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

HUA SHEN, Pennsylvania State University, USA
CHIEH-YANG HUANG, Pennsylvania State University, USA
TONGSHUANG WU, Carnegie Mellon University, USA
TING-HAO (KENNETH) HUANG, Pennsylvania State University, USA

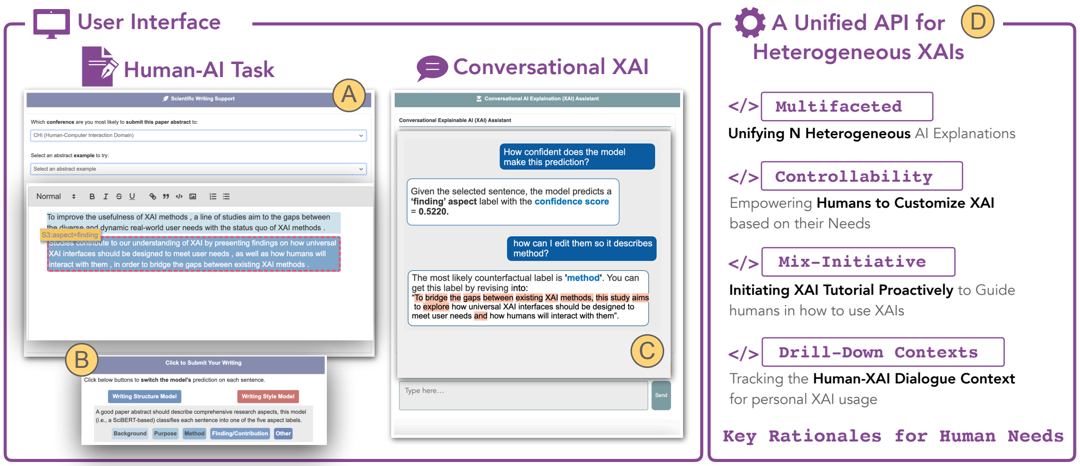

Fig. 1. An overview of ConvXAI to support human-AI scientific writing with heterogeneous AI explanations via dialog. ConvXAI includes a front-end User Interface to (A) support human-AI collaborative task interaction, (B) check AI models and predictions, and (C) inquire about heterogeneous AI explanations 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.

Despite a surge collection of XAI methods, users still struggle to obtain required AI explanations. Previous research suggests chatbots as dynamic solutions, but the effective design of conversational XAI agents for practical human needs remains under-explored. This paper focuses on Conversational XAI for AI-assisted scientific writing tasks. Drawing from human linguistic theories and formative studies, we identify four design rationales: "multifaceted", "controllability", "mix-initiative", "context-aware drill-down". We incorporate them into an interactive prototype, ConvXAI[1], which facilitates heterogeneous AI explanations for scientific writing through dialogue. In two studies with 21 users, ConvXAI outperforms a GUI-based baseline on improving human-perceived understanding and writing improvement. The paper further discusses the practical human usage patterns in interacting with ConvXAI for scientific co-writing.

1 Introduction

Despite the potential of a surge collection of eXplainable AI (XAI) methods, a number of studies show that applying state-of-the-art XAI methods to real-world human tasks could not always help users better simulate model predictions, understand AI model mistakes, etc [2, 8, 11]. To resolve the issues, researchers have explored the mismatch between real-world user demands and existing XAI methods. Shen and Huang [9], for instance, compare practical user questions 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 [5, 6, 13, 14]. Addressing this array of questions necessitates an integration of heterogeneous AI explanations. Prior work

[1] See the ConvXAI system code at: https://github.com/huashen218/convxai.git.

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 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.

has envisioned the concept of "explainability as a dialogue" to accommodate diverse user needs and mitigate cognitive load [5, 12]. 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 scientific writing, where we propose a conversational XAI system, ConvXAI. ConvXAI incorporates ten types of AI explanations into a unified dialog interface that empowers users to interactively ask various XAI questions about the AI predictions. Particularly, we augment ConvXAI with four design rationales collected from empirical formative studies with 7 users of diverse backgrounds and theoretic linguistic properties of human conversation: address various user questions ("multi-faceted"), actively provide XAI tutorials and suggestions ("mix-initiative"), empower users to dig into AI explanations ("context-aware drill-down"), and make flexible customization with details on-demand ("controllability").

