Handling the Follow-up Question: Conversational Explanations for Image Classification

Anonymous ACL submission

Abstract

Explainable AI (XAI) aims to provide insights into decisions made by deep neural networks. 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. The proposed system, fEw-shot Multi-round ConvErsational Explanation (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, EMCEE outperforms baseline models in improving users' comprehension, acceptance, trust, and collaboration with static explanations by large margins. To the best of our knowledge, this is the first conversational explanation method that can answer arbitrary user questions that follow from static explanations.

1 Introduction

Despite the high accuracy of deep neural networks (DNNs), in high-stake and mission-critical applications like healthcare, finance, and law enforcement, it remains necessary for human domain experts to verify the DNN decisions and examine the reasoning process in order to prevent catastrophic failures (Caruana et al., 2015; Powles and Hodson, 2017). To this end, in recent years, much research has been devoted to eXplainable Artificial Intelligence, or XAI (e.g., Selvaraju et al. 2017; Lundberg and Lee 2017; Chen et al. 2021).

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, which are not met by existing XAI techniques (Liao et al., 2020; Liao and Varshney, 2021; Zhang et al., 2023). The lack of customization causes insufficient understanding of model behavior and undermines human-AI collaboration (Zhang et al., 2023). Indeed, recent studies found that the end users and domain experts with limited machine learning knowledge still struggle to understand and use the XAI explanations (Ehsan et al., 2021; Wang and Yin, 2021).

Conversational explanations have been suggested as a suitable solution for providing customized explanations to users (Liao et al., 2020; Feldhus et al., 2022; Lakkaraju et al., 2022; Zhang et al., 2023), as they allow XAI systems to answer arbitrary follow-up questions from the user after they see the static explanation. Lakkaraju et al. (2022) discover that human decision makers have a strong preference for explanations in the form of natural language dialogue. They argue that conversational explanations can provide personalized responses and information based on users' conversational histories. Zhang et al. (2023) show that answering user questions following the static explanations can significantly improve participants' 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 on two types of static explanations (Zhang et al., 2023). To date, existing conversational explanations are based on human-authored templates, which can cope only with a limited and predefined range of user questions (Slack et al., 2023; Shen et al., 2023).

To handle data scarcity, a natural thought is to generate synthetic conversations using large vision language models (VLMs), which may answer technical questions to a degree (Hellas et al., 2023). However, training with synthetic data encounters two primary challenges: the lack of data diversity and model hallucination.

The first challenge, the lack of data diversity, arises as generative models tend to overrepresent high-frequency content (Schwarz et al., 2021; Shumailov et al., 2024; Briesch et al., 2023) 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 (Lee et al., 2022; Ji et al., 2023; Dai et al., 2023; Zheng et al., 2023; Berglund et al., 2024). To mitigate the hallucinated, factually incorrect answers, we train 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 conduct both automatic and human evaluations on the proposed system, fEw-shot Multi-round ConvErsational Explanation (EMCEE). The automatic evaluation is conducted on the only existing conversational explanation dataset (Zhang et al., 2023). For the human evaluation, we evaluate user comprehension, acceptance and trust in XAI, and user's ability to choose the best AI models using only the explanations. Empirical results show that EMCEE outperforms the baseline LLaVa-1.5 model in both automatic and human evaluations. Repeated training on self-generated data leads to data degeneration in diversity and quality (Briesch et al., 2023). We demonstrate that the proposed repetition penalty and hallucination detection can slow down the data degeneracy in training with synthetic data. In practice, our model significantly improves participant's comprehension, acceptance, trust, and collaborative performance.

Our contributions can be summarized as follows.

- To the best of our knowledge, we propose the first conversational explanation that can answer free-form follow-up questions after providing static explanations to the user.
- We propose a repetition penalty to enhance data diversity and a hallucination detector to reduce erroneous information in synthetic data.
- The proposed method EMCEE outperforms the baseline model in both automatic and human evaluation by large margins.

2 Methodology

The overall workflow of EMCEE is illustrated as Figure 1 and outlined in Algorithm 1. Starting from a pretrained VLM V_1, we generate a set of synthetic conversations D_1, while using the repetition penalty to encourage data diversity. Each conversation may contain 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 apply 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 finetune the VLM on D_1^{clean}, leading to the next VLM V_2, from which we start another round of generation-filter-finetuning. This process is repeated multiple times. We do not reuse synthetic data from previous rounds.

We design 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 a number of 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 synthetic data generation and mode finetuning, we randomly choose 0 or 1 demonstration and keep it consistent for each mini-batch. During model inference and evaluation, the number of demonstrations ranges between zero and three.

2.1 Repetition Penalty

The repetition penalty encourages the VLM to generate more diverse conversations by discounting the logits of tokens seen in previous conversation turns.

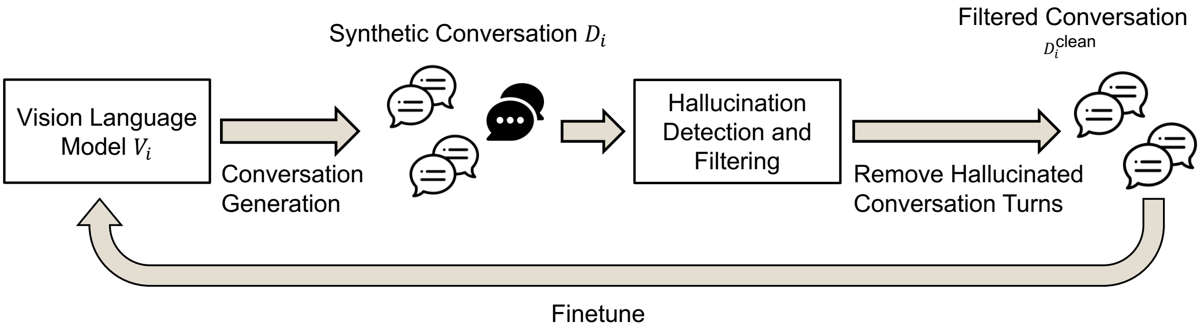

Synthetic Conversation D_i
Filtered Conversation D_i^{clean}
Vision Language Model V_i
Conversation Generation
Hallucination Detection and Filtering
Remove Hallucinated Conversation Turns
Finetune

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

```
for $r$ in $1 \ldots R$ do
    $\mathcal{D}_r \leftarrow$ generate $N$ conversations from $V_r$;
    $D_i^{\text{clean}} \leftarrow \{(\boldsymbol{x}, \boldsymbol{y}) \in D_r \mid f_h(\boldsymbol{y}) \neq 1\}$;
    $V_{r+1} \leftarrow$ finetune $V_r$ on $D_i^{\text{clean}}$;
end for
```

Specifically, given the logits z_i for each token i in the vocabulary, the probability p_i of predicting token i is computed as,

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

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.

2.2 Hallucination Detection and Filtering

VLMs often generate convincing but factually incorrect statements, especially when answering questions that require reasoning and logical deduction (Lee et al., 2022; Ji et al., 2023; Dai et al., 2023; Zheng et al., 2023; Berglund et al., 2024). 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 integrate a hallucination detector into the training process, which identifies and removes hallucinated conversation turns. To train the hallucination 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 a number of students studying machine learning. The remaining 1,000 sentences are about XAI knowledge; we use GPT-4-turbo-2024-04-09 to generate 500 factually correct sentences about XAI and subsequently altered them be incorrect. All generated sentences have been rigorously validated by XAI experts. Examples of sentences included in the dataset are displayed in Appendix E. 80% of the collected data are used for training, whereas 20% data are reserved for validation and testing.

3 Experiment

3.1 Experimental Protocol

We used LLaVa-1.5 (Liu et al., 2023b,a) 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 (Liu et al., 2023b,a).

We focus on the image classification task on the ImageNet dataset and train three classification models with different top-1 classification accuracies: Swin Transformer (84.1%), VGG-16 (71.6%), and AlexNet (56.5%). To generate explanations for model predictions, we adopt four explanation techniques from feature attribution methods: LIME (Ribeiro et al., 2016), Grad-CAM (Selvaraju et al., 2017), Integrated Gradients (Sundararajan et al., 2017), and SHAP (Lundberg and Lee, 2017). 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 (Kim et al., 2023).

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 temperature is set to 1.2 and the repetition penalty ratio is set to 1.1.

For finetuning LLaVa-1.5, we use LoRA (Hu et al., 2021) to only finetune the language model with the vision encoder and the 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.

For the hallucination detector, we train a Bert-base model (Devlin et al., 2019) using the SGD optimizer with a learning rate of 0.01, batch size of 16, and weight decay for 100 epochs. The hallucination detector received 79.5% accuracy on the held-out test set.

3.2 Evaluation

We conduct both automatic and human evaluations to demonstrate the effectiveness of the proposed model. For automatic evaluations, we conduct few-shot evaluations with 0 to 3 demonstrations. We leverage BLEU (Papineni et al., 2002) and ROUGE (Lin and Och, 2004) scores to measure word overlaps between the generated response text and ground truth text.

