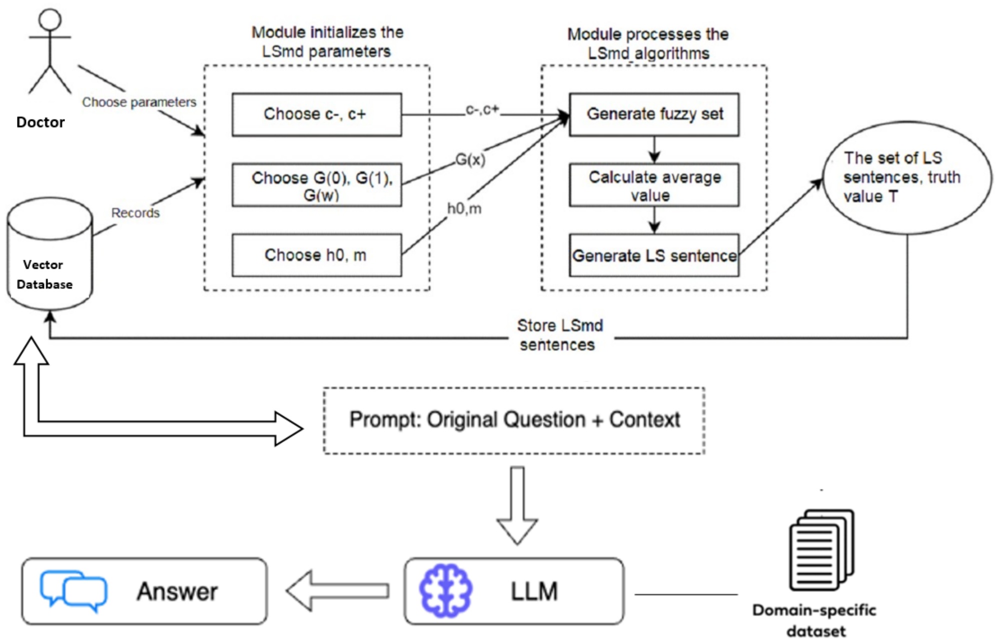


Figure 5. System architecture to extract information from heart disease database.

4.2. Proposed LSmd Algorithm

This section introduces the proposed LSmd algorithm in order to generate LS sentences of the form "Q F y is/have S", and the truth value T of each LS sentence, which is quantified from hidden rules in large datasets. These LS sentences will be updated to a vector database for LLMs of GenAI application.

Step 1: Select parameters for the HA architecture corresponding to f_k

Let c−, c+ be the negative and positive generating elements, respectively, F_0 is the basic level frame of cognition, "0" is the label with the smallest semantic value, "W" is the label with the average semantic value, "1" is the one with the greatest semantic value, H is the set of labels, h_0 is the measure of fuzziness of the average label, G_x is the fuzzy calculation range of the label x, and m is the calculation level.

Step 2: Generate a frame of cognition

Corresponding to the trapezoid of the fuzzy set representing the label x, we denote h_0x as the semantic core of x, $L_{bot}(x)$ as the left vertex ordinate of the big bottom, $R_{bot}(x)$ is the ordinate of the top right of the big bottom, $L_{top}(x)$ is the ordinate of the top left of the small bottom, $R_{top}(x)$ is the ordinate to the right of the small bottom, $S(x)$ is the ordinate interval between the two small bottom peaks. $Pre(x)$, $Pos(x)$ are the labels immediately before and after x in the ordinal set under consideration, respectively.

Call the m-level frame of cognition of fk F_m, for each label $x \in F_m$, the fuzzy set of labels x is denoted as A_x. To determine A_x, the four vertices of the trapezoid need to be determined: $A_x = \left\{R_{bot}(x), L_{bot}(x), R_{top}(x), L_{top}(x)\right\}$.

Step 3: Calculate the average value of f_k corresponding to each label as described in Algorithm 1

Let M_x be the average value of label x in the frame of cognition calculated over all records that satisfy the filter condition.

Algorithm 1: Calculating average value of term

Input: F_m, Set A = $\{A_x, \forall x \in F_m\}$, Set of records that have passed the filtering condition D_f
Output: Set M = $\{M_x, \forall x \in F_m\}$
foreach *x in* F_m **do**
  let M_x = 0;
  foreach *d in* D_f **do**
    M_x = M_x + $E_{A_x}(fk(d))$;
  end
  M_x = M_x / length(D_f;
end

Step 4: Calculate the truth value of the conclusion corresponding to each quantifier

Let LSs be the set of conclusion sentences, T(LS_i) is the truth value of the result sentence LS_i, Q is the frame of cognition of the quantifier, q is a label in Q, and E_q is the membership function of fuzzy set q.

Step 5: Indicate all results of the sentences are updated to the vector database, which interacts with prompts through LLMs

4.3. Case Study of Chatbot Dialog between Doctor and Heart Disease Patient

Thus, a set of LS sentences has been generated along with their corresponding truth values. To facilitate visualization, consider the following example. Given a set list of ages of 10 patients with heart disease as shown in Table 1.

Table 1. Ages of 10 patients.

No.	1	2	3	4	5	6	7	8	9	10
Age	52	53	70	61	62	58	58	55	46	54

In this example, the LSmd algorithm will be applied to the patient's "age" attribute.

Step 1: Select parameters for the traffic police architecture corresponding to the attribute "point"

The parameters are chosen as follows:

- Select "c−" = "low", "c+" = "high", "0" = "very low", "W" = "medium", "1" = "very high", $F_0 = \{0, c-, W, c+, 1\}$.

- Select fuzzy calculation intervals $G_x(x \in F_0)$: $G_0 = [0, 40], G_{c-} = [40, 55], G_W = [55, 60], G_{c+} = [60, 75], G_1 = [75, 100]$.
- Select the set of variables H = (L—little, V—very).
- Select the measure of fuzziness of the neutral hedge $h_0 = 1/3$.
- Select the calculation of level m = 0.

Step 2: Generate a frame of cognition

Applying the algorithm in step 2, we obtain the coordinates of the trapezoidal vertices of the fuzzy sets, as shown in Table 2

Table 2. Coordinates of 4 vertices of trapezoidal fuzzy set of terms.

Word Class	0	c−	W	c+	1
$L_{top}(x)$	0	45	55	65	75
$R_{top}(x)$	40	50	60	70	100
$L_{bot}(x)$	0	40	50	60	70
$R_{bot}(x)$	45	55	65	75	100

The following graph of fuzzy sets of the point perception framework will help us visualize more easily.

Step 3: Calculate the average value of "age" for each term

From the fuzzy sets of terms in F_0, we have a table of membership values of "age" corresponding to each patient as shown in Table 3.

Table 3. The membership of each age attribute to each term.

No.	1	2	3	4	5	6	7	8	9	10
Age	52	53	70	61	62	58	58	55	46	54
E_0	0	0	0	0	0	0	0	0	0	0
E_{c-}	0.6	0.4	0	0	0	0	0	0	1	0.2
E_W	0.4	0.6	0	0.8	0.6	1	1	1	0	0.8
E_{c+}	0	0	1	0.2	0.4	0	0	0	0	0
E_1	0	0	0	0	0	0	0	0	0	0

From Table 3, the average values can be calculated: $M_0 = 0/10 = 0, M_{c-} = 2.2/10 = 0.22, M_W = 6.2/10 = 0.62, M_{c+} = 1.6/10 = 0.16, M_1 = 0/10 = 0$.

Step 4: Calculate the truth value of the conclusion corresponding to each quantifier

We have the frame of cognition of the quantifier Q, as shown in Figures 6 and 7 (this is the default value):

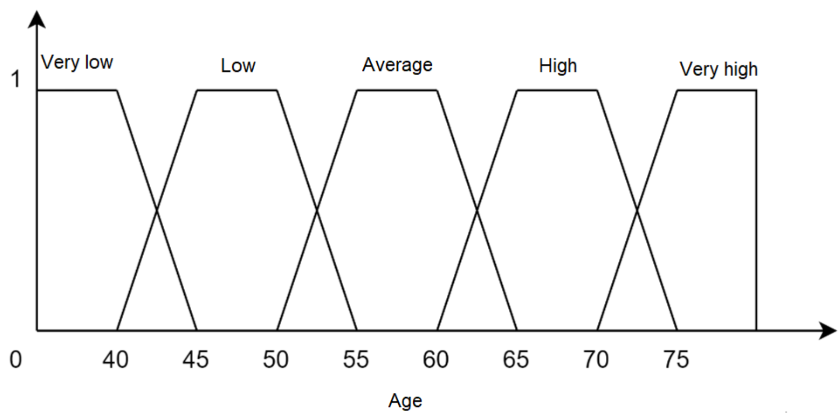


Figure 6. Fuzzy sets of terms in the frame of cognition "age".

