Adaptive Explainable AI: Designing User-Centric Explanation Systems for Enhanced Interaction

1st Xinyi Liu
xinyi.liu@utexas.edu

I. Introduction

Explainable Artificial Intelligence (XAI) is crucial for enhancing trust and understandability in AI applications. However,traditional XAI approaches often apply a uniform method to explain AI decisions, which may not suit all users or scenarios. Recognizing the diversity in users' informational needs, this work proposes a novel system that dynamically selects and provides the most appropriate XAI explanation based on specific user queries.

The system integrates a chatbot interface which engages users in selecting an image and receiving an AI-generated prediction, which users can further question to obtain deeper insights. The system supports multiple explanation modalities, including SHAP, LIME, show-more and counterfactuals, facilitating a broad range of user interactions and enhancing the personal relevance of the explanations.

II. Related Work

A. Techniques and Algorithms

Common XAI techniques include:

SHAP (SHapley Additive exPlanations) [1]: Originating from cooperative game theory, SHAP quantifies the contribution of each feature to a prediction. This method allows for both local and global explanations and is applicable across different types of machine learning models, from deep neural networks to simpler logistic regression models. Despite its versatility and depth, SHAP can be computationally intensive, especially when dealing with large datasets or models with numerous features.

LIME (Local Interpretable Model-agnostic Explanations) [2]: LIME explains the predictions of any classifier in an interpretable and faithful manner by approximating it locally with an interpretable model. By perturbing the input data and observing the resultant changes in outputs, LIME highlights influential features at the local level. While LIME provides valuable insights, its dependency on local perturbations can sometimes lead to inconsistent or misleading explanations.

Counterfactual Explanations [3]: Counterfactual explanations focus on identifying the smallest change needed to an input to change the output of a model. Essentially, they provide an example of an alternative scenario that would lead to a different decision. For instance, in a loan application model, a counterfactual explanation might show that decreasing the loan amount or increasing the applicant's annual income by a certain amount would result in the loan being approved, given all other factors remain constant.

B. Human-Centered XAI Study

In response to the need for more intuitive and user-friendly XAI, recent research has emphasized the importance of human-centered approaches for language models or writing tasks. Lai et al. [4] introduced the concept of selective explanations, where AI systems generate explanations aligned with user preferences by incorporating human input. This method addresses the gap between how AI systems and humans interpret explanations, making AI decisions more comprehensible and relevant to users. Similarly, Shen et al. [5] explored the effectiveness of delivering AI explanations through conversational interfaces in ConvXAI, which supports human-AI collaborative tasks such as scientific writing.

C. Challenges in Existing XAI Approaches

1) Hard to Personalization: Most existing XAI solutions employ a one-size-fits-all approach, providing standardized explanations that do not account for the *diverse user needs* (technical expertise) and *contextual relevance* (varied backgrounds). For the selective prediction [4] and ConvXAI methods [5], they require collecting questions from humans to build the prediction model, which is time-consuming and expensive.

Identify applicable funding agency here. If none, delete this.

2) Lack of Specificity in Explanations: A further limitation in conventional XAI systems is the generic nature of explanations, which fail to provide specific, actionable insights tailored to the unique aspects of the input data. Traditional methods like LIME or SHAP generally produce visualizations or attribute importance scores that indicate which features influenced a model's decision. However, these explanations typically lack detail on how specific attributes of these features—such as appearance, color, or spatial relationships in image data—contribute to the model's output. This generalization can make it difficult for users to understand and act upon these explanations effectively, particularly in domains where precise and detailed interpretative feedback is crucial.

D. Adaptive Systems in AI

Adaptive systems in AI are designed to dynamically modify their operations based on feedback or changes in their environment. This adaptability enables these systems to offer personalized experiences that are tailored to the individual needs and preferences of users. By leveraging techniques such as machine learning and data analytics, adaptive systems can learn from interactions and improve their accuracy and relevance over time, enhancing user engagement and satisfaction. Examples include adaptive learning systems in education that adjust content based on a student's learning pace, and personalized recommendation systems in e-commerce that tailor suggestions to individual user preferences.

E. Solution - Adaptive XAI (XAI4Adapt)

Inspired by adaptive AI systems and current challenges in existing XAI approaches, this report proposes **"XAI4Adapt"** which is designed to overcome these limitations by providing not only tailored explanations that align with the specific expertise and contextual needs of each user but also enhancing the specificity of the explanations it generates. By leveraging the power of Large Language Models such as GPT-4, we can easily generate a large number of possible questions and use them to train a prediction model, eliminating the huge human survey cost. Also, the explanation for images can be further adapted to Vision GPT, where a detailed, personalized sentence is generated for supporting non-language input such as images. Therefore, by integrating advanced interpretation techniques that delve deeper into the attributes of influential features, XAI4Adapt offers users detailed insights into the 'why' and 'how' of model decisions.