For human evaluations, we evaluate the practical effects of different conversational explanation models in improving participants' comprehension (Cheng et al., 2019), acceptance (Davis, 1989), and trust (Yang et al., 2017b) in static explanations. Based on the results of automatic evaluations, we use 2 demonstrations for conversational explanations on Grad-CAM and LIME. Due to the lack of real human conversations, we do not use demonstrations for Integrated Gradients and SHAP. We recruited $N = 40$ participants from 14 majors. Each participant engaged in the study only once. We first presented them with the static explanations for the image classification task and measure their objective understanding and subjective perceptions of static explanations. After that, half of the participants went through an online textual conversation with the pretrained LLaVa-1.5 model, during which they could seek to clarify any doubts. The other half interacted with our models. Details of the online textual conversation platform are in Appendix B.

We asked the participants to choose one model from three candidate classification models that would be the most accurate on unobserved test data and use the selection accuracy as a measurement of their objective understanding of the static explanations. The three classification models made identical decisions on 5 images. The only differences between the three networks lay in their explanations. Hence, to select the best model, the participants must rely on the explanations. The details of how the explanation images are selected and the full set of images are in Appendix C.

To measure participants' subjective perception of static explanations, we use the same set of 13 self-reporting questions in the previous study (Zhang et al., 2023). These self-reporting questions probe participants' comprehension, acceptance, and trust in explanations. All questions utilize a 7-point Likert scale for responses. The full list of the questions is in Appendix D.

3.3 Dataset

We conducted our automatic evaluation using the only existing dataset from human-human conversational XAI (Zhang et al., 2023), gathered in a Wizard-of-Oz (WoZ) setting (Kelley, 1984). Participants interacted with what they believed was an autonomous dialogue system, which was actually operated by a human expert of machine learning and XAI. Participants were recruited from 19 different disciplines. 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 do not use this dataset for training. We employ 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. The rest 52 conversations are used for testing.

3.4 Results of Automatic Evaluation

Table 1 presents the automatic evaluation results of both the pretrained LLaVa-1.5 model and our EMCEE model when we prompt them with 0 to 3 example conversations. Our method exhibits substantial improvements over the pretrained LLaVa-

Table 1: 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

1.5 in terms of both BLEU and ROUGE scores. Specifically, our model shows an increase of 81.6% in BLEU scores and 80.5% in ROUGE scores compared to the pretrained LLaVa-1.5. These results suggest that our model, which has been 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.

3.5 Results of Human Evaluation

Table 2 presents human evaluation results, comparing the pretrained LLaVa-1.5 model and EMCEE across four static explanation methods, LIME, Grad-CAM, Integrated Gradients, and SHAP.

Participants' objective understanding improves with both LLaVa-1.5 and EMCEE on all static explanation; however, participants interacting with our model consistently demonstrate greater increase in the model selection accuracy post-conversation, demonstrating strong positive effects of training on synthetic data in assisting participants collaborating with static explanations.

We observe varied objective performance among LIME, Grad-CAM, Integrated Gradients, and SHAP. Grad-CAM has the highest accuracy of objective decision accuracy and Integrated Gradients has the lowest accuracy. A potential reason might be the inherently intuitive nature of the explanations produced by Grad-CAM compared to others.

In terms of participants' subjective understanding, participants who receive conversational explanations from EMCEE report a significantly greater improvement than those who interacted with LLaVa-1.5, across all four static explanation methods. Initially, there is no notable difference in the participants' self-reported understanding of static explanations. Participants using the EMCEE model report a higher level of understanding than those who interacted with the LLaVa-1.5 model.

For acceptance of explanations, we observed similar patterns in participants' subjective understanding. Participants' perceived usefulness, perceived ease of use, and behavioral intention all increase after interacting with LLaVa-1.5 or EMCEE, but the improvements brought by LLaVa-1.5 are much smaller than EMCEE. We hypothesize that the ability to resolve confusion with EMCEE partially causes the participants to perceive greater usefulness, ease of use, and tendency to use the static explanations.

For the trust measurement, we observed a marked rise in participants' trust levels across all four static explanation methods after interaction with our model. According to theories of trust (McKnight et al., 1998; Lim et al., 2009; Hoffman et al., 2018), 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 improved understanding of static explanations, as indicated by earlier results.

3.6 Ablation Study with Automatic Evaluation

We create 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 3 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

Table 2: Results of human evaluations before and after conversations. Each score is presented as mean ± standard deviation and the change $\delta = \text{after} - \text{before}$. * indicates that change δ caused by our model is statistically higher than that from the baseline model, LLaVa-1.5, with $p < 0.05$ using the Student's t-test.

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	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 ± 1.30	5.60 ± 0.60	5.20 ± 0.60	5.20 ± 0.82	4.30 ± 0.52
		δ	0.08	0.80	0.40	0.80	0.30	0.20
	EMCEE (Ours)	before	0.36 ± 0.09	4.20 ± 1.30	5.33 ± 0.80	4.53 ± 0.92	5.00 ± 0.65	4.20 ± 0.45
		after	0.52 ± 0.11	5.20 ± 0.55	5.93 ± 0.87	5.60 ± 0.68	5.60 ± 0.76	4.80 ± 0.42
		δ	**0.16***	**1.00***	**0.60***	**1.07***	**0.60***	**0.60***
Grad-CAM	LLaVa-1.5	before	0.76 ± 0.17	4.00 ± 1.41	5.33 ± 0.41	4.87 ± 0.60	5.50 ± 0.35	4.40 ± 0.29
		after	0.84 ± 0.09	4.80 ± 0.45	5.60 ± 0.44	5.13 ± 0.38	5.80 ± 0.27	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 ± 1.09	5.30 ± 0.69	4.15 ± 0.72
		after	0.92 ± 0.11	5.40 ± 0.89	6.13 ± 0.61	5.40 ± 0.93	6.10 ± 0.45	5.25 ± 0.90
		δ	**0.12**	**1.40***	**1.00***	**0.60***	**0.80***	**1.10***
Integrated Gradients	LLaVa-1.5	before	0.24 ± 0.09	3.80 ± 0.45	4.73 ± 0.28	3.87 ± 0.77	4.40 ± 1.08	3.85 ± 0.42
		after	0.28 ± 0.18	4.00 ± 1.10	5.00 ± 0.84	4.40 ± 1.60	4.70 ± 1.20	3.85 ± 0.38
		δ	0.04	0.20	0.27	0.53	0.30	0.00
	EMCEE (Ours)	before	0.20 ± 0.14	3.80 ± 0.55	4.87 ± 0.89	3.60 ± 0.64	4.50 ± 0.79	3.85 ± 0.55
		after	0.44 ± 0.09	4.60 ± 0.45	5.20 ± 0.61	4.73 ± 0.60	5.50 ± 0.67	4.40 ± 0.80
		δ	**0.24***	**0.80***	**0.33**	**1.13***	**1.00***	**0.55***
SHAP	LLaVa-1.5	before	0.48 ± 0.11	3.80 ± 1.79	5.40 ± 0.60	4.87 ± 1.73	5.00 ± 1.06	4.20 ± 1.47
		after	0.60 ± 0.14	5.40 ± 0.84	5.60 ± 0.55	5.67 ± 0.78	5.20 ± 0.91	4.60 ± 0.84
		δ	0.12	1.60	0.20	0.80	0.20	0.40
	EMCEE (Ours)	before	0.50 ± 0.48	3.75 ± 1.89	5.43 ± 0.58	4.58 ± 1.77	5.00 ± 0.71	4.25 ± 1.14
		after	0.80 ± 0.16	5.50 ± 1.29	6.13 ± 0.82	6.00 ± 0.47	5.78 ± 0.48	5.31 ± 0.94
		δ	**0.30***	**1.75**	**0.70***	**1.42***	**0.78***	**1.06***

h

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 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.

3.7 Effects of Multiple Generation-Training Iterations

In the training of EMCEE, we repeat the generation-training process multiple times. We investigate how iterations affect the performance of EMCEE and ablated versions of EMCEE in BLEU-4 and ROUGE-L scores, as shown in Figure 2.

We observe that the ablated versions of EMCEE improve in the first few iterations and decrease afterwards. This is similar to the findings of Briesch et al. (2023), who show that repeatedly training models with self-generated data initially causes performance gains but, after a few iterations, results in degenerate synthetic data with low diversity and eventual performance drop. This is especially apparent when we remove the repetition penalty or the hallucination filter, as both BLEU-4 and ROUGE-L decrease drastically after the third and fifth iteration, respectively.

However, with both the repetition penalty and the hallucination filter of EMCEE, the performance drops become substantially milder. For BLEU-4, a small drop is observed after the fifth iteration. For Rouge-L, the performance effectively plateaus around the sixth and seventh iteration. We conclude that the proposed techniques, including the repetition penalty and the hallucination filter successfully slows down degeneracy in training with synthetic data.

3.8 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

Table 3: 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

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

presented in Table 4. 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 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.

4 Related Work

4.1 Static XAI

Explainable Artificial Intelligence (XAI) refers to techniques that explain the learning process or the predictions of AI (Yang et al., 2019). Most existing techniques are static XAI, which provides a one-time explanation with no capability for further user interaction. Two groups of static XAI include self-explanatory models and post-hoc methods. Self-explanatory models are inherently transparent, offering clarity in their decision-making processes (Lakkaraju et al., 2016; Rudziński, 2016; Yang et al., 2017a; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). The majority of recent XAI methods are post-hoc XAI methods, applied to already developed models that lack inherent transparency (Selvaraju et al., 2017; Ribeiro et al., 2016; Chen et al., 2021; Adadi and Berrada, 2018; Bodria et al., 2023). There are two main groups of methods in post-hoc XAI, i.e., feature attribution methods and example-based methods.

