The Pennsylvania State University
The Graduate School

TOWARDS USEFUL AI INTERPRETABILITY FOR HUMANS VIA INTERACTIVE AI EXPLANATIONS

A Dissertation in
Information Sciences and Technologies
by
Hua Shen

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2023

The dissertation of Hua Shen was reviewed and approved* by the following:

Ting-Hao 'Kenneth' Huang
Assistant Professor of Information Sciences and Technology
Dissertation Advisor, Chair of Committee

Mary Beth Rosson
Professor of Information Sciences and Technology

C. Lee Giles
David Reese Professor of Information Sciences and Technology

S. Shyam Sundar
James P. Jimirro Professor of Media Effects

Sherry Tongshuang Wu
Assistant Professor of School of Computer Science, Carnegie Mellon University
Special Member

Jeffrey Bardzell
Professor of Information Sciences and Technology
Program Head

*Signatures are on file in the Graduate School.

Abstract

Advancements in deep learning have revolutionized AI systems, enabling collaboration between humans and AI to enhance performance in specific tasks. AI explanations play a crucial role in aiding human understanding, control, and improvement of AI systems regarding various criteria such as fairness, safety, and trustworthiness. Despite the proliferation of eXplainable AI (XAI) approaches, the practical usefulness of AI explanations in human-AI collaborative systems remains underexplored. This doctoral research aims to evaluate and enhance the usefulness of AI explanations for humans in practical human-AI collaboration. I break down the research goal of investigating and improving human-centered useful AI explanations into three research questions: RQ1: Are cutting-edge AI explanations useful for humans in practice (Part I)? RQ2: What's the disparity between AI explanations and practical user demands (Part II)? RQ3: How to empower useful AI explanations with human-AI interaction (Part III)? We examined the three research questions by conducting four projects. To answer RQ1, we deployed two real-world human evaluation studies on analyzing computer vision AI model errors with post-hoc explanations and simulating NLP AI model predictions with inherent explanations, respectively. The two studies unveil that, surprisingly, AI explanations are not always useful for humans to analyze AI predictions in practice. This motivates our research for RQ2 – gaining insights into disparities between the status quo of AI explanations and practical user needs. By surveying over 200 AI explanation papers and comparing with summarized real-world user demands, we observe two dominating findings: *i)* humans request diverse XAI questions across the AI pipeline to gain a global view of AI system, whereas existing XAI approaches commonly display a single AI explanation that can not satisfy diverse XAI user needs; *ii)* humans are widely interested in understanding what AI systems can not achieve, which might lead to the need of interactive AI explanations that enable humans to specify the counterfactual predictions. In light of these findings, we deeply deem that, instead of designating user demands by XAI researchers during AI system development, empowering users to communicate with AI systems for their practical XAI demands is critical to unleashing useful AI explanations (RQ3). To this end, we developed an interactive XAI system via conversations that improved the usefulness of AI explanations in terms of human-perceived performance in AI-assisted writing tasks. Overall, we summarize this doctoral research by discussing the limitations and challenges of human-centered useful AI explanations.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful for every person and experience that has come into my life; they helped me to, fortunately, become who I am today.

Pursuing a Ph.D. degree has been a challenging journey, filled with obstacles, difficulties, and stress. However, it has also brought me immense joy, excitement, and hope. I have been fortunate and thankful to be a part of the CrowdAI lab, working under the guidance of the exceptional advisor, Ting-Hao 'Kenneth' Huang. I am appreciative of his invaluable advice and unwavering support throughout the journey of my Ph.D life. His insightful perspectives on the fields inspire me to delve deeper into my critical thinking, while his high standards for research encourage me to deliver results of my utmost quality. The invaluable training I have received under his guidance during the past four years has not only shaped my current endeavors but will undoubtedly have a lasting impact on my future research as well.

I would also like to express my deepest gratitude to Many Beth Rosson, an incredible kind and esteemed female professor whom I admire greatly. It is because of her that I discovered my passion for HCI and got the enlightenment of joining CrowdAI lab. I'm not sure if she realizes the profound impact her advice and encouragement have had on not only my research but also the trajectory of my life. Whenever I think of her, I am overwhelmed with feelings of gratitude, warmth, and support. She has my greatest thanks and blessings all the time.

Among the individuals I feel incredibly fortunate to meet and admire is Sherry Tongshuang Wu, a rising star in our research field. Not only is she exceptionally talented, but she is also a warm, caring, and supportive friend. I am grateful that I have learned much from her constructive suggestions, impressive visions, and research capabilities. Her excellence serves as an example for me to advance and progress as a researcher so that we can support one another.

It is with great honor that I have C. Lee Giles on my thesis committee. He is a giant in the field of AI, particularly renowned for his contributions to information science search engines and scholarly big data. I am filled with gratitude for enhancing the quality of my research under his insightful guidance and feedback, which plays a crucial role in improving my work.

I am also incredibly honored to have S. Shyam Sundar as my committee advisor, who is a prominent figure in the field of Responsible AI. His renowned interdisciplinary perspectives and influential contributions have earned my admiration. I am grateful for his valuable guidance and insightful suggestions that play a pivotal role in advancing my work.

During my doctoral years, I had the privilege of completing three internships with the excep-

tional Amazon Alexa AI and Google AI research teams. These experiences opened a brand new door to speech processing research for me. I feel so lucky to have worked alongside brilliant scientists, including Andreas Stolcke, Jasha Droppo, Ariya Rastrow, Ivan Bulyko, Yuguang Yang, Jari Kolehmainen, Yi Gu, Ankur Gandhe, Denis Filimonov, Qi Luo, Aditya Gourav, Guoli Sun, Ryan Langman, Eunjung (Christine) Han from Amazon Alexa AI teams, and the prominent scientists Vicky Zayats, Dan Walker, Dirk Padfield, and Johann Rocholl from Google AI Research. They have consistently offered me with their support and encouragement in both intellectual and computational resources. I extend my heartfelt thanks to all of them for providing me with invaluable internship experiences.

Also, I express sincere gratitude to a number of professors and distinguished researchers who have advised the projects I have undertaken – Daniel S Weld, Jeffrey Heer, Marco Tulio Ribeiro, Brian D. Davison, Hung-yi Lee, Shang-Wen (Daniel) Li, Frank E. Ritter, Ting Wang– as their constructive insights and expertise have greatly enriched my learning experience.

Collaborating and communicating with other researchers has been an enriching experience from which I continually learn. I would like to express my gratitude to my collaborator, Wenbo Guo, for his support, encouragement, and perseverance throughout our extensive project. I appreciate his patience and insightful suggestions whenever we encountered obstacles. Besides, I extend my thanks to my collaborators, Chieh-Yang Huang, Jooyoung Lee, and Adaku Uchendu. Their extensive knowledge and expertise in various techniques are impressive, and I am deeply grateful for their constant support. I am grateful for my mentees and friends, Yuxin, Ruchi, Reuben, whom I collaborate on interesting research and learn much knowledge from.

Furthermore, I am extremely grateful for my dear friends, including Tiffany, Shaurya, Wesley, Edward, Yujie, Saranya, Zhiyu, Limeng, Qian, and many more. They have brought me immense joy and provided constant mental and physical support throughout my graduate life. I am filled with gratitude to have them in my life. I would also like to thank all my fellow labmates from the CrowdAI lab, including Alan, Zeyu, Sandy, Avon, and others. The time I have spent with them during events like Bubble teas, Hotpots, and BBQs has been truly wonderful.

Above all, I am especially grateful for my dear parents. They have always been there for me, offering their constant backing and being proud of me, regardless of the challenges I faced. I would not have reached this stage without their selfless love and support. They are my greatest motivation along the life path. Last but not least, I would like to express my heartfelt thanks to my boyfriend – for his unconditional love and support during the past nine years and in the future. His presence has been a constant source of strength, helping me navigate both joyful and challenging moments. Love knows no boundaries of time and distance. I am fortunate to have him in my life, and sincerely appreciate everything he had done for me.

To everyone in my life, wish light and love continue to illuminate our paths.

Chapter 1
Introduction

1.1 Motivation

Deep learning advancements have led to breakthroughs in numerous artificial intelligence (AI) systems [154, 215, 216]. Therefore, humans collaborate with the superior capability of AI systems to achieve complementary performance on specific tasks [8, 213]. The complex applications of human-AI collaborative systems have led to a surge of interest in developing systems, which are not only optimized for task performance but also require catering to other vital criteria such as fairness for demographic groups, safety on attacks, human trustworthiness [215, 272], etc. For human-AI collaborative systems, ensuring these auxiliary criteria is of great importance. However, these auxiliary criteria, unlike the conventional task performance metrics (*e.g.,* accuracy), are commonly qualitative measures that are difficult to be quantified. Here is why we need AI interpretability criterion – interpretability per se is not our goal, instead, AI interpretability is a fallback to be used by humans to gauge the AI model reasoning and assess the auxiliary measurements [47].

Therefore, a surge of eXplainable AI (XAI) approaches has been developed and validated to faithfully reflect the model behavior with the automatic metrics such as *faithfulness* [44, 87] and be plausible to humans assessed by the metrics like *plausibility* [11]. However, taking one step further to fulfill the practical XAI roles, it is still under-explored in terms of how humans can leverage AI explanations, as the auxiliary criteria, to boost human-AI collaborative systems in practical applications such as debugging the model or simulating model prediction, etc [63].

In this paper, the **overall objective** is to **evaluate and improve the usefulness of AI explanations** for humans-AI collaborative systems in real-world practice. I break down the research objective into investigating three research questions:

- **RQ1**: Are cutting-edge AI explanations useful for humans in practice? (Part I)
- **RQ2**: What's the disparity between AI explanation and practical user demands? (Part II)
- **RQ3**: How to empower useful AI explanation with human-AI interaction? (Part III)

We examined the three research questions by conducting five projects. To answer RQ1

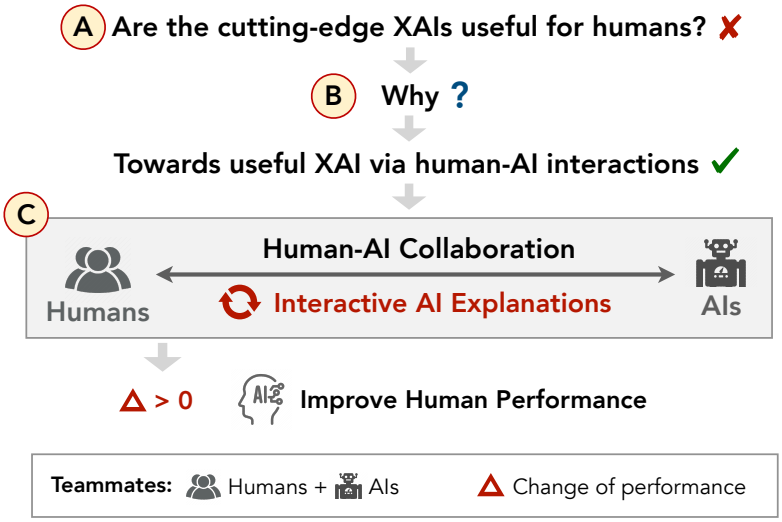


Figure 1.1: An overview of this doctoral thesis. To evaluate and improve the XAI usefulness, we investigate: (A) are cutting-edge XAI useful for humans? [211, 214] (B) what's the disparity between XAI and practical user needs? [212] and (C) how to empower useful XAI via interactions? Particularly, we develop two interactive AI explanation systems to improve human [210].

(Fig 1.1 (A)), we deployed two real-world human evaluation studies on analyzing computer vision AI model errors with post-hoc explanations [211], and simulating NLP AI model predictions with inherent explanations [214], respectively. The two studies unveil that, surprisingly, AI explanations are not always useful for humans to analyze AI predictions in practice. This motivates our research for RQ2 (Fig 1.1 (B)) – gaining insights into disparities between the status quo of AI explanations and practical user needs [212]. By surveying over 200 AI explanation papers and comparing them with summarized real-world user demands [138], we observe two dominating findings: *i)* humans request diverse XAI questions across the AI pipeline to gain a global view of AI system, where existing XAI approaches commonly display a single AI explanation that can not satisfy diverse XAI user needs; *ii)* humans are widely interested in understanding what AI systems can not achieve, which might lead to the need for interactive AI explanations that enable humans to specify counterfactual predictions.

In light of these findings, we deeply deem that, instead of designating user demands by XAI researchers during AI system development, **empowering users to communicate with AI systems for their practical demands is critical to unleashing useful AI explanations** (RQ3, Fig 1.1 (C)). To this end, we developed two interactive XAI systems that improved the usefulness of AI explanations in terms of human-perceived performance in AI-assisted writing tasks [210] (Fig 1.1 (C1)). Overall, we summarize this doctoral research by discussing the limitations and challenges of human-centered useful AI explanations.

1.2 Dissertation Outline

Following the three central themes in the aforementioned section, this thesis consists of three parts – *PART I* Human Evaluation on the Usefulness of AI Interpretability, *PART II* Disparity Between AI Explanations and Practical User Demands, and *PART III* Empower Useful AI Explanations with Human-AI Interactions.

First, I introduce the thesis introduction in Chapter 1 and an overview of AI explanations, evaluation and usefulness in Chapter 2.

PART I focuses on investigating if AI interpretability can be useful for humans in practice. We examine the usefulness of existing XAI methods, including both self-explaining models and post-hoc explanations techniques in diverse AI applications, such as image classification task in the computer vision domain and text classification tasks in NLP domain. Detailedly,

In Chapter 3, we present an investigation on whether showing visual explanations for humans helps them understand the model mistakes in image classification tasks [211].

In Chapter 4, we examine self-explaining models of NLP tasks, and conduct human evaluations on *if and when* the model generated rationales are useful for human predictions [214].

PART II further explores the disparities between the status quo of AI explanations and real-world user demands. Particularly,

In Chapter 5, we surveyed over 200 NLP XAI studies to uncover the gaps between XAI methods and practical user needs. We then present the potential directions to mitigate this misalignment [212].

PART III presents the proposed solution to improve human-centered useful interpretability – conversational AI explanations.

In Chapter 6, to bridge the gap between user demands and existing XAI methods, a conversational XAI system is proposed. This system incorporates multiple types of AI explanations into a unified interface, allowing users to ask various XAI questions through dialogue. Practical user demands for XAI are identified, and a prototype system called ConvXAI is developed based on these demands. ConvXAI is evaluated in the context of scientific writing and proves to be more useful in understanding AI feedback and improving writing quality compared to traditional GUI-based XAI systems. Insights into human usage patterns and core ingredients of useful XAI systems are also provided.

I finally conclude the dissertation in Chapter 7.

1.3 Contributions

The contributions of this thesis are summarized as: *i)* We provided diverse human evaluations of state-of-the-art XAI algorithms, which cover both post-hoc explanations and intrinsic explanations, on a range of computer vision and NLP tasks. *ii)* We further investigated the gaps between human demands and current XAI algorithms, and propose reasons and potential solutions to improve useful interpretability. *iii)* We are proposing a conversational AI interpretation system, named CONVXAI, which leverages back-and-forth conversations between humans and XAI assistants to address various user demands in supporting humans' scientific writing when collaborating with AI writing models.

Chapter 2

An Overview of AI Interpretability, Evaluation, and Usefulness

2.1 Scope and Definitions

> *Causal explanation takes the form of conversation and is thus subject to the rules of conversation.*
>
> *Denis J. Hilton*

In the realms of philosophy, cognitive science, social psychology, and machine learning, there exists an extensive body of research investigating AI explanation and interpretability. Prior to delving further into the thesis, it is imperative to expound upon the scope and definitions of the studies and terminologies in this dissertation.

Hilton [91] argues that causal explanation is first and foremost a form of social interaction. One speaks of giving causal explanations, but not attributions, perceptions, comprehensions, categorizations, or memories. The verb to *explain* is a three-place predicate: **Someone** explains **something** to **someone**. With regarding to this, Miller [154] proposes that the solution to explainable AI is not just "more AI", instead, it is a **human-agent interaction** problem [[154], p. 2]. Particularly, the AI agent could act as the explainer, while the human serves as the explainee during the human-agent interaction. Furthermore, Hilton [91] argues that explanation is a conversation, which involves two stages: "the diagnosis of causality in which the explainer determines why an action/event occurred; and the explanation, which is the social process of conveying this to someone. The problem is then to "resolve a puzzle in the explainee's mind about why the event happened by closing a gap in his or her knowledge" [[91], p. 66].

Despite the prevalent usage of "AI interpretation" and "AI explanation" as interchangeable terms in prior studies, this thesis proposes a more nuanced distinction between them. Specifically, "AI interpretation" pertains to the information derived from the AI agent (*i.e.,* explainer) that discerns the causality underlying its predictions. Conversely, "AI explanation" refers to the

information intended for human recipients (*i.e.,* explainees) and aims to address knowledge gaps in their understanding. Prior research and our own investigations [138, 154, 170, 212] reveal that much of the work in this domain relies on researchers' intuition to develop a "good interpretation" rather than a genuinely "helpful explanation" for humans. This observation serves as the impetus to establish clear terminological distinctions and explore the practical disparities between what AI agents **interpret** for their predictions and what humans require to be **explained** to bridge their knowledge gap in comprehending AI systems.

2.2 AI Interpretation Techniques

2.2.1 What Do We Explain?

Understanding *what* to explain is the foundation of further designing better interpretations. In this section, we discuss *(i)* interpreting model linguistic phenomena and the reasoning for NLP models' decisions. The latter can be roughly divided into the *(ii) global* and *(iii) local* interpretations.

2.2.1.1 Interpret Linguistic Phenomena inside NLP models

When developing NLP models, researchers expect the models to capture linguistic phenomena as human comprehending languages [200]. Thus many works investigate *whether the NLP models encode the linguistic properties* [193, 202].

Interpret static embedding. Earlier work explains the word static embeddings (*e.g.,* Word2Vec [153], Glove [178]) by mapping distributed vectors to human understandable units (*e.g.,* semantic concepts [125], sense [25]) [99, 227]. We can either interpreting the word embedding in its original space, such as explain each coordinate value to specific concepts [173, 206, 230] and rotating word dimensions to identify phenomena (*e.g.,* embedding bias) [53, 195]. Or we can map original embedding to an interpretable space, such as semantic concept space, for explanation [202, 205]. **Interpret contextual representations.** As large pre-trained language models

(*e.g.,* BERT [43], GPT [182]) achieve great success in NLP tasks [55], a growing number of research investigates their linguistic phenomena [17, 256]. A line of works aim to localize linguistic knowledge inside language models, such as different layers [85, 89, 108, 141, 143, 236] and different heads [33, 93, 122, 152, 193]. They commonly find that BERT tends to capture hierarchical linguistic knowledge with lower layer for word order information and higher layer for task-specific information. Additionally, many works examine what type of linguistic knowledge is learned [56, 108, 238, 270]. They find that language models might implicitly learn tree-like structures, capturing linguistic signals from surface to semantic features.

2.2.1.2 Interpret the Model Behavior as a Whole

A number of works aim to approximate the implicit decision rules of the NLP models for interpretation. This is considered as a *global explanation*, which explains model behavior as a whole [49]. Studies often achieve *global explanation* by creating interpretable proxy models (*e.g.,* a deterministic interpretation model [74, 275]) or learning explicit decision rules to approximate NLP model behavior [81, 166]. A representative work is *Anchors* [191], which learns the "If-Then rules" to explain model behavior. In other words, the model will give expected predictions with high precision when all the conditions in the rules are met. Furthermore, [261] uses the interpretation trees to explicitly represent the most important global decision rules contained in the NLP model. Additionally, *global explanation* can be represented in various approximate rules, such as syntactic or semantic rules [217, 240] and executable logical form [167, 204].

2.2.1.3 Interpret Instance-wise Model Prediction

In contrast to the *global interpretation*, the majority of NLP interpretation methods lie in generating *local interpretation*, also named *instance-wise interpretation* [66]. Here we examine extracting local interpretations from input instances and from dataset samples. **Identifying**

evidence within input instance. Given the input instance, this line of work aims to answer what the model reasoning process is or what part in input instance leads to the specific prediction. To reason model prediction process, studies often decompose NLP models' inside information flow [165, 166] or neurons [38, 45] into interpretable factors [76, 103]. Or they directly investigate model reasoning process paragraph [189], knowledge base [159], commonsense causal knowledge [160] or via rules [198] to understand model prediction. To examine what instance part is responsible for prediction, we often highlight the most informative fraction as "saliency maps" to show their importance. This is investigated in various NLP tasks like Summarization [265], Named Entity Recognition [67], Relation or Event Extraction [167, 235], Machine Translation [46, 244], NLI [20, 72], text classification [142, 250], hate speech classification [247], QA [41, 233] and conversational AI systems [115, 145, 197]. **Identifying**

responsible samples in datasets. Another subject of local interpretation is to find the most similar samples the model has "seen" during training stage to account for the given prediction. Influence function method [121] traces the model's prediction back to its training data to point out most responsible examples for the specific prediction. [83] applies influence function to BERT for NLP tasks, including sentiment analysis and natural language inference (NLI). An alternative method is *representer point selection* [263], which is more efficient than *influence function*.

2.2.2 Forms of Interpretations

Interpretation can be demonstrated in various formats [79]. We overview the interpretation forms in the 182 publications in a similar way as that of user types. Figure 2.1 shows the

statistics. Here, we describe nine forms of interpretations in NLP.

❶ **Visualization.** The majority of interpretation methods express NLP interpretability using visualization, including saliency map on text [27] and multimodal interpretation [186]. *Saliency map* interpretation means highlighting the sub-sequences in text (*e.g.,* tokens [126, 134, 162] or sentences [262]) to explain model's prediction using computed importance scores . *Multimodal* interpretation finds visual evidences from VQA related datasets to explain text using image features [136, 185, 234, 253, 259]. This is partially motivated by human language learning process using visual pointing [15].

❷ **Space Map.** Space map interpretation aims to project high-dimensional representations inside NLP models to human-understandable patterns, such as mapping vectors to decision boundary space [254] or the 2D space using dimensionality reduction algorithms (*e.g.,* t-SNE [147], UMAP [150]), then clustering them to find semantic meaning [228].

❸ **Free Text.** Free text interpretation means using natural language to explain model behavior. It contains more contextual explanations and is not constrained in input instances [126**?**]. Studies usually get free text explanations by human annotation [**?**], retrieval [260] or generative models [137].

❹ **Concept or Sense.** Concept or sense interpretation converts NLP model representations into human-interpretable concepts or sense [25, 202]. This is motivated by "explaining high-level concepts conforms more to human understanding language process thus might be easier interpretable by human" [119, 209]. The concepts and sense for explanation are obtained by either pre-defined before interpretation [125, 173, 202] or learning in an unsupervised manner [274].

❺ **Rules and Grammars.** Rules and grammars interpretation represents to approximately explain NLP models by executing a set of rules or grammars [86]. The formats of rules and grammars range from execution modules [112], logical operators [180], syntactic structures [217], POS tags [257] and so forth [84, 86, 231].

❻ **Tuples, Trees, and Graphs.** Tuples, trees and graphs interpretation is commonly adopted to explain model reasoning process [113, 126, 144]. Tuples depict factor relations in input texts or explanation graphs [226]. Graphs can represent more complex knowledge relations for explanation [111, 144, 260]. Trees emphasize on the hierarchical structures for interpreting model reasoning process [159, 228]. Besides, we find that graphs or trees explanations are often coupled with visualization or free text to extract key sub-sequences as their nodes of leaves [28, 261].

❼ **Examples and Demonstration.** Demonstrating examples means finding the most responsible training samples to explain the specific model prediction. This is motivated by the

assumption that training data defines the world view of a model and therefore influences its learning decision [121]. The representative methods of retrieving examples include influence functions [83, 121], representer point selections [263], cosine similarity ranking [228, 237] and so forth [94, 110].

❽ **Probing Tests.** Probing tests aim to check whether the representation in NLP models has learned specific task skills, such as syntactics [90] and semantics [141]. A set of trending studies use probing test to check language model (*e.g.,* GloVe [178]) embeddings [56, 108, 256], machine translation morphology [12], dialogue system skills [199].

❾ **Interactive Tools.** Above static interpretations can not capture target users' dynamic demands. Therefore, a number of studies propose interactive tools for NLP interpretability, such as explaining NLP models [92, 228, 229] and interpreting NLP tasks interactively [237, 246].

2.2.3 How to Interpret?

The interpretation methods can be categorized into *(i)* generating inherent interpretations inside models and *(ii)* developing post-hoc interpretability for pretrained models [79].

2.2.3.1 Inherent Interpretability

Inherent interpretability represents building self-explaining models to generate interpretations while making prediction [274]. Studies often leverages attention mechanism or neural network to generate inherent explanation in various NLP tasks such as text classification [250] and summarization [265].

Attention-based explanation. There are two contradicting viewpoints on *whether attention is explanation*. Some works use attention as the interpretation by positing that "a higher attention weight implies a greater impact on the model's prediction" [59, 157, 251]. Besides, studies discover that attentions do capture linguistic notions [174, 243], and use attention mechanism to explain NLP models [9, 72, 256]. In contrast, another line of work argues that "attention is not explanation" [80, 105]. This is supported by empirical studies of correlation between attention and gradient methods [11, 105], brittleness of perturbing attentions [105] and philosophy view [80]. Furthermore, some works give modest statement that interpretability of attention depends on input properties and model design [207, 242].

Model-generated explanation. Other methods add neural network components in NLP system to generate inherent interpretations while predicting [128, 176, 218]. These studies are investigated for various NLP tasks such as explaining commonsenseQA task [**?**], self-explaining QA models [112, 114] and dialog systems [144, 145].

2.2.3.2 Post-hoc Interpretability

The majority of interpretation methods in NLP domain aim to provide post-hoc interpretation on the pre-trained neural networks. We mainly introduce instance-wise post-hoc interpretations including *perturbation-based*, *backprop-based*, and *inner-representation-based* methods.

Perturbation-based interpretations aim to generate the interpretations by observing how the output value changes as input is "erased" [42], "marginalized" [120] or "perturbed" [42]. For instance, [134] removes the embedding of targeted sub-sequences (*e.g.,* tokens) and observe the degradation on model log-likelihood of the correct label. [5] develops a causal framework for explaining Seq2Seq model prediction. Other than perturbing to investigate importance score, [201] leverages perturbation method to check if the dialogue systems use the conversational history effectively.

Backprop-based interpretations leverage the back propagation way from outputs back to inputs to obtain the interpretation [162]. Gradient (or its variant) is commonly used to compute the importance scores [45, 133]. Besides, [133] applies layer-wise relevance propagation (LRP) and [76, 165] propose Contextual Decomposition (CD) algorithm to generate local explanations for NLP models like LSTM.

Inner-representation-based interpretations aim to explain the intermediate-layer representation of the NLP model. Particularly, the representative method in this branch is probing test [3, 75, 238], also named as "auxiliary prediction tasks" [3] or "diagnostic classifiers" [75, 98]. Studies use probing tasks for explaining both embeddings [98] and model hidden layer representations [57, 199, 217]. Probing tasks can examine both sentence-based [3, 35] as well as word-based properties [217]. Also, probing tasks can investigate linguistic properties such as verb tense or part-of-speech [12, 143], and syntactic relationships [89, 238]. However, it is debating whether the high accuracy of probing classifiers stems from representations or probe, thus control task is presented to mitigate this issue [88].

2.3 Evaluation on AI Interpretability

Although many interpretations have been proposed, how to evaluate interpretation is still a challenging problem. Here we examine NLP interpretability evaluation from both model (*i.e.,* "faithfulness" [101]) and human (*i.e.,* "plausibility" [157]) perspectives.

