User-Centered Dialogue Systems

Tong Zhang

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

07/08/2024

.....

Date

.....

Tong Zhang

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

07/08/2024

.....

Date

.....

Boyang Li

Authorship Attribution Statement

This thesis contains material from four papers published or under review in the following peer-reviewed journal/conferences in which I am listed as an author.

Chapter 2 is published as Tong Zhang, Yong Liu, Boyang Li, Peixiang Zhong, Chen Zhang, Hao Wang, Chunyan Miao. Toward Knowledge-Enriched Conversational Recommendation System. In the 4th Workshop on NLP for ConvAI. 2022.
The contributions of the co-authors are as follows:

- I designed the methodology, implemented the code, conducted experiments, and prepared the manuscript draft.
- Yong Liu, Peixiang Zhong, and I discussed the initial research direction.
- Boyang Li, Yong Liu, and Peixiang Zhong revised the manuscript.
- Hao Wang, Chunyan Miao, and Chen Zhang offered feedback on the manuscript.

Chapter 3 is published as Tong Zhang, Yong Liu, Boyang Li, Zhiwei Zeng, Pengwei Wang, Yuan You, Chunyan Miao, and Lizhen Cui. History-Aware Hierarchical Transformer for Multi-session Open-domain Dialogue System. In Findings of the Conference on Empirical Methods in Natural Language Processing. 2022.
The contributions of the co-authors are as follows:

- I designed the methodology, implemented the code, conducted experiments, and prepared the manuscript draft.
- Yong Liu, Zhiwei Zeng, and I discussed the initial research direction.
- Boyang Li, Zhiwei Zeng, and Yong Liu revised the manuscript.
- Hao Wang, Chunyan Miao, and Chen Zhang offered feedback on the manuscript.

Chapter 4 is published as Tong Zhang, X. Jessie Yang, Boyang Li. May I Ask a Follow-up Question? Understanding the Benefits of Conversations in Neural Network Explainability. International Journal of Human-Computer Interaction. 2024.
The contributions of the co-authors are as follows:

- I designed the methodology, implemented the code, conducted experiments, and prepared the manuscript draft.

- Boyang Li introduced the research direction, improved the experimental design, and revised the manuscript.
- X. Jessie Yang improved the experimental design and revised the manuscript.

Chapter 5 is submitted as Tong Zhang, Mengao Zhang, X. Jessie Yang, Boyang Li. Handling the Follow-up Question: Conversational Explanations for Image Classification. Submitted to The 2024 Conference on Empirical Methods in Natural Language Processing.
The contributions of the co-authors are as follows:

- I designed the methodology, implemented the code, conducted experiments, and prepared the manuscript draft.
- Mengao Zhang assisted in code implementation.
- Boyang Li introduced the research direction, improved the method, improved the experimental design, and revised the manuscript.
- X. Jessie Yang improved the experimental design and revised the manuscript.

07/08/2024

Tong Zhang

.....

Date

.....

Tong Zhang

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Boyang Li, for all the support he has given me over the years. I could not complete my thesis without his guidance, encouragement, and insightful advice on research. He has significantly improved my writing, presentation skills, and critical thinking abilities. I am so grateful to Prof. Li for having the best influence on this period of my life and my career.

I extend my thanks to the thesis advisory committee and qualifying examination committees: Prof. Shafiq Joty, Prof. Guoan Bi, Prof. Yiping Ke, Prof. Cheng Long, and Prof. Tianwei Zhang for their valuable comments and suggestions on my research.

Since entering NTU, I have been blessed to be surrounded by a group of outstanding mentors and colleagues. I would like to thank Zhiwei Zeng, X. Jessie Yang, Di Wang, Yong Liu, Peixiang Zhong, Yinan Zhang, Zilin Du, Haoxin Li, Miaoyu Li, Mengao Zhang, Anthony Meng Huat Tiong, Devaansh Chandra Gupta, Jiayun Luo, Yidan Sun, Qin Chao, and Junqi Zhao for their support to my research.

There are so many fantastic friends I have made over the last few years - Yuxiao Lu, Yichen Wang, Yongjie Wang, Xu Guo, Chang Liu, Yidan Hu, and Zebang Deng. These people have enriched my life in all kinds of ways.

I would also like to thank my parents and grandparents for their unwavering love and support over the years.

Finally, and most importantly, I want to thank my wife, Yanci Zhang, for her strong support and encouragement as I undertake my Ph.D. I cannot imagine how I would have gotten through my doctoral journey without her support and companionship.

"Every genuine test of a theory is an attempt to falsify it, or to refute it."

—Karl Popper

To my dear family

Abstract

A dialogue system is an intelligent machine that can converse coherently with humans using natural language. Dialogue systems aim to fulfill users' different needs. However, many existing dialogue systems do not consider user needs from user perspectives, which can lead to suboptimal interactions and reduced user satisfaction. In this thesis, we propose to design user-centered dialogue systems that prioritize the needs and preferences of end-users. In this thesis, we identify three foundational aspects of user-centered dialogue systems: Personalization, Continuity, and Support for diverse users.

Personalization refers to the system's ability to adapt its responses to individual users' goals, preferences, and communication styles. It can enhance user engagement, satisfaction, and relevance to dialogue systems. To explore better personalization, we develop a user-centered conversational recommendation system capable of providing personalized item recommendations along with contextual information. We hypothesize that generating naturalistic conversations and offering background details helps users better evaluate the recommendations and thus improve the effectiveness of personalized recommendations. To achieve this, we propose the Bag-of-Entities loss and alignment loss. These two losses encourage generated utterances to mention concepts related to the item being recommended, such as the genre or director of a movie. Experiments on the large-scale REDIAL dataset demonstrate that our model outperforms state-of-the-art baselines by 48% in automatic evaluation and by 19.5% in human evaluation. Human evaluation demonstrates that our model can generate more fluent, relevant, and informative recommendations than baseline models.

Continuity refers to the system's capacity to retain and utilize information from past interactions over time. Human conversations often span multiple sessions and draw on shared context to ensure coherence and relationship building. A user-centered dialogue

system should similarly be able to remember historical interactions to maintain context and build long-term engagement. To this end, we propose the History-Aware Hierarchical Transformer (HAHT) for multi-session open-domain dialogue. HAHT encodes historical conversation sessions into memory and leverages this information to understand current context and generate relevant responses. Experimental results on a large-scale MSC dataset suggest that HAHT consistently outperforms baseline models. Human evaluation results support that our HAHT can better leverage history conversations to generate fluent, history-relevant responses than baseline models.

Support for diverse users involves designing dialogue systems that can effectively interact with people from a wide range of backgrounds, including users with varying levels of domain knowledge, linguistic styles, cognitive abilities, and expectations. This principle is particularly critical when dialogue systems are used to mediate users' understanding of complex or opaque technologies, such as AI models in high-stakes domains. To this end, we build a user-centered dialogue system that can answer users' arbitrary follow-up questions after providing static explanations. Most existing explainable AI methods only provide only one-time, static explanations, which cannot cater to users' diverse knowledge levels and information needs. We first prove that free-form conversations can enhance users' comprehension of static explanations in image classification, improve acceptance and trust in the explanation methods, and facilitate human-AI collaboration. We then build the fEw-shot Multi-round ConvErsational Explanation (EMCEE) system. We train EMCEE with synthetic data and mitigate two main challenges of training models with synthetic data: lack of data diversity and hallucination in the generated data. 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.

Contents

List of Figures

List of Tables

List of Publications

- Tong Zhang, Yong Liu, Boyang Li, Peixiang Zhong, Chen Zhang, Hao Wang, Chunyan Miao. "Toward Knowledge-Enriched Conversational Recommendation Systems". The 4th Workshop on NLP for Conversational AI, 2022.
- Tong Zhang, Yong Liu, Boyang Li, Zhiwei Zeng, Pengwei Wang, Yuan You, Chunyan Miao, and Lizhen Cui. "History-Aware Hierarchical Transformer for Multi-session Open-domain Dialogue System". Findings of the Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP). 2022.
- Tong Zhang, X. Jessie Yang, Boyang Li. "May I Ask a Follow-up Question? Understanding the Benefits of Conversations in Neural Network Explainability". International Journal of Human-Computer Interaction. 2024.
- Tong Zhang, Mengao Zhang, Wei Yan Low, X. Jessie Yang, Boyang Li. "Handling the Follow-up Question: Conversational Explanations for Image Classification". Accepted to the ACM Conference on Intelligent User Interfaces (ACM IUI). 2025

Chapter 1

Introduction

Language is the cornerstone of human intelligence. Conversations or dialogues are one of the most prevalent uses of human language, in which people exchange knowledge, socialize, and sustain their relationship with each other [2, 3].

A dialogue system, also known as a conversational agent, is an intelligent machine that can converse coherently and engagingly with humans using natural language. Building dialogue systems has been one of the fundamental objectives of artificial intelligence (AI) research [4, 5]. Early dialogue systems are mainly designed based on hand-built rules and templates [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. For instance, ELIZA [6], the first dialogue system developed in the 60's, simulates a Rogerian psychotherapist by hand-crafted pattern matching and substitution rules. These rule-based and template-based methods can only handle a pre-defined range of queries and provide fixed responses based on rules or templates. In recent years, with the rapid development of deep learning methods and the availability of big conversation corpus, dialogue systems based on deep neural networks [15, 16, 17, 18] are emerging and have become the dominant approach in dialogue systems. These deep learning-based dialogue systems have been shown to generate flexible, human-like conversations across a wide range of topics [19].

Despite their progress, many existing dialogue systems are not designed with the user at the center. They often prioritize short-term objectives, such as completing a task or producing locally relevant responses. They usually do not take into account who the user is, what the user knows, or how the user has interacted with the system over time. As a result, these systems tend to be rigid, shallow, and difficult to engage with over extended or evolving interactions.

To provide more effective and user-friendly dialogue systems, it is vital to design dialogue systems in a user-centered way [3]. The user-centered design requires designers to build dialogue systems that prioritize the needs and preferences of end-users. In this thesis, we identify three foundational aspects of user-centered dialogue systems:

- Personalization: the system's ability to adapt its responses to individual users' goals, preferences, and styles. Personalization increases user engagement, satisfaction, and relevance—especially in scenarios where one-size-fits-all responses are insufficient [20]. A user-centered dialogue system should actively learn from user inputs and tailor responses accordingly.

- Continuity: the ability to remember and utilize prior interactions over time. Without memory, chatbots often fail to re-engage users meaningfully when previously discussed topics reemerge [21]. The apparent forgetfulness limits the chatbots' ability to establish and maintain long-term relationships with users.

- Support for diverse users: this aspect requires dialogue systems to be accessible and adaptable to individuals with different levels of expertise, language proficiency, and goals. Many users, especially non-experts, struggle to understand complex scientific concepts without assistance [22, 23]. User-centered systems should offer explanations and adapt responses to meet varied user needs.

By designing from these perspectives, I aim to build dialogue systems that enhance user satisfaction and promote sustained interaction. In this thesis, I explore how these three principles can be operationalized through the design of dialogue systems in three main scenarios of using dialogue systems.

For improving the **personalization** of dialogue systems, this thesis studies the conversational recommendation task, where users receive item suggestions through natural language interaction. While existing systems can generate personalized recommendations, their responses often lack specific information about the recommended items, such as actors or plots in a movie recommendation. This lack of informativeness limits users' ability to assess whether the recommendations meet their needs, and thus undermines the effectiveness of personalized recommendations. To address this, we propose leveraging knowledge graphs to generate more informative responses. Specifically, we introduce

a novel Bag-of-Entities loss that encourages the model to include related concepts in the responses. We also propose an alignment loss to better integrate knowledge graph entities into the generation process. Experiments on the large-scale REDIAL dataset show that our system consistently outperforms state-of-the-art baselines. Human evaluation further demonstrates that our model generates more relevant and informative recommendations.

To improve the **continuity** of dialogue systems, this thesis explores the multi-session conversation task, where dialogue systems engage with users across multiple sessions over time. To handle long-term memory, we propose the History-Aware Hierarchical Transformer (HAHT) for multi-session open-domain dialogue. HAHT maintains a long-term memory of previous sessions and uses this information to better understand the current context and generate more contextually relevant responses. It first encodes historical sessions hierarchically into a history memory, then incorporates this memory into the current context using attention-based mechanisms. Finally, HAHT uses a history-aware decoder that can generate words from either a generic vocabulary or a history-aware vocabulary. Experimental results on a large-scale multi-session dataset show that HAHT outperforms strong baselines. Human evaluation confirms that HAHT produces more coherent, context-relevant, and history-aware responses.

To provide **support for diverse users**, this thesis focuses on conversational explanations. Research in explainable AI (XAI) aims to provide insights into the decision-making process of opaque AI models. To date, most XAI approaches provide only one-time, static explanations, which cannot cater to users' diverse knowledge levels and information needs. To explore whether conversation can improve explanation, we conduct a user study involving 120 participants. Half interact with a human expert in a conversational setting, while the other half read static explanations. We measure objective and subjective comprehension, trust, and acceptance. Results show that conversation significantly improves users' understanding, trust, and acceptance of explanations, while static explanations alone do not.

After proving conversational explanations as an effective method to customize XAI explanations, we developed the fEw-shot Multi-round ConvErsational Explanation (EMCEE) system. Due to the scarcity of training data, we propose to train a vision language

model with synthetic data. However, training with synthetic data faces two main challenges: lack of data diversity and hallucination in the generated data. To address these, we introduce a repetition penalty to improve diversity and a hallucination detector to filter out untruthful responses .EMCEE achieves significant improvements over baseline models—81.6% in BLEU and 80.5% in ROUGE—and maintains high-quality conversations even after multiple training rounds. Human evaluations show that EMCEE substantially improves users' understanding, trust, and collaboration with AI systems. To the best of our knowledge, EMCEE is the first system capable of answering arbitrary user questions following static explanations.

1.1 Contribution

In this thesis, we study to design dialogue systems in a user-centered way. We develop new dialogue systems from user's perspectives in three different conversational tasks. Our contributions are summarized as follows:

- **Knowledge-Enriched Conversational Recommendation Systems (KECRS)**: We propose a novel dialogue system named KECRS, which employs the Bag-of-Entity (BOE) loss and alignment loss to effectively integrate KG with CRS for generating more diverse and informative responses. BOE loss encourages the generated utterances to mention concepts related to the item being recommended. The alignment loss integrates KG entities into the response generation network. Experiments demonstrate that the proposed two losses improve model performance. KECRS consistently outperforms state-of-the-art CRSs on the large-scale REDIAL dataset.

- **History-Aware Hierarchical Transformer (HAHT) for multi-session open-domain dialogue**: we propose a novel dialogue system HAHT, which can effectively leverage history conversations to conduct more engaging multi-session conversations. HAHT maintains a long-term memory to store historical conversational contexts, which is updated when a new session is conducted. Based on the long-term memory and the context in the current session, relevant tokens

in historical contexts are selected to adapt the current response. We show that HAHT outperforms baseline models in various evaluation metrics. Human evaluation results support that HAHT generates more readable, context-relevant, and history-relevant responses than baseline models.

- **Understanding the Benefits of Conversations in Neural Network Explainability**: We demonstrate that free-form conversations can significantly enhance users' comprehension of static explanations in image classification, improve acceptance and trust in the explanation methods, and facilitate human-AI collaboration. We are the first to study how free-form conversations may facilitate neural network explainability in a computer vision task. Our findings highlight the importance of customized model explanations in the format of free-form conversations and provide insights for the future design of conversational explanations.

- **fEw-shot Multi-round ConvErsational Explanation (EMCEE)**: We propose the first conversational explanation EMCEE 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.

1.2 Thesis Outline

Figure 1.1 depicts a high-level organization of the thesis. The detailed structure of this thesis is organized as follows:

Chapter 1 introduces the background, scope, and contribution of the thesis. Chapter 2 introduces the Knowledge-Enriched Conversational Recommendation Systems (KECRS) that provide diverse and informative responses to users.

Chapter 3 introduces the History-Aware Hierarchical Transformer (HAHT) for multi-session open-domain dialogue that can generate responses based on users' history conversations.

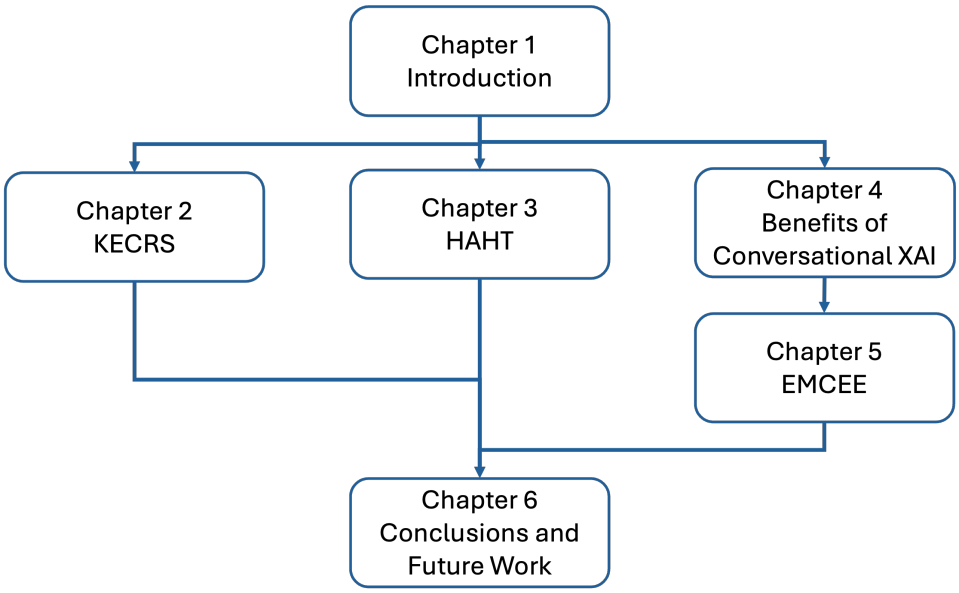


Figure 1.1: Thesis organization. KECRS represents the Knowledge-Enriched Conversational Recommendation System. HAHT represents the History-Aware Hierarchical Transformer. EMCEE represents the fEw-shot Multi-round ConvErsational Explanation.

Chapter 4 introduces how free-form conversations can significantly enhance users' comprehension of static explanations in image classification, improve acceptance and trust in the explanation methods, and facilitate human-AI collaboration.

Based on the findings of Chapter 4, Chapter 5 designs the conversational explanation system, EMCEE that can answer users' free-form follow-up questions after providing static explanations to the user.

Chapter 6 concludes this thesis and discusses future research directions.

Chapter 2

Knowledge-Enriched Conversational Recommendation Systems

2.1 Overview

Conversational recommendation systems (CRS) have received increasing attention from the Natural Language Processing community [24, 25, 26, 27, 28, 29, 30]. CRS aims to recommend items, such as movies or songs, in naturalistic interactive conversations with the user. This interactive form allows the system to provide recommendations tailored to the preferences provided by the user at the moment.

A crucial issue of CRS is to extract user preferences from the conversation, which often requires background information provided by knowledge graphs (KGs). As an example, in Figure 2.1, the user mentions two movies that belong to the horror genre. To this end, some existing studies [25, 26] leverage knowledge graphs to understand user intentions. However, when recommending items to users, existing methods usually only mention items name and lack of relative information about the recommended items. This lack of informativeness limits users' ability to assess whether the recommendations meet their needs, which affects the effectiveness of personalized recommendations.

We observe that when humans recommend items to friends, they usually describe attributes of the item. It would be more convincing and improve user's receptivity. For example, to recommend a movie, they may mention the director or actors. Such information can be easily extracted from the knowledge graph but has not been well utilized by existing approaches.

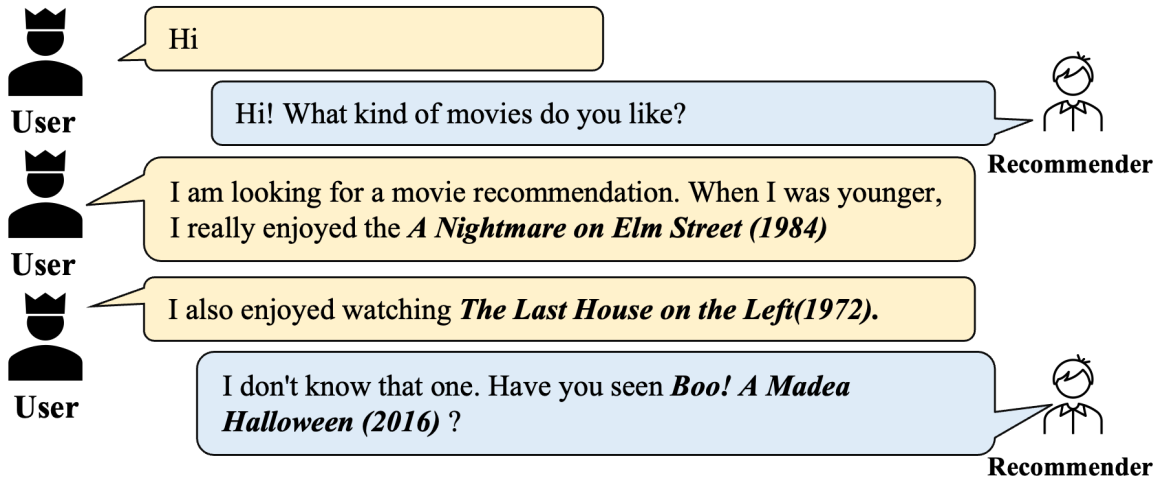


Figure 2.1: An example of a conversation between a user and the Chatbot for movie recommendation.

To emulate naturalistic conversations and provide convincing recommendations, we propose the Knowledge-Enriched Conversational Recommendation System (KECRS). Specifically, we propose the Bag-of-Entity (BOE) loss to encourage the generated utterances to mention concepts related to the item. Moreover, we propose an alignment loss that ties the word embeddings to the entity embeddings.

Experiments demonstrate that the proposed two losses improve model performance. The proposed KECRS consistently outperforms state-of-the-art CRSs on the large-scale REDIAL dataset [24].

2.2 Related work

With the developments of dialogue systems [31, 32, 33] tasks, conversational recommendation systems become an appealing solution to capture users' dynamic preferences. One group of existing works focus on the attribute-based conversational recommendation systems [34, 35, 36, 37, 38, 39, 40, 41]. They aim to provide high-quality recommendations within the shortest number of conversation turns but do not pay much attention to generate human-like responses. Thus, the conversation utterances are mostly simulated.

Another group of existing works studies the chit-chat-based conversational recommendation systems [24, 25, 26, 27, 28, 29, 30]. Most studies in this category focus on both

giving accurate recommendations and generating natural and human-like responses. Li et al. [24] release a conversational recommendation dataset in the movie domain and propose an HRED-based [42] baseline model. As it is hard to understand user's intentions only from utterances, Chen et al. [25] introduce KG into Chit-chat-based CRS and propose a knowledge-based recommender dialog system. Based on [25], Sarkar et al. [27] explore different sizes of KGs in the recommendation module. To better understand the user's preferences, Zhou et al. [26] leverage the entity-oriented KG (*i.e.,* DBpedia) and the word-oriented KG (*i.e.,* ConceptNet). To make the recommendation proactively and naturally, Liu et al. [28] and Zhou et al. [29] use topics to guide dialogue from non-recommendation to recommendation and propose a new topic-guided conversational recommendation task. Finally, research on conversational characters for e-commerce has the broad goal of building a complete shopping assistant that can answer a variety of questions in addition to recommendation [43, 44, 45, 46].

After the work of this chapter is developed and published, large language models (LLMs) began to dominate natural language processing tasks. LLMs have demonstrated strong capabilities in understanding item semantics from textual descriptions and modeling user-item interactions based on historical data [47, 48]. Motivated by these advances, many recent approaches propose leveraging LLMs to jointly model conversation and recommendation tasks within a unified framework. UniCRS [49] adopts DialoGPT [50] with task-specific soft prompts [51], but requires a three-stage optimization pipeline: semantic fusion pre-training, conversation tuning, and recommendation tuning. UniMIND [52] extends this paradigm with BART [53], unifying multiple goals through prompt-based multi-stage training. RecInDial [54] augments DialoGPT's vocabulary with item tokens and introduces a pointer mechanism to jointly predict words and items. MESE [55] replaces knowledge graphs with item metadata, incorporating it into the dialogue context and training GPT-2 [56] for end-to-end learning. Although these models strive to unify conversation and recommendation, they often depend on auxiliary modules (e.g., R-GCN [57], DistilBERT [58]) and involve complex, multi-stage training. In contrast, PECRS [59] formulates the CRS task as a single-stage natural language processing problem by integrating item descriptions and dialogue context directly.

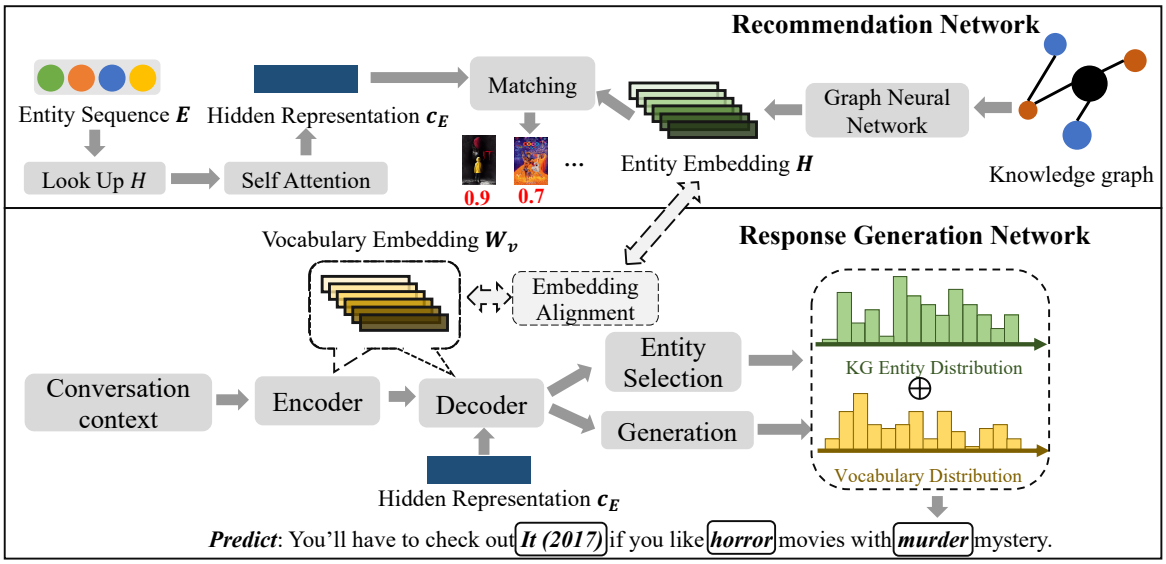


Figure 2.2: The overall framework of the proposed KECRS model.

2.3 Approach

The overall goal of a conversational recommendation system is to identify an item (e.g., a movie, a song, or a piece of merchandise) that the user will likely interact with and suggest the item to the user in the form of natural language conversations.

Formally, we represent the historic conversation $X = \langle x_1, x_2, ..., x_n \rangle$ as a sequence of n utterances x_i. The knowledge graph $\mathcal{G} = \{(v_h, r, v_t)\}$ is a set of entities E and a set of relationships r between the head entity $v_h \in E$ and the tail entity $v_t \in E$.

The conversational recommendation task is to predict the next utterance x_{n+1} using the recommendation network $f(X, \mathcal{G})$ and the response generation network $g(X, \mathcal{G}, f(X, \mathcal{G}))$. $f(X, \mathcal{G})$ predicts the next item to recommend to the user, whereas $g(X, \mathcal{G}, f(X, \mathcal{G}))$ predicts the utterance x_{n+1} one word at a time.

Figure 2.2 shows the overall structure of our proposed method, the Knowledge-Enriched Conversational Recommendation System (KECRS).

2.3.1 Recommendation Network

First, we exhaustively match each word in the conversational history X with the name of each entity in the KG. In this way, we identify K entities from the history and sequence them according to their original positions. Next, we apply a graph convolutional

network, R-GCN [57] to encode the entire KG and obtain embeddings for each KG entity node. The D-dimensional entity embeddings of the K entity form the matrix $\boldsymbol{H}_E \in \mathbb{R}^{K \times D}$. Subsequently, we apply an attention operation where the attention vector $\boldsymbol{\alpha}$ is computed by 2 fully connected (FC) layers.

$$\begin{aligned} \boldsymbol{\alpha} &= \text{softmax}\big(\mathbf{W}_k \tanh(\mathbf{W}_q \mathbf{H}_E^\top)\big), \\ \mathbf{c}_E &= \boldsymbol{\alpha} \mathbf{H}_E, \end{aligned} \tag{2.1}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable parameters. The resulting $\mathbf{c}_E \in \mathbb{R}^D$ is a condensed representation of entities appearing in the conversational history.

The recommendation module classifies $\boldsymbol{c}_E$ directly into one of the items. We directly take the entity embedding $\boldsymbol{e}_i$ from the R-GCN network as the representation of the item. The probability of recommending item i is computed with softmax:

$$P_{rec}(i) \propto \exp(\boldsymbol{c}_E^\top \boldsymbol{e}_i). \tag{2.2}$$

The module is trained using the cross-entropy loss. To avoid the model recommending the same movie that the user might have just mentioned, we only consider as a ground-truth recommendation the movie that is first time to be mentioned by the recommender in the conversation.

2.3.2 Response Generation Network

The response generation module predicts the utterance to the user word by word. We use the classic encoder-decoder Transformer architecture [60], where the encoder encodes the entire conversational history word by word.

At decoding time step j, the output of the Transformer decoder $\boldsymbol{s}_j$ is concatenated with the entity representation $\boldsymbol{c}_E$ and goes through two fully connected layers before the softmax function. The probability distribution over the vocabulary is

$$P_{res} = \text{softmax}\big(\mathbf{W}_v \boldsymbol{W}_a [\mathbf{s}_j; \mathbf{c}_E] + \boldsymbol{b}\big), \tag{2.3}$$

where $\mathbf{W}_v$ is the word embedding matrix shared with the encoder. $\mathbf{W}_a$ is a trainable linear projection to align the dimensions, and $\boldsymbol{b}$ is the bias. We train the module using cross-entropy at every decoder time step.

