Explainable AI: A Retrieval-Augmented Generation Based Framework for Model Interpretability

Devansh Guttikonda, Deepika Indran, Lakshmi Narayanan, Tanishka Pasarad and Sandesh B. J.
PES University, Bengaluru, India

Keywords: Explainable AI, RAG (Retrieval-Augmented Generation), LLMs (Large Language Models), Interpretability, Data Chunking, Embedding, Query Processing, Response Generation.

Abstract: The growing reliance on Machine learning and Deep learning models in industries like healthcare, finance and manufacturing presents a major challenge: the lack of transparency and understanding of how these models make decisions. This paper introduces a novel Retrieval-Augmented Generation (RAG) based framework to tackle this issue. By leveraging Large Language Models (LLMs) and domain-specific knowledge bases, the proposed framework offers clear, interactive explanations of model outputs, making these systems more trustworthy and accessible for non-technical users. The framework's effectiveness is demonstrated across healthcare, finance and manufacturing, offering a scalable and effective solution that can be applied across industries.

1 INTRODUCTION

There is a rapid adoption of machine learning and deep learning models across industries such as Healthcare, Finance and Manufacturing. The main concern is that most of these models often function as "black boxes" where their internal decision making process is opaque to the end users as showcased in figure 1. This lack of transparency in decision making creates a challenge for non technical users who are increasingly relying on these models to automate several tasks but struggle to interpret model results and outputs, raising concerns about the accountability and usability.

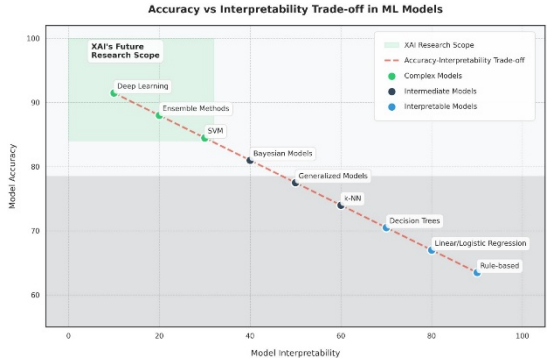

Figure 1: Interpretability of machine learning models.

Advancements have been made in this area with methodologies like SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and Partial Dependence Plots that make these complex models more interpretable as seen in Rao et al. (2022), Salih et al. (2024). These frameworks focus on post-hoc explanations that visualise the impact of individual features on the model's predictions. These methods, albeit effective, are often limited by their technical complexity and remain inaccessible to non-technical users who may not fully understand the underlying mechanics that drive the model and may not be convinced by the results. Table 1 presents a comparative evaluation of Retrieval-Augmented Generation (RAG) based Explainable AI and other traditional methods.

Industries like finance, healthcare, and manufacturing work with sensitive, proprietary data that requires strict confidentiality. AI solutions need to be secure and private for widespread adoption in these sectors. This need can be addressed by safeguarding data with private knowledge bases filled with information specific to a company's operations, clients, or research. By using private chatbots coupled with Retrieval-Augmented Generation (RAG), companies can receive insights tailored to their needs.

Guttikonda, D., Indran, D., Narayanan, L., Pasarad, T. and B. J., S.
Explainable AI: A Retrieval-Augmented Generation Based Framework for Model Interpretability.
DOI: 10.5220/0013241300003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 948-955
ISBN: 978-989-758-737-5; ISSN: 2184-433X

Table 1: Comparative Analysis of RAG-Based XAI and other methods.