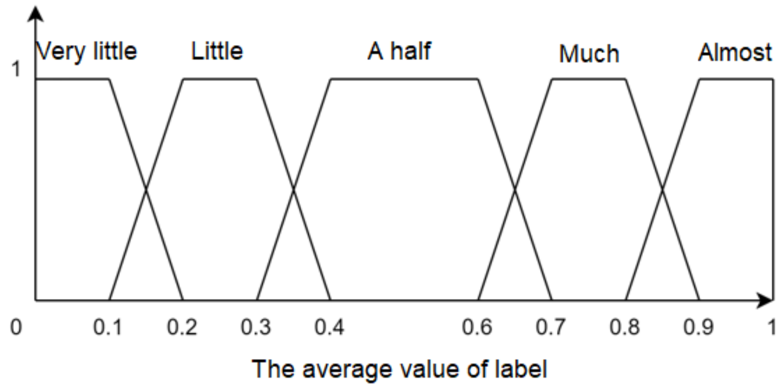


Figure 7. A frame of cognition of Q quantifiers.

Hence, for each term of "age", the membership degree corresponding to the average value of each term of the quantifier is as shown in Table 4.

Table 4. Dependence with quantifiers.

Value	M_0	M_{c-}	M_W	M_{c+}	M_1
$E_{veryFew}$	1	0	0	0.4	1
E_{few}	0	1	0	0.6	0
E_{aHalf}	0	0	0.8	0	0
E_{many}	0	0	0.2	0	0
E_{almost}	0	0	0	0	0

Thus, LS sentences are generated and their truth values T are:

- Very few patients have very low age (T = 1);
- Few patients with low age (T = 1);
- Half of the patients were of average age (T = 0.8);
- Few patients with advanced age (T = 0.6);
- Very few patients are very old (T = 1).

5. Results

The database used is the database of patients with heart disease [60]. This database includes 1025 records and 14 attribute fields (age, sex, resting blood pressure, $\cdots$). This dataset dates back to 1988 and includes four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to the use of a subset of 14 of these attributes. The "target" field refers to the presence of heart disease in the patient. It has the integer value 0 = no disease and 1 = disease.

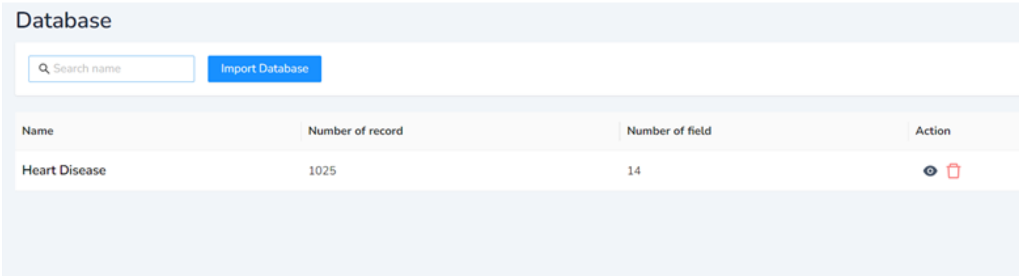
14 sub-attributes are used:

- Age;
- Gender (termed sex in the database);
- Chest pain type (four values);
- Resting blood pressure;
- Serum cholesterol in mg/dL;
- Fasting blood sugar > 120 mg/dL;
- Resting electrocardiographic results (values 0,1,2);
- Maximum heart rate achieved;
- Exercise induced angina;
- Oldpeak = ST depression induced by exercise relative to rest;
- The slope of the peak exercise ST segment;

- Number of major vessels (0-3) colored by flouroscopy;
- Thal: 0 = normal; 1 = fixed defect; 2 = reversible defect;
- Target.

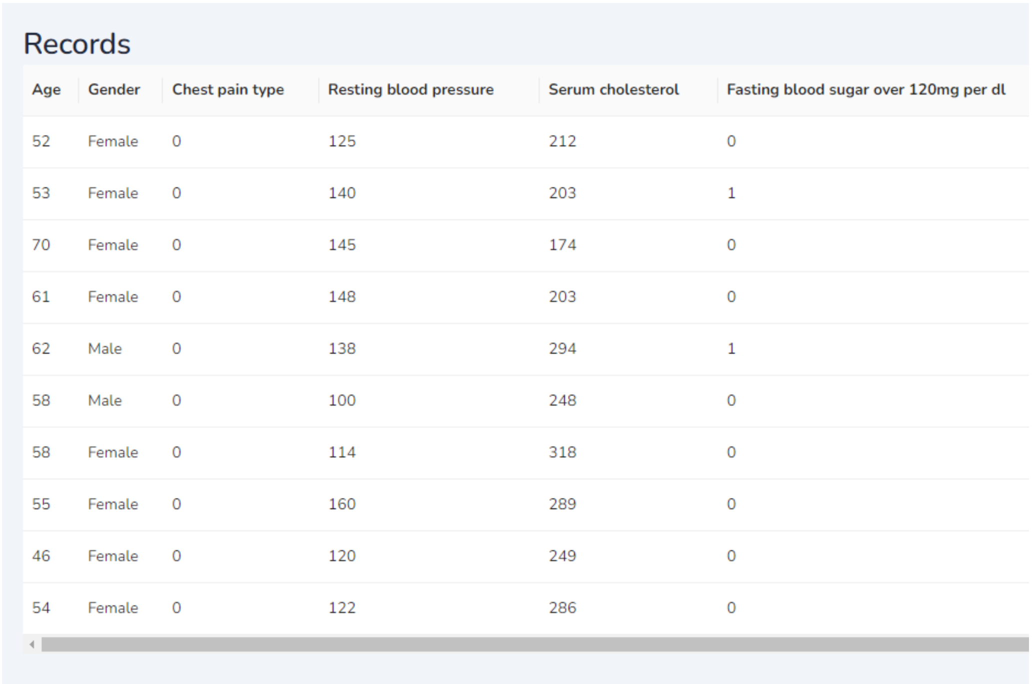
Data descriptions: The database is denoted by D; let $\{f1, f2, f3, ...f14\}$ be the fields of database D , $d_i = \{f1_i, f2_i, ...fn_i\}, i \in [1, 1025]$ is the ith record of D, $fk_i(k \in [1, n])$ is the fk field value of record d_i.

Figure 8 is a screen shot of the interface to add information to the database and Figure 9 shows a screen shot of the list of records in the database.



Name	Number of record	Number of field	Action
Heart Disease	1025	14	

Figure 8. A list of records in the database.



Age	Gender	Chest pain type	Resting blood pressure	Serum cholesterol	Fasting blood sugar over 120mg per dl
52	Female	0	125	212	0
53	Female	0	140	203	1
70	Female	0	145	174	0
61	Female	0	148	203	0
62	Male	0	138	294	1
58	Male	0	100	248	0
58	Female	0	114	318	0
55	Female	0	160	289	0
46	Female	0	120	249	0
54	Female	0	122	286	0

Figure 9. List of records in the database.

In the attribute fields, "target" plays the role of a key filtering condition for information about the generated LS sentences. This is a Boolean value field and corresponds to "target = 0", the patient does not have heart disease, and to "target = 1", the patient has heart disease. Numeric attribute fields that can apply LSmd include "age" "resting blood pressure", "serum cholesterol in mg/dL", and "maximum heart rate achieved". The remaining attribute fields can be used as filter conditions.

5.1. Evaluation Parameters

Because the nature of the paper is to serve the medical field, the most important information is whether, with such health parameters, the patient has heart disease. Therefore, the parameter for the filter condition is selected as F = "target = 1".

Based on the records that satisfy the filter condition F, we proceed to build the experimental parameter sets [1, 2, 3, 4] as shown in Tables 5, 6, 7, and 8, respectively; The parameters will include the applicable attribute LSmd f_k, two generators c+, c− and the corresponding to f_k, set G_x of fuzziness intervals, average fuzziness measure h_0, and calculation of level m. The default quantifiers that will be used for all experiments are (L,V) and the frame of cognition of the default quantifier is as discussed in Section 3.5.

Parameter set 1:

Table 5. Parameter set 1.

Parameter	**Property**	c−	**c+**	G_0	G_W	G_1	h_0	**m**
Value	Age	young	old	[0, 10]	[40, 50]	[80, 100]	0.4	0

Parameter set 2:

Table 6. Parameter set 2.

