Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.7392
This work is licensed under a Creative Commons Attribution 4.0 International License.

Multimodal Context-Aware XAI: Enhancing Explainability through VLMs and Robust Evaluation

Renato Okabayashi Miyaji [**Escola Politécnica da Universidade de São Paulo** | ***re.miyaji@usp.br***]
Arthur Thomas Luz [**Escola Politécnica da Universidade de São Paulo** | ***athomasluz@gmail.com***]
Pedro Luiz Pizzigatti Corrêa [**Escola Politécnica da Universidade de São Paulo** | ***pedro.correa@usp.br***]

Escola Politécnica, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 380 - Butantã, São Paulo - SP, 05508-010, Brazil.

Received: 14 January 2026 • **Published:** 27 November 2026

Abstract Explainable AI (XAI) techniques are essential for model interpretability but often present a technical gap that limits accessibility for non-expert users who struggle with complex numerical and graphical outputs. This study proposes an enhanced multimodal multi-agent architecture designed to transform technical XAI data into human-centric narratives. Our system integrates Retrieval-Augmented Generation (RAG) and real-time web search to ground explanations in domain-specific knowledge. A key advancement is the integration of Vision-Language Models (VLMs) to directly interpret visual artifacts from SHAP and LIME. Experimental results on the UCI Adult dataset demonstrate that our approach improves Context Awareness by 6.5% compared to standard LLM baselines. Furthermore, we validate the LLM as a Judge paradigm using the G-Eval framework with Chain-of-Thought reasoning, achieving over 80% correlation with human expert evaluations. These results highlight the effectiveness of multimodal multi-agent systems in bridging the gap between technical transparency and functional understanding in high-stakes environments.

Keywords: Explainable Artificial Intelligence (XAI), Large Language Models (LLMs), Vision Language Models (VLMs), Machine Learning

1 Introduction

Most modern Machine Learning (ML) models that achieve high predictive performance are often classified as "black-box" models, particularly in the case of neural networks [Spitzer *et al*., 2024]. At best, some models, such as decision trees and Support Vector Machines (SVMs), are considered interpretable [Doran *et al*., 2017]. The lack of transparency in these models raises significant concerns, especially in high-stakes domains where decision-making processes must be explainable and accountable [Arrieta *et al*., 2020; Ryo *et al*., 2021]. Consequently, efforts in Explainable Artificial Intelligence (XAI) have gained traction to bridge this gap by providing mechanisms that offer insights into how models arrive at their predictions [Doran *et al*., 2017].

Even with popular explanation techniques such as feature importance [James *et al*., 2013], users still require deep technical knowledge in ML to fully understand these model provisions [Bhatt *et al*., 2020]. This technical barrier limits the accessibility of ML explanations, making it challenging for non-expert users to effectively leverage the insights provided by these models. To address this issue, the research community has increasingly focused on developing techniques that enhance the interpretability of ML models without requiring extensive domain expertise [Arrieta *et al*., 2020].

More recently, XAI techniques such as Shapley Additive Explanations (SHAP) [Lundberg and Lee, 2017] and Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro *et al*., 2016] have been widely adopted to provide local and global model interpretability. However, understanding the output of these techniques still demands significant technical knowledge. Non-technical users often struggle to interpret correctly not only the numerical values but also the visual representations—such as force plots and summary charts—provided by these methods [Arrieta *et al*., 2020].

This challenge is particularly relevant for end-users who are domain experts but not necessarily ML specialists. For example, medical professionals rely on ML models for disease diagnosis [Borys *et al*., 2023], financial analysts use predictive models for risk assessment [Černevičienė and Kabašinskas, 2024], and legal experts consult AI-based tools for case analysis [Richmond *et al*., 2024]. In these cases, an intuitive and human-centric explanation system is crucial, as traditional XAI methods may not sufficiently bridge the gap between ML-generated insights and user understanding. Without proper interpretability, the adoption of ML in these domains remains restricted [Ryo *et al*., 2021].

Given these limitations, initial research efforts have explored the potential of Large Language Models (LLMs) to enhance the explainability of ML models and XAI techniques [Spitzer *et al*., 2024] [Zytek *et al*., 2024]. The primary motivation behind this approach is the ability of LLMs to engage in human-like communication, making complex technical concepts more accessible [Caseli and Nunes, 2023]. Additionally, LLMs possess the capacity to acquire domain-specific knowledge, which can further enrich explanations by contextualizing them within a given field of application [Zhu *et al*., 2024]. These characteristics suggest that LLMs could play a transformative role in making ML explanations more interpretable for non-technical users.

Preliminary studies have investigated the use of LLMs and techniques such as In-Context Learning (ICL) [Zytek *et al*., 2024] and Retrieval-Augmented Generation (RAG) [Spitzer *et al*., 2024] to improve ML explainability. By leveraging these approaches, researchers aim to make model explanations more comprehensible by integrating relevant background knowledge and dynamically adapting explanations based on user queries. Despite these promising directions, several challenges remain unaddressed, particularly regarding the contextual relevance and quality of LLM-generated explanations [Zytek *et al*., 2024].

Two primary challenges hinder the broader adoption of LLM-based explanations: (1) ensuring that explanations are contextually relevant, considering the specific domain of application, and (2) developing a robust evaluation framework to assess the quality of generated explanations. The lack of standardized benchmarks for explanation quality further complicates the validation of LLM-enhanced XAI approaches [Zytek *et al*., 2024]. Thus, there is a pressing need for innovative solutions that can refine and evaluate LLM-generated explanations more effectively.