III. XAI4ADAPT

A. Design Goal of XAI4Adapt

The rapid advancement of artificial intelligence in various domains has heightened the necessity for systems that not only make predictions but also explain them in a manner accessible to all users. The XAI4Adapt system is designed to meet this need by providing tailored, understandable explanations based on questions related to the AI prediction results from users. Therefore, the design goals of XAI4Adapt include the following features:

1) *Prediction*: XAI4Adapt aims to accurately predict labels based on user inputs through a specific AI model. A model trained on a focused dataset is employed as the predictive engine in XAI4Adapt. Users select images from a test set to assess the prediction accuracy of the system.
2) *Explanation*: XAI4Adapt is able to give the user the reason why it outputs the specific prediction label. This is achieved by feeding the input figure, the AI model, and the output prediction into different explanation methods. To cater to diverse user questions with appropriate explanations, XAI4Adapt is equipped with multiple explanatory techniques, guided by a fine-tuned classification language model. This model directs the system to select the most suitable explanation method in response to each user query. To showcase its capability in supporting multiple explanation methods, XAI4Adapt includes the following algorithms:
   - *LIME*: Identify the part of the figure that contributes most to the final prediction label.
   - *SHAP*: For each small pixel region of the input figure, find out the contribution (both positive and negative) of this region to the final prediction.
   - *Counterfactuals*: Given the input figure and its prediction label of the AI model, this algorithm will output the most possible label for a false-prediction.
   - *Show-more*: This algorithm will show the user with more figures with the same label.

B. Chatbot Interaction Workflow

Figure 1 illustrates the workflow of the XAI4Adapt chatbot. It is developed with the streamlit framework [6] and appears as a chatbot. It utilizes a gallery of images picked from "Bird 525 Species" dataset [7]. Users are presented with a selection of 6 images from this gallery, from which they choose one to use as the input of AI prediction model. Following their selection, the InceptionV3 model [8]is employed to predict the label of the chosen bird figure. After the user gets the prediction result from the AI model, they can ask any questions about the AI prediction result through a chatbox inside the chatbot. To classify the user's input, I fine-tuned the DistilBert [9] model with a generated dataset about possible questions from users. For each specific user query, the system will pick the most appropriate XAI explanation algorithm from the four XAI explanation algorithms (LIME, SHAP, Counterfactuals and Show-more) to answer the question.

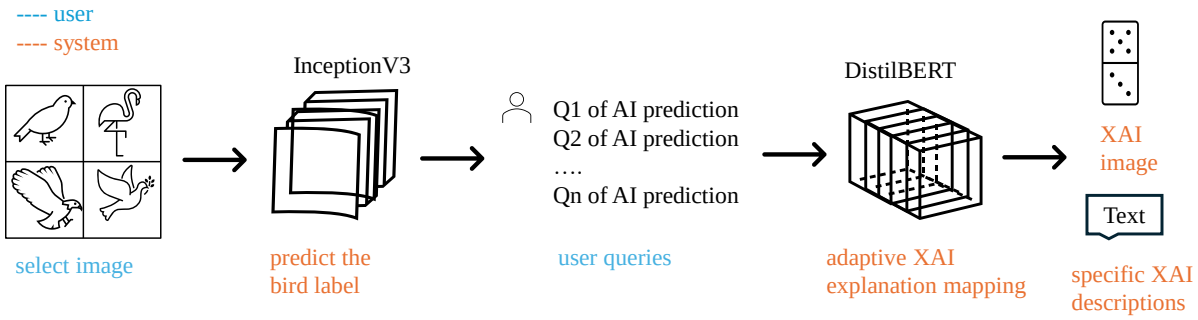

Fig. 1. XAI4Adapt Design Workflow

C. XAI4Adapt Design Details

In this section, I will show the design details of my proposed XAI4Adapt. I will first introduce the bird dataset details as well as how to select the subset which are used in XAI4Adapt. Then I will introduce the InceptionV3 and DistilBERT model as well as how these models are trained and used in XAI4Adapt. Finally, I will illustrate the demonstrations for each explanation algorithm adopted in XAI4Adapt.

1) Dataset and Analyses: The "Birds 525 Species" image classification dataset from Kaggle was chosen to demonstrate the capabilities of XAI4Adapt within the context of a chatbot interface. This dataset is particularly suitable because it presents complex, real-world data that can benefit greatly from enhanced explainable AI techniques. Image classification serves as an ideal use case for this system due to its broad applicability in various AI applications and its inherent challenges that necessitate detailed explanations for model decisions. By utilizing images, the system can leverage visual explanations, which are intuitive and easy for users to understand, thereby effectively demonstrating the adaptability and effectiveness of XAI4Adapt.

For the purpose of this report, the dataset was narrowed down to 100 species, encompassing 15,849 images in training set and 500 images in test set. This subset provides a manageable yet diverse collection of images, allowing the system to demonstrate robustness across varied examples while maintaining computational efficiency. For demonstration in XAI4Adapt, we randomly pickup 2 images per class and used them as demonstration dataset.