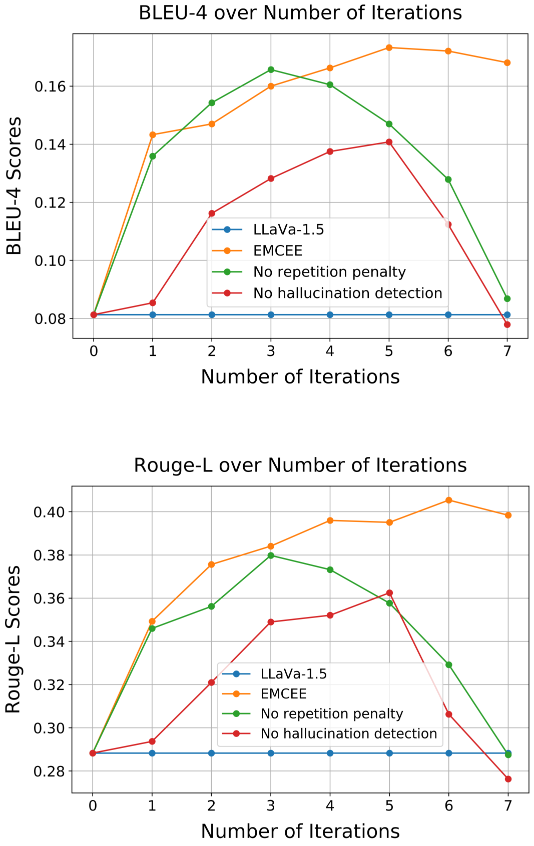


Figure 2: BLEU-4 and Rouge-L scores over the number of training iterations for LLaVa-1.5, EMCEE and different ablated version of EMCEE.

Feature Attribution. Feature attribution methods explain model predictions by investigating the importance of input features to final predictions

(Adadi and Berrada, 2018; Danilevsky et al., 2020). There are two main types of feature attribution methods, gradient-based methods (Cortez and Embrechts, 2013; Sundararajan et al., 2017; Selvaraju et al., 2017; Simonyan et al., 2013; Lundberg and Lee, 2017; Wang et al., 2024; Kokalj et al., 2021; Li et al., 2016) and surrogate methods (Ribeiro et al., 2016; Hu et al., 2018; Alvarez-Melis and Jaakkola, 2017; Liu et al., 2018; Shih et al., 2018; Ignatiev et al., 2019). 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 linear model, to locally approximate the complex neural network.

Example-based Methods. Example-based methods explain AI predictions by identifying a selection of data instances (Adadi and Berrada, 2018; Danilevsky et al., 2020; Nguyen et al., 2024). These instances may be training data points the most influential to the parameters of a prediction model (Chen et al., 2021; Guo et al., 2021), counterfactual examples that alter predictions with minimal changes to inputs (Wachter et al., 2017; Mothilal et al., 2020; Yin and Neubig, 2022; Ye et al., 2021; Ross et al., 2021; Wu et al., 2021), or prototypes that contain semantically similar parts to input instances (Croce et al., 2019; Jeyakumar et al., 2020; Kim et al., 2016).

4.2 Conversational XAI

Research into Conversational XAI is still at an early stage with limited methods being developed so far. Shen et al. (2023) apply conversational explanations to scientific writing tasks, observing improvements in productivity and sentence quality. Slack et al. (2023) design dialogue systems to help users better understand machine learning models on diabetes prediction, rearrest prediction, and loan default prediction tasks. Despite these advances, the conversations in these studies are generated based on templates and can only cope with limited predefined user queries. Our work represents the first system that can deliver free-form explanatory conversations with users about static explanations.

4.3 Training with Synthetic Data

The exceptional performance of Large Language Models (LLMs) and Vision Language Models (VLMs) in generating human-like text has led researchers to explore their use as training data generators (Meng et al., 2022; Ye et al., 2022a; Guo and Chen, 2024; Gao et al., 2023; Meng et al., 2023; Ye et al., 2022b). For example, SuperGen (Meng et al., 2022) uses LLMs conditioned on label-descriptive prompts to generate training data for text classification tasks. FewGen (Meng et al., 2023) fine-tune an LLM on few-shot samples and use it to generate synthetic data for seven classification tasks in the GLUE benchmark.

To mitigate the detrimental effects of noisy and low-quality synthetic data from LLMs and VLMs (Schwarz et al., 2021; Zhang et al., 2024; Kirk et al., 2021; Esiobu et al., 2023; Lee et al., 2022; Ji et al., 2023), several methods have been proposed (Gao et al., 2023; Guo and Chen, 2024; Meng et al., 2023; Ye et al., 2022b). For example, ProGen (Ye et al., 2022b) adjusts the importance of generated data points with regard to the validation loss, using influence function (Koh and Liang, 2017). 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 (Grosse et al., 2023).

In our work, we apply data generation to conversational explanations and utilize generated data to train the original VLM. We improve the quality of the generated data and significantly slow down model degeneracy after many generation-training iterations (see §3.7).

4.4 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 security, we train the EMCEE with synthetic data. We first use a vision language model (VLM) 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 recycle the generation-filter-finetuning process multiple times. Both automatic and human evaluation demonstrated that EMCEE outperforms baseline models by a large margin. In practice, EMCEE significantly improved users' comprehension, acceptance, trust, and collaboration with static explanations.

4.5 Limitations

We identify three 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, we mainly focus on removing factuality hallucinations, while not considering faithfulness hallucinations (Huang et al., 2023). 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.

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.

References

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

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*, pages 9525–9536.

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*, pages 412–421.

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*.

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).

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*.

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*, page 1721–1730.

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*, volume 35, pages 7081–7089.

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*, pages 1–12.

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

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)*, pages 4037–4046.

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*, pages 2136–2148.

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*, pages 447–459.

Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

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*, pages 4171–4186.

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*.

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*, pages 3764–3814.

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

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*.

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*.

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*, pages 10333–10350.

Xu Guo and Yiqiang Chen. 2024. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*.

Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the responses of large language models to beginner programmers' help requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research*, pages 93–105.

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

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*.

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*.

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*.

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*, pages 1511–1519.

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*, pages 4198–4205.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

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*, volume 33, pages 4211–4222.

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. *ACM Computing Surveys*, 55(12):1–38.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1):26–41.

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*, page 2288–2296.

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*.

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*, pages 267–280.

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*, 34:2611–2624.

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

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*, pages 16–21.

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*, pages 1675–1684.

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*.

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*, 35:34586–34599.

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*, pages 681–691.

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*, page 1–15.

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

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*, pages 2119–2128.

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)*, pages 605–612, Barcelona, Spain.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

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*, pages 1812–1820.

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

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

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*, 35:462–477.

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*, pages 24457–24477.

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*, pages 607–617.

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*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Julia Powles and Hal Hodson. 2017. Google deepmind and healthcare in an age of algorithms. *Health and Technology*, 7(4):351–367.

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*, pages 1135–1144.

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*, pages 3840–3852.

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

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

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*, pages 618–626.

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*, page 384–387.

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

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*.

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

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):873–883.

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

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

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*, page 318–328.

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*.

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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

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)*, pages 6707–6723.

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

Hongyu Yang, Cynthia Rudin, and Margo Seltzer. 2017a. Scalable bayesian rule lists. In *International conference on machine learning*, pages 3921–3930.

X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017b. 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*, pages 408–416.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683.

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*, pages 5496–5512.

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*, pages 184–198.

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.

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.

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

A VLM Prompts

The prompt contains an instruction to generate a conversation, the background information about the conversation, and a number of demonstration conversations. Example prompts for LIME, Grad-CAM, Integrated Gradients, and SHAP are shown in Figure 3, 4, 5, and 6 respectively. The input images are randomly selected from ImageNet and the explanations are generated by the corresponding XAI method.

B Oneline Textual Conversation Platform

Our study is conducted on a web-based platform where participants can remotely finish the whole procedure of the experiment. The web-based platform will ensure that all communications between users and conversational agents are text-based and recorded. Figure 7 displays an example screenshot of the web page where participants discuss static explanations with different conversational agents. There are two sections on the page. On the left, the user sees 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. Within the chatbox on the right, the user can converse with the conversational agent to clarify the explanation. Through a conversation, a user can ask any questions or provide any comments related to the explanation on the left side.

C Objective Evaluation

The objective evaluation aims to evaluate users' objective understanding of static explanations. Participants are presented with 5 input images, on which the three classification models make the same decisions. The only differences between the three models lie in their explanations. Participants need to choose the one that would be the most accurate on unobserved test data. Hence, to make the correct selection, the participants must understand the explanations. We use the accuracy of selecting the correct model to measure participants' objective understanding of static explanations. The full set of images listed in the objective evaluation for LIME, Grad-CAM, Integrated Gradients, and SHAP are shown in Figure 8, 9, 10, and 11 respectively.

We observe that static explanations do not always faithfully reflect the actual workings of classification models (Adebayo et al., 2018; Kindermans et al., 2019; Jacovi and Goldberg, 2020) 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 indeed have more reasonable explanations. This approach allows users to easily pick the best classification models if they understand the static explanations well. We deem an explanation more reasonable when it focuses more on discriminative features that are unique to the predicted class and less on spurious features that are irrelevant to the class. A good model should have explanations that rely on multiple types of discriminative features. This is because a model relying on multiple features is robust and makes the correct decision even if some discriminative features are absent or occluded.