2.3.1 Faithfulness to Models

Faithfulness metric captures how accurately the interpretation can reflect model behavior [87, 251]. A common philosophy of evaluation faithfulness, borrowed from image interpretation, is to compute *smallest sufficient region* (*i.e.,* the smallest instance region that supports a correct

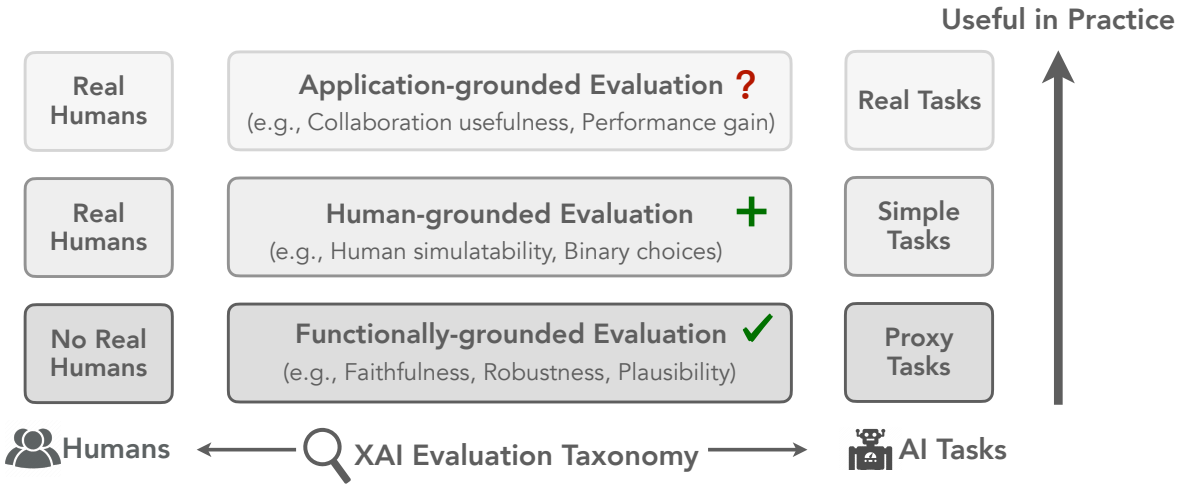


Figure 2.1: The taxonomy of evaluating AI explanations.

prediction) or *smallest destroying region* (*i.e.,* the smallest instance region that when removed, converts to an incorrect prediction) [36, 65].

Similarly in NLP tasks, [44] proposes two corresponding faithfulness metrics as "sufficiency" and "comprehensiveness". Particularly, "sufficiency" measures when only keeping important tokens for prediction, how model confidence on correct label drops (*i.e., lower the better*); "comprehensiveness" examines when removing important tokens, how corresponding model confidence degrades (*i.e., higher the better*). A number of studies adopts faithfulness metrics to evaluate their interpretation methods on NLP tasks like QA and VQA [62], visual [252] and text reasoning tasks [82]. As an alternative method, [45] computes the ablation score for the interpreted sub-sequences, and conducts the Pearson correlation between the ablation scores and interpretation scores to check its faithfulness.

2.3.2 Plausibility to Humans

Plausibility metric examines how convincing the interpretation is to humans [101]. Nevertheless, compared with faithfulness, plausibility gets much less concern [11, 42].

A feasible metric to measure plausibility is "human rationales alignment" evaluation, which compares how much the model's explanation is well agreed with human's rationales [273]. Based on whether the interpretation and human rationales are discrete selection or continuous score, this metric is divided into "hard-alignment" and "soft-alignment" evaluation, respective [44]. **Hard-alignment** metric measures overlapping degree between machine-generated interpretations and human rationales, such as using Intersection-Over-Union (IOU) score [44]. **Soft-alignment** metric compares the two vectors of interpretation and human rationale using similarity or correlation methods [168]. For instance, [40] introduces a human attention dataset for VQA task. It further quantitatively measures the human and machine attentions via rank-order correlation. [204] conducts a similar work on text classification task, which introduces a

set of similarity metrics (*e.g.,* PairwiseSim). Besides, Area Under the Precision-Recall curve (AUPRC) is adopted by [44] to obtain the evaluation. However, these alignment metrics rely on the ground truth of human rationales, which are often lacking in practice. Therefore, more human involved evaluation protocols are needed to mitigate this issue, such as designing word intrusion test [230], simulatability test [86] or model-user interaction [59, 204] for evaluating explanation.

Overall, there is a lack of studies on *how automatic metrics correlate to human evaluation* [273]. Although [168] shows several automatic measures moderately correlate with human evaluators' accuracy, more comprehensive studies are needed for connecting faithfulness and plausibility related measures.

2.4 Challenges on Useful AI Explanations for Humans

[1] categorizes interpretation utilities as justifying, controlling, improving and discovering. We review interpretation utility, from human-centered perspective, to see how NLP interpretability impacts human and how human-generated interpretations boost NLP models.

2.4.1 Boost the Performance of AI Models

Many works use human-generated interpretations to augment NLP models. The majority studies aim to use "human annotated rationales" as augmented data other than labels to train better models [266]. For instance, [126] and [51] build datasets, which incorporate natural language or selected sub-sequences explanations, for various NLP tasks such as relation extraction and text classification. However, leveraging natural language (NL) explanations for improving models meets challenging: NL explanations are unstructured which need further semantic extraction and they often involve linguistic variants [249]. To remedy, NL explanations are converted to executable logical forms [249] or programmatic labeling functions [84] before generating augmented data.

Interpretations are also used in debugging NLP model mistakes and debiasing NLP representations. For instance, [123] proposes *Explanatory Debugging* approach on text classification tasks. It allows users to provide any correction feedback after viewing interpretations. Similarly, [171] presents *Pandora*, a set of hybrid human-machine methods to describe, explain and characterize image captioning system failures. Besides, since NLP models show potential bias stemmed from large data corpus, such as gender [16], racial [69] and religious bias [148], studies also leverage interpretation to detect bias [17, 53]. For example, [5] proposes an interpretation method to explain causally related input-output tokens. Then it is used to diagnose biased translations in machine translation systems.

2.4.2 Make an Impact on Humans

Interpretation methods are widely motivated by improving *human understanding* [211], *simulatability* [191], *trust* [41, 223] and *human task performance* [31]. However, whether interpretations indeed make these impacts on human is less investigated.

[223] conducts human studies on evaluating several human perception metrics on NLP interpretations, which include trustworthiness, comprehension, accuracy, feedback importance and so on. [86] measures *human simulatability* by showing human only input and its explanation and ask human to its predict model output. [60] propose an interpretation evaluation method on domain experts participated task (*i.e.,* Quizbowl), in which it designs a human-computer cooperative game to measure how much the interpretation can effectively improve human performance. [73] investigates the explanation impacts on the annotation quality and the annotators' experience in a human-in-the-loop active learning setting. It conducts empirical study comparing the model learning outcome, human feedback content and annotator experience with/without explanation.

Nevertheless, several existing studies show negative impacts of model interpretations on human performance. Evaluated on email classification tasks, [223] finds that when model quality is low (*i.e.,* with accuracy as 76.5%), highlighting important words for interpretations reduces human trust and acceptance and increases frustration. With higher quality model (*i.e.,* 94.4% accuracy), explanations increased human understanding. Similarly, by evaluating five explanation methods on human simulation tests of text and tabular data, [86] finds only LIME for tabular data and the proposed Prototype method improve human simulatability. While only a few NLP works exist in this area, some human studies in visual interpretations also show negative effects [31, 211].

Part I

Human Evaluation on the Usefulness of AI Interpretability

Chapter 3

Human Evaluations on AI Error Analysis Using Post-hoc Interpretations

3.1 Introduction

Explaining to users why automated systems make certain mistakes is important. As deep neural network technologies achieve higher performance, they have been applied to important domains, influencing important decisions in healthcare, transportation, and education. However, due to the non-linear, complicated structures of neural models, the high performance of deep neural networks is achieved at the cost of interpretability. In response, researchers have proposed ways to explain the inner workings of deep neural networks by automatically producing interpretations [151, 190, 203]. Such machine-generated interpretations help various stakeholders [229]: researchers, who develop new deep-learning architectures; machine-learning engineers, who train and optimize existing networks; product engineers, who apply general-purpose pre-trained networks to various tasks; and the general users, who want to understand system outputs [32, 203, 222]. This paper focuses on the **end users** – who may not understand the mechanism of the underlying deep neural networks, but are most influenced by their outputs – to investigate whether machine-generated interpretations can help users make sense of errors made by algorithms.

We use the image-classification task as our test bed. Neural image classifiers generate interpretations through two approaches: designing proxies, which are inherently interpretable (*e.g.*, decision tree), to substitute the black-box deep neural networks [151]; or generating post-hoc interpretations outside the deep neural network workflow [203], which is where our work will focus. Most post-hoc interpretations are in the form of instance-wise interpretation – for example, saliency maps of input images. A saliency map highlights the most informative region of the image with respect to its classification label, unveiling post-hoc evidence of the neural network prediction. This line of work was in part motivated by the need of "end users" [50, 170], "non-expert users" [190], or "untrained users" [203], and the generated interpretations were often evaluated by how much they could boost users' trust of deep neural networks. However, it is still unclear how **useful** these interpretations are in helping users make sense of automated system errors.

The need for interpretability arises due to *Incompleteness* in the problem formalization, making it difficult to make further judgements or optimizations [48]. When a user observed a few cases where the automated system incorrectly labeled his/her images, it was difficult for the user to decide what to do. Did the errors occur because the system's accuracy level is low? If so, should the user switch to another system? Are the images too complicated for computers, in which case users should not expect reliable image labels? Did the underlying algorithms have biases that worsened with certain types of images? We argue that errors *expose* existing incompleteness in the problem formalization, requiring users to seek interpretations. Namely, an important use case of interpretations is to help users figure out what is going on when they get certain errors. Researchers have proposed evaluations to assess how much an interpretation reflects the model's behavior (also known as "fidelity") [151] or boosts users' trust in automated systems [190, 203]. However, it is unclear how *useful* these interpretations are in helping users figure out why they are getting an error.

This paper introduces a method that uses crowd workers from Amazon Mechanical Turk (MTurk) to directly evaluate the usefulness of interpretations in helping users to reason about the errors of deep neural networks[1]. Figure 3.1 overviews the workflow. In this task, each worker is presented with an image and told that the deep neural network incorrectly predicted its label. The worker may also be presented with a set of interpretations (*e.g.*, saliency maps) that explain how the deep neural network "perceives" this image and makes the final prediction. The worker is then asked to **guess the incorrectly predicted label** from five options, four of them being distractors. If an interpretation effectively explains how the underlying deep neural network model works to users, the people who were presented with the interpretation should be better at predicting the model's outputs than those who were not.

This paper tried to answer two research questions: First **(RQ1)**, do machine-generated visual interpretations help human users better identify predicted labels? Second **(RQ2)**, when do (and when do not) the visual interpretations help?

3.2 Related Work

3.2.1 Interpretation Methods

Our work focuses on post-hoc interpretations. These methods generate saliency maps to indicate where the neural networks "look" in the images for their predictions' evidence. Existing methods can be categorized into four lines: *Backprop-Based*: computes the gradient (or variants) of the neural network output to score the importance of each input pixel, such as SmoothGrad [221]; *Representation-Based*: uses the feature maps at intermediate layer of neural networks to generate saliency maps, like GradCAM [203]; *Meta-Model-Based*: trains a meta-model to predict the saliency map for any given input in a single feed-forward pass, such as RTS [37]; *Perturbation-Based*: finds the saliency map by perturbing the input with

[1]The code and interface are available via GitHub:
`https://github.com/huashen218/GuessWrongLabel`

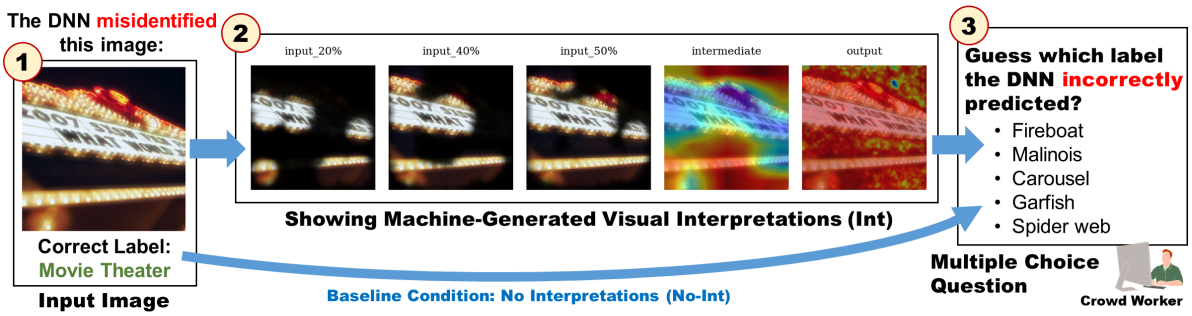


Figure 3.1: The workflow of "Guessing the Incorrectly Predicted Label" task. Each worker is presented with an image and told that the deep neural network incorrectly predicted its label (Step 1). The worker may also be presented with visual interpretations (Step 2). The worker is then asked to guess the incorrectly predicted label ("Carousel" in this example) from five options, four of them being distractors (Step 3). If an interpretation effectively explains how the underlying deep neural network model works to users, the people who were presented with the interpretation should be better at predicting the model's outputs.

minimum intervention and observing the change in model prediction, like ExtremalPerturb [64] and SHAP [146]

3.2.2 Evaluating Interpretations

Evaluating the effectiveness of interpretations is critical in practice. Existing evaluations answer two questions: whether the interpretations genuinely reflect neural network behavior [2], and whether the interpretations are useful for users. To answer the latter question, a set of metrics are proposed to involve human evaluation. For instance, trust assessment and user satisfaction is verified in [222] by surveying general users. Mental model evaluations designed by [19] and [32] measure whether general users can understand and predict model outputs. [61] creates a human-computer cooperative task to measure how much interpretation improves human performance. However, more study is needed to investigate how general users perceive and predict neural networks' failure cases, which is of vital importance in building trust and correcting model behavior.

3.2.3 Human-AI Collaboration

Although human computation has traditionally played a data annotation role in deep learning systems, there is increasing interest in incorporating it into diverse stages of human-AI hybrid systems [170]. Due to its goal of building human understanding and trust in black-box neural networks, interpretation is inherently a human-centric problem. Related efforts involve human perception of different types of interpretation representations in visual interfaces [196], etc.

3.3 Method

3.3.1 Human Study Design

We used a deep neural network to label images and employed several interpreters to generate visual interpretations for the images. We showed each image the deep neural network had labeled incorrectly to a group of online crowd workers and asked them to guess which images the deep neural network had mistakenly labelled. Only the workers in the control group were presented with the visual interpretations. We detail the procedure of the study in this section.

Step 1: Labeling Images We trained an image classifier on ImageNet dataset, with its TOP-1 accuracy reaching 78.67% [258]. We randomly selected images whose labels were incorrectly identified by the classifier.

Step 2: Generating Instance-Wise Interpretations For each image in the misclassified subset, we used three existing interpreters – *i.e.,* input perturbation [64], intermediate feature extraction [203], and output backpropagation [221] – to explain three aspects of this image. **Input perturbation interpretation (column 2-4 in Figure 3.2)** observes how the output value changes as input is "deleted" in different sub-regions. We used *ExtremalPerturb*, which aims to find a small pixel subset that, when preserved, are sufficient to keep model output stable. Moreover, ExtremalPerturb allows researchers to explicitly constrain the percentage of preserved pixels. We provided three levels of percentage: $a = \{20\%, 40\%, \text{and } 50\%\}$. **Inter-Feature extraction interpretation (column 5 in Figure 3.2)** looks at intermediate layers of the neural network to indicate the discriminative image regions used by the model for prediction. We used *GradCAM*, which extracts the gradient information flowing into the last convolutional layers, to explain the importance of each pixel. **Output backpropagation interpretation (column 6 in Figure 3.2)** leverages backpropagation to track information from the model's output back to its input to generate the saliency map. We used *SmoothGrad*, which samples similar images by adding noise to the original image and using the average of the resulting heatmaps to obtain the final interpretation. We eventually generated *(i)* three saliency maps from input perturbation view with 20%, 40% and 50% percentages respectively, *(ii)* one saliency map from intermediate feature extraction view, and *(iii)* one saliency map from the output backpropagation view.

Step 3: Having Crowd Workers Guess the Incorrectly Predicted Label Next, we recruited crowd workers on MTurk to complete tasks[2]. The workers were shown the image and its correct label, and were informed that "a computer algorithm misidentified this image as something else." Only the workers in the control group, as shown in Figure 3.1, were presented with the visual interpretations. On the interface, we explained that the visual interpretations are "visualizations that try to show how the algorithm *sees* this image," and provided comprehensive descriptions for each interpretation. For example, we explained "input

[2]Each Human Intelligence Task (HIT) contained one image, and multiple workers were recruited to answer the question. The price of a HIT is $0.05. Four built-in MTurk qualifications are used: Locale (US Only), HIT Approval Rate ($\geq$98%), Number of Approved HITs ($\geq$3000), and the Adult Content Qualification.

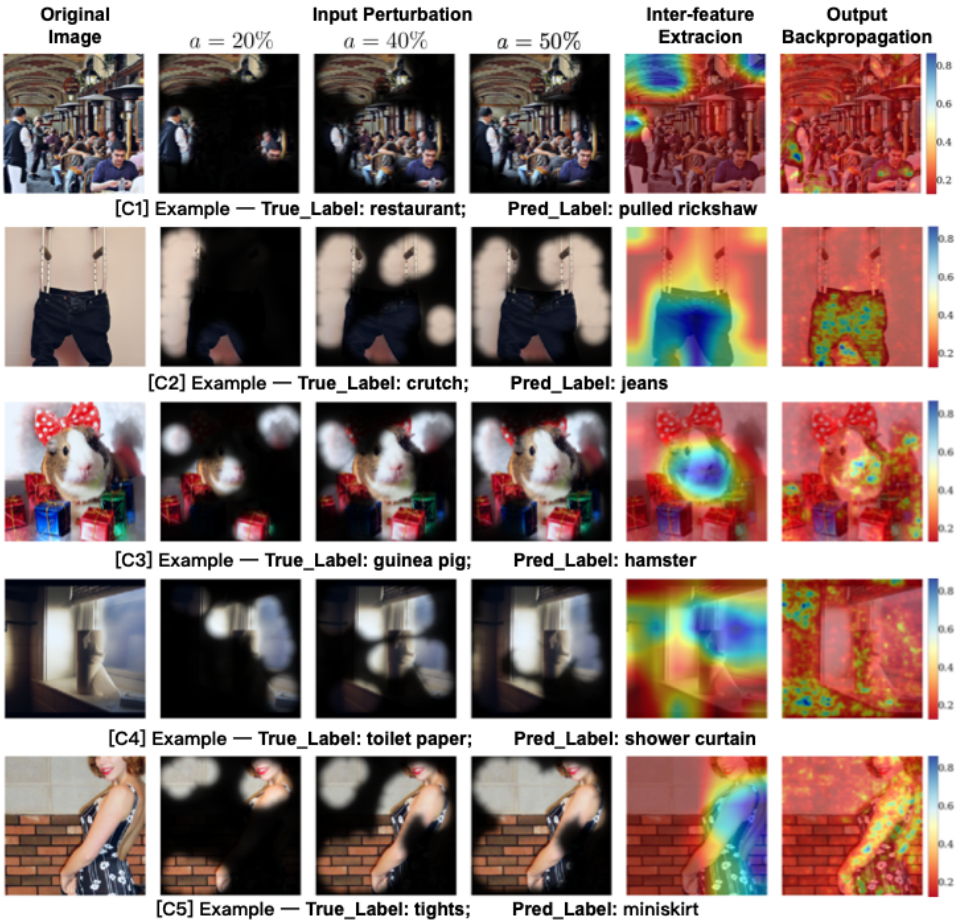


Figure 3.2: Examples of five types of errors in image classification. The visual interpretations are generated by three existing interpreters (see "Step 2" in the Method section.)

perturbation interpretation" with a 20% mask (column 2 in Figure 3.2) as "We only allow the algorithm to see 20% of the image and ask the algorithm to choose which 20% is the most important region. The black mask blocks the regions the algorithm pays less attention to." The workers are then asked to **guess the incorrectly predicted label** from five options. One of the options was the incorrect label predicted by the deep neural network model, and the remaining four were randomly selected from the whole label set of ImageNet (*i.e.,* 1,000 labels), excluding the correct gold-standard label.

The assumption is that if the visual interpretations effectively explain how the deep neural network works, the workers who were presented with the interpretations should distinguish the predicted label better than those who were not. Humans alone are sufficient to guess the *correct* label, but it requires workers to take the mechanism of deep neural networks into account to guess the *incorrect* label predicted by deep neural networks. MTurk workers are appropriate participants because they represent general users who do not necessarily understand deep neural network models nor are trained for reasoning about these models' errors.

3.3.2 Usefulness of Fine-grained Model Error Cases

To inspect usefulness of interpretation in fine-grained model failure scenarios (RQ2), we further manually categorized the model error cases and check the explanations' usefulness for each case. Particularly, the authors inspected 1,000 misclassified images and categorized them into five types of errors (Figure 3.2), in part based on the literature [6].

1. **Local Character Inference (C1):** The model arrives at wrong prediction by looking at only part of the object. For instance, in Figure 3.2(C1), the error might be due to the model partially capturing the restaurant dome, which looks similar to the canopy of a pulled rickshaw.
2. **Multiple Objects Selection (C2):** For images with multiple objects, the model makes a prediction by choosing another object rather than the labeled one, as in Figure 3.2(C2).
3. **Similar Appearance Inference (C3):** The model misclassifies the object in the image into another class with a similar appearance, as shown in Figure 3.2(C3).
4. **Correlation Learning (C4):** The model exploits correlational relationships in training data to apply an incorrect label to the image. For example, in Figure 3.2(C4), the model predicts a "shower curtain" by identifying the bathroom context, even if no curtain is in the image.
5. **Incorrect Gold-Standard Labels (C5):** The true label of the images might be incorrect in the ImageNet. Figure 3.2(C5) shows an example.

3.4 Experimental Results

3.4.1 Experiment 1: Testing Two Conditions in the Same Batch of HITs

Experiment 1 had two conditions: [Interpretation] (*i.e.*, [Int]) and [No-Interpretation] (*i.e.*, [No-Int]). The only difference is that HITs in the [No-Int] group do not show the interpretations to workers in interfaces. We evenly divided 200 randomly selected image samples into two groups. We posted these 200 images in a same batch of HITs at the same time on MTurk, where each HIT recruits nine different workers. A total of 1,800 submissions (900 submissions in each condition) were contributed by 41 workers in [Int] and 40 workers in [No-Int] conditions respectively. We did not control the workers' participation, so a worker could participate in both groups. Thirty-six out of 45 workers participated in both conditions.

Surprisingly, in Experiment 1, **showing the workers machine-generated visual interpretations *reduced* their average accuracy in guessing the incorrectly predicted labels.** We calculated the accuracy as the percentage of correctly inferring the classifier's prediction among all 900 submissions in each condition. The accuracy collected in [Int] was 0.73, while the accuracy in [No-Int] was 0.81. The difference was statistically significant (unpaired t-test,

	C1	C2	C3	C4	C5	Overall
Int	0.77	0.83	0.71	0.54	0.71	0.73
#images	29	23	28	15	5	100
No-Int	0.76	0.77	****0.87**	****0.75**	0.78	***0.81**
#images	25	10	47	12	6	100

Table 3.1: Results of Experiment 1. Showing the workers machine-generated visual interpretations *reduced* their average accuracy in guessing the incorrectly predicted labels. (Unpaired t-test. *: $p<0.05$, **: $p<0.01$.)

$p<0.05$, N=100). Based on the results, the machine-generated interpretation did not help, but instead hurt, the workers' ability to guess the incorrectly predicted labels. The by-category analysis (Table 3.1) shows that displaying interpretations significantly lowers human accuracy in cases where the errors were probably caused by similar appearances between items (C3) or by mistakenly learning from the background or scenes of the image (C4).

3.4.2 Experiment 2: Testing with Two None Overlapping Sets of Workers

Experiment 2 was controlled more strictly. We randomly selected another 200 images (different from those used in Experiment 1), and used the same photo in both [Int] and [No-Int] conditions. We used custom MTurk qualifications to control the participants: workers who participated in one condition could not accept HITs in the other condition. We recruited 10 different workers for each image, in which five workers were in the [Int] group and the other five were in the [No-Int] group. A total of 2,000 submissions (with 1,000 submissions in each condition) were collected, contributed by 42 workers in the [Int] condition and 63 workers in the [No-Int] condition respectively.

In Experiment 2, the machine-generated visual interpretation again **reduced the average human accuracy in inferring model misclassification** (Table 3.2.) The accuracy of [Int] was 0.63, whereas accuracy in [No-Int] condition was 0.73. The difference was again statistically significant (paired t-test, $p<0.01$, N=200). On average, humans do not benefit from interpretations when inferring incorrect predictions in image classification tasks. Similarly to Experiment 1, the by-category analysis showed that displaying interpretations significantly lowers human accuracy in C3 and C4 (Table 3.2) errors. We also noticed that the accuracy for C1 and C2 images increased in both experiments when showing visual interpretations, although the differences were not statistically significant.

	C1	C2	C3	C4	C5	Overall
Int	0.57	0.74	0.66	0.41	0.67	0.63
No-Int	0.52	0.71	****0.84**	***0.59**	0.77	****0.73**
#images	44	20	112	18	6	200

Table 3.2: Results of Experiment 2. The machine-generated visual interpretation again *reduced* the average human accuracy in inferring model misclassification. (Paired t-test. *: $p<0.05$, **: $p<0.01$.)

3.5 Discussion

Our experiments showed that, in the case of image classification, machine-generated visual interpretations are not necessarily useful in helping users understand deep neural network failures. It could even be harmful, as in the cases where the errors were probably caused by similar appearances between items (C3) or by mistakenly learning from the background or scenes of the images (C4). System designers should use caution when displaying machine-generated interpretations to users.

3.5.1 Why it did not help?

More research is required to discover why showing interpretations was ineffective. Here, we submit several of our hypotheses with the goal of helping future explorations. First, the interpreters are not good enough to help humans. The representational power – including the correctness, sensitivity, etc., of the interpretation model – might not be sufficient to augment human reasoning about errors. Although machine-generated interpretations captured some of the deep neural network's behaviors, it may not be good enough to help humans. Second, the format is insufficient. The saliency maps may not be the most efficient format to convey information to humans. For example, when a saliency maps model changes an inner parameter, this change might not be obvious enough to be noticeable by humans, but could still affect the final predictions. Third, the interpreters may work poorly in cases where the image classifier failed.

3.5.2 Limitations

We are aware that this work has several limitations. First, the sample size was relatively small. Given that classifiers incorrectly labelled more than 10,000 images in the ImageNet validation set alone, 200 images are relatively small portion of the data. Second, we only tested three particular types of interpretations, and also presented the interpretations together on the same page. This experimental setup introduces the possibility of missing out on the "best" interpretations, or different interpretations might affect each other and reduce their effectiveness. Third, we recruited MTurk workers with certain qualifications to simulate general users. It is difficult to eliminate data noise stemmed from workers' misunderstanding or incognizance

of images or options. Finally, we only tested visual interpretations for image classifiers. It requires more research to study if similar effects could be generalized to other tasks.

3.6 Conclusion

The goal of this study was to evaluate the usefulness of machine-generated visual interpretations for general users' reasoning about model errors. To this end, we utilized the "guess incorrectly predicted labels" task to examine the usefulness of visual interpretations. Our two sets of control experiments, with 3,800 submissions contributed by 150 online crowd workers, suggest that showing the interpretations does not increase, but rather *decreases*, the average accuracy of human guesses by roughly 10%.

Chapter 4

Human Evaluations on Simulating AI Predictions with Intrinsic Interpretations

4.1 Introduction

While neural networks have recently led to large improvements in NLP, most of the models make predictions in a black-box manner, making them indecipherable and untrustworthy to human users. In an attempt to faithfully explain model decisions to humans, various work has looked into extracting *rationales* from text inputs [106, 176], with *rationale* defined as the "shortest yet sufficient subset of input to predict the same label" [10, 128]. The underlying assumption is two-fold: (1) by retaining the label, we are extracting the texts used by predictors [106]; and (2) short rationales are more readable and intuitive for end-users, and thus preferred for human understanding [241]. Importantly, prior work has knowingly traded off some amount of model performance to achieve the shortest possible rationales. For example, when using less than 50% of text as rationales for predictions, Paranjape et al. [176] achieved an accuracy of 84.0% (compared to 91.0% if using the full text). However, the assumption that the shortest rationales have better human interpretability has not been validated by human studies [212]. Moreover, when the rationale is too short, the model has much higher chance of missing the main point in the full text. In Figure 4.1A, although the model can make the correct positive prediction when using only 20% of the text, it relies on a particular adjective, "life-affirming," which is seemingly positive but does not reflect the author's sentiment. These rationales may be confusing when presented to end-users.