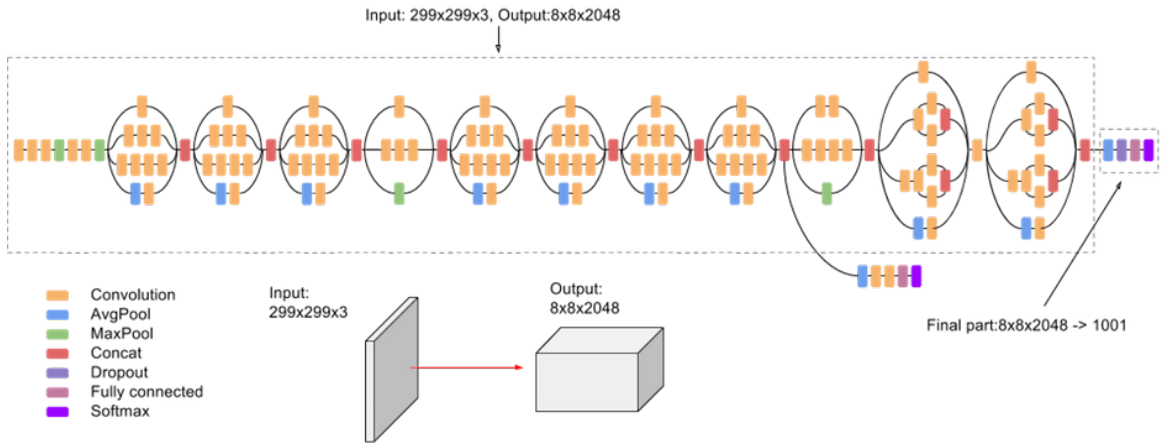

Fig. 2. InceptionV3 model architecture

2) InceptionV3: InceptionV3 is a powerful convolutional neural network that excels in image classification tasks. It is a convolutional neural network that uses a complex architecture of inception modules which include various sized convolution filters and pooling layers within the same module. The detailed model architecture is shown in figure 2. This architecture not only helps in capturing features at various scales but also significantly reduces the number of parameters, making it efficient. The model's ability to handle intricate image data makes it an excellent choice for a system that aims to provide deep insights into the AI's decision-making process.

Since the demonstration dataset only contains 100 sub-classes, the original pre-trained model which outputs 525 classes is not accurate within some explanation algorithms, such as Counterfactuals. To resolve this issue, I changed the original