This article presents an extended version of the research originally discussed in Miyaji and Corrêa [2025], addressing a critical gap in the practical application of AI: domain experts, such as clinicians or financial analysts, are frequently presented with model predictions accompanied by cryptic technical outputs from XAI tools but lack the specialized expertise to interpret them. Building upon our previous multi-agent architecture—which utilizes Retrieval-Augmented Generation (RAG) and web search to transform technical data into context-aware narratives—this study introduces two significant advancements to further bridge this gap. First, we employ Vision-Language Models (VLMs) to enable the system to interpret visual outputs from SHAP and LIME directly, moving beyond text-only feature importance to handle complex graphical charts that are often the primary medium for XAI insights. Second, we implement a more robust evaluation framework using G-Eval [Liu *et al*., 2023] for the LLM as a Judge paradigm [Gu *et al*., 2024], providing a more detailed and consistent assessment of explanation quality through Chain-of-Thought (CoT) reasoning [Wei *et al*., 2022]. This enhanced framework is rigorously validated against human evaluations to ensure that the generated narratives provide the transparency and trust necessary for non-technical users to effectively leverage ML predictions.

Our main contributions are threefold: (1) an enhanced multi-agent architecture that integrates Vision-Language Models (VLMs) with Retrieval-Augmented Generation (RAG) and web search to facilitate the multimodal interpretation of XAI outputs, effectively transforming both technical feature importance and graphical charts into context-rich narratives for domain experts; (2) a rigorous validation of the LLM as a Judge paradigm through the G-Eval framework, providing a scalable and consistent assessment of explanation quality that leverages Chain-of-Thought reasoning to achieve high alignment with human expert evaluations; and (3) the integration of these components into a unified system that bridges the gap between technical XAI provisions and meaningful insights. Collectively, these advancements aim to make machine learning predictions more accessible, interpretable, and actionable for a broader range of users, ensuring that complex model outputs are transparent and trustworthy in high-stakes environments.

2 Related Works

2.1 Explainable Artificial Intelligence (XAI)

Doran et al. (2017) categorize explainability in ML into three levels: opaque, interpretable, and comprehensible models. Opaque models, such as deep neural networks, operate as ”black boxes,” making it difficult to understand their decision-making processes. Interpretable models, such as decision trees and linear models, allow direct inspection of their parameters and decision rules, making them easier to analyze [James *et al*., 2013]. Comprehensible models, on the other hand, provide explanations in a way that is understandable to human users, bridging the gap between technical and non-technical audiences. While most ML models fall within the first two categories, achieving comprehensibility remains a significant challenge, particularly for deep learning and ensemble methods [Doran *et al*., 2017] [Dafali *et al*., 2023].

One of the most widely adopted techniques to enhance model interpretability is SHAP [Lundberg and Lee, 2017]. SHAP is based on cooperative game theory and assigns importance scores to input features based on their contribution to the model's predictions. By leveraging Shapley values, SHAP provides a globally consistent and locally accurate explanation of how features influence the output. This technique enables even highly complex models, such as deep neural networks and gradient boosting machines, to become more interpretable. However, SHAP explanations still require technical expertise to interpret correctly, limiting their accessibility for non-expert users [Arrieta *et al*., 2020].

LIME [Ribeiro *et al*., 2016] is another popular XAI technique that improves model interpretability by approximating the decision boundary of a complex model with a simpler, interpretable surrogate model. LIME perturbs the input data by making small variations and observes the changes in the model's predictions. It then trains a local, interpretable model (such as a linear regression or decision tree) to approximate the black-box model's behavior in the vicinity of a specific instance. By doing so, LIME generates local explanations that help users understand how a model makes decisions for individual predictions [Ribeiro *et al*., 2016]. However, similar to SHAP, interpreting LIME explanations correctly requires a solid understanding of ML concepts.

However, Mumuni and Mumuni [2025] point out that traditional XAI outputs, which are often purely numerical or technical, still pose a significant barrier for non-expert users. There is a critical need to bridge these visual and technical features with semantically meaningful explanations. This study addresses this gap by not only using technical data but also incorporating visual artifacts from these tools through multimodal analysis.

2.2 Large Language Models and XAI

Recent research has explored integrating XAI with LLMs to enhance comprehensibility [Burton *et al*., 2023]. These efforts have primarily leveraged ICL strategies, including Prompt Engineering [Zytek *et al*., 2024], Context-Augmented Generation (CAG) [Spitzer *et al*., 2024] and RAG [Spitzer *et al*., 2024]. The goal of these approaches is to enable LLMs to generate more natural, human-understandable explanations for ML model predictions by incorporating relevant context and domain-specific knowledge.

Beyond mere translation, recent advancements emphasize the role of narrative-driven XAI. Martens *et al*. [2025] introduce the concept of "XAIstories", demonstrating that LLMs can effectively convert technical outputs like SHAP values and counterfactual explanations into persuasive narratives. Their findings suggest that over 90% of general audiences find these narrative justifications more convincing and easier to act upon than raw data. This highlights that LLMs do not only bridge the technical gap but also enhance the overall user experience by framing model behavior within a storytelling context that aligns with human cognitive processing.

The challenge of tailoring explanations to non-experts is further characterized as the "XAI Gap", where technical transparency fails to translate into functional understanding for the end-user. Paraschou *et al*. [2025] propose a human-centered framework aimed at democratizing XAI, arguing that LLMs must act as intermediaries that filter out technical noise while preserving decision fidelity. By focusing on user-centric design, such frameworks ensure that the explanations are not only accurate but also respect human values and decision-making contexts, which is crucial for the adoption of AI in sensitive societal roles.

CAG and RAG are two key techniques that improve LLM performance in explanation tasks. Both techniques enhance LLM responses by retrieving external knowledge from unstructured documents, databases, or web sources. Instead of relying solely on pre-trained knowledge, RAG dynamically integrates real-time information, improving the relevance and accuracy of generated explanations [Zhu *et al*., 2024]. While these techniques have shown significant potential, there remain unexplored opportunities to generate contextually relevant responses [Zytek *et al*., 2024]. One key challenge is tailoring explanations to specific domains and user expertise levels. Current approaches often lack the adaptability to generate responses that effectively align with different user needs.