Metric	RAG Based- XAI	Traditional XAI Frameworks
Focus	Broader, contextual, and interactive insights	Focused on specific features or instances
Scale	Efficient with vector retrieval for scalability	Struggles with high-dimensional datasets
Data	Can handle multi-modal data	Primarily, tabular data
Bias	Proactively addresses bias with curated embeddings	Limited attention to bias
Usage	Dynamic and conversational	Static and non-dynamic outputs

In response to this challenge, this paper introduces a novel approach using Retrieval-Augmented Generation (RAG) systems. This framework facilitates the development of Explainable Artificial Intelligence (XAI) chatbots that can be used to provide interactive, context-aware explanations of machine learning models. Additionally, by creating private knowledge bases, this approach allows industries with sensitive data to integrate tailored, confidential AI solutions. Utilising this technique, we aim to bridge the gap between technical complexity and practical usability, enabling non technical professionals to interpret and utilise machine learning models more effectively while safeguarding sensitive information.

2 RELATED WORK

There has been a huge surge in the research being conducted in the fields of Explainable Artificial Intelligence (XAI) and model interpretability in the last few years. Researchers have been exploring various methodologies such as LIME and SHAP to ensure models are more transparent and interpretable for various business use cases. With the widespread adoption of GPTs, emphasizing XAI is crucial to improve transparency and build trust in their decision-making processes. Hassija et al. (2024) outlines several approaches including salient maps, feature centrality scores and counterfactual explanations that can be applied to GPT models to improve their interpretability.

The need for different knowledge bases for different domains has led to the interest and exploration of Retrieval-Augmented Generation (RAG) systems to enhance AI interpretability, particularly in Question-Answering (QA) tasks. Novel evaluation metrics such as BLEU and ROUGE focus on the formal aspects of a generated response providing insights into text similarity and measuring n-gram overlap as mentioned in Oro et al. (2024). The paper also explores BERT and BEM scores to evaluate what the generated responses mean. These metrics provide users with an understanding of the generated outputs making them more interpretable.

Fine-tuning LLM's for QA applications is an important aspect to achieve adequate results. Some of the fine-tuning techniques highlighted by VM et al. (2023), are supervised fine tuning which leverages task-specific datasets to train models on the formats of common questions and the expected structure of answers. Prompt tuning is another approach that refines the input questions, helping the model focus on pertinent context which leads to precise answers. Responses can be enhanced by retrieving real-time data from external knowledge sources using a hybrid RAG-LLM approach.

Recent research underscores XAI's growing importance across diverse industries, providing tailored interpretability for complex, high-stakes decisions. In healthcare, XAI techniques such as SHAP, LIME, and Grad-CAM have been applied to medical imaging and diagnostics supporting clinicians as seen in Band et al. (2023). The use of XAI has been explored in finance for ensuring transparency and regulatory compliance where post-hoc explanation methods clarify model decisions in Weber et al. (2024). Branco et al. (2023) analyses different approaches for AI transparency and interpretability in manufacturing and introduces an innovative platform for manufacturing users to develop insightful XAI pipelines. These advancements indicate the rapid growth and need of XAI in various industries.

3 PROBLEM STATEMENT AND OBJECTIVES

Current Explainable AI methods like SHAP and LIME fall short in making machine learning models truly accessible to non-technical users, particularly

in achieving both model and data explainability. This presents a critical barrier to the widespread adoption and implementation of models that can help automate several tasks and provide accurate predictions. The gap lies in providing intuitive, interactive explanations that users can easily engage with. To address this gap we need a more accessible framework that can provide actionable insights, bridging the gap between technical intricacies and practical usability.

The objectives of this research are as follows :

1. To identify and analyze the key factors that hinder the adoption and trust of machine learning models among non-technical stakeholders in healthcare, finance and manufacturing.
2. To provide non-technical users with a conversational interface that enables them to understand the inner workings of models, promoting trust and transparency in decision-making.
3. To develop a framework utilising RAG technique, by embedding relevant information in Pinecone and leveraging LLMs to explain the models in a user-friendly manner , enabling the integration of private knowledge bases containing confidential, industry-specific information.
4. To demonstrate the effectiveness of this RAG based approach in simplifying model explanations and their application in healthcare, finance and manufacturing industries.

4 PROPOSED METHODOLOGY

The proposed methodology adopts a systematic approach in creating a question-answering application with domain specific data to enable model interpretability in the finance, healthcare and manufacturing domain. Figure 2 showcases a high level architecture view needed to implement this process.