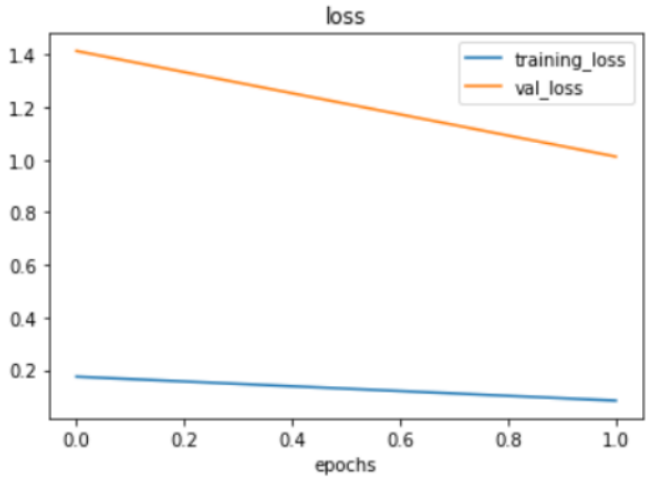

Fig. 3. Loss of Inceptionv3 Finetune

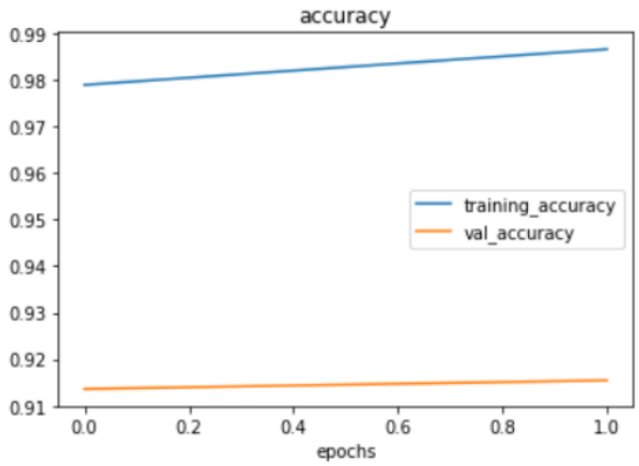

Fig. 4. Accuracy of Inceptionv3 Finetune

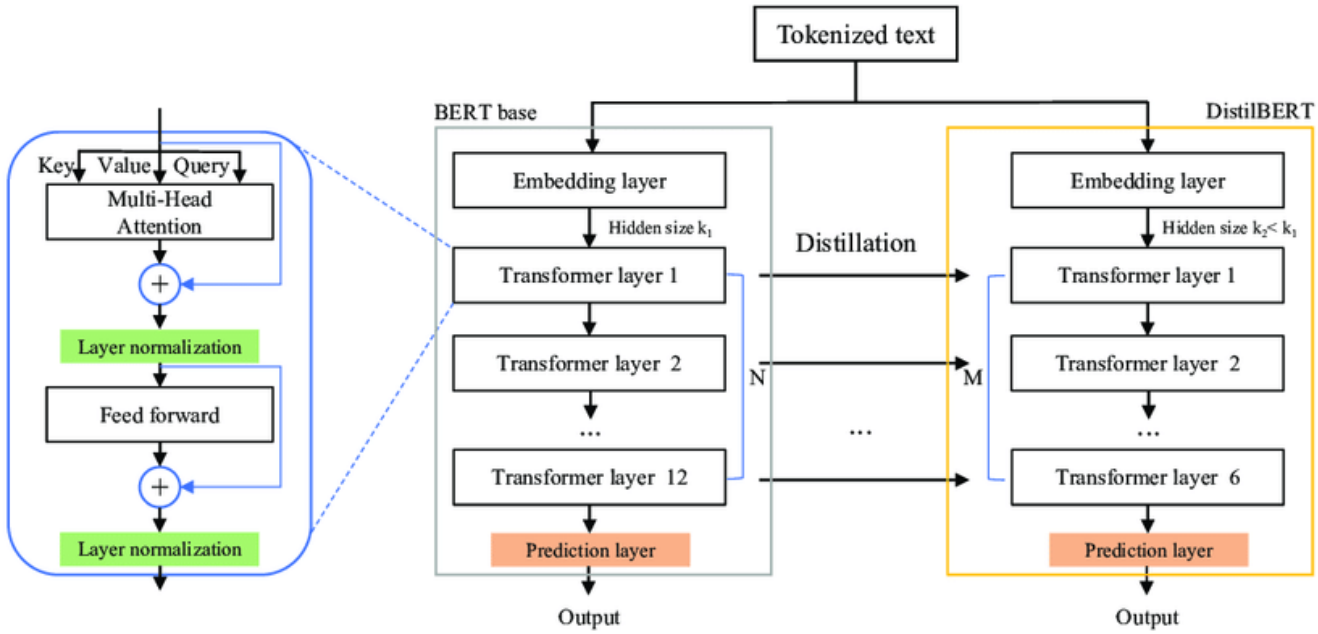

Fig. 5. DistilBERT model architecture

pre-trained InceptionV3 model by changing the last layer to a classifier with 100 classes. Weights of the hidden layers are frozen and only the last layer (the new 100-class classifier) is set to trainable. Finally, the model is fine-tuned on my training dataset as stated before to generate a better result under these 100 classes. The performance of finetune is shown in figure 3 and figure 4

3) DistilBert: DistilBERT is a distilled version of BERT (Bidirectional Encoder Representations from Transformers), which revolutionized the field of natural language processing with its deep bidirectional architecture shown in Figure 5. DistilBERT was created to provide a lighter, faster alternative to BERT, retaining most of its predecessor's ability to understand the context of words in a sentence while significantly reducing the model size and computational demands. This makes DistilBERT an excellent choice for applications that need real-time performance without a substantial loss in effectiveness, such as this XAI4Adapt system. The model's architecture and training process involve a technique called distillation, where DistilBERT is trained to replicate BERT's performance using less computational resources.

4) Adaptive XAI Explanation Mapping: The adaptive explanation component of XAI4Adapt is crucial for tailoring responses to user queries effectively. To enhance the system's capability in this area, GPT-4, the state-of-the-art language model, was

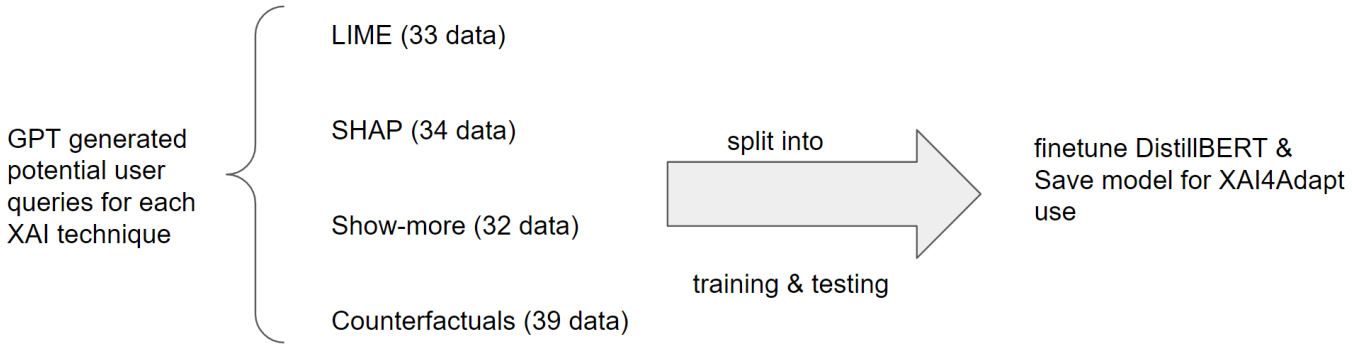

Fig. 6. DistilBERT Model Finetune

employed to generate a comprehensive set of potential user queries corresponding to each XAI explanation. A total of 131 distinct queries were created, reflecting the diverse ways in which users might seek explanations about the predictions made by the system.