We conducted two within-subject user studies with 21 users to compare with SelectXAI– the traditional GUI-based XAI system that displays all XAIs in a collapsible manner. Results show that users perceived ConvXAI to be more useful in understanding AI writing feedback and improving human writings. The results also validated the less cognitive load and effectiveness of the four user-oriented design principles. This work contributes insights into the design space of useful XAI in practice, reveals humans' XAI usage patterns and identifies opportunities for future XAI works.

2 ConvXAI System

ConvXAI is a novel system to facilitate human-AI collaboration by providing conversational AI explanations, which empowers users to request heterogeneous AI explanations via a unified dialog interface. We apply ConvXAI on human-AI scientific writing task in this study and briefly describe its design and implementation below[2] .

2.1 Delivering Heterogeneous Explanations via a Unified Dialogue Interface

Users ask for diverse XAI questions covering the whole AI lifecycle (*e.g.,* data, modeling, and evaluation) to understand the AI system [6, 9], while preferring the interactive interface to be concise for minimal cognitive load [7, 13]. Therefore, we design ConvXAI to deliver heterogeneous explanations via a unified conversational interface. Specifically, we develop this conversational XAI pipeline based on the task-oriented Dialogue-State Architecture [1], where it mainly consists of four modules. Firstly, a *Natural Language Understanding* module parses any XAI user question into a pre-defined XAI user intent (*e.g.,* "explain top-k features", "explain similar examples"), which is then mapped to a specific XAI function out of 10 XAI types elaborated in Table 1. Secondly, the *AI Explainers* module will generate AI explanations using cutting-edge XAI algorithms corresponding to the above XAI user intent. The third module, *Natural Language Generation*, then converts the generated explanation to a free-text response via the pre-defined natural language

[2]See a full paper of ConvXAI study at: https://arxiv.org/pdf/2305.09770.pdf.

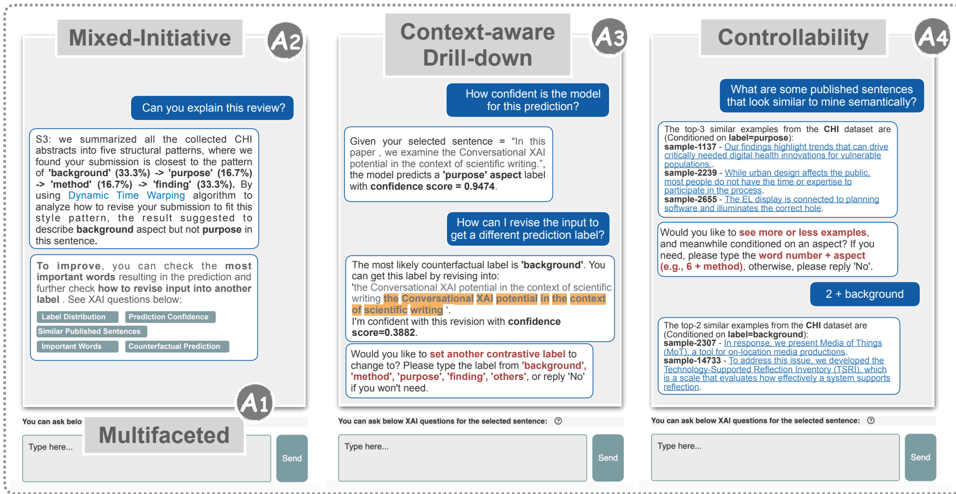

Fig. 2. An overview of four design rationales tailored for human use needs in ConvXAI. The ConvXAI dialogue flows are designed to follow the four principles of *"multifaceted"* (A_1), *"mixed-initiative"*(A_2), *"context-aware drill-down"*(A_3) and *"controllability"*(A_4).

templates, and sends the response back to users via the dialogue interface. 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. Overall, we design the conversational XAI pipeline to be model agnostic and XAI algorithm agnostic, which enables ConvXAI to be generalized to various AI and XAI methods.