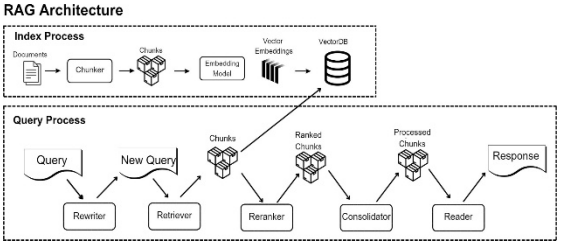

Figure 2: Proposed Methodology.

4.1 Data Chunking and Embedding

The first stage is appropriately breaking down the data into chunks. The next step is to embed these generated chunks into a vectorized format, creating a numeric representation of the data. This representation retains context and semantic meaning, which will be used for retrieval and response generation. The methods used to create these vector embeddings depends on the structure of the data used in building the knowledge base.

Textual data, such as financial documents, is often divided into chunks based on headings, content structure, and logical flow. Transformer models like GPTs and sentence transformers are commonly used to generate embeddings from these chunks, preserving the context and meaning for more effective processing and retrieval. Images on the other hand are processed using vision transformers as highlighted in Kameswari et al. (2023), where each embedding represents a different section of the image, capturing visual patterns and enabling image-based retrieval. Rows, columns and sets of attributes are used to chunk tabular data and models such as XGBoost and CatBoost are used to take feature sets and transform them into vectors that capture numerical relationships. Shwartz-Ziv et al. (2022) compares XGBoost and CATboost with deep learning models for analysing tabular data.

During the embedding process, each chunk is associated with metadata which contains important contextual information. The metadata is attached to each embedding for improved retrieval of contextual information during the retrieval phase. The final embeddings with their associated metadata are then stored in vector databases such as Pinecone or ChromaDB which are optimised for efficient retrieval and quick scalable searches.

4.2 Query Processing and Retrieval Ranking

The next phase in our methodology involves transforming user queries into a format that can be compared with our stored embeddings. This stage is a multistep process that involves query rewriting, embedding retrieval and finally, reranking embeddings to ensure that the most relevant chunks are presented to the large language model for response generation. The user's natural language query is analysed and if required, rewritten to improve clarity and alignment with our data.

This phase ensures the query adheres to the structure and domain specific terminologies present

in our knowledge base. The updated rewritten query is then transformed into a vector embedding using the same model we used on our documents. This embedding acts as a search vector, which will be compared against the instances in our vector database. Cosine similarity is one of the most widely used similarity metrics, which measures the cosine of the angle between two vectors. Other metrics such as dot product similarity, Euclidean and Manhattan distances are also used for further validation as showcased in Magara et al. (2018).

Vector databases enable fast and efficient similarity matching and are particularly useful when handling large scale data. Once the top k chunks are retrieved they are reranked and evaluated once more based on the closeness to the queries intent. This ensures that the results are ordered based on relevance and the responses generated are unlikely to be hallucinations.

4.3 Response Generation Using LLMs and Embeddings

The final step is giving a structured relevant response back to the user. This is done by synthesising the retrieved data chunks into a coherent response with the help of LLMs. A pivotal role is played by LLMs in generating natural language responses that are not only informative but are contextually aligned with the query provided by the user. The retrieved chunks are not just concatenated together but these models utilise them to generate a cohesive response. Rasool et al. (2024) validates the usage of LLM's in our use case by evaluating document-based QA on a CogTale dataset.

The enormous corpus of data that LLMs are trained on allows them to handle different data formats and queries. They are also powerful enough to eliminate contradictory or overlapping information across multiple documents by selecting the most relevant information. GPT-4 is one of the most widely used models for RAG applications because of the sheer size of the data it has been trained on in different domains. It is one of the most powerful models currently released and can handle diverse, multistep queries. LLama models are designed for efficiency and provide excellent performance with low computational complexity. They are designed to scale and can handle large volumes of queries while maintaining adequate high quality responses.