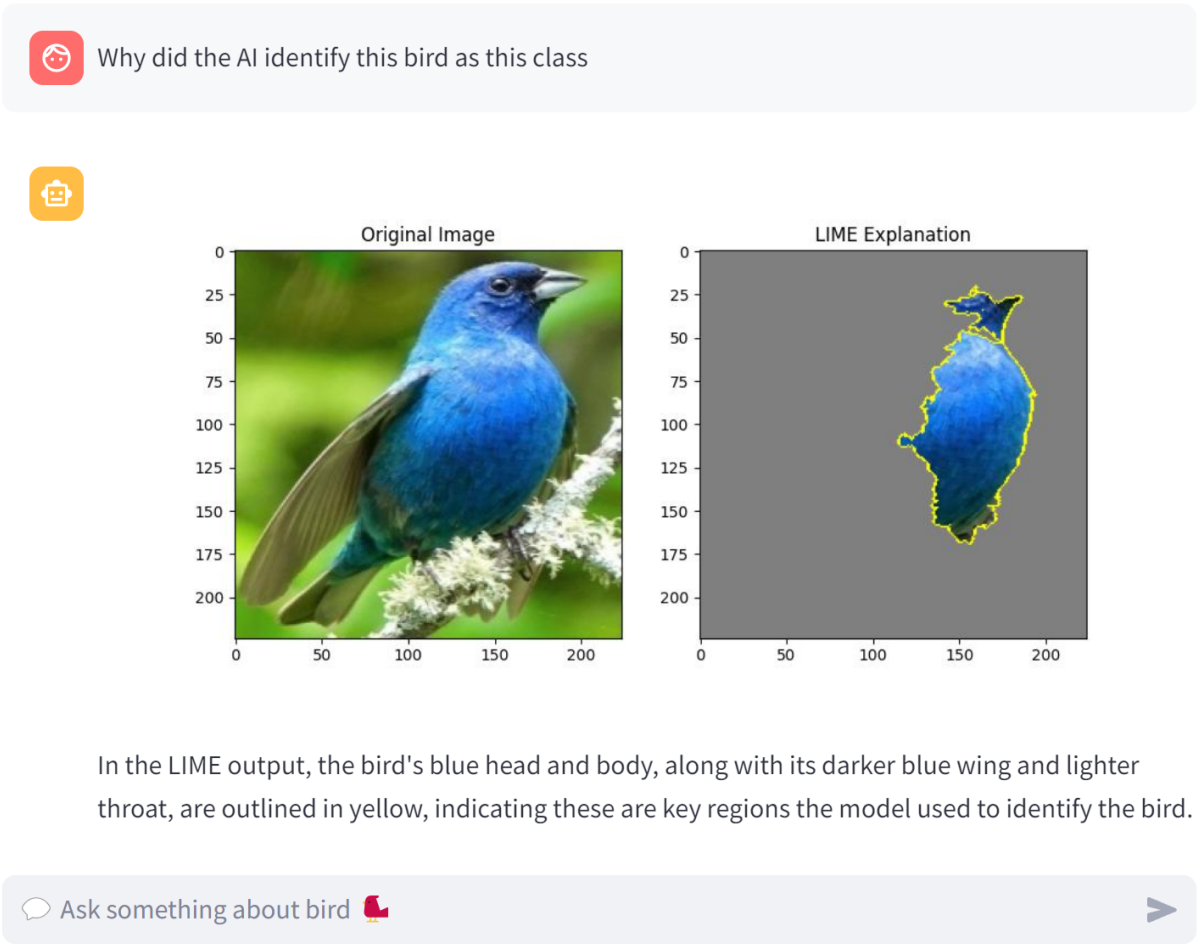

Fig. 7. LIME output with the user query

These generated queries were then used to finetune DistilBERT, ensuring that the model could accurately associate each query with the most suitable type of XAI explanation, which is shown in figure 6. This finetune process involved labeling each GPT-generated query with the corresponding explanation type—such as LIME, SHAP, Counterfactuals, and Show-more. By integrating these labeled queries into the training dataset, DistilBERT was fine-tuned to enhance its ability to discern and

match the nuances of user inquiries to the appropriate explanation type.

The effectiveness of this approach was quantitatively assessed, with the fine-tuned DistilBERT achieving an accuracy of approximately 90% in correctly mapping user queries to the corresponding XAI explanation types.

*5) **Results - Specific XAI Explanation**:* The system includes four types of XAI explanations: SHAP, LIME, show-more, and counterfactuals. These explanations are integrated into the system to enhance user understanding by providing visually and contextually rich information about the model's predictions.

LIME: Figure 7 shows the demonstration of LIME model. XAI4Adapt will display a modified version of the image where important features for the model's prediction are highlighted. This includes outlining or coloring specific parts of a bird that led to its classification. When LIME is selected based on the user's query, the system not only shows these localized importance maps but pairs them with a descriptive analysis provided by the Vision GPT API, offering a deeper understanding of the localized features that guide the AI's decisions.