Furthermore, the capability of LLMs is being explored not just as a wrapper for existing tools, but as standalone post hoc explainers. Research into various prompting strategies—including perturbation-based and instruction-based In-Context Learning—indicates that LLMs can identify feature importance and model logic on par with established methods like Integrated Gradients and SHAP [Kroeger *et al*., 2023]. This suggests a shift towards next-generation explainers that do not require white-box access to the underlying model, potentially reducing the computational overhead and complexity traditionally associated with high-fidelity XAI techniques.

Another major challenge lies in ensuring that LLM-generated explanations are not only coherent and relevant but also technically and contextually accurate [Zytek *et al*., 2024]. Existing methods do not always guarantee that explanations faithfully reflect the true reasoning behind ML model decisions. For instance, if an LLM generates an explanation that misrepresents the underlying XAI technique's outputs, it could lead to incorrect conclusions or misplaced trust in the model's predictions. Addressing this issue requires established metrics and more sophisticated mechanisms for validating explanation quality.

To improve the reliability of LLM-based explanations, recent research has explored the use of evaluation frameworks that assess the correctness and relevance of generated explanations [Zytek *et al*., 2024]. Some studies propose using LLM as a Judge frameworks [Gu *et al*., 2024], where an LLM evaluates the explanations based on predefined criteria [Wang *et al*., 2025]. However, this approach remains controversial, as it introduces potential biases and raises questions about the reliability of self-assessment. A more robust strategy involves comparing LLM-generated explanations with human evaluations and expert feedback [Wang *et al*., 2025] [Zytek *et al*., 2024].

Despite these challenges, the potential of combining LLMs with XAI represents a promising direction for improving the interpretability and comprehensibility of ML models. Thus, this study focus on refining evaluation frameworks and optimizing retrieval-based techniques to ensure that LLM-assisted explanations are not only accurate but also truly useful for non-expert users in real-world applications.

2.3 Vision-Language Models (VLMs) for Enhanced XAI

The emergence of Vision-Language Models (VLMs) has introduced a new paradigm for interpretability, shifting from purely numerical feature attributions to semantically grounded visual explanations. Traditional saliency-based methods, such as Class Activation Maps (CAM) and Grad-CAM, often suffer from limited discriminative information because they rely on one-hot categorical labels, which may lead to only partial correspondence between the image and the generated explanation [Lei *et al*., 2023]. Recent advancements, such as the Language-Image COnsistency (LICO) model, address this by correlating learnable linguistic prompts with corresponding visual features in a coarse-to-fine manner [Lei *et al*., 2023]. By enforcing consistency between textual descriptions and visual evidence, these models ensure that the explanations are not only localized but also semantically meaningful, a principle that is fundamental when interpreting complex XAI plots where visual patterns must map directly to domain-specific concepts.

Furthermore, the integration of contrastive learning frameworks, such as CLIP (Contrastive Language-Image Pre-training), has proven instrumental in refining the boundaries of visual explanations. As demonstrated by the Cross Language Image Matching (CLIMS) framework, natural language supervision can be leveraged to suppress background noise and falsely activated regions that often plague traditional XAI outputs [Xie *et al*., 2022]. In the context of

weakly supervised semantic segmentation, CLIMS utilizes language-derived constraints to ensure that visual activations remain faithful to the object of interest rather than co-occurring background elements [Xie *et al*., 2022].

Our work adapts these computer vision principles to the interpretation of structured XAI artifacts, such as SHAP force plots and LIME charts. By treating these technical visualizations as "visual documents," our Multimodal Explainer Agent leverages the cross-modal alignment capabilities to bridge the gap between graphical trends and the linguistic context retrieved through our RAG and web search modules.

3 Methodology

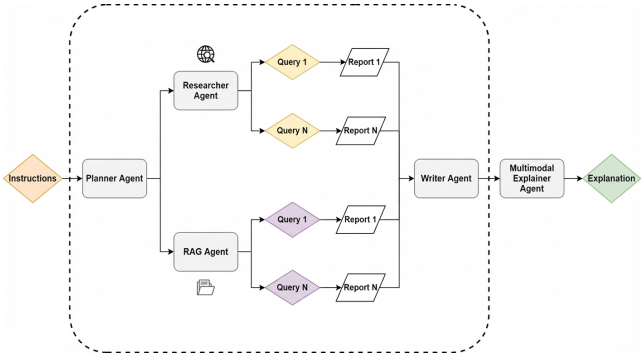

Figure 1. Proposed Multi-Agent Architecture to enrich LLM-generated explanations.

The proposed methodology extends our previous multi-agent architecture by incorporating multimodal capabilities and a more rigorous evaluation framework, specifically designed to bridge the gap between technical XAI outputs and domain-specific understanding. Our approach tests a core hypothesis: that the quality and trustworthiness of AI-generated explanations can be significantly enhanced by moving beyond single-prompt generation to a coordinated multi-agent reasoning process involving specialized agents. We posit that by systematically combining curated foundational knowledge from RAG modules, dynamic real-time information from web search, and visual insights from Vision-Language Models (VLMs), it is possible to produce narratives that are not only descriptively accurate but also causally richer and more contextually grounded than traditional methods. The following subsections detail the enhanced multi-agent architecture, the multimodal integration, the G-Eval-based evaluation methodology, and the experimental setup used to validate our findings.

3.1 Multi-Agent Architecture and Multimodal Integration

As illustrated in Figure 1, the architecture begins with the Planner Agent, which serves as the central decision-making component and orchestrator. This agent is designed to be modular, allowing for different configurations of agents depending on the query's complexity. Equipped with an LLM possessing strong reasoning capabilities, the Planner receives a structured input consisting of the human query alongside the technical context of the prediction. Based on this, it formulates a structured plan for information retrieval and synthesis. This plan determines which specific agents should be activated—whether the RAG Researcher Agent for internal domain-specific knowledge or the Researcher Agent for external web-based data. By defining these task sequences, the Planner ensures that the Explainer Agent receives a refined context to generate a narrative that is both technically accurate and human-centric.

The Researcher Agent is responsible for retrieving real-time information from external sources using a web search tool, performing targeted searches to identify and extract the most relevant content to support the explanation. By integrating dynamic, up-to-date knowledge, this component ensures that the explanations remain contextually rich and aligned with current real-world data.