D Subjective Evaluation

The subjective evaluation measures participants' self-reported perception of the static explanations, including their comprehension, acceptance, and trust. We use the same 13 questions as the previous study (Zhang et al., 2023). All questions utilize a 7-point Likert scale for responses. The full list of the questions is in Figure 12.

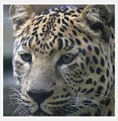
E Examples of Sentences in our Hallucination Dataset

To train the hallucination detector in MGCEE, we have collected a hallucination dataset about machine learning and XAI techniques. Table 5 displays 12 example sentences with labels in our dataset.

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 5: 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.

Instruction: A chat about explainable AI (XAI) between a curious human USER and an AI ASSISTANT. The human USER is well educated but may need help understanding how AI and XAI work. The USER asks questions to understand AI's decision-making process better. The USER's question should be diverse and related to AI and XAI. The ASSISTANT gives helpful, concise, detailed, and polite answers to the human's questions. Here is the background information for the conversation:
Task: Image classification
Given an image and 1000 predefined categories (goldfish, dog, bird, cat, etc), the algorithm identifies which category the image falls into.
Image classification model: swin transformer
Model's input:



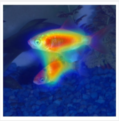
Model's prediction: Leopard
Explanation for the prediction:

Explanation method: LIME
Description of LIME:
LIME (Local Interpretable Model-Agnostic Explanations) is a technique used in machine learning to help explain the predictions made by complex AI models.
LIME works by creating a simpler, more interpretable model that approximates the behavior of the complex model in a small region around a particular data point. This simpler model is then used to explain why the complex model made a certain prediction for that data point. Regions of the image that are most important for the model's prediction are highlighted.
<**Demonstrations**>
The conversation starts:
USER:

Figure 3: The VLM prompt about LIME.

Instruction: A chat about explainable AI (XAI) between a curious human USER and an AI ASSISTANT. The human USER is well educated but may need help understanding how AI and XAI work. The USER asks questions to understand AI's decision-making process better. The USER's question should be diverse and related to AI and XAI. The ASSISTANT gives helpful, concise, detailed, and polite answers to the human's questions. Here is the background information for the conversation:
Task: Image classification
Given an image and 1000 predefined categories (goldfish, dog, bird, cat, etc), the algorithm identifies which category the image falls into.
Image classification model: swin transformer
Model's input:

Model's prediction: Leopard
Explanation for the prediction:



Explanation method: Grad-CAM
Description of Grad-CAM:
The Grad-CAM method is a technique used in computer vision to understand which parts of an image a deep learning model focuses on to make its prediction. It generates a heatmap that highlights the regions of the image that are most important for the prediction.
The heatmap is generated by weighting the activations of the final convolutional layer by their corresponding gradients and averaging the resulting weights spatially. The resulting heatmap is overlaid on the original image to provide a visual representation of the model's reasoning for its prediction. The heatmap is generated using a color gradient that ranges from blue to red. Bluer colors are used to represent areas of low importance, while redder colors indicate areas of high importance.
<**Demonstrations**>
The conversation starts:
USER:

Figure 4: The VLM prompt about Grad-CAM.

Instruction: A chat about explainable AI (XAI) between a curious human USER and an AI ASSISTANT. The human USER is well educated but may need help understanding how AI and XAI work. The USER asks questions to understand AI's decision-making process better. The USER's question should be diverse and related to AI and XAI. The ASSISTANT gives helpful, concise, detailed, and polite answers to the human's questions. Here is the background information for the conversation:

Task: Image classification

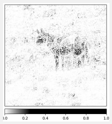
Given an image and 1000 predefined categories (goldfish, dog, bird, cat, etc), the algorithm identifies which category the image falls into.

Image classification model: swin transformer

Model's input:

Model's prediction: Leopard

Explanation for the prediction:



Explanation method: Integrated Gradients

Description of Integrated Gradients:

Integrated Gradients is a post-hoc technique used in machine learning to explain the predictions of deep learning models.

Integrated Gradients works by assigning a score to each feature in the input, representing its importance to the model's prediction. It calculates these scores by looking at how much the model's output changes when each part of the input changes. It does this by comparing the actual input to a baseline input (like a black image) and looking at all the intermediate inputs in between. Pixels with dark colors indicate greater importance for the model's prediction.

<**Demonstrations**>

The conversation starts:

USER:

Figure 5: The VLM prompt about Integrated Gradients.

Instruction: A chat about explainable AI (XAI) between a curious human USER and an AI ASSISTANT. The human USER is well educated but may need help understanding how AI and XAI work. The USER asks questions to understand AI's decision-making process better. The USER's question should be diverse and related to AI and XAI. The ASSISTANT gives helpful, concise, detailed, and polite answers to the human's questions. Here is the background information for the conversation:

Task: Image classification

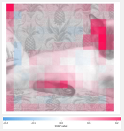
Given an image and 1000 predefined categories (goldfish, dog, bird, cat, etc), the algorithm identifies which category the image falls into.

Image classification model: swin transformer

Model's input:

Model's prediction: Leopard

Explanation for the prediction:



Explanation method: Integrated Gradients

Description of Integrated Gradients:

SHAP (SHapley Additive exPlanations) is a post-hoc explanation approach to explain the output of any machine learning model.

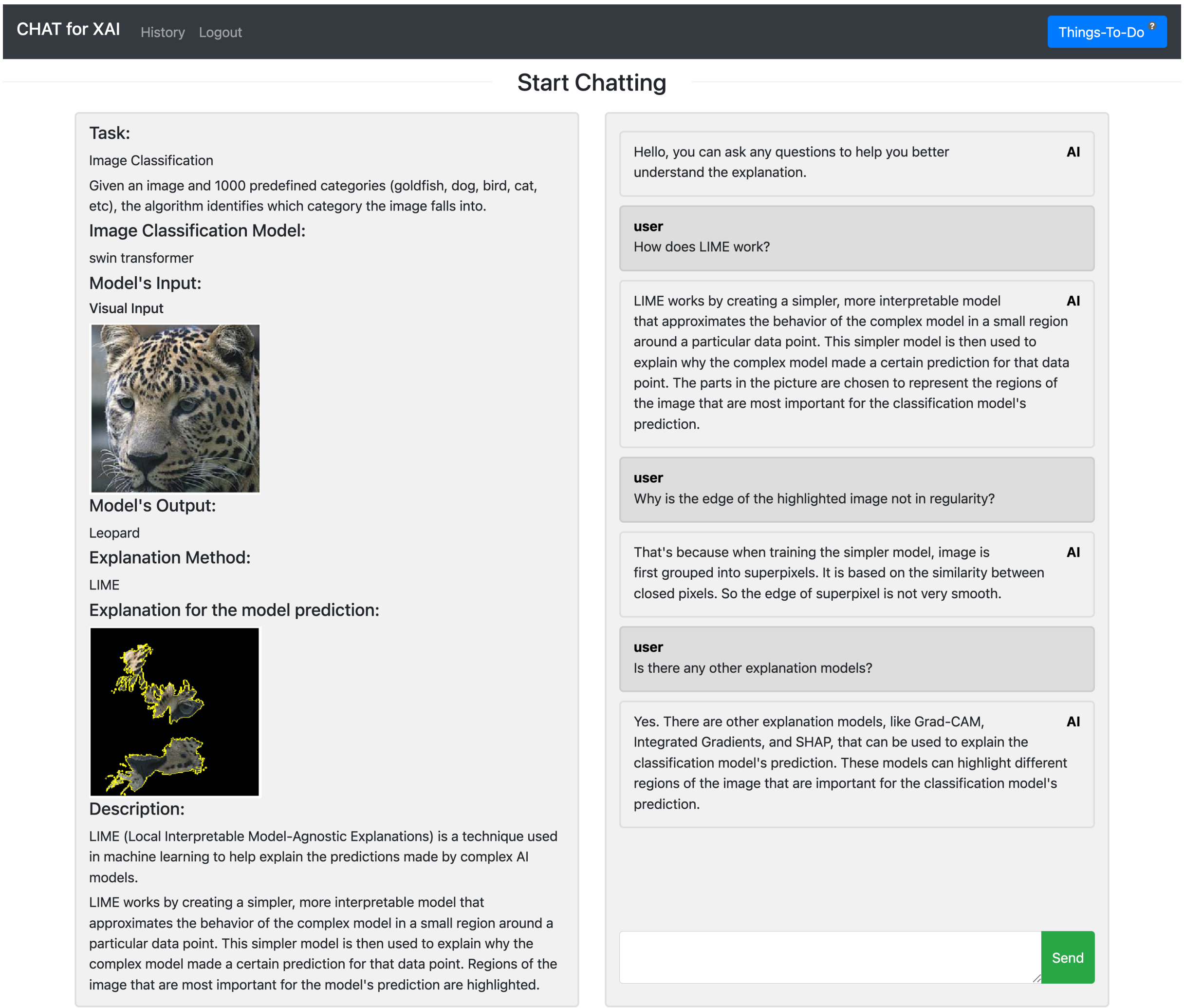
SHAP works by highlighting the regions of the image that are most important for the prediction. Each pixel in the explanation image refers to the importance value of pixels in the same location as the input image. Red pixels indicate that the pixels increase the probability of the particular class, truck. Blue pixels, on the other hand, decrease the probability of the class. Pixels with higher absolute values have higher importance in the classification.