Parameter	**Property**	c−	**c+**	G_0	G_W	G_1	h_0	**m**
Value	Age	young	old	[0, 10]	[40, 50]	[80, 100]	0.4	3

Parameter set 3:

Table 7. Parameter set 3.

Parameter	**Property**	c−	**c+**	G_0	G_W	G_1	h_0	**m**
Value	RBP	low	high	[0, 80]	[110, 130]	[150, 200]	0.4	0

Parameter set 4:

Table 8. Parameter set 4.

Parameter	**Property**	c−	**c+**	G_0	G_W	G_1	h_0	**m**
Value	RBP	low	high	[0, 80]	[110, 130]	[150, 200]	0.4	3

5.2. Experiments in Fuzzy Rules and LLMs in Extracting Datasets

The doctor's interaction is associated with rules of dialogue, as shown in Figures 10–19 of the following.

Parameter set 1:

Select parameters:

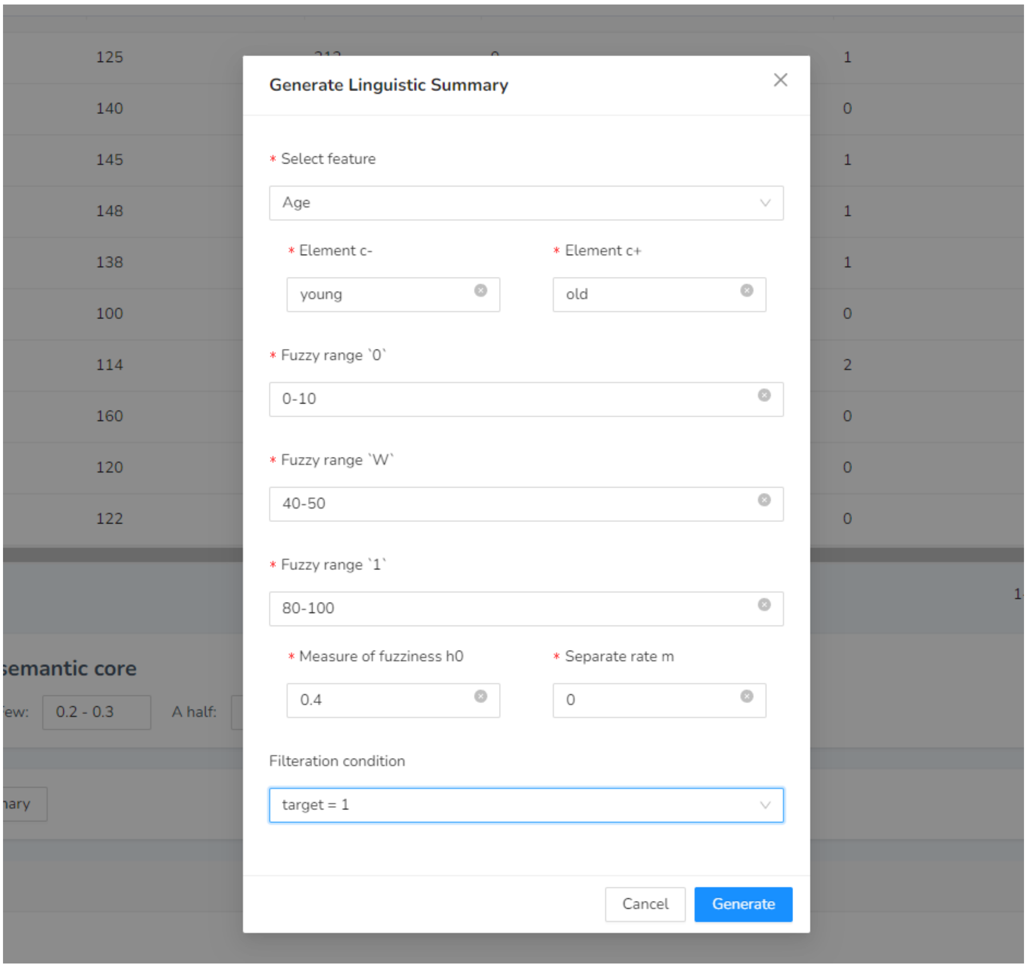


Figure 10. Select parameter set 1.

System output: LS sentences and "truth value" accuracy.

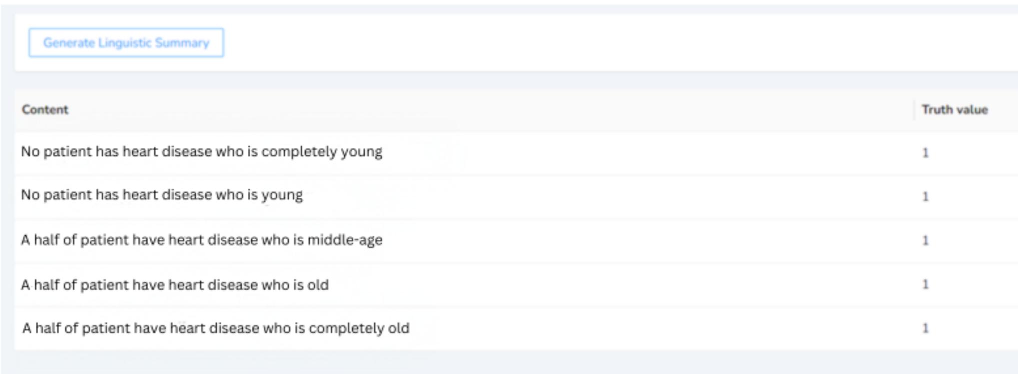


Figure 11. LS sentences of parameter set 1.

Parameter set 2:

Select parameters:

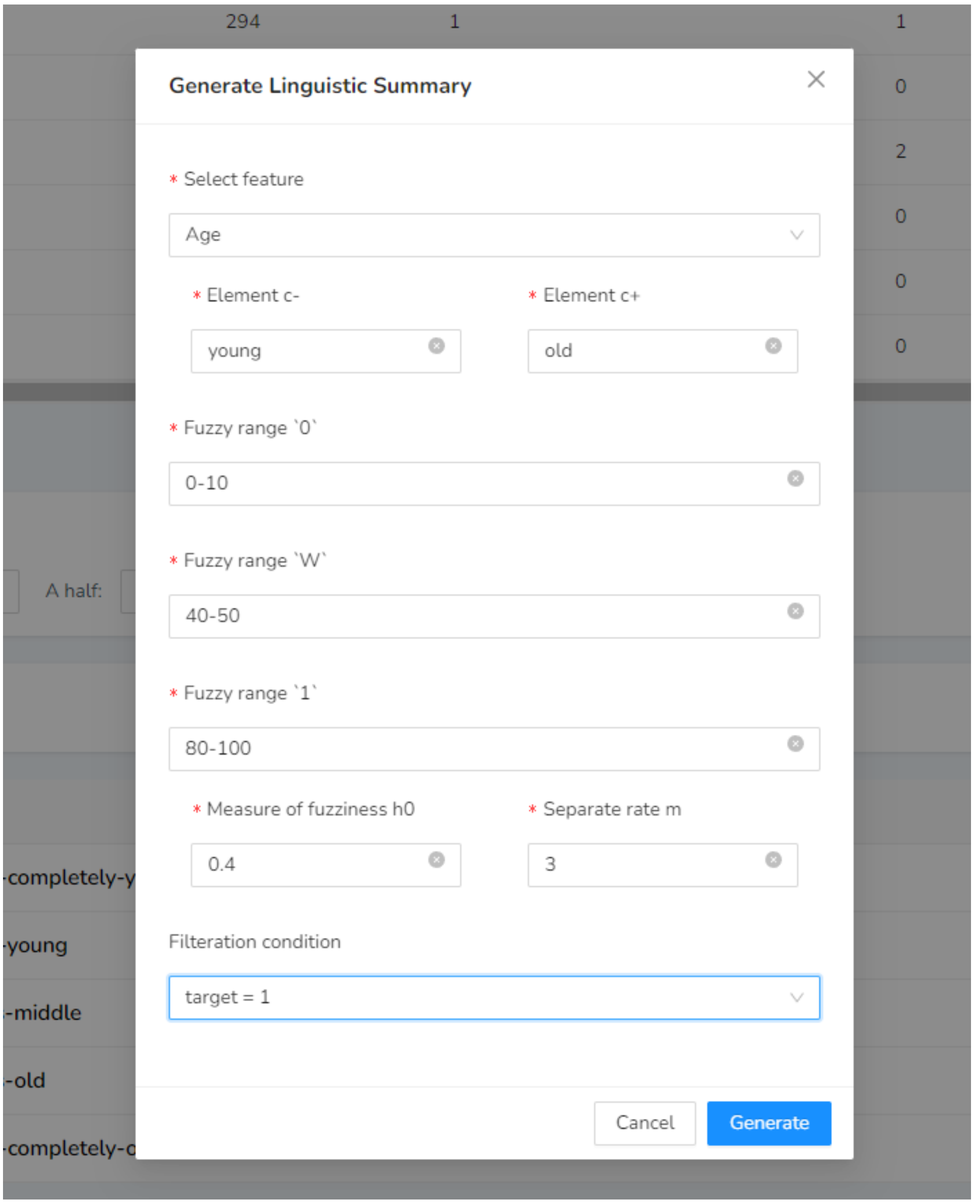
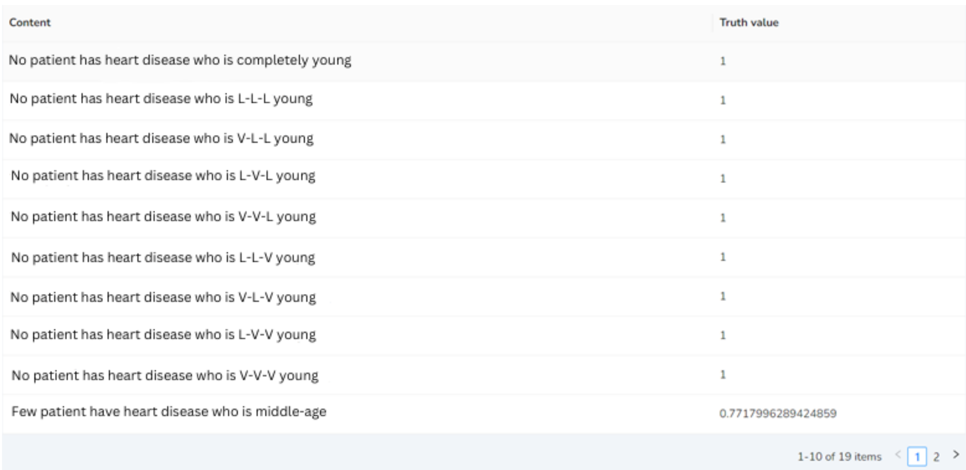


