XAI FOR ALL : CAN LARGE LANGUAGE MODELS SIMPLIFY EXPLAINABLE AI?

Philip Mavrepis[1] Georgios Makridis[1] Georgios Fatouros[1] Vasileios Koukos[1]
Spyros Theodoropoulos[1] Maria Margarita Separdani[2] and Dimos Kyriazis[1]

[1]Department of Digital Systems, University of Piraeus, Athens, Greece
[2]Department of Maritime Studies, University of Piraeus, Athens, Greece

ABSTRACT

Explainable Artificial Intelligence (XAI) is essential for making AI models transparent and understandable. However, existing XAI methods often cater to users with strong technical backgrounds, creating barriers for non-experts to comprehend these techniques. Addressing this challenge, our paper introduces "x[plAIn]," a novel approach that enhances the accessibility of XAI through a custom Large Language Model (LLM) developed using ChatGPT Builder. The objective is to design a model capable of generating clear and concise summaries of various XAI methods, tailored to different audiences such as business professionals and academics. The key novelty of our work lies in the model's ability to adapt explanations to match each audience's knowledge level and interests, providing timely insights that facilitate informed decision-making. Results from our use-case studies demonstrate that "x-[plAIn]" effectively delivers easyto-understand, audience-specific explanations regardless of the XAI method employed. This adaptability not only improves the accessibility of XAI but also bridges the gap between complex AI technologies and their practical applications. Our findings indicate a promising direction for leveraging LLMs to make advanced AI concepts more accessible to a diverse range of users.

KEYWORDS

XAI, Human-Centric Explainable AI, LLM, GPT Builder, AI

1. INTRODUCTION

In the contemporary era, often referred to as the Digital or Information Age, we witness an unprecedented proliferation of sophisticated computational systems generating vast amounts of data daily. This period is further defined by the digital transformation within industrial sectors, culminating in the emergence of the fourth industrial revolution, Industry 4.0 [1]. Artificial Intelligence (AI) serves as the cornerstone of this revolution, acting as the pivotal facilitator that fosters the development of innovative tools and processes [2].

As AI systems become more pervasive, there is a growing emphasis on the need for transparency and interpretability in AI models, giving rise to the field of Explainable Artificial Intelligence (XAI). XAI focuses on providing intelligible explanations for the inferences and decisions formulated by machine learning algorithms, enabling users to understand, trust, and effectively manage AI systems.

David C. Wyld et al. (Eds): NLAI, IBCOM, MLDS, GridCom, SCAI, ICCSEA, SPPR – 2024
pp. 127-144, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142208

Despite advancements in XAI, a significant challenge persists: existing XAI methods often cater to users with strong technical backgrounds, creating barriers for non-experts to comprehend these techniques. This limitation hampers the broader adoption and understanding of AI technologies across diverse professional contexts. The complexity of quantifying explainability and the subjective nature of what constitutes a satisfactory explanation for different stakeholders further complicate this issue.

The problem is particularly pronounced in complex domains such as Time Series Classification and Vessel Route Forecasting (VRF) [3,4], where AI interpretability is critical yet challenging due to the intricate nature of the data and models involved. Extensive surveys conducted in these areas highlight the inherent subjectivity in explainability, with individuals having varied preferences and understandings. The findings underscore a strong preference for visualization techniques among non-IT end users, such as overlaying predicted versus actual trajectories, indicating the importance of flexible, user-centric approaches in designing explainability communication.

To address these challenges, there is a need for innovative, human-centered XAI approaches that make advanced AI concepts more accessible to a diverse range of users. Leveraging advancements in Large Language Models (LLMs), specifically GPT-based models, presents a promising direction in this endeavor.

This paper aims to operationalize human-centered perspectives in XAI at the conceptual, methodological, and technical levels toward Human-Centered Explainable AI (HC-XAI) models. We enhance cutting-edge XAI approaches for explaining machine learning models and deep neural networks by shaping the final output of black-box models while considering context and biases, allowing for user feedback and adjustment. Our research is motivated by findings from extensive surveys exploring challenges and preferences in XAI, emphasizing the need for flexible, user-centric communication. Our objectives are:

1. Audience-Adaptive Explanations: Develop a model capable of producing concise, easily digestible summaries of complex XAI methods, tailored to align with the varying expertise levels and interests of diverse audience groups, ranging from business professionals to academic researchers. This customization enhances user engagement and understanding across different sectors.

2. XAI Methodology Agnosticism: Create a model with an agnostic approach to XAI methods, ensuring broad applicability and relevance across a wide spectrum of XAI techniques and knowledge domains without necessitating specific training or adaptation for each distinct method.

3. Decision-Making Facilitation: Enhance the model's capacity to provide timely, clear, and contextually relevant explanations to significantly augment decision-making processes for endusers, particularly in scenarios where comprehension of AI outputs is essential but hindered by technical complexity.

4. Empirical Validation through Use-Case Studies: Validate the practical efficacy of the LLM through empirical evidence gathered from use-case studies, demonstrating the model's effectiveness in delivering audience-specific explanations that are comprehensible and relevant. Based on these objectives, we propose an integrative approach that combines the strengths of visual and textual explanations. This approach aims to make XAI results more human-centered by providing user-friendly interfaces such as chatbot-based human-AI

interactions. It ensures that the design of explanations and interfaces aligns with the specific needs and preferences of different user groups. This strategy not only demystifies AI decisions but also enriches the user's understanding by offering context-rich, detailed insights into the AI's decision-making process.

The remainder of the paper is organized as follows: Section 2 presents the background and motivation of our research. Section 3 provides a comprehensive literature review pertinent to this study. Section 4 outlines the proposed methodological approach, introduces the overall implementation, and offers details regarding the datasets used and the evaluation procedure. Section 5 delves into the results of the conducted research and the corresponding survey. Finally, Section 6 concludes the paper with recommendations for future research and discusses the potential impact of the current study.

2. Background

Our research's underlying motivation is illuminated through an introduction to the foundational concepts of Image Classification, XAI methods, and adversarial attacks.

2.1. eXplainable AI (XAI)

Standard Interpretability or explainability in Machine Learning (ML) models refers to the ability to describe and understand an ML model's workings Choo & Liu [4]. This is particularly vital in Deep Neural Networks (DNN), which are inherently complex and thus perceived as "black boxes" Zahavy et al. [5]. The burgeoning field of research addressing the opacity of these ML "black boxes" is known as XAI Gunning [7].