To separate movie names from other words in the conversation, for every movie name we create specialized tokens in the vocabulary. For example, the token for the movie name *It* is separate from the word token *it*. This is feasible as the dataset, REDIAL, has explicitly represented movie names with special strings.

2.3.3 Bag-of-Entities Loss

Although the response generation module trained using per-step cross-entropy is capable of recommending items, it rarely mentions concepts related to the recommended item. We postulate that mentioning related entities will produce natural conversations. For example, when recommending the movie *It*, one may want to mention that it is a horror movie based on a book by Stephen King.

For this purpose, we introduce the Bag-of-Entity (BOE) loss, which encourages the decoder state $[\mathbf{s}_j; \mathbf{c}_E]$ to contain additional information about first-order neighbors of the ground-truth recommendation on the KG.

First, at every time step, we compute a score $\boldsymbol{r}_j \in \mathbb{R}^M$ for all M entities in the knowledge graph,

$$\boldsymbol{r}_j = \mathbf{H}\mathbf{W}_b[\mathbf{s}_j; \mathbf{c}_E] + \mathbf{b}_{ent}, \tag{2.4}$$

where $\mathbf{H}$ contains the embeddings of all KG entities, as produced by the R-GCN. $\mathbf{W}_b$ is a trainable matrix for dimension alignment and $\mathbf{b}_{ent}$ the bias.

As we do not constrain exactly which word in the response should contain the information, we sum up the word-level scores and then apply the component-wise sigmoid function. The probability that entity m is mentioned in the response is thus

$$P_{\text{BOE}}(m) = \text{sigmoid}(\sum_{j=1}^{L} r_{jm}), \tag{2.5}$$

where L is the length of the response and r_{jm} is the m^{th} component of $\boldsymbol{r}_j$.

We apply a binary cross-entropy loss for each KG entity. The ground-truth label is 1 if the entity is a first-order neighbor of the recommended item on the knowledge graph and 0 otherwise.

2.3.4 Aligning Word and Entity Embeddings

We create two types of tokens in the vocabulary V of the response generation network. The first type corresponds to a plain word appearing in the conversation text. The second type represents an entity that appears in the conversation and in the knowledge graph.

To tie the token embeddings of the second type to the R-GCN encoding of the knowledge graph, we propose the alignment loss. For a conversation, we use the entity representation $\mathbf{c}_E$ in Eq. (2.1) to represent all entities in the conversation and calculate the similarity score between $\mathbf{c}_E$ and each word embedding,

$$\mathbf{s} = \mathbf{W}_{v[E]}\mathbf{W}_c\mathbf{c}_E + \mathbf{b}_{align}, \tag{2.6}$$

where $\mathbf{W}_{v[E]}$ is the matrix resulting from selecting the rows of $\mathbf{W}_v$ corresponding to entity tokens only. $\mathbf{W}_c$ is a trainable matrix and $\mathbf{b}_{align}$ is the bias. The alignment loss is the mean square error between the $\mathbf{s}$ and an indicator vector $\mathbf{q} \in \{0,1\}^{|E|}$.

$$L_{align} = \|\boldsymbol{s} - \boldsymbol{q}\|^2 \tag{2.7}$$

Specifically, if an entity e exists in the conversation, the corresponding component of $\mathbf{q}$ is set to 1. Otherwise, the component is 0.

Finally, to learn the parameters of generation module, we minimize the following objective function:

$$L_{total} = L_{gen} + \lambda_1 L_{\text{BOE}} + \lambda_2 L_{align}, \tag{2.8}$$

where λ_1 and λ_2 are two hyperparameters. In the testing procedure, the probability distribution over the vocabulary at time step j is calculated as follows,

$$P_{all} = P_{res} + \lambda_3 P_{boe}, \tag{2.9}$$

where λ_3 is a hyperparameter.

Model	KG	Recall@K (%)			Precision@K (%)			NDCG@K (%)	
		$K=1$	$K=5$	$K=10$	$K=1$	$K=5$	$K=10$	$K=5$	$K=10$
HRED-CRS	-	0.41	7.04	11.69	0.55	1.82	1.51	3.73	5.25
KBRD	DBpedia	1.40	8.52	13.54	2.03	2.14	1.78	5.06	6.89
KBRD+	DBpedia	1.40	8.73	13.53	1.75	2.11	1.85	5.03	6.94
KGSF	DBpedia & ConceptNet	1.91	9.19	13.42	1.96	1.90	1.49	5.64	7.38
KECRS	DBpedia	1.52	8.69	13.63	2.11	2.19	1.78	5.11	6.91
KBRD	TMDKG	2.15	9.24	14.87	2.33	2.24	1.82	5.26	7.54
KBRD+	TMDKG	2.09	9.31	14.93	2.29	2.31	1.83	5.19	7.61
KGSF	TMDKG & ConceptNet	2.13	7.82	13.68	2.13	1.56	1.35	5.01	6.82
KECRS	TMDKG	**2.25***	**9.45***	**15.66***	**2.77***	**2.39***	**1.99***	**5.74***	**7.92***

Table 2.1: Recommendation performances of different methods based on different knowledge graphs. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*.

2.4 Experiments

2.4.1 Dataset

We use the REDIAL dataset [24], which includes 10,006 conversations and 182,150 utterances related to 51,699 movies. Following [24, 25, 26], we split REDIAL into training, validation, and testing sets with the ratio 8:1:1.

For the knowledge graphs (KGs), we experiment with all models on two different KGs, *i.e.,* DBpedia [61] and TMDKG. The DBpedia is the same as the one used in previous work [25, 27, 26]. We build the TMDKG, from The Movie Database[1], which contains 15822 entities and 15 types of relations.

2.4.2 Evaluation Metrics

We use precision, recall, and normalized discounted cumulative gain (NDCG) to evaluate the top-K item recommendation performance (respectively denoted by Recall@K, Precision@K, and NDCG@K). In the experiments, K is empirically set to 1, 5, and 10 as users may not want to be recommended too many movies in each conversation turn.

Following [25, 26], we use Distinct n-gram (n=2, 3, 4) to measure the diversity of generated responses. To better evaluate the performance of generated responses, we

[1]https://www.themoviedb.org/

Model	Automatic			Human		
	Dist-2	Dist-3	Dist-4	Fluency	Relevancy	Informativeness
HRED-CRS	0.10	0.18	0.24	1.92	1.62	1.05
Transformer	0.15	0.31	0.46	2.03	1.73	1.36
KBRD	0.31	0.38	0.52	2.10	1.72	1.32
KGSF	0.38	0.61	0.73	2.32	2.11	1.56
KECRS(Ours)	**0.48***	**0.91***	**1.23***	**2.56***	**2.29***	**2.18***

Table 2.2: Automatic and human evaluation results of the response generation achieved by different methods. Human evaluation scores are from 0-3. Dist-2,3,4 is short for Distinct-2,3,4. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*

adopt human evaluation. We randomly sample 100 multi-turn conversations from the test set and invite three annotators to score responses generated by different models from the following aspects: 1) **Fluency**: whether responses are fluent;2) **Relevancy**: whether responses are correlated with contexts;3) **Informativeness**: whether responses contain rich information of recommended items. Each aspect is rated in $[0, 3]$, and the final scores are the average of all annotators. For all evaluation metrics, the higher value indicates better performance.

2.4.3 Baseline Methods

We compare KECRS with the following baseline methods: 1) **HRED-CRS** [24]: This is a basic CRS based on HRED[42]; 2) **Transformer** [60]: This is a basic transformer model that generates responses only from utterance text and does not contain a separate recommendation module; 3) **KBRD** [25]: This is a knowledge-based CRS that employs DBpedia to understand the user's intentions and leverage KG information as a bias for generation; 4) **KGSF** [26]: This method exploits both entity-oriented and word-oriented KGs to enrich the data representations. It adopts two KG-enriched decoder layers for the generation.

2.4.4 Recommendation Performance

Table 2.1 shows the recommendation performance achieved by different methods. As shown in Table 2.1, KBRD, KBRD+, KGSF, and KECRS outperform HRED-CRS by

introducing external KGs to understand user's intentions. When using DBpedia, KGSF performs best among all models in most metrics, because it leverages two different KGs (*i.e.,* ConceptNet and DBpedia). The proposed KECRS model outperforms all other baselines that only use one KG. When using TMDKG, we note that the recommendation performance of all models excluding KGSF is improved, and our model outperforms all baseline models including KGSF. This may be because TMDKG is built in the movie domain while DBpedia here is a subgraph of an open-domain KG. KGSF does not perform well in TMKDG. The potential reason is that KGSF cannot well integrate the information of TMDKG and ConceptNet. In summary, the results under two different KGs indicate that KECRS can perform well in the recommendation task, based on the knowledge information from the external KG.

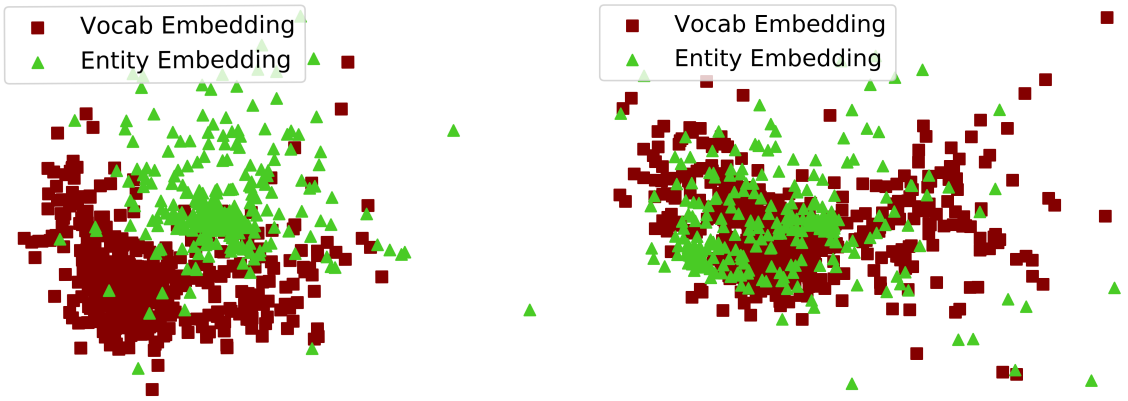
Note that the conversational recommendation problem studied in this chapter is under the cold-start setting [62], where the user's historical behaviors on items (*i.e.,* movies) are unavailable. Thus, it is very challenging to learn user preferences on items. The recommendation accuracy achieved by all models (including our KECRS) is relatively low *e.g.,* KECRS only achieves 2.25% for Recall@1.

2.4.5 Response Generation Performance

The automatic and human evaluation results of different methods are shown in Table 2.2. We note that Transformer performs better than HRED-CRS, which demonstrates that Transformer is powerful to understand and generate natural language. KBRD performs better than Transformer, because it adds a vocabulary bias to fuse knowledge from KG into the generated responses. Among all the baseline models, KGSF generates the most diverse responses, by exploiting both TMDKG and ConceptNet [63]. The potential reason is that KGSF employs two additional KG-based attention layers to make the generative model focus more on items and relevant entities in TMDKG and ConceptNet. Moreover, the proposed KECRS model outperforms all baseline methods with a large margin in terms of all evaluation metrics. This demonstrates that the proposed BOE loss and alignment loss can work jointly to better leverage KG and generate more diverse and informative responses.

Model	Dist-2	Dist-3	Dist-4
KGSF	0.38	0.61	0.73
$\text{KECRS}_{\text{w/o BOE}}$	0.31	0.64	0.87
$\text{KECRS}_{\text{w/o align}}$	0.36	0.69	0.95
KECRS	**0.48***	**0.91***	**1.23***

Table 2.3: Response generation performances of KGSF and different variants of KECRS. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*



(a) KECRS w/o infusion loss (b) KECRS w infusion loss

Figure 2.3: 2D plot of word embeddings in response module and entity embeddings in recommendation module after PCA. Red points represent word embeddings and green points represent entity embeddings.

For human evaluation, we note that *Fluency* is relatively higher compared to *Informativeness* and *Relevancy* for all models. This indicates that responses generated by these models are fluent and can be understood by human judges. However, responses generated by baseline models are more likely to be generic responses (*e.g.,* "I haven't seen that one"). By including additional supervision signals and aligning embeddings of word and entities, the proposed KECRS model alleviates this issue. Overall, KECRS can understand the dialogue context and generate fluent, relevant, and informative responses.

2.4.6 Ablation Study

To better understand the effectiveness of each component in KECRS, we study the performances of the following two variants of KECRS: 1) $\mathbf{KECRS_{w/o\ BOE}}$, which removes the BOE loss, and 2) $\mathbf{KECRS_{w/o\ align}}$, which removes the infusion loss.

Table 2.3 summarizes the response generation performance in terms of Distinct n-gram (n=2,3,4). Distinct n-gram measures the diversity of sentences by calculating the number of distinct n-gram in generated responses. KECRS outperforms $\text{KECRS}_{\text{w/o BOE}}$, which indicates the proposed BOE loss can help the model learn to generate responses not only from conversations but also from the knowledge graph. Moreover, $\text{KECRS}_{\text{w/o align}}$ performs poorer than KECRS. This indicates that aligning the word embeddings and entity embeddings also helps improve the model performances. Compared with KGSF, both ablated versions of KECRS can achieve better performances in terms of most metrics. This again demonstrates that encouraging models to mention concepts related to the recommended items and aligning word embeddings with KG entity embeddings both can help models generate more diverse responses.

To further study the effect of the infusion loss, we draw the 2D plots of word embeddings in the response module and entity embeddings in the recommendation module after PCA in Figure 2.3. By using the infusion loss, two embeddings tend to cluster together, which indicates the infusion loss can bridge the gap between two representation spaces.

2.4.7 Parameter Sensitivity Study

λ_1 and λ_2 are two important hyper-parameters used to determine the weights of different losses when training the response generation module. We conduct experiments to study the impacts of these two hyper-parameters as shown in Figure 2.4. We note that, with the increase of λ_1, the performances of KECRS are improved first and start to drop when λ_1 is larger than 1.5. As the primary objective of KECRS is learning from the ground truth responses instead of KG, too large λ_1 may lead to negative impacts on the performances of KECRS. For λ_2, when increasing it, the infusion loss may oversmooth the word embedding and reduce the response generation performances achieved by KECRS. λ_3 is a hyper-parameter only used in the testing phase. It is used to introduce

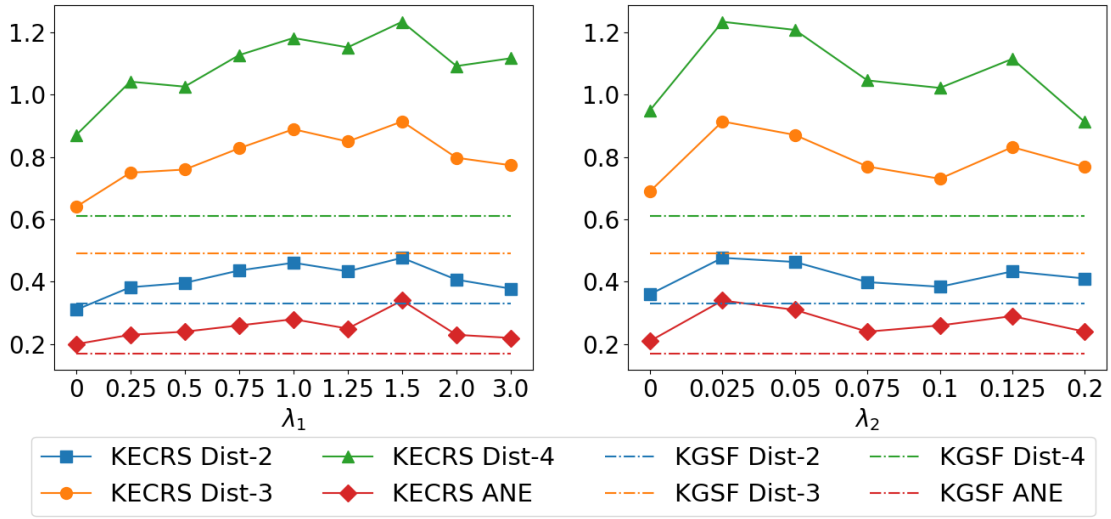


Figure 2.4: Response generation performance trends of KECRS for different λ_1 and λ_2.

position-irrelevant entities into the generated responses. As too large of λ_3 may hurt the fluency of generated responses, we empirically set it to 0.1.

2.4.8 Case Study

In Table 2.4, we present a qualitative comparison of the responses generated by different models. The conversation is selected from the dataset REDIAL. When the user expresses preferences on "*There's Something About Mary*", KECRS infers the user may like romantic comedy movie. Thus, KECRS recommends another romantic comedy movie

User:	Hi there, how are you?
Recommender:	I'm doing great, how about you?
User:	Fine thanks. I 'd love to see something similar to ***There's Something About Mary***. That movie always cracks me up !
Transformer:	I haven't seen that one.
KBRD:	I haven't seen that one. What about ***My Best Friend's Wedding***?
KGSF:	I recommend ***The Other Woman***?
KECRS (Ours):	I love ***Meet the Parents***. It 's a ***classic***. It's a little ***older*** , but still ***funny*** and ***romantic***.

Table 2.4: Case Study.

"*Meet the Parents*" and provides an informative and natural response *"It 's a classic. It's a little older , but still funny and romantic"*.

2.5 Limitations

Despite its promising results, the proposed KECRS model has three main limitations. First, while integrating a knowledge graph (KG) enhances conversational recommendations, the model relies on manually curated external KGs, which may be incomplete or contain irrelevant information. This limitation restricts the model's ability to generate fully informative responses. Second, the recommendation network and the response generation network are trained separately, with their alignment depending solely on the infusion loss. This indirect alignment may be insufficient, as it does not explicitly optimize the two representation spaces together. Consequently, entities relevant to recommendations may not naturally appear in responses. Moreover, separate training prevents the model from leveraging end-to-end feedback—improvements in response generation do not refine the recommendation process, limiting the system's ability to dynamically adapt to user interactions. Third, although our model is applicable to conversational recommendation in any domain, due to the limited availability of public datasets, the experiments were conducted only on movie recommendations.

2.6 Summary

In this chapter, we propose a novel Knowledge-Enriched Conversational Recommendation System (KECRS). Specifically, we develop the Bag-of-Entity (BOE) loss and the alignment loss to improve the response generation performances. The experimental results on REDIAL demonstrate that the proposed BOE loss can guide the model to generate more knowledge-enriched responses by selecting entities in KG, and the alignment loss can tie the word embeddings to the entity embeddings. Overall, KECRS achieves superior response quality than state-of-the-art baselines.

Chapter 3

History-Aware Multi-session Open-domain Dialogue Systems

3.1 Overview

In the previous chapter, we focused on personalization, where the dialogue system generates responses tailored to users' individual preferences and goals. While personalization enhances immediate user satisfaction, sustaining personalized interactions over time requires the system to retain and utilize information from previous sessions. Human conversations rarely occur in isolation; rather, they span multiple sessions, during which speakers draw upon shared history to ensure coherence, efficiency, and relationship building. A user-centered dialogue system should similarly be able to recall what the user said previously, maintain context, and build upon earlier sessions. In this chapter, we propose to build an open-domain dialogue system that can remember and actively utilize the previously shared history to enhance users' engagement and long-term experience with dialogue systems.

Open-domain dialogue systems, also known as chatbots, are designed to chat with and engage users on any topic with the aim of establishing, maintaining, and strengthening long-term relationships [64, 16]. With the availability of large dialogue datasets, the latest open-domain dialogue systems built based on large-scale generative pre-trained models, *e.g.,* Meena [65], BlenderBot [66], and DialogueGPT [50], have improved the performance of chatbots to a large extent.

However, most existing chatbots are designed to interact with users in a single conversation session. When the current session ends, the chatbot forgets its contents and

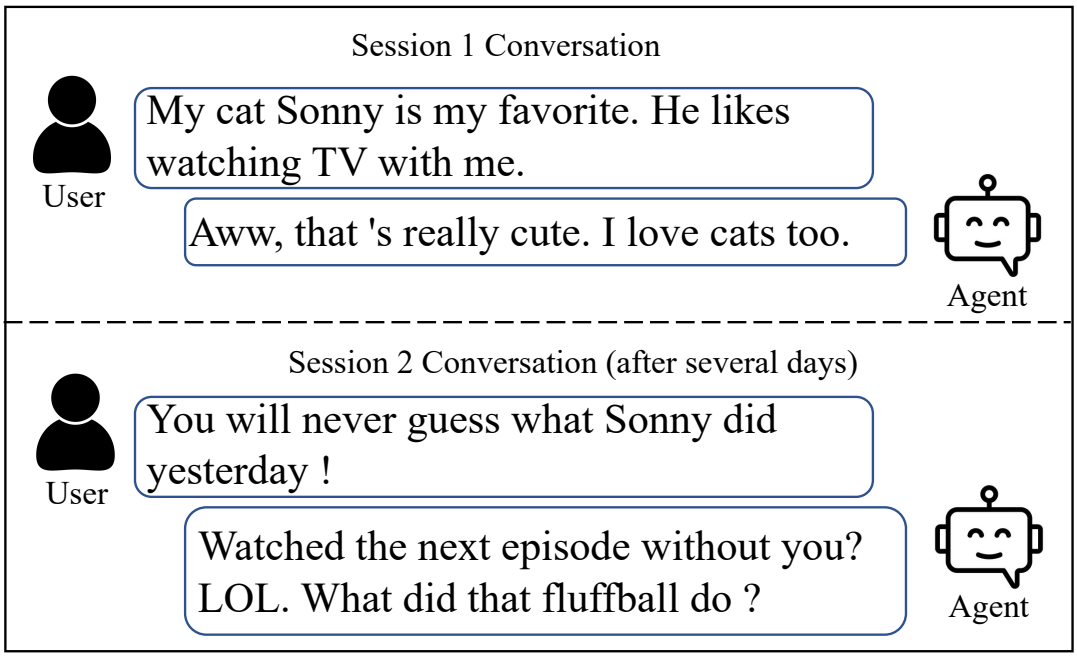


Figure 3.1: An illustrated example of a two-session conversation between a user and an agent.

will commence a new independent session with the same user next time. When previously discussed topics reemerge, such chatbots often appear ignorant and fail to reengage users appropriately. The apparent forgetfulness limits the chatbots' ability to establish and maintain long-term relationships with users.

We argue that, to better engage users in multi-session conversations (MSCs), a chatbot should maintain a long-term memory of historical contexts, which allows the chatbot to reengage the user appropriately when similar contexts reemerge. By learning from historical conversations, the chatbot should gradually refine its understanding of and deepen its relationship with the user. Figure 3.1 shows an example of a two-session conversation between a user and a chatbot. In the second session, the chatbot infers that Sonny is a cat and generates the response based on the history information that Sonny likes watching TV with the user.

History-aware chatbots will be able to generate more well-informed and context-relevant responses, which can help to elicit long-term commitments and develop emotional attachments from users to sustain close relationships over time. To this end, we propose the History-Aware Hierarchical Transformer (HAHT) for multi-session open-

domain dialogue systems, which can effectively leverage history conversations to conduct more engaging MSCs. HAHT maintains a long-term memory to store historical conversational contexts, which is updated when a new session is conducted. Based on the long-term memory and the context in the current session, relevant tokens in historical contexts are selected to adapt the current response.

Specifically, as the number of tokens in a conversation utterance and the number of turns in a conversation are usually not very long[1], we first encode the history conversation hierarchically into the history memory using Transformer [60]. The history memory serves as a high-level representation of history conversations. Secondly, as history conversations usually can facilitate the understanding of the current conversation context, we design a history-aware context encoder. The context encoder encodes conversation context, considering both history conversations and the current conversation, by adopting the transformer attention over the history memory and current conversation context. Then, the context encoder also updates the history memory based on the current conversation context. Finally, we design a history-aware decoder to fuse learned history information into the response generation process. The history-aware decoder can switch between two strategies, *i.e.,* generating a word from the generic vocabulary or directly copying a word from history conversations.

Experimental results on the large-scale Facebook MSC dataset show that the proposed HAHT model outperforms previous multi-session open-domain dialogue systems in various evaluation metrics. Human evaluation results support that HAHT generates more readable, context-relevant, and history-relevant responses than baseline models. In addition, the ablation study confirms that both the hierarchical encoding of history conversations and the history-aware decoder contribute greatly to HAHT's performance on MSCs and help it leverage historical information more effectively.

3.2 Related Work

Open-domain dialogue systems aim to perform chit-chat without task and domain restrictions [67] and establish long-term relationships with users [64, 16]. They are generally divided into two groups: generation-based systems and retrieval-based systems.

[1]On average, conversations have 13 turns and conversation utterances have 16 tokens in Facebook MSC dataset.

Retrieval-based systems seek to find a suitable response from a large response candidate set [68, 69, 70, 71, 72], whereas, generation-based systems focus on generating responses from scratch based on the dialogue history [73, 74, 65, 16, 21]. In this chapter, we focus on generation-based systems.

Early approaches to response generation include template-based generation methods [75] and statistical machine translation (SMT) methods [67]. With the development of deep learning, sequence-to-sequence (Seq2seq) models have been applied to generation-based dialogue systems and achieved great performance [76, 15, 77]. Recently, with the increasing availability of large-scale dialogue datasets [78, 79, 31, 4], Transformer-based language models pretrained with large-scale corpora, such as Meena [65], BlenderBot [66], DialogueGPT [50], and PLATO [80], have made significant progress in the area of open-domain dialogues.

Despite the advancements in the field, current state-of-the-art generative pre-trained models are designed for and trained on large datasets of single-session conversations with a small number of turns. As a result, most existing models employ short token truncation lengths, such as 128 tokens for Meena [65], and are unable to encode and utilize historical contexts in MSCs effectively. In addition, there is also a lack of public MSC datasets. Xu et al. [21] released the first multi-session conversation dataset, *i.e., Facebook MULTI-SESSION CHAT (Facebook MSC)*, and explored different retrieval-augmented generative models on the dataset [81, 82], which achieved better results than the standard Transformer [60]. However, the experimental results demonstrate that their methods need to retrieve a very large portion of history conversations to achieve better results than the standard Transformer. In addition, these models still need to concatenate the retrieved raw history conversation text with the current conversation context, yielding concatenations that are still much longer than the 128 token truncation lengths. Therefore, the incorporation of historical contexts in these methods is still limited by the short token truncation lengths of pre-trained models.

With recent advancements in model architectures and increased GPU availability, modern LLMs such as Gemini and GPT-4 have demonstrated strong capabilities in processing extended contexts. For instance, Gemini 1.5 can handle up to 1 million tokens, compared to just 128 tokens in HAHT. Consequently, most subsequent works have shifted from dense to text-based memory. For example, LD-Agent [83] leverages LLMs to generate event summaries and user personas from conversational history as long-term memory. ComPeer [84] detects and reflects on significant events during dialogue to strategically plan when and what kind of proactive support to provide. Similarly, COMEDY [85] compresses memory into concise events, fine-grained user profiles, and dynamic relationship trajectories across sessions. To better model historical events, THEANINE [86] represents memory as timelines of causally linked events. Beyond LLMs' capabilities, several works also draw inspiration from human cognition. For instance, MemoryBank [87] applies the forgetting curve to dynamically manage memory based on elapsed time and memory relevance. To enable better evaluation of long-term dialogue agents, Maharana et al. [88] introduce a benchmark (LOCOMO) for assessing very long-term memory in models. It includes tasks like question answering, event summarization, and multimodal dialogue generation, offering a comprehensive testbed for evaluating long-term conversational capabilities.

3.3 The Proposed Method

In general, a *Multi-Session Conversation (MSC)* consists of a current conversation session and several history conversation sessions that happen before the current one, all between the same two interlocutors. A multi-session open-domain dialogue system aims to generate natural, well-informed, and context-relevant responses to the user's utterances based on all history conversation sessions and the current conversation context.

Formally, we denote the MSC dataset D by a list of N conversations in the format of (H, X, y). Here, $X = \{x_1, x_2, \cdots, x_{n_x}\}$ denotes n_x context utterances of the current conversation session. $H = \{H^1, H^2, \cdots, H^M\}$ denotes M history conversation sessions, where $H^i = \{h^i_1, h^i_2, \cdots, h^i_{n_i}\}$ denotes n_i chronologically ordered utterances of the i-th history conversation session. y is the ground truth response to X under the background

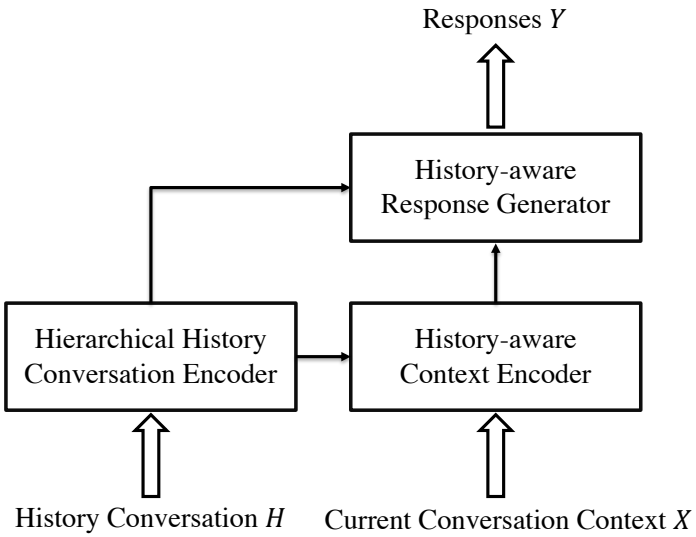


Figure 3.2: The overall structure of the proposed HAHT model, which contains 1) hierarchical history conversation encoder, 2) history-aware context encoder, and 3) history-aware response generator. The details of each component are shown in Figure 3.3, 3.4, 3.5, respectively.

of H. The MSC task can be formulated as learning a function $f(H, X)$ to predict the next utterance x_{n_x+1} based on H and X.