The conversation starts:

<**Demonstrations**>

USER:

Figure 6: The VLM prompt about SHAP.



Start Chatting

Task:

Image Classification

Given an image and 1000 predefined categories (goldfish, dog, bird, cat, etc), the algorithm identifies which category the image falls into.

Image Classification Model:

swin transformer

Model's Input:

Visual Input

Model's Output:

Leopard

Explanation Method:

LIME

Explanation for the model prediction:

Description:

LIME (Local Interpretable Model-Agnostic Explanations) is a technique used in machine learning to help explain the predictions made by complex AI models.

LIME works by creating a simpler, more interpretable model that approximates the behavior of the complex model in a small region around a particular data point. This simpler model is then used to explain why the complex model made a certain prediction for that data point. Regions of the image that are most important for the model's prediction are highlighted.

AI
Hello, you can ask any questions to help you better understand the explanation.

user
How does LIME work?

AI
LIME works by creating a simpler, more interpretable model that approximates the behavior of the complex model in a small region around a particular data point. This simpler model is then used to explain why the complex model made a certain prediction for that data point. The parts in the picture are chosen to represent the regions of the image that are most important for the classification model's prediction.

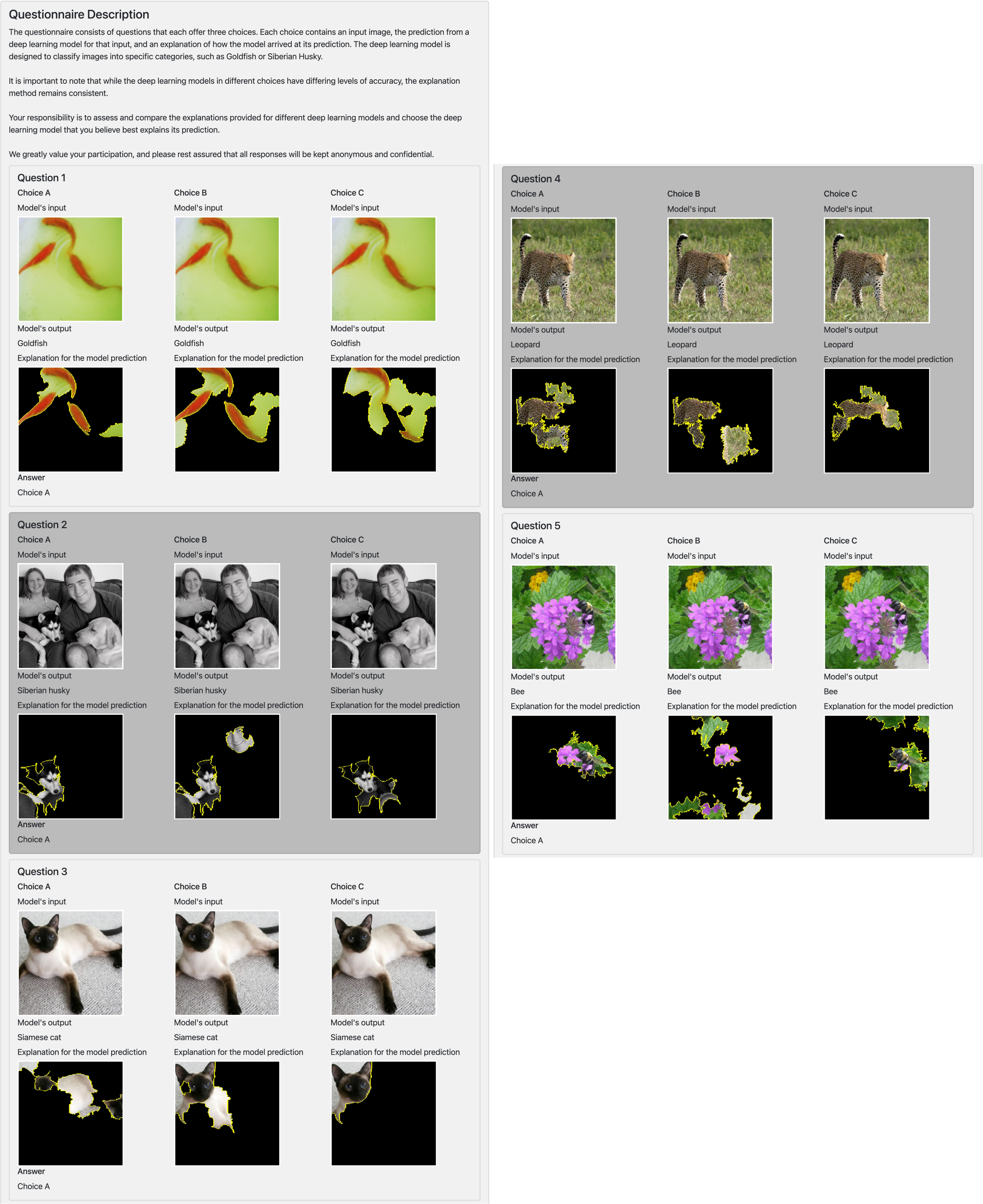
user
Why is the edge of the highlighted image not in regularity?

AI
That's because when training the simpler model, image is first grouped into superpixels. It is based on the similarity between closed pixels. So the edge of superpixel is not very smooth.

user
Is there any other explanation models?

AI
Yes. There are other explanation models, like Grad-CAM, Integrated Gradients, and SHAP, that can be used to explain the classification model's prediction. These models can highlight different regions of the image that are important for the classification model's prediction.

Figure 7: The web page where users can discuss static explanations with a conversational agent.



Questionnaire Description

The questionnaire consists of questions that each offer three choices. Each choice contains an input image, the prediction from a deep learning model for that input, and an explanation of how the model arrived at its prediction. The deep learning model is designed to classify images into specific categories, such as Goldfish or Siberian Husky.

It is important to note that while the deep learning models in different choices have differing levels of accuracy, the explanation method remains consistent.

Your responsibility is to assess and compare the explanations provided for different deep learning models and choose the deep learning model that you believe best explains its prediction.

We greatly value your participation, and please rest assured that all responses will be kept anonymous and confidential.

Question 1

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Goldfish	Goldfish	Goldfish
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer
Choice A

Question 2

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siberian husky	Siberian husky	Siberian husky
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer
Choice A

Question 3

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siamese cat	Siamese cat	Siamese cat
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer
Choice A

Question 4

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Leopard	Leopard	Leopard
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

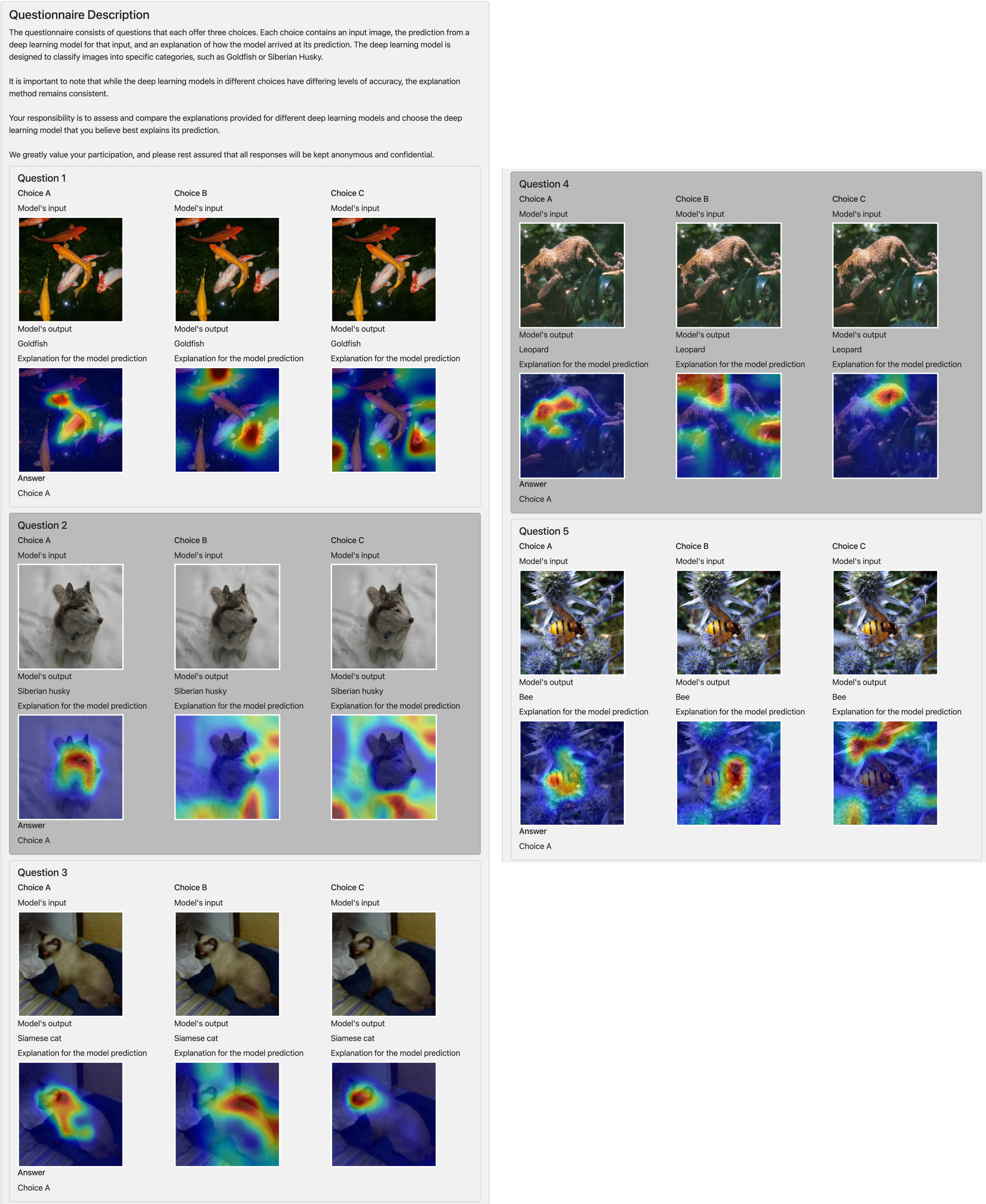
Answer
Choice A

Question 5

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Bee	Bee	Bee
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer
Choice A

Figure 8: Objective evaluation questions used for LIME.