Herein, XAI assumes a critical yet sensitive role, acting as a conduit between intricate DL models and those without IT expertise. Consequently, XAI methodologies must be precise and comprehendible to domain experts, fostering a sense of "trust" in real-time settings. Over the past few years, several XAI methods, strategies, and frameworks have emerged. For our research, we categorize XAI methods based on their simplicity, the degree of interpretability, and the dependency level on the analysed ML/AI model, as illustrated in Figure 1.

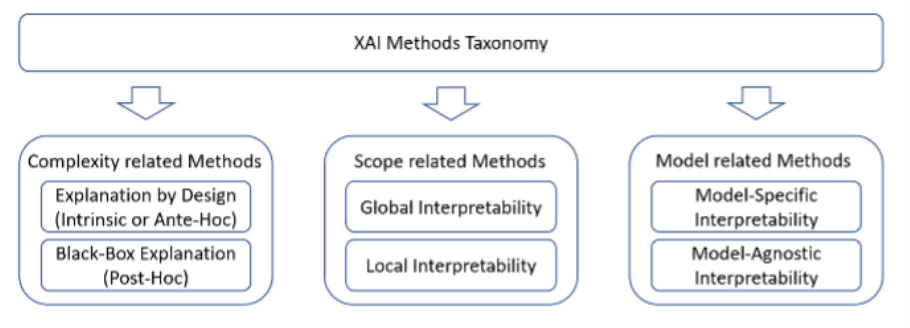

XAI Methods Taxonomy
Complexity related Methods
Explanation by Design (Intrinsic or Ante-Hoc)
Black-Box Explanation (Post-Hoc)
Scope related Methods
Global Interpretability
Local Interpretability
Model related Methods
Model-Specific Interpretability
Model-Agnostic Interpretability

Figure 1: Taxonomy of XAI Methods

Moreover, complexity-related methods in XAI can be bifurcated into i) intrinsically explainable (Ante-Hoc) models, also known as transparent or glass box approaches, and ii) black-box (posthoc) models, which necessitate deciphering the reasoning steps behind predictions for explainability purposes. Additionally, these methods can be categorized based on their scope: i) global explainability methods, which scrutinize the algorithm as a whole, including training data and proper algorithm usage, and ii) Local explainability, which pertains to the system's ability to elucidate specific decision-making processes.

Lastly, it's crucial to differentiate between model-specific and model-agnostic XAI approaches. The key difference lies in whether the XAI method depends on the underlying ML model or if it can be universally applied.

2.2. From Explainability to Interpretability

In scholarly discussions, a notable discrepancy persists regarding the precise definitions of "explainability" and "interpretability." While these terms are often used interchangeably, some scholars distinguish between them, as noted in Arrieta et al. [8] and Chakraborty et al. [9]. This analysis adheres to the differentiation between explainability and interpretability as explicated in Saeed & Omlin [10]. According to this reference, explainability entails the provision of insights tailored to satisfy a specific requirement of a designated audience, while interpretability is concerned with the extent to which these insights are comprehensible and relevant within the framework of the audience's specialized knowledge base.

Explainability is defined by three fundamental elements, as outlined in the aforementioned source: the nature of the insights provided, the specific audience targeted, and the underlying necessity for these insights. These insights emanate from various methodologies in explainability, such as textual descriptions, the significance of features, or localized elucidations, and are intended for a diverse audience including sector-specific professionals, individuals directly impacted by the outcomes of the model, and experts in the field of model development.

In the realm of interpretability, the focus shifts to the congruence and logicality of the explanations about the targeted audience's pre-existing knowledge base. This includes assessing whether the explanations are coherent and meaningful to the audience, if the audience is capable of employing these explanations in their decision-making processes, and whether the explanations provided offer a rational basis for the decisions made by the model.

2.3. Challenges in Communicating AI Concepts

Author Communicating the concepts of AI to a broad audience encompasses a multitude of challenges, stemming from the inherently complex and rapidly evolving nature of AI technology. These challenges are amplified when discussing the domain of XAI, where the goal is to make AI decision-making processes transparent and understandable to various stakeholders.

Although the various available open-source XAI algorithms, such as LIME (Local Interpretable Model-agnostic Explanations) Ribeiro et al. [11], SHAP (SHapley Additive exPlanations) Lundberg & Lee [12], and Gradient-weighted Class Activation Mapping (Grad-Cam) model Selvaraju et al. [13], examples of XAI in real-world applications still need to be discovered. The root cause of this is that SotA XAI algorithms aim to assist the developer of the AI system instead of the end-user. Developing XAI applications needs human-centered approaches that align technical development with people's explainability needs and define success by human experience, empowerment, and trust. Furthermore, AI algorithms can exhibit various forms of bias Klein [14], including social, racial, and gender prejudices. XAI and Exploratory Data Analysis Torralba & Efros [15]). However, implementing bias mitigation and XAI techniques in a larger situational context (i.e., explaining multiple AI models that perform a single task) becomes increasingly more complicated. Cutting-edge XAI approaches are rigorously disconnected, with just a local input view linked with each particular AI model utilized throughout the overall (global) reasoning process Jan et al. [16]. Moreover, existing techniques usually lack reasoning semantics and remain detached from the broader process context.

Other challenges rely significantly on the utilized interface between the human and machine/software. An effective HMI should consider various aspects such as the level of autonomy, user expertise, use case/domain Lim & Dey [17], as well as security and trust Virtue [18]. Despite the extended research, many works suggest that designers need more guidance in designing interfaces for intelligent systems Baxter [19] that could be used by the non-IT-savvy public.

3. LITERATURE REVIEW

The following review synthesizes current research on the critical aspects of artificial intelligence, with a particular emphasis on transparency and interpretability. This foundation provides a comprehensive understanding of the evolving landscape of Explainable AI.

3.1. Explainable AI: A Technical Overview

As complex predictive models are increasingly integrated into areas traditionally governed by human judgement, there is a growing demand for these models to offer more clarity in how they reach decisions Susnjak [20]. This transparency is vital for building trust and meeting regulatory compliance, especially in international legal contexts where explaining automated decisions affecting people is becoming a legal necessity. According to Wachter et al. [21], it’s also crucial that individuals can challenge decisions made by these systems and understand what changes in their data could lead to different outcomes. Technologies like Counterfactuals have been developed to provide insights into minimal changes needed for different predictions by these models.

This need for clarity has given rise to the field of XAI or Interpretable Machine Learning. This area aims to create methods that make complex predictive models more understandable and tools that explain how these models formulate their conclusions (Molnar et al. [22]). Additionally, there’s growing interest in prescriptive analytics, which focuses on using data to create actionable insights Lepenioti et al. [23].

