Conversational Explanations: Discussing Explainable AI with Non-AI Experts

Tong Zhang
College of Computing and Data Science
Nanyang Technological University
Singapore, Singapore
tong024@e.ntu.edu.sg

Mengao Zhang
College of Computing and Data Science
Nanyang Technological University
Singapore, Singapore
zh0024ao@e.ntu.edu.sg

Wei Yan Low
Nanyang Technological University
Singapore, Singapore
p180026@e.ntu.edu.sg

X. Jessie Yang
College of Engineering
University of Michigan
Ann Arbor, Michigan, USA
xijyang@umich.edu

Boyang Albert Li
College of Computing and Data Science
Nanyang Technological University
Singapore, Singapore
libo0001@gmail.com

Abstract

Explainable AI (XAI) aims to provide insights into the decisions made by AI models. To date, most XAI approaches provide only one-time, static explanations, which cannot cater to users' diverse knowledge levels and information needs. Conversational explanations have been proposed as an effective method to customize XAI explanations. However, building conversational explanation systems is hindered by the scarcity of training data. Training with synthetic data faces two main challenges: lack of data diversity and hallucination in the generated data. To alleviate these issues, we introduce a repetition penalty to promote data diversity and exploit a hallucination detector to filter out untruthful synthetic conversation turns. We conducted both automatic and human evaluations on the proposed system, fEw-shot Multi-round ConvErsational Explanation (EMCEE). For automatic evaluation, EMCEE achieves relative improvements of 81.6% in BLEU and 80.5% in ROUGE compared to the baselines. EMCEE also mitigates the degeneration of data quality caused by training on synthetic data. In human evaluations ($N = 60$), EMCEE outperforms baseline models and the control group in improving users' comprehension, acceptance, trust, and collaboration with static explanations by large margins. Through a fine-grained analysis of model responses, we further demonstrate that training on self-generated synthetic data improves the model's ability to generate more truthful and understandable answers, leading to better user interactions. To the best of our knowledge, this is the first conversational explanation method that can answer free-form user questions following static explanations.

This work is licensed under a Creative Commons Attribution 4.0 International License.
IUI '25, Cagliari, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712143

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Collaborative and social computing systems and tools**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Explainable AI (XAI), Conversational AI, Human-AI Interaction

ACM Reference Format:
Tong Zhang, Mengao Zhang, Wei Yan Low, X. Jessie Yang, and Boyang Albert Li. 2025. Conversational Explanations: Discussing Explainable AI with Non-AI Experts. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3708359.3712143

1 Introduction

Despite the high accuracy of deep neural networks (DNNs), it remains necessary for human domain experts to verify the DNN decisions and examine the reasoning process to prevent catastrophic failures in high-stake and mission-critical applications like healthcare, finance, and law enforcement. [9, 69]. To this end, much research in recent years has been devoted to eXplainable Artificial Intelligence, or XAI (e.g., [10, 59, 76]).

However, most current XAI techniques provide one-off, static explanations that are not customized to the user. As users differ in their knowledge levels, as well as tasks or goals that they try to accomplish, they will inherently have different information needs [21, 30, 51, 86]. Existing static XAI methods fail to address these diverse needs, leading to users' insufficient understanding of model behavior and undermining human-AI collaboration [51, 52, 67, 98]. Indeed, recent studies found that the end users and domain experts with limited machine learning knowledge struggle to understand and use the XAI explanations [21, 86].

Conversational explanations, or conversational XAI, have been suggested as a suitable solution for providing customized explanations to users [23, 48, 51, 98]. Lakkaraju et al. [48] discovered that human decision-makers have a strong preference for explanations in the form of natural language dialogue. They argued that

conversational explanations can provide personalized responses and relevant information based on users' conversational histories. Zhang et al. [98] showed that answering users' follow-up questions after providing static explanations significantly improves their comprehension, acceptance, trust, and collaborative decision-making with AI. While the need for conversational XAI has been recognized, building such systems is hindered by data scarcity, partially due to the difficulty of collecting high-quality conversations about AI explanations. As far as we are aware, there is only one dataset of 60 conversations focused on two types of static explanations [98]. To date, existing conversational explanations rely on human-authored templates, which can only handle a limited and predefined range of user questions [77, 82].

To handle data scarcity, we propose a novel method in this work to develop conversational explanations by training large vision language models (VLMs) on synthetic conversations. However, training with synthetic data encounters two primary challenges: the lack of data diversity in self-generated data [8, 73, 79], and the hallucinations generated by VLMs [6, 14, 39, 49, 99]. The first challenge, lack of data diversity, arises as generative models tend to overrepresent high-frequency content [8, 73, 79] and suppress the tails of the data distribution. To alleviate this issue, we introduce a repetition penalty that reduces the frequency of tokens existing in previously generated conversations. The other obstacle is the hallucination in generated conversations. VLMs often suffer from generating untruthful information, referred to as hallucination [6, 14, 39, 49, 99]. To mitigate the hallucinated, factually incorrect answers, we trained a hallucination detector to filter out such conversation turns after data generation. To train the detector, we collected a hallucination dataset of 750 factual and 750 incorrect statements about basic machine learning and XAI methods.

We conducted both automatic and human evaluations on the proposed system, fEw-shot Multi-round ConvErsational Explanation (EMCEE), to assess its performance. The automatic evaluation is conducted on the only existing conversational explanation dataset [98]. We assessed the performance of different conversational XAI systems by measuring the word overlap between generated responses and ground truth texts.

For the human evaluation, we evaluated how different conversational XAI systems assist users in understanding static explanations of image classification models, improving acceptance and trust in XAI methods, and choosing the best AI models using only the explanations. We recruited a total of 60 participants and randomly divided them into three groups of equal size. One group interacted with our EMCEE model regarding static explanations, another group engaged with the baseline LLaVa-1.5 model, and the control group independently reviewed materials about the static explanations. Before and after the conversation or reading session, we measured their objective understanding and subjective perceptions of the provided static explanations. Based on the results, we estimated the effectiveness of the different conversational explanation systems in improving users' comprehension and usage of explanations.

Empirical results showed that our EMCEE outperforms the baseline LLaVa-1.5 model in both automatic and human evaluations by a large margin. In the automatic evaluation, EMCEE achieved relative improvements of 81.6% in BLEU and 80.5% in ROUGE compared to the baseline. While repeated training on self-generated data often leads to reduced diversity and quality [8], we showed that the proposed repetition penalty and hallucination detection can slow down the data degeneracy in training with synthetic data. In the human evaluation, participants who interacted with EMCEE reported a better understanding of static explanations, felt that the explanations enhanced their experience with AI models, were more inclined to use explanations in the future, trusted the explanations more, and demonstrated that they could collaborate better with AI systems using the explanations.

To further investigate how training on self-generated synthetic data enhances user interactions with the conversational XAI system, we conducted a fine-grained analysis of model responses to different types of user questions. We manually classified the questions asked during the human evaluation into three categories: generic AI/XAI questions, questions related to the provided explanations, and extended questions. We sampled 10 questions from each category for both the baseline and EMCEE models. Three well-educated annotators rated the responses on factual correctness and understandability. Results showed that EMCEE consistently provides more accurate and truthful answers across all question types compared to the baseline. The improvement in factual correctness highlights the effectiveness of the hallucination detector in filtering out incorrect statements from the synthetic data and reducing model hallucinations. In terms of understandability, EMCEE outperforms the baseline, particularly for questions related to the provided explanations. This suggests that training on synthetic conversations helps EMCEE better grasp the conversational context of explanations, leading to more understandable responses for users.

Our contributions can be summarized as follows.

- To the best of our knowledge, we propose the first conversational explanation system that can answer free-form follow-up questions after providing static explanations to users.
- We introduce a novel method to train conversational explanation systems on self-generated synthetic data. To enhance data quality, we propose a repetition penalty to boost data diversity and a hallucination detector to reduce erroneous information in synthetic data.
- We validate the effectiveness of our conversation explanation system, EMCEE, through both automatic and human evaluation ($N = 60$). Results show that EMCEE significantly outperforms baseline models in helping non-AI experts understand and utilize AI explanations.
- We analyze model responses to user questions and demonstrate that training on self-generated synthetic data improves the model's ability to generate more truthful and understandable responses, leading to enhanced user interactions with the system.

2 Related Work

2.1 Static XAI

Explainable Artificial Intelligence (XAI) refers to techniques that explain the learning process or the predictions of AI [90]. Most existing techniques are static XAI, which provides a one-time explanation with no capability for further user interaction. These techniques can be broadly divided into two categories: self-explanatory

models and post-hoc methods. Self-explanatory models are inherently transparent, offering clarity in their decision-making processes [37, 47, 72, 88, 91]. The majority of recent XAI methods are post-hoc XAI methods, applied to already developed models that lack inherent transparency [1, 7, 10, 70, 76]. There are two main groups of methods in post-hoc XAI, i.e., feature attribution methods and example-based methods.