Figure 12. Select parameter set 2.

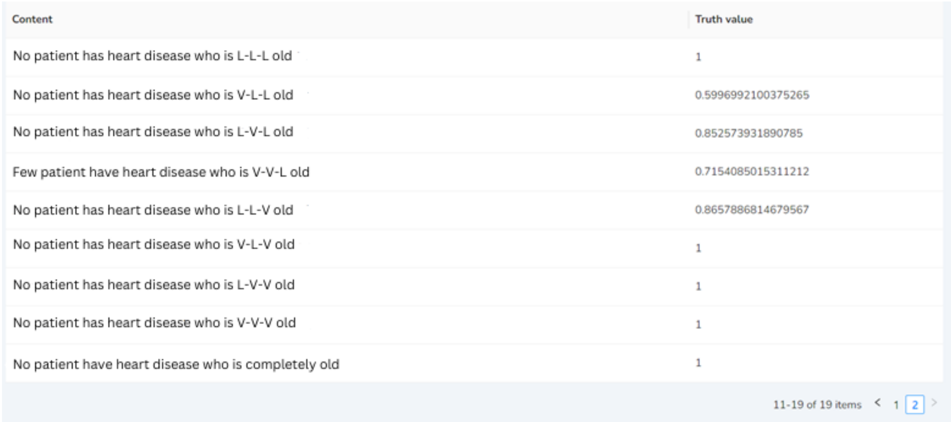
System output: LS sentences and "truth value" accuracy.



Content	Truth value
No patient has heart disease who is completely young	1
No patient has heart disease who is L-L-L young	1
No patient has heart disease who is V-L-L young	1
No patient has heart disease who is L-V-L young	1
No patient has heart disease who is V-V-L young	1
No patient has heart disease who is L-L-V young	1
No patient has heart disease who is V-L-V young	1
No patient has heart disease who is L-V-V young	1
No patient has heart disease who is V-V-V young	1
Few patient have heart disease who is middle-age	0.7717996289424859

1-10 of 19 items < 1 2 >

Figure 13. LS sentences of parameter set 2—page 1.



Content	Truth value
No patient has heart disease who is L-L-L old	1
No patient has heart disease who is V-L-L old	0.5996992100375265
No patient has heart disease who is L-V-L old	0.852573931890785
Few patient have heart disease who is V-V-L old	0.7154085015311212
No patient has heart disease who is L-L-V old	0.8657886814679567
No patient has heart disease who is V-L-V old	1
No patient has heart disease who is L-V-V old	1
No patient has heart disease who is V-V-V old	1
No patient have heart disease who is completely old	1

11-19 of 19 items < 1 2 >

Figure 14. LS sentences of parameter set 2—page 2.

Parameter set 3:
Select parameters:

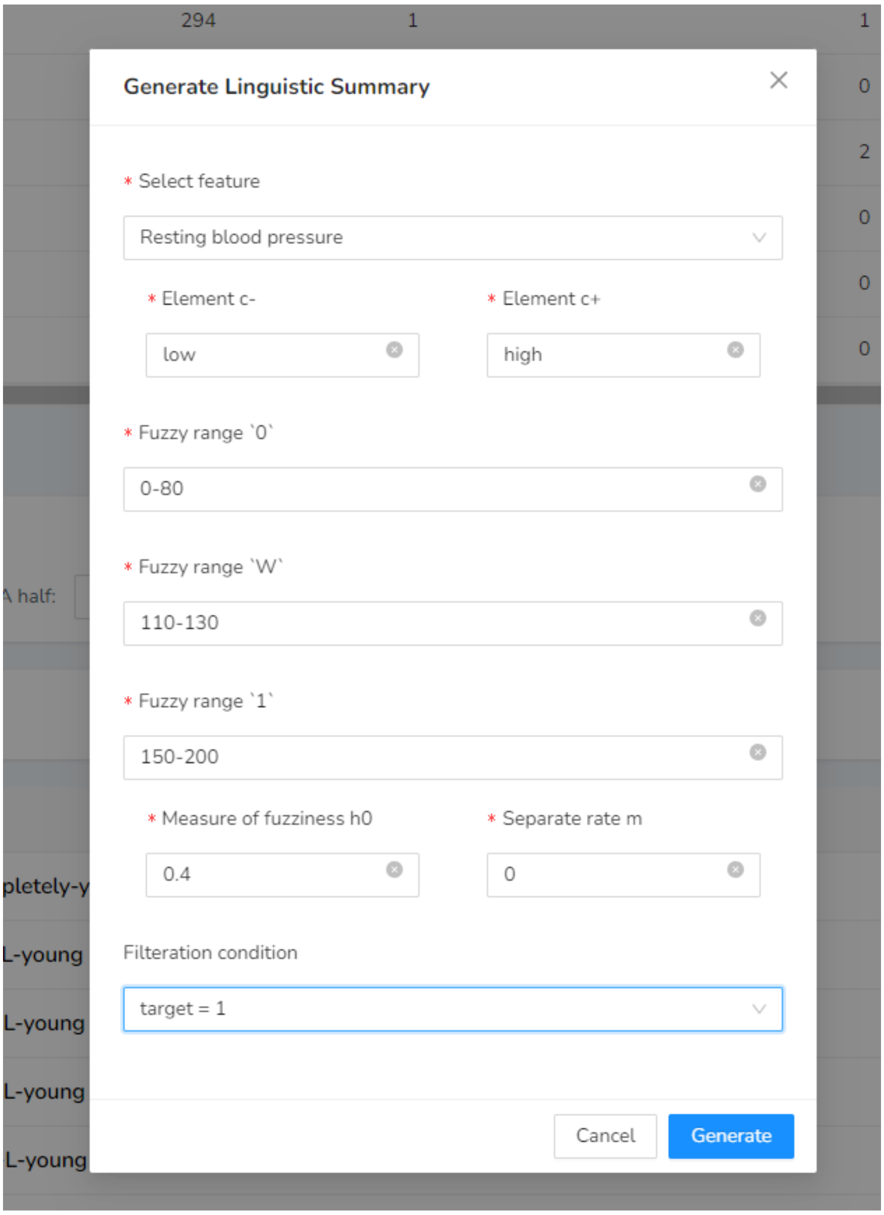
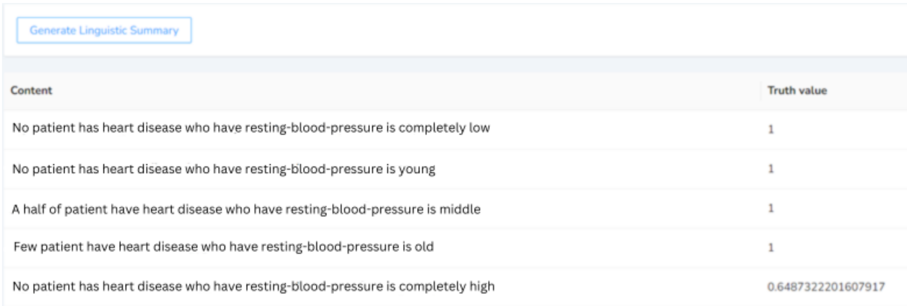


Figure 15. Select parameter set 3.