From a technical standpoint, model interpretability involves understanding the internal workings of a machine learning model post-training, generally at a broad, global level. Conversely, model explainability delves into understanding the rationale behind a model’s prediction for a specific instance, known as local level explainability. Both are important: interpretability allows institutions to broadly explain how a model works to stakeholders, while local-level explainability facilitates validating specific predictions and providing detailed feedback to those affected, like students identified as at-risk.

In the pursuit of model transparency, tools like SHAP, recognized as a leading visualization technique in XAI, provide insight into both global and local-level transparency Gramegna & Giudici [24]. The Anchors technique (Ribeiro et al. [25]) offers a high degree of local-level explainability through human readable, rule-based models. Furthermore, advanced Counterfactuals not only enhance predictive analysis but also enable prescriptive suggestions, helping learners understand the changes needed for a different outcome. This study showcases the application of these technologies across various stages of the proposed prescriptive analytics framework.

3.2. Language Models in AI

Following the success of GPT, a range of LLMs have been developed, exhibiting impressive capabilities in various Natural Language Processing (NLP) tasks, including those in finance. One standout model in this domain is BloombergGPT, created by Bloomberg's AI team and trained on an extensive collection of financial texts. It has shown exceptional proficiency in financial NLP tasks (Wu et al. [26]). However, as of May 2023, BloombergGPT remains largely for internal use at Bloomberg, lacking a publicly accessible API. Google's Bard, a key competitor to ChatGPT, is another notable LLM. Powered by Google's LAMDA (Language Model for Dialogue Applications), it merges aspects of BERT and GPT to facilitate engaging, contextually aware conversations (Thoppilan et al. [27]). Like BloombergGPT, Bard also doesn't offer an open API as of this writing. BLOOM, an open-source contender to GPT-3 (Scao et al. [28]), has also gained attention in the LLM space. While it's open-source, effectively using BLOOM requires considerable technical know-how and computing power, and it lacks a version fine-tuned for conversational tasks, a feature where models like ChatGPT excel.

Since ChatGPT's introduction, numerous LLMs have emerged targeting specific functions, such as code completion (Dakhel et al. [29]), content generation, and marketing. These models offer specialized utility, expanding the scope and impact of LLMs. ChatGPT continues to lead in the field (JasperAI [30]), thanks to its open API, extensive training data, and versatility across various tasks. Despite ChatGPT's broad application in fields like healthcare and education (Sallam [31]), its direct use in financial sentiment analysis is relatively uncharted. Fatouros et al. [32] presents evidence that ChatGPT, even when applied with zero-shot prompting, can understand complex contexts requiring advanced reasoning capabilities. In addition, MarketSense-AI, a real-world financial application, leverages GPT-4 with Chain-of-Thought (CoT) to effectively explain investment decisions Fatouros et al. [33].

3.3. Large Language Models in XAI

Significant advancements have been made in AI and LLMs based on transformers, which now exhibit near-human proficiency in text generation and discourse. This progress is largely attributed to their ability to understand longrange dependencies and contextual nuances in texts, thanks to self-attention mechanisms. Models like Google's BERT (Devlin et al. [34]) and OpenAI's latest GPT series have set new benchmarks in various natural language processing tasks, including text generation (Brown et al. [35]). OpenAI's most recent development, the ChatGPT model, exemplifies these advancements by effectively translating complex analytical outputs into user-friendly, actionable language, aiding learners and advisors.

In the realm of cybersecurity, HuntGPT utilizes the capabilities of LLMs and XAI to enhance network anomaly detection. It integrates a Random Forest classifier with the KDD99 dataset Stolfo et al. [36], advanced XAI frameworks, and the power of GPT-3.5 Turbo. HuntGPT not only detects threats with remarkable accuracy but also conveys them in a clear, understandable format, greatly improving decision-making for cybersecurity experts Ali & Kostakos [37]. While, Chun & Elkins [38] delves into the fusion of XAI with Computational Digital Humanities. It investigates diachronic text sentiment analysis and narrative generation using advanced LLMs like GPT-4. Additionally, it introduces an innovative XAI grey box ensemble. This ensemble combines top-tier model performance with superior interpretability and privacy, underpinned by novel local and global XAI metrics.

4. METHODOLOGY

Initially, an introduction to state-of-the-art XAI techniques, such as LIME, SHAP, and GradCAM, will take place. These methods will then be adapted and integrated into a customized LLM infrastructure, focusing on generating natural language explanations delivered via the AI chat interface. This integration aims to transform complex XAI visualizations into user-friendly narratives and insights that are easily interpretable by end users.

4.1. Baseline XAI approaches

Our innovative approach in Explainable AI (XAI) builds upon existing methods by integrating outputs from LIME, SHAP, and Grad-CAM into a GPT-powered framework that enhances both textual and data analysis. This combined methodology leverages the strengths of each tool to provide more comprehensive and interpretable insights.

LIME (Local Interpretable Model-Agnostic Explanations) works by approximating complex models with simpler, interpretable models on a local scale. It highlights which features are most influential in driving specific predictions. However, while LIME offers valuable insights, it can suffer from instability due to its reliance on local approximations, which may not always generalize well across the entire dataset.

SHAP (SHapley Additive exPlanations) is grounded in game theory and provides a global assessment of feature importance. It offers consistent and theoretically sound interpretations by distributing the prediction's value among all features in a manner akin to the Shapley value in cooperative game theory. SHAP’s interpretations are generally considered fair and unbiased, but they come with a trade-off in terms of computational cost and can sometimes be challenging to intuitively understand due to their complex nature.

Grad-CAM (Gradient-weighted Class Activation Mapping) is a visualization technique used primarily in the context of convolutional neural networks for image classification tasks. It identifies and highlights important regions in an image that contribute to the model's predictions, making the decision process more transparent. Grad-CAM is appreciated for its ability to maintain a high degree of interpretability while improving both the versatility and accuracy of model explanations.

4.2. Role of GPT-Builder in LLM Development

The development of LLMs such as GPT variants has revolutionized the field of natural language processing (NLP). A critical component in this evolution is the role of tools like GPT-Builder [39], a sophisticated framework for constructing, custom-purpose GPT models. GPT-Builder itself is a GPT model created by OpenAI with instructions and an action that allows it to write (update) to the fields of the GPT that is being built (by the user). GPT-Builder serves as a pivotal element in LLM development, offering a blend of user-friendly interfaces and powerful backend processes that streamline the creation and management of these complex models GPT Builder.

On top, GPT-Builder plays an instrumental role in democratizing access to LLM technology. It empowers organizations and individual developers to build custom LLMs tailored to specific needs or domains. This customization is crucial in scenarios where a standard GPT model may not provide optimal performance, such as in specialized professional fields or for languages and dialects with limited representation in mainstream models. Although the foundational model used is not being finetuned or retrained, since the only thing changing is the behavior, GPT-Builder simplifies the process of enabling in a robust and strict manner desirable behaviors and answer patterns.