In this work, we ask: *Are shortest rationales really the best for human understanding?* To answer the question, we first design LIMITEDINK[1], a self-explaining model that flexibly extracts rationales at any target length (Figure 4.1A). LIMITEDINK allows us to control and compare rationales of varying lengths on input documents. Besides **controls on rationale length**, we also design LIMITEDINK's sampling process and objective function to be **context-aware** (*i.e.,* rank words based on surrounding context rather than individually, Figure 4.1B_2) and **coherent** (*i.e.,* prioritize continuous phrases over discrete tokens, Figure 4.1C_2). Compared to existing

[1]Find open-source code at: `https://github.com/huashen218/LimitedInk.git`

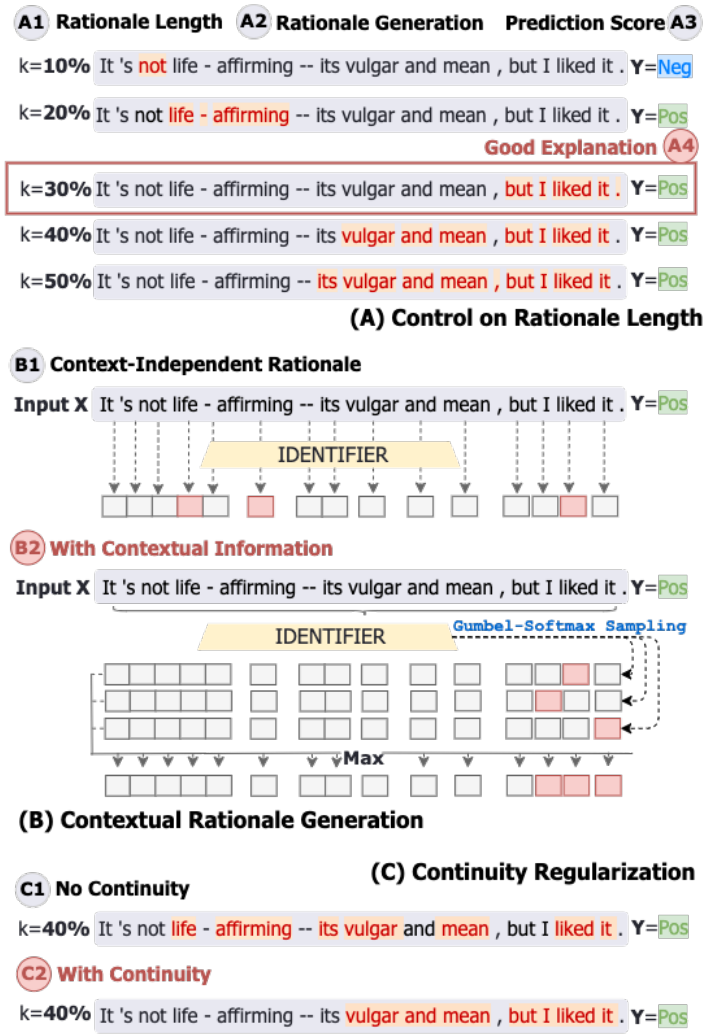


Figure 4.1: LIMITEDINK's rationale generation with length control: (A) control rationale generation with different lengths; (B) incorporating contextual information into rationale generation; (C) regularizing continuous rationale for human interpretability. Examples use the SST dataset for sentiment analysis [224].

baselines (*e.g.,* `Sparse-IB`), LIMITEDINK achieves compatible end-task performance and alignment with human annotations on the ERASER [44] benchmark, which means it can represent recent class of self-explaining models.

We use LIMITEDINK to conduct user studies to investigate the effect of rationale length on human understanding. Specifically, we ask MTurk participants to predict document sentiment polarities based on only LIMITEDINK-extracted rationales. By contrasting rationales at five different length levels, we find that shortest rationales are largely not the best for human understanding. In fact, humans do not perform better prediction accuracy and confidence better than using randomly masked texts when rationales are too short (*e.g.,* 10% of input texts). In summary, this work encourages a rethinking of self-explaining methods to find the right balance

Method	Movies				BoolQ				Evidence Inference				MultiRC				FEVER			
	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1
Full-Text	.91	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41
Sparse-C	.82	.17	.36	.23	.44	.15	.11	.13	.41	.03	.15	.05	.62	.15	**.41**	.22	.83	.35	.52	.42
Sparse-IB	.84	.21	.42	.28	.46	**.17**	.15	.15	.43	.04	.21	.07	.62	.20	.33	.25	.85	**.37**	.50	**.43**
LIMITEDINK	**.90**	**.26**	**.50**	**.34**	**.56**	.13	**.17**	**.15**	**.50**	**.04**	**.27**	**.07**	**.67**	**.22**	.40	**.28**	**.90**	.28	**.67**	.39
Length Level	50%				30%				50%				50%				40%			

Table 4.1: LIMITEDINK performs compatible with baselines in terms of end-task performance (**Task**, weighted average F1) and human annotated rationale agreement (**P**recision, **R**ecall, **F1**). All results are on test sets and are averaged across five random seeds. For LIMITEDINK, we report results for the best performing *length level*.

between brevity and sufficiency.

4.2 LIMITEDINK

4.2.1 Self-Explaining Model Definition

We start by describing typical self-explaining methods [10, 128, 176]. Consider a text classification dataset containing each document input as a tuple $(\mathbf{x}, y)$. Each input $\mathbf{x}$ includes n features (*e.g.,* sentences or tokens) as $\mathbf{x} = [x_1, x_2, ..., x_n]$, and y is the prediction.

The model typically consists of an *identifier* $\mathbf{idn}(\cdot)$ to derive a boolean mask $\mathbf{m} = [m_1, m_2, ..., m_n]$, where $m_i \in \{1, 0\}$ indicates whether feature x_i is in the rationale or not. Note that the mask $\mathbf{m}$ is typically a binary selection from the *identifier*'s probability distribution, i.e., $\mathbf{m} \sim \mathbf{idn}(\mathbf{x})$. Then it extracts rationales $\mathbf{z}$ by $\mathbf{z} = \mathbf{m} \odot \mathbf{x}$, and further leverages a *classifier* $\mathbf{cls}(\cdot)$ to make a prediction y based on the identified rationales as $y = \mathbf{cls}(\mathbf{z})$. The optimization objective is:

$$\min_{\theta_{\mathbf{idn}}, \theta_{\mathbf{cls}}} \underbrace{\mathbb{E}_{\mathbf{z} \sim \mathbf{idn}(\mathbf{x})} \mathcal{L}(\mathbf{cls}(\mathbf{z}), y)}_{\text{sufficient prediction}} + \underbrace{\lambda \Omega(\mathbf{m})}_{\text{regularization}} \tag{4.1}$$

where $\theta_{\mathbf{idn}}$ and $\theta_{\mathbf{cls}}$ are trainable parameters of *identifier* and *classifier*. $\Omega(\mathbf{m})$ is the regularization function on mask and λ is the hyperparameter.

4.2.2 Generating Length Controllable Rationales with Contextual Information

We next elaborate on the definition and method of controlling rationale length in LIMITEDINK Assuming that the rationale length is k as prior knowledge, we enforce the generated boolean mask to sum up to k as $k = \sum_{i=1}^{n}(m_i)$, where $\mathbf{m} = \mathbf{idn}(\mathbf{x}, k)$. Existing self-explaining

methods commonly solve this by sampling from a Bernoulli distribution over input features, thus generating each mask element m_i independently conditioned on each input feature x_i [176]. For example, in Figure 4.1B_1), "life affirming" is selected independent of the negation context "not" before it, which contradicts with the author's intention. However, these methods potentially neglect the contextual input information. We leverage the concrete relaxation of subset sampling technique [29] to incorporate contextual information into rationale generation process (see Figure 4.1B_2), where we aim to select the top-k important features over all n features in input $\mathbf{x}$ via Gumbel-Softmax Sampling (*i.e.,* applying the Gumbel-softmax trick to approximate weighted subset sampling process). To further guarantee precise rationale length control, we deploy the *vector and sort* regularization on mask $\mathbf{m}$ [64]. See more model details in Appendix A.0.1.

4.2.3 Regularizing Rationale Continuity

To further enforce coherent rationale for human interpretability, we employ the Fused Lasso to encourage continuity property [10, 106]. The final mask regularization is:

$$\Omega(\mathbf{m}) = \lambda_1 \underbrace{\sum_{i=1}^{n} |m_i - m_{i-1}|}_{\text{Continuity}} + \lambda_2 \underbrace{\| \text{vecsort}\,(m) - \hat{m}\|}_{\text{Length Control}} \tag{4.2}$$

For BERT-based models, which use subword-based tokenization algorithms (*e.g.,* WordPiece), we assign each token's importance score as its sub-tokens' maximum score to extract rationales during model inference (see Figure 4.1C).

4.3 Model Performance Evaluation

We first validate LIMITEDINK on two common rationale evaluation metrics, including end-task performance and human annotation agreement.

4.3.1 Experimental Setup

We evaluate our model on five text classification datasets from the ERASER benchmark [44]. We design the *identifier* module in LIMITEDINK as a BERT-based model, followed by two linear layers with the ReLU function and dropout technique. The temperature for Gumbel-softmax approximation is fixed at 0.1. Also, we define the *classifier* module as a BERT-based sequence classification model to predict labels. We train five individual self-explaining models of different rationale lengths with training and validation sets, where we set the rationale lengths as $\{10\%, 20\%, 30\%, 40\%, 50\%\}$ of all input text. Then we select one out of the five models, which has the best weighted average F1 score, to compare with current baselines on end-task performance and human annotation agreement on test sets. Note that we use all models with five rationale lengths in human evaluation described in Section 4.4.

Figure 4.2: Key components of the User Interface in the MTurk *task* HITs. Note that each HIT contains five reviews with different rationale lengths.

Baselines. We compare LIMITEDINK with four baselines. `Full-Text` consists of only the *classifier* module with full-text inputs. `Sparse-N` enforces shortest rationales by minimizing rationale mask length [10, 128]. `Sparse-C` controls rationale length by penalizing the mask when its length is less than a threshold [106]. `Sparse-IB` enables length control by minimizing the KL-divergence between the generated mask with a prior distribution [176]. See Appendix A.0.1 for more model and baseline details.

4.3.2 Evaluation Results

End-Task Performance. Following metrics in [44], we report the weighted average F1 scores for end-task classification performance. Among five LIMITEDINK models with different rationale lengths, Table 4.1 reports the model with the best end-task performance on the test set. We observe that LIMITEDINK performs similarly to or better than the self-explaining baselines in all five datasets. See ablation studies in Appendix A.0.2.

Human-Annotated Rationale Agreement. We calculate the alignment between generated rationales and human annotations collected in the ERASER benchmark [44]. As also shown in Table 4.1, we report the Token-level F1 (F1) metric along with corresponding Precision (P) and Recall (R) scores. The results show that LIMITEDINK can generate rationales that are consistent with human annotations and comparable to self-explaining baselines in all datasets.

4.4 Human Evaluation

Equipped with LIMITEDINK, we next carry out human studies to investigate the effect of rationale length on human understanding.

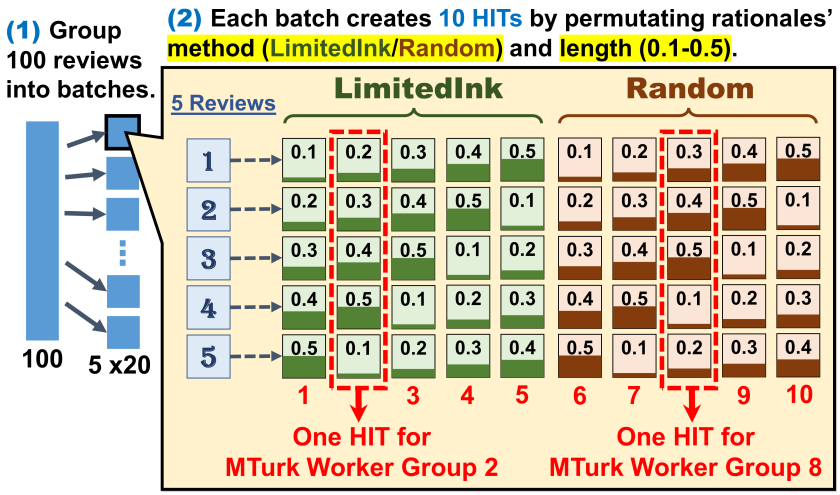


Figure 4.3: The human evaluation's workflow. We (1) divide 100 movie reviews into 20 batches and (2) produce 10 HITs from each batch for ten worker groups.

4.4.1 Study Design

Our goal is to quantify human performance on predicting the labels and confidence based solely on the rationales with different lengths. To do so, we control LIMITEDINK to extract rationales of different lengths, and recruit Mechanical Turk (MTurk) workers to provide predictions and confidence.

Dataset & rationale extraction. We focus on sentiment analysis in user study, and randomly sample 100 reviews from the Movie Reviews [269] test set that have correct model predictions. Then, we extract five rationales for each review using **LIMITEDINK**, with lengths from 10% to 50%, with an increment of 10%.

Since human accuracy likely increases when participants see more words (*i.e.,* when the lengths of rationales increase), we also create a **`Random`** rationale baseline, where we randomly select words of the same rationale length on the same documents (10% to 50%) while taking the continuity constraint into consideration. More details of `Random` baseline generation are in Appendix A.0.3.

Study Procedure. The study is completed in two steps. First, we posted a *qualification* Human Intelligence Tasks (HITs, $0.50 per assignment) on MTurk to recruit 200 qualified workers.[2] Next, the 200 recruited workers can participate the *task* HIT ($0.20 per assignment, 7 assignments posted) which contains five distinct movie reviews, with varying rationale lengths (10%-50%). In *task* HIT, as key components shown in Figure 4.2, we only display the rationales and mask all other words with ellipses of random length, such that participants can not infer the

[2]In addition to our custom qualification used for worker grouping, three built-in worker qualifications are used in all of our HITs: HIT Approval Rate (≥98%), Number of Approved HITs (≥ 3000), and Locale (US Only) Qualification.

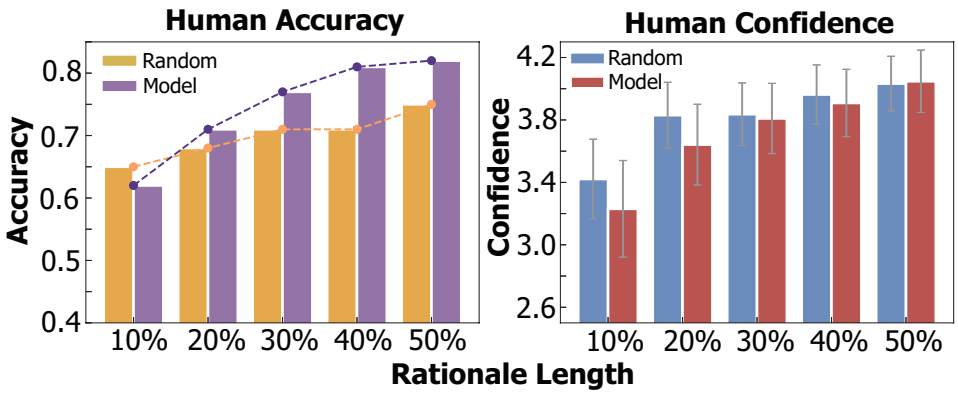


Figure 4.4: Human accuracy and confidence on predicting model labels given rationales with different lengths.

actual review length. Then participants are asked to guess the sentiment of the full review, and provide their confidence level based on a five-point Likert Scale [139]. The full user interface is in Appendix A.0.3.

Participants recruiting and grouping. With each review having ten distinct rationales (five from LIMITEDINK and five `Random`), if these rationale conditions were randomly assigned, participants are likely to see the same review repeatedly and gradually see all the words. We carefully design our study to eliminate such undesired learning effect. More specifically, we group our 100 reviews into 20 batches, with five reviews in each batch (Step 1 in Figure 4.3). For each batch, we create five HITs for LIMITEDINK and `Random`, respectively, such that all the rationale lengths of five reviews are covered by these 10 HITs (Step 2 in Figure 4.3). Further, we make sure each participant is only assigned to one unique HIT, so that each participant can only see a review once. To do so, we randomly divide the 200 qualified workers into 10 worker groups (20 workers per group), and pair one worker group with only one HIT in each batch. This way, each HIT can only be accomplished by one worker group. As our participant control is more strict than regular data labeling tasks on MTurk, we keep the HITs open for 6 days. 110 out of 200 distinct workers participated in the main study, and they completed 1,169 of 1,400 assignments.

4.4.2 Results

We show the human prediction accuracy and confidence results in Figure 4.4. We find that the best explanations for human understanding are largely not the shortest rationales (10% length level): here, the human accuracy in predicting model labels is lower than for the random baseline (0.61 vs. 0.63), indicating that the shortest rationales are not the best for human understanding. There is a significant difference in human predicted labels (*i.e.,* "positive"=1,"negative"=2) between LIMITEDINK (M=1.24,SD=0.71) and `Random` (M=1.32,SD=0.54); t(1169)=2.27, p=0.02. Table 4.2 shows human performance for each category.

Additionally, notice that the slope of our model's accuracy consistently flattens as the

length level (%) & Extract. method		**Negative** P / R / F1	**Positive** P / R / F1
10%	LIMITEDINK	0.66 / 0.56 / / 0.61	**0.70** / 0.58 / 0.64
	Random	**0.67 / 0.57 / 0.62**	0.66 / **0.70 / 0.68**
20%	LIMITEDINK	**0.75 / 0.61 / 0.67**	**0.71 / 0.77 / 0.74**
	Random	0.69 / 0.60 / 0.64	0.68 / 0.74 / 0.71
30%	LIMITEDINK	**0.74 / 0.76 / 0.75**	**0.81 / 0.78 / 0.79**
	Random	0.72 / 0.61 / 0.66	0.72 / 0.78 / 0.75
40%	LIMITEDINK	**0.84 / 0.76 / 0.80**	**0.78 / 0.85 / 0.81**
	Random	0.79 / 0.63 / 0.70	0.65 / 0.79 / 0.71
50%	LIMITEDINK	**0.78 / 0.78 / 0.78**	**0.85 / 0.84 / 0.85**
	Random	0.77 / 0.63 / 0.70	0.75 / 0.84 / 0.79

Table 4.2: Human performance (*i.e.,* Precision / Recall / F1 Score) on predicting model labels of each category in the Movie Reviews dataset.

rationale increases, whereas the random baseline does not display any apparent trend and is obviously lower than our model at higher length levels (*e.g.*, 40%). We hypothesize that this means our model is (1) indeed learning to reveal useful rationales (rather than just randomly displaying meaningless text), and (2) the amount of information necessary for human understanding only starts to saturate at around 40% of the full text. This creates a clear contrast with prior work, where most studies extract 10-30% of the text as the rationale on the same dataset [106, 176]. The eventually flattened slope potentially suggests a sweet spot to balance human understanding on rationales and sufficient model accuracy.

4.5 Discussion

By examining human prediction performance on five levels of rationale lengths, we demonstrate that the shortest rationales are largely not the best for human understanding. We are aware that this work has limitations. The findings are limited to Movie Reviews dataset, and we only evaluate human performance with rationales generated by the proposed LIMITEDINK. Still, our findings challenge the "shorter is better" assumption commonly adopted in existing self-explaining methods. As a result, we encourage future work to more cautiously define the best rationales for human understanding, and trade off between model accuracy and rationale length. More concretely, we consider that rationale models should find the right balance between brevity and sufficiency. One promising direction could be to clearly define the optimal human interpretability in a measurable way and then learn to adaptively select rationales with appropriate length.

To investigate if the shortest rationales are best understandable for humans, this work presents a self-explaining model, LIMITEDINK, that achieves comparable performance with current self-explaining baselines in terms of end-task performance and human annotation agreement. We further use LIMITEDINK to generate rationales for human studies to examine

how rationale length can affect human understanding. Our results show that the shortest rationales are largely not the best for human understanding. This would encourage a rethinking of rationale methods to find the right balance between brevity and sufficiency.

4.6 Related Work

4.6.1 Self-explaining models.

Self-explaining models, which condition predictions on their rationales, are considered more trustworthy than post-hoc explanation techniques [5, 146, 183, 190]. However, existing efforts often enforce minimal rationale length, which degrades the predictive performance [10, 106, 266]. [176] improves this by proposing an information bottleneck approach to enable rationale length control at the sentence level. In this paper, LIMITEDINK further enables length control at the token level to allow more flexibility needed for our human studies. However, it generates sentence-level rationales without considering contextual information, which cannot work well on short inputs (*e.g.*, one sentence input) or for extracting coherent token-level rationales. We present a method to incorporate contextual information into generating rationales, which can be applied to both sentence and token levels.

4.6.2 Human-grounded evaluation.

A line of studies evaluated model-generated rationales by comparing them against human-annotated explanations [22, 176]. Some other studies collect feedback from users to evaluate the explanations, such as asking people to choose a preferred model [190] or to guess model predictions only based on rationales [130, 211]. In this work, we evaluate LIMITEDINK using both human rationale alignment and empirical human evaluation. To our best knowledge, this is the first study that limitedinkatically varies rationale length and measures differences in human understanding.

4.7 Conclusion

To investigate if the shortest rationales are best understandable for humans, this work presents a self-explaining model, LIMITEDINK, that achieves comparable performance with current self-explaining baselines in terms of end-task performance and human annotation agreement. We further use LIMITEDINK to generate rationales for human studies to examine how rationale length can affect human understanding. Our results show that the shortest rationales are largely not the best for human understanding. This would encourage a rethinking of rationale methods to find the right balance between brevity and sufficiency.

Part II

Disparity Between AI Interpretability and Practical Human Demands

Chapter 5

Gauging Explainable AI Gaps with User Demands Using XAI Forms

5.1 Introduction

Researchers have attempted to produce model interpretations for deep neural networks [158] under the broader umbrella of Explainable Artificial Intelligence (XAI). The primary objective of this line of research is two-fold [102]: to create interpretations that faithfully characterize the models' behavior (*i.e.*, are *faithful*), and to improve user trust or understanding of black-box algorithms (*i.e.*, appear *plausible*). However, this objective does not always align with the practical needs of users. Recent studies reveal that a faithful or plausible model interpretation can still be useless, or even harmful, to its users. For example, our previous work found that showing users visual explanations (saliency maps) decreased — *not* increased — users' ability to make sense of the mistakes made by neural image classifiers [211]. Another study showed that visual explanations may not alter human accuracy or trust in the model [31]. Recent work in XAI has begun to mitigate this misalignment [54]; one example is collecting algorithm-informed user demands from real-world practices [138].

This paper takes a closer look into the gap between user need and current XAI. Specifically, we survey the common *forms* of explanations, such as feature attribution [27, 128], decision rule [112, 198], or probe [56, 141], used in 218 recent NLP papers, and compare them to the 43 questions collected in the XAI Question Bank [138]. We use the forms of the explanations to gauge the misalignment between user questions and current NLP explanations.

5.2 Gauging Explainable AI Gaps Using Forms

Liao *et. al* [138] developed the XAI Question Bank, a set of prototypical questions users might ask about AI systems. This paper investigates how well these questions are answered by current XAI work in NLP. We collected 218 recent NLP papers about interpretability, analyzed the *forms* of interpretations these papers researched (*e.g.*, feature attribution, decision rules, etc.),

and used these forms to associate each paper to the questions it tried to answer. This section overviews our two-step procedure.

5.2.1 Step 1: Survey the Forms of Interpretations in NLP Papers

We first reviewed 218 explanation studies published in the NLP field between 2015 and 2020, and came up with 12 common XAI forms. We defined a paper as an NLP explanation study if: *(i)* its motivation was to explain or analyze NLP models, tasks, or datasets; or *(ii)* it aimed to develop more explainable NLP models, tasks, or datasets; or *(iii)* the explanation format is natural language. Given those definitions, we decided on a set of search keywords (*e.g.*, "explain", "interpretation"), a list of top-tier publications and conference proceedings (*e.g.*, ACL, EMNLP), and a range of publication years. Within the venues and years, we collected all papers whose titles or abstracts contained those keywords. Then we read each paper and added the related papers that it cited about "NLP explanation" into our collections. Our ultimate list of papers covered various conferences, workshops, and other research fields (*e.g.*, human-computer interaction).

Our definition of "interpretation form" is *how the study represents its explanation results*. In this paper, we present 12 different interpretation forms. We started with four commonly used forms, including "feature attribution [27, 128]," "tuple/graph [159, 226]," "free text [126**?**]," and "example [83, 263]." Then we read each paper, assigned a form to it, and added new forms into our scheme as needed.

We present the 12 forms with their abbreviations, format weight, brief definitions, representative work, and one typical question in sidebars on page 2 to 4. We released our data[1], which contains the list of the 218 NLP explanation papers with each paper's title, year, venue, and form annotations.

We then computed what percent of the 218 XAI papers used each type of interpretation form (*i.e.,* format weight). We gave each paper a weight of 1. If the paper used only one form type, we assigned 1 to the form. If the paper used multiple interpretation forms, we assigned all its applicable forms an equal weight totaling 1. To obtain the final percentage of each form type, we added up its scores among all papers and divided by the count of papers.

As shown in the sidebars, the most common form of current NLP explanations — around 44% of related studies — is to highlight features (*e.g.*, tokens or sentences) within input text. Approximately 10% of NLP explanation work leverages a tuple, rule, or concept format to demonstrate the model's reasoning process. Other studies use a probe to diagnose what information the model representation can embed, or directly explains model behavior with free text. Less than 5% of algorithms use training data examples, projection space, or output confidence scores to visualize NLP explanations. Fewer than 5 papers explain NLP models with word cloud, trigger, or image formats.

[1]Please see details at `https://human-centered-exnlp.github.io`.

5.2.2 Step 2: Compare Against User Questions in the XAI Question Bank

The XAI Question Bank collected user questions for AI explanations from real-world user needs [138]. It consists of 43 questions within 7 categories about AI systems, as detailed in Figure 5.1 and Table 5.1. The prototypical questions are identified by analyzing current XAI algorithms and interviewing UX and design practitioners in IBM product lines. We annotated each user question in the XAI Question Bank with all applicable forms identified in Step 1. The principle we used to annotate a user question with forms was *if the format had ever answered similar questions among our collected studies*. Specifically, in the first step, we noted typical questions the form answered in the literature exemplified in the sidebars. For instance, the "example" form primarily answers "What are the training instances most responsible to support this prediction?" [83, 121] Then we inspected each question in the XAI Question Bank and looked for similar questions we collected for the 12 forms. We labeled the user question with its corresponding form when the form was used to answer similar questions in the literature. For instance, to answer the user question "How does the system make a prediction?" we can explain AI systems to users using executable logic rules (*i.e.*, RUL), decision-reasoning graphs (*i.e.*, TUP), or by showing each class's representative examples (*i.e.*, EXP). Afterwards, we calculated each user question's weight by adding all its labeled formats' weights. The user question weight roughly approximates the proportion of published NLP papers that can answer this question. This resulted in the weighted XAI Question Bank as shown in Figure 5.1, which provides intuitive visualization of the NLP research's attention to answering user questions. Note that XAI forms may evolve rapidly due to proliferation of XAI studies, but we can extend the collected XAI forms and repeat the gap-gauging process easily.

5.3 The Need to Explain the Road Not Taken

5.3.1 Users need AI explanations of a global view of AI systems

While 9 out of 43 questions in the XAI Question Bank are about how AI systems **can** provide specific predictions (*i.e.*, Q19-21,23-25,36-38), 16 questions are about what AI systems **cannot** achieve and why (*e.g.*, Q5-6,11,15-16,26- 35,41). Many of these under-answered or unanswered questions are *counterfactual* questions, such as "Why did the model predict P instead of Q for this instance?" These questions can probably be answered by a trigger (TRG), but only three papers out of the surveyed 218 focused on counterfactual explanations [62, 201, 245]. Furthermore, we speculate that many of these questions assume one or more well-defined, seemly similar legitimate counterpart labels (*e.g.*, *positive* versus *negative*, *dog* versus *cat*), in which the user wonders why the system choose one over the other. More fundamentally, the fact that users want to know both *why* and *why not* the AI system made certain predictions may suggest that users' goals are often to **gain a *global view* of how the AI system works**.