RESTful API calls are used to connect cloud based LLMs to applications as seen in Song et al. (2023). This allows easy integration, making these applications scalable and flexible in various industries. This method is also cost effective as the pricing is based on usage, making it a pay-as-you-go system. Existing infrastructure can also be integrated with minimal disruption making the RAG-LLM architecture highly desirable.

5 IMPLEMENTATION AND USE CASES

This section outlines the practical implementation of the RAG chatbot framework and the specific use cases across multiple industries namely- Healthcare, Finance and Manufacturing. In all the 3 chatbots, Gradio was utilised for the frontend interface and LLama-70b-8192 via Groq was used to generate human-like responses. Pinecone was selected for the storage of all our metadata and vector embeddings for our knowledge base, to aid with retrieval.

ChromaDB, FAISS, and Milvus are powerful vector databases but require self-hosting. Pinecone stands out as it is a fully managed service which eliminates the user having to maintain infrastructure.

Pinecone is easier to scale and automatically configures itself to handle large scale data, whereas FAISS and Milvus require fine-tuning and manual configuration. J. Bhattacharyya (2024) highlighted the qualities that made pinecone the superior choice when selecting our storage service due to its unmatched simplicity, scalability, and performance for handling large-scale vectorized data retrieval.

Gradio, an open-source Python library was used to create a user-friendly interface that seamlessly interacts with machine learning models. Its real time interactiveness is particularly useful in the case of chatbots. While selecting a large language model, we considered factors such as performance, scalability, ease of integration, and cost. GPT-3, as well as other popular models like BERT, T5, and PaLM are computationally intensive and are not efficient for large scale implementations. Jagdishbhai, N. and Thakkar (2023) explore the capabilities and limitations of GPT in natural language processing. LLama-70b-8192 was chosen for its efficiency and scalability. It also boasts 70 billion parameters and retains context particularly well for long text sequences. Llama models also excel in generative tasks especially in answering questions and Conversational AI. Table 2 provides an overview of the different models and their specifications.

Table 2: Large Language Models Specifications.

Model	Parameters	Open Source	Access
GPT-3	175 Billion	No	API
BERT	340 Million	Yes	Hugging Face
T5	11 Billion	Yes	Hugging Face
PaLM	540 Billion	No	API
LLama-70 b	70 Billion	Yes	API (Groq) / Open Source

5.1 Healthcare

In the healthcare domain, we implemented a classification model to distinguish between different types of brain tumours, namely glioma, meningioma, pituitary, and non-tumour. We employed Xception as the base model which is pre-trained on the ImageNet dataset. Xception's architecture is well-suited due to its efficient use of depthwise separable convolutions. The model was trained on a labeled dataset of MRI images with a balanced representation of all the classes.

The images used in the classification process were embedded using a Vision Transformer. These embeddings capture critical visual features needed for tumour classification. The embeddings capture both local features (such as textures, edges or small anomalies) and global structures (tumour regions, shape of the brain) (Shah D et al., 2022). Next, we used a Sentence transformer 'all-MiniLM-L6-v2' to generate the vector embeddings for the code chunks and markdown portions. These embeddings further enrich the knowledge base with relevant documentation and code explanations. All the embeddings were stored in Pinecone, with separate indexes for 768-dimensional image embeddings and 384-dimensional text embeddings.

The trained Xception model was stored as an HDF5 file. This provided a seamless user experience, which allowed healthcare professionals to interact with the model, upload MRI images, and receive both classification results and explanations.Users can query the chatbot for explanations, such as "How was this classification made?" In response, the system retrieves the relevant markdown documentation and code that explain how the Xception model processes images and identifies features for tumour classification. The chatbot utilized LLama-70b-8192 via Groq to generate human-like responses. It efficiently synthesized information retrieved from Pinecone to provide context-aware explanations to healthcare professionals, ensuring clarity for those without deep technical knowledge.