In this chapter, we propose a novel model, namely HAHT, for the MSC task. Figure 3.2 shows the overall structure of HAHT, which consists of three main components: 1) hierarchical history conversation encoder, 2) history-aware context encoder, and 3) history-aware response generator. We present the details of each component of HAHT as follows.

3.3.1 Hierarchical History Conversation Encoder

The main challenge in encoding history conversation sessions is the limited maximum input length imposed by pre-trained dialogue systems. If all history conversations are simply concatenated and fed into the pre-trained dialogue system, the length of the concatenation will exceed the maximum input length. Thus, most parts of the input will be truncated. To preserve more information in the history conversation, we encode different history conversation sessions separately and encode each history conversation hierarchically.

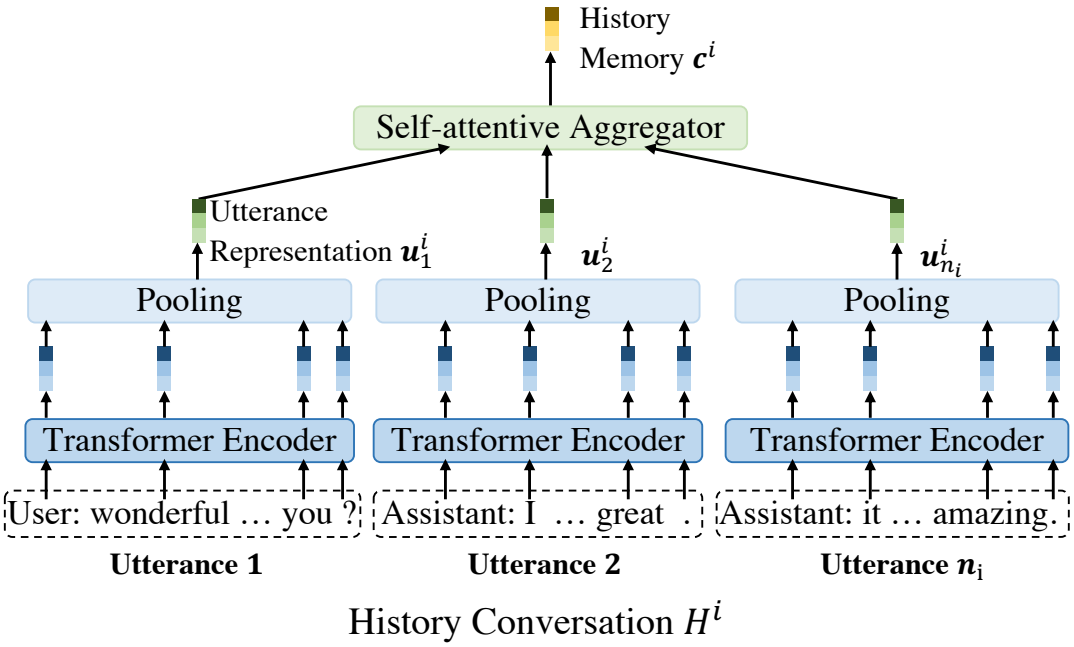


Figure 3.3: The structure of the hierarchical history conversation encoder in HAHT.

Specifically, for a history conversation session $H^i = \{h_1^i, h_2^i, \cdots, h_{n_i}^i\}$, we first prepend a special token "User:" or "Assistant:" to each utterance h_j^i in H^i depending on the role of the utterance speaker, and then pad all utterances to the same length l_{utter}. For each utterance h_j^i, we apply an embedding layer E_m, n_{enc} Transformer encoder layers, and a Max-pooling layer to obtain its dense representation as follows,

$$\mathbf{u}_j^i = \text{Max-pooling}\big(\text{Transformer}_{n_{enc}}(E_m(h_j^i))\big), \tag{3.1}$$

where $\mathbf{u}_j^i \in \mathbb{R}^d$. Moreover, we denote all the utterance representations in the history conversation H^i by $\mathbf{U}^i = \{\mathbf{u}_1^i, \mathbf{u}_2^i, \cdots, \mathbf{u}_{n_i}^i\} \in \mathbb{R}^{n_i \times d}$, where n_i is the turn number of H^i. Next, we apply a conversation aggregator F_c to aggregate all utterance representations $\mathbf{U}^i$ into the condensed history memory $\mathbf{c}^i$,

$$\mathbf{c}^i = F_c(\mathbf{U}^i). \tag{3.2}$$

The conversation aggregator is developed based on the following self-attentive mechanism [89],

$$\begin{aligned} F_c(\mathbf{U}^i) &= \alpha \mathbf{U}^i, \\ \alpha &= \text{softmax}\big(\mathbf{W}_k \tanh(\mathbf{W}_q \mathbf{U}^{i\top})\big), \end{aligned} \tag{3.3}$$

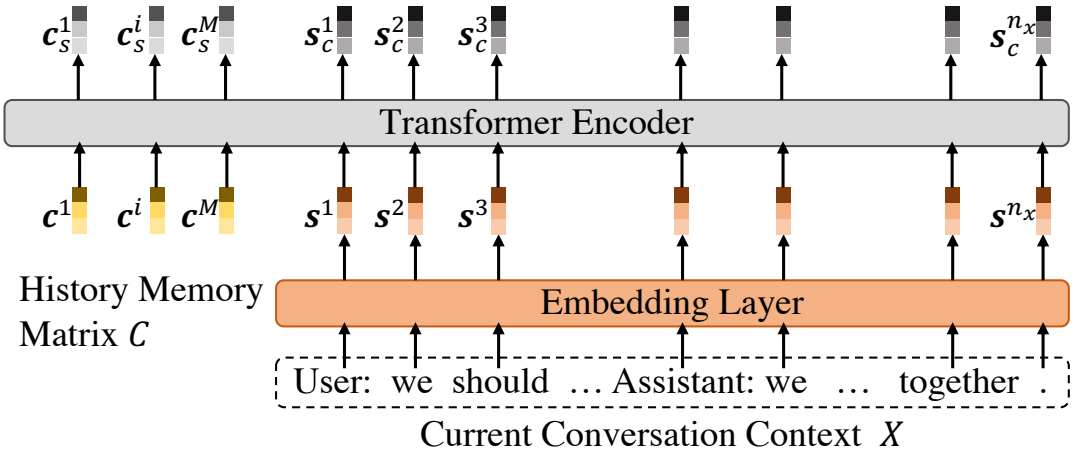


Figure 3.4: The structure of the history-aware context encoder in HAHT.

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable parameters. $\alpha \in \mathbb{R}^{n_i}$ is the importance vector of the history conversation utterances in H^i.

After applying previous steps to all history conversations H, we will finally obtain a history memory matrix $\mathbf{C} \in \mathbb{R}^{M \times d}$ containing a history memory for each history conversation, where M is the number of history conversation sessions.

3.3.2 History-aware Context Encoder

History conversation sessions usually contain the background stories (*e.g.,* interlocutors' profiles or previous discussions between them) that bring out the current conversation session. Leveraging the history conversations will help the model to better understand the current conversation context and respond properly. On the other hand, the current conversation context can help the model update the history memories. Thus, we encode the history memory $\mathbf{C}$ together with the current conversation context by adopting the transformer attention between them.

For the current conversation context X, we also prepend a special token "User:" or "Assistant:" to each utterance depending on the role of the utterance speaker and concatenate all utterances into a single sentence. Then, we adopt the embedding layer E_m to obtain a sequence of context token embeddings $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{n_x}\}$, where n_x is the length of the context sequence. Next, we concatenate the history memory matrix $\mathbf{C} \in \mathbb{R}^{M \times d}$ with $\mathbf{S} \in \mathbb{R}^{n_x \times d}$ over the first dimension and apply n_{enc} Transformer encoder layers.

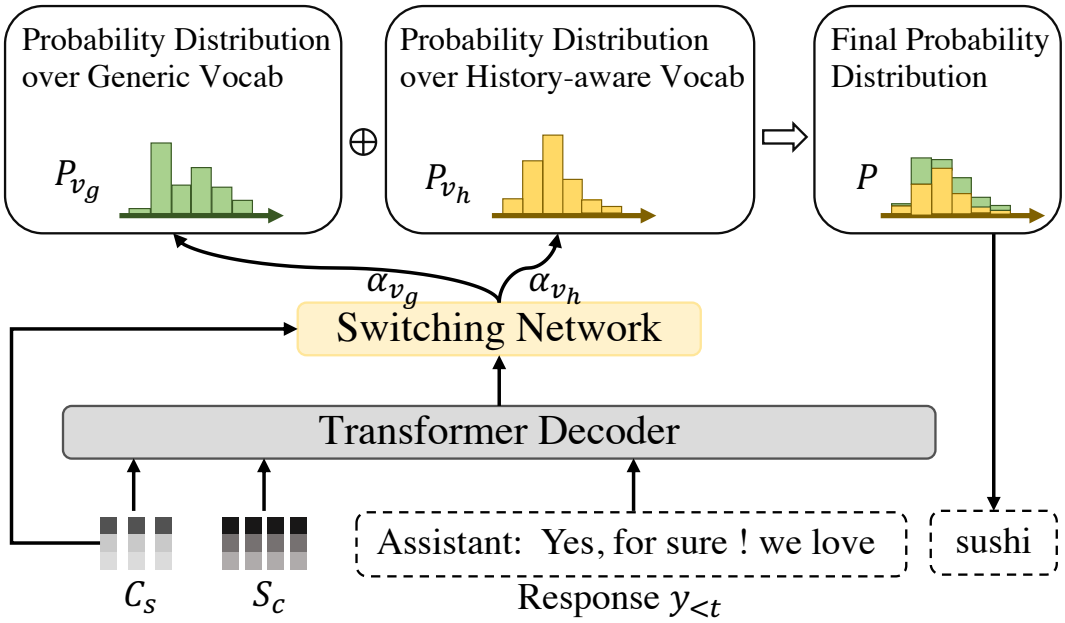


Figure 3.5: The structure of the history-aware response generator in HAHT.

By employing attention in the transformer encoder layers, our model can understand the conversation context by attending to all context token embeddings and history conversation memories. We denote this history-aware context encoding by $\mathbf{S}_c \in \mathbb{R}^{n_x \times d}$. After context encoding, history conversation memories are updated based on the latest information from the current conversation context. We denote this context-updated history memory as $\mathbf{C}_s \in \mathbb{R}^{M \times d}$. The concatenation of $\mathbf{C}_s$ and $\mathbf{S}_c$ over the first dimension will become the input of the response generator.

3.3.3 History-aware Response Generator

Inspired by CopyNet [90], we construct two vocabularies, *i.e.,* generic vocabulary V_g and history-aware vocabulary V_h, to better generate history-aware responses. The generic vocabulary V_g contains the words that appear in all the training dataset, and the history-aware vocabulary V_h only contain the words that appear in the history conversations H. To generate a word of the response, the response generator will choose to generate a generic word from V_g or directly copy a word from V_h based on the switching mechanism [91].

Specifically, at each decoding time step t, we feed $\mathbf{C}_s$, $\mathbf{S}_c$ and the ground truth word sequence before j into n_{dec} Transformer decoder layers and obtain a hidden representation vector $\mathbf{o}_j \in \mathbb{R}^d$. The probability distribution over the generic vocabulary V_g at the decoding time step t is computed as,

$$P_{v_g} = \text{softmax}\big(\text{FC}_1(\mathbf{o}_t)\big), \tag{3.4}$$

where FC_1 is a fully connected layer.

To calculate the probability distribution over the history-aware vocabulary V_h, we adopt a max-pooling layer over the context-updated history memory $\mathbf{C}_s$, a fully connected layer, and a softmax function as follows,

$$P_{v_h} = \text{softmax}\big(\text{FC}_2(\text{max-pooling}(\mathbf{C}_s))\big), \tag{3.5}$$

where FC_2 is a fully connected layer.

The final word probability distribution at time step t is computed by using a switching mechanism between P_1 and P_2 as follows,

$$P = \alpha_{v_g} * P_{v_g} + \alpha_{v_h} * P_{v_h}, \tag{3.6}$$

where α_{v_g} and α_{v_h} is the switching probability of generating from generic vocabulary or copying from history conversations. α_{v_g} and α_{v_h} is calculated as follows,

$$\begin{aligned}[\alpha_{v_g},\, \alpha_{v_h}] = \\ \text{softmax}\big(\text{FC}_3([o_j; \text{max-pooling}(\mathbf{C}_s)])\big),\end{aligned} \tag{3.7}$$

where FC_3 is a fully connected layer, and [;] is a concatenation operation over the last dimension.

3.3.4 Model Training

We train the model to maximize the generation probability of the target response, given the current conversation context and history conversations in an end-to-end manner. The loss function of HAHT is defined as,

$$\mathcal{L} = -\sum_{t=1}^{n_y} \log\big(P(y_j|X, H, y_{<t})\big), \tag{3.8}$$

where X denotes the current conversation context, H denotes all history conversations, $y_{<t}$ denotes tokens before time step t, and n_y denotes the length of the ground truth response.

Session number	Train		Valid		Test	
	Conv.	Utter.	Conv.	Utter.	Conv.	Utter.
1*	8939	131,438	1000	7,801	1015	6,634
2	4000	46,420	500	5,897	501	5,939
3	4000	47,259	500	5,890	501	5,924
4	1001	11,870	500	5,904	501	5,940
5	-	-	500	5,964	501	5,945
Total	-	236,987	-	31,456	-	30,382

Table 3.1: The statistics of Facebook Multi-Session Chat (Facebook MSC) Dataset. Session number i indicates there are i-1 history conversation sessions that happen before the last conversation session.*: Session 1 does not contain history conversation sessions.

Model	Session 2			Session 3			Session 4			Session 5		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
BlenderBot	2.79	0.65	13.73	2.41	0.45	13.06	2.14	0.39	12.76	2.26	0.45	12.75
BlenderBot$_{\text{msc}}$	4.76	1.51	16.18	5.03	1.61	16.39	4.78	1.49	15.56	4.98	1.48	16.10
FID-RAG	4.82	1.54	16.53	5.04	1.61	16.42	4.84	1.48	15.89	5.06	1.57	16.01
HAHT (ours)	**5.07***	**1.57**	**16.90***	**5.27***	**1.67**	**16.72***	**5.00***	**1.55***	**15.97***	**5.16***	**1.60***	**16.42**

Table 3.2: Automatic evaluation results of different models on all session data. Session i indicates there are i-1 history conversation sessions. B-2, B-3, and R-L denote BLEU-2, BLEU-3, and Rouge-L respectively. The best results are in **boldface**. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*.

3.4 Experimental Settings

In this section, we introduce the experimental dataset, evaluation metrics, baseline methods, and model settings.

3.4.1 Experimental Dataset

The experiments are performed on a large dataset, *i.e.,* Facebook MULTI-SESSION CHAT (Facebook MSC) [21]. It is a crowdsourced dataset consisting of multiple chat sessions, where the speaking partners learn about each other's interests and discuss the things they have obtained from past sessions. The number of history conversations in Facebook MSC varies from 1 to 4. Session number i indicates there are i-1 history conversations happening before the last conversation session. The statistics of the Face-

Model	Readability	Context Relevancy	History Relevancy
BlenderBot	1.78	1.13	0.09
$\text{BlenderBot}_{\text{msc}}$	1.82	1.56	0.13
RAG-FID	1.89	1.84	0.21
HAHT (ours)	**2.05***	**2.03***	**0.33***

Table 3.3: Human evaluation of the response generation by different methods. All scores are rated in four levels 0/1/2/3. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*.

book MSC dataset are summarized in Table 3.1. As session 1 does not contain history conversations, we evaluate our model from session 2-5.

3.4.2 Evaluation Metrics

We conduct both automatic and human evaluations to demonstrate the effectiveness of the proposed model. For automatic evaluations, we leverage BLEU-2, BLEU-3 [92], and ROUGE-L [93] to measure word overlaps between the generated response text and ground truth text.

Moreover, we also randomly sample 50 MSCs from the test set to conduct human evaluations. We present all the history conversation sessions, current conversation context, and the generated responses to three well-educated annotators. The annotators will evaluate the quality of the generated responses from the following three aspects: 1) **Readability**, which measures whether the generated responses are natural and fluent, 2) **Context Relevancy**, which measures whether the generated responses are correlated with the current conversation context, 3) **History Relevancy**, which measures whether the generated responses are correlated with history conversations. Each aspect is rated in four different levels 0/1/2/3, and the final score of each aspect is the average of the scores given by all annotators. For all evaluation metrics, the higher value indicates better performance.

3.4.3 Baseline Methods

We compare the proposed HAHT model with the following baseline methods.

Model	Session 2			Session 3			Session 4			Session 5		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
BlenderBot	4.71	1.47	18.20	3.85	0.93	17.10	3.69	0.83	16.78	4.00	1.19	17.19
$\text{BlenderBot}_{\text{msc}}$	6.39	2.56	19.30	5.82	1.93	18.67	5.30	1.76	17.9	6.10	2.30	18.65
FID-RAG	6.41	2.51	19.82	5.83	1.95	18.38	**5.81**	1.85	**18.44**	6.02	2.27	18.52
HAHT (ours)	**6.69***	**2.73***	**20.02***	**6.03***	**2.20**	**18.70***	5.48	**1.95**	18.00	**6.38***	**2.51***	**19.18***

Table 3.4: Automatic evaluation results of different models on session-opening data. Session i indicates there are i-1 history conversation sessions. B-2, B-3, and R-L denote BLEU-2, BLEU-3, and Rouge-L respectively. The best results are in **boldface**. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*.

- **BlenderBot** [66]: This is a large-scale open-domain dialogue model pre-trained on the dialogue data scraped from social discussions on the web.
- **$\text{BlenderBot}_{\text{msc}}$**: This is the BlenderBot model finetuned on the MSC dataset.
- **FID-RAG** [82] : In this method, RAG-trained retriever [81] is used to retrieve top-N history conversations, and Fusion-Decoder (FiD) [94] is adopted to generate a final response considering the retrieved history conversations and current conversations. Following [21], N is empirically set to 5.

3.4.4 Model Settings

In this chapter, all the evaluated methods are trained following the same settings. Due to the limitation of computation resources, we use the BlenderBot model with 90M parameters as the initial pre-trained model and finetune it on the Facebook MSC dataset. The input length truncation is set to 256. The number of Transformer encoder layers n_{enc} and encoder layers n_{dec} are both set to 12. For model training, we use the Adamax optimizer [95] with a learning rate of 1×10^{-6} , batch size of 16, dropout ratio of 0.1, and early stopping patience of 10.

3.5 Experimental Results

This section presents the experimental results of automatic evaluation, human evaluation, evaluation on session openings, ablation study, and case study.

Model	Session 2			Session 3			Session 4			Session 5		
	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L	B-2	B-3	R-L
HAHT	**5.07**	**1.57**	**16.90**	**5.27**	**1.67**	**16.72**	**5.00**	**1.55**	**15.97**	**5.16**	**1.60**	**16.42**
$\text{HAHT}_{\text{w/o HIER}}$	5.00	1.57	16.72	5.19	1.63	16.61	4.86	1.49	15.90	5.10	1.57	16.21
$\text{HAHT}_{\text{w/o HIST}}$	4.98	1.50	16.81	5.09	1.58	16.51	4.75	1.45	15.51	5.10	1.49	16.24
$\text{HAHT}_{\text{w/o SW}}$	5.01	1.56	16.86	5.19	1.61	16.46	4.87	1.55	15.88	5.07	1.55	16.17

Table 3.5: The performance achieved by HAHT and different HAHT variants. Session i indicates there are i-1 history conversation sessions. B-2, B-3, and R-L denote BLEU-2, BLEU-3, and Rouge-L respectively. The best results are in **boldface**. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*.

3.5.1 Automatic Evaluation

The automatic evaluation results of different models are shown in Table 3.2. It can be observed that $\text{BlenderBot}_{\text{msc}}$ performs much better when finetuned on the MSC dataset. FID-RAG performs better than $\text{BlenderBot}_{\text{msc}}$. The potential reason is that RAG can retrieve important history conversations, and FID can combine the retrieved conversations with current conversations to generate better responses. Moreover, the proposed HAHT model consistently outperforms baseline methods in terms of all the evaluation metrics. This indicates that HAHT can better encode the history conversations, leverage history conversations to understand the current conversation context, and generate more human-like responses.

3.5.2 Human Evaluation

Table 3.3 summarizes the human evaluation results on the Facebook MSC dataset. Generally, HAHT outperforms all the baseline methods in terms of all perspectives. This observation is consistent with the automatic evaluation results shown in Table 3.2. In particular, we find that HAHT performs much better than other baselines in terms of history relevancy. This demonstrates that HAHT can better leverage the history conversation sessions and engage the user more in the on-going session with the history memory. HAHT also performs better than other baselines in terms of readability and context relevancy. This indicates that HAHT can better understand the current conversation context with the help of the history memory.

3.5.3 Evaluation on Session Openings

In the MSC task, the session opening is the first conversation turn of the current conversation. According to our observation and the similar observation in [21], the opening conversation turn is categorically different from other conversation turns. It typically involves a statement or question that aims to reengage the other speaker based on the known information that has been exchanged in history conversations. Therefore, the performance on the session opening data can further demonstrate the model's capability in understanding and leveraging history conversations.

We compare all models on these opening responses and show the results in Table 3.4. We observe that the proposed HAHT model achieves the best performance in terms of most metrics. Especially, when there are 4 history conversations, HAHT outperforms FID-RAG and $\text{BlenderBot}_{\text{msc}}$ by 10.6% and 9.1% in terms of BLUE-3. This indicates that the proposed HAHT can better leverage conversation history to reengage the user into a new conversation session.

3.5.4 Ablation study

To better understand the effectiveness of each main component of HAHT, we conduct ablation study for HAHT. Specifically, we consider the following variants of HAHT.

- $\textbf{HAHT}_{\textbf{w/o HIER}}$: In this variant, we do not encode the history conversations hierarchically. Instead, we concatenate all the utterances of history conversations into a long sentence and directly encode it using the transformer encoder.
- $\textbf{HAHT}_{\textbf{w/o HIST}}$: In this variant, we remove the history encoder from HAHT.
- $\textbf{HAHT}_{\textbf{w/o SW}}$: In this variant, we remove the switching mechanism from the response generator of HAHT.

Table 3.5 summarizes the results achieved by different HAHT variants, in terms of BLEU-2, BLEU-3, and Rouge-L. From Table 3.5, we note that HAHT outperforms $\text{HAHT}_{\text{w/o HIER}}$, which indicates that hierarchically encoding the history conversations can help the model reserve more history memory to generate more human-like responses.

History Conversation Sessions	
...	
User:	... Have you had fish lately?
Agent:	Yes! I had salmon the other night with ***my partner***!
User:	That's great! Salmon is one of my favorite types of fish. ...
Agent:	Yes! ***Salmon*** is our ***favorite fish*** as well! We also love it in sushi!
...	
Agent:	Not in a while. It is expensive and ***my partner*** and I are saving up to buy a house!
...	
Current Conversations	
User:	We should go eat seafood with our significant others.
Responses from Different Dialogue Systems	
BlenderBot:	Sounds like the perfect plan!
$\text{BlenderBot}_{\text{msc}}$:	That sounds like a great idea! I'm sure we'll have a great time.
FID-RAG:	That sounds like a great idea! I'm not sure if we'll be able to get the seafood to work, but I'd love to go.
HAHT (ours):	That sounds great! I'll bring my husband along. We all love seafood.

Table 3.6: Case Study.

Moreover, HAHT achieves better performance than $\text{HAHT}_{\text{HIST}}$. This observation indicates that removing the history encoder causes the most decline in all metrics. This result confirms the necessity to leverage history conversations to understand the current conversation and generate the response. In addition, the performance degradation caused by removing the switching mechanism shows that directly copying words from the history conversation can help the model generate more history-aware responses.

3.5.5 Case Study

Table 3.6 shows a case study about the multi-session conversations generated by different models. Compared with baseline models, the proposed HAHT model can better leverage the history conversations to understand the current conversation and generate more history-aware responses. As shown in Table 3.6, when the user expresses the intention to "*eat seafood with significant others*", HAHT can remember information mentioned in the history conversation that it has a partner, the partner is her husband, and they

both like seafood. HAHT can leverage those history memory and generate more human-like, context relevant, and history-aware responses: *"That sounds great! I'll bring my husband along. We all love seafood."*.

3.6 Limitations

One limitation of this work is that HAHT has only been evaluated on one dataset. However, to the best of best of our knowledge, Facebook MSC is, by far, the only large-scale multi-session conversation dataset available. Nevertheless, our proposed model consistently outperforms baseline models on conversations with different numbers of history sessions in Facebook MSC. A potential solution to this limitation is to construct more MSC datasets in open-domain or in specific-domain that may benefit from the awareness of history conversations, *e.g.,* conversational recommendation or automatic medical assistants.

3.7 Summary

In this chapter, we propose the History Aware Hierarchical Transformer (HAHT) model for the multi-session open-domain dialogue system. Differing from previous works that concatenate history conversation sessions with the current conversation context, the proposed HAHT model hierarchically encodes the history conversation sessions into history memory. Then, HAHT encodes conversation context leveraging history memory and updates the history memory based on the current conversation context. Finally, the switching mechanism is used to explicitly generate history-ware responses. Experimental results obtained under both the normal multi-session conversation setting and the session opening multi-session conversation setting demonstrate that HAHT can generate more humanized and history-aware responses than state-of-the-art models.

Chapter 4

Understanding the Benefits of Conversations for Explainable AI

4.1 Overview

While personalization and continuity focus on adapting to users over time and across sessions, another critical aspect of user-centered dialogue systems is their ability to serve users with diverse backgrounds and knowledge levels. A personalized and memory-aware system may still fall short if it assumes uniform user expertise or communication preferences. To truly support all users, dialogue systems must be designed to be accessible and understandable for individuals from different domains and with varying levels of familiarity with complex topics.

When users encounter relevant complex scientific content, they may struggle to understand it. A dialogue system that offers personalized responses to users' questions can significantly facilitate their understanding of such complex scientific material. However, users are often from different domains with different knowledge levels. A user-centered dialogue system should be accessible and adaptable to those diverse individuals. In this chapter, we focus on Explainable AI, which is an important but complex scientific topic for lay users to understand [22, 23]. We conduct Wizard-Of-Oz experiments to explore the potential benefits of conversations for explainable AI.

The rapid advancement of Artificial Intelligence (AI) is largely powered by opaque deep neural networks (DNNs), which are difficult to interpret by humans [96]. The lack of transparency prevents verification of AI decisions by human domain experts and

is especially concerning in areas of high-stake decisions, such as healthcare and law enforcement, where erroneous algorithmic decisions could lead to severe consequences [97, 98, 99] and erosion of public trust [100, 101]. To improve the explainability of AI models, numerous eXplainable Artificial Intelligence (XAI) methods have been proposed (for detailed reviews, we refer readers to [102, 103, 96]). It has been reported that explainability enhances user understanding [104] and trust [105, 106] in AI models, improves human-AI collaboration in decision-making [107, 108], and helps AI developers identify and rectify model errors [109, 110]. Despite these successes, a number of recent studies find that the explanations often do not resolve user confusion regarding the neural networks they are purported to explain [22, 111, 104, 112, 113, 114, 115, 23]. These seemingly conflicting findings warrant further investigation.

We postulate that two major factors contribute to the ineffectiveness of AI explanations. First, the explanations do not properly account for average users' knowledge of machine learning, which may be insufficient to establish causal relations between the explanations and the model behaviors [22, 116, 117, 118]. Communication theory posits that effective communication requires the senders and receivers to establish common ground [119, 120]. However, experts usually find it hard to accurately estimate what laypeople know [121, 122, 123]. To make matters worse, underestimating and overestimating the receivers' knowledge level are equally detrimental to communication [121, 113]. As a result, the explanations designed by experts are almost always at a mismatch with the laypersons' actual knowledge level.

Second, users of XAI have diverse intentions and information needs [112, 124, 114, 118]. For example, Liao and Varshney [125] identify five different objectives of users of explanations, including model debugging, assessing the capabilities of AI systems, making informed decisions, seeking recourse or contesting the AI, and auditing for legal or ethical compliance. One static explanation usually cannot satisfy all objectives and purposes. Therefore, researchers have suggested injecting interactivity to model explanations in order to establish common ground, address knowledge gaps, and create customized explanations that adapt to the users [113, 126, 127, 128, 129, 130].

Existing work on interactive explanations can be broadly categorized into two types. The first type, interactive machine learning [131, 132], allows users to provide feedback

and suggestions to the machine learning model using model explanations. Their primary goal is to improve machine learning performances, rather than explaining model behaviors to layperson users. In this setting, explanations have been shown to improve user satisfaction [133] and feedback quality [134, 135]. The second type aims to elucidate model behaviors by allowing users to freely modify input features and observe how outputs change while showing feature attribution explanations [129, 136, 137, 138]. This type of interactivity has been shown to improve user understanding [129] and perceived usefulness [136] of AI models. However, the effective use of these interactive approaches still requires a rudimentary understanding of machine learning, such as the generic relation between input and output, or what model properties the interpretations reveal. These interactive explanations cannot answer most types of follow-up questions laypeople may have.

Free-form conversations that accompany static explanations are arguably the most versatile mode of interaction as they allow users to ask arbitrary follow-up questions and receive explanations tailored to their backgrounds and needs [114, 139, 113]. Through interviews with decision-makers, Lakkaraju et al. [113] discover that they have a strong preference for explanations in natural language dialogue. They argue that conversational explanations satisfy five requirements of interactive explanations and are ideal for users with limited machine learning knowledge. With the progress in conversational characters [140, 141, 142], especially knowledge-based question answering [143, 144, 145] powered by large language models [146, 147, 148], AI systems that can answer questions about their own decisions appear to be within our reach in the near future. However, before investing effort to develop such a chatbot, it would be beneficial to empirically quantify the effects of conversational explanations.

In the current study, we conduct Wizard-of-Oz experiments to investigate how conversations assist users in understanding static explanations of image classification models, improving acceptance and trust in XAI methods, and selecting the best AI models based on explanations. Specifically, a total of 120 participants join our experiments. We first present them with static explanations for an image classification task and measure their objective understanding and subjective perceptions of static explanations. After that, half of the participants, who are assigned to the experimental group, seek to clarify

any doubts with an online textual conversation with an AI system, played by human XAI experts. The other half of the participants, assigned to the control group, read materials about the static explanations independently. After the conversation or reading session, participants complete the same pre-session measurements. From the results, we estimate the effects of conversational explanations.