In contrast to the external focus of the web search, the RAG Researcher Agent operates within a controlled knowledge environment. It retrieves information from a pre-curated set of documents related to Machine Learning, XAI, and the specific domain of application, ensuring that explanations are grounded in reliable, peer-reviewed references. Once the information is gathered, the Writer Agent plays a critical summarization role, processing the outputs from both research agents. This agent filters redundant or irrelevant content and synthesizes key insights into a structured, coherent summary, preparing the well-organized contextual foundation necessary for the final interpretation.

Finally, the transition from technical data to human-centric narrative is handled by the Multimodal Explainer Agent. In this extended version of our work, the Explainer Agent is significantly enhanced by the integration of a Vision-Language Model (VLM). Beyond merely processing textual feature importance lists, the VLM-powered agent can directly analyze graphical outputs from SHAP and LIME, such as force plots, summary charts, and local explanation graphs. This capability allows the system to capture visual nuances and graphical trends that are often lost in text-only representations. By integrating the summarized contextual information from the Writer Agent with these multimodal XAI insights, the Explainer Agent translates complex model behavior into intuitive narratives. This ensures that the final output is not only technically sound and faithful to the model's logic but also accessible and actionable for non-expert users in real-world applications.

3.2 Evaluation and LLM as a Judge

To assess the quality of the generated explanations, we adopted four key evaluation criteria based on the framework proposed by Zytek *et al*. [2024]. These metrics are designed to capture both technical accuracy and communicative effectiveness: (1) Soundness, which measures the factual correctness of the information; (2) Fluency, which evaluates the extent to which the narrative sounds natural and human-like; (3) Completeness, which assesses whether the explanation provides sufficient detail and covers all relevant aspects of the prediction; and (4) Context Awareness, which determines the degree to which the narrative "explains the explanation" by providing external contextual grounding. Each explanation was scored on a scale from 1 to 5 for each criterion to ensure a

high-granularity measurement, where higher scores indicate superior performance. The detailed scoring rubric used for this assessment is presented in Table 1.

A key enhancement in this study is the transition to the G-Eval framework for the LLM as a Judge paradigm [Liu *et al*., 2023]. Unlike previous approaches that relied on direct scoring, G-Eval employs a more robust, criteria-based process utilizing Chain-of-Thought (CoT) reasoning [Wei *et al*., 2022]. For each metric, the G-Eval judge is provided with the detailed rubric and a set of instructions to first generate a step-by-step reasoning path before assigning the final score. This approach significantly minimizes bias and increases the consistency of the automated assessment, providing an efficient means of benchmarking explanation quality for large-scale evaluations [Wang *et al*., 2025].

To validate the reliability of this automated framework, we conducted a parallel human assessment. An expert in ML and XAI independently reviewed a subset of the explanations using the same 1–5 scale. The robustness of the LLM-based scores was analyzed by calculating the agreement with human evaluations through Cohen's Kappa, alongside Pearson and Spearman correlation coefficients. For the latter two, an average was calculated independently across the four criteria to ensure a statistically sound validation of our automated judge [James *et al*., 2013].

3.3 Case Study

The experiments were conducted on the UCI Adult dataset [Becker and Kohavi, 1996]. The dataset, also known as the Census Income dataset, is derived from the US Census and consists of approximately 15,600 records with 14 demographic and employment-related attributes. This dataset has become a benchmark in the ML community due to its diverse feature set, which includes age, education, occupation, and marital status, among others. Its structured yet real-world nature makes it a valuable resource for studying various predictive modeling techniques. The dataset's relevance in ML extends beyond mere classification tasks. Researchers frequently use it to explore fairness and bias in algorithms, as the data inherently contains sensitive attributes. This makes it an ideal candidate for experiments aiming to evaluate and mitigate discriminatory biases in predictive models [Sena and Machado, 2024] [Girhepuje, 2023].

In the literature, studies involving the UCI Adult dataset often focus on interpretability and fairness [Sena and Machado, 2024]. Researchers use this dataset to investigate how different machine learning models handle these fairness criteria, and to quantify bias [Girhepuje, 2023]. Moreover, the dataset serves as an excellent testbed for explainability research in machine learning [Mindlin *et al*., 2024]. Techniques such as SHAP and LIME are frequently applied to models trained on the UCI Adult dataset to provide insights into feature importance [Lundberg, 2018]. Overall, the dataset's combination of complexity and accessibility makes it a powerful tool for advancing research in ML [Sena and Machado, 2024].

3.4 Experiments

This section details the implementation of the experiments. As in established benchmarks, we selected XGBoost [Chen and Guestrin, 2016] as the classifier for the UCI Adult dataset, given its superior predictive performance reported in the literature [Becker and Kohavi, 1996]. Additionally, XGBoost is widely used as a baseline model for applying XAI techniques, making it an ideal candidate for evaluating explainability methods such as SHAP and LIME [Lundberg, 2018].

To conduct the experiments, we trained the XGBoost model on the training set of the UCI Adult dataset using optimized hyperparameters based on grid search and cross-validation. After model training, we generated predictions for 100 randomly selected test samples to serve as the basis for our explainability analysis. Each of these samples was then subjected to an interpretability assessment using SHAP [Lundberg and Lee, 2017] and LIME [Ribeiro *et al*., 2016].

The generated outputs of both SHAP and LIME serve as input for the proposed Multi-Agent System. To construct the Multi-Agent Architecture, we utilized the LangGraph library, which facilitates the implementation of multi-agent workflows by enabling structured interactions between agents [LangGraph, 2025]. The selection of LLMs was tailored to the specific demands of each task within the system, ensuring an optimal balance between reasoning capabilities and computational efficiency.