5.2 Finance

Financial documents are often very text intensive. We created an application that determines whether an individual is eligible for a loan or not based on financial statements and other attributes such as marital status, employment status, age etc. We concluded that the Logistic Regression machine learning model would be the most effective approach for predicting loan eligibility. as there is a binary outcome and the goal is to determine the probability of the outcome for each instance (Costa e Silva et al., 2020).

To determine the variables we would actually use in training the model we conducted exploratory data analysis on our dataset using libraries such as Seaborn and MatPlotLib to discover various insights through graph plots, heat maps and other visualisations. The most important features in our dataset were 'Credit_History', 'Education', and 'Gender'. The model was trained and its parameters were tweaked, yielding satisfactory results with an accuracy of 83% on our test dataset. The model was then stored in a pickle file ready for deployment.

For the generation of embeddings, we initially used BERT, a transformer model that captures the meaning of words with the context of the text surrounding it. However, this model is extremely computationally intensive and is not suitable for large datasets as it has a longer processing duration. RoBERTa and DistilBERT are variations of BERT that offer improvements for optimized outputs and enhanced performance. However, they still did not meet our specific requirements for this use case.

We discovered that the Sentence Transformer 'all-MiniLM-L6-v2' strikes an ideal balance between efficiency and performance, making it well-suited for our use case (Yin, C. & Zhang, Z., 2024). It is a model that runs on a 6 layer architecture and is designed for fast inference, which makes it highly suitable for RAG applications. The visualisations that were created were also inserted into our knowledge base using a vision transformer. Queries from users were taken in and the application provided results with a high degree of context on why Logistic Regression was selected for this application as well as the inner workings of the model for this specific use case.

5.3 Manufacturing

In the manufacturing domain, particularly Aerospace manufacturing, we developed a cost estimation tool that would help production managers accurately estimate the costs of producing aerospace parts. This is an extremely competitive domain where it is important to maintain margins and optimize bids. The tool has 2 machine learning models incorporated, one for estimating the Raw material dimensions, given the Part dimensions and another model for estimating the Price per kilogram. This is essential for reducing material wastage and optimising costs, given the high expense of materials like Titanium and Aluminium in aerospace applications.

We trained an XGBoost Model to predict the Raw Material dimensions, where it iteratively improved upon the predictions of weaker models by minimizing the prediction error through boosting. The model achieved an R^2 value of 0.91 Traditionally, given the Part dimensions, the Raw Material dimensions are estimated using the industry knowledge of the estimator. This leads to variations and discrepancies in the estimates. Standardizing this process by utilising the power of historical data and machine learning can help streamline operations and minimize human errors.

In the next phase, we utilized a Random Forest model to predict the Price per kilogram. Random Forest's ability to handle both categorical and numerical data efficiently made it an ideal choice for predicting continuous values such as manufacturing costs (Kiangala et al., 2021). The input features included the dimensions, Alloy, Spec, Temper, Quantity, Weight, and Material. The model was trained on historical manufacturing data and achieved an impressive R^2 value of 0.96.

The frontend of the tool was built using React JS and the backend utilized Flask to make function calls to the respective models. The model predictions were stored in an Excel format for further use. To support transparent and detailed explanations of cost predictions, we incorporated an XAI chatbot. The knowledge base of the RAG-based chatbot consisted of CSV files, IPYNB notebooks, and project documentation. All the textual data were embedded on Pinecone using Sentence Transformer 'all-MiniLM-L6-v2' which provided 384-dimensional embeddings. The trained XGBoost and Random Forest models were stored as pickle files. Users could interact with the chatbot powered by Llama-70b-8192 and ask several technical questions regarding the working of the tool and its overall reasoning. The main challenge we faced was LLM hallucination, where the model lacked precise context on which data corresponded to which model, leading to inaccurate or irrelevant explanations. To address this issue, we introduced additional project documentation and markdown chunks into the knowledge base. By embedding detailed markdowns that clearly outlined the scope, data sources, and specific models used for each task, the LLM was able to reference the correct context and provide accurate, relevant, and reliable responses. This not only reduced the hallucinations, but also improved the overall explainability of the system.