System output: LS sentences and “truth value” accuracy.



Generate Linguistic Summary

Content	Truth value
No patient has heart disease who have resting-blood-pressure is completely low	1
No patient has heart disease who have resting-blood-pressure is young	1
A half of patient have heart disease who have resting-blood-pressure is middle	1
Few patient have heart disease who have resting-blood-pressure is old	1
No patient has heart disease who have resting-blood-pressure is completely high	0.6487322201607917

Figure 16. LS sentences of parameter set 3.

Parameter set 4:
Select parameters:

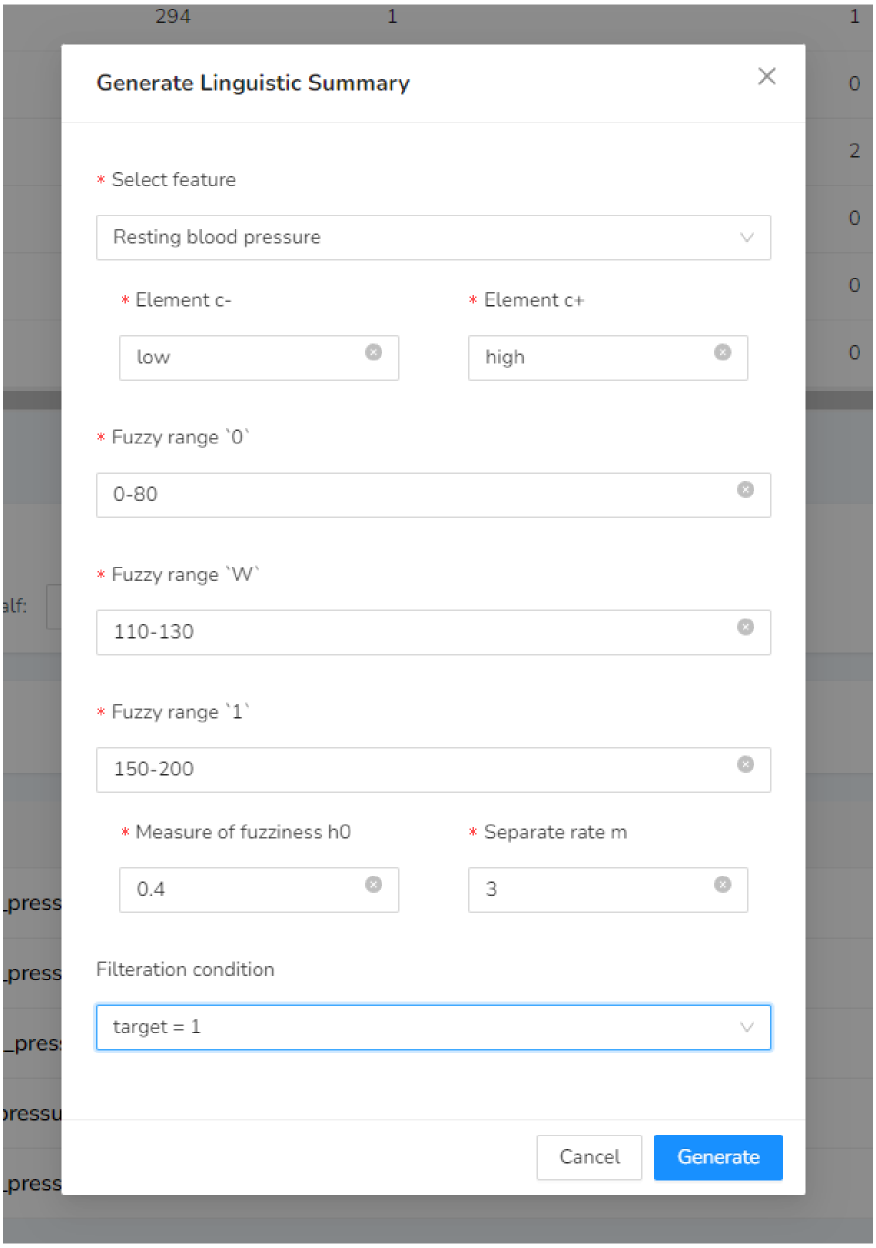


Figure 17. Select parameter set 4.

System output: LS sentences and "truth value" accuracy.

Content	Truth value
No patient has heart disease who have resting-blood-pressure is completely low	1
No patient has heart disease who have resting-blood-pressure is L-L-L young	1
No patient has heart disease who have resting-blood-pressure is V-L-L young	1
No patient has heart disease who have resting-blood-pressure is L-V-L young	1
No patient has heart disease who have resting-blood-pressure is V-V-L young	1
No patient has heart disease who have resting-blood-pressure is L-L-V young	1
No patient has heart disease who have resting-blood-pressure is V-L-V young	1
No patient has heart disease who have resting-blood-pressure is L-V-V young	1
No patient has heart disease who have resting-blood-pressure is V-V-V young	1
A half of patient have heart disease who have resting-blood-pressure is middle	1

1-10 of 19 items < 1 2 >

Figure 18. LS sentences of parameter set 4—page 1.

Content	Truth value
No patient has heart disease who have resting-blood-pressure is L-L-L old	1
No patient has heart disease who have resting-blood-pressure is V-L-L old	1
No patient has heart disease who have resting-blood-pressure is L-V-L old	1
No patient has heart disease who have resting-blood-pressure is V-V-L old	0.8211354924497338
No patient has heart disease who have resting-blood-pressure is L-L-V old	1
No patient has heart disease who have resting-blood-pressure is V-L-V old	1
No patient has heart disease who have resting-blood-pressure is L-V-V old	1
No patient has heart disease who have resting-blood-pressure is V-V-V old	1
No patient has heart disease who have resting-blood-pressure is completely high	0.7198515769944341

11-19 of 19 items < 1 2 >

Figure 19. LS sentences of parameter set 4—page 2.

In the outputs of the experiments, LS sentences are updated to the vector database, which are also instantly prompts to LLMs.

6. Experimental Results

6.1. Dataset in Experiments

'BLOOM' [57] is used as an instruction dataset to train the model; it consists of 498 Hugging Face datasets and 46 natural languages [61]. It is used to evaluate the testing regime, and the experimental results derived in the case study of English and Vietnamese languages.

The baseline datasets have been investigated for the proposed model using ***Vicuna*** [62], which is a dataset used to validate the tests in an evaluation benchmark. 'Vicuna' consists of 80 questions categorized into eight distinct groups. The benchmark dataset has been tested for the proposed model using the language model's capacity.

6.2. Training Model with Optimal Approach Using Low-Rank Adaption

Low-rank adaption (LoRA), as a key method in natural language processing, is training on general domain data and adapting to specific tasks or fields. With large models, full fine-tuning, which involves retraining all the model's parameters, becomes challenging due to memory issues. For a model with over 100 billion parameters, training becomes prohibitively expensive due to high hardware requirements. For instance, with weights expressed in a '16-bit floating-point' format, the memory required to load a 100 billion parameter model is $100 \times 10^9 \times 2$ bytes bytes, approximately 372 GB. It is therefore clear that no reasonably priced GPU can meet such VRAM requirements. Therefore, the LoRA technique is proposed to address this problem, which relates to the limited resources available to an SME. The benefits are derived from the use of LoRA of the following:

- Firstly, a pre-trained model can be shared and used to build LoRA modules for a variety of tasks. It freezes the model weights and dynamically converts matrices A and B. In addition, it reduces the cost of storage and conversion between tasks.
- Secondly, LoRA makes the training process efficient while reducing the hardware limitations by up to 3 times (being optimized dynamically). It can be calculated for gradients or maintain an optimal state for all parameters.
- Thirdly, to optimize the parameters of the inserted low-level matrix, conventional fine-tuning methods often encounter the problem of inference latency, which can be simply applied at the time it takes to process and respond to the model after being trained. However, LoRA can be designed to help with training matrices, without causing inference latency in the full fine-tuned model.
- Finally, LoRA is independent of many methods which can be completely combined with each other. In addition, LoRA can be applied to limit the model's performance, since it learns from a small number of parameters. To improve the model's performance, full-parameter fine tuning can be employed, although it may lead to significant training resource consumption.