Feature Attribution. Feature attribution methods explain model predictions by investigating the importance of input features to final predictions [1, 15]. There are two main types of feature attribution methods, gradient-based methods [12, 46, 50, 59, 76, 81, 83, 87, 87] and surrogate methods [3, 33, 35, 57, 70, 78]. Gradient-based methods employ gradients to evaluate the contribution of a model input on the model output. Surrogate methods leverage a simple and inherently interpretable model, such as a linear model, to locally approximate the behavior of the complex neural network.

Example-based Methods. Example-based methods explain AI predictions by identifying a selection of data instances [1, 15, 66]. These instances may be training data points with the most influence on the parameters of a prediction model [10, 27], counterfactual examples that alter predictions with minimal changes to inputs [64, 71, 85, 89, 95, 96], or prototypes that contain semantically similar parts to input instances [13, 38, 41].

In this work, we focus primarily on feature attribution methods, as they directly highlight the importance of input features, making the decision-making process of models more intuitive for laypeople [42]. Specifically, we select Grad-CAM, Integrated Gradients, and SHAP from gradient-based methods, as well as LIME from surrogate methods, to evaluate the effectiveness of different conversational evaluation systems.

2.2 Conversational XAI

Human-Computer Interaction (HCI) researchers have recently proposed that XAI methods should involve conversation, aligning with the natural way humans explain to each other. Specifically, Lombrozo [58] argues that explanations emerge from a conversational interaction between an explainer and an explainee. Similarly, Miller [63] emphasizes that explanations should include an interactive communication process, where the explainer provides the necessary information for the explainee to understand the causes of an event through dialogue. Building on this perspective of human explanations, recent works have introduced the concept of "explainability as dialogue," aiming to make explanations more accessible to a wide range of non-expert users [23, 48, 51].

Despite much exploration of the role of conversation in explainability, the practical development of conversational XAI is still in its early stages, with limited methods available so far. Shen et al. [77] applied conversational explanations to scientific writing tasks, finding improvements in productivity and sentence quality. Likewise, Slack et al. [82] designed dialogue systems that help users better understand machine learning models in tasks like diabetes prediction, rearrest prediction, and loan default prediction. However, these systems rely on template-generated conversations and can only handle a limited set of predefined queries. Our work represents the first system capable of delivering free-form explanatory conversations about static explanations.

2.3 Training with Synthetic Data

The exceptional performance of Large Language Models (LLMs) and Vision Language Models (VLMs) in generating human-like text has encouraged researchers to explore their use as training data generators [25, 28, 61, 62, 93, 94]. For example, SuperGen [61] uses LLMs conditioned on label-descriptive prompts to generate training data for text classification tasks. FewGen [62] finetunes an LLM on few-shot samples and uses it to generate synthetic data for seven classification tasks in the GLUE benchmark. While LLMs and VLMs have shown promise in generating human-like texts, they still face the challenge of producing noisy and low-quality synthetic data. This may lead to decreased performance or perpetuated biases in the model trained on the data [22, 39, 44, 49, 73, 97].

To mitigate the detrimental effects of noisy and low-quality synthetic data from LLMs and VLMs, several methods have been proposed [25, 28, 62, 94]. For example, ProGen [94] adjusts the weight of generated data points with regard to its influence on the validation loss, using influence function [45]. However, these strategies have primarily focused on generating data for classification tasks and on training small-scale task-specific models. Techniques such as applying the influence function to weigh data points are effective for smaller models. They present challenges and require a special design when adapted to LLMs [26].

In our work, we apply data generation to conversational explanations and utilize generated data to train the original VLM. We improved the quality of the generated data and significantly slowed down model degeneracy after multiple generation-training iterations (see §5.1.3).

3 Methodology

The overall workflow of EMCEE is illustrated as Figure 1 and outlined in Algorithm 1. Starting from a pretrained VLM V_1, we generated a set of synthetic conversations D_1, while using the repetition penalty to encourage data diversity. Each conversation contains multiple turns, denoted as $\langle(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots\rangle$, where the human turn is $\boldsymbol{x}_i$ and the machine response is $\boldsymbol{y}_i$. Then, we applied a hallucination detector f_h, which filters out hallucinated conversation turns. That is, if we detect hallucination from the machine response (*i.e.,* $f_h(\boldsymbol{y}_i) = 1$), $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is removed from the conversation. This process yields cleaned data D_1^{clean}. Afterwards, we finetuned the VLM on D_1^{clean}, leading to the next VLM V_2, from which we start another round of generation-filtering-finetuning. This process is repeated multiple times, without reusing any synthetic data from previous rounds.

We designed a prompt that is used across all stages, *i.e.,* data generation, model fine-tuning, and model inference. The prompt includes an instruction, background information about the AI model and XAI method, and several demonstration conversations. The instruction specifies the purpose of the conversation, which is to enhance user comprehension of static explanations. The background information includes details about the prediction task, the machine learning model, the XAI technique, and an example explanation. Details of the prompts are in Appendix A.

The number of demonstration conversations utilized varies in different stages. During data generation and model finetuning, we randomly chose zero or one demonstration and kept it consistent

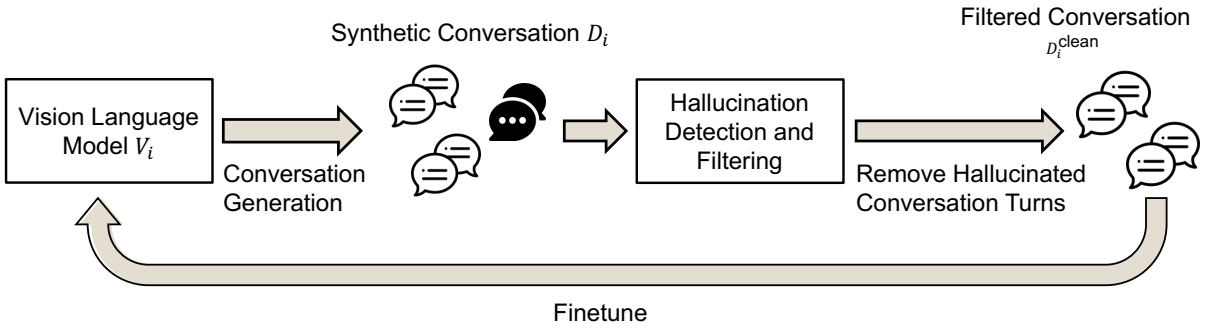


Figure 1: The Overall Workflow of EMCEE. V_i denotes the VLM and D_i denotes the synthetic conversation data in the i-th iteration. Starting from a pretrained VLM V_1, we first generate diverse synthetic conversations D_1 with the repetition penalty. Next, we use a hallucination detector to clean synthetic data, producing cleaned data D_1^{clean}. We then finetune the VLM on D_1^{clean}, which creates V_2, and this process repeats.

Algorithm 1 EMCEE

Input: a pretrained VLM V_1; a hallucination detector f_h, $f_h(\boldsymbol{y}) = 1$ if $\boldsymbol{y}$ is deemed hallucination; number of conversations to generate per round N; maximum number of rounds R.
Output: a finetuned model V_R

1: **for** r **in** $1...R$ **do**
2: $\quad\mathcal{D}_r \leftarrow$ generate N conversations from V_r;
3: $\quad D_i^{\text{clean}} \leftarrow \{(\boldsymbol{x}, \boldsymbol{y}) \in D_r \mid f_h(\boldsymbol{y}) \neq 1\}$;
4: $\quad V_{r+1} \leftarrow$ finetune V_r on D_i^{clean};
5: **end for**

for each mini-batch. During model inference and evaluation, the number of demonstrations ranged between zero and three.

3.1 Repetition Penalty

The repetition penalty encourages the VLM to generate more diverse conversations by discounting the logits of tokens seen in previous conversation turns. Specifically, given the logits z_i for each token i in the vocabulary, the probability p_i of predicting token i is computed as,

$$p_i = \frac{\exp(z_i/(T + \theta \cdot \mathbb{1}(i \in G)))}{\sum_j \exp(z_j/(T + \theta \cdot \mathbb{1}(j \in G)))}, \tag{1}$$

where T is the temperature. θ is the ratio of the repetition penalty. G is the set of words existing in generated conversations in the current round, and $\mathbb{1}$ is an indicator function. When the token i exists in G, $\mathbb{1}(i \in G)$ is 1, otherwise, $\mathbb{1}(i \in G)$ is 0.

3.2 Hallucination Detection and Filtering

VLMs often generate convincing but factually incorrect statements, especially when answering questions that require reasoning and logical deduction [6, 14, 39, 49, 99]. Conversational explanations are mainly about explaining the causal relationship between static explanations and AI predictions, which involves significant reasoning. Therefore, hallucination is a major concern in this use case.

To reduce hallucination, we integrated a hallucination detector into the training process, which identifies and removes hallucinated conversation turns. To train the detector, we constructed a dataset comprising 1,500 sentences about machine learning and XAI methods. The dataset is balanced, containing 750 factually correct sentences and 750 factually incorrect ones. It includes 500 sentences on general machine learning knowledge, sourced from students studying machine learning. Among 500 sentences, 250 sentences were picked from the class notes by students. After that, the students were asked to create 250 incorrect sentences based on the correct ones. The remaining 1,000 sentences are about XAI knowledge; we used GPT-4-turbo-2024-04-09 to generate 500 factually correct sentences about XAI and subsequently alter them to be incorrect. All generated sentences have been rigorously validated by XAI experts. To prevent data duplication, we manually removed all duplicated sentences. Example sentences from this dataset are displayed in Table 1. We used 80% of the dataset for training the detector, with the remaining 20% reserved for validation and testing.