6 RESULTS

To evaluate the effectiveness of our RAG LLM chatbots across different domains- healthcare, finance, and manufacturing, we conducted a comprehensive evaluation using a variety of metrics (Hu T. & Zhou XH., 2024).

BERTScore evaluates the semantic similarity between the chatbot's responses and the reference answers, using pre-trained transformer models to capture embeddings of generated text and the ground truth. Table 3 outlines the results we achieved using BERTScore.

Table 3: BERT Scores.

Industry	Query	F1 Score
Healthcare	Explain how The model classifies Non-tumour cases?	0.8546
Finance	Which data visualization frameworks have been used to analyze the dataset?	0.7879
Manufacturing	Provide me with the most commonly occurring raw material used to manufacture parts in the given dataset.	0.5684

We used the ROUGE metric to evaluate the overlap of n-grams, word sequences, and longest common subsequences between the chatbot's generated responses and reference texts. Table 4 presents the results we achieved using the ROUGE metric.

Table 4: ROUGE Scores.

Industry	Query	Rouge 1 Score	Rouge 2 Score
Healthcare	Can you explain the accuracy metrics of the model and the cases it can misclassify ?	0.7134	0.6423
Finance	Which models are suitable for loan eligibility prediction	0.7755	0.5957
Manufacturing	Give me the densities for all available materials.	0.9189	0.8938

Another critical metric was response time, as the efficiency of retrieval and generation impacts the user experience, especially in high-stakes domains like healthcare and finance. Table 5 presents the results obtained from evaluating our framework.

Table 5: Response Time.

Industry	Time
Healthcare	4.2s
Finance	7.7s
Manufacturing	5.6s

We also utilized Perplexity, which is a commonly used metric in natural language processing to measure how well a language model predicts a given sequence of text. It measures the inverse probability of the model's generated responses, normalized by the number of words in the response. Equation 1 provides the mathematical calculation for Perplexity and Table 6 outlines the results achieved.

$$Perplexity = 2^{-\frac{1}{N}\Sigma log\, P(w_i)} \quad (1)$$

7 CONCLUSION AND FUTURE SCOPE

The integration of RAG with XAI allows the development of applications that generate accurate and contextually relevant responses that provide valuable business insight. With this architecture, industries can have greater trust in AI driven decisions as demonstrated through applications such as cost estimation in manufacturing, brain tumour classification in healthcare, and loan eligibility predictions. The proposed framework achieved satisfactory results over various metrics such as perplexity, BERT and ROUGE scores with adequate overlap between the generated responses and the reference text provided for each prompt. This cross-industry exploration demonstrates the flexibility and potential of RAG models.

Table 6: Perplexity Scores.

Industry	Query	Score
Healthcare	Explain how the model classifies the different types of tumours?	10.85
Finance	What has been the methodology used to predict loan eligibility? Explain.	11.30
Manufacturing	Describe the dataset in your knowledge base and the models used for price prediction.	6.79

Expanding the applicability of our proposed framework to big data and larger datasets could introduce challenges such as increased hallucinations and biased embeddings. This occurs due to the LLM's reliance on imperfectly retrieved embeddings. Hierarchical clustering, optimized vector-based retrieval methods, and incorporating additional contextual metadata can combat the issues faced when handling larger and more complex datasets.

Some of the key methods that can be employed to combat the complexity of large datasets are Approximate Nearest Neighbor (ANN) search techniques, such as FAISS and ScaNN, which improve retrieval efficiency by reducing the search space through clustering or quantization. Hierarchical clustering works in a similar fashion by narrowing the search space using tree-based indexing methods like Ball Trees or KD-Trees. Additionally, query rewriting and preprocessing enhances the alignment of user queries with the stored embeddings, ensuring more accurate and contextually relevant results.