The experimental measurements include both an objective component and a subject component of the users' understanding and perception. In the objective evaluation, from three candidate neural networks, the users need to choose one network that would be the most accurate on test data so far unobserved, using information from the static explanations. This task, known as model selection, is one of the most fundamental tasks for machine learning practitioners [149]. By design, the three candidate networks make exactly the same predictions on the same inputs but have different rationales for the predictions, as revealed by the static explanations. Hence, the only way for the users to make the right choice is to correctly understand the explanations. The subjective evaluation contains 13 questions requiring users to self-report three aspects of their perceptions of the static explanations: comprehension, acceptance, and trust.

Results show that free-form conversations with XAI experts in the Wizard-of-Oz setting significantly improve comprehension, acceptance, trust, and collaboration with static explanations. Our study underscores the effects of free-form conversations on neural network explainability in practice and provides insights into the future development of conversational explanations. To the best of our knowledge, this is the first study of how free-form conversations may facilitate neural network explainability in practice.

4.2 Related Work

In this section, we review three bodies of research that motivate our study. First, we explore the existing work of static Explainable Artificial Intelligence (XAI). Second, we discuss interactive explanations, especially the limitations of existing methods and the need for conversations to enhance explainability. Lastly, we examine different types of human-AI collaboration and the design of the subjective evaluation during collaboration.

4.2.1 Static Explanation

Explainable Artificial Intelligence (XAI) refers to those models that can explain either the learning process or the outcome of AI predictions to human users [102]. Static XAI involves models that provide a fixed, one-time explanation, without the capability for further user interaction or exploration. They are usually categorized into two groups: self-explanatory models and post-hoc methods. Post-hoc methods can be categorized into feature attribution methods and example-based methods. Self-explanatory models are inherently transparent, offering clarity in their decision-making processes and facilitating explainability [103, 96]. Examples of such models include linear regression, logistic regression, decision trees, Naive Bayes, attention mechanism [150], decision sets [151], rule-based models [152, 153], among others. However, the requirements of self-explanatory models place constraints on model design, which may cause them to underperform in complex tasks. Conversely, the majority of recent XAI methods are post-hoc XAI methods, which can be used for an already developed model that is usually not inherently transparent [154, 155, 156, 157, 158, 96]. These methods often do not attempt to explain how the model works internally, but instead, employ separate techniques to extract explanatory information. Post-hoc XAI methods can be viewed as reverse engineering processes that employ other independent explanatory models or techniques to extract explanatory information without altering, elucidating, or even understanding the inner workings of the original black-box model. There are two main groups of methods to generate post-hoc XAI explanations, i.e., feature attribution methods and example-based methods.

4.2.1.1 Feature Attribution Methods

Feature attribution methods [159, 154, 160, 161, 162, 155, 163, 164, 165, 166, 167] explain model predictions by investigating the importance of different input features to final predictions. There are two main types of feature attribution methods, gradient-based methods [160, 159, 154, 161, 162] and surrogate methods [155, 163, 164, 165, 166, 167]. Gradient-based methods use gradients/derivatives to evaluate the contribution of a model input on the model output. An example method is Grad-CAM [154]. It superimposes a heatmap on the regions of important input features by weighting the activations

of the final convolutional layer by their corresponding gradients and averaging the resulting weights spatially. Besides directly calculating the importance score of input features, several methods propose to use a simple and understandable surrogate model, e.g., a linear model, to locally approximate the complex deep neural model. Surrogate models can explain the predictions from the complex deep neural model due to their inherent interpretable nature. LIME and its variants are typical methods for generating local surrogate models. LIME [155] builds a linear model locally around the data point to be interpreted and generates an instance-level explanation for the output.

4.2.1.2 Example-based Methods

Example-based methods [156, 157, 168, 169, 170, 171] refer to those that explain predictions of black-box models by identifying and presenting a selection of similar or representative instances. Those examples can be selected or generated from different perspectives, such as training data points that are the most influential to the parameters of a prediction model or the predictions themselves [172, 156, 173], counterfactual examples that are similar to the input query but with different predictions [157, 174, 175, 176, 168, 169, 170], or prototypes that contain semantically similar parts to input instances [177, 178, 171, 179, 180, 181].

In this chapter, we mainly focus on feature attribution methods as they directly highlight the importance of input features, making the decision-making process of models more intuitive [182] than example-based methods for laypeople. Specifically, we select Grad-CAM from gradient-based methods and LIME from surrogate methods to conduct conversational explanations with participants.

4.2.2 Interactive Explanation

Several studies emphasize the need for interactivity in XAI methods [113, 126, 127, 128]. For instance, Lakkaraju et al. [113] find that decision-makers strongly prefer interactive explanations. Similarly, a literature analysis by Abdul et al. [126] suggests that interactions can help users progressively explore and gather insights from static explanations. Rohlfing et al. [128] reason that explanations should be co-constructed in an interaction

between the explainer and the explainee, adapting to individual differences since the human understanding process is dynamic. From an interdisciplinary perspective, Schmid and Wrede [127] underscore the necessity of user-XAI interactions to adapt to diverse information requirements.

To integrate interactivity and explainability, two primary methodologies emerge. One group of methods focuses on using explanations to help users provide feedback about improving machine learning models. In these methods, the interactivity lies in the cycle of model explanation, user feedback, and model improvement. Explanations aim to help users better understand model decisions and provide valuable feedback. As a result, machine learning models can be incrementally trained with additional loss terms from explanatory feedback [135, 183, 134, 184, 185, 133] or with added data points [186, 187, 188, 189]. However, these methods are aimed at machine learning practitioners who can well understand and utilize explanations. Another group focuses on enhancing user understanding of explanations by allowing them to modify the model input and observe changes in the corresponding output. Such interactivity has been shown to improve user comprehension and the perceived utility of AI models [129, 136]. For instance, Tenney et al. [137] and Hohman et al. [138] propose different user interfaces that allow for debugging and understanding machine learning models by examining input-output relationships. However, a rudimentary understanding of machine learning is still required for effective utilization of these interfaces, such as the generic relation between input and output, or what model properties the interpretations reveal.

HCI researchers have recently proposed that XAI methods should align with the ways humans naturally explain mechanisms. Specifically, Lombrozo [190] argues that an explanation is a byproduct of a conversational interaction process between an explainer and an explainee. Miller [123] argues that explanations should contain a communication process, where the explainer interactively provides the information required for the explainee to understand the causes of the event through conversations. Building on this perspective of human explanations, recent works envision "explainability as dialogue" to provide explanations suitable for a wide range of layperson users [114, 139, 113]. While there is much theoretical analysis about the significance of conversations in explainability, practical investigations into their impact on users remain limited. In this context,

two previous works have investigated the practical effect of conversations for explainability [115, 23]. Shen et al. [23] apply conversational explanations to scientific writing tasks, observing improvements in productivity and sentence quality. Slack et al. [115] 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 cope with limited predefined user intentions. In this study, we explore the role of free-form conversations in enhancing users' comprehension of static explanations, and how they affect users' acceptance, trust, and collaboration with these explanations.

4.2.3 Human-AI Collaboration

Human-AI collaboration is an emerging research area, which explores how humans and AI systems can work together to achieve shared goals [191, 192, 182]. Prior studies within this domain have investigated collaborations between humans and various AI systems, from robots [193, 194, 195, 196, 197] to virtual agents [198, 98, 199, 200]. The tasks involved span a broad scope, including text [104] and image [182] classifications, medical diagnosis [98], deception detection [108] and cooperative games [193, 198, 201]. An area of particular interest within these collaborations is the role of explanations in influencing human-AI decision-making [202, 107, 108, 104].

Our study aligns with existing work on human-AI collaboration [202, 107, 108, 104, 201]. In our work, participants need to collaborate with explanations to choose the most accurate neural networks among others. Instead of exploring the role of explanations in collaboration, we mainly examine the potential of conversations in aiding users to effectively use explainability techniques and understand their outputs.

4.3 Method

Our study aims to investigate the impact of conversations on the explainability of AI models by observing participants' comprehension, acceptance, trust of the static explanations, and ability to use the explanations to select the most accurate neural networks before and after the conversation. Our study has received approval from the Institutional Review Board at Nanyang Technological University (#IRB-2023-254).

Table 4.1: Academic disciplines of our participants and the number of participants in each group. There are 120 participants from 4 different discipline groups.

Academic Discipline	**Number of Participants**
Business	23
Engineering	16
Humanities	55
Science	26

4.3.1 Participants

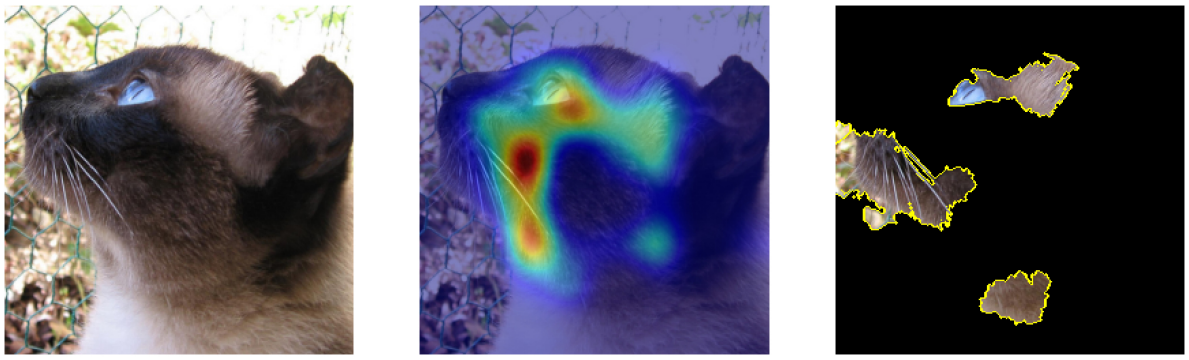
A total of 120 participants joined our study. All were 21 years old or older, fluent in English, and had not been involved in research about XAI previously. We recruited our participants in two ways: by posting advertisements on an online forum and by emailing students and staff across various departments and schools. They are from a wide range of disciplines to promote diversity. For ease of reporting, we categorize their disciplines into four groups:

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

Table 4.1 shows statistics of the academic disciplines that the participants enrolled in.

4.3.2 Experimental Task

In our study, we focus on the image classification task on the ImageNet dataset [203]. Image classification task is a cornerstone in the field of computer vision (CV) that has been the subject of various human-AI collaborative studies [204, 171]. We train three classification models with different top-1 classification accuracies: Swin Transformer [205]



(a) Input to the classification model (Swin Transformer)

(b) Explanation generated by Grad-CAM

(c) Explanation generated by LIME

Figure 4.1: Example explanations generated by Grad-CAM and LIME. (a) is the input to the classification model (Swin Transformer), (b) is the explanation generated by Grad-CAM, and (c) is the explanation generated by LIME. The predicted class of the model is "Siamese cat".

(84.1%), VGG-16 [206] (71.6%), and AlexNet [207] (56.5%). To generate explanations for model predictions, we select two explanation techniques from two main categories of feature attribution explanation methods: LIME [155] (a surrogate method) and Grad-CAM [154] (a gradient-based method). We focus on feature attribution explanations as we believe the relationship between input features and model predictions is more intuitive to understand than example-based methods for laypeople [182]. Figure 4.1 displays example explanations generated by these two explanation methods.

To conduct the study, we design and build a web-based platform where participants can remotely finish the whole procedure of the experiment. After users log into the platform, we first evaluate their objective and subjective understanding of static explanations. The objective explanations require participants to choose, from three classification models, the most accurate on unobserved test data. The three classification models yield identical decisions on 5 images. The only differences between the three networks lie in their explanations. Hence, to select the best model, the participants must rely on the explanations. Figure 4.2 presents an example question, including the original image, the model outputs, and the explanations. The full set of questions used in the study is listed in Figure 4.11 and 4.10.

The subjective evaluation measures participants' self-reported perception of the static explanations, including their comprehension [208, 129], acceptance [209, 210, 211, 212],

Table 4.2: Detailed questions in the subjective evaluation. The user will respond to each question using a 7-point Likert scale.

Aspect	**Question**
Comprehension	How much do you think you understand the explanations provided for predictions of deep learning models?
Perceived Usefulness	Using explanations would improve my understanding of deep learning models' predictions.
	Using explanations would enhance my effectiveness in understanding predictions of deep learning models.
	I would find explanations useful in understanding predictions of deep learning models.
Perceived Ease-of-Use	I become confused when I use the explanation information.
	It is easy to use explanation information to understand predictions of deep learning models.
	Overall, I would find explanation information easy to use.
Behavioral Intention	I would prefer getting explanation information as long as it is available when getting predictions from deep learning models.
	I would recommend others use explanation information to understand predictions of deep learning models.
Trust	How would you rate the competence of the explanation method? - i.e. to what extent does the explanation method perform its function properly?
	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?
	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?
	How would you rate your overall trust in the explanation method?

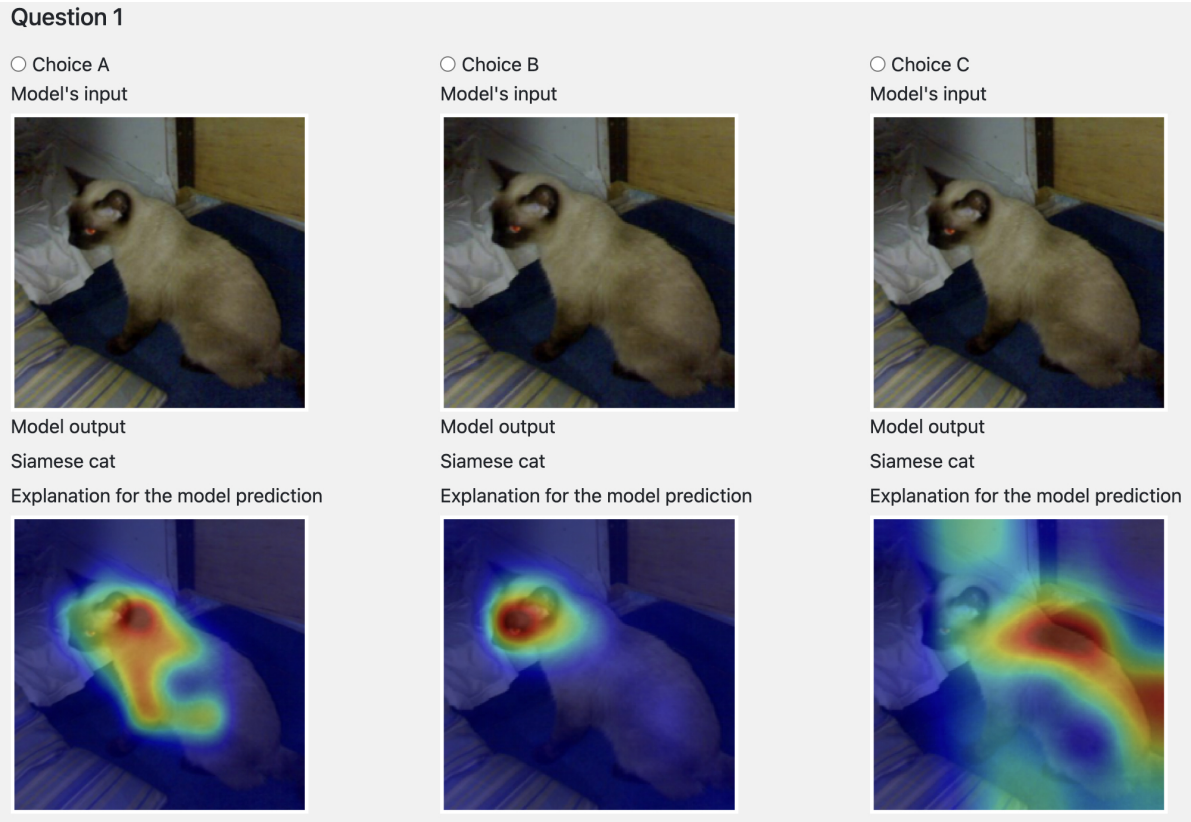


Figure 4.2: An example of the objective evaluation. The objective evaluation aims to objectively measure participants' comprehension of static explanations. Each choice contains a prediction from a different classification model, paired with its respective static explanation. Participants need to choose the best model based on the explanations.

and trust [209, 210, 211, 213, 214]. Based on an in-depth review of existing literature, we chose the questions from those that have been validated in prior research. The subjective evaluation contains a total of 13 questions, each utilizing a 7-point Likert scale for responses. Table 4.2 lists all the questions we used. Labels of the 7-point Likert scale are listed in Figure 4.11 and 4.10.

After these two evaluations, participants are divided into two groups, i.e., the control group and the experimental group. Participants in the control group read static explanations for 15 minutes. Participants in the experimental group conduct conversational explanations with participants in the Wizard-of-Oz (WoZ) setting [215]. They interact with a dialogue system that they believe to be autonomous but is actually operated by a human expert on machine learning.

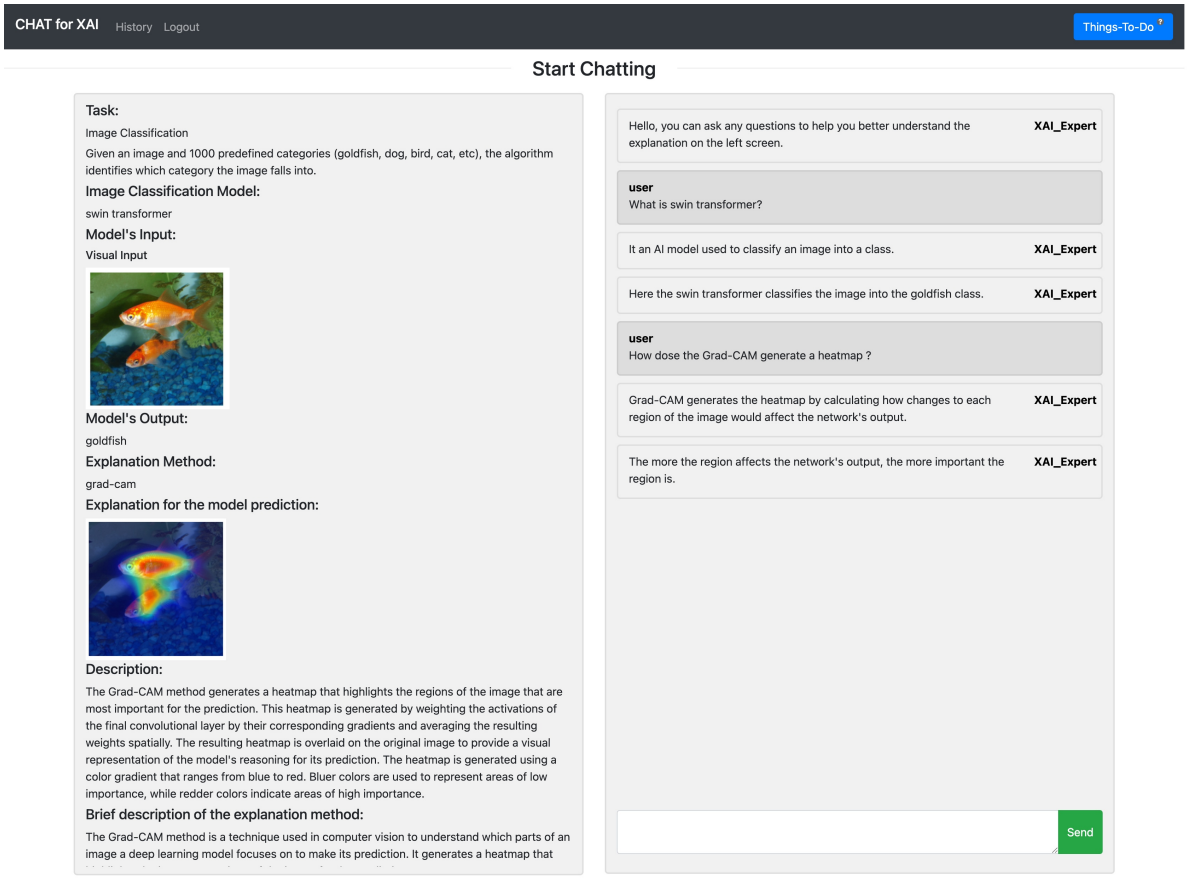


Figure 4.3: The web page where users can discuss static explanations with an expert.

To support the WoZ experiment, we built a conversation page with a two-section structure, as depicted in Figure 4.3. On the left, the page shows a task description, a textual description of the prediction model, a textual description of the explanation technique, an example input image, the model prediction on the input image, a static explanation for the prediction, and a textual description of the explanation. On the right, the interface enables users to converse with XAI experts, seeking clarifications and posing questions about the explanation. For the users in the control group, we replace the textual chat user interface with a 15-minute timer. Once the timer reaches zero, users are allowed to proceed to the post-evaluations. Users from both groups receive the same post-evaluations, which are identical to the pre-evaluations. We discuss the evaluations below.

4.3.3 Experimental Design

There are two independent variables and two categories of dependent variables. The independent variable in the experiments is the explanation method: LIME or Grad-CAM and the method of understanding static explanations: conversation with human experts or reading static explanations. As we devise both subjective and objective evaluations before and after conversations or readings, two categories of dependent variables were collected in the experiment: the model selection accuracy and the self-reported perception scores.

4.3.3.1 Objective Evaluation – Selection of Classification Models

The evaluation aims to objectively evaluate participants' understanding of the static explanations. Participants are presented with 5 input images, on which the three neural networks make the same decisions. The only differences between the three networks 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.

We recognize that existing explanation techniques are not always faithful to the underlying model [216, 217, 218] and do not always provide actionable information for model selection. As our goal is to test if the users can understand the static explanations *when* they do provide actionable information, rather than evaluating the static explanations themselves, we selected input images where better classification models indeed have more reasonable and intuitive 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. In addition, good models 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 missing or occluded. In the example in Figure 4.2, Model B is better than Model A or Model C as Model B utilizes both the head and the body of the cat for classification. In addition,

unlike Model A, Model B does not focus on the background, which is irrelevant to the predicted class, Siamese Cat.

4.3.3.2 Subjective Evaluation

We also measure participants' subjective perception of static explanations, including their comprehension, acceptance, and trust. The subjective evaluation contains a total of 13 questions listed in table 4.2. All questions utilize a 7-point Likert scale for responses.

- Comprehension [208, 129]: Participants' subjective perceptions of their understanding of explanations. It complements the objective evaluation, providing a holistic perspective on participants' understanding of static explanations.

- Perceived Usefulness [209, 210, 211]: The degree to which participants feel that the explanations enhance their experience with deep learning models. Along with *perceived ease of use* and *behavioral intention*, these three aspects measure participants' acceptance of static explanations. They are derived from the Technology Acceptance Model (TAM) [209, 210, 211], a widely applied theory for understanding individual acceptance and usage of information systems. As the explanations are used by end-users, investigating their acceptance of the explanations is very important.

- Perceived Ease of Use [209, 210, 211]: Participants' assessment of the simplicity and clarity of the explanations.

- Behavioral Intention [209, 210, 211]: The tendency of participants to utilize the explanation information in the future.

- Trust [219, 220]: Participants' confidence in the explanation methods keeping functioning as intended. Trust has been recognized as an important factor in human-AI collaboration as it mediates the human's reliance on AI models, thus directly affecting the effectiveness of the human-AI team [221, 222, 223, 224, 225].

  The literature demonstrated that static explanations have inconsistent effects on users' trust in AI systems. On one hand, several studies have demonstrated that

detailed explanations [226, 227, 225], contrastive explanations [228], and example-based explanations [229] can enhance user trust in systems. On the other hand, studies showed that static explanations do not have strong effects on user trust in AI systems [129, 230, 112, 111].

One main reason for these inconsistent reports is that trust is mediated by the users' understanding of the static explanations [230, 112, 111], and such understanding is often absent. According to theories of trust [231, 232, 208], the ability to build a mental model of AI systems is the key for user trust in AI. Unsurprisingly, studies on the effects of static explanations for laypersons show that users with limited knowledge of machine learning struggle to understand static explanations and the decision-making processes they are supposed to explain. Consequently, these users do not exhibit increased trust in AI systems after receiving static explanations [111, 112].

With this chapter, we quantitatively investigate whether customized conversations about static model explanations can enhance user understanding and improve trust. The conversational approach toward explanations has been advocated by previous studies [226, 233, 234, 139, 113] but never experimentally verified. For example, through interviews with decision-makers, Lakkaraju et al. [113] found that decision-makers strongly prefer conversational explanations that allow them to ask follow-up questions.

4.3.4 Detailed Study Procedure

Before participation, individuals are required to sign an informed consent form that outlines the objectives and procedures of the study. The form also clarifies compensation details and assures both the anonymity and confidentiality of data collected during the study. Upon signing the consent, participants receive an email that guides them to access the study platform.

After logging in, a pop-up prompt provides an overview of the tasks ahead. Participants then complete pre-experiment objective and subjective evaluations of the static explanations. The objective evaluation measures participants' understanding of static

explanations by letting them choose, from three classification models, the most accurate on unobserved test data. There are 5 explanation examples in the objective evaluation. The subjective evaluation, with 13 self-reporting questions, probes the perceived comprehension, acceptance, and trust towards the static explanations. Following these evaluations, participants in the experimental group engage in a WoZ discussion about static explanations. During the conversation, one example image is displayed on the screen. The example image is different from those used in the evaluations; however, the explanation methods remain the same. Participants are motivated to understand the explanations as they need to select the best-performing classification model using explanations only when doing objective evaluation. Our XAI experts faithfully answer the user's questions based on their knowledge, trying to help the user gradually understand the explanation. For participants in the control group, they read the static explanation for 15 minutes which is the average conversation time of the experimental group. After the conversation or 15-minute reading, participants complete the same set of evaluations as before. All evaluation outcomes and conversation records are documented. Upon study completion, each participant receives a $10 reward.

Table 4.3: Results of the experimental group before and after conversations, and the control group before and after 15-minute reading. Each score is presented as mean ± standard deviation and the change δ before and after. * $p < 0.001$

Explanation Methods	Group	Evaluation Timing	Objective Understanding (Decision-Making Accuracy)	Subjective Understanding	Perceived Usefulness	Perceived Ease of Use	Behavioral Intention	Trust
LIME	experimental	before	0.38 ± 0.20	4.03 ± 1.35	5.09 ± 1.07	4.48 ± 0.94	5.25 ± 0.95	4.15 ± 0.88
		after	0.53* ± 0.16	5.30* ± 0.88	5.92* ± 0.66	5.28* ± 0.84	5.83* ± 0.81	4.92* ± 0.73
	control	before	0.37 ± 0.17	4.57 ± 1.43	5.67 ± 0.95	4.87 ± 1.26	5.73 ± 0.69	4.37 ± 0.90
		after	0.40 ± 0.20	4.60 ± 1.16	5.33 ± 0.96	4.48 ± 1.26	5.27 ± 1.08	4.36 ± 1.05
Grad-CAM	experimental	before	0.82 ± 0.21	4.17 ± 0.91	5.49 ± 0.97	4.71 ± 0.95	5.52 ± 0.65	4.40 ± 1.00
		after	0.92* ± 0.11	5.43* ± 0.97	6.12* ± 0.60	5.58* ± 0.82	6.08* ± 0.79	5.19* ± 0.80
	control	before	0.81 ± 0.20	4.07 ± 1.34	5.58 ± 0.59	4.36 ± 1.15	5.45 ± 0.71	4.22 ± 0.96
		after	0.79 ± 0.19	4.40 ± 1.28	5.46 ± 0.69	4.70 ± 1.21	5.33 ± 0.83	4.42 ± 0.87

4.4 Results & Discussion

Table 4.3 tabulates the mean and standard deviation (SD) for all the measures. As explanation methods (LIME vs. Grad-CAM) and group (experimental vs. control) are

between-subjects variables and time (before vs. after) is a within-subject variable, we conduct a three-way Analysis of Variance (ANOVAs).

4.4.1 Effects of explanations on objective decision accuracy and subjective measures

Results show significant main effects of group ($F(1, 116) = 5.60, p = .02$), method ($F(1, 116) = 218, p < .001$) and time ($F(1, 116) = 12.51, p < .001$). The experimental group, the Grad-CAM method, and the after-conversation condition display a higher objective decision accuracy. We also find a significant interaction effect between group and time ($F(1, 116) = 11.3, p = .01$), as displayed in the figure 4.4. In the participant's initial decision, there were no significant differences between the experimental and control conditions. During participants' final decision, those who interact with the XAI expert (i.e., experimental condition) have better decision accuracy. This phenomenon highlights the effectiveness of conversational explanations in enhancing the objective understanding of static explanations of users.