Questionnaire Description

The questionnaire consists of questions that each offer three choices. Each choice contains an input image, the prediction from a deep learning model for that input, and an explanation of how the model arrived at its prediction. The deep learning model is designed to classify images into specific categories, such as Goldfish or Siberian Husky.

It is important to note that while the deep learning models in different choices have differing levels of accuracy, the explanation method remains consistent.

Your responsibility is to assess and compare the explanations provided for different deep learning models and choose the deep learning model that you believe best explains its prediction.

We greatly value your participation, and please rest assured that all responses will be kept anonymous and confidential.

Question 1

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Goldfish	Goldfish	Goldfish
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction
Answer		
Choice A		

Question 2

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siberian husky	Siberian husky	Siberian husky
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction
Answer		
Choice A		

Question 3

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siamese cat	Siamese cat	Siamese cat
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction
Answer		
Choice A		

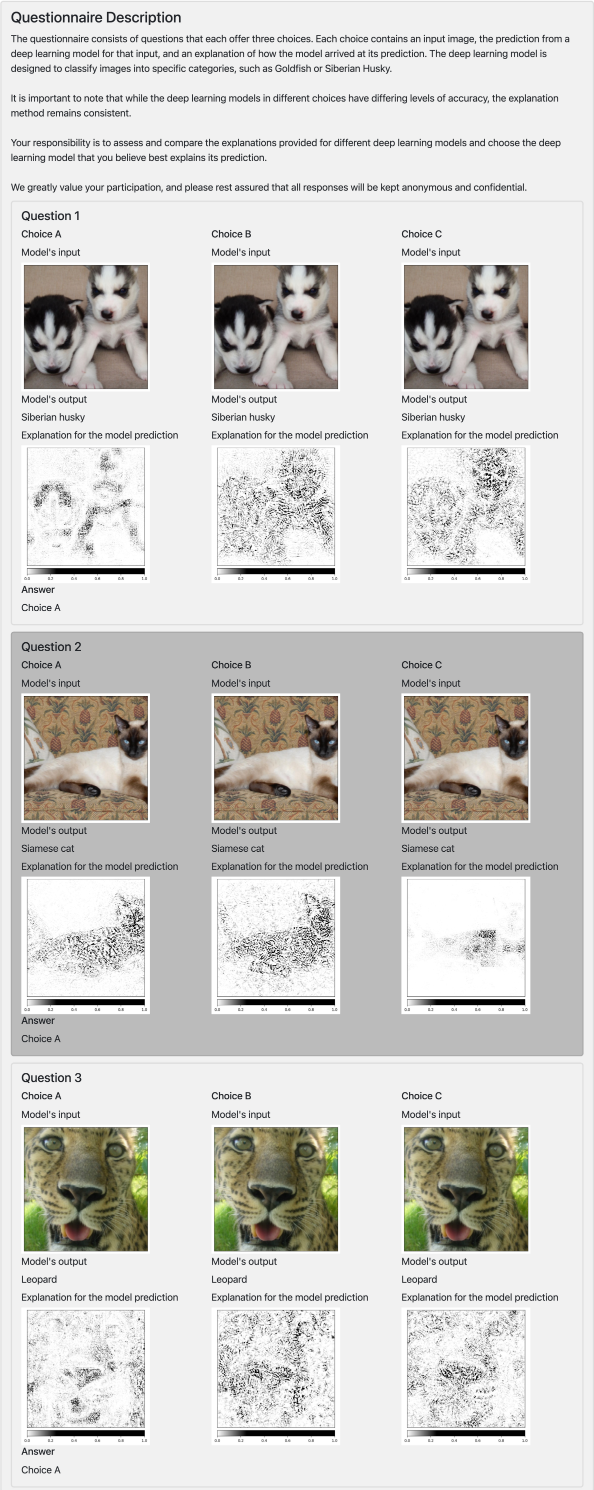
Question 4

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Leopard	Leopard	Leopard
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction
Answer		
Choice A		

Question 5

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Bee	Bee	Bee
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction
Answer		
Choice A		

Figure 9: Objective evaluation questions used for Grad-CAM.



Questionnaire Description

The questionnaire consists of questions that each offer three choices. Each choice contains an input image, the prediction from a deep learning model for that input, and an explanation of how the model arrived at its prediction. The deep learning model is designed to classify images into specific categories, such as Goldfish or Siberian Husky.

It is important to note that while the deep learning models in different choices have differing levels of accuracy, the explanation method remains consistent.

Your responsibility is to assess and compare the explanations provided for different deep learning models and choose the deep learning model that you believe best explains its prediction.

We greatly value your participation, and please rest assured that all responses will be kept anonymous and confidential.

Question 1

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siberian husky	Siberian husky	Siberian husky
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

Question 2

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siamese cat	Siamese cat	Siamese cat
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

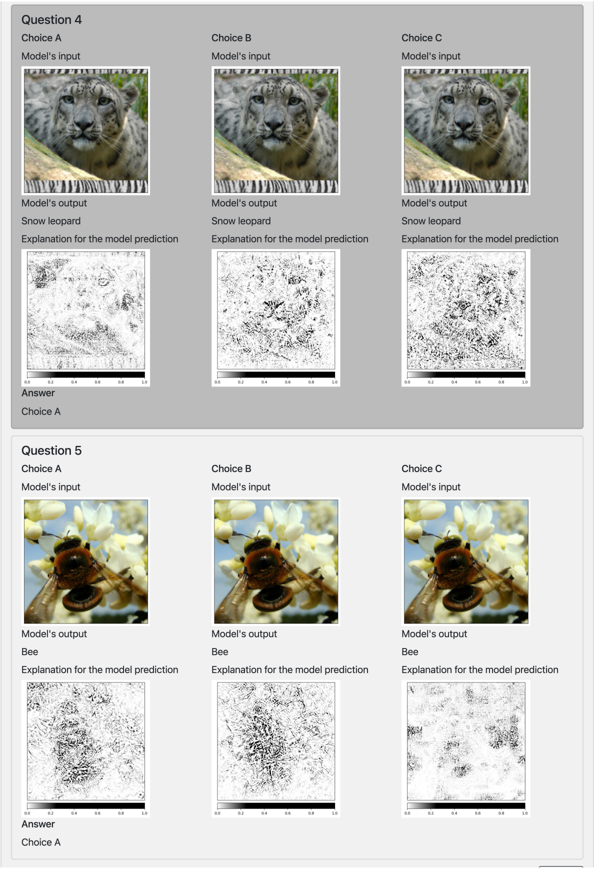
Choice A

Question 3

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Leopard	Leopard	Leopard
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A



Question 4

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Snow leopard	Snow leopard	Snow leopard
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

