Enhancing ML Explainability with Multi-Agent LLMs: A Context-Aware XAI Approach

Renato O. Miyaji[1], **Pedro L. P. Corrêa**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

{re.miyaji,pedro.correa}@usp.br

Abstract. *Explainable AI (XAI) techniques enhance ML interpretability but often require technical expertise. We propose a multi-agent architecture that improves LLM-generated explanations through structured reasoning and contextual retrieval. Our system integrates web search, Retrieval-Augmented Generation (RAG), and XAI outputs via specialized agents. Experiments on the Adult dataset show that our approach outperforms standard LLM explanations by 7% in Context Awareness. Additionally, we validate LLM as a Judge, achieving over 80% correlation with human evaluations. Different LLMs, including OpenAI's GPT-4o and GPT-4o-mini, were tested, highlighting the effectiveness of multi-agent systems in ML explainability.*

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 in ML [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. These techniques allow even opaque models to become more transparent by offering insights into feature attributions and decision boundaries [Lundberg and Lee 2017] [Ribeiro et al. 2016]. However, even with these advancements, understanding the output of XAI techniques still demands significant ML and XAI knowledge. Non-technical users often struggle to interpret the provided explanations correctly,

limiting their practical applicability in real-world decision-making [Arrieta et al. 2020] [Bhatt 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 study addresses a critical gap in the practical application of AI: domain experts, such as clinicians or financial analysts, are often presented with a model's prediction alongside cryptic technical outputs from XAI tools, yet lack the expertise to interpret them. To bridge this gap, we propose a multi-agent architecture designed to transform these outputs into meaningful, context-aware narratives. By integrating a web search tool for real-time information with a RAG module accessing domain-specific documents, our system enriches explanations with the context necessary for a non-technical user to understand ML predictions. Furthermore, we explore the feasibility of using an LLM-based evaluation system (LLM as a Judge) [Gu et al. 2024] to assess the quality of explanations, comparing its performance with human evaluations.

Our main contributions are therefore: (1) a novel multi-agent architecture that translates technical XAI outputs into robust, context-rich explanations for domain experts; and (2) the validation of the LLM as a Judge paradigm as a scalable and reliable method for evaluating explanation quality. Ultimately, these contributions aim to make model insights more accessible, interpretable, and actionable for a broader range of users.

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.

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.

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.

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.

3. Method

Our methodology is designed to test a core hypothesis: that the quality and trustworthiness of AI-generated explanations can be significantly enhanced by moving beyond single-prompt generation to a multi-agent reasoning process. We posit that by systematically combining curated, foundational knowledge with dynamic, real-time information, it is possible to produce narratives that are not only descriptively accurate but also causally richer and more contextually grounded. This section details the multi-agent architecture, the evaluation methodology, the case study, and the experiments.

3.1. Multi-Agent Architecture

To leverage the reasoning and tool-use capabilities of LLMs [Wang et al. 2023], we propose a multi-agent architecture designed to enrich LLM-generated explanations with relevant contextual information. The architecture consists of five specialized agents, each responsible for a distinct task that contributes to the generation of comprehensive and human-like explanations. Figure 1 presents the proposed Multi-Agent Architecture to enrich LLM-generated explanations.

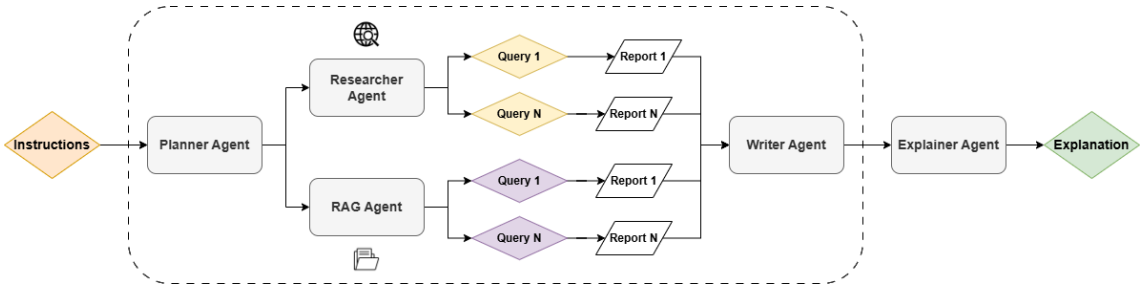

Figure 1. Proposed Multi-Agent Architecture to enrich LLM-generated explanations.