For complex tasks requiring advanced reasoning and planning, we adopted OpenAI's o3-mini, a model designed for structured problem-solving and decision-making [OpenAI, 2025b]. This choice ensures that the agents responsible for orchestrating multi-step workflows can generate well-structured and contextually relevant outputs. Specifically, the Planner Agent, responsible for devising retrieval and explanation strategies, the Writer Agent, tasked with synthesizing contextual information, and the Explainer Agent, which generates the final user-facing explanations, all employ OpenAI's o3-mini. Furthermore, for the Multimodal Explainer Agent, we utilized OpenAI's GPT-5, leveraging its multimodal capabilities to directly interpret and synthesize graphical XAI outputs—such as SHAP and LIME charts—alongside textual data.

Conversely, for lighter summarization tasks, we opted for OpenAI's GPT-4o-mini [OpenAI, 2025a], which offers a more efficient solution without compromising quality. The Researcher Agent, responsible for retrieving real-time information via web search, and the RAG Agent, which retrieves domain-specific documents for contextual enrichment, utilize GPT-4o-mini to summarize and preprocess retrieved content before passing it to downstream agents.

The Researcher Agent was equipped with a Web Search Tool provided by Tavily [Tavily, 2025], which enables queries across multiple search engines while also retrieving contextual information from news articles and web pages through web scraping techniques. The output of this tool consists of a structured string containing the detailed content of each retrieved webpage, ensuring that relevant information is effectively extracted and made available for downstream processing within the Multi-Agent System.

Table 1. Evaluation criteria to assess the generated explanations Zytek *et al*. [2024].

Metric	1	3	5
Soundness	The explanation includes one or more objective errors	The explanation includes one or more misleading statements	The explanation contains no errors
Fluency	The explanation is very unnatural or confusingly worded	The explanation is somewhat natural	The explanation sounds very natural as though written by a human
Completeness	Some features given were missed entirely, or the direction was not specified	All features were described, but exact feature or contribution was not given	All feature values are given, and all contributions are described with directions
Context-awareness	No context information is is provided	The answer includes references and comparisons to the average feature values	The answer includes further explanations of what may cause a specific contribution

Explainer Agent Prompt for SHAP

```
You are helping users understand an ML model's prediction. The model predicts the probability of whether the individual earns more than 50K based on census variables (UCI Adult Dataset).
I will give you feature contribution explanations, generated using SHAP, in (feature, feature_value, contribution, average_feature_value) format.
Convert the explanations in simple narratives. Do not use more tokens than necessary.
Example of house pricing: (Second floor square feet, 854, 12757.83, 583.0), (Construction date, 2003, 9115.72, 1976.1), (Basement area in sq. ft., 856, -6157.86, 1073.6)
Think step by step. Interpret and justify the obtained results in terms of causality, based on facts, if they exist.
```

"LLM as a Judge" Prompt

```
You are an expert evaluator assessing the quality of AI-generated explanations.
Rate the given explanation based on the following criteria using a scale from 1 to 5.

### **Evaluation Criteria:**
1 **Soundness (The correctness of the information included in the narrative.)**:
2 **Fluency (The extent to which the narrative sounds "natural" or like it was generated by a human peer in conversation.)**:
3 **Completeness (The amount of information included in the narrative.)**:
4 **Context-awareness ( The degree to which the narrative "explains the explanation" by providing external context.)**:
```

Figure 2. Prompts for the Explainer Agent [Zytek *et al*., 2024] and for LLM as a Judge.

For the RAG Agent, we adopted Chroma [Chroma, 2025] as the vector database, which served as a repository for domain-specific knowledge. This database was composed of scientific articles and books cited in this study, covering topics related to ML, XAI, and the case study dataset. Additionally, we included relevant prior research that has explored the UCI Adult dataset, ensuring that the retrieval process could leverage established knowledge to generate more informative and context-aware explanations.

To enhance the retrieval process, we employed a Recursive Chunking strategy with 2048 characters [Zhu *et al*., 2024]. At query time, the system retrieved the top-5 most relevant documents, ensuring that explanations were constructed based on highly pertinent and structured sources. This strategy prevented excessive fragmentation of information while maintaining a coherent contextual flow for subsequent summarization.

Upon retrieving relevant content, both the Researcher Agent and the RAG Agent generated summarized reports based on the search query. These reports distilled key insights from the retrieved sources, structuring the extracted knowledge in a way that facilitated its integration into the explanation generation pipeline. By combining real-time web search with a curated scientific knowledge base, this approach ensured that explanations were grounded in both up-to-date and authoritative sources, enhancing their credibility, completeness, and domain relevance.

Based on the combined report generated from the Researcher Agent's web search results and the RAG Agent's retrieved documents, which was then summarized by the Writer Agent, the Explainer Agent received two key inputs: the outputs from the XAI model and the best-performing prompt identified in prior research [Zytek *et al*., 2024], as presented in Figure 2. The Explainer Agent's primary objective was to generate a final explanation that would be clear, informative, and accessible to the end user.

To systematically assess the contribution of each component within the proposed Multi-Agent Architecture, we conducted a three-stage evaluation. First, we implemented a Zero-Shot baseline, where the Explainer Agent used only the prompt from Figure 2, without any additional contextual information from external sources. This allowed us to establish a reference performance level based purely on the prompt's effectiveness.

Next, we evaluated a RAG approach, where the Explainer Agent was supplemented with contextual information exclusively from the RAG Agent. This step enabled us to assess the impact of domain-specific knowledge retrieval in improving the quality of generated explanations.

Finally, we tested the full system, incorporating both the RAG Agent's retrieved knowledge and the Researcher Agent's real-time web search results. This configuration aimed to leverage complementary information sources, combining scientific literature and real-time updates to enhance the context-awareness, relevance, and completeness of the explanations. By comparing these different setups, we systematically measured how each component contributed to improving the interpretability and user comprehension of the ML model's predictions.

The explanations generated by the different system configurations (Zero-Shot, RAG-only, and RAG with Web Search) were evaluated based on the criteria presented in Table 1. These criteria aimed to assess key aspects, ensuring a comprehensive evaluation of the generated explanations. Each explanation was scored on a 1 to 5 scale, following the established evaluation framework used in prior studies [Zytek *et al*., 2024].