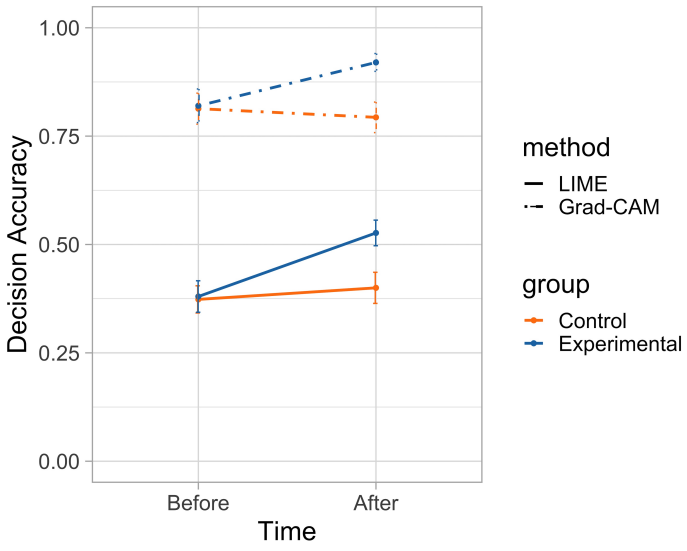
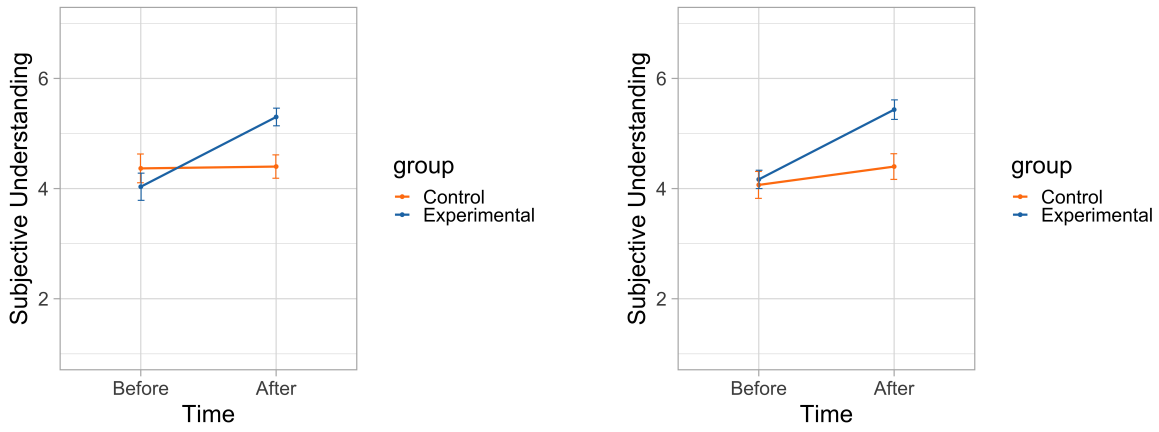


Figure 4.4: Objective decision accuracy for different groups before and after conditions.

We observe varied objective performance between LIME and Grad-CAM ($F(1, 116) = 218, p < .001$). Grad-CAM has a higher accuracy of objective decision accuracy com-

pared to LIME. A potential reason might be the inherently intuitive nature of the explanations produced by Grad-CAM compared to LIME.



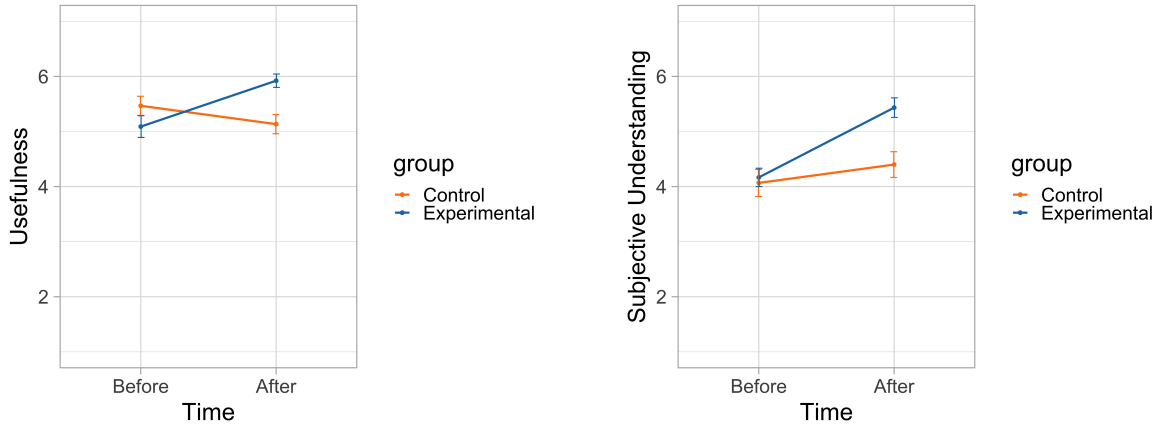
(a) LIME (b) Grad-CAM

Figure 4.5: Subjective understanding score for (a) LIME and (b) Grad-CAM before and after conditions.

In terms of participants' subjective understanding, we find a significant main effect of the evaluation timing ($F(1, 116) = 4.08, p < .001$). Participants receiving conversational explanations have a significantly larger improvement in subjective understanding. We also observe a significant interaction effect between group and time ($F(1, 116) = 37.3, p < .001$), shown in figure 4.5. Initially, there is no significant difference in the participants' self-reported understanding of static explanations between the experimental and control groups. After the conditions, participants in the experimental group demonstrate a higher self-report understanding compared to those in the control group.

The main effect of the explanation method ($F(1, 116) = .72, p = .40$) is not significant for participants' subjective understanding, contrasting with its effect on objective understanding. Even though participants can intuitively choose the best classification model based on the heatmap in the objective evaluation, participants' initial self-reporting understanding score of Grad-CAM is just slightly larger than 4 (average understanding). This shows that participants still feel confused about how Grad-CAM works and how it explains models' predictions, even though they can perform well in the objective

evaluation. This also demonstrates that subjective and objective evaluations measure participants' understanding of static explanations from complementary aspects. Self-reporting scores can be influenced by personal biases, while the objective evaluation might not capture users' feelings about understanding. Combining both methods can provide a comprehensive view of participants' understanding of static explanations.

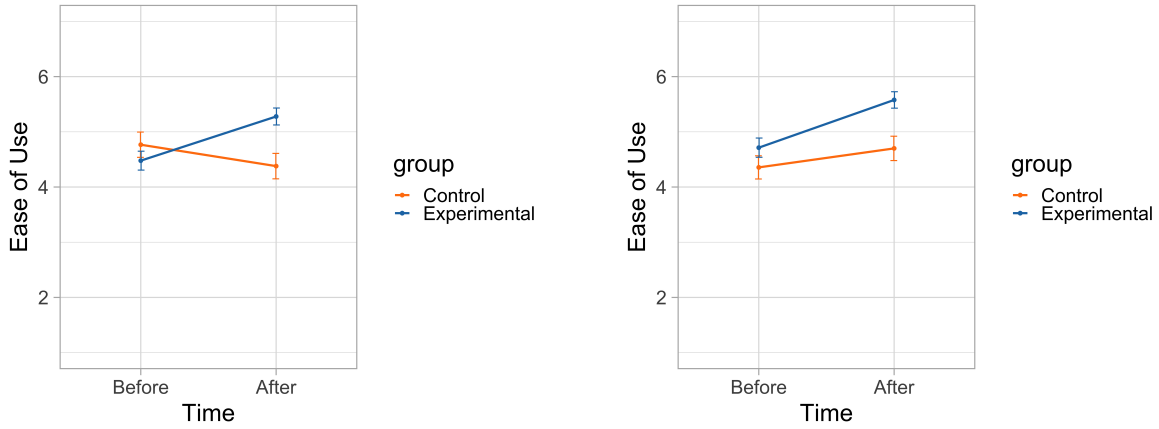


(a) LIME (b) Grad-CAM

Figure 4.6: Participants' self-report usefulness score for (a) LIME and (b) Grad-CAM before and after conditions.

For the perceived usefulness, results show a significant main effect of time ($F(1, 116) = 14.6, p < .001$), as well as a significant interaction effect between group and time ($F(1, 116) = 52.9, p < .001$), as depicted in figure 4.6. The experiment group (i.e., receiving conversational explanation) results in a larger increment of perceived usefulness. For the control group, Grad-CAM increases perceived ease of use when participants are given more time to view the static explanation. However, a reversed trend is observed for the LIME method in the control group – the perceived ease of use drops after additional time is provided.

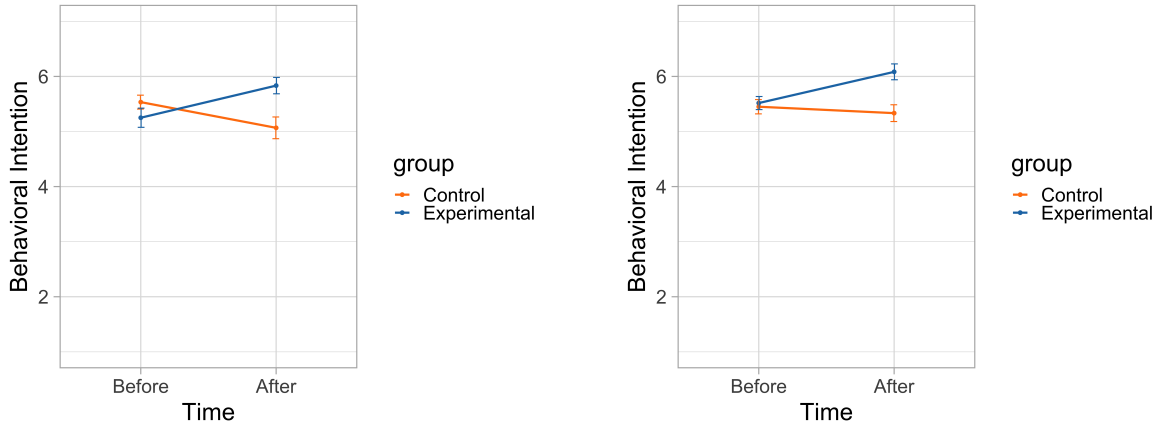
Similar results are observed for participants' perceived ease of use. There are significant main effects of group ($F(1, 116) = 5.19, p = .002$) and of time ($F(1, 116) = 30.3, p < .001$), as well as a significant interaction effect between group and time ($F(1, 116) = 33.7, p < .001$). The perceived ease of use increases largely for the experiment group after interacting with XAI experts. For the control group, the Grad-CAM



(a) LIME (b) Grad-CAM

Figure 4.7: Participants' self-report ease of use score for (a) LIME and (b) Grad-CAM before and after conditions.

method increases perceived ease of use while LIME methods decrease it when giving participants more time to view the static explanation.



(a) LIME (b) Grad-CAM

Figure 4.8: Participants' self-report behavioral intention score for (a) LIME and (b) Grad-CAM before and after conditions.

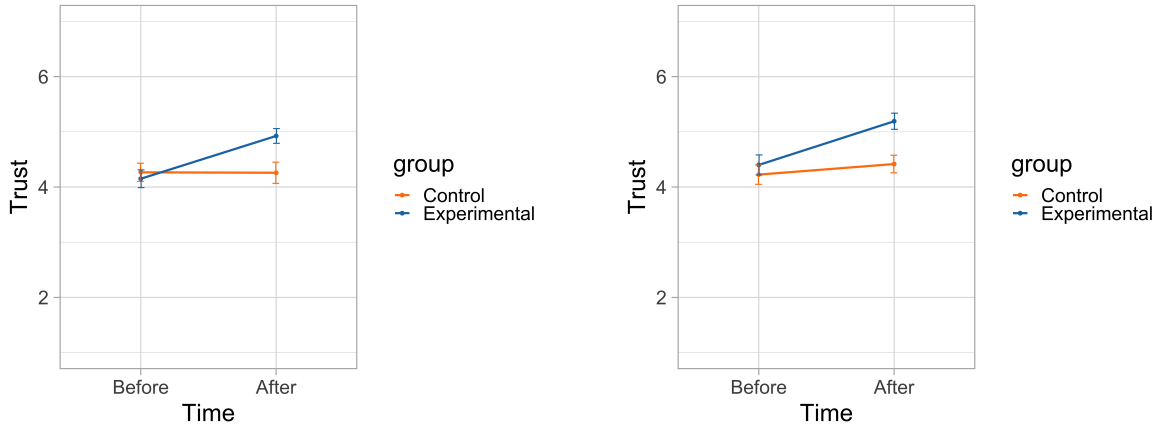
For the behavioral intention, results show a significant main effect of the time ($F(1, 116) = 3.92, p = .005$) and a significant interaction effect between group and time ($F(1, 116) = 3.92, p < .001$) as shown in figure 4.8. Participants increase their behavioral intention and are more inclined to use explanations in future scenarios after

receiving conversational explanations. On the contrary, the behavioral intention of the control group decrease for both Grad-CAM and LIME.

The boost in usefulness, ease of use, and behavioral intention for the experimental group can be attributed to the increased understanding of static explanations. Prior to the expert interactions, participants might have had limited knowledge or even misconceptions about the explanation methods. Experiment results show that participants gain a clearer understanding of how the XAI methods function, after the participants' questions are addressed in the conversations. Consequently, they report perceiving the static explanations as more useful and easier to use, and report higher inclination to use the static explanations in future tasks.

The perceived usefulness, ease of use, and behavioral intention of the control group all decrease after reading static explanations for a longer time. This trend suggests a decreased willingness to utilize explanations in future scenarios. The reluctance may be attributed to the frustration the control group faced in attempting to comprehend the static explanations on their own. Research by Carolin Ebermann and Weibelzahl [235] on the impact of cognitive fit and misfit in the acceptance of AI system usage highlights this phenomenon. They found that users experiencing a cognitive misfit with the AI system often report negative moods, which in turn, reduce their perceived usefulness, ease of use, and behavioral intention of the AI systems. The contrary results of the control group and the experimental group also underscore the importance and effectiveness of conversations in enhancing user behavioral intentions of static explanations.

For the trust, results show significant main effects of group ($F(1, 116) = 4.31$, $p = .04$) and time ($F(1, 116) = 70.0, p < .001$). The experimental group and the after condition display a higher trust score of participants. We also find a significant interaction effect between group and time ($F(1, 116) = 43.7, p < .001$), as displayed in the figure 4.9. Initially, there were no significant differences in trust scores between the experimental and control conditions. During participants' final decision, those who interact with the XAI expert (i.e., experimental condition) report a higher trust score. The enhancements of the experimental group, contrasted with the unchanged trust score of the control group indicate that informativeness and clarity through conversations can help static explanations gain more trust from users. While there exist numerous studies on how



(a) LIME (b) Grad-CAM

Figure 4.9: Participants' trust for (a) LIME and (b) Grad-CAM before and after conditions.

explanations of AI predictions can influence users' trust in AI predictions [129, 236, 230, 111, 116, 182], to our knowledge, this is the first experiment designed explicitly to gauge the impact of conversations on enhancing participants' trust in explanations.

4.4.2 Analysis of Collected Conversations

We collect 60 free-form conversations between XAI experts and participants from 4 different discipline groups. On average, each conversation had 27.4 turns, with each turn comprising approximately 14.4 tokens. By analyzing the users' questions, we divide them into six categories:

- Basic concepts in machine learning: Questions about basic terms and concepts in machine learning that lay people may not know, e.g., what is a deep learning model, what is accuracy, the model structure, and the training data, etc.

- Application and performance of machine learning models: Questions about the ability, accuracy, and limitations of machine learning.

- Diagram reading: Questions about the explanation diagram generated by Grad-CAM or LIME, e.g., what different colors represent in the heatmap.

- Basic concepts in explainable AI: Questions about basic concepts of explanation methods, e.g., what are explanation methods?
- Mechanism of explanation methods: Questions about how explanation methods work and how the provided explanation is generated.
- Other explanations: Questions that require the generation of other types of explanations on the current predictions, explanations for different predictions, or comparisons between the provided explanation and other explanation methods.

Based on this categorization, we build a repository for questions that could occur in the conversations. In total, we collected 397 questions from the six different categories. Table 4.4 contains examples and the number of questions in each category. As observed in Table 4.4, the questions of participants mainly revolve around basic concepts in machine learning, the fundamentals of explanation methods, and their underlying mechanisms. This trend might be attributed to the multi-disciplinarity of the participants. It suggests that many participants may not be familiar with machine learning models and explanation methods, which is aligned with the real application of explanation methods. Therefore, it's crucial to tailor responses to these questions to help users better understand explanations. Furthermore, we note a marked interest in new explanations. This could indicate that as users become more familiar with provided explanation examples, they exhibit curiosity about alternative explanation methods and how models might behave under specific scenarios. Concurrently, the diagram reading category contains only 16 questions, implying that explanations generated by Grad-CAM and LIME were relatively straightforward and easy to understand. The diverse range of questions sourced from our conversations underscores that static, one-off explanations are often insufficient for users to understand them. Engaging in dialogue can provide more dynamic and tailored explanations to users, hence deepening their understanding of static explanations.

Having well internalized their knowledge, experts are often unable to estimate what laypeople know [121]. This phenomenon is also referred to as the "curse of knowledge" [237]. As a result, experts tend to overlook potential areas of confusion or make unwarranted assumptions about what is "common knowledge". While analyzing the

Table 4.4: Overview of Collected Questions. Including categories of questions, examples, and the count of questions in each category.

Question Category	Question Examples	Num
Basic concepts in machine learning	• What is a deep learning model? • What is the image classification task? • How does the model know what features to extract?	85
Application, performance, and limitations of machine learning models	• How about the precision of the classification model? • Where has this Swin Transformer classification method been used in practical applications? • Will the different species of an animal affect the classification model categorizing the animal?	68
Diagram reading	• Are regions colored in red areas that have been identified as containing key features for the animal? • What are the yellow line spots for (in LIME explanations)? • What do the red and blue colors mean (in Grad-CAM explanations)?	16
Basic concepts of explanation methods	• What is the explanation model used for? • Can LIME be used without the internet? • What are some limitations of the Grad-CAM (LIME) method?	95
Mechanism of explanation methods	• Why does the (LIME) explanation not highlight all the parts of the leopard? • How LIME model recognize the most important parts for the model prediction? • Seems like the Classification Model and the Explanation Model are trained separately - how can we be sure that the underlying logic of making a prediction is the same for both models?	91
Other explanations	• Can you list other visualization methods? • Is there anything special about the Grad-CAM (or LIME) method that is different from others? • What if there are both fishes and humans in an image? How should this image be classified, and can you provide such explanations?	42

collected conversations, we often find ourselves unable to anticipate the user questions, which corroborates the literature. We describe a few examples below.

Several participants misunderstood the idea of the heatmap produced by Grad-CAM as depicting literal heat dissipating from objects. They infer that the model uses the temperature of objects to perform classification. In reality, a heatmap is just a metaphor that visualizes numerical values distributed spatially, which refers to the feature importance in our case. This misconception leads to questions about how the heat of objects is measured and why non-living objects are warmer than their environment. Some example utterances from participants include: "*So the Grad-cam method basically just refers to the usage of generating a heatmap to capture living matters correct? ... based on the parts of the image that generate more heat?*" –P36, "*basically using heat to predict what is the input right?...how will we know what is the animal or input simply based on heat?*" – P47, "*if these are pictures, how do they figure out the heat since the animal isn't generating heat*" – P49, "*So a heat sensor is not required? A heatmap is automatically generated from each photo and analyzed using the model.*" – P52.

A second common misconception is the conflation between the post-hoc explanation technique and the classification models. Some example user questions include: "*is the explanation method what the model uses to classify & predict what the image is supposed to be?*" – P6, "*Swin transformer uses LIME model? ... what are the differences between lime model and Swin transformer?*" – P8. Furthermore, participants face challenges in understanding certain terms commonly used in AI and XAI, even though these terms are frequently used and understood within academic communities. Many participants asked questions about basic concepts in machine learning, such as: "*what is the explanation method?*" – P7, "*how do you classify the image?*" – P17, "*what is the algorithm? does it mean lime? what are deep neural networks?*" – P32, "*How would you explain the term 'perturbations of images' to a five-year-old?*" – P46.

The observations from the interactions between XAI experts and layperson users demonstrate the importance of conversations for users to understand static explanations as they bridge the knowledge gap between the two groups. Conversations can reveal the specific areas of misunderstanding, such as incorrect implicit assumptions the users make and knowledge they lack. Hence, conversational explanations may help the AI system communicate with and bring genuine understanding to the users.

4.4.3 Implications for building dialogue systems to explain static explanations

Our study indicates the impact of conversational explanations on user comprehension, acceptance, and trust of static explanations. Static explanations, while informative, may not cater to users with varied backgrounds and expertise. Engaging in conversational explanations provides a dynamic and interactive medium for users to seek clarifications, ask questions, and thereby facilitate a deeper and more personalized understanding.

The emergence of advanced conversational agents [140, 141, 142], especially knowledge-based question-answering [143, 144, 145] powered by large language models [146, 147, 148] paves the way toward conversational agents that can explain model decisions and discuss static explanations. Our study suggests the following desiderata for such agents.

- *Extensive knowledge of AI and XAI.* As observed in our study, a large portion of user questions are related to core concepts of machine learning models and explanation methods. To answer those questions, conversational agents need to be trained on a comprehensive corpus encompassing AI and XAI concepts. Besides, in our study, participants also are curious about the applications, performances, and limitations of machine learning models and explanation methods. Therefore, besides answering abstract questions, dialogue systems also should relate them to real-world applications and limitations.

- *Capability to generate new explanations as needed.* As an improved understanding of the provided explanations, participants in our study exhibit curiosity about alternative explanation methods and explaining different predictions. Dialogue systems should provide new explanations to users when requested. For instance, if a user is curious about how changing a feature would affect the model output, the system should generate a new explanation with the new feature, which showcases the effect.

- *Capability to interpret scientific diagrams and visualizations.* A significant portion of AI and XAI explanations often comes in the form of diagrams [154, 155], such as heatmaps or feature importance visualizations. Our study reveals that users

have questions related to understanding these diagrams. Answering these questions usually requires an understanding of specific regions of the diagrams, such as answering what parts of the object are highlighted by the yellow line in LIME explanations. Therefore, future dialogue systems should have visual processing capabilities, understanding and interpreting diagrams contextually. For instance, they should be able to recognize colors, patterns, and other graphical elements in heatmaps or charts and relate them to users' questions. The recent development in multimodal large language models [238, 239, 240] is a promising direction to achieve this goal.

4.5 Limitations

Despite the insights gained, there are several limitations that should be acknowledged. First, the static explanations used in our study are limited. Our experiments focused on feature attribution explanation methods. The applicability of our findings to other explanation methods, such as example-based explanation methods, remains an open question. Second, as our main objective was to discern the effects of free-form conversational explanations, we did not delve into the comparative performance of different explanation methods. In our experiments, we intentionally selected explanation examples where the best classification model yielded the most reasonable explanations. The explanation examples discussed by participants and XAI experts were chosen such that they reasonably explain the predictions of the classification model. Future work would be to extend these conversations to include explanations that might be less reliable. Third, we explore how conversations foster user trust in explanations in our study. Nevertheless, previous studies [112, 111, 227] have shown that humans may trust AI models even if they make wrong decisions. We do not explore whether users' trust in our study is misplaced, which we leave for future work. Fourth, we use AI to classify the images. Previous studies [241, 242] found that participants favor humans over AI decision-makers when their decisions directly affect participant welfare. In our study, AI decisions do not directly affect participant welfare. We also did not investigate if the participants preferred conversations with humans or AI chatbots or if their trust

in the explanations was affected by that variable. Finally, our research is confined to one geographical region and includes only students and staff from the university. Factors such as cultural backgrounds and age-related differences could potentially influence user interactions with XAI and how they seek to clarify confusion. Future studies could involve recruiting participants from diverse countries, regions, and age groups.

4.6 Summary

In this chapter, we conduct Wizard-of-Oz experiments to investigate how free-form conversations assist users in understanding static explanations, promoting trust, and making informed decisions about AI models. Participants engage in conversational explanations with XAI experts to understand how the provided static explanation explains the model decision. To evaluate the effects of conversations, we design objective and subjective measurements. We observe a notable improvement in users' comprehension, acceptance, trust, and collaboration after conversations. From collected conversations, we find that participants' questions and confusions are diverse and unanticipated. Based on the findings in this chapter, we develop a dialogue system in the subsequent chapter to provide personalized explanations to laypeople.

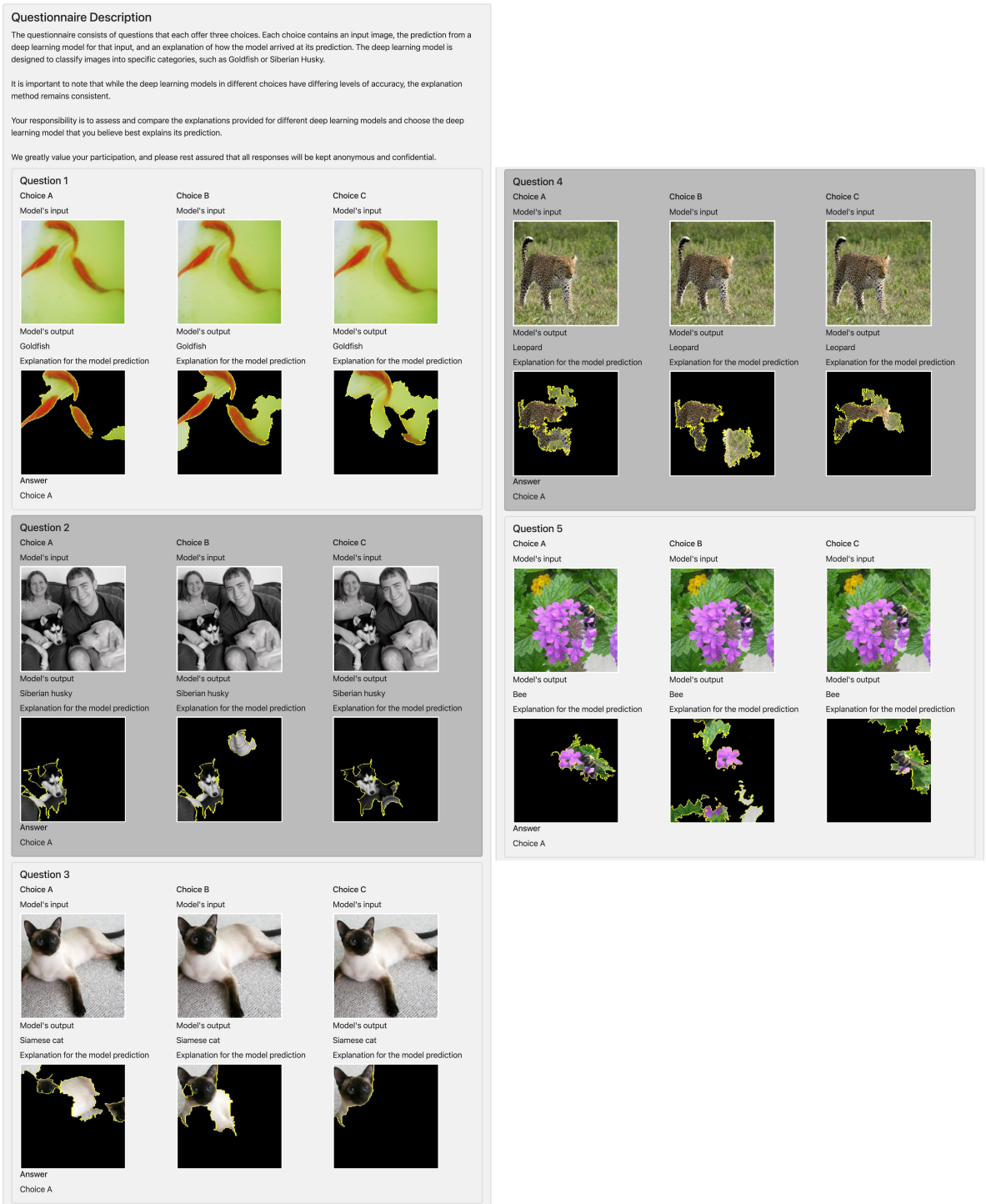


Figure 4.10: Objective evaluation questions used for LIME.

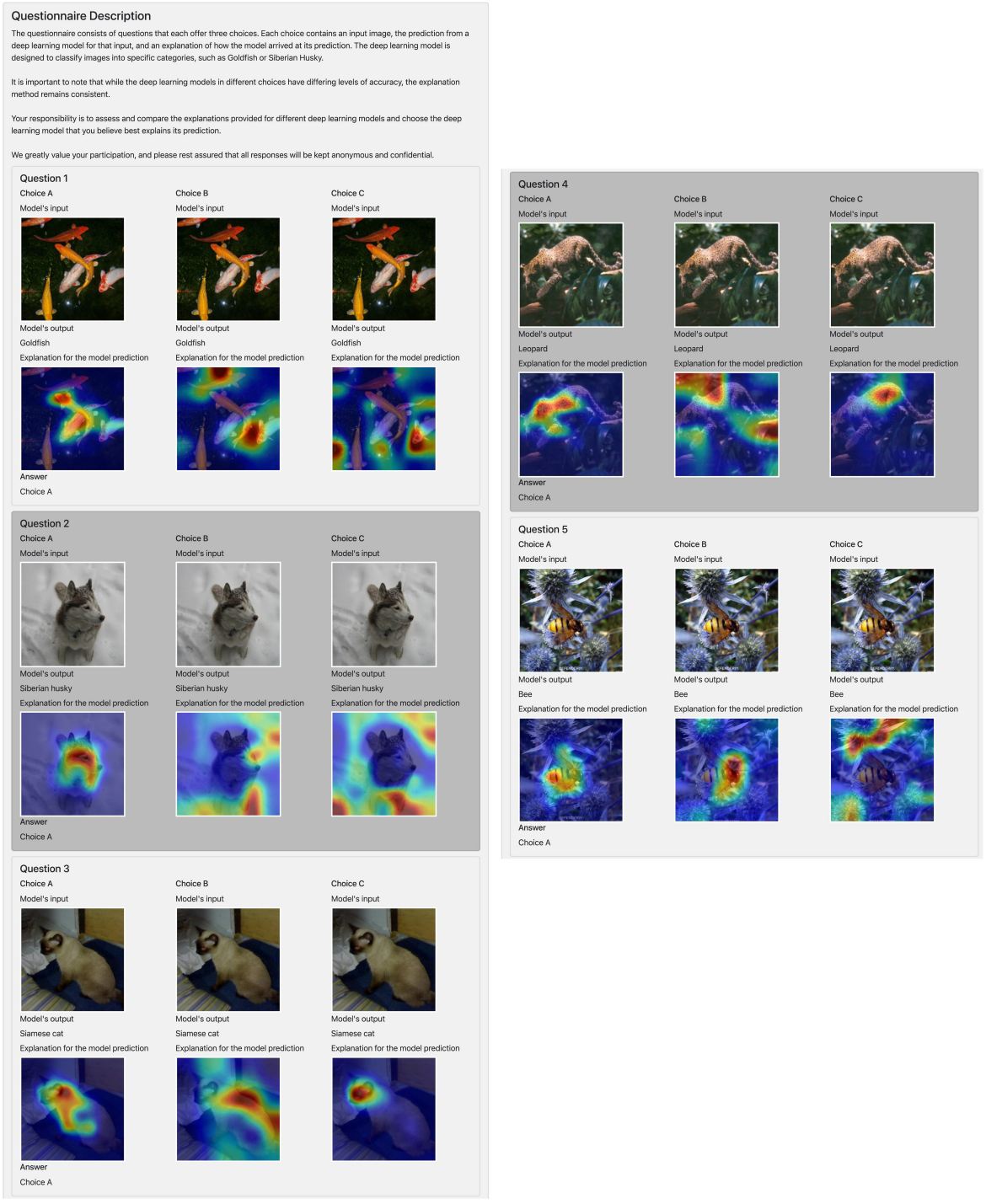


Figure 4.11: Objective evaluation questions used for Grad-CAM.