3.3 Implementation

We used LLaVa-1.5 [55, 56] as our base vision language model. LLaVa-1.5 is an end-to-end trained large multimodal model that combines a vision encoder and an LLM for general-purpose visual and language understanding. We chose LLaVa-1.5 for its high performance in answering scientific questions and proficiency in visual chat scenarios [55, 56].

For the data generation process, the number of generated conversations N at each round is set to 2000, with 500 conversations for each static explanation method. The number of training iterations of the generation network is empirically tuned on the validation set and set to five. The temperature is set to 1.2 to encourage diverse generations while maintaining coherence. The repetition penalty ratio is set to 1.1. For finetuning LLaVa-1.5, we used LoRA [32] to only finetune the language model while keeping the vision encoder and projector frozen. The rank of the LoRA parameter is set to 128, the batch size is 32, and the learning rate is 2×10^{-4} with cosine annealing. In each generation-filtering-finetuning round, we finetuned the LLaVa-1.5 for 3 epochs. Finally, for the hallucination detector, we trained a Bert-base model [18] using the SGD optimizer with a learning rate of 0.01, batch size of 16, and weight decay for 100 epochs. The hallucination detector achieved an accuracy of 79.5% on the held-out test set.

Sentence	Label
When the amount of data stays the same, the more parameters, the more difficult to estimate the parameters accurately.	0
When the amount of data stays the same, increasing the number of parameters can improve the accuracy of their estimates.	1
XAI is less important in systems where decisions are not critical.	0
XAI is only relevant in non-critical systems.	1
Grad-CAM can be applied to any convolutional layer of a network, not just the final layer.	0
Grad-CAM is restricted to analyzing the input and output layers of a network.	1
LIME can explain any machine learning model as long as it can probe the model with perturbed inputs.	0
LIME can only explain models that are specifically designed to work with its framework.	1
The path taken from baseline to input in Integrated Gradients is typically linear.	0
The path taken is randomly generated in each run of Integrated Gradients.	1
SHAP values can be computed for any data point in the dataset, providing versatile insights.	0
SHAP values can only be computed for a limited set of predefined data points.	1

Table 1: Examples of sentences with labels in our hallucination dataset. Label 0 means the sentence is factually correct; label 1 means the sentence is factually incorrect.

4 Evaluation Methodology

In this section, we present the evaluation methodology used to assess the performance of our proposed EMCEE model. We employed two evaluation methods: automatic and human evaluations. The automatic evaluation is crucial for objectively measuring the model's ability to generate responses that align with ground truth explanations, using established metrics. Since our conversational XAI system is designed to help users better understand and utilize static explanations, human evaluation is necessary to assess its real-world impact. We examined the system's effectiveness by observing participants' comprehension, acceptance, trust in the static explanations, and their ability to collaborate with these explanations, both before and after interacting with the system.

4.1 Automatic Evaluation Metrics and Dataset

For automatic evaluations, we conducted few-shot evaluations with zero to three demonstrations. We leverage BLEU [68] and ROUGE [54] scores to measure word overlaps between generated response text and ground truth text. Higher BLEU and ROUGE scores indicate better alignment between the generated and human-written texts, reflecting the model's ability to produce more accurate and contextually appropriate outputs. These two metrics are commonly used in natural language processing (NLP) evaluation, as they are easy to compute and comparable across different papers.

We conducted our automatic evaluation using the only existing dataset of human-human conversational XAI interactions, which was collected in previous work by Zhang et al. [98]. This dataset was gathered using a Wizard-of-Oz (WoZ) setting [40]. Participants interacted with what they believed was an autonomous dialogue system, which was actually operated by a human expert in machine learning and XAI. The dataset includes 30 conversations on the LIME method and another 30 on the Grad-CAM method. On average, each conversation contains 27.4 utterances, with each utterance averaging 14.4 words. Due to its small size, we did not use this dataset for training. We employed one conversation per static explanation method (LIME and Grad-CAM) as a demonstration in the data generation prompt and six conversations for demonstrations in the few-shot evaluation. In the remaining 52 conversations, 10 conversations were used for validation and 42 were used for testing.

Table 2: Academic disciplines of our participants and the number of participants in each group. There are 60 participants from 4 different discipline groups.

Academic Discipline	Number of Participants
Business	14
Engineering	10
Humanities	18
Science	18

Although BLEU and ROUGE are useful and widely used, they have limitations. High n-gram overlap with human-written references does not necessarily guarantee that users can understand the generated responses or that the responses are coherent within the conversation. Therefore, we also conducted human evaluations to assess the effectiveness of different conversational models in helping users understand, accept, and trust static explanations.

4.2 Human Evaluation Protocol

For the human evaluation, we evaluated the effect of different conversational XAI methods by observing participants' objective understanding and subjective perception of static explanations, before and after interacting with different conversational XAI methods. Our study has received approval from our Institutional Review Board (#IRB-2023-254).

4.2.1 Participants. We recruited 60 participants for our study. All were 21 years old or older, fluent in English, and had not been involved in research about XAI previously. We recruited our participants in two ways: by posting advertisements on an online forum

and by emailing students and staff across various departments and schools. To ensure diversity, participants came from a broad range of disciplines. For ease of reporting, we categorize their disciplines into four groups:

- Business, including Business and Accountancy.
- Engineering, including Civil and Environmental Engineering, Electrical and Electronics Engineering, Chemical Engineering and Biotechnology
- Humanities, including Psychology, Economics, Communication Studies, Linguistics and Multilingual Studies, and Sociology.
- Science, including Biology, Chemistry, Sport Science & Management, and Physics.

Table 2 shows statistics of the academic disciplines that the participants enrolled in.

4.2.2 Experimental Task. We focused on the image classification task on the ImageNet dataset and trained three classification models with different top-1 classification accuracies: Swin Transformer (84.1%), VGG-16 (71.6%), and AlexNet (56.5%). We chose image classification because it requires minimal domain-specific expertise, making it well-suited for crowdsourcing among participants from diverse domains. To generate explanations for model predictions, we adopted four feature attribution explanation methods: LIME [70], Grad-CAM [76], Integrated Gradients [83], and SHAP [59]. For a more comprehensive evaluation, we extended the two XAI methods used in automatic evaluation to these four attribution explanation methods. The focus is on feature attribution as we believe the relationship between input features and model predictions is more intuitive to understand for laypeople than, for example, data attribution [42].

4.2.3 Experimental Interface. Our study was conducted on a web-based platform allowing participants to complete the entire procedure remotely. This platform ensures that all communication between users and conversational agents is text-based and recorded. Figure 2 displays an example screenshot of the interface. There are two sections on the page. On the left (Figure 2 Part A), participants see a task description, a description of the prediction model, a model input, a model output, an explanation generated by the explanation model, and a description of the explanation. On the right within the chatbox (Figure 2 Part B), participants engage in a text-based conversation with the agent to clarify the provided explanation. Participants can ask any questions or provide comments related to the explanation on the left. In the control group, we replaced the chatbox with a 15-minute timer. Once the timer reached zero, participants were allowed to proceed to the post-measurements.

4.2.4 Experimental Design. There are two independent variables and two categories of dependent variables. The first independent variable is the explanation method: LIME, Grad-CAM, Integrated Gradients, or SHAP. The second independent variable is the participant's group: participants either have a conversation with our EMCEE, a conversation with the baseline LLaVA-1.5, or read the static explanations. We measure participants' objective understanding and subjective perceptions of explanations before and after conversations or readings. Two sets of dependent variables are collected in the experiment: the model selection accuracy and the self-reported perception scores.

Previous work [21] indicates that participants' prior knowledge of AI may influence their perceptions of explanations, potentially introducing a confounding factor. To mitigate any potential confounding variable, participant assignments to the three groups are completely random. Additionally, the results in Section 5.2 showed no significant difference ($p = 0.44$) amongst self-reported pre-conversation understanding of the three groups.

4.2.5 Measurement of Users' Objective Understanding – Selection of Classification Models. We assessed users' understanding of static explanations by measuring their performance in a model selection task. Model selection is a fundamental task for machine learning practitioners [4]. Specifically, participants were presented with 5 input images, on which the three classification models make identical decisions. The only differences between these models are their explanations. Participants must choose the model that they believe will perform most accurately on unobserved test data. Hence, to make the correct selection, the participants must understand the explanations. We measured participants' objective understanding of static explanations by their accuracy in selecting the correct model. The complete set of images used for LIME, Grad-CAM, Integrated Gradients, and SHAP is detailed in Appendix B.

We observe that static explanations do not always faithfully reflect the actual workings of classification models [2, 36, 43] and do not always contain actionable information for model selection. In our study, model selection is used to determine whether users can comprehend static explanations *when* the explanations do have actionable information for selection, rather than assessing the explanations themselves. For this, we chose images that models with high accuracy can provide more reasonable explanations. An explanation is deemed more reasonable when it highlights features that are unique to the predicted class and avoids irrelevant features. A good model should have explanations that rely on multiple types of discriminative features, making the model more robust. Consequently, the model makes the correct decision even if some discriminative features are absent or occluded. Hence, this approach allows users to easily pick the best classification models if they understand the static explanations well.