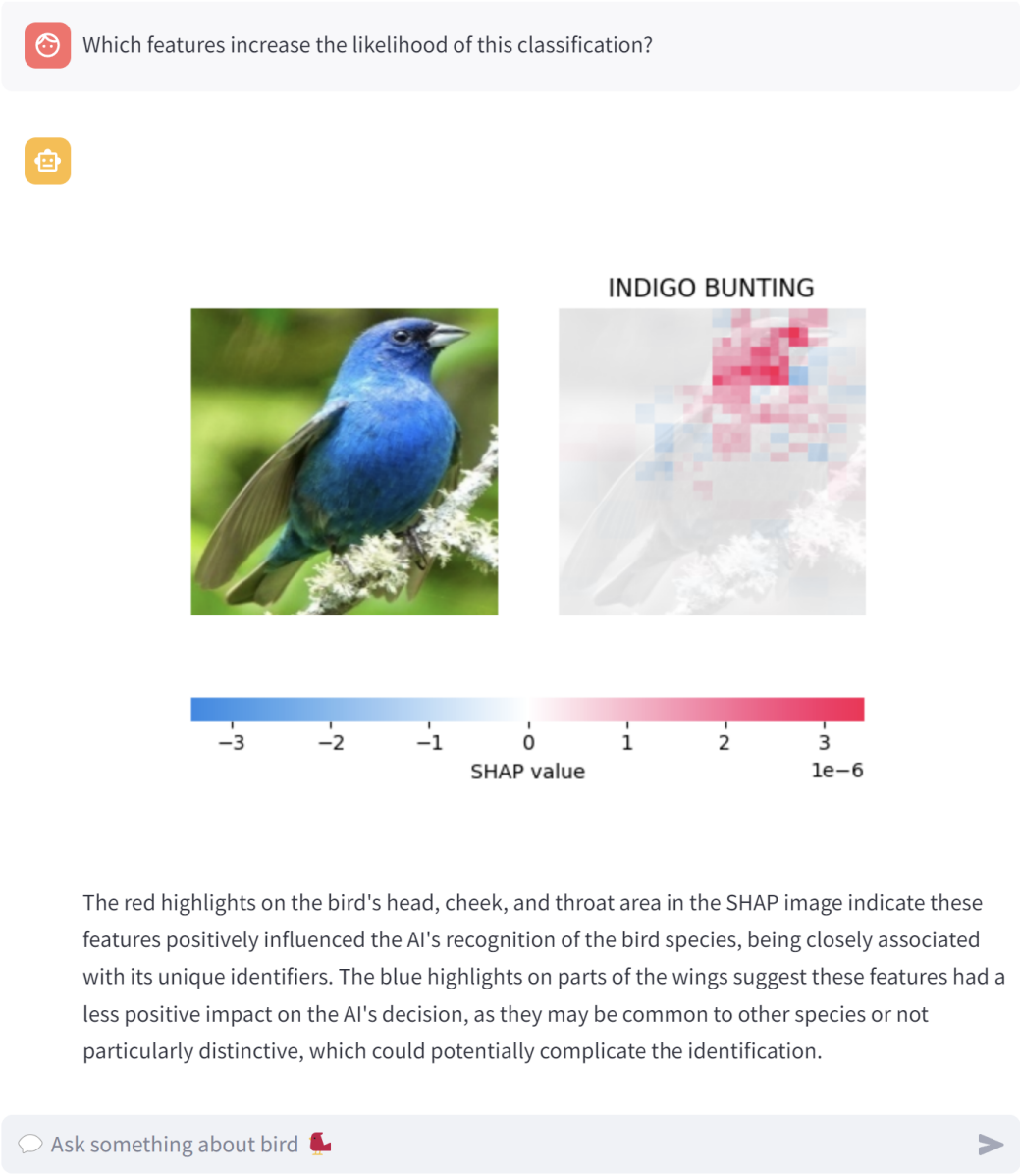

Fig. 8. SHAP output with the user query

SHAP: Figure 8 illustrates the demonstration of SHAP model. SHAP generates visual outputs highlighting the important features that influenced the model's prediction. Areas marked in red indicate features that positively influence the prediction—meaning these features push the model's output towards a particular class. Areas in blue negatively impact the prediction,