It is worth noting that more NLP work has begun to generate counterfactual examples

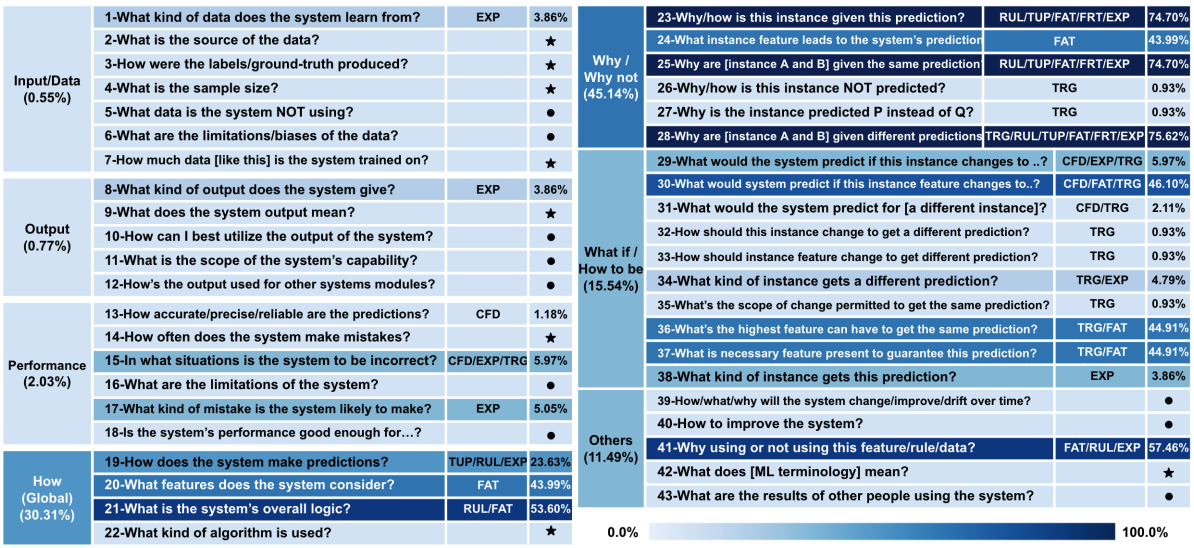


Figure 5.1: The questions in XAI Question Bank, heat-mapped by the estimated percentage (%) of NLP XAI studies attempting to answer them. (●: questions that can *not* be answered by most NLP XAI studies; ⋆: questions that can likely be answered by the AI system's meta information.)

(*i.e.*, "contrastive sets"), often with the purpose of learning robust NLP models [68, 117, 255]. These methods could be extended to generating counterfactual explanations. As counterfactual explanations have been explored in other domains, such as computer vision [23], tabular data classification [161], and interactive tools [77], recent NLP work has begun to focus more attention on developing counterfactual explanations [104, 194][2].

5.3.2 Users need AI Explanations for what AI can not do

Developing counterfactual explanations in NLP can be challenging. It is not always easy to tell **which counterfactual predictions** should be explained. Jacovi *et al.* submitted a good example [104]: When people ask "Why did the AI system choose to hire Person X?" they could mean either "Why did the AI system choose to hire Person X rather than not hire Person X?" or "Why did the AI system choose to hire Person X rather than hire Person Y?" Liao *et al.* suggested that AI explanations can be provided in an *interactive* manner, allowing people to "explicitly reference the contrastive outcome and ask follow-up *what if* questions" [138]. As ambiguous and underspecified language can be common, more research is required to help users spot the meaningful counterfactual predictions they actually care about.

[2]We did not include these recent studies in our paper collection because they were published after our paper-collecting and analysis process.

5.4 Discussion and Limitation

5.4.1 User Questions Beyond the Scope of the Current XAI.

In another finding included in Figure 5.1, we observed 8 questions (*i.e.*, labeled $\star$) that can be addressed by the *meta information* in AI algorithms (such as "What is the source of the data?") but that XAI forms do not answer. However, we find 10 questions (*i.e.*, labeled $\bullet$) that the XAI forms cannot address well. These questions mainly inquire about the **limitation, potential utility, or capability scope** of AI systems (*e.g.*, "What are the limitation/biases of the data?"), which are seldom introduced in XAI studies. We posit XAI algorithm developers should use these questions to develop corresponding XAI methods or to clarify capability scope, system utility, and limitation in the methods.

5.4.2 Limitations

We are aware of several limitations of our work. First, this paper focuses on NLP applications, but the XAI Question Bank captures user questions for a broader spectrum of AI systems. Second, the XAI Question Bank provides an in-depth analysis of lay users' needs, while the user population for the NLP papers included in our study are broader, such as domain experts [59, 275] and AI practitioners [11, 190]. Finally, using forms of interpretation to associate papers with user questions inevitably overlooks some information. For instance, the "probing" form does not appear in the XAI Question Bank. This could be caused by the fact that some particular forms of interpretations, such as probing methods, are primarily developed for AI practitioners rather than lay people.

5.5 Conclusion

Our analysis explicates the gaps between what users want and the current focus of XAI research in NLP. Questions like "Why is this instance given this prediction?" were studied extensively, and can be answered by five different interpretation formats (*i.e.*, "rule/grammar," "tuple/graph," "feature importance," "free text," and "example"). Meanwhile, 16 out of 43 user questions in the XAI Question Bank are relevant to counterfactual inquiries, such as "Why did the model predict P instead of Q for this instance?", but only a handful of papers have tried to produce counterfactual explanations. We learned that users want to know the decision scope of AI systems, including what the AI system can and cannot achieve.

XAI researchers can collaborate with user-experience (UX) designers to mitigate this misalignment. In particular, XAI algorithm developers can produce more counterfactual explanations for answering global and local counterfactual questions, or directly generate AI explanations that can explain both *can* and *cannot* questions (*e.g.*, tree-based rules). On the other hand, one XAI form may not be enough to satisfy practical user demands for understanding *can* and *cannot* questions simultaneously. Therefore, XAI UX designers can combine multiple

forms and algorithms to meet real-world user requirements. Since awareness of new explainable AI forms can change user demand [138, 140], perhaps XAI researchers can leverage the variety of forms to to respond more effectively to real-world user needs.

Format Name	Abbr.	Percent	Definition	Question Example
Feature Attribution	FAT	43.99%	highlight the sub-sequences in input texts [27, 128]	*How can we attribute the systems' predictions to input features?* [162]
Tuple / Graph	TUP	10.15%	explain model reasoning process with tuples/ trees/ graphs [159, 226]	*How does the system use reasoning graphs to arrive at the answer?* [28]
Concept / Sense	CPT	9.72%	convert to human interpretable concepts or terminologies [25, 202]	*What sense does the system's intermediate representation make?* [173]
Rule / Grammar	RUL	9.61%	extract executable rules or logic for model decisions. [112, 180]	*How can we explain the system's behavior with executable rules?* [191]
Probing	PRB	7.79%	classify representation with specific diagnostic dataset [56, 141]	*What linguistic properties does the system's representation have?* [88]
Free Text	FRT	7.09%	use natural language to explain model behavior [126]	*How can we explain a system's decision using natural language justification?* [21]
Example	EXP	3.86%	find most responsible training samples as explanations [83, 263]	*How can we trace the system's prediction back to the training sample(s) most responsible for it?* [121]
Projection Space	PSP	3.82%	project dense vectors into low-dimensional space [228, 254]	*How can we project the system's high-dimensional representation to a human-understandable space?* [7]
Confidence Score	CFD	1.18%	leverage model prediction probability to show confidence [86, 92]	*How much uncertainty does the system have on its prediction?* [60]
Word Cloud	WCL	1.16%	generate word cloud using model representations [27, 175]	*What are the input patterns that activate the system prediction?* [129]
Trigger	TRG	0.93%	make change to trigger models to produce counterfactual predictions [62, 201]	*What are the token sequences that trigger a model to produce a different prediction?* [245]
Images	IMG	0.70%	visualize model representations by token-related images [172]	*How to map the system's language tokens to their related images?* [234]

Table 5.1: The AI Explanation Formats collected into 12 XAI formats and the corresponding definitions and question of examoels.

Part III

Empower Useful AI Explanations with Humans-AI Interactions

Chapter 6

ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing

6.1 Introduction

The advancement of deep learning has led to breakthroughs in a number of artificial intelligence systems (AI). Yet, the superior performance of AI systems is often achieved at the expense

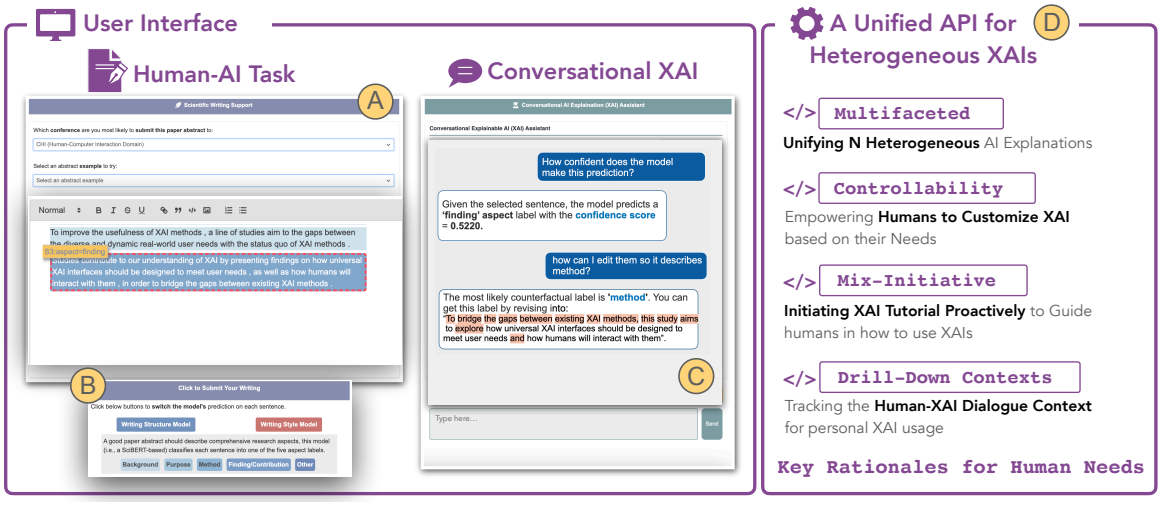


Figure 6.1: An overview of ConvXAI to support human-AI scientific writing with heterogeneous AI explanations via dialog. ConvXAI includes a front-end User Interface to (A) support human-AI collaborative task interaction, (B) check AI models and predictions, and (C) inquire about heterogeneous AI explanations via dialogue. Also, ConvXAI involves a back-end deep learning server to generate AI predictions and explanations, which is embedded with (D) a unified API for generating heterogeneous AI explanations that are designed to cater to practical human use needs.

of the interpretability of deep learning models [154]. To address this challenge, researchers have developed a collection of eXplainable AI (XAI) methods that aim to enhance human understanding of AI from various perspectives [211, 212]. These methods typically focus on answering specific XAI questions of interest to users. For example, saliency maps and feature attributions [146, 190] highlight key rationales behind AI predictions to address "why" questions, while counterfactual explanations perturb input to explore "why X not Y" scenarios that affect model behavior [154, 255].

Despite their potential, the usefulness of XAI methods in real-world applications has yielded inconsistent findings [8, 181]. While some studies demonstrate that different explanations can support specific use cases, such as model debugging [131] and human-AI collaboration [78], others reveal limitations in enhancing users' ability to simulate model predictions [214] or understand AI errors [211]. To bridge this gap, researchers have explored the mismatch between real-world user demands and existing XAI methods. [212], for instance, compare practical user questions [138] with over 200 XAI studies and identify a bias in current methods towards certain types of XAI questions, neglecting others. Additionally, users also tend to have *multiple*, *dynamic* and sometimes *interdependent* questions on AI explanations [124, 239].

Addressing this array of questions necessitates an integration of heterogeneous AI explanations. Taking inspiration from the flexibility of dialog systems [58, 116], prior work has envisioned the concept of "explainability as a dialogue" to accommodate diverse user needs and mitigate cognitive load [124, 220]. For instance, [124] discovered that decision-makers strongly prefer interactive explanations in the form of natural language dialogue. However, there is a dearth of exploration regarding the design of conversational XAI systems to meet practical user needs and understand user reactions.

In this paper, we investigate the potential of conversational XAI in the context of practical human-AI collaborative writing. Through formative user studies on a preliminary system and a review of human conversation characteristics, we identify four design rationales for conversational XAI: addressing various user questions ("multi-faceted"), actively suggesting and accepting follow-up questions ("mix-initiative" and "context-aware drill-down"), and providing on-demand details ("controllability"). Guided by these rationales, we develop a conversational XAI prototype system called ConvXAI, which incorporates the four user-oriented XAI principles. Moreover, we evaluate the potential of ConvXAI in the realm of human-AI scientific writing, where writers leverage ConvXAI to improve their paper abstracts for submission to top-tier research conferences. In this use case, ConvXAI assists users in interacting with two AI writing models that assess the structure and quality of abstracts at the sentence level. Users can engage in dialogue with ConvXAI to comprehend the writing feedback and enhance their papers with the aid of heterogeneous AI explanations.

We conducted two within-subject user studies to evaluate the ConvXAI system. We compared ConvXAI with SelectXAI, a traditional GUI-based universal XAI system that displays all XAIs on the interface in a collapsible manner (Figure 6.4). In the first user study, involving an open-ended writing task with 13 participants, we found that the majority of users perceived ConvXAI to be more useful in understanding AI writing feedback and improving

their own writing. These results further confirmed the reduced cognitive load and effectiveness of the four user-oriented design principles. Additionally, in the second user study, which focused on a well-defined writing task with 8 rejoining participants, we collected the users' writing artifacts generated using both CONVXAI and SELECTXAI systems. We evaluated these artifacts using both human evaluators and auto-metrics. The analysis revealed that both CONVXAI and SELECTXAI assisted users in producing better writing based on the built-in auto-metrics, with CONVXAI proving particularly valuable for improving writing quality. However, we observed a misalignment between the measurements of the human evaluator and the auto-metrics, indicating the importance of designing AI model predictions to align with human expectations. Building upon these studies and findings, we further contribute insights into the practical human usage patterns of XAI in CONVXAI and core ingredients of useful XAI systems for future XAI work. We conclude this work by discussing its limitations and outlining future research directions.

6.2 Related Work

6.2.1 Human-Centered AI Explanations

Earlier studies in the fields of Explainable Artificial Intelligence (XAI) primarily focus on developing different XAI techniques, which aims to explain *why the model arrives at the predictions.* This line of studies can be broadly categorized into generating post-hoc interpretations for well-trained deep learning models [79] and designing self-explaining models [128, 214, 218]. In specific, the majority of XAI methods aim to provide post-hoc interpretations either for each input instance (*i.e.,* named "local explanations") [42, 120, 201] or for providing a global view of how the AI model works (*i.e.,* named "global explanations") [191], where our study covers both of them. Additionally, XAI approaches are also divided into different formats [212], including example-based [60], feature-based [190], free text-based [21, 184], rule-based explanations [191], etc, where our study covers a range of XAI formats.

Despite the increasing number of XAI approaches have been proposed, evaluating AI with humans is still a challenging problem. Doshi-Velez and Kim [48] propose a taxonomy of interpretability evaluation including "application-grounded", "human-grounded" and "functionally-grounded" evaluation metrics based on different levels of human involvement and application tasks. The majority of the proposed XAI approaches are commonly validated effectively using the "functionally-grounded" evaluation methods [87, 101, 251], which seek for automatic metrics (*e.g.,* "plausibility") on proxy tasks without real human participations [11, 157, 273].

Furthermore, we can see burgeoning efforts being put around involving real humans in evaluating AI explanations under the theme of "human-centered explainable AI". The state-of-the-art XAI methods are applied to real human tasks, such as assessing human understanding [211], human simulatability [191, 214], human trust and satisfaction on AI predictions [41, 223], and human-AI teamwork performance [31], etc [60, 73, 86]. However, many human studies show that AI explanations are not always helpful for human understanding in tasks such as simulating

model prediction [214], analyzing model failures [211], human-AI team collaboration [8]. For instance, [8] conducted human studies to investigate if XAI helps achieve complementary team performance and showed that none of the explanation conditions produced an accuracy significantly higher than the simple baseline of showing confidence.

In response, a line of work dives deep into the gaps between real-world user demands and the status quo XAI methods. Their findings reveal that users tend to ask *multiple*, *dynamic*, and sometimes *interdependent* questions on AI explanations, whereas state-of-the-art XAI methods are mostly unable to satisfy. Although GUI-based XAI systems, which integrate multiple XAI into one interface, can potentially mitigate this issue, they inevitably suffer from the drawbacks, such as cognitive overload, frequent UI updates, etc.

Therefore, prior studies envision the potential of "Explainability as a Dialogue" to balance the cognitive load with the diverse user needs [124, 149, 220, 232, 239]. For example, through interviews with healthcare professionals and policymakers, [124] found that decision-makers strongly prefer interactive explanations with natural language dialogue forms and thereby advocated for interactive explanations. Nevertheless, there has been little exploration of how a conversational XAI system should be designed in practice and how users might react to it. Our studies aim to resolve this problem by incorporating practical user needs into the conversational XAI design, propose a user-oriented conversational universal XAI interface and investigate human behaviors during using these systems.

6.2.2 Conversational AI Systems

Our work is situated within the rich body of conversational AI or chatbots studies, which entails a long research history in the NLP [135, 187] and HCI fields [58, 208]. Jurafsky [116] proposes that conversation between humans is an intricate and complex join activity, which entails a set of imperative properties: *multiple turns*, *common grounding*; *dialogue structure*, *mixed-initiative*. By incorporating these properties, conversational interactions are also shown to significantly contribute to establishing long-term rapport and trust between humans and systems [14]. User interaction experience can be improved by a set of factors from the conversational AI systems [208]. For example, Chaves and Gerosa [30] describe how human-like social characteristics, such as conversational intelligence and manners, may benefit the user experience.

These principles and theories inform us to design a conversational AI explanation system that fulfills the diverse user needs in practice. Our study is deeply rooted in the conversational explanations in XAI – the users request their demanded explanations through the chatbot-based AI assistants [225, 239]. Previous studies have explored the effectiveness of interactive dialogues in explaining online symptom checkers (OSCs) [232, 239]. For example, Tsai et al. [239] intervened in the diagnostic and triage recommendations of the OSCs with three types of explanations (*i.e.,* rationale-based, feature-based and example-based explanations) during the conversational flows. The findings yield four implications for future OSC designs, which include empowering users with more control, generating multifaceted and context-aware explanations, and being cautious of the potential downsides.

Stage	XAI Goal	User Question Samples	XAI Formats	Algorithm
1	Understand Data	1.What data did the system learn from?	Data Statistics	Data Sheets
		2.What's the range of the style quality scores?		
		3.How are the structure labels distributed?		
	Understand Model	4.What kind of models are used?	Model Description	Model Card
	Understand Instance	5.How confident is the model for this prediction?	Prediction Confidence	Model probability score
		6.What are some published sentences similar to mine semantically?	Similar Examples	NN-DOT
		7.Which words in this sentence are most important for prediction?	Feature Attribution	Integrated Gradient
	Improve Instance	8.How can I revise the input to get a different prediction label?	Counterfactual	GPT3 In-context Learning
2	Understand Data	9.What's the statistics of the sentence lengths?	Data Statistics	Data Sheets
	Understand Suggestion	10.Can you explain this sentence review?	XAI Tutorial	Template

Table 6.1: CONVXAI covers ten types of user questions (*i.e.,* Data Statistic, Model Description, Feature Attribution, etc.) serving to five different XAI goals (*e.g.,* Understand Model, Understand Data, Improve Instance, etc.). Stage (1) shows eight XAIs included in the formative study, and Stage (2) indicates two added XAIs in CONVXAI.

.

However, these existing conversational AI explanation systems are still in the preliminary stage, which only provides one type of explanation and disables users from selecting different explanation types. Also, these are far from being able to incorporate user feedback into producing AI explanations (*e.g.,* enable users to choose counterfactual prediction foil) and produce personalized explanations for users' individual needs. In addition, these conversational AI explanation systems are primarily applied to improve system transparency and comprehensibility, thus helping users understand and build trust in the systems. Little attention has been paid to examining *if* and *how* conversational AI explanations can be indeed useful for users to improve their performance in human-AI collaborative tasks.

Our work improves the conversational AI explanation systems from two perspectives: i) we focus on AI tasks where the human's goal is to improve their task performance (*i.e.,* scientific writing) rather than merely gain an understanding of the AI predictions; ii) we identify four design principles and incorporate them into the empirical system design for further evaluation with human tasks. Our work aims to further unleash the capability of conversational AI explanations and make them more useful for human tasks.

6.2.3 AI Writing Support Tools

The improvements in large language models (LMs) like GPT3 [18] and Meena [4] have provided unprecedented language generation power. This leads to an increasing interest in how these new technologies may support writers with AI-assisted writing support tools [127]. In these human-AI collaborative writing tasks, the writers interact with AI writing support tools not only for understanding its assessment but also aim to leverage its feedback to improve the human writing output [95]. A few technologies are developed to support human writing. Many of them focused on *lower-level linguistic improvement*, such as proofreading, text generation, grammar correction, auto-completion, etc. For instance, Roemmele and Gordon [192] proposed a Creative Help system that uses a recurrent neural network model to generate suggestions for the next sentence. Furthermore, a few studies propose AI assistants that leverage the

generation capability of the language models to *generate inspirations* to assist the writers' ideation process [34, 71, 248]. For instance, Wordcraft [34] is an AI-assisted editor proposed for story writing, in which a writer and a dialogue system collaborate to write a story. The system further supports natural language generation to users including planning, writing and editing the story creation.

In addition, there are a number of studies that design AI assistants to *provide assessment and feedback* to help improve human writings iteratively [52, 213]. For example, Huang et al. [97] argue that writing, as a complex creative task, demands rich feedback in the writing revision process. They present Feedback Orchestration to guide writers to integrate feedback into revisions by a rhetorical structure. More studies are proposed for AI-assisted peer review [26]. For example, Yuan et al. [267] automate the scientific review process that uses LLMs to generate reviews for scientific papers.

In this work, we apply conversational AI explanations to human-AI scientific writing tasks, in which humans submit their writings to the system and iteratively make a sequence of small decision-making processes based on AI feedback and explanations. As *writing is a goal-directed thinking process* [71]. The goal of the CONVXAI system is to support writers to *understand the feedback* and further *improve their writing* outputs. Therefore, we aim to evaluate the effects of conversational AI explanations in terms of not only helping users understand the AI prediction but also improving writing performance.

6.3 Understanding Practical User Demands in Conversational XAI

Due to the unique characteristics of AI-assisted human scientific writing tasks and the early status of conversational XAI systems, we see a lack of established designs and techniques of conversational AI explanations that can cater to user needs in scientific writing support tasks. Therefore, we first analyze the practical user demands of conversational XAI by conjecturing a system walkthrough in a usage scenario with a student submitting her CHI paper (Section 6.3.1), and then conducting a formative study with seven users of diverse backgrounds (Section 6.3.2). We summarize the resulting four design rationales in Section 6.3.3.

6.3.1 Example User Scenario

Gloria is a Ph.D. student in the CHI research field. While she has already finished a paper draft, she wants to use the system to receive more paper review feedback on her paper abstract writing, so that the paper would get a higher chance of being accepted by the CHI conference. She is especially curious about *What would be the review feedback of my paper abstract*? *Why would the system give me this feedback*? *How should I improve my writing to get a better paper abstract*? To answer these questions, Gloria starts to interact with the system with these questions in mind. First, she is asked to choose the target conference she wants to submit

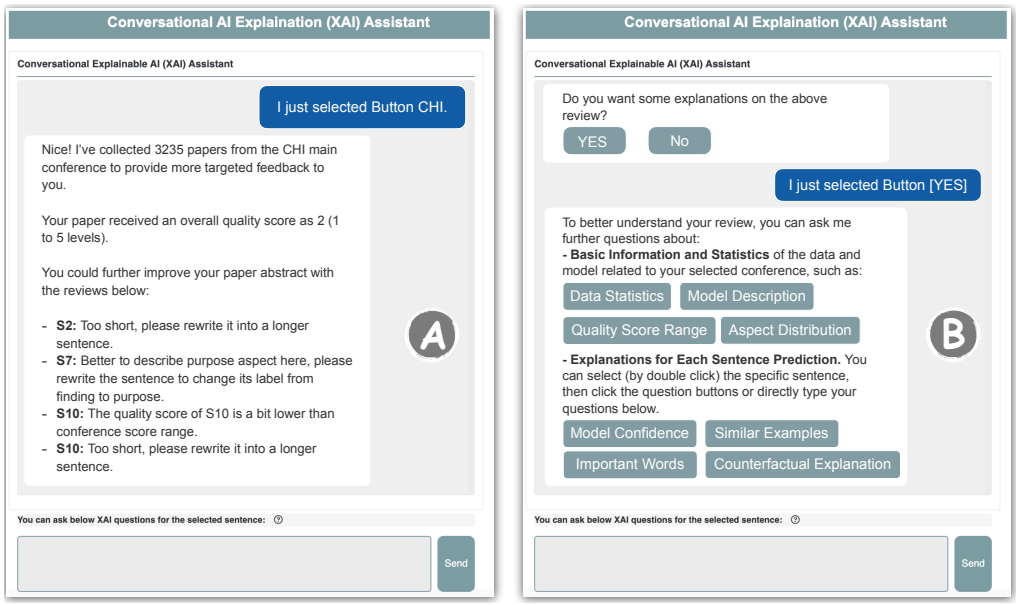


Figure 6.2: An overview of User Interface (UI) for the pilot study. (A) shows the recommended edits from the writing models, and (B) displays a range of XAI buttons for users to choose from for viewing AI explanations.

her paper abstract. After choosing CHI as the target conference, Gloria can see the abstract example options and writing editor panel, so that Gloria can *edit her abstract content* and then submit her abstract to get AI assistant assessments on each sentence.

For example, one piece of writing review that Gloria received is "Sentence 3: Based on the sentence labels' percentage and order in your abstract, it is suggested to write your *background* at this sentence, rather than describing *purpose* here.". Before diving deeper into understanding the predictions, Gloria first wants to assess if she should trust the models by understanding how the model and data work in this ConvXAI system. So she asks *"What data did the system use?"* and *"What kind of models is the* ConvXAI *using?"*. After learning that the ConvXAI is using the state-of-the-art language models and the data is the collection of the latest five years from CHI, Gloria decides to trust the system and proceed with the AI explanations. At the next stage, Gloria is wondering *why the system suggests she describe the background instead of purpose in the sentence 3*. By asking *"What words make the assistant think it is describing the purpose?"*, she learns that the "purpose" aspect prediction is attributed to the top 6 important words, including "examine", "paper", "conversational", "xai", "scientific", "writing" (feature attribution explanation). Furthermore, Gloria wants to know *"how can I edit them so it describes background"* and is suggested to remove the "in this paper, we examine" words at the beginning and add "is yet to be explored" in the end. (counterfactual explanation) After interacting with the XAI agent with multi-turn dialogues, she *understands the system, predictions and reviews* better. Finally, Gloria *revises the sentence* based on her understanding with the help of XAI agent and *re-submit the abstract*. The structure review is successfully resolved. Gloria can then move on to the next sentence.

6.3.2 Formative Study

In the early phase of the project, we conducted a formative study to inform ourselves about how humans leverage AI explanations to achieve their AI-assisted scientific writing tasks, and the common limitations and needs necessary for enhancing human performance. This is primarily to help us develop a set of design rationales listed in Section 6.3.3 to motivate system designs.

6.3.2.1 AI Tasks and AI Explanation Design.

To form the human-AI interactive writing scenario, we develop two AI writing models to generate writing structure and style predictions, respectively. The writing structure model gives each sentence a research aspect label, indicating which aspect the sentence is describing among the five categories (*i.e.,* background, purpose, method, contribution/finding, and others). On the other hand, the writing style model provides each sentence a style quality score assessing "how well the writing style of this sentence can match well with the published sentences of the target conference". Based on the predictions of all sentences, we further use algorithms to integrate all sentences' predictions into the writing reviews.