2.2 Enabling User-oriented Multifaceted XAI with Interactive Customization

By combining the feedback from a formative study with 7 users of diverse backgrounds and the human conversational linguistic theories [3, 4], we embed four user-oriented design rationales into ConvXAI. As shown in Figure 2, the ConvXAI requires to be *Multifaceted*, *Mixed-initiative*, *Context-aware drill-down*, and *Controllability*.

Concretely, ConvXAI enables users to ask ten types of **multi-faceted** AI explanations (shown in Table 1) in the conversation input panel (Figure 2A_1). To design ConvXAI to be **mixed-initiative**, we start the explanation dialog with a review summary of the AI writing structure model and style model's outputs. The users can select any one sentence in this review list to ask instance-wise XAI questions and start a conversation session on the sentence. Uniquely, to make it serve as proactive guidance towards more sophisticated XAI methods, ConvXAI adds an additional explanation type, *understand suggestion* – to explain AI suggestions and provide brief XAI tutorials. Also, ConvXAI initiates a prompt message *"to improve..."* (Figure 2A_2) with a subset of relevant XAIs, based on "guessing" what users would want to improve their writing at this point. To enable **context-aware drill down** (Figure 2A_3), ConvXAI leverages Global XAI State Tracker to store the history of user and XAI agent dialog, and generate the XAI response based on the previous dialog accordingly. Still, given the default XAI responses may not satisfy users' needs on customizing their own explanations in some cases, we, therefore, make the XAI agent proactively present hints for human **controllablity**, *e.g.*, "would you like to..." (at the bottom of Figure 2A_3. Figure 2A_4) for humans to customize XAIs based on their needs. By embedding these user-oriented XAI rationales into ConvXAI design, the system can thus be more useful for humans in practical AI-assisted writing tasks.

2.3 Applying Conversational XAI to Human-AI Scientific Writing

We apply ConvXAI to human-AI collaborative writing on scientific papers, as the human-AI co-writing process consumes complex cognitive loads that can potentially inspire users to enquire more AI explanations. As shown in Figure 1,

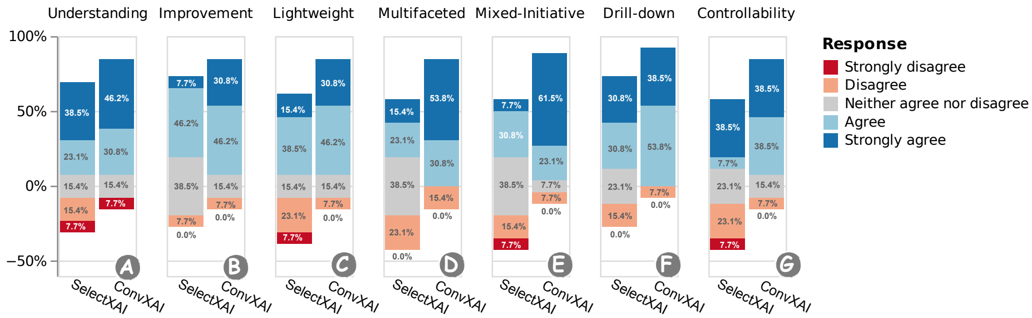

Fig. 3. Analyses on users' self-ratings on their experiences playing with ConvXAI and SelectXAI in Task One. They self-rated ConvXAI to be better on all dimensions, and most significantly on the usefulness of mix-initiative and multifaceted functionality.