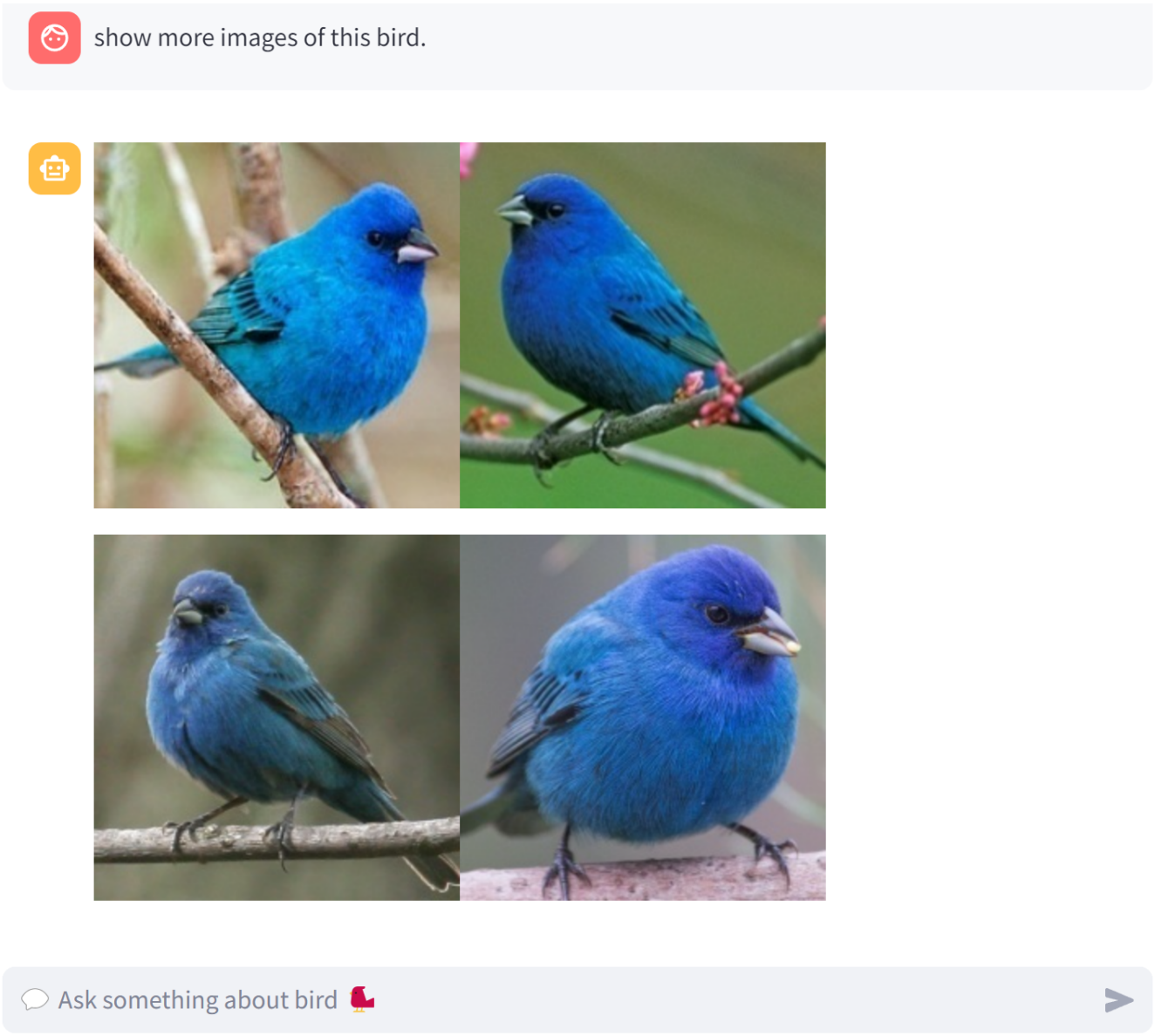

Fig. 9. Show-more output with the user query

pushing the output away from a particular class. Each colored region is associated with a SHAP value, quantifying the magnitude of impact. High absolute values (whether positive or negative) signify a strong influence on the model's decision. If SHAP is selected based on the user's query, the system not only displays the feature importance visualization but also pairs it with the original image of the bird. The combined visual output is then processed by the Vision GPT API, which provides a descriptive analysis of the important features identified in the XAI visualization. This dual display of image and descriptive text enhances the user's understanding of the factors contributing to the AI's decision.

Counterfactuals: Figure 10 illustrates the demonstration of the Counterfactuals algorithm. When selected, this explanation shows the most similar bird species that could have been predicted instead, effectively offering insight into what might change if the model's predictions were different. This helps users understand the model's decision boundaries and alternative outcomes.

Show-more: Figure 9 illustrates the demonstration of the Show-more algorithm. This option is triggered if the user expresses interest in seeing more images of the same bird species. It provides additional images from the dataset, offering a broader visual context and aiding in a deeper appreciation of the species' variability.