4.3. Use Cases

In our study, we developed a custom GPT model, the x-[plAIn] GPT which can be found here. x[plAIn] model underwent extensive testing across a diverse range of XAI methods and problem definitions. However, for the interactive component involving end users, we focused on five specific use cases presented through a questionnaire. We made a concerted effort to select XAI implementations that spanned various sectors and catered to different levels of technical expertise.

4.3.1. Use Case 1

The first use-case featured in our study was derived from Makridis et al. [40] which investigated the detection of boar taint. In this research, the authors identified significant factors contributing to the boar-taint phenomenon, employing SHAP values among other methods. This particular implementation of SHAP values was incorporated into our questionnaire shown in Figure 2.

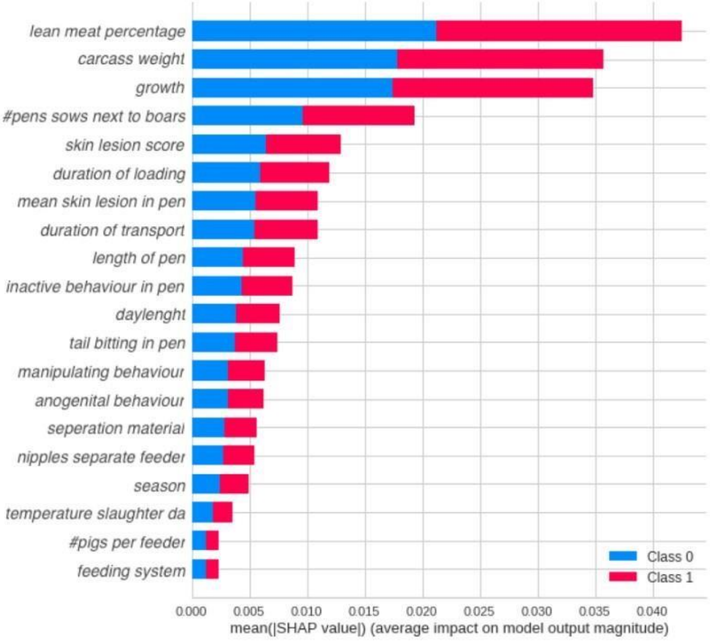
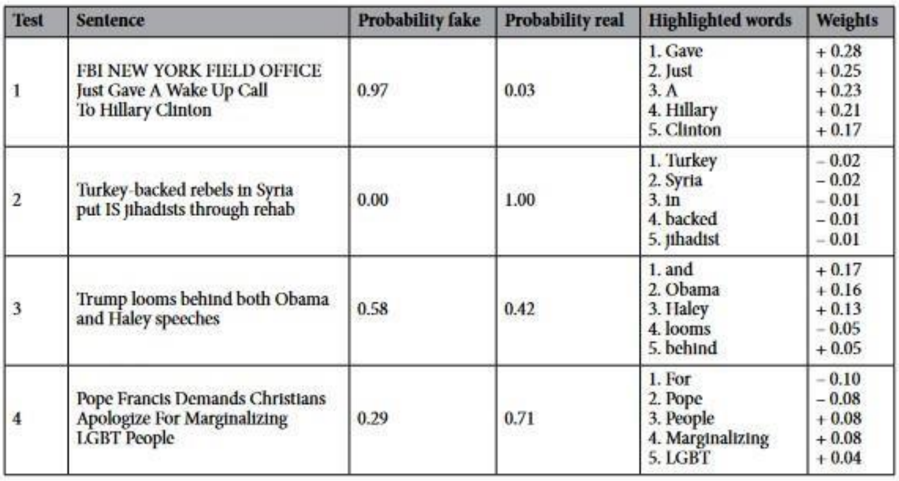


Figure 2: Plot for SHAPley values evaluation in Makridis et al. [40].

4.3.2. Use Case 2

The second use-case was based on Szczepanski et al. [41] where the authors explored the use of LIME and Anchors (XAI methods) for generating explainable visualizations in the context of fake news detection. This study represented another facet of XAI application, showcasing its utility in media and information analysis. Figure 3 demonstrates the output of the applied method in a set of sentences.

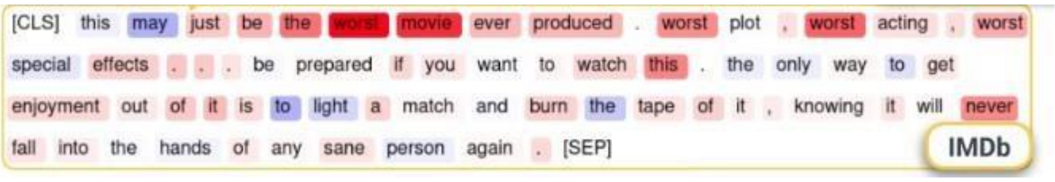


Test	Sentence	Probability fake	Probability real	Highlighted words	Weights
1	FBI NEW YORK FIELD OFFICE Just Gave A Wake Up Call To Hillary Clinton	0.97	0.03	1. Gave 2. Just 3. A 4. Hillary 5. Clinton	+ 0.28 + 0.25 + 0.23 + 0.21 + 0.17
2	Turkey-backed rebels in Syria put IS jihadists through rehab	0.00	1.00	1. Turkey 2. Syria 3. in 4. backed 5. jihadist	− 0.02 − 0.02 − 0.01 − 0.01 − 0.01
3	Trump looms behind both Obama and Haley speeches	0.58	0.42	1. and 2. Obama 3. Haley 4. looms 5. behind	+ 0.17 + 0.16 + 0.13 − 0.05 + 0.05
4	Pope Francis Demands Christians Apologize For Marginalizing LGBT People	0.29	0.71	1. For 2. Pope 3. People 4. Marginalizing 5. LGBT	− 0.10 − 0.08 + 0.08 + 0.08 + 0.04

Figure 3: Plot for LIME evaluation of fake news from Szczepanski et al. [41].

4.3.3. Use Case 3

The third use-case in our study involved the visualizations developed by Feldhus et al. [42]. The authors employed the Integrated Gradients feature attribution method to represent the predictions made by a BERT model. Building on this, they created a model-free and instructed (GPT-3.5) Saliency Map Verbalization (SMV) explaining the prediction representations. Figure 4 shows a depiction of the SMV, where words that seem more important are highlighted in shades of red.