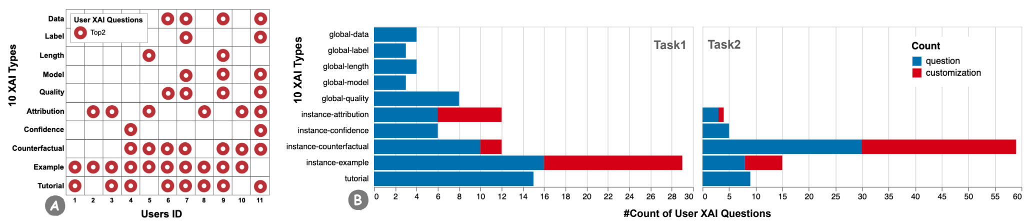

Fig. 4. User demands analysis during using ConvXAI to improve scientific writing in Task 1 and Task 2. Particularly, (1) We ranked the top-2 most frequently requested XAI methods by each user ID in Task 1(A). (2) We compute all the users' question amount for each of the 10 XAI methods in (B) Task 1 and Task 2.

humans can submit their drafts (Figure 1A) to the editor to get AI models' feedback (Figure 1B). Then they can leverage the conversational AI explanations to *understand the writing models' feedback* (Figure 1C) (*i.e.,* including AI model integrated feedback), and further *improve and resubmit their writings iteratively*. Here, we incorporate two AI models to provide AI predictions on writing structure and style, respectively, and further integrate them into writing reviews.

3 User Studies

We conducted two within-subjects human evaluation studies with 13 participants in the Task One and 8 participants rejoining in the Task Two. The users were recruited from university mailing list and required to have research writing experience. We asked users to compare ConvXAI against SelectXAI, a GUI-based XAI system that statically displays all the XAI formats at one-time. The user study aimed to investigate *if* ConvXAI *can help users better understand the AI writing feedback and further improve the writing artifacts accordingly*. We asked each participant to edit two paper abstracts with the help of ConvXAI and SelectXAI, respectively. Participants were then asked to rate their experience using 5-point Likert scale in the survey. We particularly designed the Task One to be an open-ended writing task to evaluate the effectiveness of user-oriented design in the system, and Task Two as a well-defined writing task to investigate how systems can help users improve their scientific writing process and output in practice [10].

We summarized participants' ratings in Task One on the two systems, ConvXAI and SelectXAI, in Figure 3. We performed the non-parametric Wilcoxon signed-rank test to compare users' nominal ratings. The results indicated that participants rated ConvXAI higher in terms of helping them understand the reasons behind the reviews they received for their writings (ConvXAI 4.07 ± 1.18 vs. SelectXAI 3.69 ± 1.37, $p = 0.036$, Figure 3A). They also felt that ConvXAI helped them more in improving their writing (4 ± 0.91 vs. 3.53 ± 0.77, $p = 0.019$, Figure 3B). Additionally, participants rated ConvXAI better in terms of the lighter cognitive load and four design aspects. Furthermore, by analyzing the user demands and usage patterns during the tasks (Figure 4), we found that different users prioritize different AI explanations and orders for their needs, their needs changed over time, and XAI customization is important for their needs.

References

[1] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. 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. https://openreview.net/forum?id=BJh6Ztuxl

[2] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[3] Ian Hutchby and Robin Wooffitt. 2008. *Conversation analysis*. Polity.

[4] Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

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

[6] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[7] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. (2018).

[8] Hua Shen and Ting-Hao Huang. 2020. 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*, Vol. 8. 168–172.

[9] Hua Shen and Ting-Hao'Kenneth' Huang. 2021. Explaining the Road Not Taken. *ACM CHI 2022 Workshop on Human-Centered Explainable AI* (2021).

[10] Hua Shen and Tongshuang Wu. 2023. Parachute: Evaluating interactive human-lm co-writing systems. *arXiv preprint arXiv:2303.06333* (2023).

[11] Hua Shen, Tongshuang Wu, Wenbo Guo, and Ting-Hao Huang. 2022. 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)*. Association for Computational Linguistics, Dublin, Ireland, 10–19. https://doi.org/10.18653/v1/2022.acl-short.2

[12] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations. (2022).

[13] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[14] Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. 2023. ScatterShot: Interactive In-context Example Curation for Text Transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 353–367.