3.2. Evaluation and LLM as a Judge

The Planner Agent serves as the central decision-making component of the architecture. Equipped with an LLM with strong reasoning capabilities, this agent receives the human query and formulates a structured plan for information retrieval. The plan outlines which sources should be queried, whether external (via web search) or internal (via a curated document repository), ensuring that the generated explanation is as relevant and informative as possible.

The Researcher Agent is responsible for retrieving real-time information from external sources using a web search tool. This agent performs targeted searches based on the Planner Agent's instructions, identifying and extracting the most relevant content to support the explanation. By integrating dynamic, up-to-date knowledge, this component ensures that the generated explanations remain relevant and contextually rich.

The RAG Researcher Agent operates within a controlled knowledge environment, retrieving information from a pre-curated set of documents related to ML, XAI, and the specific domain of application. Unlike the Researcher Agent, which focuses on real-time external sources, the RAG Researcher Agent ensures that explanations are grounded in reliable, authoritative references. This is particularly important for technical fields where accuracy is critical.

The Writer Agent plays a summarization role, processing the retrieved information from both the Researcher and RAG Researcher Agents. This agent filters redundant or irrelevant content and synthesizes key insights into a structured, coherent summary. The goal is to prepare well-organized contextual information that will serve as a foundation for the final explanation.

Finally, the Explainer Agent integrates the summarized contextual information with the outputs from XAI techniques to generate human-like explanations. This agent ensures that the final explanation is not only technically sound but also accessible and understandable to non-expert users. By leveraging LLMs' natural language generation capabilities, the Explainer Agent translates complex model behavior into intuitive narratives.

To assess the quality of the generated explanations, we adopted four key evaluation criteria based on Zytek et al (2024): Soundness, which measures the correctness of the information included in the explanation. Fluency, which evaluates the extent to which the narrative sounds "natural" or like it was generated by a human peer in conversation.

Completeness that assesses whether the explanation provides sufficient detail and covers all relevant aspects of the prediction. Context Awareness that determines the degree to which the narrative "explains the explanation" by providing external context.

Each explanation was scored from 1 to 5 for each criterion to ensure a higher granularity measurement, based on the evaluation framework proposed by Zytek et al (2024). Higher scores indicate better performance in each category. Table 1 presents the evaluation criteria.

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

To automate the evaluation process, we employed an LLM as a Judge approach [Gu et al. 2024], where an LLM was tasked with assessing the generated explanations based on the predefined evaluation criteria. The LLM assigned scores from 1 to 5 for each category, ensuring consistency and efficiency in the evaluation process. This method provides an automated means of benchmarking explanation quality for large-scale assessments [Wang et al. 2025]. To validate the reliability of the LLM-based evaluation, we also conducted a human assessment of the explanations. An expert in ML and XAI independently reviewed a subset of generated explanations and rated them using the same criteria. The agreement between human evaluations and the LLM as a Judge scores was analyzed to determine the robustness of the automated evaluation framework using Cohen's Kappa, alongside Pearson and Spearman correlation coefficients, with the latter two representing an average of correlations calculated independently for each of the four criteria [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. Specifi-

cally, 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.

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.

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.

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.

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 employed a high-reasoning-capacity model, specifically OpenAI's o3-mini [OpenAI 2025b], as it has been optimized for complex evaluation tasks. To standardize and enhance consistency in the judgment process, we used a carefully designed evaluation prompt, detailed in Figure 2. This prompt guided the model in scoring each explanation across the defined evaluation criteria while ensuring that both qualitative and quantitative aspects were taken into account. Additionally, the LLM-based scoring was compared against human evaluations to measure the correlation between automated and human assessments, ensuring the reliability of this approach for future studies.