[CLS] this may just be the worst movie ever produced . worst plot , worst acting , worst special effects . . . be prepared if you want to watch this . the only way to get enjoyment out of it is to light a match and burn the tape of it , knowing it will never fall into the hands of any sane person again . [SEP]

Figure 4: Plot for Saliency Map Verbalization (SMV) from Feldhus et al. [42].

4.3.4. Use Case 4

The fourth application incorporated XAI techniques as implemented by Moujahid et al. [43]. In their study, Grad-CAM was employed to identify regions of interest pertinent to the prediction of COVID-19 in lung X-ray images, utilizing various network architectures. Figure 5 shows a sample of the result of Grad-CAM to the X-rays where regions of importance are highlighted.

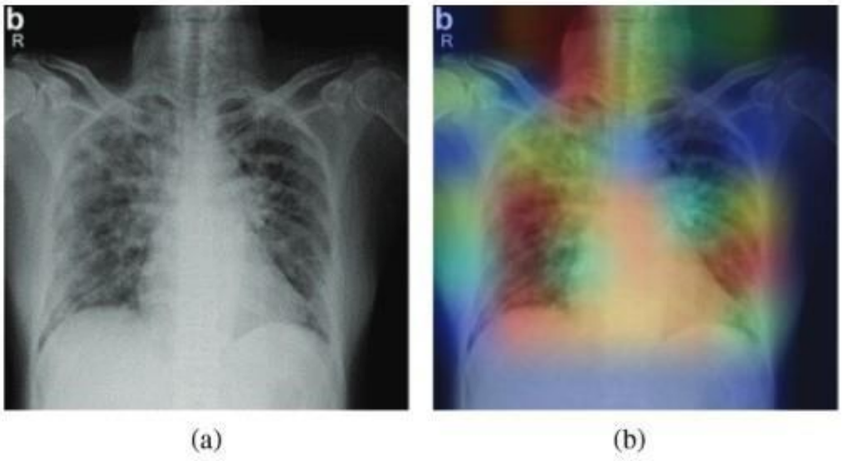


Figure 5: Plot for Grad-CAM explanations from Moujahid et al. [43].

4.3.5. Use Case 5

Lastly, the fifth use-case presented substantial technical complexities. The researchers in Moosbauer et al. [44] employed Partial Dependence Plots (PDP), an infrequently used method within XAI, for the purpose of hyperparameter optimization. Consequently, they generated and scrutinized plots to exhibit robust and trustworthy Partial Dependence (PD) estimates across an intelligible subset of the hyperparameter space, considering a variety of model parameters.

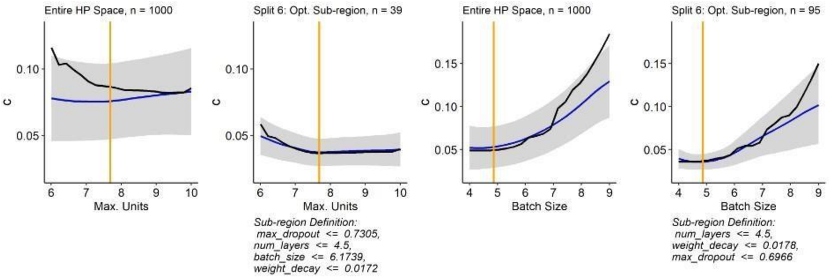


Figure 6: PDP (blue) and confidence band (grey) of the Gaussian Process for hyperparameterhyperparameter max. number of units (batch size) on the left (right) side. The black line shows the PDP of the meta surrogate model representing the true PDP estimate

4.4. LLM Enhanced XAI Explainer

In the development of our GPT-based XAI explainer, a rigorous, technical approach was employed. Initially, we utilized a specialized interface for defining the explainer's core objectives, focusing on advanced interpretability for AI decision-making processes. The configuration phase was comprehensive, involving precise customization of the model's parameters, including its naming, operational descriptions, and initialization prompts tailored for nuanced AI explanations. The development of the prompt engineering process is delineated in Table 1. This progression adheres to the prompt-engineering guidelines provided in the official documentation of OpenAI, accessible via this hyperlink. Additionally, acknowledging the paramount significance of causal explanations and insight generation for HC-XAI, we incorporated a CoT approach into our final prompt. Given the absence of universally correct responses for this task, we eschewed the methodology detailed in Wei et al. [45]. Instead, we adopted a more straightforward strategy, integrating the phrase "Let's think step by step." This inclusion has proven to be notably effective, as substantiated by Tan [46]. This technical methodology ensured the creation of a GPT model specifically fine-tuned for the complexities and demands of XAI, enhancing its effectiveness in delivering clear, understandable insights into AI decisions.

Table 1: Gradual Development of ChatGPT Prompts for XAI Method Simplification

Prompt Version	**Prompt**	**Benefit Over Previous Version**
P1	Provide summaries of insights from XAI methods, focusing on clarity and relevance. Avoid including obvious elements unless specifically requested.	Introduces the basic concept of summarizing XAI insights with a focus on clarity and relevance.
P2	Summarize insights from XAI methods like LIME and SHAP. Focus on clarity and relevance, and begin to consider the user's context in your summaries. Exclude obvious elements unless they are explicitly asked for.	Specifies XAI methods (LIME and SHAP) and introduces the concept of tailoring summaries to the user's context.
P3	Generate clear and relevant summaries from XAI methods such as LIME and SHAP, tailored to the user's context. Begin to integrate actionability into the insights and ask the user for their expertise level (beginner, intermediate) before responding.	Adds the element of actionability and the need to adjust responses based on the user's expertise level.
P4	Provide clear, relevant, and actionable summaries of insights from XAI methods like LIME and SHAP. Tailor the content to the user's expertise level and specific inquiries. Include practical suggestions or conclusions and avoid obvious elements from the input unless requested.	Emphasizes the customization of summaries to specific user inquiries and the inclusion of practical suggestions or conclusions.
P5	Objective: Deliver concise summaries of insights from XAI methods like LIME and SHAP, tailored to user's context and expertise level. Focus on clarity, relevance, actionability, and responsiveness. Ask the user for their expertise level and tailor your response accordingly. Avoid including obvious input elements unless explicitly asked. Provide practical suggestions or conclusions.	Introduces the concept of responsiveness to user-specific inquiries and further emphasizes tailoring content based on expertise.
P6	Objective: Provide concise, userfriendly summaries of insights derived from XAI methods. If multiple insights can be drawn from a single input try to combine them into a larger context. Let's think step by step about how the final insight is reached. Output Expectations: Clarity: Deliver straightforward and easily comprehensible summaries. Relevance: Ensure insights are directly applicable	Fully integrates all elements including clarity, relevance, actionability, responsiveness, and user-specific customization, creating a comprehensive and detailed approach to summarizing XAI insights.

	to the user’s context or domain. Actionability: Focus on providing practical suggestions or conclusions. Responsiveness: Tailor summaries to answer user-specific inquiries based on the XAI analysis. DO NOT include obvious elements (numbers, text) from the given input unless EXPLICITLY asked. BEFORE ANSWERING 1. Ask the user for his expertise level (beginner, intermediate). 2. Ask the user’s domain of expertise (if none is provided assume it aligns with the domain of the input provided) If the user is intermediate provide information about how he should understand the provided input and then the insight(s). If the user is a beginner DO NOT provide information about the provided input PROVIDE ONLY THE INSIGHT. Tailor responses to a user’s technical and domain expertise. Provide examples that match the expertise of the user in analogous manner to explain the insights.	