For the training of the weight matrix $W_0 \in \mathbb{R}^{d \times k}$, the parameter is δW, which is created by dimensions compared to the pre-trained weight: A, compression matrix; and B, decompression matrix. These matrices can be updated by the latter with a low-rank decomposition, as in Equation (4). Figure 20's models show the relationship(s) and the process.

$$W_0 + \delta W = W_0 + BA \tag{4}$$

where $\left(B \in \mathbb{R}^{d \times r}\right)$, $\left(A \in \mathbb{R}^{r \times k}\right)$, and $(r \ll \min(d,k)))$.

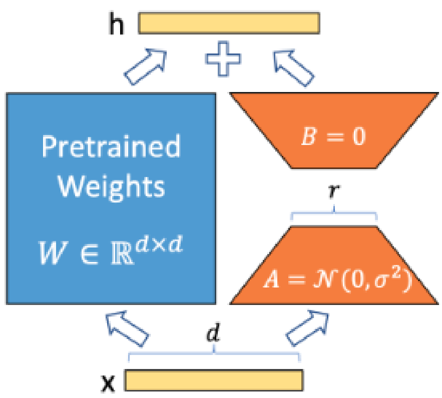


Figure 20. The operational mechanism of LoRA is delineated through the flow depicted in the image.

During the training process, (W_0) presents gradient updates, while (A) and (B) consist of trainable parameters. Consider $x \in R^k$ which can be input to the model, both (W_0) and $(\delta W = BA)$ are multiplied with the same input and output vectors. The hidden state of x-h in the model is expressed by Equation (5):

$$h = W_0 x + \delta W x = W_0 x + BAx \tag{5}$$

The proposed model uses 'DeepSpeed' [59] which is a deep learning optimization library providing advanced techniques for improvement of the performance of deep learning models. The proposed model uses its large-scale model training with time redundancy. The significant distributed training libraries (e.g., *torchrun* or *accelerate*) allow for loading of data in real time.

In the experiments, we conducted tests using the NVIDIA A100 40 GB GPU. Moreover, DeepSpeed with a batch_size of 1 can lower VRAM usage, to effectively utilize the GPU. Two model training techniques are used, LoRA [58,63] and 'DeepSpeed'. LoRA is to reduce the training time by adapting training layers and freezing the backbone for optimal computational costs of the GPU in the training process. Table 9 shows a comparison of approaches in the following: training time, batch size, and memory consumption for full fine-tuning, using LoRA combined with DeepSpeed. In the experiments, the 'BLOOM' model with the 7 billion-parameter set (BLOOM-7B1) was identified as the most suitable model.

Table 9. Comparison of methods.

	Time/Epoch	Batch Size	Memory
Proposed model (BLOOM)	54.5 h	1	3.59 GB
Proposed model (BLOOM) + LoRA	4 h	1	39.5 GB
Proposed model (BLOOM) + LoRA + DeepSpeed	4 h	1	36.5 GB
Proposed model (BLOOM) + LoRA + DeepSpeed	3 h	2	39.5 GB

6.3. Prompting

For inference processes the question–answer pairs are shown in Figure 21. The input prompt enables the *GenAI-Algebra* and Phoenix models to assist, which is also required to respond to these questions.

```
A chat between medical advice and our proposed GenAI-Algebra in heath
care questions.
Human:  <s> Question / Mining hidden rules from Mass datasets  </s>
GenAI-Algebra Assistant:  <s> Answer </s>.
```

Figure 21. Prompt is applied to testing for *GenAI-Algebra* in heath care questions.

The detailed prompt shown in the case study of the heart dialog are shown in pairs of questions for prompts in Figure 22. The input prompt enables the *GenAI-Algebra* to find hidden rules in extracting the mass datasets, quantified by fuzzy rules as shown in Figures 10–19.

```
A chat between heart problem advice and the proposed GenAI-Algebra model
in heath care questions.
Human:  <s> What do hidden rules consider in UCI datasets of the heart
domain  </s>
GenAI-Algebra Assistant:  <s>  No patient has heart dissease who have
resting blood pressure which is completed high  </s>.
```

Figure 22. Prompt is applied to LS sentence of the *GenAI-Algebra* in heart questions.

7. Experimental Results

In this section, the proposed model tested using a case study. In the case study we set out an evaluation of the *GenAI-Algebra* model based on the English and Vietnamese languages.

The proposed *GenAI-Algebra* model is compared to the Phoenix [35] with common features of the proposed model generally improving on the Phoenix model. Phoenix was created by fine-tuning 'BLOOM' with the datasets as follows:

- **Multilingual instruction**: Uses the Alpaca instruction dataset with 'gpt-3.5-turbo' API to generate answers.
- **User-centered instruction**: The 'gpt-3.5-turbo' API is used to generate answers for each sample.

- **Conversation**: It consists of conversation histories.

7.1. Evaluation

To validate the proposed model, the evaluation criteria consist of the question and quality of the responses in terms of accuracy. Equation (6) was used to evaluate the Phoenix model [35], where the equation was used for comparison with other language models.

$$P = \frac{\sum_{i=1}^{n} score_i^X}{\sum_{i=1}^{n} score_i^Y} \tag{6}$$

where the performance (P) of model X and model Y is given by using Formula (6), with (n) being the total questions. In a general case, $\left(score_i^j\right)$ is the score for the ith question of model j.

7.2. Comparison of GenAI-Algebra and Phoenix Method Using English Benchmark

In our evaluation of *GenAI-Algebra*, it performed well on the English benchmark. The experimental results are shown in Table 10. The experimental results show that *GenAI-Algebra* has a better performance than Phoenix on the English benchmark. In the evaluation, the scores for answers are obtained from the 'gpt-3.5-turbo' API, calculated by (6). Table 11 shows the performance ratio results between *GenAI-Algebra* and Phoenix and calculated by Equation (6).

Table 10. Details of the number of wins for each model over the categories in both English and Vietnamese. The bold numbers indicate the model that won in each category.

	English			Vietnamese	
Category	**Phoenix**	***GenAI-Algebra***	**Total**	**Phoenix**	***GenAI-Algebra***
Heart common	2	**5**	7	3	**4**
Health sense	3	**6**	10	4	**6**
Health care	4	**6**	10	**5**	5
Consultant	4	**6**	10	**6**	4
Generic	3	**7**	10	**6**	4
Knowledge	3	**7**	10	3	**7**
Math	**6**	4	10	**6**	4
Heart dialog	4	**6**	10	4	**6**
Common sense	**6**	4	10	**5**	**5**
Total wins	12	**43**	87	18	**21**

Table 11. Performance ratio (%) of *GenAI-Algebra* compared to Phoenix in the comparison on the English benchmark.

Performance Ratio	**English**	**English in Specification**
Phoenix	97.89	95.72
GenAI-Algebra	97.50	96.70

In experimental testing, the proposed *GenAI-Algebra* achieves performance results in the range of 96.70–97.50% compared to Phoenix, which achieved results in the range of 95.72–97.89% compared to ChatGPT on the Vietnamese and English benchmarks. Furthermore, the training time was reduced with its limited hardware resources with respect to the normal training method.

7.3. Comparison of GenAI-Algebra and Phoenix Method in Winning Cases

The experimental results shown in Table 10 use both English and Vietnamese as benchmarks. In a comparative results evaluation we can see the *GenAI-Algebra* and Phoenix models.

The *GenAI-Algebra* and Phoenix models have been tested in total of 80 categories. The experimental results in Table 10 show the *GenAI-Algebra* model achieves a significant improvement in 43 categories and performs as well in 21 categories for the Vietnamese benchmark. In summary, the results for both the English and Vietnamese benchmarks are as follows:

- The *GenAI-Algebra* model showed a better performance than Phoenix in some categories in the healthcare, heart domain.
- The overall results for the two models were similar when using the English and Vietnamese benchmarks for testing.