Given this AI task, we deem that conversational XAI system should be prepared to **answer a wide range of knowledge gaps between the users and the AI models** [155]. That says – the conversational XAI system is able to answer a variety of XAI questions that cover different perspectives of the system, including AI models, datasets, training and inference stages and even system limitations, etc [212]. Therefore, we design the XAI questions around four explanation goals, as illustrated in Table 6.1 (1), (a) *understanding data*, which uses data to help contextualize users' understanding of where they abstract sit in the larger distribution; (b) *understanding model*, which provides information on the underlying model structure so users can assess the model reliability; (c) *understand instance*, which allows users to ask questions that dive into, each individual prediction unit (*i.e.,* sentence). (d) *improve instance*, which goes one step further than understanding, and targets the goal of helping people to *improve* their writing by suggesting potential changes.

Embodied with the aforementioned two AI writing models and 8 types of AI explanations, we build up a preliminary system of conversational AI explanations for scientific writing support. The front-end user interface looks similar to Figure 6.1, which includes a *human-AI task* panel on the left where users can inspect and edit their abstracts, and a *conversational XAI* panel on the right where users interact with the XAI agent. In the **human-AI writing task** panel, users can iteratively edit their abstracts, and submit them to receive AI assessments on their writing structure and style.[1] As for the **conversational XAI** panel, at the initial entry, the panel provides a summary of the recommended edits (Figure 6.2A). Then, as participants dive into each individual sentence, we allow them to select XAI methods they might find suitable by clicking on the corresponding buttons (Figure 6.2B). The button-based design is inspired by the standard interface for service chatbots [264], while participants were still allowed to

[1]As the writing models of the preliminary and formal conversational XAI systems are identical, we encourage readers to refer to Section 6.4.2 for more details of all the writing models and reviews.

just type their own questions. This setting is also similar to the existing XAI interactive dialogue systems [232, 239], where they provide different formats of AI explanation for the same prediction and evaluate human assessment on different explanations.

6.3.2.2 Participants and Study Procedure.

We recruited seven participants with diverse research backgrounds and experiences in the formative study: 1 assistant professor, 2 Ph.D. students, 3 industry scientists or engineers, and 1 master's student working on HCI, NLP, and AI research (refer to Table B.1 for detailed demographic statistics). The formative studies are conducted virtually via virtual conference calls on Zoom. During this study, participants were asked to either bring one of their abstract drafts or use one example provided by us. We conducted a semi-Wizard-of-Oz (WoZ) process where we encouraged users to think aloud during asking AI explanations to the XAI agent, with keeping in mind the goal of improving their abstract writing. One researcher, who had several years of HCI and algorithmic AI explanation experience, acted as the XAI agent in this WoZ setting. We collected users' reflections on the system and summarized them into design rationales below.

6.3.3 Design Rationales

While formative study participants all appreciated the access to multiple XAI methods, merely listing all XAI options for human use is not enough. Instead, they were frequently overwhelmed by the large number of options available. We combine their feedback with theoretic linguistic properties of human conversation [100, 116], and propose the following for design requirements for ConvXAI systems:

R.1 **Multifaceted**: ConvXAI system should provide diverse types and formats of AI explanations for users to choose from, and use multi-modal visualization techniques to display the explanations efficiently. As we have argued in Section 6.3.2), to satisfy diver users needs [138, 212], it is imperative to **provide multiple XAI types and formats**. Nevertheless, some formative study participants noticed that having all the explanations displayed at once is overwhelming, and preferred to have a "overview first, details on demand" structure [219]. I-6 discussed that *"I can tell the system knows a variety of AI explanations. However, it can be too much for me to understand all these explanations at once. I would prefer to know the 'big picture' first, and then drill down with 'some options' as I need to dive deeper."*

R.2 **Mixed-initiative**: ConvXAI system should enable both user and XAI agent to initiate the conversation. Especially, it should proactively speculate the XAI user needs and prompt with next-step suggestions. One unique characteristic of conversations is mixed-initiative, *i.e.,* who drives the conversation [116]. Just as many existing conversational systems, we aim to mimic human-human conversations where initiative shifts back and forth between the human and the ConvXAI. This way, not only can the system answer users' questions, but it can also occasionally steer the conversation in different directions. In our study, we

also found this to be quite essential, especially when users do not have a clear goal in mind (*e.g., "Which sentence in the abstract should I look into first?"*).

R.3 **Context-aware drill-down**: CONVXAI system should allow users to drill down AI explanations with multi-turn conversations with awareness of the context. Linguistic theories model human conversation as a sequence of turns, and conversational analysis theory [100] describes the complex dialogues as joining the basic units, named adjacency pairs. This was also empirically validated in our pilot study. For instance, I-2 discussed potentially switching between explanations based on current observations: *"I might directly ask the system how to rewrite the sentence to change this sentence into the background aspect (*i.e., *"counterfactual explanation"). But if its rewritten sentences are not good enough, I would check the most similar examples of background aspects to learn their style and write on my own then (*i.e., *"similar examples")"*. Carrying over context throughout the conversation without users repeating themselves too much is useful for making the conversation natural and continuous.

R.4 **Controllability**: CONVXAI system should be able to generate customized AI explanations that can satisfy the user needs and context. This includes both only displaying explanations that are relevant to their questions (*e.g.,* answer "why this prediction" with feature attribution), and adjusting the explanation settings (*e.g.,* number of important words to highlight). As I-7 said – *"I spent too much time on figuring out what each XAI means, then I forget what I want to write in the abstract. It would be great if to give me the AI explanations targeting my question and enable me to input some variables to generate the XAIs I want.* At the same time, users still preferred to have a default explanation first and then provide options to control the variables or diver deeper into details, so they only need to pay attention to parts that are worthy of personalization.

6.4 CONVXAI

Based on the use scenario and design principles, we present CONVXAI, a system that applies conversational AI explanations on scientific writing support tasks, which incorporates the four rationales into the system design. The system aims to leverage conversational AI explanations on the AI writing models to improve human scientific writing. We extend the system developed in the formative study, which consists of a writing panel and an explanation panel. The writing panel is similar to the formative study, which can enable users to iteratively submit their paper abstract and check the writing model predictions for each sentence. We introduce more details of the scientific writing task and how the two writing models generate predictions and reviews in Section 6.4.2. On the other hand, we significantly improve the conversational AI explanation panel by incorporating the four design rationales described above (Section 6.4.1). Below, we elaborate on the ten formats of AI explanations included in our CONVXAI system, how we design the conversational XAI with the four principles, and the implementation of the system pipeline with details (Section 6.4.3).

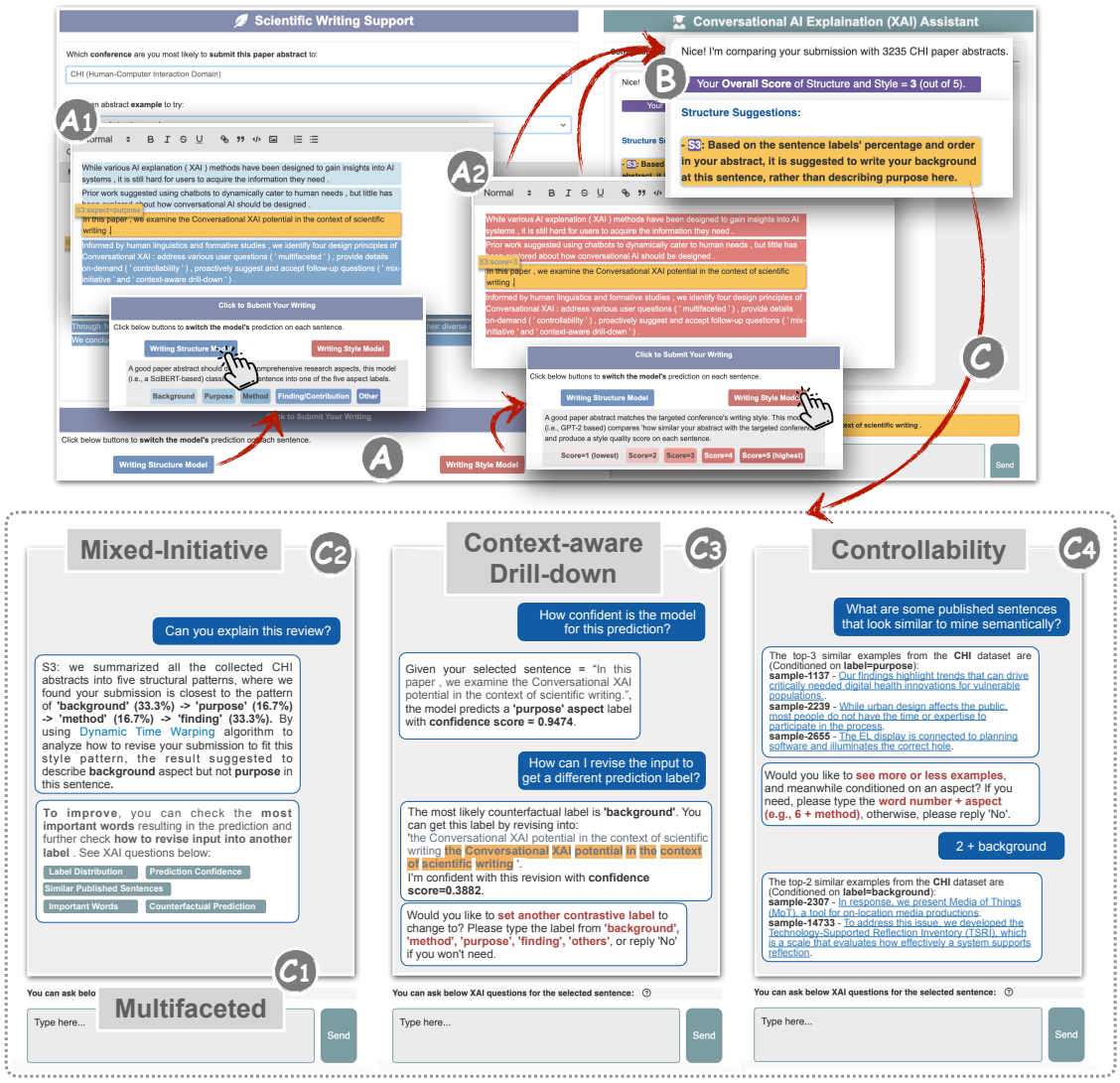


Figure 6.3: An overview of ConvXAI system. ConvXAI includes two writing models (A) to generate writing structure predictions (A1) and writing style (A2) predictions. Furthermore, the XAI agent in ConvXAI provides integrated writing review (B) followed by conversations with users to explain the writing predictions and reviews. Especially, the dialogue flows are designed to follow the four principles of *"multifaceted"* (C_1), *"mixed-initiative"*(C_2), *"context-aware drill-down"*(C_3) and *"controllability"*(C_4).

6.4.1 Overview of User-Oriented ConvXAI Design

The final ConvXAI user interface is illustrated in Figure 6.3. We significantly revise the underlying dialog mechanism based on the preliminary system according to the four design rationales, so users can interact more smoothly with the XAI agent to cater to user demands. We use Figure 6.3C to demonstrate the design.

To design ConvXAI to be **mixed-initiative** (R.2), we start the explanation dialog with a

review summary of the writing structure model and style model's outputs (Figure 6.3B). The users can select any one sentence (in this case, the third sentence with the sentence id `S3`) in this suggestion list to dive in, and start a conversation session on the sentence. Uniquely, to maintain **multifaceted** explanations (R.1) without overwhelming users, we add an additional explanation type, *understand suggestion* — answering questions like *"Can you explain this review"* — which provides general contextualization on a given suggestion (Figure 6.3C_2). To make it serve as proactive guidance towards more sophisticated XAI methods, the agent also initiates a prompt message *"to improve..."* with a subset of relevant XAIs, based on the "guess" that users would want to improve their writing at this point.

To enable **context-aware drill down** (R.3), the user questions as well as agent answers are considered subsequently. For example, in Figure 6.3C_3, the user receives a review suggesting to describe *background* aspect instead of *purpose* aspect for the selected S3. The user firstly wants to know *how confident the model makes this prediction.* Given the model confidence is quite high (around 0.95), she wanted to know how much she has to change in order to receive a different label. The agent directly contextualizes these questions based on the suggested change in Figure 6.3C_2 ("suggested to describe **background**"), and responds with a rewrite for the label *background* without having to double-check with the user first.

Still, the default may not reflect users' judgment in some cases. To mitigate potential wrong contextualization, we make the agent always proactively initiate hints for **controllablity** (R.4), *e.g.,* "would you like to..." at the bottom of Figure 6.3C_3. Figure 6.3C_4 provides a more concrete example: when the user asks for similar sentences published in the targeted conference, the XAI agent responds to the top-3 similar examples conditioned on the predicted aspect (*i.e., purpose*) by default. However, as the user is suggested to rewrite this sentence into *background*, she requests for the top-2 similar sentences which have *background* labels by specifying "2 + background", so to use those examples as gold ground truths for improving her own writing.

6.4.2 Human-AI Scientific Writing Task

We aim to provide two sets of writing support: (1) whether the abstract follows the typical semantic structure of the intended submission conferences, and (2) whether the abstract writing style matches with the conference norm. To do so, we leverage two large language models to generate predictions for each abstract sentence.

First, we use a **writing structure** model to assess the semantic structure by assessing if the abstract sufficiently covers all the required research aspects (*e.g.,* provide background context, describe the proposed method, etc.) [96] (Figure 6.3A_1). We create the model by finetuning SciBERT-base [13], a pre-trained model specifically captures scientific document contexts, on the CODA-19 datasets [96], which annotates each sentence in 10,000+ abstracts by their intended aspects, including Background, Purpose, Method, Finding/Contribution, and Other in the COVID-19 Open Research Dataset. The model achieves an F1 score of over 0.62 for each aspect and an overall accuracy of 0.7453. The model performance is demonstrated in Appendix B.0.2A.

While this model provides per-sentence predictions, the quality of an abstract depends more on the *sequence* of sentence structures. For example, "background" sentences should not be too many and should be primarily before "purpose" and "method". To support abstract improvement, we further implement a pattern explanation wrapper on top of the model, which suggests writers change some sentences' aspects to reach a better aspect pattern. For example, "background" sentences should not be too many and should be primarily before "purpose" and "method". Therefore, we provide structure *pattern* assessment, which suggests writers change some sentences' aspects to reach a better aspect pattern. Specifically, for each conference (*e.g.,* ACL), we clustered all abstracts in the conference into five groups and extracted the centers' structural patterns as the benchmark (*e.g.,* "background" (33.3%) -> "purpose" (16.7%) -> "method" (16.7%) -> "finding" (33.3%)). Afterward, we compare the submitted abstract's structural pattern with the closest pattern using the Dynamic Time Warping [164] algorithm to generate the structure suggestion for writers. See the extracted structural patterns for all conferences in Appendix B.0.2B.

Second, we use a **writing style model** to predict the style quality score for each sentence, and check if the writing style matches well with the target conference. As we intend first to support abstract improvement in the CS domain, we collect 9935 abstracts published during 2018-2022 from three conferences with relatively diverse writing styles, namely ACL (3221 abstracts), CHI (3235 abstracts), and ICLR (3479 abstracts), which are representatives of the top-tier conferences in Natural Language Processing, Human-Computer Interaction, and Machine Learning domains. More data statistics of the three conferences are in Appendix B.0.2C. To represent raw writing style match, we use the style model to assign a perplexity score [109] for each sentence, which is a measurement that approximates the sentence likelihood based on the training data. Further, since the perplexity score is quite opaque, we add a normalization layer for better readability. Specifically, we categorize the quality scores into five levels (*i.e.,* score = 1 (lowest) to 5 (highest)), which is similar to the conference review categories that writers are familiar with. To achieve these five levels, for each conference, we got the distribution of all sentences' perplexity scores, and computed the [20-th, 40-th, 60-th, 80-th] percentiles of all the scores, then divided all scores based on these percentiles. See the quality score distribution in Appendix B.0.2D.

To provide better overviews, we further offer an overall, abstract-level assessment by averaging its "overall style score" and "overall structure score". The "overall style score" is computed by averaging all sentences' quality scores. Whereas we compute the "overall structure score" as $\texttt{overall structure score} = 5 - 0.5 * \texttt{\#structure comments}$, where `#structure comments` means the number of structure reviews.

6.4.3 A Unified Interface for Heterogeneous XAIs via Conversations

6.4.3.1 CONVXAI conversational XAI pipeline.

We develop the CONVXAI system to include a web server to host the User Interface (UI), and a deep learning server with GPUs to host both the writing language models and AI explanation models. We mainly describe our implementation of the conversational XAI agent module below. Specifically, we develop the conversational XAI pipeline from scratch based on the Dialogue-State Architecture [3] from the task-oriented dialogue systems. The pipeline consists of four modules including a *Natural Language Understanding* module that classifies each XAI user question into a pre-defined user intent, which is mapped into one type of XAI algorithm. The second module, named *AI Explainers* is for generating ten types of AI explanations. Then the output is connected to the third module, named *Natural Language Generation*, to generate natural language responses that are friendly to users. On top of the pipeline, we include a Global XAI State Tracker, to record users' turn-based conversational interactions, including user intent transitions and the users' customization on AI explanations. We introduce more implementation details below.

- **Natural Language Understanding (NLU).** This module aims to parse the XAI user question and classify the user intent into which types of AI explanations they may need. We currently design the intent classifier to be a combined model of a rule-based classifier and a Deberta-based model. We trained the Deberta-based classifier [**?**] to do the intent classification, where we classify each user question into one of the eleven pre-defined XAI user intents (*i.e.,* ten user intents and the "others" type).
- **AI Explaners (XAIers).** Based on the triggered XAI user intent, this module selects the corresponding AI explainer algorithm to generate the AI explanations. Currently, we implemented the **AI Explainers** to include ten XAI methods to answer the ten XAI user questions listed in Table 6.1 correspondingly. Furthermore, we design a unified API to generate heterogeneous AI explanations to implement this *AI Explainer*, which can incorporate the four principles discussed above. For example, the *AI Explainers* enables users to input the personalized variable (*e.g.,* how many similar examples to explain) they need, and the *AI Explainers* will feed the "user-defined" variable into the AI algorithm to generate "user-customized" AI explanations.
- **Natural Language Generation (NLG).** Given the outputs from the *AI Explainers*, we leverage a template-based NLG module to convert the generated AI explanations into natural language responses. Note that we especially design the NLG templates to be multi-modal, so that it enables both free-text responses and visual-assisted responses (*e.g.,* heatmap to explain feature attributions) to meet users' needs.
- **Conversational XAI State Tracker.** As our CONVXAI empowers users to choose from multiple types of XAI methods, drill down to AI explanations and make XAI

customizations. We specifically design the global Conversational XAI State Tracker to record users' turn-based conversational interactions. Particularly, we record the turn-based user intent transitions and the users' customization on AI explanations.

Overall, we design the conversational XAI pipeline to be model agnostic and XAI algorithm agnostic. This enables the CONVXAI system to be naturally generalized to various AI task models and AI explanation methods.

6.4.3.2 Embodying Heterogenous AI Explanations in CONVXAI.

Here, we provide technical details on all the explanation methods enumerated in Table 6.1. First, **understanding data and model** requires more global explanations that summarize the training data distribution as well as the model context. For the data, we include data sheets [70] for the datasets used. We further compute important attribution distributions, including the quality scale mentioned above, the structure label distribution, and the sentence length. Such information also helps users contextualize where their abstract sits on the distribution. Similarly, for providing sufficient model information, we incorporate model cards [156] for SciBERT and GPT-2, and adjust them based on our finetuning data.

Second, for understanding and improving models, we leverage the state-of-the-art XAI algorithms to generate local AI explanations. This includes:

- **Prediction confidence**, which is the probability score after the softmax layer of the SciBERT model reflecting model prediction certainty. This explanation is only provided for the writing structure model.
- **Similar examples**, which retrieves semantically similar sentences published in the target conference to be referenced. We assess this with the dot product similarity of the sentence embeddings [179] (derived from the corresponding writing assistant models). This is provided for both writing structure and style models. [2]
- **Important words**, which aims to highlight the top-K words that attribute the writing model to the sentence prediction. We leverage the *Integrated Gradient approach* [163] to generate the word importance score (*i.e.,* attribution).
- **Counterfactual Predictions**, which re-writes the input sentence with a desired aspect while keeping the same meaning. We design an in-context learning approach using GPT3 [18] to re-write sentences. Given an input sentence, we first retrieve the top-5 semantically similar sentences for each of the five aspects from the collected CS-domain abstracts (the semantic similarity between sentences is measured by the cosine similarity over sentence embeddings [188]). A total of 25 examples would be extracted dynamically and form a prompt using the template "{example sentence} is labeled {aspect}". After

[2]Note that we deem similar examples useful mostly because users also tend to learn about the writing academic writing styles through mimicking published papers, but whether such reference counts as (or encourages) plagiarism is an open question that needs investigation.

showing 25 examples, we add "Rewrite {input sentence} into label {desired aspect}" to the prompt. GPT3 then follows the instruction to generate a modified sentence with the desired aspect label.

Finally, as described in Section 6.4.1, we further add **understanding suggestions** to answer the general question of "*how did the system generate the suggestions?*", and provide pointers to other finer-grained explanations methods. We create "suggestion explanations" for each piece of writing feedback. Particularly, we create one template for writing structure review, writing style review, and sentence length review, respectively. In each template, we describe how we compare all predictions in the abstract with the target conference data statistics to generate the corresponding review. Then we initiate an "improving message" aiming to guide users in how to use XAI to improve their writing, this message includes the buttons of potential XAI methods that we deem users might use for resolving this review (as one example shown in Figure 6.3).

6.4.4 Implementation Details

We develop CONVXAI as a stand-alone system independent of any platforms. The front-end of CONVXAI is built on the open-source Flask codebase with HTML, CSS, and Javascript codes hosted on a web server. On the other hand, the back-end of CONVXAI is a deep learning server with GeForce RTX 2080 GPUs hosting AI writing models and the conversational pipeline to generate heterogeneous AI explanations in Python and PyTorch. We also refer to ParlAI [**?**] to develop the conversational AI pipeline in CONVXAI. The front-end and back-end of CONVXAI communicate with the WebSocket protocol using the Socket.IO library and save all CONVXAI data in the MongoDB database. Around 4,300 lines of font-end codes and 6,500 lines of back-end codes are added, resulting in around **10,800 lines** of code in the final CONVXAI. Furthermore, to better generalize the unified API for conversational XAI for future study, we **extract the core unified API in CONVXAI into a Notebook**[3] for further research reference.

6.5 User Studies

We conducted two within-subjects human evaluation studies, where we compare the proposed CONVXAI against SELECTXAI, a GUI-based universal XAI system. The user study aimed to investigate how users leverage the XAIs systems to better understand the AI writing feedback and improve their scientific writing. We particularly designed the study to consist of (1) an open-ended writing task to evaluate the effectiveness of user-oriented design in the system, and (2) a well-defined writing task to investigate how systems can help users improve their scientific writing process and output in practice. Specifically, we pose the following research questions:

[3]See the unified API of conversational XAI at: `https://github.com/huashen218/convxai/blob/main/notebook_unified_XAI_API/convxai_unified_api.ipynb`

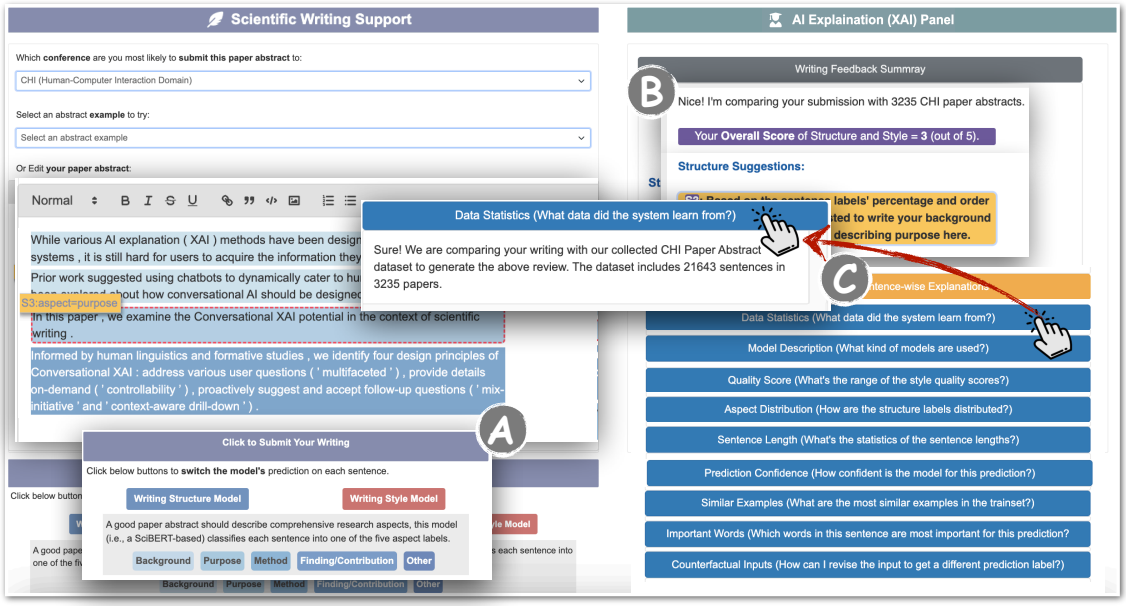


Figure 6.4: An overview of SELECTXAI system. Similarly, it includes (A) two writing models to generate writing structure predictions, and (B) integrated writing review followed by (C) static XAI buttons to show and hide the explanations.

- **RQ1**: Can user-oriented design in CONVXAI help humans better understand the AI feedback and perceive improvement in writing performance?
- **RQ2**: Can the CONVXAI be useful for humans to achieve a better writing process and output?
- **RQ3**: How do humans leverage different AI explanations in CONVXAI to finish their practical tasks?

6.5.1 Task1: Open-Ended Tasks for System Evaluation

Can CONVXAI *help users to better understand the writing feedback and improve their scientific writing? What designs support this purpose?* With these questions kept in mind, we conduct a within-subject user study comparing CONVXAI with a SELECTXAI baseline interface. Following the study, we ask participants to comment on the systems and examine how they use the CONVXAI to improve their writing by observing their interaction process.

6.5.1.1 Study Design and Procedure

Participants and* SELECTXAI *System.

We recruited 13 participants from university mailing lists. All the participants had research writing experience, resided in the U.S. and were fluent in English. The group has no overlap with the formative study participants, none of them had used CONVXAI prior to the study.

Each study lasted for one and a half hours. The participant was compensated with $40 in cash for their participation time.

We ask each participant to compare CONVXAI with a baseline system, named SELECTXAI, shown in Figure 6.4. The SELECTXAI system also consists of all the AI explanation formats included in CONVXAI. However, it statically displays all the XAI formats on the right-hand view panel instead of using dynamic conversations to convey XAIs. To display all the XAI for each sentence, users can select a sentence from the left writing editor panel to be explained, then generate all XAI formats by clicking a trigger button at the right panel. As a result, users can view all XAI formats with each having a button to control hiding and showing the AI explanations results. In other words, SELECTXAI remains multifaceted (R.1) and somewhat controllable (R.4), but does not have drill-down (R.3) or mixed-initiative properties (R.2).

Study Procedure.

We conducted *within-subjects study* where we have the same users to interact with both the proposed CONVXAI system and SELECTXAI baseline system. Each user study consists of three steps where *i)* we first instruct each user *how to use the* CONVXAI *and* SELECTXAI *systems* by showing them a live demo or recorded videos. They can stop the instruction anytime and ask any questions about the tutorials. *ii)* After the system tutorials, we invited the users to explore both CONVXAI and SELECTXAI systems with the pre-defined order. Particularly, we randomized the orders of all 13 studies. As a result, we ask 7 participants to start with the CONVXAI group, and 6 participants to start with the SELECTXAI group. *iii)* Finally, we ask the users to fill in a post-hoc survey including two demographic questions and 14 questions rating their user experience on 5 points Likert scale. We further ask them three open-form questions after the survey to interview their opinions about the CONVXAI and SELECTXAI systems.

During the step *ii)* and *iii)*, we recorded the video of the process, and encouraged them to think aloud. Besides, we designed the users to evaluate two systems either both with their own papers or both with the examples we provide. We encouraged users to use their own paper drafts where users had more incentives to improve their writing. As a consequence, 12 out of 13 users submit their own drafts or published papers.