4.5. Audience Analysis and Content Customization

This tool primarily serves two key demographics: end-users of XAI methods and AI developers, notably data scientists, who utilize XAI methods for model understanding. The former category, end-users, typically possesses limited technical knowledge but may exhibit considerable domain expertise. Their primary interest lies in deriving insights from the XAI methods, rather than comprehending the technical intricacies of how these methods operate or the underlying model training processes. Conversely, highly technical users, such as AI developers, leverage this tool to gain deeper insights into their models. They focus on understanding the training mechanisms of the models, identifying potential biases highlighted by the XAI methods, and exploring strategies to address these issues.

The design of this tool is interactive and user-centric, enabling it to evaluate the user’s proficiency in AI and XAI methodologies, as well as any domain specific expertise they may possess. Following this assessment, the tool adeptly tailors its responses, adjusting the focus of its answers to align with the user’s knowledge level. This approach ensures that the insights provided are not only relevant and actionable but are also derived effectively from the input given by the user.

4.6. Evaluation - Feedback

In our comprehensive study to evaluate the applicability and effectiveness of our GPT-based XAI explainer, we conducted an extensive survey targeting a broad spectrum of professionals. This survey, which can be accessed here, was designed to gather insights into the various aspects of XAI in the context of our GPT model.

Survey Design and Purpose: The survey was meticulously structured to probe into the respondents’ understanding and experiences with AI, Machine Learning (ML), and Deep Learning (DL), as well as their exposure to and perceptions of XAI methods. This allowed us to gauge the baseline knowledge of our audience, which is crucial in tailoring the XAI components of our GPT model.

Assessing User Familiarity and Application of AI: One of the key objectives was to understand how familiar the respondents were with AI, especially in the context of using AI for specific tasks. This information is vital to ensure that our GPT XAI explainer is accessible to users with varying levels of AI expertise.

Understanding Preferences in Data Description: The survey extensively examined various applications of XAI methods, including LIME, SHAP, and GradCAM, each presented with two distinct descriptions. The first was the original description from the research papers, selectively modified to provide the end user with essential information. In contrast, the second description was generated by the x-[plAIn] GPT model in response to the query, ”What are the top insights from this picture?” Notably, in instances where the problem definition was not evident or deducible from the input, the model's query included the specific research problem addressed in the original paper. This approach enabled a comprehensive understanding of user preferences regarding textual explanations, including aspects like structure, length, and formality, allowing for subsequent fine-tuning of the model.

4.7. Limitations

This tool demonstrates a remarkable ability to interpret various outputs from XAI methods, offering insightful and targeted explanations. However, it has been observed that there are instances in which the model mistakenly attempts to explain the provided image rather than focusing on the XAI output. An illustrative case of this behavior can be found in use case #3 (SMV), as detailed in Section 4.2. In this particular example, the model states: Highlighted High-Impact Negative Phrases:

- The phrases “the worst movie ever produced”, “worst plot”, “worst acting” and “worst special effects” are strongly emphasized in the saliency map. This implies that these phrases are key elements the model associates with a negative review.
- The recommendation to “light a match and burn the tape” is exceedingly negative, indicating a high level of dissatisfaction.

While the second observation remains accurate and provides valuable insights, it is not directly related to the actual XAI output since those words are not prominently highlighted by the saliency map.

In general, the model's contextual interpretation can yield generalized insights that may extend beyond the strict confines of the XAI output. Striking the right balance between the precision of the response and the breadth of the insights provided poses a challenge. The most effective approach is to engage an active end-user who employs critical thinking and maintains an open minded approach to understanding the results.

5. RESULTS AND DISCUSSION

In our endeavour to delve into the usability and effectiveness of x-[plAIn], we administered a meticulously crafted survey to an eclectic mix of partners and participants, that can be found here. Drawing from real-world scenarios, we simulated a context where AI models transition from mere decision-support mechanisms to primary decision-makers, emphasizing their paramount need for transparency and trustworthiness. The survey pivoted on two primary axes: gauging participants' baseline familiarity with AI, ML, and DL; and discerning their perception of key data interpretation based on their experience on AI enhanced decision-making. By gathering feedback through this structured lens, we aimed to carve out a roadmap for the subsequent development and refinement of x-[plAIn] that we plan to offer via the GPT Store.

Through the conducted questionnaire, it was revealed that a significant majority of participants, exceeding 70%, expressed a satisfaction level below 60% concerning their comprehension of AIbased decision models. This is further underscored by the fact that a mere 30% of participants are actively employing XAI methodologies. This finding raises critical questions about the prevalence and perceived efficacy of XAI techniques within the industry. When it comes to scenario-based preferences, over 80% of participants favored x-[plAIn] descriptions in comparison to the conventional descriptions derived from original papers associated with XAI methods, particularly in decision-making contexts and image comprehension.

The feedback on enhancing x-[plAIn] predominantly revolved around the brevity of responses, given the tendency of GPT models towards verbosity, and the need for tailoring explanations to suit the specific background or domain of the end-user. While the model inherently possesses the capability to customize responses based on domain-specific information, this feature was not fully showcased to the participants due to the absence of domain information in the prompts, which was intentionally omitted to ensure a level playing field in the assessment. This suggests a potential area for refinement in future iterations, where the model's adaptive response generation can be demonstrated more effectively to respondents.