Other objective measurements to assess user understanding [11, 86] are not suitable for our study. These methods measure users' accuracy on the same prediction tasks as the classifier, such as student admission [11] or recidivism prediction [86]. Their main idea is to compare users' predictive accuracy with and without explanations. However, the current work focuses on the task of image classification, a task users can easily perform without any explanation. Even when the user cannot make correct predictions, explanations highlighting relevant portions of the image is unlikely to provide any assistance. Instead, we used model selection as the objective measurement. To ensure that the correct model selection reflects users' understanding of XAI, we carefully curated a set of images so that users could only make the correct choice if they comprehended the explanations provided.

4.2.6 Measurements of Users' Subjective Perception. We also measured participants' self-reported perception of the static explanations, including their comprehension, acceptance, and trust. There are a total of 13 questions. All questions utilize a 7-point Likert scale for responses. The full list of the questions is in Appendix C.

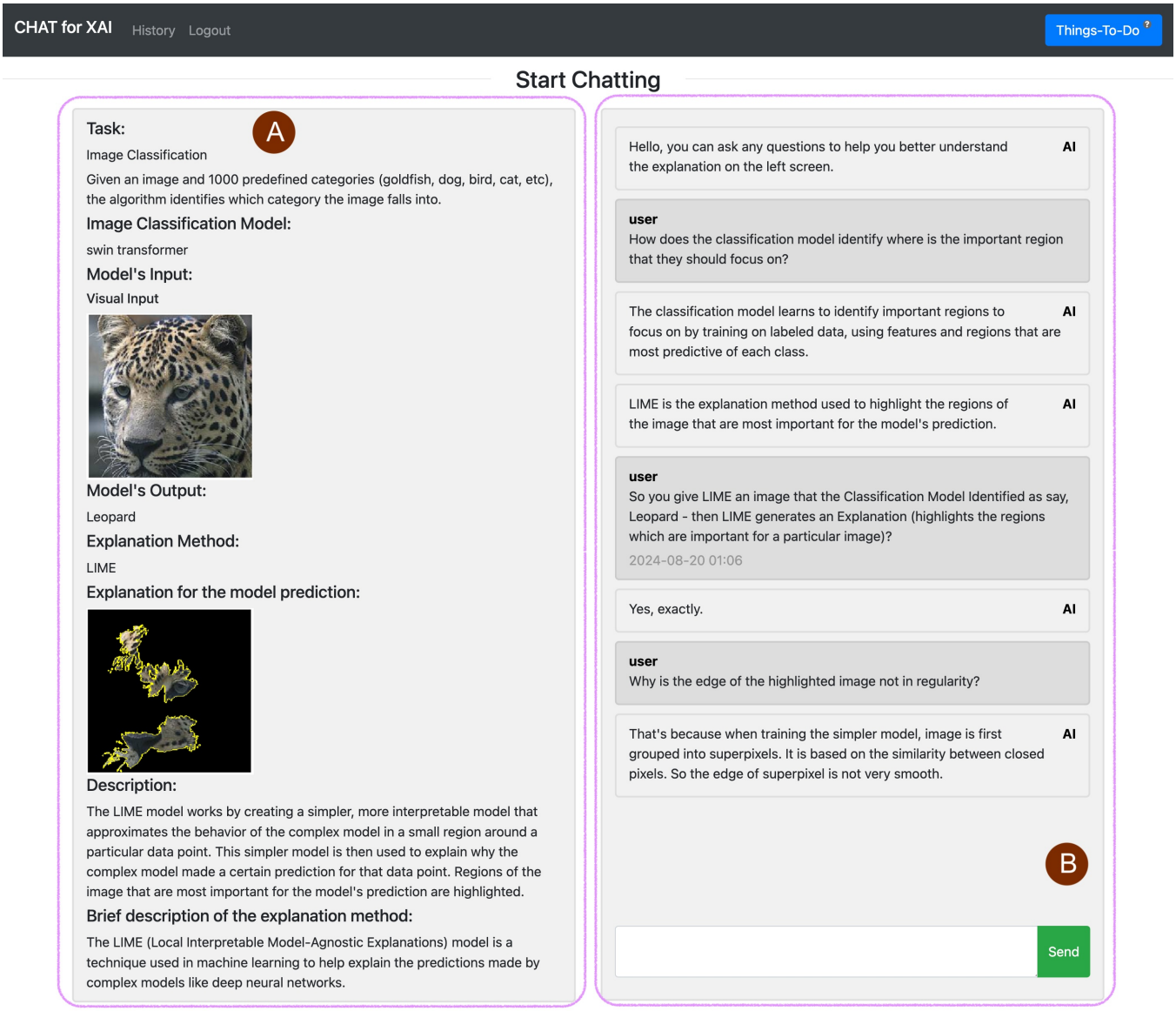


Figure 2: The interface where users discuss static explanations with a conversational agent. *Part A*: Information about static explanations, including a task description, a description of the prediction model, a model input, a model output, an explanation generated by the explanation model, and a description of the explanation. *Part B*: A chatbox where users converse with a conversational agent to clarify the explanation.

- Comprehension [11, 31]: Participants' subjective perceptions of their understanding of explanations. It serves as a supplement to objective assessments, offering a more comprehensive view of how well participants understand static explanations.
- Perceived Usefulness [16, 17, 19]: The degree to which participants feel that the explanations enhance their experience with deep learning models. Together with *perceived ease of use* and *behavioral intention*, these three factors measure participants' acceptance of static explanations. They are derived from the Technology Acceptance Model (TAM) [16, 17, 19], a widely applied theory for understanding individual acceptance and usage of information systems. Investigating users' acceptance of the explanations is very important, as the explanations are intended for end-users.
- Perceived Ease of Use [16, 17, 19]: Participants' judgment of the simplicity and clarity of the explanations.
- Behavioral Intention [16, 17, 19]: The tendency of participants to utilize the explanation information in the future.

- Trust [5, 65]: Participants' confidence in the reliability of the explanation methods to perform as intended. Trust has been recognized as a key factor in human-AI collaboration, as it influences how much humans rely on AI models, thus directly affecting the effectiveness of the human-AI team [20, 29, 74, 75, 80, 84, 92].

4.2.7 Experimental Procedure. Before participating in the study, participants signed an informed consent form that details the study's objectives and procedures. The form also explained the compensation and ensured both anonymity and confidentiality of the collected data. After signing, participants received an email with instructions to access the study platform. Once logged in, a pop-up window provided a brief overview of the tasks. Participants then began with pre-experiment measurements for their objective understanding and subjective perceptions of static explanations. Objective understanding was assessed by letting participants choose the most accurate of three classification models on unseen test data, using 5 explanation examples. The subjective perception was measured through 13 self-reporting questions. These questions probed participants' perceived comprehension, acceptance, and trust in the explanations.

After these initial measurements, we randomly assigned participants into three groups of equal size. One-third engaged in an online text-based conversation with our EMCEE model to ask questions and clarify doubts. Another third conversed with the baseline LLaVA-1.5 model. The remaining third, serving as the control group, spent 15 minutes reviewing the static explanations independently. The duration matched the average time spent in conversations by the other two groups. The information provided to participants at this stage is displayed in Figure 2. Since this phase focuses on explaining XAI methods to users rather than testing their understanding, explanations for just one classifier were sufficient. We used the Swin Transformer as the classifier due to its higher classification accuracy.

After the conversation or reading session, participants repeated the same measurements of objective understanding and subjective perceptions as in the pre-session phase. All results and conversation records are documented. Upon completing the study, participants received a $10 reward.

5 Results & Discussion

This section presents the experimental results from both automatic and human evaluations of the baseline and our EMCEE model. For the automatic evaluation, we report results on an existing dataset of human-human conversational XAI interactions and include an ablation study to show the effectiveness of different components in our method. For the human evaluation, we present results on participants' objective understanding and subjective perceptions of static explanations, measured before and after different conditions. We also analyze the collected conversations and provide insights into why our system can improve users' understanding, acceptance, and trust in static explanations.