6.5.1.2 Study Results

We first look into the overall usefulness of CONVXAI, and answer the question: is CONVXAI useful for users' ultimate goal of understanding and improving their abstract quality (*RQ1*)? We summarize participants' ratings on the two systems, CONVXAI and SELECTXAI, in Figure 6.5. We performed the non-parametric Wilcoxon signed-rank test to compare users' nominal Likert Scale ratings and found that participants self-perceived CONVXAI to **help them to better understand why their writings were given the corresponding reviews** (CONVXAI 4.07 ± 1.18 vs. SELECTXAI 3.69 ± 1.37, $p = 0.036$, Figure 6.5A). . They also felt that **CONVXAI helped them more in improving their writing** (4 ± 0.91 vs. 3.53 ± 0.77,

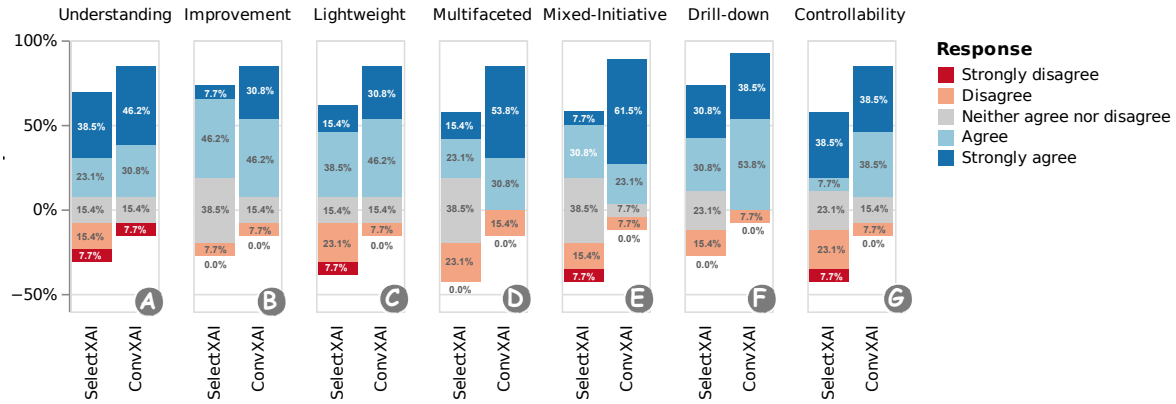


Figure 6.5: Analyses on users' self-ratings on their experiences playing with CONVXAI and SELECTXAI. They self-rated CONVXAI to be better on all dimensions, and most significantly on the usefulness of mix-initiative and multifaceted functionality.

$p = 0.019$, Figure 6.5B). The helpfulness are likely because participants can more effectively find answers to their diverse questions, which we detail in Section 6.5.1.2.

Besides their promising self-reflection, 3 out of 13 participants actually edited and iterated their abstracts in CONVXAI. They all successfully addressed the AI-raised issue (*i.e.,* the corresponding suggestion disappeared when they re-evaluated the edited version). However, the other 10 participants showed low incentive to revise the published abstracts. Through interviews, we summarize some challenges they faced in interacting with the current CONVXAI in Section 6.6.2. Through the study observations and free-form question interviews with users, we obtained that 9 out of 13 participants prefer to use CONVXAI than SELECTXAI system for improving their scientific writing. We conjecture that this might primarily result from CONVXAI's ability to answer user questions more *sufficiently*, *efficiently*, and *diversely*. More specifically, the benefit comes from three dimensions:

First, **CONVXAI reduces users' cognitive load digesting the available information**. 9 participants were overwhelmed by SELECTXAI, and complained that they had to manually click through all the available buttons before they realize all of them contain explanations in the exact same sentence. In contrast, CONVXAI releases the *same* information more *gradually* through the back-and-forth conversations. Participants especially appreciated that the initial suggestions from CONVXAI (mixed-initiative, **R2**), as it enables them to interact with the system without having to understand its full XAI capability (unlike in SELECTXAI). For example, P12 pointed out, *"it is very helpful that the XAI agent can give me some hints on using the AI explanations. Especially when I'm a novice of scientific writing and AI explanation knowledge, this helps me get involved in the system more quickly."* Indeed, this is also reflected in participants' ratings: in Figure 6.5E, participants found CONVXAI helped them figure out how to inquiry about a sentence (CONVXAI 4.23 ± 0.83 vs. SELECTXAI 3.77 ± 1.09, $p = 0.001$). Additionally, it is important that the CONVXAI is robust in detecting user intents, such as being tolerant of user input typos. As P1 and P2 mentioned, "I really like the CONVXAI that allows my typos by only capturing the keywords, so that I don't need to memorize much

knowledge for using the system."

Second, **CONVXAI enables users to pinpoint the XAI questions efficiently.** We quantified the types of questions participants frequently asked, and found 9 out of 13 participants had explicit preferences for using some specific AI explanations formats. Among these 9 users, 66.67%, 55.56%, and 33.33% participants primarily used *counterfactual explanation*, *similar example*, and *feature attribution* explanations, respectively. This suggests that, indeed, people have different kinds of questions and XAI needs. Participants liked that they could take the initiation and prioritize their own needs, and simply query the associated XAI through the dialog, whereas in SELECTXAI, "I just go over all the explanations and read everything, for some of the explanations I just don't care, this is somehow a bit overwhelming to me." (P3) This also means they were much less likely to be distracted by duplicate details (*e.g.,* P1: *"I only need to understand the general information about the model and data at the very beginning, after that, I don't need to check it repeatedly every time for each sentence."*), or explanations irrelevant to their questions. As a result, they rated CONVXAI to provide explanation more easily and more naturally (CONVXAI 4.0 ± 0.91 vs. SELECTXAI 3.3 ± 1.25, $p = 0.008$, Figure 6.5C).

Interestingly, having users to self-initiate questions brought an unexpected benefit — it helps users think through the writing and what they actually want to understand. As P6 said, *"Compared with* SELECTXAI*,* CONVXAI *slows down the interaction and gives me the time and incentive to think about what I want the robot to explain."* P4 also pointed out, *"The follow-up hints inspire me to think more about how to use the XAI for my writing."* This somewhat echoes prior work that showed pairing humans with slower AIs (that wait or take more time to make recommendations) may provide humans with a better chance to reflect on their own decisions [177].

Third, **CONVXAI provides sufficient AI explanations crafted for user need**. Interestingly, though CONVXAI and SELECTXAI implemented the same amount of explanation types and participants were overwhelmed by SELECTXAI, they still rated CONVXAI to have a more sufficient amount of explanations (multi-faceted, CONVXAI 4.23 ± 1.09 vs. SELECTXAI 3.31 ± 1.03, $p = 0.007$, Figure 6.5D). CONVXAI's controllability (CONVXAI 4.08 ± 0.95 vs. SELECTXAI 3.46 ± 1.45, $p = 0.014$, Figure 6.5G) played an important role here (CONVXAI 4.07 ± 0.95 vs. SELECTXAI 3.46 ± 1.45, $p = 0.001$, Figure 6.5E). Participants mentioned that it is essential for them to customize *how* their questions were answered, and were satisfied that they could customize the level of details in one XAI type (*e.g.,* number of similar words in feature attribution, targeted label in counterfactual prediction, etc.), whereas SELECTXAI did not provide the same level of control (as per *status-quo*). We observe all (13 out of 13) participants performed the personalized control on generating AI explanations during the user study.

The ability to drill down was equally important. We saw users performing different kinds of follow-ups based on their current explorations. For instance, as P5 mentioned, *"I would first check the* model confidence *explanation, if the confidence score is low, I would directly ignore this sentence prediction which makes my writing much easier. However, if the confidence*

A Condition	Edit-Distance ↑	Normalized-ED ↑	# Submission ↑
SelectXAI	39.75 (±22.44)	0.204 (± 0.148)	5.38 (±1.922)
ConvXAI	**56.88 (±25.02)**	**0.276 (± 0.131)**	**10.75 (±4.062)**

B Condition	Overall Writing	Writing Structure	Writing Quality
SelectXAI	3.25 (±1.035)	3.375 (±1.302)	3 (±1.195)
ConvXAI	**4.25 (±1.389)**	**4.375 (±1.408)**	**4 (±1.414)**

C Condition	Grammarly (1-100)		Model Quality (1-5)		Model Structure (1-5)		Human Quality (1-10)		Human Structure (1-10)	
	Original	Improved	Original	Improved	Original	Improved	Original	Improved	Original	Improved
SelectXAI	84.8 (±10.4)	85.1 (±5.52)	2.82 (± 0.75)	3.05 (± 0.64)	4.19 (± 0.37)	**4.75 (± 0.38)**	6.5 (±1.69)	**6.50 (±1.30)**	6.5 (±1.07)	**6.63 (±1.19)**
ConvXAI		**86.6 (±6.50)**		**3.18 (± 0.71)**		4.31 (± 0.46)		6.38 (±0.93)		**6.63 (±1.19)**

Figure 6.6: Evaluation of **Productivity** (A), **Perceived Usefulness** (B), and **Writing Performance** (C) measurements to assess users' writing performance in Task2. (A) We deploy **Productivity** with three auto-metrics including "Edit Distance", "Normalized-Edit-Distance", and "Submission Count". (B) We ask users to rate their perceived system usefulness for improving "Overall Writing", "Writing Structure", and "Writing Quality". (C) We evaluate writing outputs using both auto-metrics (*i.e.,* "Grammarly", "Model Quality", and "Model Structure"), and human evaluation (*i.e.,* "Human Quality" and "Human Structure").

score is high, I will use the counterfactual explanation *to check how to revise this sentence."* Participants also mentioned "the function of enabling users to generate these personalized explanations are the most important features" resulting in why they prefer CONVXAI over SELECTXAI systems. Like P8 pointed out, *"I think* SELECTXAI *has the advantage of easier to use because the learning curve is short. However, I would still prefer* CONVXAI *because it can provide me with much more explanations that I need."* To better understand users' preferences on explanations, we summarize some use patterns in CONVXAI in the next section.

6.5.2 Task2: Well-defined Tasks for Writing Evaluation

To answer RQ2, we further evaluate participants' productivity and writing output quality to assess the usefulness of CONVXAI and SELECTXAI on human writing performance in Task 2.

6.5.2.1 Study Design and Procedure

Participants and Grouping. We recalled 8 users, who have joined Task1 and been familiar with the system, to participate in Task2 again. There are two reasons to recruit the same group of users again: i) the experience in Task 1 could help users reduce their learning curve and cognitive load on familiarizing the XAIs and systems. Therefore, users can focus more on the writing process; ii) this design can potentially provide a temporal change in user behaviors on leveraging the systems. To conduct rigorous human studies, we divide 8 users into 4 pair of groups, with groups' research domains lying in "NLP", "HCI", "AI", and "AI", respectively.

Study design and paper selection. Similar to Task1, we also conducted a within-subjects study, but with the objective of evaluating users' scientific writing outputs with the help of CONVXAI and SELECTXAI systems. For each group of two users, we ask them to rewrite the same two papers asynchronously, with a reverse order of system assistants. For instance, within the same group, user1 rewrites with 'paper1-CONVXAI' followed by 'paper2-SELECTXAI' settings, whereas user2 rewrites with 'paper1-SELECTXAI' and 'paper2-CONVXAI' settings

successively. Hence, these settings eliminate the correlations between papers and system types and orders. Afterward, we evaluate the users' writing outputs and experience with a set of metrics, including a real-human editor evaluation, a set of auto-metrics, and a post-survey.

For a fair comparison, we pre-selected eight papers (*i.e.,* 2 papers * 4 domain groups) for users to rewrite, which are recently submitted to arXiv (*i.e.,* around Nov/29/2022) within the domains of Artificial Intelligence[4], Computation and Language[5], and Human-Computer Interaction[6]. Also, we followed a set of rules during paper selection: i) The papers are not in the top-5 best papers ranked by the editor and accepted by journals or conferences; ii) Users don't need specialized domain knowledge to improve writing. (e.g., no need to read the whole paper's contents to improve the writing); iii) The AI aspect labels and quality score predictions are correct (checked by the authors). During the study, we also recorded a video of the process and encouraged the participants to think aloud.

6.5.2.2 Study Results.

We evaluate participants' scientific writing performance quantitatively in terms of *productivity* and *writing performance* (*i.e.,* how many changes have been made and whether the improved writing outputs are scored better). Akin to Task1, we also qualitatively assess participants' *perceived usefulness* with 5 points likert scale from the post-survey.

Productivity. We evaluate *productivity* with respect to the "Edit-Distance" and the "Normalized-Edit-Distance" ("Normalized-ED") between the original paper abstract and the modified version from participants. We leverage Damerau–Levenshtein edit distance [39, 132] and its normalized version [268] to compute these two metrics. From Table 6.6 (A), we observe that participants' edit distance using the ConvXAI is 43.09% (*i.e.,* M=56.88 vs. M=39.75) higher than that using SelectXAI in average, meanwhile, the normalized edit distance is 35.29% (M=0.276 vs. M=0.204) higher comparing ConvXAI and SelectXAI as well. This demonstrates that the ConvXAI is potentially useful to help users make more modifications to writing than that using the SelectXAI system.

Besides, we also record the "Submission" counts representing how many time the users modified their draft and re-submitted to the systems. Table 6.6 (A) shows participants submitted 99.81% more times with ConvXAI than using SelectXAI during the writing, with a statistically significant difference (p=0.0045). This result also indicates users tend to interact and submit more with ConvXAI than SelectXAI for rewriting the abstracts.

These findings are consistent with the users' think-aloud notes, in which most of them preferred to use the ConvXAI than SelectXAI for improving writing. Like P5 (who uses SelectXAI first followed by ConvXAI) mentioned, "I somehow struggled with using the

[4]https://arxiv.org/list/cs.AI/recent.
[5]https://arxiv.org/list/cs.CL/recent
[6]https://arxiv.org/list/cs.HC/recent

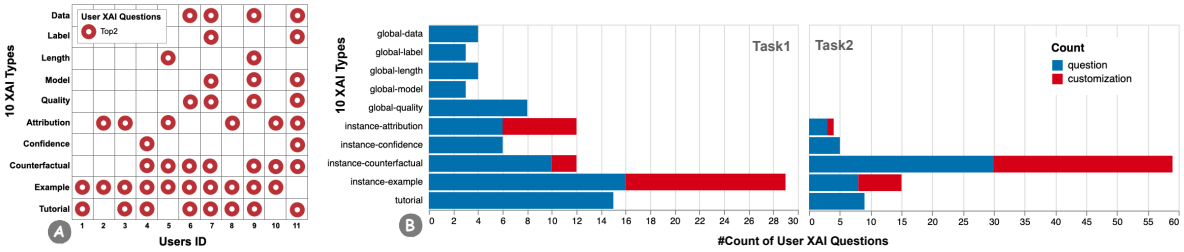


Figure 6.7: User demands analysis during using CONVXAI to improve scientific writing in Task 1 and Task 2. Particularly, (1) We ranked the top-2 most frequently requested XAI methods by each user ID in Task 1(A). (2) We compute all the users' question amounts for each of the 10 XAI methods in (B) Task 1 and Task 2.

SELECTXAI system because it provides very limited help. But I kind of started enjoying the writing process with the help of CONVXAI. "

Writing Performance. To understand whether CONVXAI can actually help users improve writing outputs, we compare the abstracts before (*i.e.,* Original) and after (*i.e.,* Improved) editing with CONVXAI and SELECTXAI as shown in Table 6.6. We evaluated abstracts using three different measurements: (*i*) Grammarly, (*ii*) CONVXAI's built-in models, and (*iii*) human evaluation. To measure the abstract quality with Grammarly, we set Grammarly's suggestion goal as audience = expert and formality = formal, manually copy-and-paste all the abstracts to Grammarly, and record the scores. Besides, we also adopt the two CONVXAI's built-in models, including the writing style model and the writing structure model. We leverage them to measure abstracts' language quality and abstract structure, respectively. These scores are also the AI scoring feedback for users during their writing tasks. For human evaluation, we hire one professional editor to rate abstracts' quality in terms of language quality and abstract structure. Note that it is difficult to find an expert who is experienced in reviewing abstracts of all "NLP", "HCI", and "AI" domains. Therefore, we are also aware of the limitation of these human evaluations.

All scores are demonstrated in Table 6.6 (C). We can observe that, by comparing with *Original* scores, **both CONVXAI and SELECTXAI are useful for humans to improve their auto-metric writing performance**, including the "Grammarly", "Model Quality", and "Model Structure" scores. Furthermore, CONVXAI specifically outperforms SELECTXAI on Grammarly and writing quality metrics, indicating that **CONVXAI can potentially help users to write better grammar-based and style-based sentences** in scientific abstracts than SELECTXAI. On the other hand, the human editor's evaluation shows inconsistent results, where **CONVXAI and SELECTXAI can both improve the writing Structure** evaluations, but not in the Quality metric. To probe the inconsistency between human and auto-metric evaluations, we further compute the Pearson correlation between the model scores and the human ratings and find that both quality and structure are negatively correlated or not correlated (quality: -0.0311 and structure: -0.1150), showing that there is a misalignment between humans and models.

Therefore, we posit that both universal XAI systems, including ConvXAI and SelectXAI, are useful to improve human writing performance under auto-metric evaluations. Particularly, ConvXAI can outperform SelectXAI in terms of grammar and style-based writing quality. Besides, as the human is not aligned with model evaluations based on Pearson correlations, the improvement failed in the human quality metric. This negative finding actually provides valuable insights into the importance of aligning the human judgment and model objective in AI tasks, so that users can use the systems to effectively reach both improvement goals.

Perceived Usefulness. In the post-survey, we also ask users to rate their perception of system usefulness in terms of assisting their abstract writing. We particularly measured the users' perceived usefulness on "Overall Writing", "Writing Structure" improvement, and "Writing Quality" improvement. We design these three metrics to be consistent with the feedback from the AI writing models. Shown in Table 6.6 (B), we can see participants perceived ConvXAI to be 1 (out of 5) point higher than SelectXAI in terms of use on all writing aspects. At the end of the survey, we further ask which AI explanations or system functions they perceived to be most useful, we elaborate on this finding in Sec 6.5.3 below.

6.5.3 Usage Patterns with ConvXAI

We propose ConvXAI based on the statement that universal XAI interfaces are important for satisfying user demands in real-world practice. In this section, we provide practical evidence to support that the *universal XAI interface is indeed a necessary design of useful XAI for real-world user needs*.

By reviewing all 11 (from Task 1) and 8 (from Task 2) recorded study videos, we collected all the users' XAI question requests when they leverage ConvXAI to improve writing. In total, there are **95** and **92** XAI user requests in Task 1 and Task 2, respectively. Based on analyzing these XAI user requests., we demonstrate Figure 6.7 to provide detailed insights on practical user demands. More specifically, in Figure 6.7 (1), we visualize each individual user's top-2 priority in using the different XAI methods. In Figure 6.7 (2), we accumulate all users' requests on each XAI method to visualize the usage distribution among the ten XAI methods. We also separately visualize Task 1 and Task 2 in order to observe the temporal usage patterns on XAI methods. We summarize our findings in detail below.

6.5.3.1 Different users prioritize different AI explanations and orders for their needs.

First, focusing on the same task but with different users, we observe that *different users often prioritize different types of AI explanations even within the same task*. In specific, for Task 1 shown in Figure 6.7 (A1), although 9 users (*i.e.,* 1,2,3,4,6,7,8,9) prioritize using "Examples" explanations, the other 2 users (*i.e.,* 5,11) leverage "Attributition" and "Confidence" explanations most in their writing task 1. Besides, the 2nd-popular AI explanations of the 11 users are scattered among all the 10 XAI types without a unified pattern.

Additionally, in Task 2 with Figure 6.7 (B1), we can see users' top2 explanations are converging into instance-wise explanations (*i.e.,* "Attribution", "Counterfactual", etc). In

specific, 7 out of 8 users prioritize "Counterfactual" and the other one leveraged "Example" explanation the most. This is also consistent with the user's think-aloud observation. For instance, P5 lacks an AI background and didn't understand what "Prediction Confidence" means in this situation, whereas P11 mentioned *"model confidence is the first explanation I'll ask to decide whether I'll ignore the prediction or continue the explanations."*

Furthermore, we accumulate the users' XAI request counts for each XAI type and show the results of Task 1 and Task 2 in Figure 6.7 (A2) and (B2), respectively. We can observe that although user needs are often dominated by one XAI type ("Example" and "Counterfactual" in Task 1 and 2, respectively), users also leverage ConvXAI to probe a wide range of other XAI types, such as "XAI tutorial", "Confidence", "Attribution", etc.) In short, these findings validate that it is important to use the universal XAI interface like ConvXAI, which can **accommodate different users' backgrounds and practical demands**.

6.5.3.2 User demands are changing over time.

In addition, we focus on the changes of user demands over time. We specifically compare the same user group's XAI needs in the two Tasks. By comparing Figure 6.7 (A1) vs. (B1), we can see that the top of users' XAI demands is gradually converging into the instance-wise explanations, including "Counterfactual", "Example", "Confidence", "Tutorial" and "Attribution" explanations.

This can be further verified by comparing Figure 6.7 (A2) vs. (B2). We can see that i) user demands in Task 2 are highly skewed to "Counterfactual" explanations, which are two times more than the "Example" explanation ranked as top in Task 1. ii) Users leverage much less and even no global information explanations (*e.g.,* "Data", "Model", "Length", etc) in Task 2. This is also consistent with the user think-aloud notes, where P4 pointed out "After I know these data and model information, I might not need them again a lot, unless I need this information to analyze each sentence's prediction later."

This again shows that it is important to design XAI systems to be a universal yet flexible XAI interface, as ConvXAI, to **capture the dynamic changes of user needs over time**.

6.5.3.3 Proactive XAI tutorials are imperative to improve the XAI usefulness.

Both our pilot study and the two tasks illustrate that **providing users with instructions on how to use XAI is crucial**. Particularly, echoing the "Mixed-initiative" design principle, we proactively give hints of XAI use patterns (*i.e.,* how to use AI explanations) for improving writing during the conversations. In Table 6.2, we exemplify a set of user patterns to resolve different AI writing feedback.

From Figure 6.7 (A1) and (B1), we can observe that 72.73% (8 out of 11) users and 37.5% (3 out of 8) users prioritize "Tutorial" explanations as top-2 during Task 1 and 2, respectively. Similarly, in Figure 6.7 (A2) and (B2), the accumulated counts of "Tutorial" explanations also ranked within top-3 in both Task 1 and 2, indicating a high user demand for checking tutorial/hints of XAI usage patterns.

Improvement Goal	Usage Patterns	Explanation
Change Structure Label	**Pattern1:** Counterfactual Explanation (use target-label).	Ask GPT-3 model to rewrite the sentence into the target aspect.
	Pattern2: Similar Examples (label: target-label, rank: quality_score).	Refer to similar examples with the target labels and high-quality scores.
Lengthen/Shorten Length	**Pattern1:** Similar Examples (rank: short).	Refer to similar short examples for rewriting.
	Pattern2: Rewrite while keeping important_words.	Find Important Words, then keep them during rewriting to keep the correct aspects.
Improve Quality	**Pattern1:** Counterfactual Explanation (use same label).	Ask GPT-3 model to paraphrase the original sentence.
	Pattern2: Similar Examples (rank: quality_score).	Refer to similar examples with high-quality scores.

Table 6.2: Examples of Use Patterns shown in the "Tutorial" explanations suggested by the CONVXAI system.

Furthermore, we also observe a decreasing trend of "Tutorial" explanation needs over time by comparing Task 1 and Task 2. This potentially indicates that users are gradually being more proficient in using AI explanations for their own needs.

6.5.3.4 XAI Customization is crucial.

By observing the think-aloud interviews in the two tasks, we deem one fundamental reason that CONVXAI outperforms SELECTXAI is that it provides much more flexible customization for the user request. This corresponds to the "Controllability" design principle derived from the pilot study as well. Note that we only design 3 out of 10 AI explanations to enable XAI customization. Particularly, we allow users to specify one variable (*i.e.,* "target-label") for generating "Counterfactual" and "Attribution" explanations, and four variables (*i.e.,* "target-label", "example-count", "rank-method", "keyword") to generate "Example" explanations.

Importantly, by visualizing Figure 6.7 (A2) and (B2), we observe that there are 22.11% and 40.22% practical user requests for XAI customization in Task 1 and 2, respectively. Besides, all users in both Task 1 and 2 requested XAI customization during their studies. These findings indicate that **enabling users to customize their personal XAI needs is crucial in practice**.

6.5.3.5 Same feedback can be resolved with different AI explanations.

Additionally, we observe that the same writing feedback can be resolved with different AI explanations. As shown in Table 6.2, we demonstrate two use pattern examples to resolve each type of AI prediction feedback as the "hints" of how to use XAI within CONVXAI systems.

Correspondingly, we also find different users choose different AI explanations to resolve similar problems. For instance, when users receive a suggestion to rewrite the sentence into another aspect label, some participants directly ask for *counterfactual explanations* to change the label (*e.g.,* P1, P7, P8), whereas others might refer to *similar examples* to understand the conference published sentences first, and then revise their own writings (*e.g.,* P2, P6, P9, P11). Further, even the same people could use different XAIs based on different scenarios. As P1 mentioned *"If time is urgent, I'll use counterfactual explanation because they are straightforward. However, when I have more time, I'll use similar example explanations because I can potentially learn more writing skills from them."*

6.6 Discussion and Limitations

In this work, we propose CONVXAI as a unified XAI interface in the form of conversations. We especially incorporate practical user demands, representing as the four design principles collected from the formative study, into the CONVXAI design. As a result, users are able to better leverage the multi-faceted, mixed-initiated, context-aware, and customized AI explanations in CONVXAI to achieve their tasks (*e.g.,* scientific abstract writing). The CONVXAI design and findings can potentially shed light on developing more useful XAI systems. Additionally, we have released the core codes of unified XAI APIs and the complete code base of CONVXAI. The CONVXAI can be generalized to a variety of applications since the unified XAI methods and interface are model-agnostic.

In this section, we further elaborate on the core ingredients for useful XAIs based on user study observations with CONVXAI, the system generalizability, and the empirical limitations. As a novel model of a unified XAI interface using conversations, we believe it provides a valuable grounding on how future conversational XAI systems should be developed to better meet real-world user demands.

6.6.1 Crucial Ingredients of Useful XAI

We design CONVXAI system as a prototypical yet potential solution of **useful AI explanation systems** in real-world tasks. The rationale is to mitigate the gaps between the practical, diverse, and dynamic user demands of existing AI explanations via a unified XAI interface in the form of conversations. Especially, we aim to probe "*what are the crucial ingredients of useful XAI systems?*" during the one formative study and two human evaluation tasks. In summary, we elaborate on our preliminary findings of useful XAI systems should potentially incorporate four factors, including: " **Integrated XAI interface + proactive XAI tutorial + customized XAIs + lightweight XAI display**". We elaborate on each ingredient with supportive evidence in our studies for more details.

6.6.1.1 Integrated XAI interface accessible to multi-faceted XAIs.

In Sec 6.5.3 and Figure 6.7, we demonstrate diverse XAI user needs and usage patterns from empirical observations. This indicates that XAI user demands are generally dynamically changed across different users and over time. Therefore, it is essential to empower users to choose the appropriate XAIs on their own preferences. Users can therefore leverage an integrated XAI interface with access to multi-faceted XAIs for their needs.

6.6.1.2 Proactive XAI usage tutorial.

From the formative study, we learned that it is difficult for users to figure out "how to leverage and combine the power of different XAI types to finish their practical goals". This finding motivates the "Mix-initiated" design principle, and resulting in designing XAI "tutorial" expla-

nations to instruct users. Moreover, the two user studies provide evidence (*i.e.,* in Sec 6.5.3.3 and Figure 6.7) that users indeed request many XAI tutorial explanations during the writing tasks, but the requested amount is gradually decreased as the users getting more proficient in using the ConvXAI system.

6.6.1.3 Customized XAI interactions.

Users commonly demand more controllability in generating AI explanations. We observed these user demands from both the formative study (*i.e.,* leading to "Controllability" design principle), and two user studies. More quantitatively, we provide evidence (in Sec 6.5.3.4 and Figure 6.7) that although only 3 out of 10 XAI types allow customization, all users leverage XAI customization to generate XAIs. Further, the demands for XAI customization increase over time.

6.6.1.4 Lightweight XAI display with details-on-demands.

By conducting user studies with both ConvXAI and SelectXAI, we observe that users prefer the XAI interface to be versatile yet simple. Regarding this, a details-on-demand approach using conversations (*e.g.,* ConvXAI) is more appropriate, as the users can directly pinpoint the expected XAI type as they need. We provide supportive evidence by comparing ConvXAI (details-on-demand) and SelectXAI (full initial disclosure) in Sec 6.5.1 and Sec 6.5.2.