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 were then 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. Each configuration represents a different level of contextual enhancement, allowing us to evaluate the impact of additional retrieved information on explanation quality. An example of explanations generated by each configuration is presented in Figure 3, illustrating how the inclusion of retrieved knowledge influences the interpretability of the model's predictions.

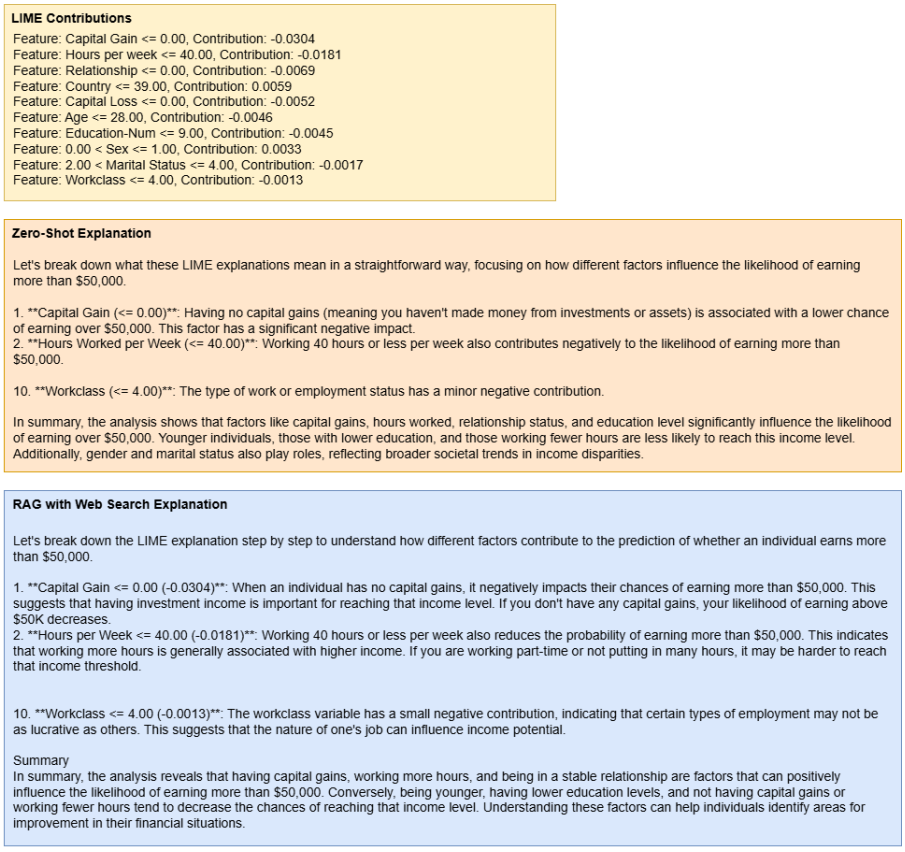

Figure 3. Example of explanations generated by Zero-Shot and RAG with Web Search configurations.

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 results indicate that all three configurations exhibit high performance across

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

Configuration	Soundness	Fluency	Completeness	Context-awareness
Zero-Shot	4.96 ± 0.02	4.98 ± 0.01	4.99 ± 0.01	4.64 ± 0.14
RAG-Only	4.94 ± 0.02	4.99 ± 0.01	4.98 ± 0.01	4.86 ± 0.12
RAG with Web Search	4.97 ± 0.01	4.99 ± 0.01	4.99 ± 0.01	4.96 ± 0.13

the Soundness, Fluency, and Completeness metrics, demonstrating that, regardless of additional contextual information, the expressive capabilities and internal knowledge of LLMs are already sufficient to translate the XAI outputs into coherent and well-structured natural language explanations. This suggests that the base LLM itself possesses strong communication capabilities, ensuring the clarity and completeness of the generated explanations.