5.1 Results of Automatic Evaluation

5.1.1 Comparison of Baseline and Our Method. Table 3 presents the automatic evaluation results of both the baseline LLaVa-1.5 model and our EMCEE model when we prompt them with 0 to 3 example

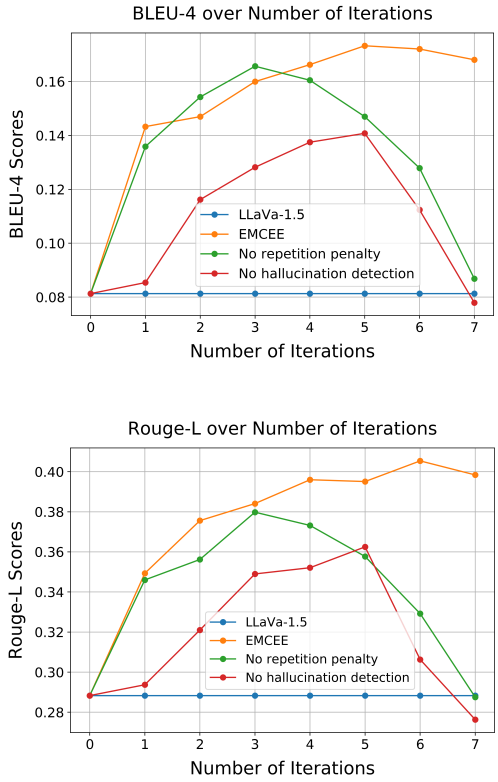


Figure 3: BLEU-4 and Rouge-L scores over the number of training iterations for LLaVa-1.5, EMCEE and different ablated version of EMCEE.

conversations. Our model exhibits substantial improvements over LLaVa-1.5 in terms of both BLEU and ROUGE scores. Specifically, EMCEE shows an increase of 81.6% in BLEU scores and 80.5% in ROUGE scores compared to LLaVa-1.5. These results suggest that our model, trained on self-generated synthetic conversations in a multi-round setting, can better explain static XAI and produce responses more aligned with human answers to users' inquiries.

5.1.2 Ablation Study. We created the following ablated versions of EMCEE: (1) No multi-round training, which performs one round of synthetic generation, filtering, and model finetuning. (2) No repetition penalty, which removes the repetition penalty. (3) No hallucination detection, which does not detect and remove hallucinated conversation turns.

Table 4 summarizes the results of different ablated versions of EMCEE. We make the following observations. First, the absence of multi-round training significantly reduces the performance across all BLEU and ROUGE metrics. This demonstrates that generating synthetic conversations and filtering out hallucination conversations in an iterative way can gradually improve the quality of generated conversations and thus improve the performance of our model. Second, the model's performance decreases when the repetition

Table 3: Automatic Evaluation of pretrained LLaVa-1.5 and our model. We prompt models with 0 to 3 example conversations.

Methods	Shot Num	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L
LLaVa-1.5	0	0.1328	0.0534	0.0235	0.0103	0.3150	0.0595	0.0179	0.2507
	1	0.1447	0.0680	0.0361	0.0196	0.2823	0.0823	0.0374	0.2324
	2	0.2160	0.1329	0.0985	0.0813	0.3365	0.1469	0.1014	0.2883
	3	0.1979	0.1265	0.0854	0.0687	0.3153	0.1339	0.0839	0.2709
EMCEE (Ours)	0	0.2394	0.1659	0.1270	0.1055	0.3918	0.2295	0.1794	0.3418
	1	0.2895	0.2186	0.1826	0.1618	0.4513	0.2854	0.2391	0.4006
	2	**0.3056**	**0.2336**	**0.1945**	**0.1721**	**0.4629**	**0.2964**	**0.2454**	**0.4054**
	3	0.2786	0.2100	0.1769	0.1571	0.4380	0.2798	0.2339	0.3881

Table 4: An ablation study of the proposed EMCEE on the conversational explanation dataset

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L
EMCEE	**0.3056**	**0.2336**	**0.1945**	**0.1721**	**0.4629**	**0.2964**	**0.2454**	**0.4054**
No Multi-round Training	0.2808	0.2079	0.1685	0.1465	0.4198	0.2608	0.2162	0.3756
No Repetition Penalty	0.2824	0.2214	0.1854	0.1657	0.4219	0.2778	0.2329	0.3798
No Hallucination Detection	0.2730	0.1977	0.1631	0.1408	0.4161	0.2375	0.1950	0.3625

penalty is removed. This result indicates that the diversity of synthetic conversations plays a crucial role in our model. Third, the most substantial performance drop occurs when the hallucination detector is removed, with a 10.7% decrease in BLEU scores and a 15.3% decrease in ROUGE scores. This result highlights the importance and necessity of filtering hallucinated synthetic data after generation.

5.1.3 Effects of Multiple Generation-Training Iterations. In the training of EMCEE, we repeated the generation-training process multiple times. We investigated how iterations affect the performance of EMCEE and ablated versions of EMCEE in BLEU-4 and ROUGE-L scores, as shown in Figure 3.

We observed that the ablated versions of EMCEE improved in the first few iterations and decreased afterward. This is similar to the findings of Briesch et al. [8], who showed that repeatedly training models with self-generated data initially caused performance gains but, after a few iterations, resulted in degenerate synthetic data with low diversity and eventual performance drop. This is especially apparent when we removed the repetition penalty or the hallucination filter, as both BLEU-4 and ROUGE-L decreased drastically after the third and fifth iterations, respectively. However, with both the repetition penalty and the hallucination filter of EMCEE, the performance drops became substantially milder. For BLEU-4, a small drop was observed after the fifth iteration. For Rouge-L, the performance effectively plateaued around the sixth and seventh iteration. We conclude that the proposed techniques, including the repetition penalty and the hallucination filter successfully slow down degeneracy in training with synthetic data.

5.2 Results of Human Evaluation

Table 5 presents the human evaluation results, comparing LLaVa-1.5, EMCEE, and the no-conversation group across four explanation methods: LIME, Grad-CAM, Integrated Gradients, and SHAP. The explanation methods (LIME, Grad-CAM, Integrated Gradients, SHAP) and participant groups (control, LLaVa-1.5, EMCEE) are between-subject variables, while time (before vs. after) is a within-subject variable. We conducted a three-way Analysis of Variance (ANOVA) to analyze the results.

To ensure the validity of the ANOVA results, we verified its underlying assumptions: independence, normality, and equal variance. For independence, participants were randomly assigned to one of the three groups and each participated in the study only once. For normality, we performed Shapiro-Wilk tests on participants' subjective understanding of different explanation methods. The resulting p-values were 0.155, 0.171, 0.062, and 0.084, indicating that the normality assumption was met. For equal variance, we conducted Bartlett's test, which yielded a p-value of 0.376, confirming that this assumption was also satisfied.

5.2.1 Effects of Different Conversational XAI Systems on Users' Objective Understanding and Subjective Perception of Static Explanations. Results showed significant main effects for group ($F(2, 48) = 3.79, p = .04$), method ($F(3, 48) = 43.25, p < .001$), and time ($F(1, 48) = 18.48, p < .001$). The EMCEE group, the Grad-CAM method, and the after-conversation condition displayed the highest objective decision accuracy. We also found a significant interaction effect between group and time ($F(2, 48) = 5.44, p = .007$), as displayed in the Figure 4. In participants' initial decisions, no significant differences were observed between the EMCEE, LLaVA-1.5, and control groups. During the final decision, participants interacting with EMCEE or LLaVa-1.5 both showed improved decision

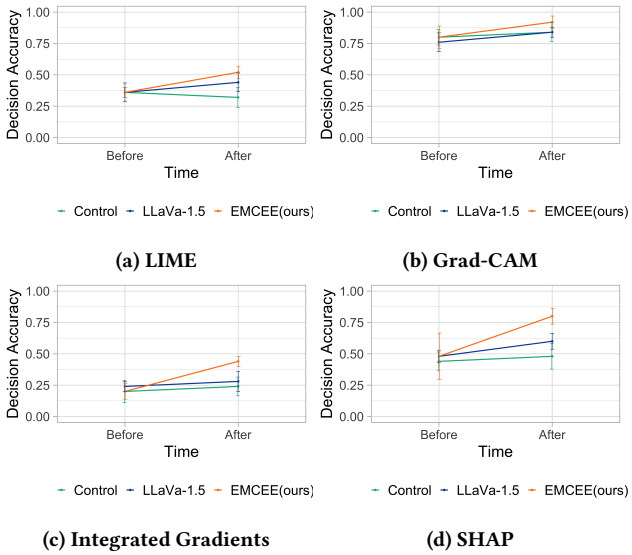
Table 5: Results of human evaluations before and after conversations. Each score is presented as mean ± standard deviation and the change δ = after − before.