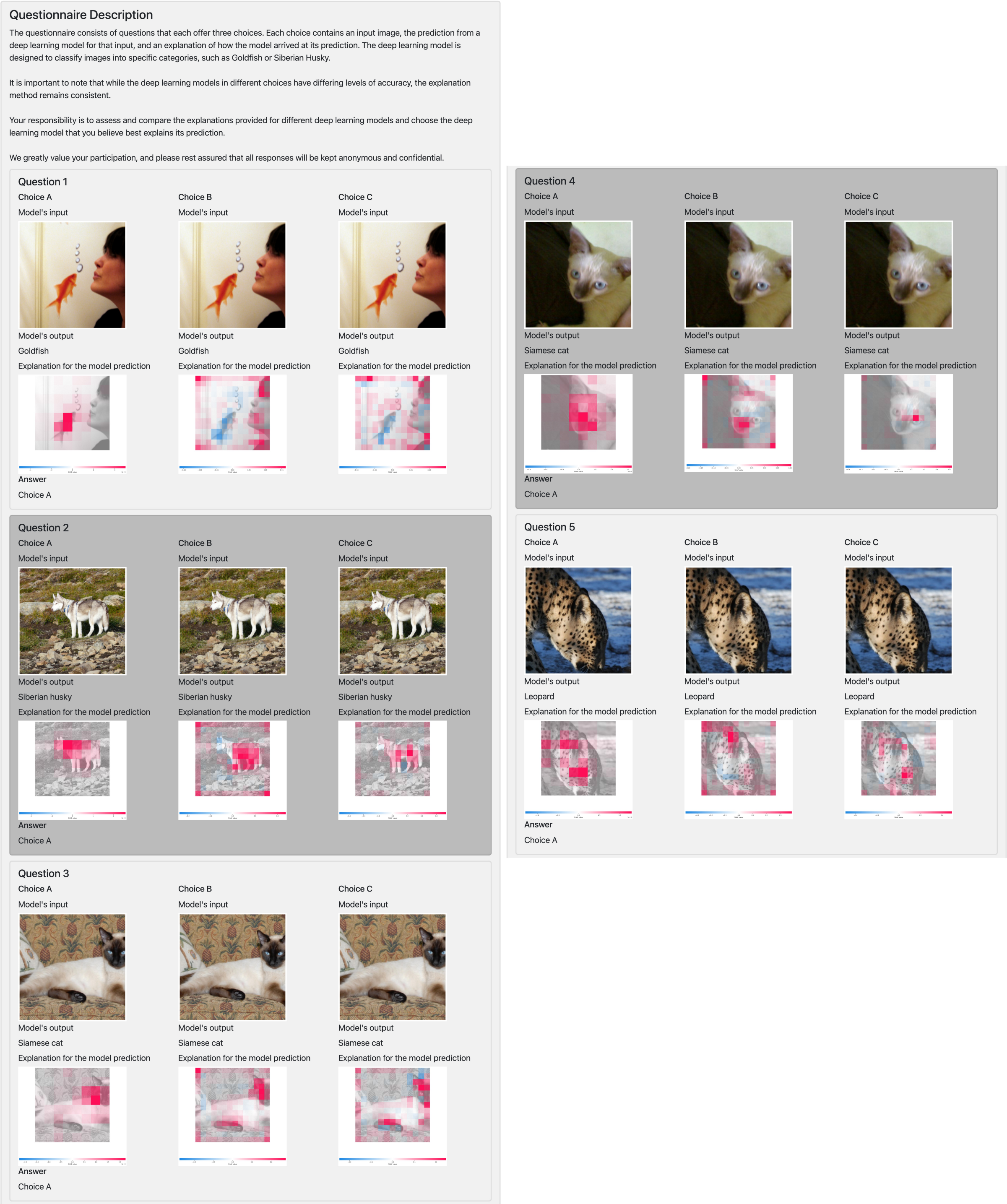
Question 5

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Bee	Bee	Bee
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

Figure 10: Objective evaluation questions used for Integrated Gradients



Questionnaire Description

The questionnaire consists of questions that each offer three choices. Each choice contains an input image, the prediction from a deep learning model for that input, and an explanation of how the model arrived at its prediction. The deep learning model is designed to classify images into specific categories, such as Goldfish or Siberian Husky.

It is important to note that while the deep learning models in different choices have differing levels of accuracy, the explanation method remains consistent.

Your responsibility is to assess and compare the explanations provided for different deep learning models and choose the deep learning model that you believe best explains its prediction.

We greatly value your participation, and please rest assured that all responses will be kept anonymous and confidential.

Question 1

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Goldfish	Goldfish	Goldfish
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

Question 2

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siberian husky	Siberian husky	Siberian husky
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

Question 3

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siamese cat	Siamese cat	Siamese cat
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

Question 4

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Siamese cat	Siamese cat	Siamese cat
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

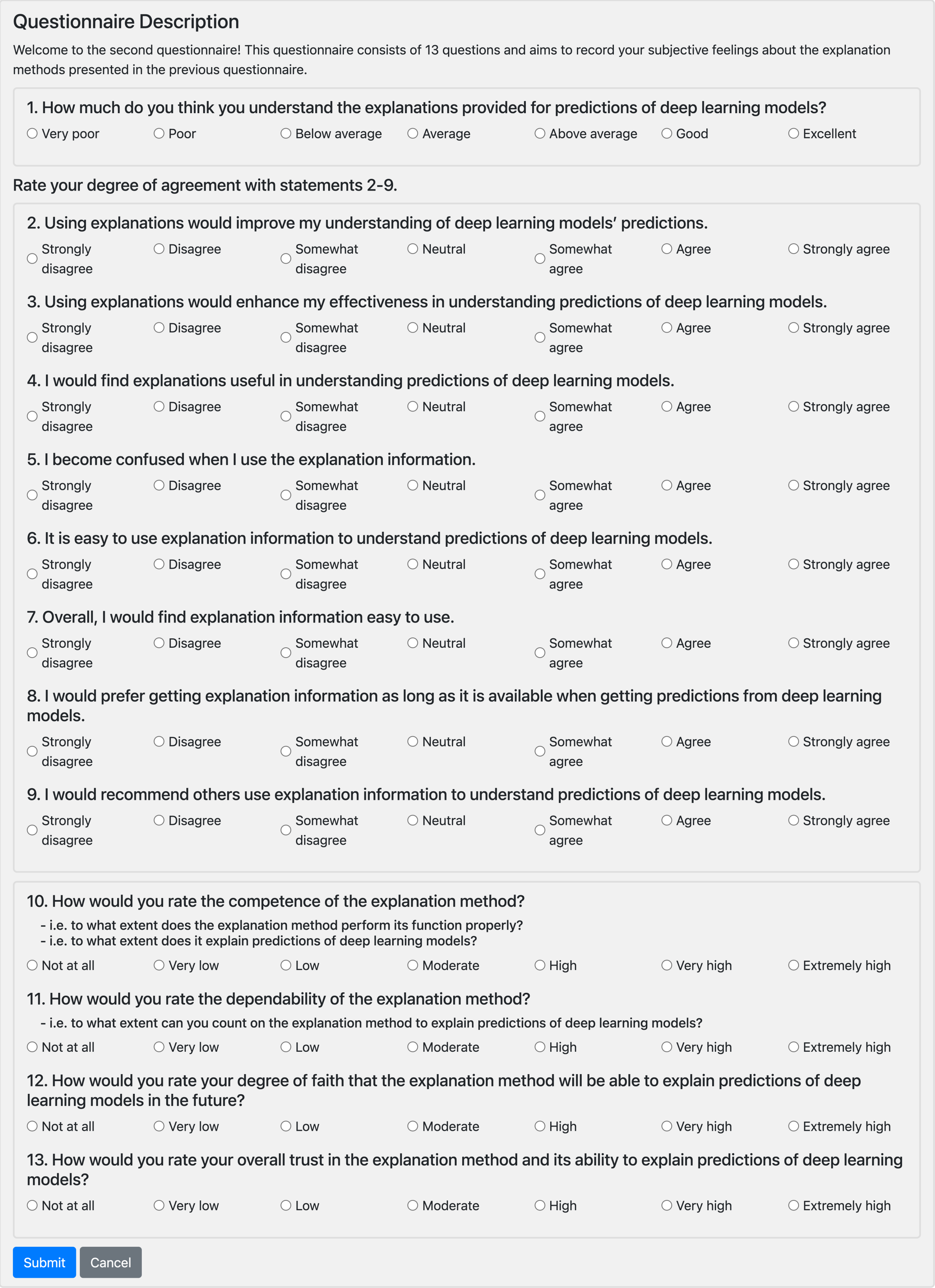
Question 5

Choice A	Choice B	Choice C
Model's input	Model's input	Model's input
Model's output	Model's output	Model's output
Leopard	Leopard	Leopard
Explanation for the model prediction	Explanation for the model prediction	Explanation for the model prediction

Answer

Choice A

Figure 11: Objective evaluation questions used for SHAP.



Questionnaire Description

Welcome to the second questionnaire! This questionnaire consists of 13 questions and aims to record your subjective feelings about the explanation methods presented in the previous questionnaire.

1. How much do you think you understand the explanations provided for predictions of deep learning models?

○ Very poor ○ Poor ○ Below average ○ Average ○ Above average ○ Good ○ Excellent

Rate your degree of agreement with statements 2-9.

2. Using explanations would improve my understanding of deep learning models' predictions.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

3. Using explanations would enhance my effectiveness in understanding predictions of deep learning models.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

4. I would find explanations useful in understanding predictions of deep learning models.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

5. I become confused when I use the explanation information.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

6. It is easy to use explanation information to understand predictions of deep learning models.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

7. Overall, I would find explanation information easy to use.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

8. I would prefer getting explanation information as long as it is available when getting predictions from deep learning models.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

9. I would recommend others use explanation information to understand predictions of deep learning models.

○ Strongly disagree ○ Disagree ○ Somewhat disagree ○ Neutral ○ Somewhat agree ○ Agree ○ Strongly agree

10. How would you rate the competence of the explanation method?

- i.e. to what extent does the explanation method perform its function properly?
- i.e. to what extent does it explain predictions of deep learning models?

○ Not at all ○ Very low ○ Low ○ Moderate ○ High ○ Very high ○ Extremely high

11. How would you rate the dependability of the explanation method?

- i.e. to what extent can you count on the explanation method to explain predictions of deep learning models?

○ Not at all ○ Very low ○ Low ○ Moderate ○ High ○ Very high ○ Extremely high

12. How would you rate your degree of faith that the explanation method will be able to explain predictions of deep learning models in the future?