6.6.2 Limitations

Although the ConvXAI performs mostly better in assisting users in understanding the writing feedback and improving their scientific writings, there are still factors and limitations to be noted when deploying ConvXAI in practice. Here, we discuss potential obstacles they faced and potential fixes to improve ConvXAI.

Users have a steeper learning curve to use ConvXAI. In interviewing the users about the advantages and disadvantages of the two systems, we found participants, especially those with less AI knowledge, experienced a steeper learning curve to use the ConvXAI– That says, participants need more effort to learn what answers they can expect from the XAI agent. In comparison, they think SelectXAI is much simpler to interact with because all the answers they can get are displayed in the interface. However, some participants also mentioned that they would like to spend the efforts to learn ConvXAI since it provides more potential explanations to be used. From the above observation, we deem that the ConvXAI system can be improved by providing the "instruction of system capability range" at the initial user interaction stages, and this learning effort will disappear when users interact more with ConvXAI in the long run.

The performance of writing models and XAI algorithms influence the user experience of ConvXAI. Another phenomenon we observed is that the under-performed model and XAI algorithm quality can influence the user experience, such as trust and satisfaction. Note that

in real-world AI tasks, humans are commonly motivated to use the XAI methods to analyze AI predictions, such as improving writing performance according to the AI writing feedback with the help of ConvXAI's XAI methods. However, there are situations that AI writing feedback is misaligned with human judgment. In these situations, users commonly ignore the misaligned feedback which can potentially reduce satisfaction and trust in the AI prediction models. To mitigate this issue, we posit two actions to resolve: i) it is important to align the AI models' predictions and feedback with human judgment before asking users to leverage analysis methods (*e.g.,* XAIs in ConvXAI) to explain or interact with the AI predictions. ii) if the AI task is difficult thus, it's inevitable to occur misalignment (*e.g.,* the scientific writing task in this study), enabling human intervention in the models' prediction outputs can alleviate the harm to user experience. For example, when P4 met the misalignment between the model output and his own judgment, he mentioned, "it would be great if I can manually make the model ignore this review so that the score can reflect my performance more fairly."

6.6.3 Future Directions

Contextualize for the right user group. During the studies, we found different users with different backgrounds requested diverse levels of AI explanation details for the same XAI question. For instance, when asking for the model description explanations, AI experts mostly looked for more model details such as the model architecture, how it was trained, etc. In contrast, participants less familiar with AI knowledge only wanted to see the high-level model information, such as who released the model and if it is reliable, etc. The observation echoes the motivating example used by [169], and indicates that users who have different backgrounds need different granularity levels of AI explanations. While most XAI methods tend to provide user-agnostic information, it might be promising to wrap them based on intended user groups, *e.g.,* with non-experts getting the simplified versions with all the jargon removed or explained. Prior work has also noted that users' perceptions on automated systems can be shaped by conceptual metaphors [118], which is also an interesting presentation method to explore.

Characterize the paths and connections between XAI methods. We observe two interesting usage patterns of XAI methods in ConvXAI: First, different XAI methods can serve different roles in a conversation. For example, explanations on the training data information and model accuracy are static enough that it is sufficient to only describe them once in the ConvXAI tutorial; feature attributions and model performance confidence tend to be treated as the basic explanation and initial exploration points, whereas counterfactual explanations are most suitable for follow-ups Second, some explanation methods can lead to natural drill-downs. For example, we may naturally consider editing the most important words to get counterfactual explanations, *after* we identify those words in feature attributions). If we more rigorously inspect the best roles of, and links between, explanation methods, we may be able to create a graph connecting them. Tracing the graph should help us understand and implement what context should be kept for what potential follow-ups.

Meanwhile, while we encourage continuous conversations, we also observe that as the conversation becomes longer, the earlier information is usually flushed out, and it becomes hard

to stay on top of the entire session. Some users suggested promising directions, one participant recommended "slicing the dialogue into sessions, where each session only discusses one specific sentence." Alternatively, advanced visual signals that reflect conversation structures [116] (*e.g.,* the hierarchical dropdown in Wikum reflecting information flow [271]) could help people trace back to earlier snippets.

Incorporate multi-modality. While our current controls and user queries tend to be explicit, prior work envisioned much more implicit control signals. For example, [124] envisioned the Natural Language Understanding unit should be able to parse sentences like "Wow, it's surprising that...", decipher users' intent on querying outlier feature importance, and provide appropriate responses. Identifying users' emotional responses to certain explanations (*e.g.,* surprised, frustrated, affirmed) could be an interesting way to point to potential control responses.

Though natural language interaction is intuitive, not all information needs to be conveyed through dialog. Inspired by SELECTXAI's flat learning curve, a combination of natural language inquiry and traditional WIMP interaction could make the system easier to grasp. Future work can survey how people might react to buttons or sliders that allow them to control the number of words or the number of similar examples to inspect.

6.7 Conclusion

In this study, we present CONVXAI, a system to support scientific writing via conversational AI explanations. Informed by linguistic properties of human conversation and empirical formative studies, we identify four design principles of Conversational XAI. That says – these systems should address various user questions ("multi-faceted"), provide details on-demand ("controllability"), and should actively suggest and accept follow-up questions ("mix-initiative" and "context-aware drill-down"). We further build up an interactive prototype to instantiate these rationales, in which paper writers can interact with various state-of-the-art explanations through a typical chatbot interface. Through 21 user studies, we show that conversational XAI is promising for prompting users to think through what questions they want to ask, and for addressing diverse questions. We conclude by discussing the use patterns of CONVXAI, as well as implications for future conversational XAI systems.

Chapter 7
Conclusion and Future Work

7.1 Conclusion

In this dissertation, I provide the readers with a thorough overview of human-centered useful Ae I explanations: the human evaluations on the usefulness of AI interpretability (PART I), the investigation of challenges (PART II), and proposed interactive AI explanation approaches (PART III).

In PART I, we focus on exploring the research question: *If the state-of-the-art AI explanation approaches are useful for humans in real-world practice*?

In Chapter 3, explaining to users why automated systems make certain mistakes is important and challenging. Researchers have proposed ways to automatically produce interpretations for deep neural network models. However, it is unclear how *useful* these interpretations are in helping users figure out why they are getting an error. If an interpretation effectively explains to users how the underlying deep neural network model works, people who were presented with the interpretation should be better at predicting the model's outputs than those who were not. This paper presents an investigation on whether or not showing machine-generated visual interpretations helps users understand the **incorrectly predicted labels** produced by image classifiers. We showed the images and the correct labels to 150 online crowd workers and asked them to select the incorrectly predicted labels with or without showing them the machine-generated visual interpretations. The results demonstrated that displaying the visual interpretations did not increase, but rather *decreased*, the average guessing accuracy by roughly 10%.

In Chapter 4, existing self-explaining models typically favor extracting the shortest possible rationales — snippets of an input text "responsible for" corresponding output — to explain the model prediction, with the assumption that shorter rationales are more intuitive to humans. However, this assumption has yet to be validated. Is the shortest rationale indeed the most human-understandable? To answer this question, we design a self-explaining model, LIMITEDINK, which allows users to extract rationales at any target length. Compared to existing baselines, LIMITEDINK achieves compatible end-task performance and human-annotated rationale agreement, making it a suitable representation of the recent class of self-explaining

models. We use LIMITEDINK to conduct a user study on the impact of rationale length, where we ask human judges to predict the sentiment label of documents based only on LIMITEDINK-generated rationales with different lengths. We show rationales that are too short do not help humans predict labels better than randomly masked text, suggesting the need for more careful design of the best human rationales.

PART II further explores the disparities between the status quo of AI explanations and real-world user demands.

In Chapter 5, It is unclear if existing interpretations of deep neural network models respond effectively to the needs of users. This paper summarizes the common *forms* of explanations (such as feature attribution, decision rules, or probes) used in over 200 recent papers about natural language processing (NLP), and compares them against user questions collected in the XAI Question Bank [138]. We found that although users are interested in explanations for *the road not taken* — namely, why the model chose one result and not a well-defined, seemly similar legitimate counterpart — most model interpretations cannot answer these questions.

PART III presents the proposed solution to improve human-centered useful interpretability – conversational AI explanations.

In Chapter 6, while various AI explanation (XAI) methods have been proposed to interpret AI systems, whether the state-of-the-art XAI methods are practically useful for humans remains inconsistent findings. To improve the usefulness of XAI methods, a line of studies identifies the gaps between the diverse and dynamic real-world user needs with the status quo of XAI methods. Although prior studies envision mitigating these gaps by integrating multiple XAI methods into the universal XAI interfaces (*e.g.,* conversational or GUI-based XAI systems), there is a lack of work investigating how these systems should be designed to meet practical user needs. In this study, we present CONVXAI, a conversational XAI system that incorporates multiple XAI types, and empowers users to request a variety of XAI questions via a universal XAI dialogue interface. Particularly, we innovatively embed practical user needs (*i.e.,* four principles grounding on the formative study) into CONVXAI design to improve practical usefulness. Further, we design the domain-specific language (DSL) to implement the essential conversational XAI modules and release the core conversational universal XAI API for generalization. The findings from two within-subjects studies with 21 users show that CONVXAI is more useful for humans in perceiving the understanding and writing improvement, and improving the writing process in terms of productivity and sentence quality. Finally, our work contributes insight into the design space of useful XAI, reveals humans' XAI usage patterns with empirical evidence in practice, and identifies opportunities for future useful XAI work. We release the open-sourced CONVXAI codes for future study.

7.2 Future Work

Putting it all together, we are really excited to contribute to this field and witness the progress that has been made in this field for the past years. Meanwhile, we also deeply believe that there

is still a long way to go toward useful AI explanations for humans in a wide range of real-world use scenarios.

One key challenge is that we still don't have well-studied benchmarks to define and measure the objective and subjective usefulness metrics for AI explanations that can represent a wide range of complicated use cases in practice. These benchmarks not only include scientific measurements to assess AI explanation usefulness on human performance – often occurs with *how much can AI explanations be useful for human accuracy, efficiency, understanding*? But also, they involve assessments on model performance, such as *how much can AI explanation help improve the model performance (*e.g., *accuracy, fairness)*? As a result, there is a lack of well-defined assessments that can fairly compare and even guide the numerous AI explanation approaches toward the objective of being practically useful for human-AI collaboration. However, we believe interactive AI explanations can serve as a start for future research to explore more space for useful AI explanations in practice.

In the future, taking one step from generating faithful and plausible AI explanations, we will have to develop more AI explanation approaches from a "human-AI-team-centered" perspective to make explanations more useful for human-AI collaborations. We also hope to encourage more researchers to work on diverse topics related to improving the usefulness of AI explanations, such as *building interactive AI explanations to fulfill practical human-AI use cases*, *developing benchmarks to evaluate AI explanation usefulness*, etc. We believe that it will lead us towards building better AI explanation, and further push into the more outperformed, fair, trustworthy, and safe human-AI collaborations in general.

Appendix A
LIMITEDINK

A.0.1 Model Details and Hyperparameters

Methodology Details

Concrete Relaxation of Subset Sampling Process. Given the output logits of *identifier*, we use Gumbel-softmax [107] to generate a concrete distribution as $\mathbf{c} = [c_1, ...c_n] \sim \text{Concrete}(\mathbf{idn}(\mathbf{x}))$, represented as a one-hot vector over n features where the top important feature is 1. We then sample this process k times in order to sample top-k important features, where we obtain k concrete distributions as $\{\mathbf{c}^1, ..., \mathbf{c}^k\}$. Next we define one n-dimensional random vector $\mathbf{m}$ to be the element-wise maximum of these k concrete distributions along n features, denoted as $\mathbf{m} = \max_j\{\mathbf{c}_i^j\}_{i=n}^{j=k}$. Discarding the overlapping features to keep the rest, we then use $\mathbf{m}$ as the k-hop vector to approximately select the top-k important features over document $\mathbf{x}$.

Vector and sort regularization. We deploy a *vector and sort* regularization on mask $\mathbf{m}$ [64], where we sort the output mask m in a increasing order and minimize the L_1 norm between m and a reference $\hat{m}$ consisting of $n - k$ zeros followed by k ones.

Model Training Details

Training and inference. During training, we select the Adam optimizer with the learning rate at 2e-5 with no decay. We set hyperparameters in Equation A.3 and 4.2 as $\lambda = 1e - 4$, $v_1 = 0.5$ and $v_2 = 0.3$ and trained 6 epochs for all models. Furthermore, we train CONVXAI on a set of sparsity levels as $k = \{10\%, 20\%, 30\%, 40\%, 50\%\}$ and choose models with optimal predictive performance on validation sets.

Details of Self-Explaining Baselines

We compare our method with state-of-the-art self-explaining baseline models.

Sparse-N (Minimization Norm). This method learns the short mask with minimal L_0 or

L_1 norm [10, 128], which penalizes for the total number of selected words in the explanation.

$$\min \ \mathbb{E}_{\mathbf{z}\sim\mathbf{idn(x)}}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda||m|| \tag{A.1}$$

Sparse-C (Controlled Norm Minimization). This method controls the mask sparsity through a tunable predefined sparsity level α [24, 106]. The mask is penalized as below as long as the sparsity level α is passed.

$$\min \ \mathbb{E}_{\mathbf{z}\sim\mathbf{idn(x)}}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda \max(0, \frac{||\mathbf{m}||}{N} - \alpha) \tag{A.2}$$

where N is the input length and $||m||$ denotes mask penalty with L_1 norm.

Sparse IB (Controlled Sparsity with Information Bottleneck). This method introduces a prior probability of $\mathbf{z}$, which approximates the marginal $p(\mathbf{m})$ of mask distribution; and $p(\mathbf{m}|\mathbf{x})$ is the parametric posterior distribution over $\mathbf{m}$ conditioned on input $\mathbf{x}$ [176]. The sparsity control is achieved via the information loss term, which reduces the KL divergence between the posterior distribution $p(\mathbf{m}|\mathbf{x})$ that depends on $\mathbf{x}$ and a prior distribution $r(\mathbf{m})$ that is independent of $\mathbf{x}$.

$$\min \ \mathbb{E}_{\mathbf{z}\sim\mathbf{idn(x)}}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda KL[p(\mathbf{m}|\mathbf{x}), r(\mathbf{m})] \tag{A.3}$$

A.0.2 Ablation Study on Model Components

We provide an ablation study on the Movie dataset to evaluate each loss term's influence on end-task prediction performance, including Precision, Recall, and F1 scores. The result is shown in Table A.1.

Setups	**End-Task Prediction**		
	Precision	**Recall**	**F1**
No Sufficiency	0.25	0.50	0.34
No Continuity	0.82	0.81	0.81
No Sparsity	0.80	0.79	0.79
No Contextual	0.83	0.83	0.83
Our Model	**0.91**	**0.90**	**0.90**

Table A.1: Ablation study of each module in our model on Movie Review dataset.

A.0.3 Additional Details of Human Study

Generating Random Baselines Human accuracy likely increases when participants can

see more words, *i.e.*, when the lengths of rationales increase. If a rationale and a random text span have the same number of words, the rationale should help readers predict the label better. We created a simple baseline that generated rationales by randomly selecting words to form the rationales. We could control (1) how many words to select and (2) how many disjointed rationales to produce. In the study, we set these two numbers to be identical to that of CONVXAI at each length level.

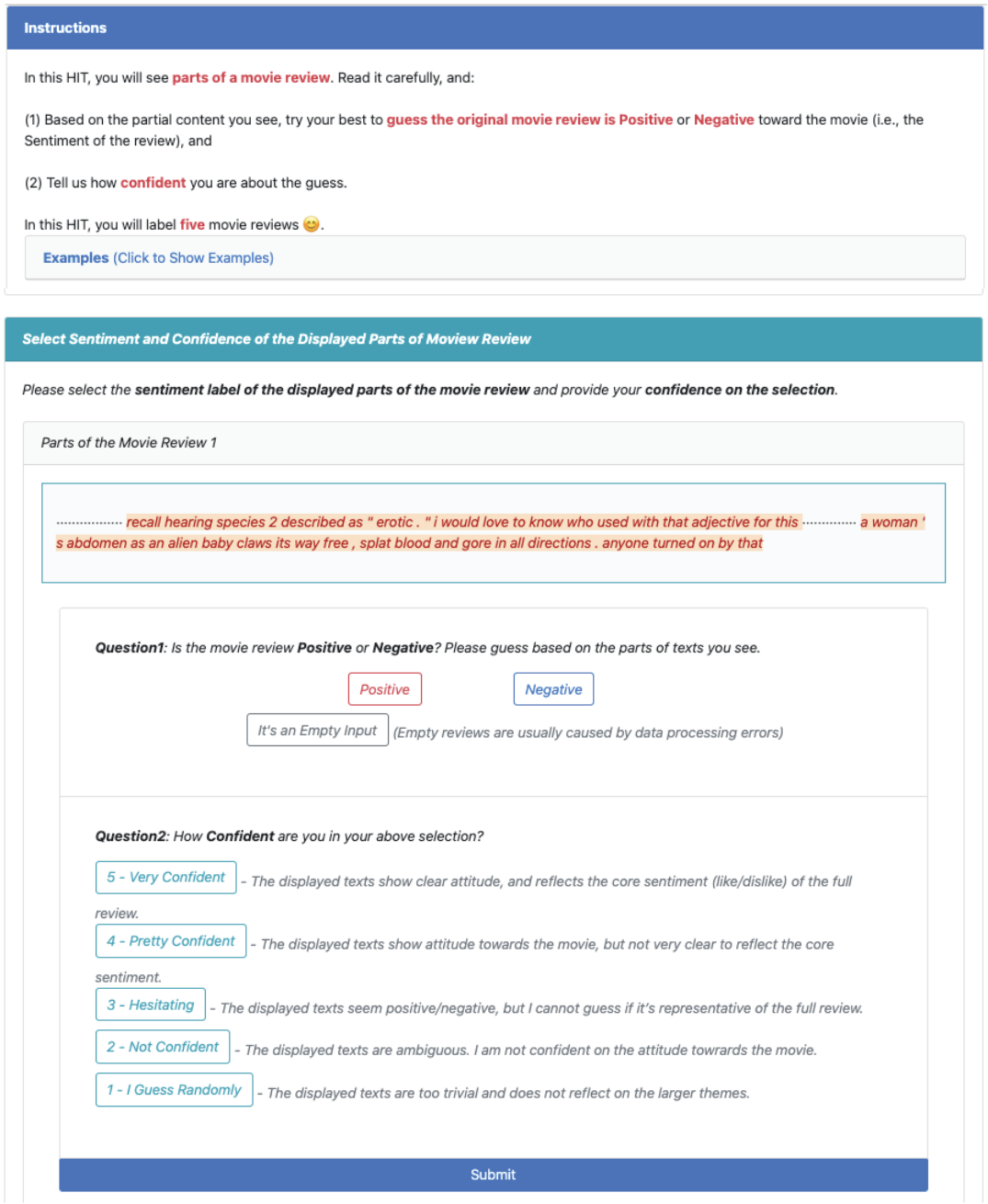
In detail, given the rationale length k, we first got the count of total tokens in rationale as #tokens = k. Next, we computed the average number of rationale segments m, which are generated by CONVXAI, over the Movie dataset. We randomly selected m spans with total tokens' count as #tokens from the full input texts, thus obtaining the random baselines. We evenly separated 10 worker groups to finish five random baseline HITs and CONVXAI HITs each. We determined that good model rationales should get higher human accuracy compared with same-length random baselines.

Human Evaluation User Interface

We provide our designed user interfaces used in the human study. Specifically, we show the interface of the human study panel in Figure A.1 (B). We also provide the detailed instructions for workers to understand our task, the instruction inteface is shown in Figure A.2.

	Review1	Review2	Review3	Review4	Review5
Worker Group 1	Our@10%	Our@20%	Our@30%	Our@40%	Our@50%
Worker Group 2	Our@20%	Our@30%	Our@40%	Our@50%	Our@10%
Worker Group 3	Our@30%	Our@40%	Our@50%	Our@10%	Our@20%
Worker Group 4	Our@40%	Our@50%	Our@10%	Our@20%	Our@30%
Worker Group 5	Our@50%	Our@10%	Our@20%	Our@30%	Our@40%
Worker Group 6	Random@10%	Random@20%	Random@30%	Random@40%	Random@50%
Worker Group 7	Random@20%	Random@30%	Random@40%	Random@50%	Random@10%
Worker Group 8	Random@30%	Random@40%	Random@50%	Random@10%	Random@20%
Worker Group 9	Random@40%	Random@50%	Random@10%	Random@20%	Random@30%
Worker Group 10	Random@50%	Random@10%	Random@20%	Random@30%	Random@40%

(A) Worker Group Assignment



(B) Worker Study Interface

Figure A.1: (A) The design of the worker group assignment in our human study. (B) The worker interface of the human study.

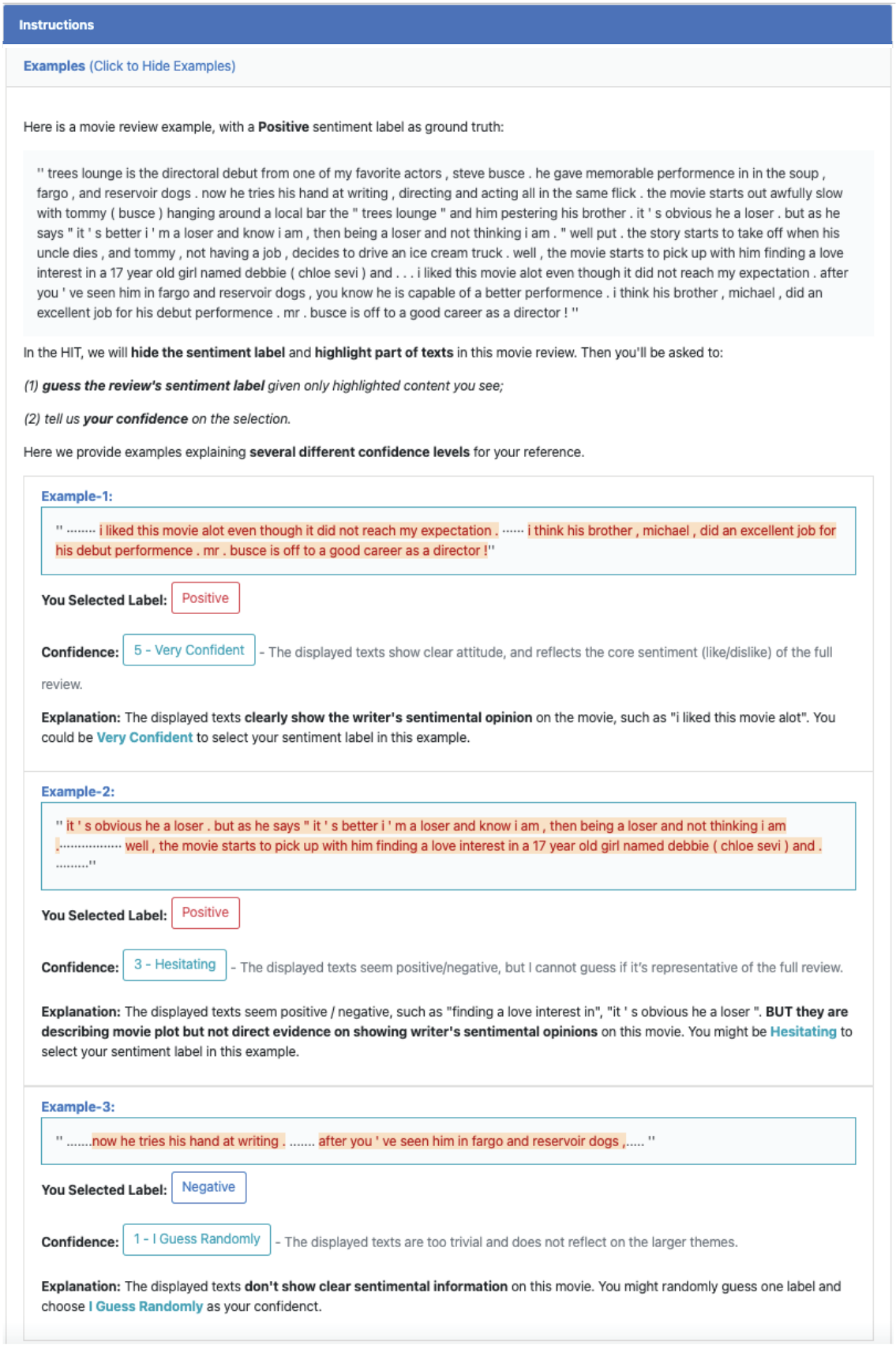


Figure A.2: User Interface of the instruction in the human study.

Appendix B
ConvXAI

B.0.1 Formative Study

Participants Details. In order to capture the user demands of conversational XAI systems from more comprehensive and representative views, we recruited seven participants with diverse backgrounds and occupations in the formative study. The demographic statistics of the seven participants are summarized in Table B.1(A). Specifically, we invited 7 participants, including 3 females and 4 males. In detail, we collected and recorded the participants' information according to the criteria as: **Writing Expr.** (*i.e.,* how many years of scientific writing experience do they have?): <1 year; 1-3 years; 3-5 years; 5-10 years; >10 years; **AI Knowlg.** (*i.e.,* what level of AI Knowledgeability would they describe themselves?): 5 - I am a machine learning expert; 4 - I know a lot about machine learning; 3 - I know some knowledge about machine learning; 2 - I know little knowledge about machine learning; 1 - I never heard about machine

A

ID	Writing Expr.	AI Knowlg.	# Paper	Occupation
1	5-10 years	5	>10	Assistant Professor
2	3-5 years	4	>10	Ph.D. candidate
3	<1 year	2	<1	Upcoming Master student
4	1-3 years	3	5-10	Ph.D. student
5	> 10 years	1	>10	Senior Applied Scientist
6	1-3 years	2	5-10	Software Engineer
7	5-10 years	4	>10	Applied Scientist

B

	Multifaceted (R1)	Mixed-Initiate (R2)	Context-aware Drill-down (R3)	Controllability (R4)
Interactive Dialogue	✘	✓	✓	✘
SelectXAI	✓	✘	✘	✓
ConvXAI	✓	✓	✓	✓

Table B.1: (A) The demographic statistics of the users in the formative study. We recruit seven participants with diverse backgrounds and occupations in order to capture the user needs for the conversational XAI system in more comprehensive views. (B) The four design principles for conversational XAI systems summarized from the formative study. We further compare the existing systems (*i.e.,* Interactive Dialogue [232, 239]), the baseline (*i.e.,* SelectXAI) and our proposed ConvXAI system, regarding these four principles.

	background	purpose	method	finding	other
#Samples	5062	821	2140	6890	562
Precision	0.714	0.610	0.720	0.789	0.811
Recall	0.783	0.637	0.623	0.758	0.857
F1	0.747	0.623	0.668	0.773	0.833

A

	# Abstract	# Sentence	Avg Sent Len
ACL	3221	20744	26
CHI	3235	21643	25
ICLR	3479	25873	27

C

	20%th	40%th	50%th	60%th	80%th
ACL	22	32	39	46	71
CHI	32	45	53	63	97
ICLR	35	52	62	74	116

D

Conf.	B Aspect Patterns
ACL	1. 'background' (25%) -> 'purpose' (12.5%) -> 'method' (37.5%) -> 'finding' (25%); 2. 'background' (33.3%) -> 'purpose' (16.7%) -> 'method' (16.7%) -> 'finding' (33.3%); 3. 'background' (42.9%) -> 'method' (28.6%) -> 'finding' (28.5%); 4. 'background' (50%) -> 'purpose' (16.7%) -> 'finding' (33.3%); 5. 'background' (25%) -> 'finding' (12.5%) -> 'method' (12.5%) -> 'finding' (50%);
CHI	1. 'background' (42.9%) -> 'purpose' (14.3%) -> 'finding' (42.9%); 2. 'background' (22.2%) -> 'purpose' (11.2%) -> 'method' (33.3%) -> 'finding' (33.3%); 3. 'background' (33.3%) -> 'purpose' (16.7%) -> 'method' (16.7%) -> 'finding' (33.3%); 4. 'background' (33.3%) -> 'method' (16.7%) -> 'finding' (50%); 5. 'background' (20%) -> 'finding' (6.7%) -> 'background' (13.3%) -> 'purpose' (6.7%) -> 'background' (13.3%) -> 'finding' (6.7%) -> 'method' (6.7%) -> 'finding' (26.7%);
ICLR	1. 'background' (33.3%) -> 'purpose' (16.7%) -> 'method' (16.7%) -> 'finding' (33.3%); 2. 'Method' (20%) -> 'finding' (80%); 3. 'background' (42.9%) -> 'purpose' (14.2) -> 'finding' (42.9%); 4. 'background' (45.5%) -> 'method' (9.1%) -> 'finding' (9.1%) -> 'method' (9.1%) -> 'finding' (27.3%); 5. 'Background' (22.2%) -> 'purpose' (11.1%) -> 'method' (33.3%) -> 'finding' (33.4%);

Table B.2: The summary of writing models' performance. The writing structure model performance (with fine-tuned Sci-BERT language model) is shown in (A); (B) shows the extracted five aspect patterns for each conference; the data statistics of three conferences in terms of abstract number, sentence number and average sentence length in (C) and the quality score distribution in (D).

learning. **# Paper** (*i.e.,* how many submitted papers do they have?): <1; 1-3; 3-5; 5-10; >10. **Occupation**: we also record the occupation of each participant.