Explanation Methods	Conversational Explanation method	Evaluation Timing	Objective Understanding (Model Selection Accuracy)	Subjective Understanding	Acceptance			Trust
					Perceived Usefulness	Perceived Ease of Use	Behavioral Intention	
LIME	Control	before	0.36 ± 0.09	4.60 ± 1.67	5.40 ± 1.19	4.73 ± 1.04	4.90 ± 0.55	4.05 ± 0.97
		after	0.32 ± 0.18	4.40 ± 1.52	5.13 ± 1.26	4.27 ± 1.30	4.60 ± 0.42	3.95 ± 1.71
		δ	-0.04	-0.20	-0.27	-0.46	-0.30	-0.10
	LLaVa-1.5	before	0.36 ± 0.17	4.00 ± 1.58	5.20 ± 1.02	4.40 ± 1.62	4.90 ± 1.02	4.10 ± 0.22
		after	0.44 ± 0.17	4.80 ± 1.48	5.60 ± 0.60	5.20 ± 0.60	5.20 ± 0.76	4.30 ± 0.54
		δ	0.08	0.80	0.40	0.80	0.30	0.20
	EMCEE (Ours)	before	0.36 ± 0.09	4.20 ± 0.84	5.27 ± 0.64	4.53 ± 0.60	5.00 ± 0.35	4.20 ± 0.37
		after	0.52 ± 0.11	5.20 ± 0.45	5.93 ± 0.64	5.60 ± 0.68	5.70 ± 0.45	4.85 ± 0.34
		δ	0.16	1.00	0.66	1.07	0.70	0.65
Grad-CAM	Control	before	0.80 ± 0.14	4.00 ± 1.00	5.27 ± 0.36	4.67 ± 0.67	5.00 ± 0.61	4.30 ± 0.37
		after	0.84 ± 0.17	4.00 ± 1.22	5.20 ± 0.84	4.27 ± 0.92	4.90 ± 0.82	4.40 ± 0.80
		δ	0.04	0.00	-0.07	-0.40	-0.10	0.10
	LLaVa-1.5	before	0.76 ± 0.17	4.00 ± 1.41	5.33 ± 0.41	4.87 ± 0.38	5.20 ± 0.57	4.40 ± 0.29
		after	0.84 ± 0.09	4.80 ± 0.45	5.60 ± 0.44	5.13 ± 0.51	5.50 ± 0.50	5.00 ± 0.47
		δ	0.08	0.80	0.27	0.26	0.30	0.60
	EMCEE (Ours)	before	0.80 ± 0.20	4.00 ± 1.22	5.13 ± 1.07	4.80 ± 0.77	5.20 ± 0.27	4.15 ± 0.72
		after	0.92 ± 0.11	5.40 ± 0.89	6.13 ± 0.61	5.40 ± 0.93	6.10 ± 0.42	5.25 ± 0.90
		δ	0.12	1.40	1.00	0.60	0.90	1.10
Integrated Gradients	Control	before	0.20 ± 0.20	3.80 ± 0.45	4.80 ± 0.50	3.87 ± 0.90	4.20 ± 0.97	3.65 ± 0.45
		after	0.24 ± 0.17	4.00 ± 0.71	4.73 ± 0.76	3.80 ± 0.77	4.00 ± 1.17	3.65 ± 0.72
		δ	0.04	0.20	-0.07	-0.07	-0.20	0.00
	LLaVa-1.5	before	0.24 ± 0.09	3.80 ± 0.45	4.73 ± 0.49	3.87 ± 0.77	4.40 ± 1.08	3.85 ± 0.55
		after	0.28 ± 0.18	4.00 ± 1.00	5.00 ± 0.71	4.40 ± 1.60	4.70 ± 1.20	3.85 ± 0.22
		δ	0.04	0.20	0.27	0.53	0.30	0.00
	EMCEE (Ours)	before	0.20 ± 0.14	3.60 ± 1.14	4.67 ± 0.71	3.60 ± 0.44	4.60 ± 0.65	3.85 ± 0.22
		after	0.44 ± 0.09	4.60 ± 0.55	5.20 ± 0.61	4.73 ± 0.55	5.50 ± 0.71	4.50 ± 0.40
		δ	0.24	1.00	0.53	1.13	0.90	0.65
SHAP	Control	before	0.44 ± 0.17	4.20 ± 0.84	5.20 ± 0.38	4.47 ± 0.56	4.80 ± 0.27	4.20 ± 0.57
		after	0.48 ± 0.23	4.00 ± 1.00	5.07 ± 0.80	4.33 ± 0.85	4.70 ± 0.76	4.30 ± 0.54
		δ	0.04	-0.20	-0.13	-0.14	-0.10	0.10
	LLaVa-1.5	before	0.48 ± 0.11	3.80 ± 1.79	5.40 ± 0.49	4.87 ± 1.73	5.00 ± 1.06	4.20 ± 1.47
		after	0.60 ± 0.14	5.20 ± 1.64	5.60 ± 0.44	5.67 ± 0.78	5.20 ± 0.91	4.60 ± 1.14
		δ	0.12	1.40	0.20	0.80	0.20	0.40
	EMCEE (Ours)	before	0.48 ± 0.41	3.80 ± 1.30	5.40 ± 0.60	4.60 ± 0.92	5.00 ± 0.79	4.20 ± 0.91
		after	0.80 ± 0.14	5.60 ± 1.14	6.13 ± 0.69	6.00 ± 0.41	5.90 ± 0.89	5.30 ± 0.82
		δ	0.32	1.80	0.73	1.40	0.90	1.10

accuracy. However, participants using our EMCEE model consistently demonstrated a greater increase in model selection accuracy after the conversation. This phenomenon highlights EMCEE's effectiveness in helping participants collaborate with static explanations.

We observed varied objective performance across explanation methods ($F(3, 48) = 43.25, p < .001$). Participants achieved the highest accuracy in the model selection task with Grad-CAM and the lowest accuracy with Integrated Gradients. A potential reason might be the inherently intuitive nature of the explanations produced by Grad-CAM compared to others [98].

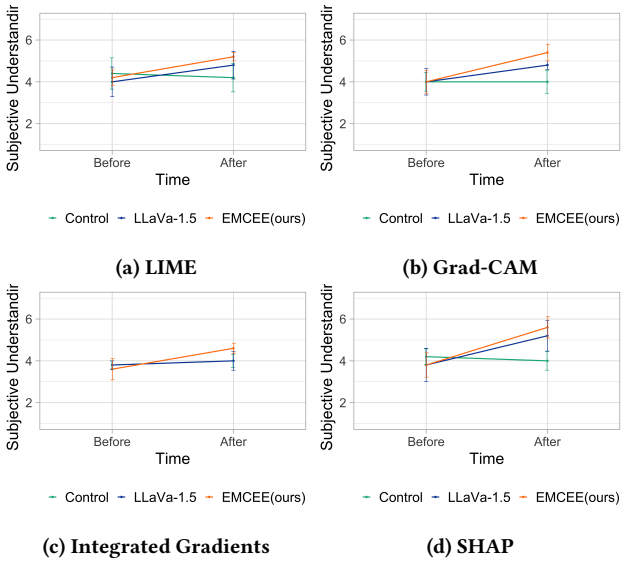
Regarding participants' subjective understanding, we found a significant main effect of evaluation timing ($F(1, 48) = 30.56, p < .001$) and a significant interaction between group and time ($F(1, 116) =$



(a) LIME (b) Grad-CAM

(c) Integrated Gradients (d) SHAP

Figure 4: Model selection accuracy for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.



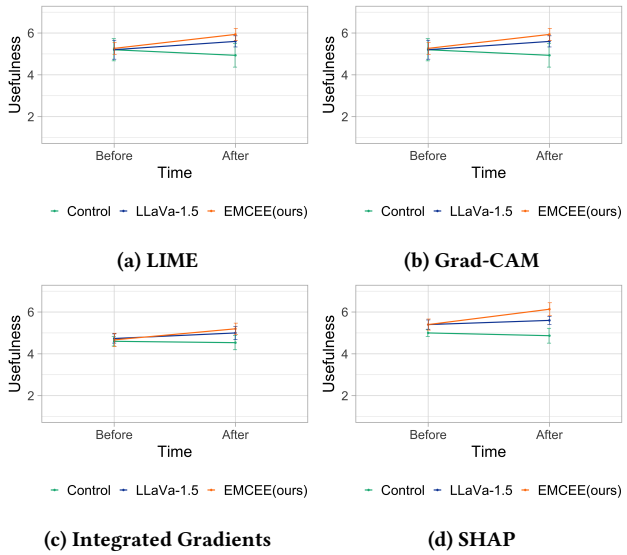
(a) LIME (b) Grad-CAM

(c) Integrated Gradients (d) SHAP

Figure 5: Subjective understanding score for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.

10.16, $p < .001$). Initially, there was no significant difference in participants' self-reported understanding of static explanations among different groups. After the conditions, participants who received conversational explanations from EMCEE reported significantly greater improvements than the other two groups across all four explanation methods.

For perceived usefulness, the results showed a significant main effect of method ($F(3, 48) = 2.86, p = .0046$) and time ($F(1, 48) =$



(a) LIME (b) Grad-CAM

(c) Integrated Gradients (d) SHAP

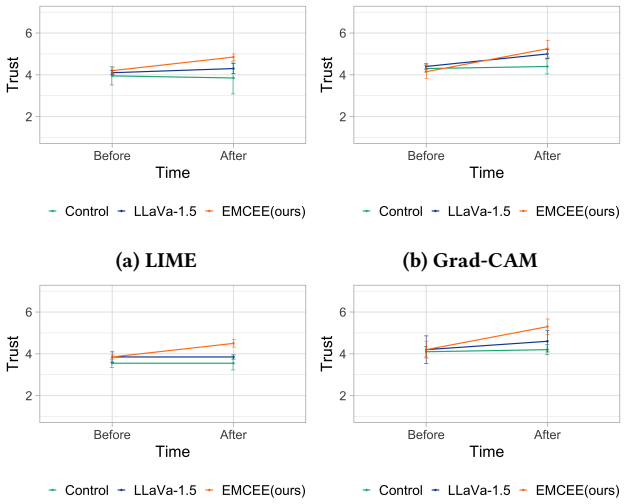
Figure 6: Participants' self-report usefulness score for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.