○ Not at all ○ Very low ○ Low ○ Moderate ○ High ○ Very high ○ Extremely high

13. How would you rate your overall trust in the explanation method and its ability to explain predictions of deep learning models?

○ Not at all ○ Very low ○ Low ○ Moderate ○ High ○ Very high ○ Extremely high

Figure 12: Questions in the subjective evaluation. The user will respond to each question using a 7-point Likert scale.

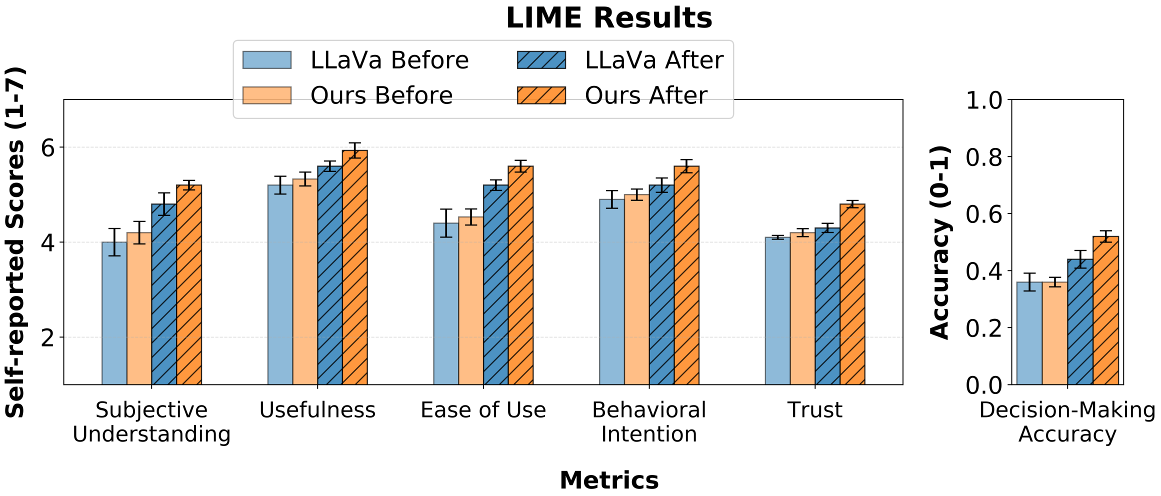


Figure 13: Results of human evaluation of LIME. We report the participants' objective understanding (decision-making accuracy), subjective understanding, perceived usefulness, ease of use, behavioral intention, and trust in static explanations, before and after conversational explanations with LLaVa-1.5 and our model. Decision-making accuracy is ranged from 0 to 1 and the rest scores are from 1 to 7.

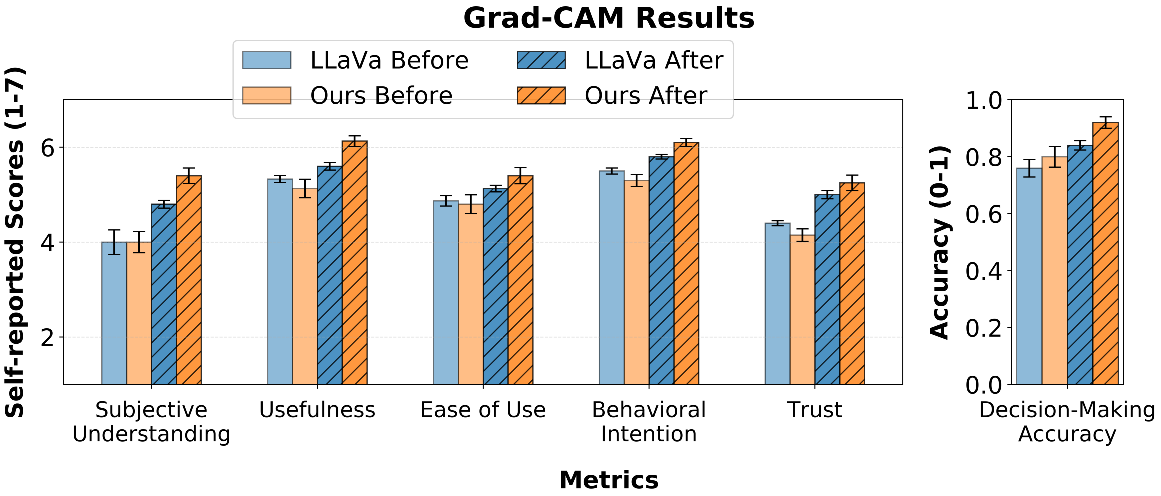


Figure 14: Results of human evaluation of Grad-CAM. We report the participants' objective understanding (decision-making accuracy), subjective understanding, perceived usefulness, ease of use, behavioral intention, and trust in static explanations, before and after conversational explanations with LLaVa-1.5 and our model. Decision-making accuracy is ranged from 0 to 1 and the rest scores are from 1 to 7.

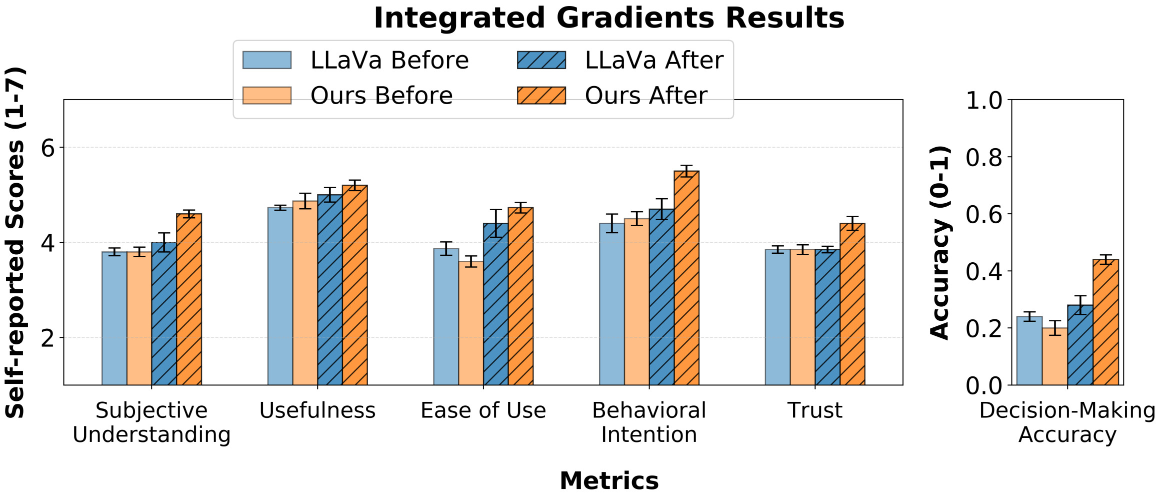


Figure 15: Results of human evaluation of Integrated Gradients. We report the participants' objective understanding (decision-making accuracy), subjective understanding, perceived usefulness, ease of use, behavioral intention, and trust in static explanations, before and after conversational explanations with LLaVa-1.5 and our model. Decision-making accuracy is ranged from 0 to 1 and the rest scores are from 1 to 7.

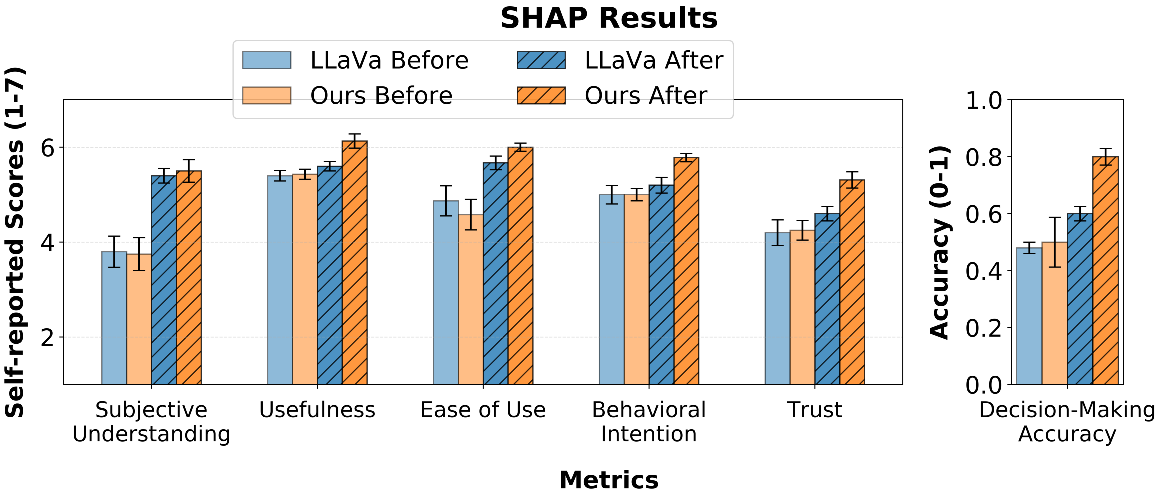


Figure 16: Results of human evaluation of SHAP. We report the participants' objective understanding (decision-making accuracy), subjective understanding, perceived usefulness, ease of use, behavioral intention, and trust in static explanations, before and after conversational explanations with LLaVa-1.5 and our model. Decision-making accuracy is ranged from 0 to 1 and the rest scores are from 1 to 7.