B.0.2 Writing Model Performance

We summarize the writing model performance in the Figure B.2. We can observe the writing structure model performance of the fine-tuned Sci-BERT language model is shown in Figure B.2A. The model accuracy is 0.7453. Figure B.2B shows the extracted five aspect patterns for each conference. Further, we can see the data statistics of three conferences in terms of abstract number, sentence number and average sentence length in Figure B.2C and the quality score distribution in Figure B.2D.

Bibliography

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

[2] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. pages 9505–9515, 2018.

[3] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

[5] 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*, pages 412–421, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[6] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[7] M. Aubakirova and M. Bansal. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2035–2041, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[8] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *CHI*, 2021.

[9] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, S. Schockaert, and H. Saggion. Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4766–4771, Brussels, Belgium, 2018. Association for Computational Linguistics.

[10] J. Bastings, W. Aziz, and I. Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy, 2019. Association for Computational Linguistics.

[11] J. Bastings and K. Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online, 2020. Association for Computational Linguistics.

[12] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, 2017. Association for Computational Linguistics.

[13] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[14] T. Bickmore and J. Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403, 2001.

[15] P. Bloom. How children learn the meanings of words. In *United Kingdom: MIT Press*, 2002.

[16] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357, 2016.

[17] R. Bommasani, K. Davis, and C. Cardie. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online, 2020. Association for Computational Linguistics.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[19] Z. Bucinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*, pages 454–464, 2020.

[20] O. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information*

Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 9560–9572, 2018.

[21] O. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572, 2018.

[22] S. Carton, A. Rathore, and C. Tan. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online, 2020. Association for Computational Linguistics.

[23] C. Chang, E. Creager, A. Goldenberg, and D. Duvenaud. Explaining image classifiers by counterfactual generation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[24] S. Chang, Y. Zhang, M. Yu, and T. S. Jaakkola. Invariant rationalization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR, 2020.

[25] T.-Y. Chang and Y.-N. Chen. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6066–6072, 2019.

[26] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi. Ai-assisted peer review. *Humanities and social sciences communications*, 8(1):1–11, 2021.

[27] H. Chen and Y. Ji. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4236–4251, Online, 2020. Association for Computational Linguistics.

[28] J. Chen, S.-t. Lin, and G. Durrett. Multi-hop question answering via reasoning chains. *ArXiv preprint*, abs/1910.02610, 2019.

[29] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of IEEE Conference on Machine Learning (ICML)*, 2018.

[30] A. Choudhry, M. Sharma, P. Chundury, T. Kapler, D. W. Gray, N. Ramakrishnan, and N. Elmqvist. Once upon a time in visualization: Understanding the use of textual narratives for causality. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1332–1342, 2020.

[31] E. Chu, D. Roy, and J. Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *ArXiv preprint*, abs/2007.12248, 2020.

[32] E. Chu, D. Roy, and J. Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

[33] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[34] A. Coenen, L. Davis, D. Ippolito, E. Reif, and A. Yuan. Wordcraft: A human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*, 2021.

[35] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, 2018. Association for Computational Linguistics.

[36] P. Dabkowski and Y. Gal. Real time image saliency for black box classifiers. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6967–6976, 2017.

[37] P. Dabkowski and Y. Gal. Real Time Image Saliency for Black Box Classifiers. 2017.

[38] F. Dalvi, N. Durrani, H. Sajjad, Y. Belinkov, A. Bau, and J. R. Glass. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6309–6317. AAAI Press, 2019.

[39] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[40] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas, 2016. Association for Computational Linguistics.

[41] R. Das, A. Godbole, M. Zaheer, S. Dhuliawala, and A. McCallum. Chains-of-reasoning at TextGraphs 2019 shared task: Reasoning over chains of facts for explainable multi-hop inference. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 101–117, Hong Kong, 2019. Association for Computational Linguistics.

[42] N. De Cao, M. S. Schlichtkrull, W. Aziz, and I. Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online, 2020. Association for Computational Linguistics.

[43] 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[44] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, 2020. Association for Computational Linguistics.

[45] K. Dhamdhere, M. Sundararajan, and Q. Yan. How important is a neuron. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[46] S. Ding, H. Xu, and P. Koehn. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy, 2019. Association for Computational Linguistics.

[47] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *ArXiv preprint*, abs/1702.08608, 2017.

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

[49] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

[50] M. Du, N. Liu, Q. Song, and X. Hu. Towards explanation of dnn-based prediction with guided feature inversion. In Y. Guo and F. Farooq, editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1358–1367. ACM, 2018.

[51] M. Du, N. Liu, F. Yang, and X. Hu. Learning credible deep neural networks with rationale regularization. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 150–159. IEEE, 2019.

[52] W. Du, Z. M. Kim, V. Raheja, D. Kumar, and D. Kang. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. *arXiv preprint arXiv:2204.03685*, 2022.

[53] P. Dufter and H. Schütze. Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1185–1191, Hong Kong, China, 2019. Association for Computational Linguistics.

[54] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*, pages 211–223, 2018.

[55] K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[56] A. Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[57] A. Ettinger, A. Elgohary, and P. Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany, 2016. Association for Computational Linguistics.

[58] E. Fast, B. Chen, J. Mendelsohn, J. Bassen, and M. S. Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[59] J. Feng, C. Shaib, and F. Rudzicz. Explainable clinical decision support from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1478–1489, Online, 2020. Association for Computational Linguistics.

[60] 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*, pages 229–239, 2019.

[61] 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 (IUI)*, pages 229–239, 2019.

[62] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium, 2018. Association for Computational Linguistics.

[63] R. Fok and D. S. Weld. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *arXiv preprint arXiv:2305.07722*, 2023.

[64] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2950–2958, 2019.

[65] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3449–3457. IEEE Computer Society, 2017.

[66] J. Fu, P. Liu, and G. Neubig. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online, 2020. Association for Computational Linguistics.

[67] J. Fu, P. Liu, and G. Neubig. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online, 2020. Association for Computational Linguistics.

[68] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang, and B. Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online, 2020. Association for Computational Linguistics.

[69] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[70] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[71] K. Gero, A. Calderwood, C. Li, and L. Chilton. A design space for writing support tools using a cognitive process model of writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 11–24, 2022.

[72] R. Ghaeini, X. Fern, and P. Tadepalli. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium, 2018. Association for Computational Linguistics.

[73] B. Ghai, Q. V. Liao, Y. Zhang, R. Bellamy, and K. Mueller. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. *ArXiv preprint*, abs/2001.09219, 2020.

[74] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[75] M. Giulianelli, J. Harding, F. Mohnert, D. Hupkes, and W. Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium, 2018. Association for Computational Linguistics.

[76] F. Godin, K. Demuynck, J. Dambre, W. De Neve, and T. Demeester. Explaining character-aware neural networks for word-level prediction: Do they discover linguistic rules? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3275–3284. Association for Computational Linguistics, 2018.

[77] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 531–535, New York, NY, USA, 2020. Association for Computing Machinery.

[78] A. V. Gonzalez, G. Bansal, A. Fan, R. Jia, Y. Mehdad, and S. Iyer. Human evaluation of spoken vs. visual explanations for open-domain qa. *arXiv preprint arXiv:2012.15075*, 2020.

[79] F. Graliński, A. Wróblewska, T. Stanisławek, K. Grabowski, and T. Górecki. GEval: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy, 2019. Association for Computational Linguistics.

[80] C. Grimsley, E. Mayfield, and J. R.S. Bursten. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1780–1790, Marseille, France, May 2020. European Language Resources Association.

[81] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[82] N. Gupta, K. Lin, D. Roth, S. Singh, and M. Gardner. Neural module networks for reasoning over text. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[83] X. Han, B. C. Wallace, and Y. Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[84] B. Hancock, P. Varma, S. Wang, M. Bringmann, P. Liang, and C. Ré. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia, 2018. Association for Computational Linguistics.

[85] Y. Hao, L. Dong, F. Wei, and K. Xu. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, Hong Kong, China, 2019. Association for Computational Linguistics.

[86] P. Hase and M. Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, 2020. Association for Computational Linguistics.

[87] B. Herman. The promise and peril of human evaluation for model interpretability. *ArXiv preprint*, abs/1711.07414, 2017.

[88] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics.

[89] J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[90] J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[91] D. J. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107:65–81, 1990.

[92] B. Hoover, H. Strobelt, and S. Gehrmann. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online, 2020. Association for Computational Linguistics.

[93] P. M. Htut, J. Phang, S. Bordia, and S. R. Bowman. Do attention heads in bert track syntactic dependencies? *ArXiv preprint*, abs/1911.12246, 2019.

[94] C.-Y. Huang, M.-H. Chen, and L.-W. Ku. Towards a better learning of near-synonyms: Automatically suggesting example sentences via fill in the blank. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 293–302, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[95] C.-Y. Huang, S.-H. Huang, and T.-H. K. Huang. Heteroglossia: In-situ story ideation with the crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[96] T.-H. Huang, C.-Y. Huang, C.-K. C. Ding, Y.-C. Hsu, and C. L. Giles. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. *arXiv preprint arXiv:2005.02367*, 2020.

[97] Y.-C. Huang, H.-C. Wang, and J. Y.-j. Hsu. Feedback orchestration: Structuring feedback for facilitating reflection and revision in writing. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 257–260, 2018.

[98] D. Hupkes, S. Veldhoen, and W. Zuidema. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926, 2018.

[99] M. Hurtado Bodell, M. Arvidsson, and M. Magnusson. Interpretable word embeddings via informative priors. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6323–6329, Hong Kong, China, 2019. Association for Computational Linguistics.

[100] I. Hutchby and R. Wooffitt. *Conversation analysis*. Polity, 2008.

[101] 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*, pages 4198–4205, Online, 2020. Association for Computational Linguistics.

[102] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *ACL'20*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[103] A. Jacovi, O. Sar Shalom, and Y. Goldberg. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium, 2018. Association for Computational Linguistics.

[104] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.

[105] S. Jain and B. C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[106] S. Jain, S. Wiegreffe, Y. Pinter, and B. C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online, 2020. Association for Computational Linguistics.

[107] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. 2017.

[108] G. Jawahar, B. Sagot, and D. Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, 2019. Association for Computational Linguistics.

[109] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

[110] Y.-H. Jen, C.-Y. Huang, M. Chen, T.-H. Huang, and L.-W. Ku. Assessing the helpfulness of learning materials with inference-based learner-like agent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3807–3817, Online, 2020. Association for Computational Linguistics.

[111] H. Jhamtani and P. Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online, 2020. Association for Computational Linguistics.

[112] Y. Jiang and M. Bansal. Self-assembling modular networks for interpretable multi-hop reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4474–4484, Hong Kong, China, 2019. Association for Computational Linguistics.

[113] Y. Jiang, N. Joshi, Y.-C. Chen, and M. Bansal. Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2714–2725, Florence, Italy, 2019. Association for Computational Linguistics.

[114] Y. Jiang, N. Joshi, Y.-C. Chen, and M. Bansal. Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2714–2725, Florence, Italy, 2019. Association for Computational Linguistics.

[115] L. Jin, D. King, A. Hussein, M. White, and D. Danforth. Using paraphrasing and memory-augmented models to combat data sparsity in question interpretation with a virtual patient dialogue system. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 13–23, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

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

[117] D. Kaushik, E. Hovy, and Z. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

[118] P. Khadpe, R. Krishna, L. Fei-Fei, J. T. Hancock, and M. S. Bernstein. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[119] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR, 2018.

[120] S. Kim, J. Yi, E. Kim, and S. Yoon. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, Online, 2020. Association for Computational Linguistics.

[121] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 2017.

[122] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

[123] 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*, IUI '15, page 126–137, New York, NY, USA, 2015. Association for Computing Machinery.

[124] 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.

[125] A. M. Lauretig. Identification, interpretability, and bayesian word embeddings. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[126] D.-H. Lee, R. Khanna, B. Y. Lin, S. Lee, Q. Ye, E. Boschee, L. Neves, and X. Ren. LEAN-LIFE: A label-efficient annotation framework towards learning from explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 372–379, Online, 2020. Association for Computational Linguistics.

[127] M. Lee, P. Liang, and Q. Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[128] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, 2016. Association for Computational Linguistics.

[129] P. Lertvittayakumjorn, L. Specia, and F. Toni. Find: human-in-the-loop debugging deep text classifiers. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[130] P. Lertvittayakumjorn and F. Toni. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5195–5205, Hong Kong, China, 2019. Association for Computational Linguistics.

[131] P. Lertvittayakumjorn and F. Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021.

[132] V. I. Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[133] 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*, pages 681–691, San Diego, California, 2016. Association for Computational Linguistics.

[134] J. Li, W. Monroe, and D. Jurafsky. Understanding neural networks through representation erasure. *ArXiv preprint*, abs/1612.08220, 2016.

[135] M. Li, J. Weston, and S. Roller. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*, 2019.

[136] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo. Tell-and-answer: Towards explainable visual question answering using attributes and captions. *arXiv preprint arXiv:1801.09041*, 2018.

[137] Q. Li, Q. Tao, S. Joty, J. Cai, and J. Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567, 2018.

[138] 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*, pages 1–15, 2020.

[139] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[140] B. Y. Lim and A. K. Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 195–204, 2009.

[141] Y. Lin, Y. C. Tan, and R. Frank. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, 2019. Association for Computational Linguistics.

[142] H. Liu, Q. Yin, and W. Y. Wang. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy, 2019. Association for Computational Linguistics.

[143] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[144] Z. Liu, Z.-Y. Niu, H. Wu, and H. Wang. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792, Hong Kong, China, 2019. Association for Computational Linguistics.

[145] J. Lu, C. Zhang, Z. Xie, G. Ling, T. C. Zhou, and Z. Xu. Constructing interpretive spatio-temporal features for multi-turn responses selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Florence, Italy, 2019. Association for Computational Linguistics.

[146] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. 2017.

[147] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[148] T. Manzini, L. Yao Chong, A. W. Black, and Y. Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[149] W. Marrakchi. *Explaining by Conversing: The Argument for Conversational Xai Systems*. PhD thesis, Harvard University, 2021.

[150] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv preprint*, abs/1802.03426, 2018.

[151] D. A. Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. 2018.

[152] P. Michel, O. Levy, and G. Neubig. Are sixteen heads really better than one? In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024, 2019.

[153] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013.

[154] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[155] T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. 2017.

[156] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[157] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, Online, 2020. Association for Computational Linguistics.

[158] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[159] S. Moon, P. Shah, A. Kumar, and R. Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854. Association for Computational Linguistics, 2019.

[160] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran, and J. Chu-Carroll. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online, 2020. Association for Computational Linguistics.

[161] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[162] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia, 2018. Association for Computational Linguistics.

[163] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere. Did the model understand the question? *arXiv preprint arXiv:1805.05492*, 2018.

[164] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[165] W. J. Murdoch, P. J. Liu, and B. Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[166] W. J. Murdoch and A. Szlam. Automatic rule extraction from long short term memory networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[167] S. Murty, P. W. Koh, and P. Liang. Expbert: Representation engineering with natural language explanations. *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[168] D. Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[169] N. Nobani, F. Mercorio, and M. Mezzanzanica. Towards an explainer-agnostic conversational xai. In *IJCAI*, pages 4909–4910, 2021.

[170] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 97–105, 2019.

[171] B. Nushi, E. Kamar, and E. Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *The 6th AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[172] A. Panchenko, F. Marten, E. Ruppert, S. Faralli, D. Ustalov, S. P. Ponzetto, and C. Biemann. Unsupervised, knowledge-free, and interpretable word sense disambiguation. *arXiv preprint arXiv:1707.06878*, 2017.

[173] A. Panigrahi, H. V. Simhadri, and C. Bhattacharyya. Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, Florence, Italy, 2019. Association for Computational Linguistics.

[174] A. Panigrahi, H. V. Simhadri, and C. Bhattacharyya. Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, Florence, Italy, 2019. Association for Computational Linguistics.

[175] N. Pappas and A. Popescu-Belis. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*, pages 455–466, 2014.

[176] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online, 2020. Association for Computational Linguistics.

[177] J. S. Park, R. Barber, A. Kirlik, and K. Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15, 2019.

[178] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics.

[179] P. Pezeshkpour, S. Jain, B. C. Wallace, and S. Singh. An empirical comparison of instance attribution methods for nlp. *arXiv preprint arXiv:2104.04128*, 2021.

[180] P. Pezeshkpour, Y. Tian, and S. Singh. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3336–3347, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[181] F. Poursabzi-Sangdeh, D. Goldstein, J. Hofman, J. W. Vaughan, and H. Wallach. Manipulating and measuring model interpretability. *CHI*, 2021.

[182] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.

[183] D. Rajagopal, V. Balachandran, E. H. Hovy, and Y. Tsvetkov. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[184] N. F. Rajani, B. McCann, C. Xiong, and R. Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy, 2019. Association for Computational Linguistics.

[185] N. F. Rajani and R. J. Mooney. Ensembling visual explanations for vqa. In *Proceedings of the NIPS 2017 workshop on Visually-Grounded Interaction and Language (ViGIL)*, 2017.

[186] N. F. Rajani and R. J. Mooney. Using explanations to improve ensembling of visual question answering systems. In *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, pages 43–47, 2017.

[187] S. Reddy, D. Chen, and C. D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[188] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[189] M. Ren, X. Geng, T. Qin, H. Huang, and D. Jiang. Towards interpretable reasoning over paragraph effects in situation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6745–6758, Online, 2020. Association for Computational Linguistics.

[190] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal,

D. Shen, and R. Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[191] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535. AAAI Press, 2018.

[192] M. Roemmele and A. S. Gordon. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*, pages 81–92. Springer, 2015.

[193] A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[194] A. Ross, A. Marasović, and M. E. Peters. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985*, 2020.

[195] S. Rothe, S. Ebert, and H. Schütze. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California, 2016. Association for Computational Linguistics.

[196] C. Roy, M. Shanbhag, M. Nourani, T. Rahman, S. Kabir, V. Gogate, N. Ruozzi, and E. D. Ragan. Explainable activity recognition in videos. In *IUI Workshops*, 2019.

[197] M. Saeidi, M. Bartolo, P. Lewis, S. Singh, T. Rocktäschel, M. Sheldon, G. Bouchard, and S. Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium, 2018. Association for Computational Linguistics.

[198] S. Saha, S. Ghosh, S. Srivastava, and M. Bansal. PRover: Proof generation for interpretable reasoning over rules. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 122–136, Online, 2020. Association for Computational Linguistics.

[199] A. Saleh, T. Deutsch, S. Casper, Y. Belinkov, and S. Shieber. Probing neural dialog models for conversational understanding. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 132–143, Online, 2020. Association for Computational Linguistics.

[200] S. C. Salveter. Natural language processing. *American Journal of Computational Linguistics*, 7(4), 1981.

[201] C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy, 2019. Association for Computational Linguistics.

[202] R. Schwarzenberg, L. Raithel, and D. Harbecke. Neural vector conceptualization for word vector space interpretation. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[203] 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 (ICCV)*, pages 618–626, 2017.

[204] P. Sen, M. Danilevsky, Y. Li, S. Brahma, M. Boehm, L. Chiticariu, and R. Krishnamurthy. Learning explainable linguistic expressions with neural inductive logic programming for sentence classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4211–4221, Online, 2020. Association for Computational Linguistics.

[205] L. K. Senel, I. Utlu, F. Şahinuç, H. M. Ozaktas, and A. Koç. Imparting interpretability to word embeddings while preserving semantic structure. *ArXiv preprint*, abs/1807.07279, 2018.

[206] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koc, and T. Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779, 2018.

[207] S. Serrano and N. A. Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[208] V. Setlur and M. Tory. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[209] W. Shalaby and W. Zadrozny. Learning concept embeddings for dataless classification via efficient bag-of-concepts densification. *Knowledge and Information Systems*, 61(2):1047–1070, 2019.

[210] H. Shen, C.-Y. Huang, T. Wu, and T.-H. Huang. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. *arXiv preprint arXiv:2305.09770*, 2023.

[211] H. Shen and T.-H. Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172, 2020.

[212] H. Shen and T.-H. Huang. Explaining the road not taken. *ACM CHI 2022 Workshop on Human-Centered Explainable AI*, 2021.

[213] H. Shen and T. Wu. Parachute: Evaluating interactive human-lm co-writing systems. *ACM CHI 2023 Workshop on In2Writing*, 2023.

[214] H. Shen, T. Wu, W. Guo, and T.-H. Huang. Are shortest rationales the best explanations for human understanding? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 10–19, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[215] H. Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, and A. Stolcke. Improving fairness in speaker verification via group-adapted fusion network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7077–7081. IEEE, 2022.

[216] H. Shen, V. Zayats, J. C. Rocholl, D. D. Walker, and D. Padfield. Multiturncleanup: A benchmark for multi-turn spoken conversational transcript cleanup. *arXiv preprint arXiv:2305.12029*, 2023.

[217] X. Shi, I. Padhi, and K. Knight. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas, 2016. Association for Computational Linguistics.

[218] X. Shi, I. Padhi, and K. Knight. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas, 2016. Association for Computational Linguistics.

[219] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier, 2003.

[220] D. Slack, S. Krishna, H. Lakkaraju, and S. Singh. Talktomodel: Explaining machine learning models with interactive natural language conversations. 2022.

[221] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. SmoothGrad: Removing Noise by Adding Noise. In *International Conference on Machine Learning Workshop on Visualization for Deep Learning*, 2017.

[222] 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*, pages 1–13, 2020.

[223] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. L. Boyd-Graber, D. S. Weld, and L. Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ML. In R. Bernhaupt, F. F. Mueller, D. Verweij,

J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, and R. Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020.

[224] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics.

[225] K. Sokol and P. A. Flach. Glass-box: Explaining ai decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *IJCAI*, pages 5868–5870, 2018.

[226] J. Stadelmaier and S. Padó. Modeling paths for explainable knowledge base completion. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 147–157, Florence, Italy, 2019. Association for Computational Linguistics.

[227] A. Stepanjans and A. Freitas. Identifying and explaining discriminative attributes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4313–4322, Hong Kong, China, 2019. Association for Computational Linguistics.

[228] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363, 2018.

[229] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2017.

[230] A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, and E. H. Hovy. SPINE: sparse interpretable neural embeddings. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4921–4928. AAAI Press, 2018.

[231] S. Subramanian, B. Bogin, N. Gupta, T. Wolfson, S. Singh, J. Berant, and M. Gardner. Obtaining faithful interpretations from compositional neural networks. *arXiv preprint arXiv:2005.00724*, 2020.

[232] Y. Sun and S. S. Sundar. Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[233] A. Sydorova, N. Poerner, and B. Roth. Interpretable question answering on knowledge bases and text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4943–4951, Florence, Italy, 2019. Association for Computational Linguistics.

[234] H. Tan and M. Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[235] Z. Tang, G. Hahn-Powell, and M. Surdeanu. Exploring interpretability in event extraction: Multitask learning of a neural event classifier and an explanation decoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 169–175, Online, 2020. Association for Computational Linguistics.

[236] I. Tenney, D. Das, and E. Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, 2019. Association for Computational Linguistics.

[237] 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*, pages 107–118, Online, 2020. Association for Computational Linguistics.

[238] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. V. Durme, S. R. Bowman, D. Das, and E. Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[239] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[240] Y.-H. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1823–1833, Online, 2020. Association for Computational Linguistics.

[241] K. Vafa, Y. Deng, D. Blei, and A. Rush. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10314–10332, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[242] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui. Attention interpretability across nlp tasks. *ArXiv e-prints*, 2019.

[243] J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, 2019. Association for Computational Linguistics.

[244] T.-T. Vu and G. Haffari. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3048–3053, Brussels, Belgium, 2018. Association for Computational Linguistics.

[245] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, 2019. Association for Computational Linguistics.

[246] E. Wallace, J. Tuyls, J. Wang, S. Subramanian, M. Gardner, and S. Singh. Allennlp interpret: A framework for explaining predictions of nlp models. 2019.

[247] C. Wang. Interpreting neural network hate speech classifiers. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 86–92, Brussels, Belgium, 2018. Association for Computational Linguistics.

[248] Y. Wang, P. Venkatesh, and B. Y. Lim. Interpretable directed diversity: Leveraging model explanations for iterative crowd ideation. In *CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

[249] Z. Wang, Y. Qin, W. Zhou, J. Yan, Q. Ye, L. Neves, Z. Liu, and X. Ren. Learning from explanations with neural execution tree. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[250] Z. Wang, Y. Zhang, M. Yu, W. Zhang, L. Pan, L. Song, K. Xu, and Y. El-Kurdi. Multi-granular text encoding for self-explaining categorization. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 41–45. Association for Computational Linguistics, 2019.

[251] 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)*, pages 11–20, Hong Kong, China, 2019. Association for Computational Linguistics.

[252] T. Wolfson, M. Geva, A. Gupta, M. Gardner, Y. Goldberg, D. Deutch, and J. Berant. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198, 2020.

[253] J. Wu, Z. Hu, and R. Mooney. Generating question relevant captions to aid visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3585–3594, Florence, Italy, 2019. Association for Computational Linguistics.

[254] J. Wu and R. J. Mooney. Self-critical reasoning for robust visual question answering. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8601–8611, 2019.

[255] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*, 2021.

[256] Z. Wu, Y. Chen, B. Kao, and Q. Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online, 2020. Association for Computational Linguistics.

[257] X. Xiao, L. Wang, B. Fan, S. Xiang, and C. Pan. Guiding the flowing of semantics: Interpretable video captioning via POS tag. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2068–2077, Hong Kong, China, 2019. Association for Computational Linguistics.

[258] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[259] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015.

[260] V. Yadav, S. Bethard, and M. Surdeanu. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4514–4525, Online, 2020. Association for Computational Linguistics.

[261] C. Yang, A. Rangarajan, and S. Ranka. Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE, 2018.

[262] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics.

[263] C. Yeh, J. S. Kim, I. E. Yen, and P. Ravikumar. Representer point selection for explaining deep neural networks. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9311–9321, 2018.

[264] S.-F. Yeh, M.-H. Wu, T.-Y. Chen, Y.-C. Lin, X. Chang, Y.-H. Chiang, and Y.-J. Chang. How to guide task-oriented chatbot users, and when: A mixed-methods study of combinations of chatbot guidance types and timings. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[265] Y. You, W. Jia, T. Liu, and W. Yang. Improving abstractive document summarization with salient information modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2132–2141, Florence, Italy, 2019. Association for Computational Linguistics.

[266] M. Yu, S. Chang, Y. Zhang, and T. Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China, 2019. Association for Computational Linguistics.

[267] W. Yuan, P. Liu, and G. Neubig. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*, 2021.

[268] L. Yujian and L. Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[269] O. Zaidan and J. Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii, 2008. Association for Computational Linguistics.

[270] F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, and F. Fallucchi. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online, 2020. Association for Computational Linguistics.

[271] A. X. Zhang, L. Verou, and D. Karger. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2082–2096, 2017.

[272] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[273] Z. Zhang, J. Singh, U. Gadiraju, and A. Anand. Dissonance between human and machine understanding. 3(CSCW), 2019.

[274] T. Zhao, K. Lee, and M. Eskenazi. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1098–1107. Association for Computational Linguistics, 2018.

[275] A. Zupon, M. Alexeeva, M. Valenzuela-Escárcega, A. Nagesh, and M. Surdeanu. Lightly-supervised representation learning with global interpretability. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 18–28, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.