21.35, $p < .001$), as well as a significant interaction between group and time ($F(2, 48) = 15.37, p < .001$), as depicted in Figure 6. Participants' perceived usefulness increased after interacting with LLaVa-1.5 or EMCEE, though the improvement is much smaller with LLaVa-1.5. In contrast, for the control group, perceived usefulness dropped after more time to the static explanations was provided.

Similar results were observed for participants' perceived ease of use. There were significant main effects of method ($F(3, 48) = 3.83, p = .002$) and of time ($F(1, 48) = 22.14, p < .001$), as well as a significant interaction effect between group and time ($F(2, 48) = 15.5, p < .001$). The interaction effect is displayed in appendix Figure 17. The perceived ease of use increased after participants interacted with EMCEE or LLaVa-1.5. EMCEE produced a greater improvement than LLaVa-1.5. On the contrary, the control group's perceived ease of use decreased after spending more time with static explanations.

For the behavioral intention, results showed significant main effects of the group ($F(2, 48) = 5.14, p = .009$), method ($F(3, 48) = 2.84, p = .004$), and time ($F(1, 48) = 18.48, p < .001$). We also observed a significant interaction effect between group and time ($F(1, 116) = 20.94, p < .001$). The interaction figure is displayed in appendix Figure 18. Participants are more inclined to use explanations in future scenarios after receiving conversational explanations from EMCEE. On the other hand, the behavioral intention of the control group decreased for all explanation methods.

The boost in perceived usefulness, ease of use, and behavioral intention after interacting with EMCEE can be attributed to the increased understanding of static explanations. Prior to the interactions, participants might have had limited knowledge or even misconceptions about the explanation methods [98]. Experiment results showed that participants gained a clearer understanding of how the XAI methods function, after the participants' questions



(a) LIME (b) Grad-CAM

(c) Integrated Gradients (d) SHAP

Figure 7: Participants' trust for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.

were addressed in the conversations with EMCEE. Consequently, they reported perceiving the static explanations as more useful and easier to use, and were more inclined to use the static explanations in future tasks.

For the trust measurement, results showed a significant main effect of time ($F(1, 48) = 40.16, p < .001$) and a significant interaction effect between group and time ($F(1, 116) = 43.7, p < .001$), as shown in Figure 7. Initially, there were no significant differences in trust scores among the groups. However, by the end, participants who interacted with EMCEE reported the highest trust scores. According to theories of trust [31, 53, 60], the ability to build a mental model of AI systems is the key to user trust in AI. The improvements in trust may be a result of an improved understanding of static explanations, as indicated by earlier results.

From Table 5, one may notice that the improvement deltas are small relative to the standard deviation of the measurements. The relatively large standard deviation is due to inherent variations in individuals' subjective perceptions. These variations arise from differences in participants' backgrounds, experiences, and understanding of explanation methods. Despite this, the deltas capture the overall shift of the entire user group before and after the study, indicating the impact of different experimental conditions. Our deltas are consistently higher than those of the baselines across all explanation methods and measurements, demonstrating the effectiveness of the proposed EMCEE.

5.2.2 Analysis of Collected Conversations. We collected 40 conversations between participants from four different discipline groups and two conversational explanation systems. On average, each conversation has 22.8 turns. By conducting a basic content analysis of the users' questions, we divide them into three categories:

- Generic questions about machine learning and explainable AI concepts: Questions about fundamental terms and concepts in machine learning and explainable AI that lay people may not know. Examples include, "What is a deep learning model?", "What is accuracy?", or "What are explanation methods?"
- Questions related to the provided explanations: Questions about the specific explanations provided during the conversation, such as how the explanation is created and how the explanation methods function. Examples include, "How does Grad-CAM produce the heatmap?" or "What do different colors represent in SHAP?".
- Extended questions: Questions that arise after users understand the provided explanations, e.g., generating other explanations for the current prediction, explanations for different predictions, or comparisons between the provided explanation and other explanation methods.

Based on this categorization, we classified all questions in our collected conversations. In total, we identified 358 questions across the three categories. Table 6 provides examples and the number of questions in each category. As observed in Table 6, a large portion of the questions revolve around basic machine learning and explainable AI concepts. This trend might be attributed to the diverse backgrounds of the participants. It suggests that many participants may not be familiar with machine learning models and explanation methods. This is consistent with the real application of explanation methods, where non-expert users often need clarification on fundamental concepts.

We also observed a significant interest of participants in questions related to the provided explanations. This suggests that explanations generated by Grad-CAM, LIME, and Integrated Gradients are not always easily understood by users. This highlights the importance of tailoring responses to users' specific questions to enhance their understanding of these explanations. Furthermore, participants demonstrated notable curiosity regarding extended questions, such as asking for new explanations or comparisons between different explanations. This indicates that as participants become more familiar with the provided explanations, they develop an interest in exploring alternative methods and understanding how models might behave in different scenarios.

To better understand the advantages of the proposed EMCEE model compared to the baseline LLaVa-1.5, we randomly selected 60 question-answer pairs from the conversations collected in the human evaluation. For each model, we selected 10 question-answer pairs from each of the three question categories. We then recruited three well-educated annotators to evaluate the answers based on two criteria: *Factual Correctness* and *Understandability*. Factual correctness assesses whether the responses are accurate, while understandability measures whether the responses are easy to comprehend. Each criterion is rated as either 0 or 1. Table 7 presents the results, which showed that the EMCEE model consistently generated more factually correct answers across all three categories, compared to the baseline model. This improvement can be attributed to the use of a hallucination detector during the training phase, which removes factually incorrect statements from the synthetic data and reduces the hallucinations in the final model. Regarding understandability, EMCEE outperforms the baseline, particularly in questions related to the provided explanations. This is likely due to the method used for generating synthetic conversations, where both questions and answers are conditioned on the explanations. As a result, when trained on this data, the EMCEE model

Table 6: Overview of Collected Questions. Including categories of questions, examples, and the number of questions in each category.

Question Category	Question Examples	Num
Generic questions about machine learning and explainable AI concepts	• What is a deep learning model? • What is Swin Transformer? • What is an explanation method?	87
Questions related to the provided explanations:	• How does SHAP determine the regions of the image that are most important for the prediction? • How does it mean by the output changes when the input changes (in Integrated Gradients)? • Would the Grad-CAM get wrong?	168
Extended questions	• Can I use grad-cam for an image containing more than 1 type of animals? • What if some important or unique parts of the animal are blocked? How should this image be classified, and can you provide such explanations? • What are the potential limitations when using SHAP in practical applications?	103

Model	Question Category 1		Question Category 2		Question Category 3	
	Understandability	Factual correctness	Understandability	Factual correctness	Understandability	Factual correctness
LLaVa-1.5	0.77	0.70	0.63	0.77	0.78	0.70
EMCEE(ours)	0.78	0.83	0.83	0.87	0.80	0.80

Table 7: Understandability and Factual Correctness of replies generated by EMCEE and LLaVa-1.5. Two scores are rated as 0 or 1. The best results are in boldface. We measure the inter-rater reliability with Fleiss' Kappa [24]. Our annotations obtain "moderate agreement" for Understandability (0.57) and "substantial agreement" for Factual Correctness (0.675).

becomes more adept at answering questions about the explanations provided.

Type	Utterance
Q:	Does the XAI provide explanations about how the classification model made the prediction?
A:	No, it only gives local explanations about what the model predicts.
Q:	Are there any limitations to the Grad-CAM?
A:	No, there are no limitations to the method.
Q:	Is the process of SHAP complicated?
A:	No, SHAP is a relatively simple process, even though it provides deeper feature attribution insights than traditional methods such as permutation importance tests.
Q:	Does LIME produce probabilities or classifications?
A:	LIME produces classifications, not probabilities.

Table 8: Examples of conversation turns that are identities as hallucinations by the detector.

5.2.3 Examples of Identified Hallucinations in Generated Conversations. To qualitatively assess the effectiveness of the hallucination detector, we selected four conversation turns identified as hallucinations by the detector, as presented in Table 8. These examples demonstrate that LLMs tend to generate untruthful responses about both fundamental machine learning concepts and various XAI techniques. The hallucination detector in our model can successfully identify and exclude such incorrect turns from the synthetic dataset. Consequently, the hallucination detection and filtering process diminishes the occurrence of hallucinations in the synthetic data and enhances the performance of models finetuned on this refined dataset.

6 Limitations

We identified five limitations of the current work. First, the static explanations used in our study are limited. Our experiments focused on feature attribution explanation methods on image classification. Even though our method is applicable to any static explanation method, the performance of our model on other types of static explanation methods, such as example-based explanation methods, or NLP tasks, is yet to be explored. Second, with 79.5% accuracy on held-out test data, the hallucination detector is not perfect. Errors from the detector may cause hallucinated responses to slip through or valid responses to be incorrectly filtered. Third, we mainly focused on removing factuality hallucinations, while not considering faithfulness hallucinations [34]. Factuality hallucinations refer to

statements that are factually incorrect or fabricated. Faithfulness hallucinations refer to statements that are not related to instructions and contextual information. In data generation, our model also may generate unrelated conversations to the static explanations. We leave building a detector or using other methods to filter these unrelated conversations for future work. Fourth, previous work [21] indicates that prior knowledge of AI may influence participants' perception of explanations. We mitigated this potential confounding factor by randomly assigning participants. However, the effect of prior knowledge on the use of conversational XAI remains an open problem. Finally, our research is confined to one geographical region. Factors such as cultural backgrounds could potentially affect how users interact with XAI and how they seek to clarify confusion. Future studies could involve recruiting participants from diverse countries and regions.