For the LLM as a Judge evaluation, we implemented the G-Eval framework [Liu *et al*., 2023] utilizing OpenAI's GPT-5.1, a high-reasoning-capacity model specifically selected for its advanced performance in complex evaluative tasks. Unlike traditional direct-scoring methods, G-Eval adopts a more robust, criteria-based approach powered by Chain-of-Thought (CoT) reasoning. To standardize and enhance consistency in the judgment process, the model was provided with a carefully designed evaluation prompt—detailed in Figure 2—which incorporated the specific rubrics for Soundness, Fluency, Completeness, and Context-Awareness described in Table 1.

4 Results

We conducted the experiments following the implementation details previously described. Initially, the XGBoost classifier underwent hyperparameter optimization using a Grid Search approach, with 10-fold cross-validation to ensure robust model selection. This process systematically explored different parameter combinations, selecting the configuration that maximized predictive performance. The best-performing model achieved an accuracy of 83.1% on the hold-out test dataset, with a ROC-AUC score of 82.4%, demonstrating strong classification capabilities.

After identifying the optimal configuration, the trained XGBoost model was used to generate predictions for 100 randomly selected samples from the test dataset. These predictions served as the basis for the subsequent explainability analysis, where SHAP and LIME were applied. Both techniques were used to compute feature attributions, providing insight into the contribution of individual features to the model's predictions.

The resulting XAI outputs—comprising both textual feature importance lists and multimodal visual artifacts such as summary plots and waterfall plots—were utilized as inputs for our Multi-Agent architecture, enabling the generation of explanations under the three proposed configurations: Zero-Shot, RAG-only, and RAG with Web Search. Notably, for the multimodal outputs, the Zero-Shot configuration was specifically tested to establish a baseline and evaluate the VLM's inherent capacity to interpret complex graphical data directly, without the aid of supplementary external context. Each configuration represents a different level of contextual enhancement, allowing us to systematically evaluate the impact of additional retrieved information on the quality and depth of the explanations. An example of the narratives generated by each configuration is presented in Figure 3, illustrating how the inclusion of retrieved knowledge influences the interpretability and semantic richness of the model's predictions. Figure 4 presents examples of explanations generated by the Multimodal Explainer Agent. It demonstrates the capability of the agent to convert visual Explainable AI plots into natural language summaries. The workflow shows the agent taking graphical XAI Outputs as inputs and generating corresponding textual explanations.

Following the explanation generation process, all outputs were systematically evaluated based on the criteria presented in Table 1, using the LLM as a Judge approach. The evaluation results are summarized in Table 2, which reports the average scores for each configuration across the two XAI techniques (SHAP and LIME).

The empirical results, summarized in Table 2, demonstrate that all three architectural configurations maintain consistently high performance across the Soundness, Fluency, and Completeness metrics. This indicates that the intrinsic reasoning capabilities and internal linguistic knowledge of the frontier LLMs utilized are robust enough to translate raw XAI data into well-structured and coherent natural language narratives, regardless of the level of external contextual augmentation.

However, a significant divergence in performance is observed in the Context-Awareness criterion, which serves as the primary differentiator for the proposed multi-agent system. In the Zero-Shot configuration, while the explanations are linguistically fluent, they often lack the causal depth and broader interpretability required by domain experts. Without access to external information, the model is limited to describing the mathematical feature importance without justifying why specific variables influence the prediction within a real-world or scientific context.

When additional contextual grounding is incorporated, the Context-Awareness scores exhibit a substantial and progressive increase. Specifically, the performance rises from a baseline of 4.65 in the Zero-Shot setup to 4.79 in the RAG-only configuration, and reaches its peak at 4.96 when both RAG and Web Search are integrated. This represents an improvement in the model's ability to provide externally

Table 2. Average scores with Zero-Shot, RAG-only, and RAG with Web Search

Configuration	Soundness	Fluency	Completeness	Context-awareness
RAG with Web Search	4.93 ± 0.02	4.89 ± 0.01	4.93 ± 0.01	4.96 ± 0.11
RAG-Only	4.92 ± 0.02	4.86 ± 0.01	4.90 ± 0.01	4.79 ± 0.12
Zero-Shot	4.94 ± 0.01	4.86 ± 0.01	4.83 ± 0.01	4.65 ± 0.11

LIME Contributions

Feature: Capital Gain <= 0.00, Contribution: -0.0304
Feature: Hours per week <= 40.00, Contribution: -0.0181
Feature: Relationship <= 0.00, Contribution: -0.0069
Feature: Country <= 39.00, Contribution: 0.0059
Feature: Capital Loss <= 0.00, Contribution: -0.0052
Feature: Age <= 28.00, Contribution: -0.0046
Feature: Education-Num <= 9.00, Contribution: -0.0045
Feature: 0.00 < Sex <= 1.00, Contribution: 0.0033
Feature: 2.00 < Marital Status <= 4.00, Contribution: -0.0017
Feature: Workclass <= 4.00, Contribution: -0.0013

Zero-Shot Explanation

Let's break down what these LIME explanations mean in a straightforward way, focusing on how different factors influence the likelihood of earning more than $50,000.

1. **Capital Gain (<= 0.00)**: Having no capital gains (meaning you haven't made money from investments or assets) is associated with a lower chance of earning over $50,000. This factor has a significant negative impact.
2. **Hours Worked per Week (<= 40.00)**: Working 40 hours or less per week also contributes negatively to the likelihood of earning more than $50,000.

10. **Workclass (<= 4.00)**: The type of work or employment status has a minor negative contribution.

In summary, the analysis shows that factors like capital gains, hours worked, relationship status, and education level significantly influence the likelihood of earning over $50,000. Younger individuals, those with lower education, and those working fewer hours are less likely to reach this income level. Additionally, gender and marital status also play roles, reflecting broader societal trends in income disparities.

RAG with Web Search Explanation

Let's break down the LIME explanation step by step to understand how different factors contribute to the prediction of whether an individual earns more than $50,000.