8. Discussion

This study addresses the creation of a chatbot utilizing GenAI and LLM(s). The novel feature in our proposed*GenAI-Algebra* model is the identification of hidden rules in large datasets with appropriate question–response interactions. Moreover, the proposed *GenAI-Algebra* method has the capability to reduce the resource requirements, thus providing an effective basis upon which an SME can implement a multilingual chatbot.

The model training using 'low-rank adaptation' contributed to a reduction in training time and computational cost. In addition, we posit that our proposed model will be used for other languages. A reinforcement learning from human feedback (RLHF) method can be designed to improve the quality and safety of chatbot responses to questions and the quality of the extracting rules.

When reviewing large datasets of projects [for example in the medical domain used in the case study], the *GenAI-Algebra* outlines a framework describing the five levels of GenAI solutions through seven different levels of complexity. By using the *GenAI-Algebra* model, organizations can clearly understand their current position in the proposed model. This understanding will help them plan specific strategies to achieve their business goals.

To align internal skills and capabilities with desired business outcomes, enterprises can realistically assess their current position according to the *GenAI-Algebra* model. They should then consider the business outcomes they aim to achieve and evaluate what needs to be achieved to reach that future maturity state. This involves technical aspects and allows for practical adjustments in initiatives, skill development, support, and build-or-buy decisions. Understanding their maturity level will assist them in transforming to realize the desired business outcomes.

GenAI-Algebra enhances data strategy, processes, sharing, and more, alongside predictive AI in deploying end-to-end applications. In the preparation of datasets, it focuses on creating, managing, and preparing data—the essential raw material for GenAI models. This involves collecting large datasets, cleaning them, and ensuring their quality and relevance for training purposes. All of the LS sentences have truth values T of each LS sentence, which are quantified from hidden rules in large datasets. These LS sentences are updated to a vector database for prompts in LLMs of the GenAI application. In multiple domains, we can set up multiple models such as *GenAI-Algebra* and GenAI models of chatbots.

8.1. Practical Managerial Significance

The experimental results show that organizations can select suitable GenAI models and create effective prompts to interact with them. Prompts are textual inputs that guide the model's outputs, and choosing the right model and prompts is crucial for achieving the desired outcomes. Additionally, this level involves serving these models, making them accessible for specific tasks, to fine-tune the *GenAI-Algebra* with proprietary or domain-specific data. Fine-tuning is the process of adjusting a pre-trained model to better suit a particular task or field, enhancing its performance and customization capabilities. This allows organizations/enterprises to tailor the model to their unique needs and requirements.

In the case study [in the medical domain] *GenAI-Algebra* is further refined through benchmarking and output evaluation, ensuring accuracy, relevance, and ethical alignment. Multi-agent systems are deployed, where multiple GenAI models collaborate under the coordination of a large language model (LLM) and use larger datasets using algebra with its fuzzy rules. This facilitates complex tasks requiring coordination and integration of diverse capabilities. By incorporating knowledge bases into the *GenAI-Algebra* model, a chatbot can learn from human feedback and adapt to doctor preferences, while also ensuring the safety and privacy of responses. A further approach to generating domain-specific knowledge bases for the domain is to deploy the model on a specific user group, and to collect large datasets.

8.2. Future Work

While this study has addressed a number of research questions relating to question–response interactions of both responses and automatic responses from rules considered from extracting large datasets, the proposed model has some weaknesses in timing processes as follows: (1) it is hard for LLMs to extract data from large datasets in various domains with raw or unstructured data; (2) the proposed models struggle to deal with various domains at the same time.

For further investigations, large datasets of various domains with raw or unstructured data can be processed as clean datasets, which transforms them to a vector database in order to extract rules for LLMs. To address problems by dealing with various domains at the same time, GenAI-Algebra models can be applied in specific domains by incorporating a knowledge base. To apply the GenAI-Algebra model with its potential application to other sectors, domain-specific models of these sectors require access to high-quality training data and expertise in the target domain. By incorporating a knowledge base into the GenAI-Algebra model, a chatbot can enhance response quality in specific domains. A further approach to generating domain-specific knowledge bases for sectors is to deploy the model in potential applications.

We have created a GenAI model for a chatbot for application to heterogeneous domains, complete with an LLM that can adapt to multiple languages such as Vietnamese and English for use in GenAI models, suitable for resource-limited SMEs. To adapt to specific domains, the *GenAI-Algebra* has been performed with large language models (LLMs); the BLOOM approach, such as LLaMA 2 and Mistral, continuously raises the performance bar. To enhance the model with specific domains, we can replace the backbone with newer, better-performing models like LLaMA or Mistral, depending on the specific use case of SMEs. For instance, the LLaMA and Mistral models excel primarily in English language tasks.

In future work, we will investigate how to create a virtual assistant that supports the quality of chatGPT with automatically mined rules in large datasets while minimizing computational costs for domains in smart cities. Further studies will investigate multiple models of extracting datasets by dealing with multimodal design for the future of ChatGPT.

However, there will as always be open research questions (ORQs) resulting from the research. Such ORQs include:

- Technical and societal issues including the impact and effect(s) resulting from the development and implementation of GenAI-driven systems.
- Such socio-technical effects of GenAI may be classified in terms of technological determinism (TD) as discussed in [37], which is characterized by delays in understanding such effects, as addressed in [64].
- In this study, we have noted the parameter “sex” to describe an individual’s gender (i.e., ‘male’ or ‘female”). However, the societal change in gender identification, as discussed in [65], forms a significant issue reflected in “transnormativity” (i.e., non-binary identity).

Addressing TD and transnormativity forms a difficult challenge.

9. Conclusions

The experimental results show the combination of hedge algebra approaches and a multilingual large language model to find hidden rules; the case study in the medical domain has shown the utility of the proposed approach. Extracting natural language knowledge from large datasets utilizing fuzzy sets and hedge algebra to extract these rules presented in meta data for ChatGPT and generative AI applications provides an effective solution for an SME to implement a GenAI-driven chatbot. This investigation contributes to the discussion on how GenAI can be leveraged to maximum effect for small and medium-sized enterprises constructively.

Author Contributions: Conceptualization, methodology, H.V.P. and P.M.; software, H.V.P.; validation, H.V.P. and P.M.; formal analysis, H.V.P.; investigation, H.V.P. and P.M.; data, H.V.P. and P.M.; writing—original draft preparation, H.V.P.; writing—review and editing, H.V.P. and P.M.; project administration, H.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology of Viet Nam under Program KC4.0, No. KC-4.0-38/19-25.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code and datasets of the research are available at https://github.com/dat-browny/Expert-B/tree/master/models (accessed on 20 December 2023).

Acknowledgments: The authors thank Nguyen Ha Thanh Dat and technical Engineering, Hanoi University of Science and Technology, doctor from Traditional Medical Hospital, and language experts participating in the experiments in ChatGPT for both English and Vietnamese languages. This work has been supported by Ministry of Scicence and Technology of VietNam under Program KC4.0, No. KC-4.0-38/19-25.

Conflicts of Interest: The authors declare no conflicts of interest. This research does not involve any human or animal participation. All authors have checked and agreed with the submission.

References

1. Christensen, C.M.; McDonald, R.; Altman, E.J.; Palmer, J.E. Disruptive Innovation: An Intellectual History and Directions for Future Research. *J. Manag. Stud.* **2018**, *55*, 1043–1078. [CrossRef]
2. Van Pham, H.; Thai, K.P.; Nguyen, Q.H.; Le, D.D.; Le, T.T.; Nguyen, T.X.D.; Phan, T.T.K.; Thao, N.X. Proposed Distance and Entropy Measures of Picture Fuzzy Sets in Decision Support Systems. *Int. J. Fuzzy Syst.* **2023**, *44*, 6775–6791. [CrossRef]
3. Pham, H.V.; Duong, P.V.; Tran, D.T.; Lee, J.H. A Novel Approach of Voterank-Based Knowledge Graph for Improvement of Multi-Attributes Influence Nodes on Social Networks. *J. Artif. Intell. Soft Comput. Res.* **2023**, *13*, 165. [CrossRef]
4. VPham, V.H.; Nguyen, Q.H.; Truong, V.P.; Tran, L.P.T. The Proposed Context Matching Algorithm and Its Application for User Preferences of Tourism in COVID-19 Pandemic. In *International Conference on Innovative Computing and Communications*; Springer Nature: Singapore, 2023; Volume 471. [CrossRef]
5. Eysenbach, G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Med Educ.* **2023**, *9*, e46885. [CrossRef]
6. Michail, A.; Konstantinou, S.; Clematide, S. UZH_CLyp at SemEval-2023 Task 9: Head-First Fine-Tuning and ChatGPT Data Generation for Cross-Lingual Learning in Tweet Intimacy Prediction. *arXiv* **2023**, arXiv:2303.01194.
7. Haleem, A.; Javaid, M.; Singh, R.P. An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Trans. Benchmarks Stand. Eval.* **2022**, *2*, 100089. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010. Available online: https://dl.acm.org/doi/10.5555/3295222.3295349 (accessed on 8 June 2024).
9. Hagendorff, T.; Fabi, S.; Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* **2023**, *3*, 833–838. [CrossRef]
10. Anam Nazir, Z.W. A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. *Meta-Radiology* **2023**, *1*, 100022. [CrossRef]
11. Chiarello, F.; Giordano, V.; Spada, I.; Barandoni, S.; Fantoni, G. Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation* **2024**, *133*, 103002. [CrossRef]