Chapter 5

Tailoring Explainable AI to Laypersons Through Conversations

5.1 Overview

As demonstrated in the last chapter, conversations with XAI experts can significantly improve the effectiveness of Explainable AI. In this chapter, we propose to build a conversational explanation system that provides personalized responses and information based on users' conversational histories.

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 [243]. To date, existing conversational explanations rely on human-authored templates, which can only handle a limited and predefined range of user questions [115, 23].

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 [244, 245, 246], and the hallucinations generated by VLMs [247, 248, 249, 250, 251]. The first challenge, lack of data diversity, arises as generative models tend to overrepresent high-frequency content [244, 245, 246] 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 [247, 248, 249, 250, 251]. 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 [243]. 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 [246], 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.

5.2 Related Work

5.2.1 Static XAI

Explainable Artificial Intelligence (XAI) refers to techniques that explain the learning process or the predictions of AI [102]. 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 [151, 152, 153, 18, 252]. The majority of recent XAI methods are post-hoc XAI methods, applied to already developed models that lack inherent transparency [154, 155, 156, 158, 96]. 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 [158, 103]. There are two main types of feature attribution methods, gradient-based methods [160, 159, 154, 161, 162, 253, 254, 255, 253] and surrogate methods [155, 163, 164, 165, 166, 167]. 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 [158, 103, 256]. These instances may be training data points with the most influence on the parameters of a prediction model [156, 257], counterfactual examples that alter predictions with minimal changes to inputs [174, 169, 258, 259, 260, 261], or prototypes that contain semantically similar parts to input instances [177, 171, 181].

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 [182]. 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.

5.2.2 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 [190] argues that explanations emerge from a conversational interaction between an explainer and an explainee. Similarly, Miller [123] 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 [114, 139, 113].

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. [23] applied conversational explanations to scientific writing tasks, finding improvements in productivity and sentence quality. Likewise, Slack et al. [115] 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.

5.2.3 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 [262, 263, 264, 265, 266, 267]. For example, SuperGen [262] uses LLMs conditioned on label-descriptive prompts to generate training data for text classification tasks. FewGen [266] finetune 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 [244, 268, 269, 270, 247, 248].

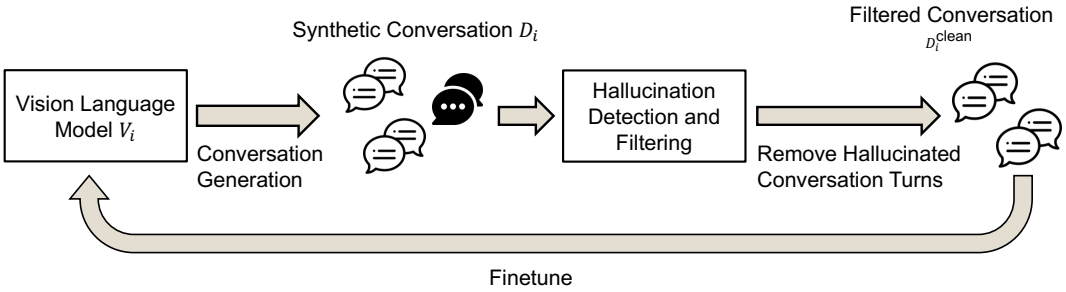


Figure 5.1: The Overall Workflow of EMCEE. V_i denotes the VLM and D_i denotes the synthetic conversation data in the i-th iteration. Starting from a pretrained VLM V_1, we first generate diverse synthetic conversations D_1 with the repetition penalty. Next, we use a hallucination detector to clean synthetic data, producing cleaned data D_1^{clean}. We then finetune the VLM on D_1^{clean}, which creates V_2, and this process repeats.

To mitigate the detrimental effects of noisy and low-quality synthetic data from LLMs and VLMs, several methods have been proposed [265, 264, 266, 267]. For example, ProGen [267] adjusts the weight of generated data points with regard to its influence on the validation loss, using influence function [271]. 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 [272].

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 §5.5.1.3).

5.3 Methodology

The overall workflow of EMCEE is illustrated as Figure 5.1 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 $\boldsymbol{D}_1^{\text{clean}}$. Afterwards, we finetuned the VLM on $\boldsymbol{D}_1^{\text{clean}}$, leading to the next VLM $\boldsymbol{V}_2$, 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.

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

5.3.1 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,

$$p_i = \frac{\exp(z_i/(T + \theta \cdot \mathbb{1}(i \in G)))}{\sum_j \exp(z_j/(T + \theta \cdot \mathbb{1}(j \in G)))}, \tag{5.1}$$

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.

5.3.2 Hallucination Detection and Filtering

VLMs often generate convincing but factually incorrect statements, especially when answering questions that require reasoning and logical deduction [247, 248, 249, 250, 251].

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:

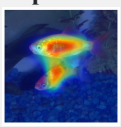
Instruction: A chat about explainable AI (XAI) between a curious human USER and an AI ASSISTANT. The human USER is well educated but may need help understanding how AI and XAI work. The USER asks questions to understand AI's decision-making process better. The USER's question should be diverse and related 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 5.2: The VLM prompt for LIME and 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:

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.
<**Demonstrations**>
The conversation starts:
USER:

Figure 5.3: The VLM prompt for Integrated Gradients and SHAP.

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** $1...R$ **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^{clean};
end for

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 5.1. We used 80% of the dataset for training the detector, with the remaining 20% reserved for validation and testing.

5.3.3 Implementation

We used LLaVa-1.5 [273, 274] 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

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 5.1: 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.

performance in answering scientific questions and proficiency in visual chat scenarios [273, 274].

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 [275] 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 [17] 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.

5.4 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.

5.4.1 Automatic Evaluation Metrics and Dataset

For automatic evaluations, we conducted few-shot evaluations with zero to three demonstrations. We leverage BLEU [92] and ROUGE [93] 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. [243]. This dataset was gathered using a Wizard-of-Oz (WoZ) setting [215]. 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. The remaining 52 conversations were reserved for testing.

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.

5.4.2 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).

5.4.2.1 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

Table 5.2: 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

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 5.2 shows statistics of the academic disciplines that the participants enrolled in.

5.4.2.2 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 [155], Grad-CAM [154], Integrated Gradients [159], and SHAP [162]. 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 [182].

5.4.2.3 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 5.4 displays an example screenshot of the interface. There are two sections on the page. On the left (Figure 5.4 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 5.4 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.

5.4.2.4 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 groups: 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 categories of dependent variables are collected in the experiment: the model selection accuracy and the self-reported perception scores.

As indicated in previous work [124], participants' prior knowledge of AI may influence their perceptions of explanations, introducing a potential confounding factor. To mitigate the effects of such factors, we randomly assigned participants to one of the three groups in the study. Additionally, the results in Section 5.5.2 showed no significant difference ($p = 0.44$) in participants' pre-conversation self-reported understanding across the different groups.

5.4.2.5 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 [149]. 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 shown in Figure 5.12, 5.13, 5.14, and 5.15 respectively.

We observe that static explanations do not always faithfully reflect the actual workings of classification models [216, 217, 218] 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.

There are also other objective measurements to assess users' understanding in previous work [112, 129]. These methods measure users' accuracy on the same prediction tasks as the classifier, such as student admission [129] or recidivism prediction [112]. However, these measurements are not suitable for our study. In our case, XAI methods are applied on top of image classification models. Users can easily identify an image's class without relying on explanations. As a result, it is not feasible to evaluate users' understanding of XAI by simply asking them to predict an image's class with and without

explanations, as was done in previous work [112, 129]. 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. These images were chosen so that users could only make the correct choice if they comprehended the explanations provided.

5.4.2.6 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 Table 4.2.

- Comprehension [208, 129]: 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 [209, 210, 211]: 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) [209, 210, 211], 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 [209, 210, 211]: Participants' judgment of the simplicity and clarity of the explanations.

- Behavioral Intention [209, 210, 211]: The tendency of participants to utilize the explanation information in the future.

- Trust [219, 220]: 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 [221, 222, 223, 224, 225, 213, 276].

5.4.2.7 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 5.4. 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.

5.5 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.

Table 5.3: 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

5.5.1 Results of Automatic Evaluation

5.5.1.1 Comparison of Baseline and Our Method

Table 5.3 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 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.

Table 5.4: 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

5.5.1.2 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 5.4 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 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.

5.5.1.3 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 5.5.

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. [246], 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.

5.5.2 Results of Human Evaluation

Table 5.5 presents the human evaluation results, comparing the LLaVa-1.5 model, EMCEE, and the control group across four explanation methods: LIME, Grad-CAM, Integrated Gradients, and SHAP. 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.

5.5.2.1 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

Table 5.5: 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

group and time ($F(2, 48) = 5.44, p = .007$), as displayed in the Figure 5.6. 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 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 [243].

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 ($F(1, 116) = 10.16, p < .001$). 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 ($F(1, 48) = 21.35, p < .001$), as well as a significant interaction between group and time ($F(2, 48) = 15.37, p < .001$), as depicted in Figure 5.8. 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 Figure 5.9. 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 Figure 5.10. 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 [243]. Experiment results showed that participants gained a clearer understanding of how the XAI methods function, after the participants' questions 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 5.11. 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 [231, 232, 208], 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 5.5, one observation to note is that the 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 differences across the entire user

Table 5.6: 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

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.

5.5.2.2 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 5.6 provides examples and the number of questions in each category. As observed in Table 5.6, 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

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 5.7: 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 [1]. Our annotations obtain "moderate agreement" for Understandability (0.57) and "substantial agreement" for Factual Correctness (0.675).

understandability measures whether the responses are easy to comprehend. Each criterion is rated as either 0 or 1. Table 5.7 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 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 5.8: Examples of conversation turns that are identities as hallucinations by the detector.

5.5.2.3 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 5.8. 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.

5.5.2.4 Addressing Key Requirements for Conversational Explanation Systems

As discussed in Chapter 4.4.3, we identified three key requirements for building effective conversational explanation systems: (1) extensive knowledge of AI and XAI, (2) the ability to generate new explanations based on user input, and (3) the capability to interpret scientific diagrams and visualizations. The proposed EMCEE model in this chapter meets the first and third requirements. EMCEE is built on LLaVa, which is pretrained on a broad range of knowledge, including AI-related content. It is further trained on synthetic conversations covering various AI and XAI topics, enabling it to answer both general and explanation-specific questions. The third requirement is also fulfilled: EMCEE is trained on synthetic conversations designed to help users understand visual elements in explanation diagrams, such as heat maps and saliency maps. Combined with LLaVa's vision-language pretraining, EMCEE is capable of interpreting scientific diagrams and visualizations for users. The second requirement, however, remains unmet. While EMCEE can clarify and elaborate on existing explanations, it cannot generate new explanations or simulate model behavior under different conditions.

5.6 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 [277]. 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 [124] 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.

5.7 Summary

In this chapter, we propose 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 recycle the generation-filter-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 interaction experience.

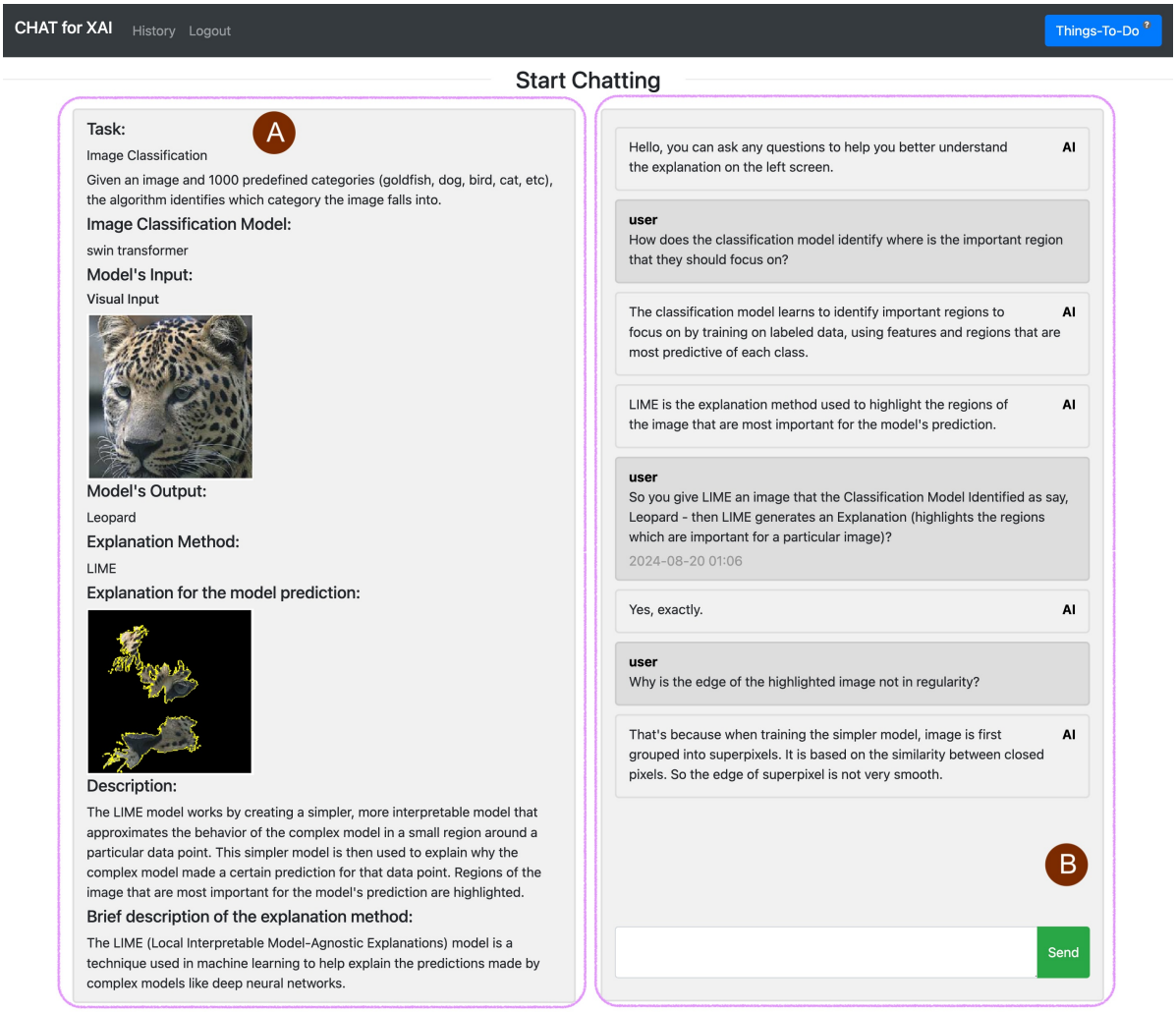


Figure 5.4: The interface where users discuss static explanations with a conversational agent. *Part A*: Information about static explanations, including a task description, a description of the prediction model, a model input, a model output, an explanation generated by the explanation model, and a description of the explanation. *Part B*: A chatbox where users converse with a conversational agent to clarify the explanation.

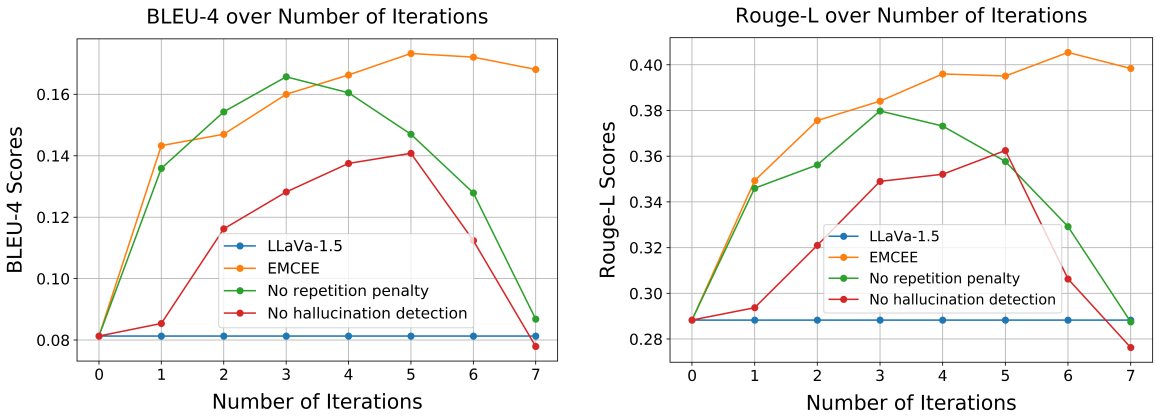
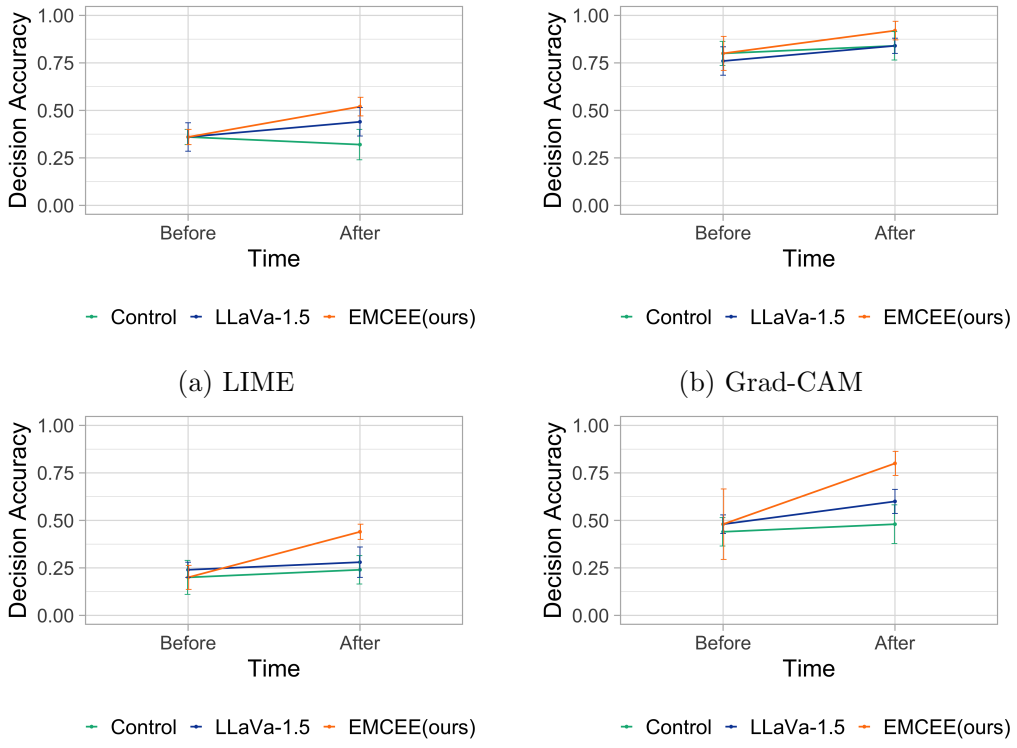


Figure 5.5: BLEU-4 and Rouge-L scores over the number of training iterations for LLaVa-1.5, EMCEE and different ablated version of EMCEE.



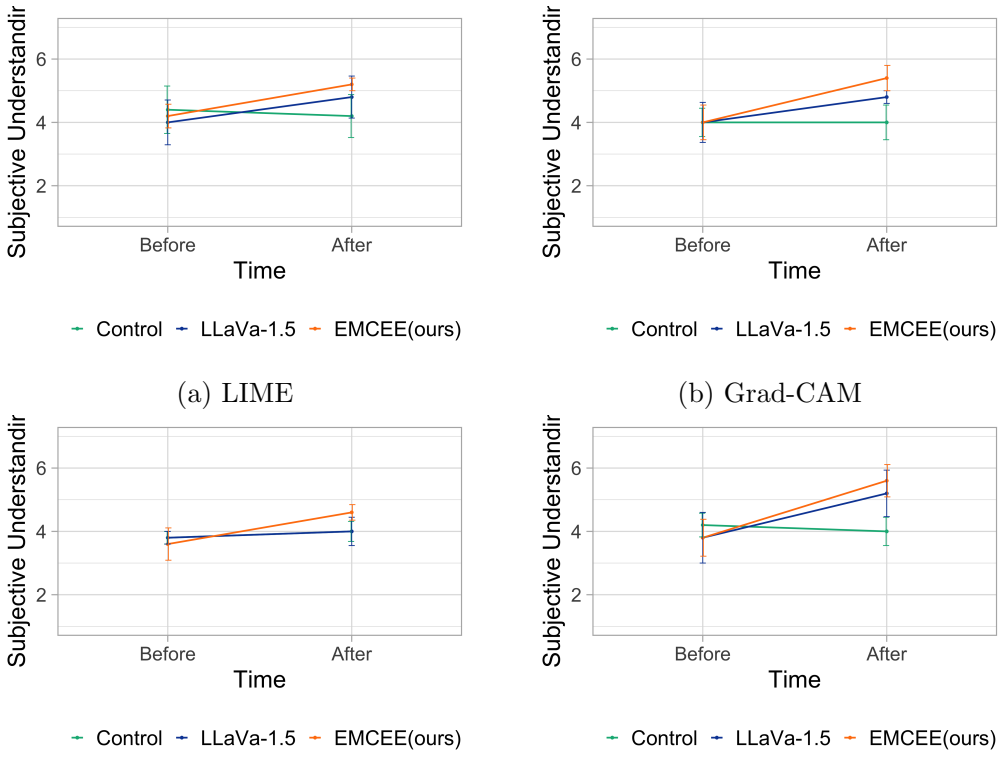
(a) LIME

(b) Grad-CAM

(c) Integrated Gradients

(d) SHAP

Figure 5.6: Model selection accuracy for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.



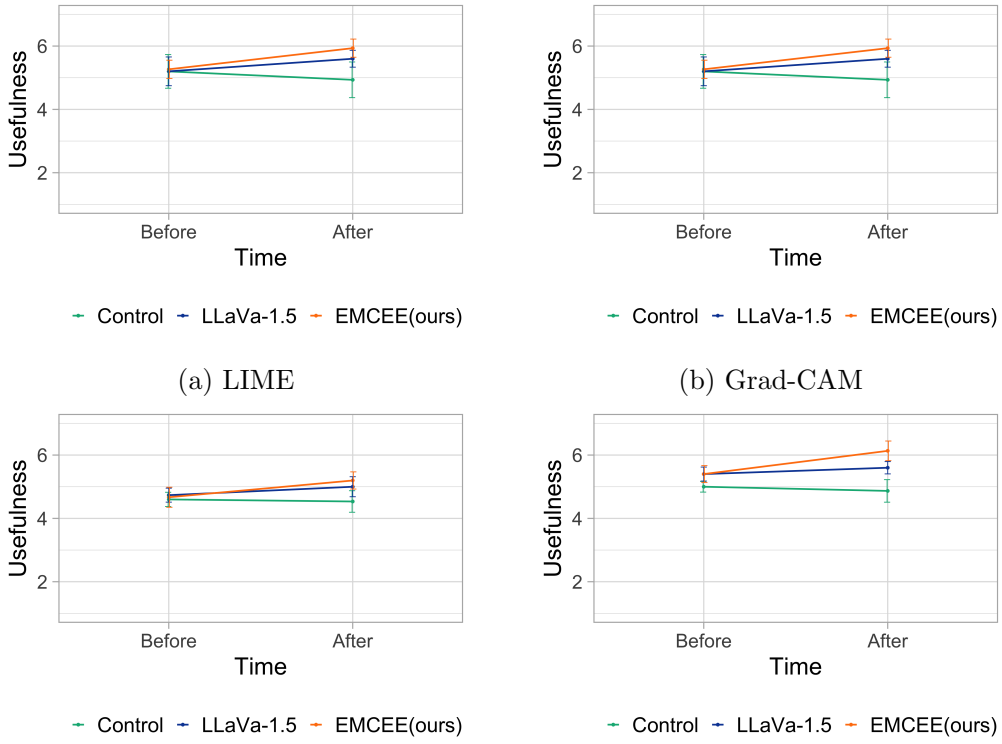
(a) LIME

(b) Grad-CAM

(c) Integrated Gradients

(d) SHAP

Figure 5.7: Subjective understanding score for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.



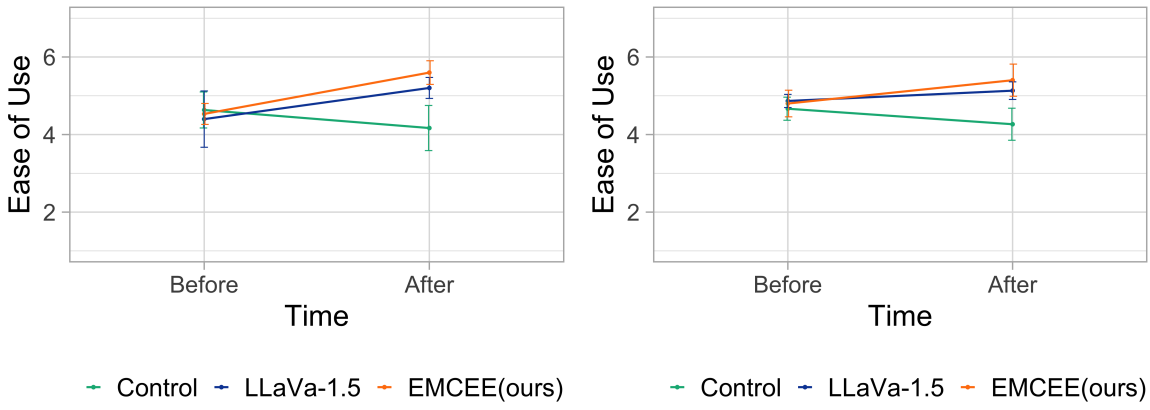
(a) LIME

(b) Grad-CAM

(c) Integrated Gradients

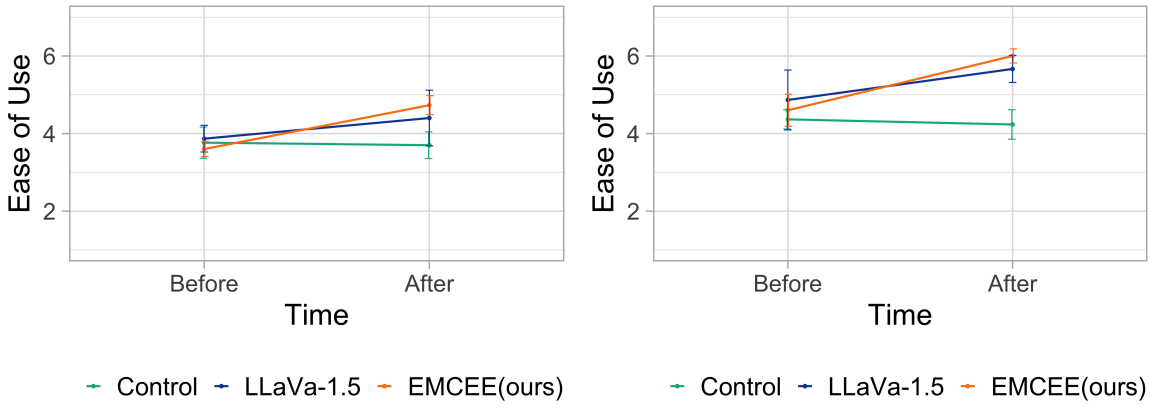
(d) SHAP

Figure 5.8: Participants' self-report usefulness score for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.



(a) LIME

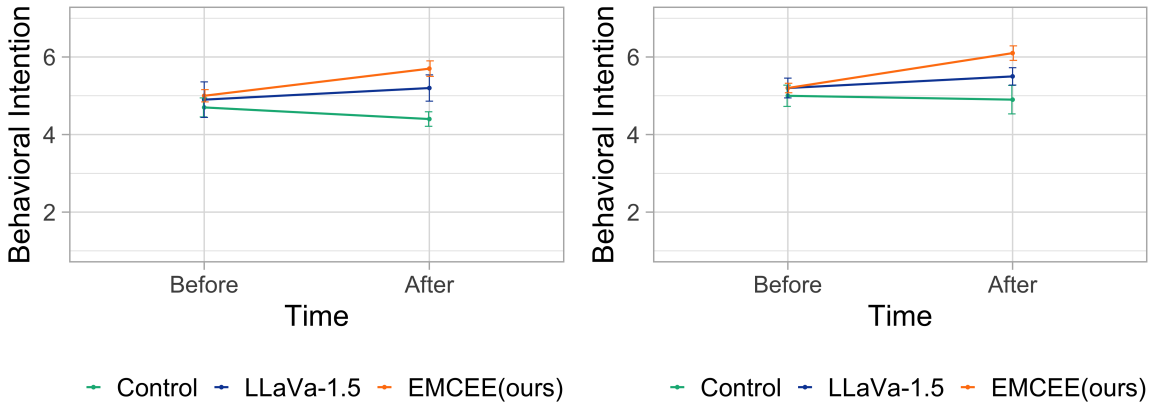
(b) Grad-CAM



(c) Integrated Gradients

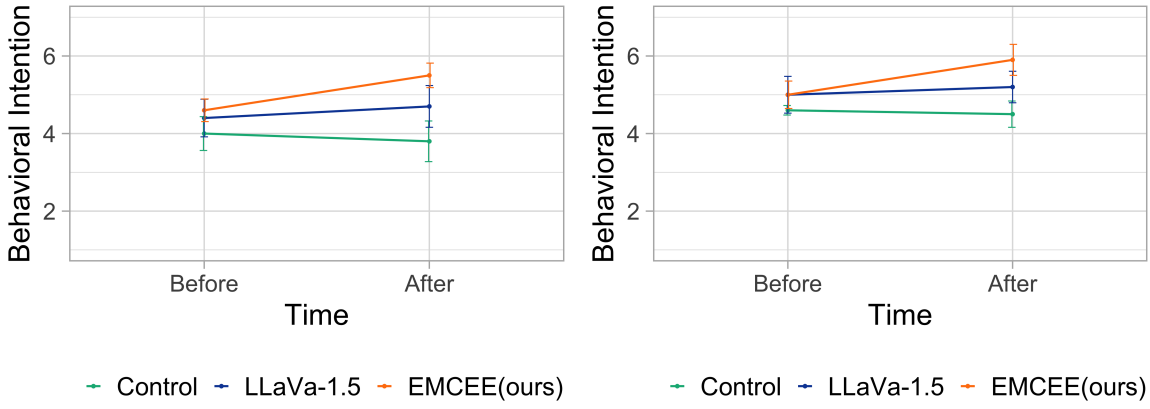
(d) SHAP

Figure 5.9: Participants' self-report ease of use score for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.