The comparative analysis of the two bar plots reveals insightful trends about the acceptability of "x-[plAIn]" across different user groups and their engagement with XAI methods. In Figure 7a, which distinguishes between end users and AI experts, a distinct pattern emerges. End users, presumably less versed in the technical aspects of AI, show varying levels of preference for "x[plAIn]" across different use cases. This variability could indicate a nuanced approach to AI explanations, where the complexity or context of each use case significantly influences their preference. On the other hand, AI experts, with their deeper technical understanding, exhibit a more consistent response pattern across the use cases. Their preferences might reflect a critical evaluation of "x-[plAIn]" against their advanced understanding of AI processes. Figure 7b, focuses on the usage of XAI methods and also provides compelling insights. Respondents who actively use XAI methods demonstrate a certain level of preference for "x-[plAIn]", which might suggest that their familiarity with XAI influences their expectations and acceptance of explanatory tools. Conversely, non-XAI users, who might not have a benchmark for comparing such tools, show differing degrees of acceptance for "x-[plAIn]", potentially guided more by the tool's clarity and usability than by its technical robustness.

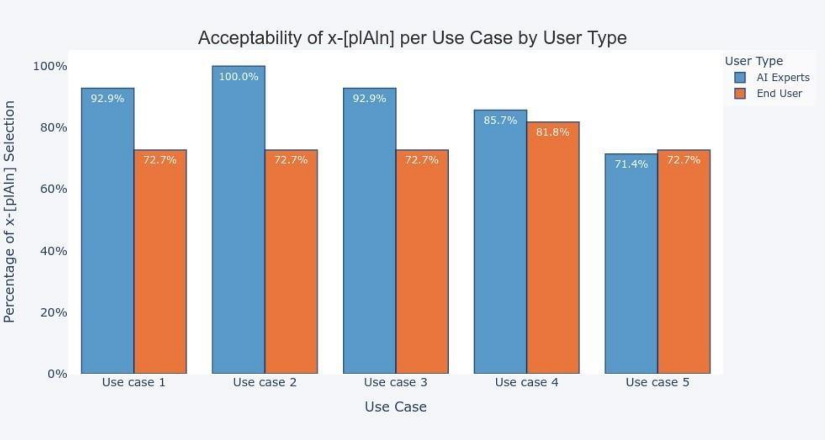


Figure 7 (a) Acceptability of x-[plAIn] concerning role of the users.

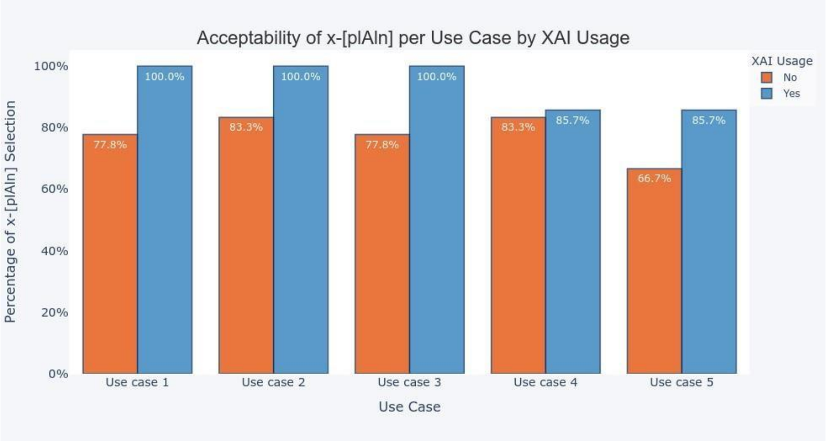


Figure (b) Acceptability of x-[plAIn] concerning the usage of XAI of the users.

Figure 8 demonstrates a notable trend concerning users' preferences in relation to their selfreported comprehension of AI model outputs. This graphical representation indicates a discernible correlation between the level of claimed understanding and the preferences exhibited by respondents. It appears that individuals professing a more profound grasp of AI model outputs tend to require less information, potentially influencing a shift in their preferences. Despite this observed trend, it is noteworthy that x-[plAIn] retains a significant degree of favorability, being the preferred choice in 75% of instances among the cohort exhibiting the highest level of understanding.

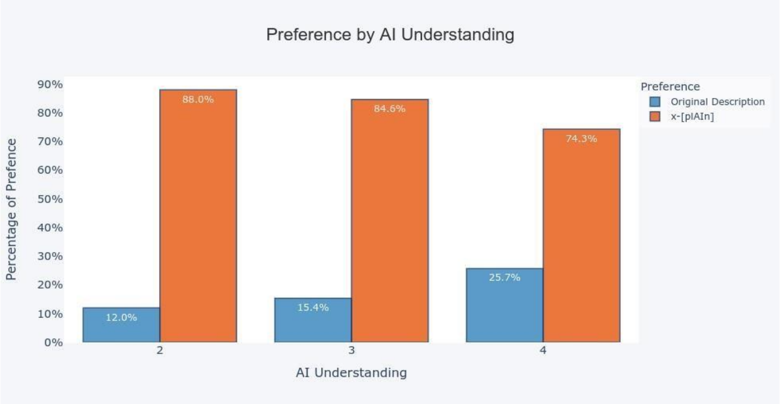


Figure 8: Description preference based on perceived AI understanding level.

6. CONCLUSION

For future enhancements, it will be crucial to implement features that allow end-users to specify (more strictly) their preference for the level of detail they require. A verbose setting could cater to those looking for in-depth understanding, while a more streamlined option would benefit users seeking brief clarifications. Such a choice empowers users, providing a user-centric approach that accommodates a wide range of use cases from novice inquiries to expert validations.

Additionally, the feedback highlights the importance of considering user experience, particularly for those unfamiliar with the subject matter. Breaking down complex topics into smaller, individually explained segments can significantly enhance comprehension. Conversely, for experienced users, lengthy and information-dense responses may prove unnecessary and timeconsuming. To this end, introducing an option to toggle between longer and shorter answer formats while shifting focus from understanding the XAI methods to the extraction of insights can be beneficial.

In future work, we aim to investigate a specific characteristic of this tool, identified during the development of x-[plAIn]. This tool holds potential for experienced AI engineers, offering a resource to pinpoint and mitigate potential biases that may emerge at different phases of the model creation pipeline. These phases include data collection, preprocessing, model training, and validation processes. By utilizing this tool, AI professionals can significantly contribute to the cultivation of AI systems that excel not only in technical prowess but also in ethical integrity and social responsibility.

ACKNOWLEDGEMENTS

The research leading to the results presented in this paper has received funding from the Europeans Union's funded Project HumAIne under grant agreement no 101120218.

REFERENCES

[1] G. Makridis, D. Kyriazis, and S. Plitsos, "Predictive maintenance leveraging machine learning for time-series forecasting in the maritime industry," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems(ITSC)*, pp. 1–8, IEEE, 2020.

[2] J. Soldatos and D. Kyriazis, "Trusted artificial intelligence in manufacturing; trusted artificial intelligence in manufacturing: A review of the emerging wave of ethical and human centric ai technologies for smart production; a review of the emerging wave of ethical and human centric ai technologies for smart production," 2021.