While our work outlines the foundational steps towards industry-specific RAG applications with XAI, it also highlights the importance of future innovations in this domain. As vision and sentence

transformers are further developed, the context and accuracy of retrieved embeddings will only improve. These improvements will enable better responses as the most similar chunks with respect to our natural language prompt, will be retrieved from our knowledge base. The research conducted in the field of LLMs has surged tremendously in the last few years and the natural language processing and Generative AI capability is expected to develop significantly offering more precise and human-like responses. As customization and fine-tuning of these models continue to advance, this architecture will be able to seamlessly integrate and cater to specialised domain specific use cases. Future work in this domain will focus on optimizing retrieval mechanisms, developing more intuitive explainability frameworks, and integrating these systems seamlessly into existing business workflows.

REFERENCES

Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, *vol. 40*, p. 101286

Bhattacharyya, J. (2024). A Brief Comparison of Vector Databases - CodeX - Medium. Medium. Retrieved October 25, 2024, from https://medium.com/codex/a-brief-comparison-of-vect or-databases-e194dedb0a80

Branco, R., Agostinho, C., Gusmeroli, S., Lavasa, E., Dikopoulou, Z., Monzo, D., & Lampathaki, F. (2023). Explainable AI in manufacturing: an analysis of transparency and interpretability methods for the XMANAI platform. In *2023 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pp. 1-8, IEEE.

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics*, *vol. 47,* pp. 2879-2894.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *vol.16*, pp. 45-74.

Hu, T., & Zhou, X. H. (2024). Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. *arXiv preprint arXiv:2404.09135*.

Jagdishbhai, N., & Thakkar, K. Y. (2023). Exploring the Capabilities and Limitations of GPT and ChatGPT in Natural Language Processing. *J. Manag. Res. Anal*, *vol. 10*, pp. 18-20.

Kameswari, C. S., Kavitha, J., Reddy, T. S., Chinthaguntla, B., Jagatheesaperumal, S. K., Gaftandzhieva, S., & Doneva, R. (2023). An overview of vision transformers for image processing: A survey. *International Journal of Advanced Computer Science and Applications*, *vol. 14*.

Kiangala, S. K., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications*, *vol. 4*, p. 100024.

Magara, M. B., Ojo, S. O., & Zuva, T. (2018). A comparative analysis of text similarity measures and algorithms in research paper recommender systems. In *2018 conference on information communications technology and society (ICTAS)*, pp. 1-5, IEEE.

Oro, E., Granata, F. M., Lanza, A., Bachir, A., De Grandis, L., & Ruffolo, M. (2024). Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models

Rao, S., Mehta, S., Kulkarni, S., Dalvi, H., Katre, N., & Narvekar, M. (2022). A study of LIME and SHAP model explainers for autonomous disease predictions. In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1-6, IEEE.

Rasool, Z., Kurniawan, S., Balugo, S., Barnett, S., Vasa, R., Chesser, C., ... & Bahar-Fuchs, A. (2024). Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset. *Natural Language Processing Journal*, p. 100083.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, p. 2400304

Shah, D., Shah, D., Jodhawat, D., Parekh, J., & Srivastava, K. (2022). Xception Net & Vision Transformer: A comparative study for Deepfake Detection. In *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*, pp. 393-398, IEEE.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *vol. 81*, pp. 84-90.

Song, Y., Xiong, W., Zhu, D., Wu, W., Qian, H., Song, M., ... & Li, S. (2023). RestGPT: Connecting Large Language Models with Real-World RESTful APIs. *arXiv preprint arXiv:2306.06624*.

VM, K., Warrier, H., & Gupta, Y. (2024). Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations. *arXiv preprint arXiv:2404.10779.*

Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, *vol. 74*, pp. 867-907.

Yin, C., & Zhang, Z. (2024). A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With "Same Semantics, Different Structure" After Fine Tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pp. 677-684. Atlantis Press.