7 Conclusion

This paper proposes the fEw-shot Multi-round ConvErsational Explanation (EMCEE) to provide customized explanations to users from diverse domains. To deal with data scarcity, we train the EMCEE with synthetic data. We first use a vision language model to generate synthetic conversations with the repetition penalty to promote the diversity of generated data. Then, to reduce hallucinations in generated data, we apply a hallucination detector to filter hallucinated conversation turns after the data generation. To iteratively improve the performance, we repeat the generation-filtering-finetuning process multiple times. Both automatic and human evaluation demonstrate that EMCEE outperforms baseline models by a large margin. In practice, EMCEE significantly improved users' comprehension, acceptance, trust, and collaboration with static explanations. By analyzing conversations, we demonstrate that EMCEE can generate more truthful and understandable responses, leading to a better user experience.

Acknowledgments

We gratefully acknowledge the support by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*. 9525–9536.

[3] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 412–421. https://doi.org/10.18653/v1/D17-1042

[4] D Anderson and K Burnham. 2004. Model selection and multi-model inference. *Springer-Verlag* 63 (2004), 512.

[5] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2022. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction* 0, 0 (2022), 1–16. https://doi.org/10.1080/10447318.2022.2138826

[6] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". In *The Twelfth International Conference on Learning Representations*.

[7] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery* 37, 5 (01 Sep 2023). https://doi.org/10.1007/s10618-023-00933-9

[8] Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. *arXiv Preprint 2311.16822* (2023).

[9] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1721–1730. https://doi.org/10.1145/2783258.2788613

[10] Yuanyuan Chen, Boyang Li, Han Yu, Pengcheng Wu, and Chunyan Miao. 2021. Hydra: Hypergradient data relevance analysis for interpreting deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7081–7089.

[11] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[12] Paulo Cortez and Mark J Embrechts. 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences* 225 (2013), 1–17.

[13] Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4037–4046.

[14] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2136–2148. https://doi.org/10.18653/v1/2023.eacl-main.156

[15] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 447–459.

[16] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340.

[17] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management science* 35, 8 (1989), 982–1003.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[19] El Bachir Diop, Shengchuan Zhao, and Tran Van Duy. 2019. An extension of the technology acceptance model for understanding travelers' adoption of variable message signs. *PLoS one* 14, 4 (2019).

[20] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[21] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).

[22] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 3764–3814.

[23] Nils Feldhus, Ajay Madhavan Ravichandran, and Sebastian Möller. 2022. Mediators: Conversational agents explaining NLP model behavior. *arXiv preprint arXiv:2206.06029* (2022).

[24] Joseph L. Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement* 33, 3 (1973), 613–619. https://doi.org/10.1177/001316447303300309

[25] Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, WEIZHONG ZHANG, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=h5OpjGd_lo6

[26] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint*

arXiv:2308.03296 (2023).

[27] Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10333–10350. https://doi.org/10.18653/v1/2021.emnlp-main.808

[28] Xu Guo and Yiqiang Chen. 2024. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. *arXiv preprint arXiv:2403.04190* (2024).

[29] Yaohui Guo and X. Jessie Yang. 2021. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics* 13 (2021), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

[30] Xin He, Yeyi Hong, Xi Zheng, and Yong Zhang. 2023. What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1519–1542. https://doi.org/10.1080/10447318.2022.2095093

[31] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[33] Linwei Hu, Jie Chen, Vijayan N Nair, and Agus Sudjianto. 2018. Locally interpretable models and effects based on supervised partitioning (LIME-SUP). *arXiv preprint arXiv:1806.00663* (2018).

[34] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).

[35] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. 2019. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1511–1519.

[36] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4198–4205.

[37] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[38] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. 2020. How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. In *Advances in Neural Information Processing Systems*, Vol. 33. 4211–4222.

[39] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[40] J. F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems* 2, 1 (1984), 26–41. https://doi.org/10.1145/357417.357420

[41] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2288–2296.

[42] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 17 pages. https://doi.org/10.1145/3544548.3581001

[43] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (Un)reliability of Saliency Methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (2019), 267–280.

[44] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems* 34 (2021), 2611–2624.

[45] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[46] Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. 16–21. https://aclanthology.org/2021.hackashop-1.3

[47] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[48] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01875* (2022).

[49] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems* 35 (2022), 34586–34599.

[50] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 681–691. https://doi.org/10.18653/v1/N16-1082

[51] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[52] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[53] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[54] Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, 605–612. https://doi.org/10.3115/1218955.1219032

[55] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning.

[56] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.

[57] Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. 2018. On interpretation of network embedding via taxonomy induction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1812–1820.

[58] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[59] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[60] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.

[61] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 462–477.

[62] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*. 24457–24477.

[63] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[64] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[65] Bonnie M Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (1996), 429–460.

[66] Giang Nguyen, Valerie Chen, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. PCNN: Probable-Class Nearest-Neighbor Explanations Improve Fine-Grained Image Classification Accuracy for AIs and Humans. *arXiv preprint arXiv:2308.13651* (2024).

[67] Giang Nguyen, Mohammad Reza Taesiri, Sunnie SY Kim, and Anh Nguyen. 2024. Allowing humans to interactively guide machines where to look does not always improve a human-AI team's classification accuracy. *arXiv preprint arXiv:2404.05238* (2024).

[68] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[69] Julia Powles and Hal Hodson. 2017. Google DeepMind and healthcare in an age of algorithms. *Health and Technology* 7, 4 (2017), 351–367. https://doi.org/10.1007/s12553-017-0179-1

[70] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[71] Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP Models via Minimal Contrastive Editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3840–3852. https://doi.org/10.18653/v1/2021.findings-acl.336

[72] Filip Rudziński. 2016. A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers. *Applied Soft Computing* 38 (2016), 118–133.

[73] Katja Schwarz, Yiyi Liao, and Andreas Geiger. 2021. On the frequency bias of generative models. *Advances in Neural Information Processing Systems* 34 (2021), 18126–18136.

[74] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *Comput. Surveys* 54, 4 (2021), 43 pages. https://doi.org/10.1145/3386867

[75] Sarah Sebo, Ling Liang Dong, Nicholas Chang, Michal Lewkowicz, Michael Schutzman, and Brian Scassellati. 2020. The influence of robot verbal support on human team members: Encouraging outgroup contributions and suppressing ingroup supportive behavior. *Frontiers in Psychology* (2020), 3584.

[76] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[77] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 384–387. https://doi.org/10.1145/3584931.3607492

[78] Andy Shih, Arthur Choi, and Adnan Darwiche. 2018. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364* (2018).

[79] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv Preprint 2305.17493* (2024).

[80] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2023. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of XAI on Human-Agent Interaction. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

[81] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[82] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* 5, 8 (2023), 873–883. https://doi.org/10.1038/s42256-023-00692-8

[83] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. 3319–3328.

[84] E. S. Vorm and David J. Y. Combs. 2022. Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM). *International Journal of Human–Computer Interaction* 38, 18-20 (2022), 1828–1845. https://doi.org/10.1080/10447318.2022.2070107

[85] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[86] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328. https://doi.org/10.1145/3397481.3450650

[87] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024. Gradient based Feature Attribution in Explainable AI: A Technical Review. *arXiv preprint arXiv:2403.10415* (2024).

[88] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[89] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6707–6723. https://doi.org/10.18653/v1/2021.acl-long.523

[90] Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating explanation without ground truth in interpretable machine learning. *arXiv preprint arXiv:1907.06831* (2019).

[91] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. 2017. Scalable Bayesian rule lists. In *International conference on machine learning*. 3921–3930.

[92] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. ACM, 408–416. https://doi.org/10.1145/2909824.3020230

[93] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 11653–11669. https://doi.org/10.18653/v1/2022.emnlp-main.801

[94] Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022. ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3671–3683. https://doi.org/10.18653/v1/2022.findings-emnlp.269

[95] Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting Attributions and QA Model Behavior on Realistic Counterfactuals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5496–5512. https://doi.org/10.18653/v1/2021.emnlp-main.447

[96] Kayo Yin and Graham Neubig. 2022. Interpreting Language Models with Contrastive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 184–198.

[97] Jieyu Zhang, Bohan Wang, Zhengyu Hu, Pang Wei W Koh, and Alexander J Ratner. 2024. On the trade-off of intra-/inter-class diversity for supervised pre-training. *Advances in Neural Information Processing Systems* 36 (2024).

[98] Tong Zhang, X Jessie Yang, and Boyang Li. 2023. May I Ask a Follow-up Question? Understanding the Benefits of Conversations in Neural Network Explainability. *International Journal of Human–Computer Interaction* (2023).

[99] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *arXiv preprint arXiv:2304.10513* (2023).