[3] G. Makridis, G. Fatouros, V. Koukos, D. Kotios, D. Kyriazis, and I. Soldatos, "Xai for time-series classification leveraging image highlight methods," arXiv preprint arXiv:2311.17110, 2023.

[4] G. Makridis, G. Fatouros, A. Kiourtis, D. Kotios, V. Koukos, D. Kyriazis, and J. Soldatos, "Towards a unified multidimensional explainability metric: Evaluating trustworthiness in ai models," in 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), pp. 504–511, IEEE, 2023.

[5] J. Choo and S. Liu, "Visual analytics for explainable deep learning," IEEE computer graphics and applications, vol. 38, no. 4, pp. 84–92, 2018.

[6] T. Zahavy, N. Ben-Zrihem, and S. Mannor, "Graying the black box: Understanding dqns," in International conference on machine learning, pp. 1899– 1908, PMLR, 2016.

[7] D. Gunning, "Explainable artificial intelligence (xai) darpa-baa-16-53," Defense Advanced Research Projects Agency, 2016.

[8] A. B. Arrieta, N. D´ıaz-Rodr´ıguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garc´ıa, S. Gil-L´opez, D. Molina, R. Benjamins, et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," Information fusion, vol. 58, pp. 82–115, 2020.

[9] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, Preece, S. Julier, R. M. Rao, et al., "Interpretability of deep learning models: A survey of results," in 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), pp. 1–6, IEEE, 2017

[10] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," Knowledge-Based Systems, vol. 263, p. 110273, 2023.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135– 1144, 2016.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

[14] A. Klein, “Reducing bias in ai-based financial services,” 2020.

[15] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in CVPR 2011, pp. 1521–1528, IEEE, 2011.

[16] S. T. Jan, V. Ishakian, and V. Muthusamy, “Ai trust in business processes: the need for processaware explanations,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13403–13404, 2020.

[17] B. Y. Lim and A. K. Dey, “Assessing demand for intelligibility in contextaware applications,” in Proceedings of the 11th international conference on Ubiquitous computing, pp. 195–204, 2009.

[18] E. Virtue, “Designing with ai,” Retrieved July, vol. 29, p. 2022, 2017.

[19] K. Baxter, “How to meet user expectations for artifcial intelligence,” Medium. Retrieved September, 2018.

[20] T. Susnjak, “Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and chatgpt,” International Journal of Artificial Intelligence in Education, pp. 1–31, 2023.

[21] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” Harv. JL & Tech., vol. 31, p. 841, 2017.

[22] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable machine learning–a brief history, stateof-the-art and challenges,” in Joint European conference on machine learning and knowledge discovery in databases, pp. 417–431, Springer, 2020.

[23] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas, “Prescriptive analytics: Literature review and research challenges,” International Journal of Information Management, vol. 50, pp. 57–70, 2020.

[24] A. Gramegna and P. Giudici, “Shap and lime: an evaluation of discriminative power in credit risk,” Frontiers in Artificial Intelligence, vol. 4, p. 752558, 2021.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision modelagnostic explanations,” in Proceedings of the AAAI conference on artificial intelligence, vol. 32, 2018.

[26] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, “Bloomberggpt: A large language model for finance,” arXiv preprint arXiv:2303.17564, 2023.

[27] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al., “Lamda: Language models for dialog applications,” arXiv preprint arXiv:2201.08239, 2022.

[28] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ili´c, D. Hesslow, R. Castagn´e, A. S. Luccioni, F. Yvon, M. Gall´e, et al., “Bloom: A 176b-parameter open-access multilingual language model,” arXiv preprint arXiv:2211.05100, 2022.

[29] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, and Z. M. Jiang, “Github copilot ai pair programmer: Asset or liability?,” Journal of Systems and Software, p. 111734, 2023.

[30] JasperAI, “The ai in business trend report,” 2023. Accessed:May 26, 2023.

[31] M. Sallam, “Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns,” in Healthcare, vol. 11, p. 887, MDPI, 2023.

[32] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, “Transforming sentiment analysis in the financial domain with chatgpt,” Machine Learning with Applications, vol. 14, p. 100508, 2023.

[33] G. Fatouros, K. Metaxas, J. Soldatos, and D. Kyriazis, “Can large language models beat wall street? unveiling the potential of ai in stock selection,” arXiv preprint arXiv:2401.03737, 2024.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[35] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are fewshot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[36] Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan, “Kdd cup 1999 data.” UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C51C7N

[37] Ali and P. Kostakos, “Huntgpt: Integrating machine learning-based anomalydetection and explainable ai with large language models (llms),” arXiv preprint arXiv:2309.16021, 2023.

[38] J. Chun and K. Elkins, “explainable ai with gpt4 for story analysis and generation: A novel framework for diachronic sentiment analysis,” International Journal of Digital Humanities, pp. 1–26, 2023.

[39] OpenAI, "GPT Builder," OpenAI Help Center, https://help.openai.com/en/articles/8770868-gptbuilder. Accessed Sep. 2, 2024.

[40] G. Makridis, E. Heyrman, D. Kotios, P. Mavrepis, B. Callens, R. Van De Vijver, J. Maselyne, M. Aluw´e, and D. Kyriazis, “Evaluating machine learning techniques to define the factors related to boar taint,” Livestock Science, vol. 264, p. 105045, 2022.

[41] Szczepanski, M. Pawlicki, R. Kozik, and M. Chora´s, “New explainabilitymethod for bert-based model in fake news detection,” Scientific reports, vol. 11, no. 1, p. 23705, 2021.

[42] Feldhus, L. Hennig, M. Nasert, C. Ebert, R. Schwarzenberg, and S. Mller,“Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods,” in Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE), pp. 30–46, 2023.

[43] H. Moujahid, B. Cherradi, M. Al-Sarem, L. Bahatti, A. B. A. M. Y. Eljialy, A. Alsaeedi, and F. Saeed, “Combining cnn and grad-cam for covid-19 disease prediction and visual explanation.,” Intelligent Automation & Soft Computing, vol. 32, no. 2, 2022.

[44] J. Moosbauer, J. Herbinger, G. Casalicchio, M. Lindauer, and B. Bischl, “Explaining hyperparameter optimization via partial dependence plots,” Advances in Neural Information Processing Systems, vol. 34, pp. 2280–2291, 2021.

[45] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., “Chain-ofthought prompting elicits reasoning in large language models,” Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837, 2022.

[46] J. T. Tan, “Causal abstraction for chain-of-thought reasoning in arithmetic word problems,” in Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pp. 155–168, 2023.