(a) LIME

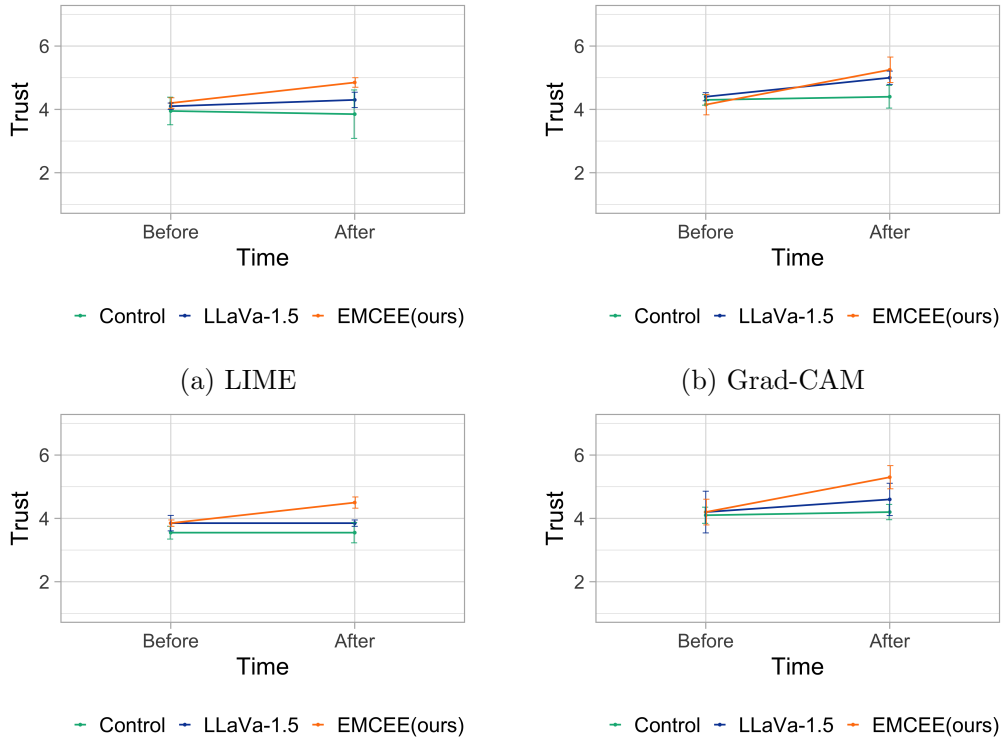
(b) Grad-CAM



(c) Integrated Gradients

(d) SHAP

Figure 5.10: Participants' self-report behavioral intention score 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 5.11: Participants' trust for (a) LIME and (b) Grad-CAM (c) Integrated Gradients (d) SHAP before and after conditions.

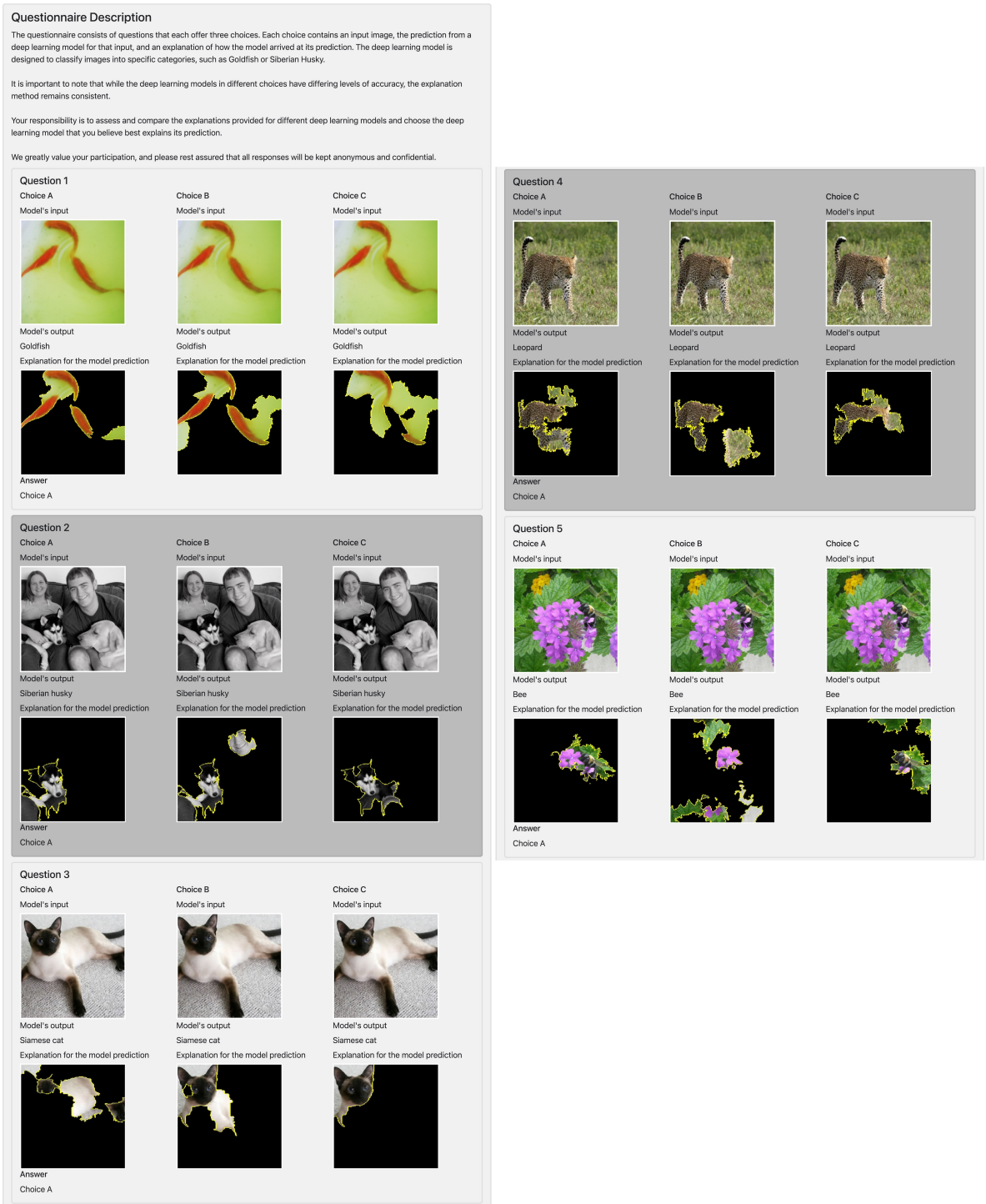


Figure 5.12: Objective evaluation questions used for LIME.

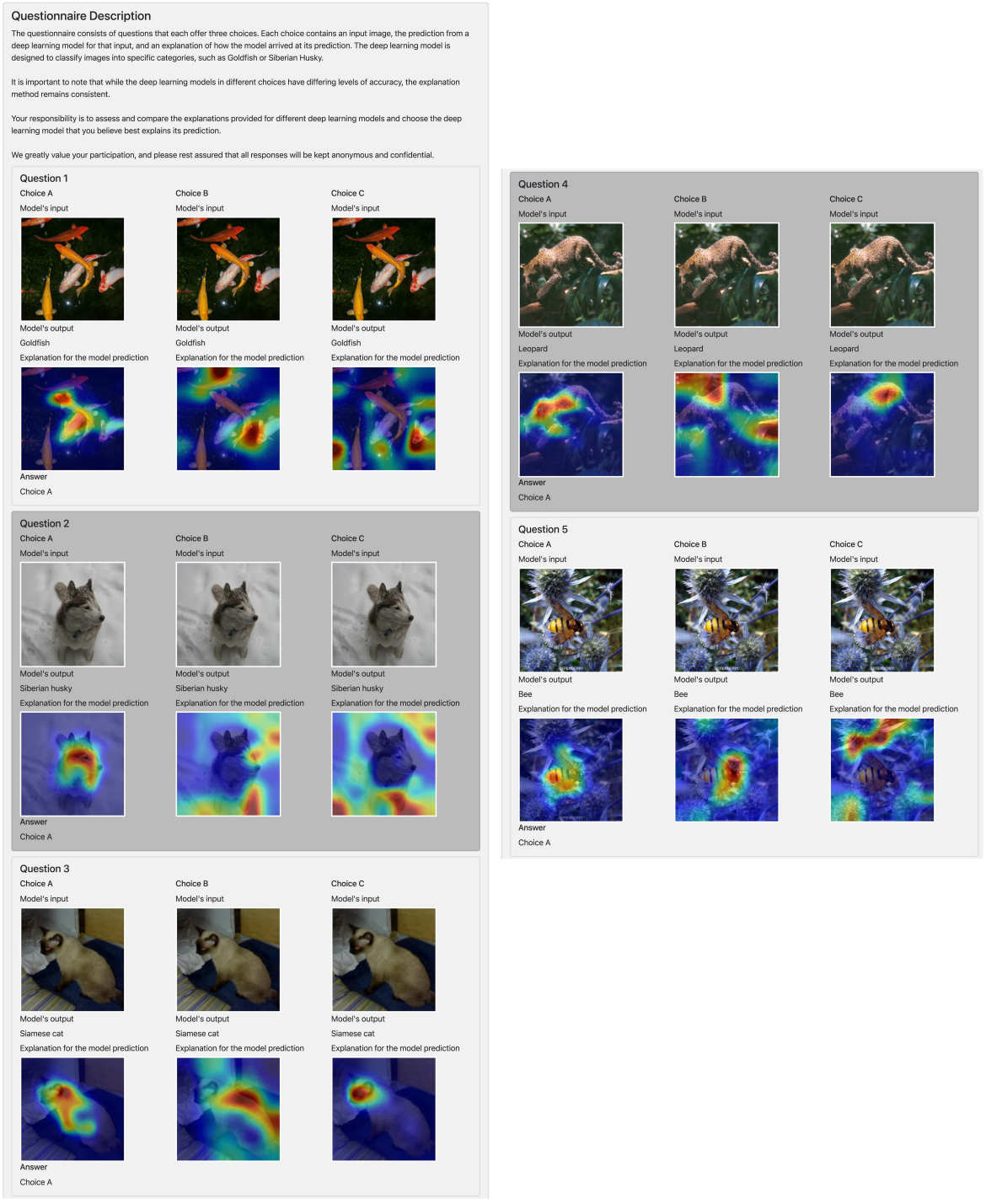


Figure 5.13: Objective evaluation questions used for Grad-CAM.

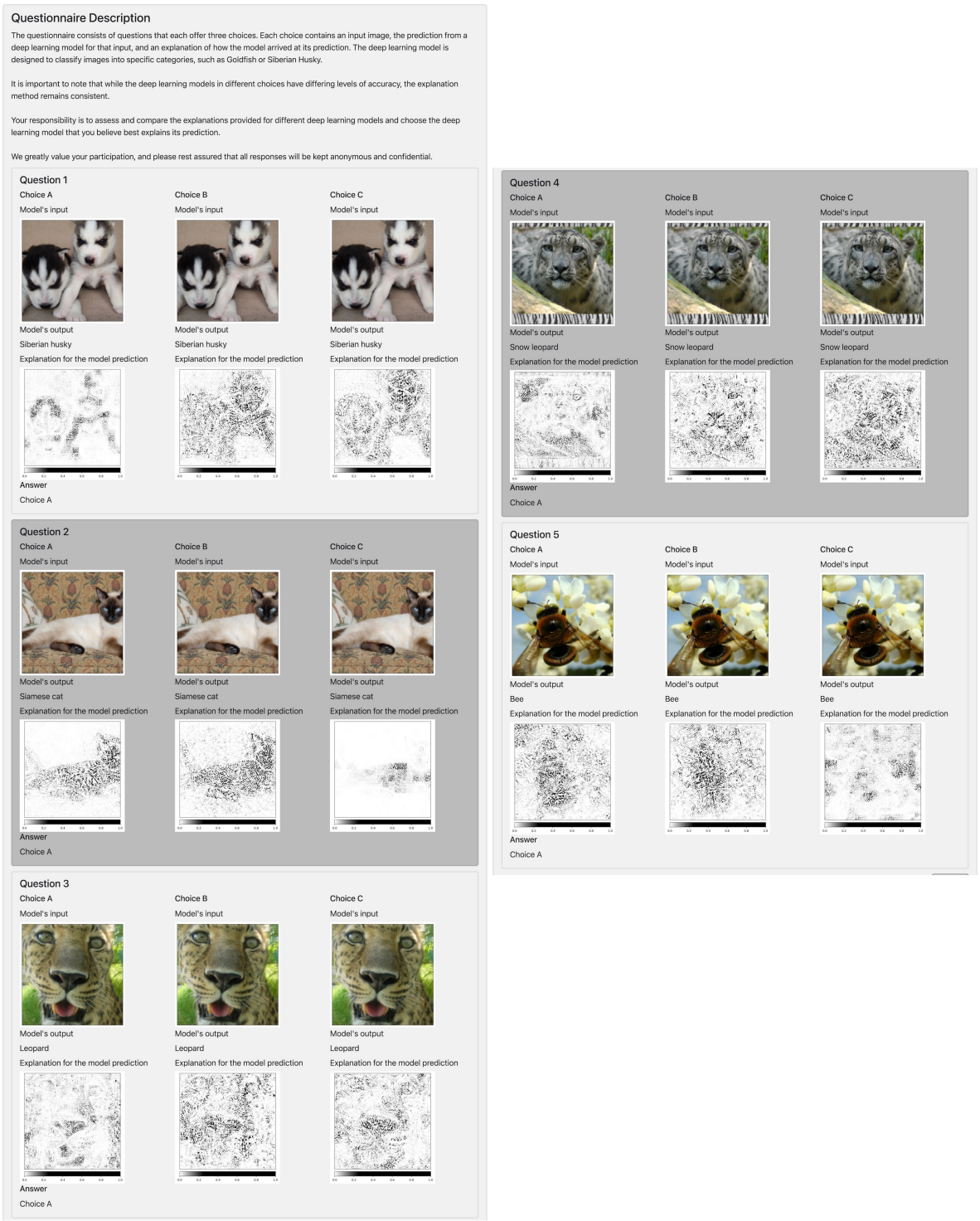


Figure 5.14: Objective evaluation questions used for Integrated Gradients

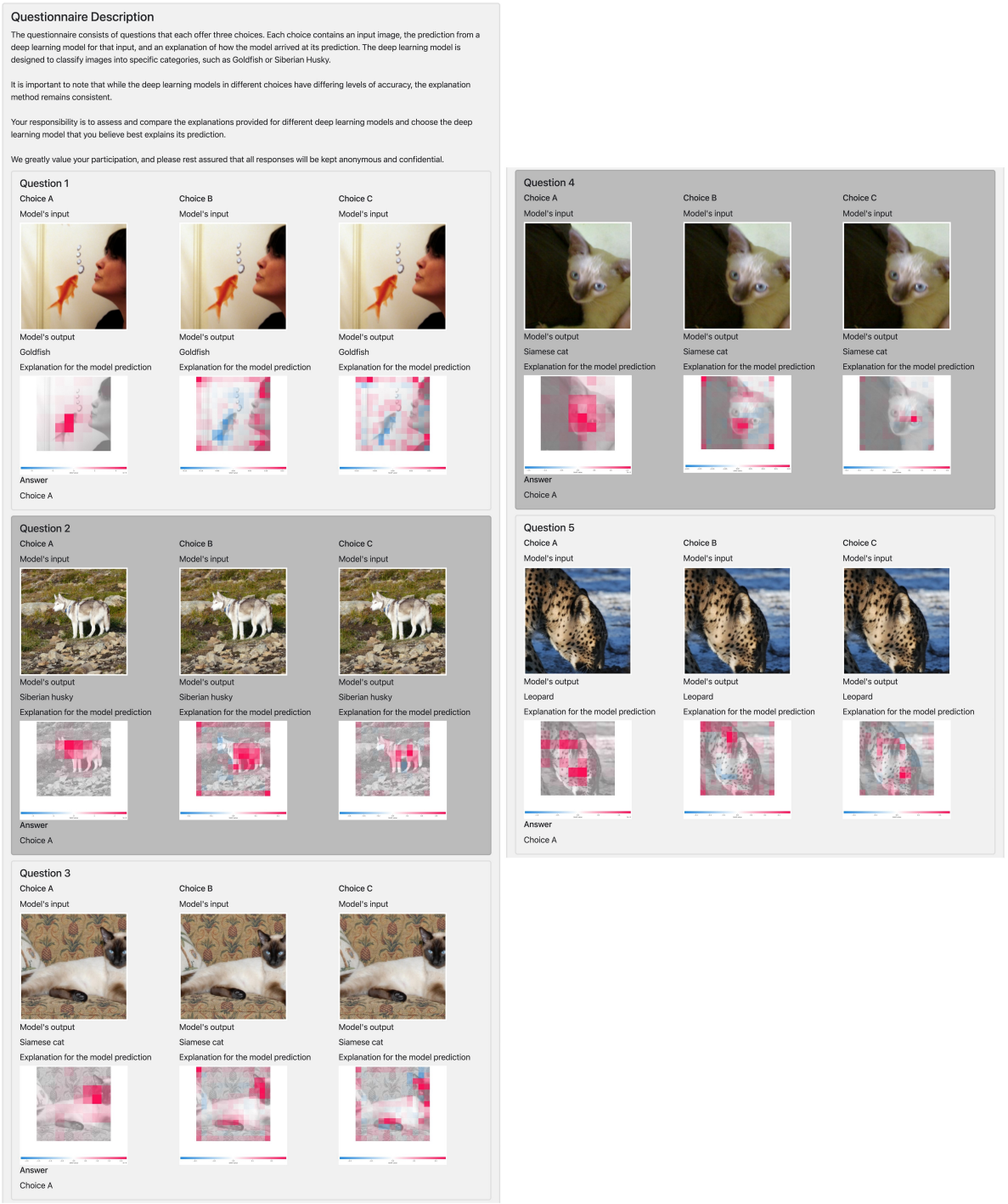


Figure 5.15: Objective evaluation questions used for SHAP.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

In this thesis, we propose to build user-centered dialogue systems from important aspects: personalization, continuity, and support for diverse users.

In chapter 2, we design a users-centered dialogue system from better personalization. We build a conversational recommendation system, KECRS that can provide recommendations with information about items. To give more convincing recommendations, we propose to utilize knowledge graphs to generate more related information about recommended items in responses. We develop the Bag-of-Entity loss and the alignment loss to generate naturalistic conversations and effectively utilize knowledge graphs containing background information about recommendations. The experimental results demonstrate that KECRS can generate more diverse, informative, and relevant recommendations than state-of-the-art baselines.

In chapter 3, we design a users-centered dialogue system with the ability to conduct long-term and continuous conversations with users. We build a multi-session open-domain dialogue system HAHT that can build long-term relationships with users by maintaining the memory of history conversations. Experimental results demonstrate that HAHT can generate more humanized and history-aware responses than state-of-the-art models. Human experiments demonstrate that HAHT can better remember and utilize history conversations with users to conduct more personalized conversations.

In chapter 4 and chapter 5, we design a users-centered dialogue system for user diverse information needs. We first conduct Wizard-of-Oz experiments to investigate

how free-form conversations assist users in understanding static explanations, promoting trust, and making informed decisions about AI models. After proving the benefits of conversations on static explanations, we propose the fEw-shot Multi-round ConvErsational Explanation (EMCEE) to provide customized explanations to users. To deal with data scarcity, we train the EMCEE with synthetic data. We tackle the two main challenges of training with synthetic data: the lack of data diversity and model hallucination. In practice, EMCEE significantly improved users' comprehension, acceptance, trust, and collaboration with static explanations. To the best of our knowledge, we are the first study of how free-form conversations may facilitate neural network explainability in a computer vision task. EMCEE is also the first conversational explanation that can answer free-form follow-up questions after providing static explanations to the user.

We hope that our works could foster more dialogue systems that can consider users' needs at the first place and are designed from the user perspective.

6.2 Future Work

In the following paragraphs, we discuss some of the potential research directions:

Improving evaluations for dialogue systems. Existing evaluations of dialogue systems rely on both automatic and human evaluations. Common automatic evaluation metrics, such as BLEU and ROUGE, measure the n-gram overlap between a system-generated text and a human-written reference text. However, there are multiple proper and correct ways to reply to the same conversation turn, which is usually not considered in existing automatic evaluations. Human evaluation can address this issue by asking human annotators to rate the appropriateness and relevance of each response within the context of conversations. However, human evaluation is very cost- and time-intensive. Furthermore, human evaluations conducted by different papers are usually difficult to compare due to different annotators. Therefore, a faster and more reliable evaluation method for dialogue systems would accelerate the research process and facilitate more consistent comparisons across models.

Unifying item recommendation network and response generation network for conversational recommendation systems. In the proposed conversational recommendation system, the item recommendation network and the response generation

network are trained separately. This separation results in a cumbersome training process and leads to semantic misalignment between the item representations in the recommendation system and the word representations in the response generation network. Such misalignment may prevent the model from recommending items that align with user expectations, as evidenced by the significant performance drop observed after removing the alignment loss (Section 2.3). Although our proposed embedding alignment mitigates this issue, it does not completely resolve it. With the rapid advancement of large language models (LLMs), LLMs have demonstrated the ability to understand items based on descriptions and user-item interaction history [47, 48]. A promising future direction is to leverage LLMs to unify item recommendation and response generation.

Summarization and retrieval for multi-Session conversations. In HAHT, we proposed maintaining conversation history by encoding different historical conversation sessions into dense representations. At the time, language models struggled to handle long-context inputs effectively. However, with recent advancements in model architectures and the increasing availability of GPU resources, LLMs can now process inputs exceeding 200k tokens. Given this, summarizing historical conversation sessions or extracting key sentences may be a more effective approach for managing multi-session conversations. However, summarizing historical conversations introduces new challenges. First, generating accurate and informative summaries without losing critical context remains an open problem, as existing methods may omit essential details or introduce hallucinations. Second, dynamically selecting the most relevant sentences from past sessions requires robust retrieval mechanisms that can effectively capture user intent and maintain coherence across sessions. Lastly, integrating these summaries into ongoing conversations without disrupting the natural flow is challenging, especially when user queries require nuanced contextual understanding. Future work should explore advanced retrieval-augmented generation (RAG) techniques, adaptive summarization strategies, and more efficient memory management mechanisms to enhance multi-session dialogue coherence and efficiency.

Expanding applications of conversational explanations. In the conversational explanations, our experiments focused on feature attribution explanation methods on image classification. The main goal of feature attribution is to understand the contributions of each input feature to a model's decision. There are also other explanation

methods, like example-based explanation methods which focus on the influence of particular data instances. All these explanation methods can provide meaningful insights for users to understand AI models and predictions. However, understanding different explanation methods requires different knowledge bases. Users may ask different questions for different explanation methods on different tasks. As we demonstrated in Chapter 4, it is hard to anticipate questions from laypersons before the study. Therefore, conversational explanations for different explanation methods and different tasks may require special design and are worth exploring.

Removing hallucinations in dialogue systems. In the conversational explanation task, to mitigate hallucinations in dialogue systems, we train a hallucination detector to identify and filter out hallucinated conversation turns during training. To build this detector, we constructed a dataset containing examples of hallucinated statements about basic machine learning and explainable AI methods. However, hallucinations are not limited to a single task; they occur across dialogue systems in various domains and settings. Therefore, developing more general and sophisticated hallucination detection methods remains an important direction for future work. One promising approach is to detect hallucinations based on uncertainty estimation [278, 279]. Uncertainty estimation leverages the model's own confidence in its predictions to identify whether outputs are unfaithful or factually incorrect. Importantly, it operates solely on the LLM-generated text, avoiding the need for additional resources, external knowledge, or further supervised training.

Dialogue systems for low-resource languages. Most existing dialogue systems are designed for widely spoken languages like English, Spanish, German, and Chinese, which require extensive data for training [280]. However, collecting large amounts of training data is challenging for most languages in the world. There are over 7000 languages around the world. The top 15 languages comprise more than 90% of online texts [281]. Future work should explore innovative strategies to build dialogue systems for languages with limited training data. Possible solutions can be adopting transfer learning from high-resource to low-resource languages [282] and employing meta-learning techniques to learn to acquire a new language with limited data [283].

References

[1] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, no. 3, pp. 613–619, 1973. [Online]. Available: https://doi.org/10.1177/001316447303300309

[2] A. J. Liddicoat, *An introduction to conversation analysis.* Bloomsbury Publishing, 2021.

[3] D. Jurafsky, *Speech & language processing.* Pearson Education India, 2000.

[4] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–32, 2020.

[5] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, "Recent advances in deep learning based dialogue systems: a systematic survey," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3055–3155, 2023.

[6] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[7] K. M. Colby, S. Weber, and F. D. Hilf, "Artificial paranoia," *Artificial Intelligence*, vol. 2, no. 1, pp. 1–25, 1971.

[8] R. S. Wallace, "The anatomy of alice," in *Parsing the turing test.* Springer, 2009, pp. 181–210.

[9] J.-P. Kruth, T. Van Ginderachter, P. I. Tanaya, and P. Valckenaers, "The use of finite state machines for task-based machine tool control," *Computers in Industry*, vol. 46, no. 3, pp. 247–258, 2001.

[10] P. J. Price, "Evaluation of spoken language systems: the ATIS domain," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990.

[11] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[12] D. A. Dahl, M. Bates, M. K. Brown, W. M. Fisher, K. Hunicke-Smith, D. S. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the atis task: The atis-3 corpus," in *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

[13] M. A. Walker, J. S. Aberdeen, J. E. Boland, E. O. Bratt, J. S. Garofolo, L. Hirschman, A. N. Le, S. Lee, S. S. Narayanan, K. Papineni *et al.*, "Darpa communicator dialog travel planning systems: the june 2000 data collection." in *INTERSPEECH*. Citeseer, 2001, pp. 1371–1374.

[14] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[15] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[16] S. Roller, Y.-L. Boureau, J. Weston, A. Bordes, E. Dinan, A. Fan, D. Gunning, D. Ju, M. Li, S. Poff *et al.*, "Open-domain conversational agents: Current progress, open problems, and future directions," *arXiv preprint arXiv:2006.12442*, 2020.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[18] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[19] J. Gao, M. Galley, L. Li *et al.*, "Neural approaches to conversational ai," *Foundations and trends® in information retrieval*, vol. 13, no. 2-3, pp. 127–298, 2019.

[20] Y.-P. Chen, N. Nishida, H. Nakayama, and Y. Matsumoto, "Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, May 2024, pp. 13 650–13 665. [Online]. Available: https://aclanthology.org/2024.lrec-main.1192/

[21] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: Long-term open-domain conversation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5180–5197.

[22] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[23] H. Shen, C.-Y. Huang, T. Wu, and T.-H. K. Huang, "ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-ai scientific writing," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 2023, p. 384–387.

[24] R. Li, S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal, "Towards deep conversational recommendations," in *NeurIPS*, 2018, pp. 9725–9735.

[25] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang, "Towards knowledge-based recommender dialog system," *arXiv preprint arXiv:1908.05391*, 2019.

[26] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, and J. Yu, "Improving conversational recommender systems via knowledge graph based semantic fusion," in *SIGKDD*, 2020.

[27] R. Sarkar, K. Goswami, M. Arcan, and J. P. McCrae, "Suggest me a movie for tonight: Leveraging knowledge graphs for conversational recommendation," in *COLING*, 2020, pp. 4179–4189.

[28] Z. Liu, H. Wang, Z.-Y. Niu, H. Wu, W. Che, and T. Liu, "Towards conversational recommendation over multi-type dialogs," *arXiv preprint arXiv:2005.03954*, 2020.

[29] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen, "Towards topic-guided conversational recommender system," *arXiv preprint arXiv:2010.04125*, 2020.

[30] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, and Z. Yu, "INSPIRED: Toward sociable recommendation dialog systems," *arXiv preprint arXiv:2009.14306*, 2020.

[31] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of Wikipedia: Knowledge-powered conversational agents," in *Proceedings of 7th International Conference on Learning Representations*, 2019.

[32] A. Madotto, C.-S. Wu, and P. Fung, "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," *arXiv preprint arXiv:1804.08217*, 2018.

[33] C.-S. Wu, R. Socher, and C. Xiong, "Global-to-local memory pointer networks for task-oriented dialogue," *arXiv preprint arXiv:1901.04713*, 2019.

[34] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua, "Estimation-action-reflection: Towards deep interaction between conversational and recommender systems," in *WSDM*, 2020, pp. 304–312.

[35] W. Lei, G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, and T.-S. Chua, "Interactive path reasoning on graph for conversational recommendation," in *SIGKDD*, 2020, pp. 2073–2083.

[36] Y. Sun and Y. Zhang, "Conversational recommender system," in *SIGIR*, 2018, pp. 235–244.

[37] K. Christakopoulou, F. Radlinski, and K. Hofmann, "Towards conversational recommender systems," in *SIGKDD*, 2016, pp. 815–824.

[38] K. Christakopoulou, A. Beutel, R. Li, S. Jain, and E. H. Chi, "Q&R: A two-stage approach toward interactive recommendation," in *SIGKDD*, 2018, pp. 139–148.