However, the most significant difference is observed in the Context-Awareness criterion. In the Zero-Shot approach, the explanations, while linguistically well-formed, lack completeness in terms of causal reasoning and deeper interpretability. Without access to external information, the model does not incorporate broader knowledge to justify why specific features contribute to the model's prediction, potentially limiting the user's understanding of the decision-making process. When additional contextual information is incorporated, whether from retrieved scientific documents (RAG-only) or a combination of RAG and Web Search, the Context-Awareness scores increase substantially. Specifically, the performance rises from 4.64 in the Zero-Shot setup to 4.86 in the RAG-only configuration and further to 4.96 when Web Search is also included. A paired t-test 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, reinforcing the value of external knowledge augmentation. This result highlights that retrieved and dynamically searched information enhances the model's ability to justify predictions with more complete and causally grounded explanations.

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.

Table 3. Average correlation coefficients between LLM evaluator and human expert scores

Pearson	Spearman	Cohen's Kappa
0.825	0.791	0.583

Table 3 presents the Pearson, Spearman, and Cohen's Kappa correlation coefficients between the scores assigned by the LLM evaluator and the human expert. The reported Pearson and Spearman coefficients are an average of the correlations calculated independently for each of the four criteria. The results indicate a high level of agreement, with a Pearson correlation coefficient of 0.825, suggesting a strong linear relationship be-

tween the LLM and human judgments. Similarly, the Spearman correlation coefficient of 0.791 further confirms that the ranking of explanations based on the evaluation metrics remains consistent between the two evaluators. In addition, the Cohen's Kappa coefficient of 0.583 indicates moderate to substantial agreement between the LLM and the human expert, considering that Kappa adjusts for random agreement.

This demonstrates that LLM as a Judge is a viable alternative for evaluating XAI-generated explanations, as its assessments align closely with those of a domain expert. These findings reinforce the feasibility of automating the evaluation process, enabling scalable assessment of explanation quality while maintaining human-level reliability.

5. Conclusions

In this study, we demonstrated that a Multi-Agent architecture can significantly enhance ML explainability by moving beyond simple translation to generate causally grounded justifications. This qualitative leap in explanation quality was quantitatively validated on the Adult dataset, where our approach outperformed standard LLM explanations by 7% in Context-Awareness. Furthermore, we present a key methodological contribution by successfully validating the LLM as a Judge paradigm, which achieved over 80% correlation with human expert evaluations, establishing it as a robust and scalable assessment tool. Our results therefore highlight the effectiveness of integrating structured reasoning with contextual retrieval to make complex model insights genuinely interpretable for non-technical users.

As this study represents an initial exploration, several future directions can be pursued. For this study, the Adult dataset and XGBoost were selected as a well-established benchmark and a strong baseline, respectively, allowing us to isolate and robustly evaluate the specific contribution of our architecture. However, there are limitations of this choice, including the dataset's age and the need for future work to validate the approach on more diverse, modern datasets and with different ML models to assess the generality of our findings. For the Multi-Agent system, future work should evaluate the performance of different LLMs, optimizing their roles within the architecture. Additionally, techniques such as fine-tuning LLMs could be explored to improve adaptability. Expanding the retrieval system with more comprehensive and structured knowledge sources could also enhance the contextual richness of explanations. Regarding the LLM as a Judge framework, further validation is necessary through a broader evaluation with more human annotators possessing diverse expertise. Investigating the consistency of different LLMs as evaluators and refining evaluation metrics could improve reliability. Future work could also address the handling of categorical variables by employing human-readable labels instead of numerical codes.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. It was made possible by the Thematic Projects of FAPESP "Life cycles and aerosol clouds in the Amazon" (2017/17047-0) and "Research Centre for Greenhouse Gas Innovation - RCG2I" (2020/15230-5).

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.

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.

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

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*.

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*.

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.

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*.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W., Wei, Z., and Wen, J. (2023). A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*

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).