1. **Capital Gain <= 0.00 (-0.0304)**: When an individual has no capital gains, it negatively impacts their chances of earning more than $50,000. This suggests that having investment income is important for reaching that income level. If you don't have any capital gains, your likelihood of earning above $50K decreases.
2. **Hours per Week <= 40.00 (-0.0181)**: Working 40 hours or less per week also reduces the probability of earning more than $50,000. This indicates that working more hours is generally associated with higher income. If you are working part-time or not putting in many hours, it may be harder to reach that income threshold.

10. **Workclass <= 4.00 (-0.0013)**: The workclass variable has a small negative contribution, indicating that certain types of employment may not be as lucrative as others. This suggests that the nature of one's job can influence income potential.

Summary
In summary, the analysis reveals that having capital gains, working more hours, and being in a stable relationship are factors that can positively influence the likelihood of earning more than $50,000. Conversely, being younger, having lower education levels, and not having capital gains or working fewer hours tend to decrease the chances of reaching that income level. Understanding these factors can help individuals identify areas for improvement in their financial situations.

Figure 3. Example of explanations generated by Zero-Shot and RAG with Web Search configurations.

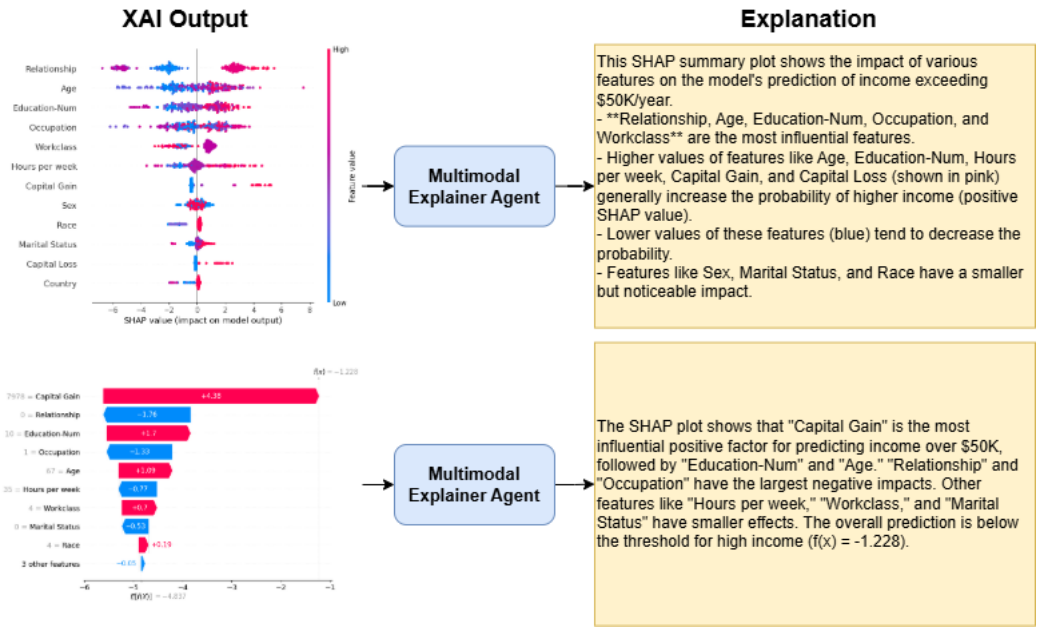

Figure 4. Illustrative examples of the narratives generated by the Multimodal Explainer Agent. The figure showcases the system's ability to interpret visual XAI artifacts and transform them into human-centric explanations.

grounded justifications. A paired t-test conducted at a 0.05 significance level confirms that the improvement from the Zero-Shot approach to the RAG with Web Search configuration is statistically significant. These findings reinforce the value of the multi-agent retrieval process, highlighting that the dynamic integration of scientific documents and real-time data is essential for transforming cryptic XAI outputs into causally grounded insights that are truly actionable for

non-expert users.

To validate the use of LLM as a Judge for evaluating the generated explanations, the same explanations were independently assessed by a human expert in ML and XAI, following the same predefined evaluation criteria. This human annotation served as a ground-truth reference to compare the reliability and alignment of the LLM-based evaluation.

The results indicate a high level of agreement across the board, starting with a Pearson correlation coefficient of 0.80, which suggests a strong linear relationship between the automated and manual judgments. Similarly, a Spearman correlation coefficient of 0.73—calculated as an average of the correlations measured independently for each of the four evaluation criteria—further confirms that the ranking of explanations remains consistent across both evaluators. To account for the possibility of random agreement, we also measured Cohen's Kappa, yielding a coefficient of 0.56. This value indicates moderate to substantial agreement, reflecting the consistency of the LLM-based scores even when adjusted for chance.

These findings demonstrate that the LLM as a Judge paradigm, particularly when leveraging high-reasoning models, is a viable alternative for evaluating XAI-generated explanations. Since the automated assessments align closely with those of a domain expert, this approach reinforces the feasibility of automating the evaluation process. This enables the scalable assessment of explanation quality across large datasets while maintaining a level of reliability that matches human-level expertise.

5 Conclusions

This study presented an enhanced multi-agent architecture designed to bridge the gap between technical XAI outputs and human-centric understanding. By integrating Retrieval-Augmented Generation (RAG) with real-time web search and utilizing Vision-Language Models (VLMs) for the direct interpretation of SHAP and LIME graphical artifacts, the system effectively transformed technical data into context-rich narratives. Experimental results on the UCI Adult dataset demonstrated that the full system configuration significantly improves context awareness by 6.5% compared to standard LLM-generated explanations. Furthermore, the validation of the G-Eval framework through the LLM as a Judge paradigm achieved over 80% correlation with human expert evaluations, providing a scalable and consistent methodology for assessing explanation quality. However, we acknowledge a limitation in the human validation process: the evaluation relied on a single expert serving as the ground truth. Given the inherent subjectivity involved in assessing explanation criteria, the absence of other human evaluator limits a more comprehensive analysis of this subjectivity.