[39] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft, "Towards conversational search and recommendation: System ask, user respond," in *CIKM*, 2018, pp. 177–186.

[40] J. Zou, Y. Chen, and E. Kanoulas, "Towards question-based recommender systems," *arXiv preprint arXiv:2005.14255*, 2020.

[41] K. Xu, J. Yang, J. Xu, S. Gao, J. Guo, and J.-R. Wen, "Adapting user preference to online feedback in multi-round conversational recommendation," in *WSDM*, 2021, p. 364–372.

[42] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," *arXiv preprint arXiv:1605.06069*, 2016.

[43] F.-L. Li, M. Qiu, H. Chen, X. Wang, X. Gao, J. Huang, J. Ren, Z. Zhao, W. Zhao, and L. Wang, "Alime assist: An intelligent assistant for creating an innovative e-commerce experience," in *CIKM*, 2017.

[44] L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen, "Response ranking with deep matching networks and external knowledge in information-seeking conversation systems," in *SIGIR*, 2018, pp. 245–254.

[45] M. Fu, J. Guan, X. Zheng, J. Zhou, J. Lu, T. Zhang, S. Zhuo, L. Zhan, and J. Yang, "ICS-Assist: Intelligent customer inquiry resolution recommendation in online customer service for large E-commerce businesses," *arXiv preprint arXiv:2008.13534*, 2020.

[46] S. Song, C. Wang, H. Chen, and H. Chen, "An emotional comfort framework for improving user satisfaction in E-commerce customer service chatbots," in *NAACL*, 2021, pp. 130–137.

[47] Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasirzadeh, and S. Milano, "A review of modern recommender systems using generative models (gen-recsys)," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24, 2024, p. 6448–6458.

[48] J. Deng, S. Wang, K. Cai, L. Ren, Q. Hu, W. Ding, Q. Luo, and G. Zhou, "Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment," *arXiv preprint arXiv:2502.18965*, 2025.

[49] X. Wang, K. Zhou, J.-R. Wen, and W. X. Zhao, "Towards unified conversational recommender systems via knowledge-enhanced prompt learning," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 1929–1937.

[50] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278.

[51] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[52] Y. Deng, W. Zhang, W. Xu, W. Lei, T.-S. Chua, and W. Lam, "A unified multi-task learning framework for multi-goal conversational recommender systems," *ACM Trans. Inf. Syst.*, vol. 41, no. 3, 2023.

[53] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[54] L. Wang, H. Hu, L. Sha, C. Xu, D. Jiang, and K.-F. Wong, "RecInDial: A unified framework for conversational recommendation with pretrained language models," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online only, 2022, pp. 489–500.

[55] B. Yang, C. Han, Y. Li, L. Zuo, and Z. Yu, "Improving conversational recommendation systems' quality with context-aware item meta-information," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 38–48.

[56] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[57] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," *arXiv preprint arXiv:1703.06103*, 2017.

[58] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[59] M. Ravaut, H. Zhang, L. Xu, A. Sun, and Y. Liu, "Parameter-efficient conversational recommender system as a language processing task," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 152–165.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[61] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, and et al., "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web Journal*, vol. 6, pp. 167–195, 2014.

[62] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[63] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *AAAI*, 2017.

[64] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, V. Wade, and B. R. Cowan, "What makes a good conversation? Challenges in designing truly conversational agents," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12.

[65] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[66] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 300–325.

[67] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 583–593.

[68] X. Zhou, D. Dong, H. Wu, S. Zhao, D. Yu, H. Tian, X. Liu, and R. Yan, "Multi-view response selection for human-computer conversation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 372–381.

[69] C. Yuan, W. Zhou, M. Li, S. Lv, F. Zhu, J. Han, and S. Hu, "Multi-hop selector network for multi-turn response selection in retrieval-based chatbots," 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).* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 111–120.

[70] P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, "Towards persona-based empathetic conversational models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Online: Association for Computational Linguistics, Nov. 2020, pp. 6556–6566.

[71] Y. Zhu, J.-Y. Nie, K. Zhou, P. Du, and Z. Dou, "Content selection network for document-grounded retrieval-based chatbots," in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR.* Berlin, Heidelberg: Springer-Verlag, 2021, p. 755–769.

[72] H. Qian, Z. Dou, Y. Zhu, Y. Ma, and J.-R. Wen, "Learning implicit user profile for personalized retrieval-based chatbot," in *Proceedings of the 30th ACM International Conference on Information and Knowledge Management.* New York, NY, USA: Association for Computing Machinery, 2021, p. 1467–1477.

[73] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.* AAAI Press, 2016, p. 3776–3783.

[74] H.-Y. Shum, X. He, and D. Li, "From Eliza to XiaoIce: Challenges and opportunities with social chatbots," *arXiv preprint arXiv:1801.01957*, 2018.

[75] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 928–939.

[76] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119.

[77] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 3295—-3301.

[78] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995.

[79] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213.

[80] G. Platonov, L. Schubert, B. Kane, and A. Gindi, "A spoken dialogue system for spatial question answering in a physical blocks world," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, Jul. 2020, pp. 128–131.

[81] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474.

[82] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3784–3803.

[83] H. Li, C. Yang, A. Zhang, Y. Deng, X. Wang, and T.-S. Chua, "Hello again! llm-powered personalized agent for long-term dialogue," *NAACL*, 2025.

[84] T. Liu, H. Zhao, Y. Liu, X. Wang, and Z. Peng, "Compeer: A generative conversational agent for proactive peer support," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '24, 2024.

[85] N. Chen, H. Li, J. Huang, B. Wang, and J. Li, "Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations," *arXiv preprint arXiv:2402.11975*, 2024.

[86] K. T.-i. Ong, N. Kim, M. Gwak, H. Chae, T. Kwon, Y. Jo, S.-w. Hwang, D. Lee, and J. Yeo, "Towards lifelong dialogue agents via relation-aware memory construction and timeline-augmented response generation," *arXiv preprint arXiv:2406.10996*, 2024.

[87] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 724–19 731.

[88] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of LLM agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 13 851–13 870.

[89] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proceedings of 5th International Conference on Learning Representations*, 2017.

[90] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1631–1640.

[91] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio, "Pointing the unknown words," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 140–149.

[92] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.

[93] C.-Y. Lin and F. J. Och, "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, Jul. 2004, pp. 605–612.

[94] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 874–880.

[95] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[96] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, Sep 2023.

[97] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "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*. ACM, 2015, p. 1721–1730. [Online]. Available: https://doi.org.remotexs.ntu.edu.sg/10.1145/2783258.2788613

[98] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, ""Hello AI": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, 2019.

[99] Y. Zheng, B. Rowell, Q. Chen, J. Y. Kim, R. A. Kontar, X. J. Yang, and C. A. Lester, "Designing human-centered AI to prevent medication dispensing errors: Focus group study with pharmacists," *JMIR Formative Research*, vol. 7, no. 1, p. e51921, 2023.

[100] T. P. Quinn, M. Senadeera, S. Jacobs, S. Coghlan, and V. Le, "Trust and medical AI: The challenges we face and the expertise needed to overcome them," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 890–894, 2021.

[101] J. Powles and H. Hodson, "Google deepmind and healthcare in an age of algorithms," *Health and Technology*, vol. 7, no. 4, pp. 351–367, 2017.

[102] F. Yang, M. Du, and X. Hu, "Evaluating explanation without ground truth in interpretable machine learning," *arXiv preprint arXiv:1907.06831*, 2019.

[103] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "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*. Association for Computational Linguistics, 2020, pp. 447–459.

[104] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021, pp. 1–16.

[105] A. V. González, G. Bansal, A. Fan, Y. Mehdad, R. Jia, and S. Iyer, "Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021, pp. 1103–1116.

[106] R. Luo, N. Du, and X. J. Yang, "Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time," *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1962–1971, 2022.

[107] G. Nguyen, M. R. Taesiri, and A. Nguyen, "Visual correspondence-based explanations improve AI robustness and human-ai team accuracy," *Neural Information Processing Systems (NeurIPS)*, vol. 35, p. 34287–34301, 2022.

[108] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2019, p. 29–38.

[109] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim, "Debugging tests for model explanations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020.

[110] M. Idahl, L. Lyu, U. Gadiraju, and A. Anand, "Towards benchmarking the utility of explanations for model debugging," in *Proceedings of the First Workshop on Trustworthy Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 68–73.

[111] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020, p. 295–305.

[112] X. Wang and M. Yin, "Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making," in *26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2021, p. 318–328.

[113] H. Lakkaraju, D. Slack, Y. Chen, C. Tan, and S. Singh, "Rethinking explainability as a dialogue: A practitioner's perspective," *arXiv preprint arXiv:2202.01875*, 2022.

[114] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, p. 1–15.

[115] D. Slack, S. Krishna, H. Lakkaraju, and S. Singh, "Explaining machine learning models with interactive natural language conversations using talktomodel," *Nature Machine Intelligence*, vol. 5, no. 8, pp. 873–883, 2023.

[116] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma, "Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[117] A. Springer and S. Whittaker, "Progressive disclosure: Empirically motivated approaches to designing effective transparency," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, p. 107–120.

[118] X. He, Y. Hong, X. Zheng, and Y. Zhang, "What are the users' needs? Design of a user-centered explainable artificial intelligence diagnostic system," *International Journal of Human–Computer Interaction*, vol. 39, no. 7, pp. 1519–1542, 2023.

[119] H. H. Clark and C. R. Marshall, "Definite knowledge and mutual knowledge," in *Elements of Discourse Understanding*. Cambridge University Press, 1981, pp. 10–63.

[120] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition*. American Psychological Association, 1991, pp. 127–149.

[121] J. Wittwer, M. Nückles, and A. Renkl, "Is underestimation less detrimental than overestimation? The impact of experts' beliefs about a layperson's knowledge on learning and question asking," *Instructional Science*, vol. 36, pp. 27–52, 2008.

[122] M. Wilkesmann and U. Wilkesmann, "Knowledge transfer as interaction between experts and novices supported by technology," *Vine*, vol. 41, no. 2, pp. 96–112, 2011.

[123] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[124] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I.-H. Lee, M. Muller, and M. O. Riedl, "The who in explainable AI: How AI background shapes perceptions of AI explanations," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024, pp. 1–32.

[125] Q. V. Liao and K. R. Varshney, "Human-centered explainable AI (XAI): From algorithms to user experiences," *arXiv preprint arXiv:2110.10790*, 2021.

[126] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2018, p. 1–18.

[127] U. Schmid and B. Wrede, "What is missing in XAI so far? An interdisciplinary perspective," *KI-Künstliche Intelligenz*, vol. 36, no. 3-4, pp. 303–315, 2022.

[128] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach *et al.*, "Explanation as a social practice: Toward a conceptual framework for the social design of AI systems," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 717–728, 2020.

[129] H.-F. Cheng, R. Wang, Z. Zhang, F. O'connell, T. Gray, F. M. Harper, and H. Zhu, "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*. Association for Computing Machinery, 2019, pp. 1–12.

[130] M. Guesmi, M. A. Chatti, S. Joarder, Q. U. Ain, R. Alatrash, C. Siepmann, and T. Vahidi, "Interactive explanation with varying level of details in an explainable scientific literature recommender system," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–22, 2023.

[131] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[132] J. A. Fails and D. R. Olsen Jr, "Interactive machine learning," in *Proceedings of the 8th international conference on Intelligent user interfaces*, 2003, pp. 39–45.

[133] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater, "No explainability without accountability: An empirical study of explanations and feedback in interactive ml," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020, pp. 1–13.

[134] W. Liang, J. Zou, and Z. Yu, "ALICE: Active learning with contrastive natural language explanations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4380–4391.

[135] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2015, p. 126–137.

[136] H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–45, 2021.

[137] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, and A. Yuan, "The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 107–118.

[138] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019, p. 1–13.

[139] N. Feldhus, A. M. Ravichandran, and S. Möller, "Mediators: Conversational agents explaining NLP model behavior," *arXiv preprint arXiv:2206.06029*, 2022.

[140] T. Zhang, Y. Liu, B. Li, Z. Zeng, P. Wang, Y. You, C. Miao, and L. Cui, "History-aware hierarchical transformer for multi-session open-domain dialogue system," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Dec. 2022, pp. 3395–3407.

[141] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey," *Artificial intelligence review*, vol. 56, no. 4, pp. 3055–3155, 2023.

[142] K. Shuster, J. Xu, M. Komeili, D. Ju, E. M. Smith, S. Roller, M. Ung, M. Chen, K. Arora, J. Lane *et al.*, "BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage," *arXiv preprint arXiv:2208.03188*, 2022.

[143] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, and J.-R. Wen, "A survey on complex knowledge base question answering: Methods, challenges and solutions," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4483–4491.

[144] M. Luo, Z. Fang, T. Gokhale, Y. Yang, and C. Baral, "End-to-end knowledge retrieval with multi-modal queries," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023, pp. 8573–8589.

[145] L. Zhang, J. Zhang, Y. Wang, S. Cao, X. Huang, C. Li, H. Chen, and J. Li, "FC-KBQA: A fine-to-coarse composition framework for knowledge base question answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023, pp. 1002–1017.

[146] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[147] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[148] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[149] D. Anderson and K. Burnham, "Model selection and multi-model inference," *Springer-Verlag*, vol. 63, p. 512, 2004.

[150] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of 3rd International Conference on Learning Representations*. openreview, 2015.

[151] H. Lakkaraju, S. H. Bach, and J. Leskovec, "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*, 2016, pp. 1675–1684.

[152] F. Rudziński, "A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers," *Applied Soft Computing*, vol. 38, pp. 118–133, 2016.

[153] H. Yang, C. Rudin, and M. Seltzer, "Scalable bayesian rule lists," in *International conference on machine learning*. JMLR.org, 2017, pp. 3921–3930.

[154] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*. IEEE, 2017, pp. 618–626.

[155] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*. Association for Computing Machinery, 2016, pp. 1135–1144.

[156] Y. Chen, B. Li, H. Yu, P. Wu, and C. Miao, "Hydra: Hypergradient data relevance analysis for interpreting deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35. MIT Press, 2021, pp. 7081–7089.

[157] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[158] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[159] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. JMLR.org, 2017, pp. 3319–3328.

[160] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences*, vol. 225, pp. 1–17, 2013.

[161] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[162] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[163] L. Hu, J. Chen, V. N. Nair, and A. Sudjianto, "Locally interpretable models and effects based on supervised partitioning (LIME-SUP)," *arXiv preprint arXiv:1806.00663*, 2018.

[164] D. Alvarez-Melis and T. Jaakkola, "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*. Association for Computational Linguistics, 2017, pp. 412–421.

[165] N. Liu, X. Huang, J. Li, and X. Hu, "On interpretation of network embedding via taxonomy induction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1812–1820.

[166] A. Shih, A. Choi, and A. Darwiche, "A symbolic approach to explaining bayesian network classifiers," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, p. 5103–5111.

[167] A. Ignatiev, N. Narodytska, and J. Marques-Silva, "Abduction-based explanations for machine learning models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 1511–1519.

[168] K. H. Tran, A. Ghazimatin, and R. Saha Roy, "Counterfactual explanations for neural recommenders," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2021, pp. 1627–1631.

[169] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 607–617.

[170] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, 2020, pp. 344–350.

[171] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, "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, 2020, pp. 4211–4222.

[172] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019.

[173] J. Yoon, S. Arik, and T. Pfister, "Data valuation using reinforcement learning," in *International Conference on Machine Learning*. JMLR.org, 2020, pp. 10 842–10 851.

[174] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[175] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, 2020, p. 166–172.

[176] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 895–905.

[177] D. Croce, D. Rossini, and R. Basili, "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)*. Association for Computational Linguistics, 2019, pp. 4037–4046.

[178] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *The Annals of Applied Statistics*, vol. 5, no. 4, pp. 2403–2424, 2011.

[179] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, pp. 1–9, 2013.

[180] F. Doshi-Velez, B. C. Wallace, and R. Adams, "Graph-Sparse LDA: A topic model with structured sparsity," in *Twenty-Ninth AAAI conference on artificial intelligence*, vol. 29. AAAI Press, 2015, pp. 2575–2581.

[181] B. Kim, R. Khanna, and O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2016, p. 2288–2296.

[182] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, ""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*. Association for Computing Machinery, 2023.

[183] P. Lertvittayakumjorn, L. Specia, and F. Toni, "FIND: Human-in-the-Loop Debugging Deep Text Classifiers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 332–348.

[184] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2662–2670.

[185] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting, "Making deep neural networks right for the right scientific reasons by interacting with their explanations," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 476–486, 2020.

[186] O. Alkan, D. Wei, M. Mattetti, R. Nair, E. Daly, and D. Saha, "FROTE: Feedback rule-driven oversampling for editing models," in *Proceedings of Machine Learning and Systems*, vol. 4, 2022, pp. 276–301.

[187] A. Biswas and D. Parikh, "Simultaneous active learning of classifiers & attributes via relative feedback," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 644–651.

[188] S. Teso and K. Kersting, "Explanatory interactive machine learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, p. 239–245.

[189] S. Teso, A. Bontempelli, F. Giunchiglia, and A. Passerini, "Interactive label cleaning with example-based explanations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 966–12 977, 2021.

[190] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[191] S. Herse, J. Vitale, and M.-A. Williams, "Using agent features to influence user trust, decision making and task outcome during human-agent collaboration," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1740–1761, 2023.

[192] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to human interaction with AI systems: New challenges and opportunities for hci professionals to enable human-centered ai," *International Journal of Human–Computer Interaction*, vol. 39, no. 3, pp. 494–518, 2023.

[193] K. I. Gero, Z. Ashktorab, C. Dugan, Q. Pan, J. Johnson, W. Geyer, M. Ruiz, S. Miller, D. R. Millen, M. Campbell, S. Kumaravel, and W. Zhang, "Mental models of AI agents in a cooperative game setting," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020, p. 1–12.

[194] R. Häuslschmid, M. von Bülow, B. Pfleging, and A. Butz, "Supportingtrust in autonomous driving," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2017, p. 319–329.

[195] C. Carissoli, L. Negri, M. Bassi, F. A. Storm, and A. D. Fave, "Mental workload and human-robot interaction in collaborative tasks: A scoping review," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–20, 2023.

[196] L. Liu, F. Guo, Z. Zou, and V. G. Duffy, "Application, development and future opportunities of collaborative robots (cobots) in manufacturing: A literature review," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–18, 2022.

[197] S. Bhat, J. B. Lyons, C. Shi, and X. J. Yang, "Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2024, pp. 32–41.

[198] Z. Ashktorab, Q. V. Liao, C. Dugan, J. Johnson, Q. Pan, W. Zhang, S. Kumaravel, and M. Campbell, "Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, 2020.

[199] S. D'Avella, G. Camacho-Gonzalez, and P. Tripicchio, "On multi-agent cognitive cooperation: Can virtual agents behave like humans?" *Neurocomputing*, vol. 480, no. C, p. 27–38, 2022.

[200] T. Numata, H. Sato, Y. Asa, T. Koike, K. Miyata, E. Nakagawa, M. Sumiya, and N. Sadato, "Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions," *Scientific reports*, vol. 10, no. 1, p. 5977, 2020.

[201] S. Feng and J. Boyd-Graber, "What can AI do for me? Evaluating machine learning interpretations in cooperative play," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2019, p. 229–239.

[202] G. Nguyen, D. Kim, and A. Nguyen, "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 422–26 436, 2021.

[203] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[204] M. R. Taesiri, G. Nguyen, and A. Nguyen, "Visual correspondence-based explanations improve AI robustness and human-ai team accuracy," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 287–34 301, 2022.

[205] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*. Computer Vision Foundation, 2021, pp. 10 012–10 022.

[206] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of 3rd International Conference on Learning Representations*. openreview, 2015.

[207] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, 2012, pp. 84–90.

[208] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[209] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[210] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Management science*, vol. 35, no. 8, pp. 982–1003, 1989.

[211] E. B. Diop, S. Zhao, and T. V. Duy, "An extension of the technology acceptance model for understanding travelers' adoption of variable message signs," *PLoS one*, vol. 14, no. 4, 2019.

[212] C. Flathmann, B. G. Schelble, N. J. McNeese, B. Knijnenburg, A. K. Gramopadhye, and K. C. Madathil, "The purposeful presentation of AI teammates: Impacts on human acceptance and perception," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–18, 2023.

[213] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "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*, 2017, pp. 408–416.

[214] Y. Guo, X. J. Yang, and C. Shi, "Enabling team of teams: A trust inference and propagation (TIP) model in multi-human multi-robot teams," in *Robotics: Science and Systems XIX*. Association for Computing Machinery, 2023.

[215] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Transactions on Information Systems*, vol. 2, no. 1, p. 26–41, 1984.

[216] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, vol. 21. Curran Associates, Inc., 2018, pp. 9525–9536.

[217] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, *The (Un)reliability of Saliency Methods*. Springer International Publishing, 2019, pp. 267–280.

[218] A. Jacovi and Y. Goldberg, "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*. Association for Computational Linguistics, 2020, pp. 4198–4205.

[219] B. M. Muir and N. Moray, "Trust in automation. part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[220] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A systematic literature review of user trust in AI-enabled systems: An HCI perspective," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–16, 2022.

[221] S. Sebo, L. L. Dong, N. Chang, M. Lewkowicz, M. Schutzman, and B. Scassellati, "The influence of robot verbal support on human team members: Encouraging outgroup contributions and suppressing ingroup supportive behavior," *Frontiers in Psychology*, p. 3584, 2020.

[222] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura, "Voice in human–agent interaction: A survey," *ACM Computing Surveys*, vol. 54, no. 4, 2021.

[223] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[224] E. S. Vorm and D. J. Y. Combs, "Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (istam)," *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1828–1845, 2022.

[225] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, and M. Gombolay, "Explainable artificial intelligence: Evaluating the objective and subjective impacts of XAI on human-agent interaction," *International Journal of Human–Computer Interaction*, vol. 39, no. 7, pp. 1390–1404, 2023.

[226] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, 2008, p. 227–236.

[227] T. Ha and S. Kim, "Improving trust in AI with mitigating confirmation bias: Effects of explanation type and debiasing strategy for decision-making with explainable AI," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–12, 2023.

[228] R. Larasati, A. D. Liddo, and E. Motta, "The effect of explanation styles on user's trust," in *2020 Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies*. Association for Computing Machinery, 2020.

[229] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt, "How do visual explanations foster end users' appropriate trust in machine learning?" in *Proceedings of the 25th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2020, p. 189–201.

[230] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019, p. 1–12.

[231] D. H. McKnight, L. L. Cummings, and N. L. Chervany, "Initial trust formation in new organizational relationships," *Academy of Management review*, vol. 23, no. 3, pp. 473–490, 1998.

[232] B. Y. Lim, A. K. Dey, and D. Avrahami, "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*. Association for Computing Machinery, 2009, pp. 2119–2128.

[233] W. Pieters, "Explanation and trust: what to tell the user in security and ai?" *Ethics and information technology*, vol. 13, pp. 53–64, 2011.

[234] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your AI: Expertise and explanations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2019, p. 240–251.

[235] M. S. Carolin Ebermann and S. Weibelzahl, "Explainable AI: The effect of contradictory decisions and explanations on users' acceptance of AI systems," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1807–1826, 2023.

[236] K. Yu, S. Berkovsky, R. Taib, J. Zhou, and F. Chen, "Do I trust my machine teammate? An investigation from perception to decision," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, p. 460–468.

[237] C. Camerer, G. Loewenstein, and M. Weber, "The curse of knowledge in economic settings: An experimental analysis," *Journal of Political Economy*, vol. 97, no. 5, pp. 1232–1254, 1989.

[238] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *The Twelfth International Conference on Learning Representations*. openreview, 2024.

[239] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "PaLM-E: an embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023.

[240] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, "MultiModal-GPT: A vision and language model for dialogue with humans," 2023.

[241] S. Bankins, P. Formosa, Y. Griep, and D. Richards, "AI decision making with dignity? contrasting workers' justice perceptions of human and AI decision making in a human resource management context," *Information Systems Frontiers*, vol. 24, no. 3, pp. 857–875, 2022.

[242] P. Formosa, W. Rogers, Y. Griep, S. Bankins, and D. Richards, "Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts," *Computers in Human Behavior*, vol. 133, p. 107296, 2022.

[243] T. Zhang, X. J. Yang, and B. Li, "May i ask a follow-up question? understanding the benefits of conversations in neural network explainability," *International Journal of Human–Computer Interaction*, 2023.

[244] K. Schwarz, Y. Liao, and A. Geiger, "On the frequency bias of generative models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 126–18 136, 2021.

[245] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "The curse of recursion: Training on generated data makes models forget," *arXiv Preprint 2305.17493*, 2024.

[246] M. Briesch, D. Sobania, and F. Rothlauf, "Large language models suffer from their own output: An analysis of the self-consuming training loop," *arXiv Preprint 2311.16822*, 2023.

[247] N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro, "Factuality enhanced language models for open-ended text generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 586–34 599, 2022.

[248] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[249] W. Dai, Z. Liu, Z. Ji, D. Su, and P. Fung, "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*, 2023, pp. 2136–2148. [Online]. Available: https://aclanthology.org/2023.eacl-main.156

[250] S. Zheng, J. Huang, and K. C.-C. Chang, "Why does chatgpt fall short in answering questions faithfully?" *arXiv preprint arXiv:2304.10513*, 2023.

[251] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans, "The reversal curse: LLMs trained on "a is b" fail to learn "b is a"," in *The Twelfth International Conference on Learning Representations*, 2024.

[252] S. Wiegreffe and Y. Pinter, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. [Online]. Available: https://aclanthology.org/D19-1002

[253] Y. Wang, T. Zhang, X. Guo, and Z. Shen, "Gradient based feature attribution in explainable ai: A technical review," *arXiv preprint arXiv:2403.10415*, 2024.

[254] E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak, and M. Robnik-Šikonja, "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*, 2021, pp. 16–21. [Online]. Available: https://aclanthology.org/2021.hackashop-1.3

[255] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "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, 2016, pp. 681–691. [Online]. Available: https://aclanthology.org/N16-1082

[256] G. Nguyen, V. Chen, M. R. Taesiri, and A. T. Nguyen, "Pcnn: Probable-class nearest-neighbor explanations improve fine-grained image classification accuracy for ais and humans," *arXiv preprint arXiv:2308.13651*, 2024.

[257] H. Guo, N. Rajani, P. Hase, M. Bansal, and C. Xiong, "FastIF: Scalable influence functions for efficient model interpretation and debugging," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10 333–10 350. [Online]. Available: https://aclanthology.org/2021.emnlp-main.808

[258] K. Yin and G. Neubig, "Interpreting language models with contrastive explanations," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 184–198.

[259] X. Ye, R. Nair, and G. Durrett, "Connecting attributions and QA model behavior on realistic counterfactuals," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 5496–5512.

[260] A. Ross, A. Marasović, and M. Peters, "Explaining NLP models via minimal contrastive editing (MiCE)," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Aug. 2021, pp. 3840–3852.

[261] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld, "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)*, Aug. 2021, pp. 6707–6723. [Online]. Available: https://aclanthology.org/2021.acl-long.523

[262] Y. Meng, J. Huang, Y. Zhang, and J. Han, "Generating training data with language models: Towards zero-shot language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 462–477, 2022.

[263] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong, "ZeroGen: Efficient zero-shot learning via dataset generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11 653–11 669.

[264] X. Guo and Y. Chen, "Generative ai for synthetic data generation: Methods, challenges and the future," *arXiv preprint arXiv:2403.04190*, 2024.

[265] J. Gao, R. Pi, L. Yong, H. Xu, J. Ye, Z. Wu, W. ZHANG, X. Liang, Z. Li, and L. Kong, "Self-guided noise-free data generation for efficient zero-shot learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=h5OpjGd_lo6

[266] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, and J. Han, "Tuning language models as training data generators for augmentation-enhanced few-shot learning," in *International Conference on Machine Learning*, 2023, pp. 24 457–24 477.

[267] J. Ye, J. Gao, Z. Wu, J. Feng, T. Yu, and L. Kong, "ProGen: Progressive zero-shot dataset generation via in-context feedback," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 3671–3683. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.269

[268] J. Zhang, B. Wang, Z. Hu, P. W. W. Koh, and A. J. Ratner, "On the trade-off of intra-/inter-class diversity for supervised pre-training," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[269] H. R. Kirk, Y. Jun, F. Volpin, H. Iqbal, E. Benussi, F. Dreyer, A. Shtedritski, and Y. Asano, "Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models," *Advances in neural information processing systems*, vol. 34, pp. 2611–2624, 2021.

[270] D. Esiobu, X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. Smith, "Robbie: Robust bias evaluation of large generative language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3764–3814.

[271] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[272] R. Grosse, J. Bae, C. Anil, N. Elhage, A. Tamkin, A. Tajdini, B. Steiner, D. Li, E. Durmus, E. Perez *et al.*, "Studying large language model generalization with influence functions," *arXiv preprint arXiv:2308.03296*, 2023.

[273] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[274] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[275] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[276] Y. Guo and X. J. Yang, "Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach," *International Journal of Social Robotics*, vol. 13, pp. 1899–1909, 2021.

[277] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[278] T. Zhang, L. Qiu, Q. Guo, C. Deng, Y. Zhang, Z. Zhang, C. Zhou, X. Wang, and L. Fu, "Enhancing uncertainty-based hallucination detection with stronger focus," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Dec. 2023, pp. 915–932.

[279] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[280] K. Wołk, A. Wołk, D. Wnuk, T. Grześ, and I. Skubis, "Survey on dialogue systems including slavic languages," *Neurocomputing*, vol. 477, pp. 62–84, 2022.

[281] (2024) [3]. usage statistics of content languages for websites. [Online]. Available: https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/

[282] X. Guo, B. Li, and H. Yu, "Improving the sample efficiency of prompt tuning with domain adaptation," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 3523–3537.

[283] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2017, p. 1126–1135.