12. Crosthwaite, P.; Baisa, V. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Appl. Corpus Linguist.* **2023**, *3*, 100066. [CrossRef]
13. Tuan, N.T.; Moore, P.; Thanh, D.H.V.; Pham, H.V. A Generative Artificial Intelligence Using Multilingual Large Language Models for ChatGPT Applications. *Appl. Sci.* **2024**, *14*, 3036. [CrossRef]
14. Khosla, M.; Anand, A.; Setty, V. A comprehensive comparison of unsupervised network representation learning methods. *arXiv* **2019**, arXiv:1903.07902.
15. Sun, Q.; Zhao, C.; Tang, Y.; Qian, F. A survey on unsupervised domain adaptation in computer vision tasks. *Sci. Sin. Technol.* **2022**, *52*, 26–54. [CrossRef]
16. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [CrossRef]
17. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 63. [CrossRef]
18. Beck, J.; Vuorio, R.; Liu, E.Z.; Xiong, Z.; Zintgraf, L.; Finn, C.; Whiteson, S. A survey of meta-reinforcement learning. *arXiv* **2023**, arXiv:2301.08028.
19. Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; Sui, Z. A survey for in-context learning. *arXiv* **2022**, arXiv:2301.00234.
20. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.L.; Tang, Y. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1122–1136. [CrossRef]
21. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.S.; Sun, L. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv* **2023**, arXiv:2303.04226.
22. Borji, A. A categorical archive of chatgpt failures. *arXiv* **2023**, arXiv:2302.03494.
23. Alkaissi, H.; McFarlane, S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **2023**, *15*, e35179. [CrossRef]
24. Cotton, D.R.; Cotton, P.A.; Shipway, J.R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **2023**, *61*, 1–12.
25. Howard, A.; Hope, W.; Gerada, A. ChatGPT and antimicrobial advice: The end of the consulting infection doctor? *Lancet Infect. Dis.* **2023**, *23*, 405–406. [CrossRef]
26. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [CrossRef]
27. Biswas, S.S. Role of chat gpt in public health. *Ann. Biomed. Eng.* **2023**, *51*, 868–869. [CrossRef] [PubMed]
28. McGee, R.W. Is Chat GPT Biased against Conservatives? An Empirical Study (15 February 2023). Available online: https://ssrn.com/abstract=4359405 (accessed on 5 May 2024).
29. Ali, M.J.; Djalilian, A. Readership awareness series—Paper 4: Chatbots and chatgpt-ethical considerations in scientific publications. In *Seminars in Ophthalmology*; Taylor & Francis: Abingdon, UK, 2023; pp. 1–2.
30. Naumova, E.N. A mistake-find exercise: A teacher's tool to engage with information innovations, ChatGPT, and their analogs. *J. Public Health Policy* **2023**, *44*, 173–178. [CrossRef]
31. King, M.R.; ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cell. Mol. Bioeng.* **2023**, *16*, 1–2. [CrossRef]
32. Thorp, H.H. ChatGPT Is Fun, but Not an Author, 2023. Available online: https://www.science.org/doi/full/10.1126/science.adg7879 (accessed on 5 May 2024).
33. Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv* **2023**, arXiv:2303.04671.
34. Li, M.; Wang, R. Chatbots in e-commerce: The effect of chatbot language style on customers' continuance usage intention and attitude toward brand. *J. Retail. Consum. Serv.* **2023**, *71*, 103209. [CrossRef]
35. Chen, Z.; Jiang, F.; Chen, J.; Wang, T.; Yu, F.; Chen, G.; Zhang, H.; Liang, J.; Zhang, C.; Zhang, Z.; et al. Phoenix: Democratizing ChatGPT across Languages. *arXiv* **2023**, arXiv:2304.10453.
36. Mackenzie, D. Surprising Advances in Generative Artificial Intelligence Prompt Amazement—and Worries. *Engineering* **2023**, *25*, 9–11. [CrossRef]
37. Evans, O.; Wale-Awe, O.; Osuji, E.; Ayoola, O.; Alenoghena, R.; Adeniji, S. ChatGPT impacts on access-efficiency, employment, education and ethics: The socio-economics of an AI language model. *BizEcons Q.* **2023**, *16*, 1–17.
38. Baidoo-Anu, D.; Owusu Ansah, L. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI* **2023**, *7*, 52–62. [CrossRef]
39. Kohnke, L.; Moorhouse, B.L.; Zou, D. Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Comput. Educ. Artif. Intell.* **2023**, *5*, 100156. [CrossRef]
40. Martínez-Plumed, F.; Gómez, E.; Hernández-Orallo, J. Futures of artificial intelligence through technology readiness levels. *Telemat. Inform.* **2021**, *58*, 101525. [CrossRef]
41. Sætra, H.S. Generative AI: Here to stay, but for good? *Technol. Soc.* **2023**, *75*, 102372. [CrossRef]
42. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

43. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
44. Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; et al. A mathematical framework for transformer circuits. *Transform. Circuits Thread* **2021**, *1*, 12.
45. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]
46. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
47. Liu, Z.; Lin, W.; Shi, Y.; Zhao, J. A robustly optimized BERT pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 471–484.
48. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763. Available online: https://dl.acm.org/doi/10.5555/3454287.3454804 (accessed on 8 June 2024).
49. Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; Schneider, N. Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, 8–9 August 2013; pp. 178–186.
50. Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O.K.; Liu, Q.; et al. Language is not all you need: Aligning perception with language models. *arXiv* **2023**, arXiv:2302.14045.
51. Chen, S.; Liu, J.; Wang, H.; Xu, Y.; Augusto, J.C. A linguistic multi-criteria decision making approach based on logical reasoning. *Inf. Sci.* **2014**, *258*, 266–276. [CrossRef]
52. Nguyen, C.H.; Tran, T.S.; Pham, D.P. Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application. *Knowl.-Based Syst.* **2014**, *67*, 244–262. [CrossRef]
53. Klir, G.K.; Yuan, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*; Prentice Hall: Hoboken, NJ, USA, 1995.
54. Berkan, R.C.; Trubatch, S.L. *Fuzzy Systems Design Principles: Building Fuzzy IFTHEN Rule Bases*; IEEE Press: Piscataway, NJ, USA, 1997.
55. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
56. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
57. Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. BLOOM A 176B-Parameter Open-Access Multilingual Language Model. *arXiv* **2022**, arXiv:2211.05100.
58. Sun, X.; Ji, Y.; Ma, B.; Li, X. A Comparative Study between Full-Parameter and LoRA-based Fine-Tuning on Chinese Instruction Data for Instruction Following Large Language Model. *arXiv* **2023**, arXiv:2304.08109.
59. Ren, J.; Rajbhandari, S.; Aminabadi, R.Y.; Ruwase, O.; Yang, S.; Zhang, M.; Li, D.; He, Y. ZeRO-Offload: Democratizing Billion-Scale Model Training. *arXiv* **2021**, arXiv:2101.06840.
60. Lapp, D. Heart Disease Dataset | Kaggle—kaggle.com. 2019. Available online: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset (accessed on 2 March 2024).
61. Laurençon, H.; Saulnier, L.; Wang, T.; Akiki, C.; Villanova del Moral, A.; Le Scao, T.; Von Werra, L.; Mou, C.; González Ponferrada, E.; Nguyen, H.; et al. The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 31809–31826
62. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685. [CrossRef]
63. Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2022**, arXiv:2106.09685.
64. Checkland, P.; Holwell, S. *Information, Systems and Information Systems: Making Sense of the Field*; John Wiley and Sons: Oxford, UK, 1997.
65. Murawsky, S. The struggle with transnormativity: Non-binary identity work, embodiment desires, and experience with gender dysphoria. *Soc. Sci. Med.* **2023**, *327*, 115953. [CrossRef]