Future research will focus on extending the evaluation of this framework to other high-stakes domains, such as medical diagnosis and financial risk assessment, to ensure its robustness across diverse datasets. Additionally, conducting user studies with domain experts would provide valuable insights into the practical impact of these narrative-driven explanations on human trust and decision-making. We also plan to investigate the optimization of the multi-agent workflow to reduce computational latency and costs, exploring the potential of specialized, fine-tuned open-source models to perform agentic tasks without compromising the quality and accuracy of the generated insights.

Acknowledgements

This study was supported by the Amigos da Poli – project 2025_021. It was also financed in part by CAPES – Finance Code 001.

Authors' Contributions

RM and PC contributed to the conception of this study. RM and AL performed the experiments. RM is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they do not have competing interests.

References

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al*. (2020). Explainable artificial intelligence (xai): : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58(C).

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J., and Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Borys, K., Schmitt, Y., Nauta, M., Seifert, C., Krämer, N., Friedrich, C., and Nensa, F. (2023). Explainable ai in medical imaging: An overview for clinical practitioners – saliency-based xai approaches. *European Journal of Radiology*, 162.

Burton, J., Moubayed, N., and Enshaei, A. (2023). Natural language explanations for machine learning classification decisions. *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*.

Caseli, H. and Nunes, M. (2023). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. Brasileiras - Processamento de Linguagem Natural.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chroma (2025). Chroma. Available on: `https://www.trychroma.com/`. Accessed in 23 March 2025.

Dafali, S. M., Kissi, M., and El Beggar, O. (2023). Comparative study between global and local explainable models.

In *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–8. DOI: 10.1109/SITA60746.2023.10373599.

Doran, D., Schulz, S., and Besold, T. R. (2017). What does explainable ai really mean? a new conceptualization of perspectives. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*.

Girhepuje, S. (2023). Identifying and examining machine learning biases on adult dataset. *ArXiv*.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Z., Gao, W., Ni, L., and Guo, J. (2024). A survey on llm-as-a-judge. *ArXiv*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, Londres.

Kroeger, N., Ley, D., Krishna, S., Agarwal, C., and Lakkaraju, H. (2023). Are large language models post hoc explainers? In *Proceedings of the NeurIPS Workshop on Robustness of Zero/Few-Shot Learning in Foundation Models (R0-FoMo)*.

LangGraph (2025). Langgraph. Available on: `https://langchain-ai.github.io/langgraph/tutorials/introduction/`. Accessed in 23 March 2025.

Lei, Y., Li, Z., Li, Y., Zhang, J., and Shan, H. (2023). LICO: Explainable models with language-image COnsistency. In *Advances in Neural Information Processing Systems*, volume 36, pages 61870–61887.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.153.

Lundberg, S. (2018). Benchmark xgboost explanations. Available on: `https://shap.readthedocs.io/en/latest/example_notebooks/benchmarks/tabular/Benchmark%20XGBoost%20explanations.html`. Accessed in 23 March 2025.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Martens, D., Hinns, J., Dams, C., Vergouwen, M., and Evgeniou, T. (2025). Tell me a story! Narrative-driven XAI with large language models. *Decision Support Systems*, 191:114402.

Mindlin, D., Robrecht, A., Morasch, M., and Cimiano, P. (2024). Measuring user understanding in dialogue-based xai systems. *Proceedings of the 27th European Conference on Artificial Intelligence*.

Miyaji, R. O. and Corrêa, P. L. P. (2025). Enhancing ML explainability with multi-agent LLMs: A context-aware XAI approach. In *Anais do XL Simpósio Brasileiro de Bancos de Dados (SBBD)*, pages 28–41, Porto Alegre. Sociedade Brasileira de Computação. DOI: 10.5753/sbbd.2025.246998.

Mumuni, F. and Mumuni, A. (2025). Explainable artificial intelligence (XAI): from inherent explainability to large language models.

OpenAI (2025a). Gpt-4o mini: advancing cost-efficient intelligence. Available on: `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Accessed in 23 March 2025.

OpenAI (2025b). Openai o3-mini. Available on: `https://openai.com/index/openai-o3-mini/`. Accessed in 23 March 2025.

Paraschou, E., Arapakis, I., Yfantidou, S., Macaluso, S., and Vakali, A. (2025). Mind the XAI gap: A human-centered LLM framework for democratizing explainable AI. In Guidotti, R., Schmid, U., and Longo, L., editors, *Explainable Artificial Intelligence, Part I*. Springer, Cham.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Richmond, K., Muddamsetty, S., Gammeltoft-Hansen, T., Olsen, H., and Moeslund, T. (2024). Explainable ai and law: An evidential survey. *Digital Society*, 3(1).

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., and F, H. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, 44:199–205.

Sena, L. and Machado, J. (2024). Evaluation of fairness in machine learning models using the uci adult dataset. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*. SBC.

Spitzer, P., Celis, S., Martin, D., Kühl, N., and Satzger, G. (2024). Looking through the deep glasses: How large language models enhance explainability of deep learning models. *Proceedings of Mensch und Computer 2024*.

Tavily (2025). Tavily. Available on: `https://docs.tavily.com/welcome`. Accessed in 23 March 2025.

Wang, B., Li, Y., Zhou, J., and Chen, F. (2025). Can llm assist in the evaluation of the quality of machine learning explanations? *ArXiv*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xie, J., Hou, X., Ye, K., and Shen, L. (2022). CLIMS: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4483–4492.

Zhu, Y., Yuan, H., Wang, S., Liu, S., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J. (2024). Large language models for information retrieval: A survey. *ArXiv*.

Zytek, A., Pidò, S., and Veeramachaneni, K. (2024). Llms for xai: Future directions for explaining explanations. *ACM CHI Workshop on Human-Centered Explainable AI*.

Černevičienė, J. and Kabašinskas, A. (2024). Explainable artificial intelligence (xai) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(216).