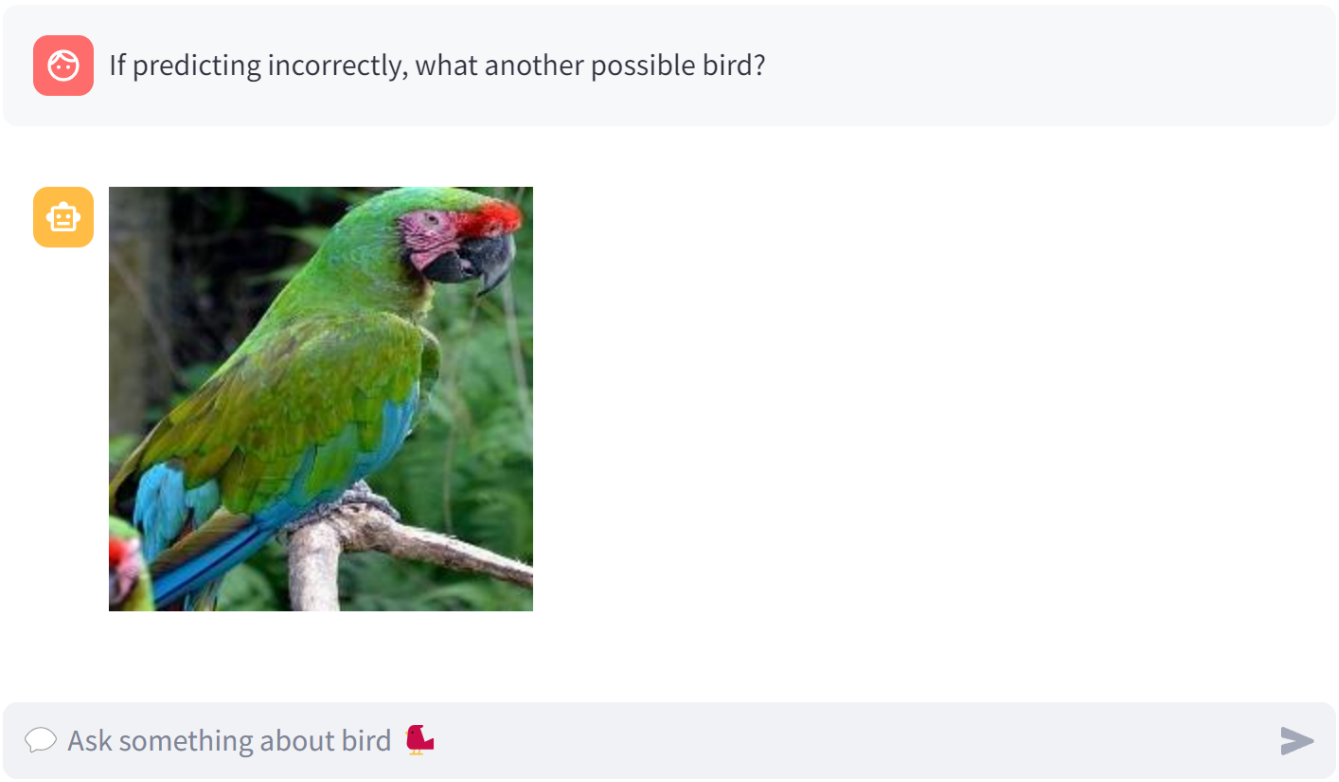

Fig. 10. Counterfactuals output with the user query

IV. Discussion

This report introduced XAI4Adapt, a novel system designed to enhance user interactions with AI through adaptive and specific explainable artificial intelligence explanations. The research aimed to address the limitations of current XAI systems, which often provide generic or overly technical explanations that may not meet the needs of all users. Besides such implications, there are also some limitations.

A. Limitations

1) Domain-Specific Focus: The study focused exclusively on one domain (bird species identification), which may not fully capture the complexities and requirements of other domains where XAI could be applied. Future research could explore the applicability of the XAI4Adapt system across various fields such as medical diagnostics, financial services, or customer service to evaluate its versatility and effectiveness in different contexts.

2) Limited Interaction Scenarios: The interactions were limited to predefined phases and may not fully represent real-world use cases where user interactions can be more dynamic and less structured. Further research could investigate more natural interaction flows and longer-term system use.

3) Data Type Limitation: The current study only utilized image classification data. This restricts the understanding of the system's effectiveness in handling other data types, such as tabular data, which is prevalent in many real-world AI applications. Future research should include experiments with tabular and possibly even text data to assess the adaptability and effectiveness of XAI techniques across various data formats.

B. Future Research Directions

1) Enhance System Responsiveness and Intelligence: Future iterations of the system could incorporate advanced natural language processing capabilities to better understand and respond to a wider range of user queries, especially those posed in non-technical language. This would make the system more accessible to users without a background in AI.

2) Expand the Knowledge Base: Integrating a broader knowledge base covering non-AI topics, as suggested by participant feedback, would enhance the system's utility and educational value. Research could focus on developing methods to seamlessly integrate diverse datasets to enrich the system's explanations.

V. Conclusion

As AI continues to permeate various aspects of daily life, the importance of making AI decisions understandable and trustworthy cannot be overstated. XAI4Adapt represents a step towards more transparent, user-friendly AI systems, but much work remains to be done. Future research should explore additional data types, domains, and user demographics to further refine and validate the approach.

This thesis ultimately argues for a paradigm shift in how we develop and implement AI systems that it is important to place user interaction and understanding at the forefront of technological advancements.

REFERENCES

[1] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.
[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.
[3] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: Challenges revisited," 2021.
[4] V. Lai, Y. Zhang, C. Chen, Q. V. Liao, and C. Tan, "Selective explanations: Leveraging human input to align explainable ai," 2023.
[5] H. Shen, C.-Y. Huang, T. Wu, and T.-H. K. Huang, "Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing," 2023.
[6] "Streamlit • A faster way to build and share data apps," Jan. 2021. [Online]. Available: https://streamlit.io/
[7] "BIRDS 525 SPECIES- IMAGE CLASSIFICATION." [Online]. Available: https://www.kaggle.com/datasets/gpiosenka/100-bird-species
[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.
[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.