Looking Through the Deep Glasses: How Large Language Models Enhance Explainability of Deep Learning Models

Philipp Spitzer
philipp.spitzer@kit.com
Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

Sebastian Celis
office@ksri.kit.de
Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

Dominik Martin
dominik.martin@partner.kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

Niklas Kühl
kuehl@uni-bayreuth.de
University of Bayreuth
Bayreuth, Bavaria, Germany

Gerhard Satzger
gerhard.satzger@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Baden-Württemberg
Germany

ABSTRACT

As AI becomes more powerful, it also becomes more complex. Traditionally, eXplainable AI (XAI) is used to make these models more transparent and interpretable to decision-makers. However, research shows that decision-makers can lack the ability to properly interpret XAI techniques. Large language models (LLMs) offer a solution to this challenge by providing natural language text in combination with XAI techniques to provide more understandable explanations. However, previous work has only explored this approach for inherently interpretable models–an understanding of how LLMs can assist decision-makers when using deep learning models is lacking. To fill this gap, we investigate how different augmentation strategies of LLMs assist decision-makers in interacting with deep learning models. We evaluate the satisfaction and preferences of decision-makers through a user study. Overall, our results provide first insights into how LLMs support decision-makers in interacting with deep learning models and open future avenues to continue this endeavor.

CCS CONCEPTS

• **Human-centered computing → Collaborative interaction**; **Empirical studies in HCI**; • **Computing methodologies → Natural language generation**.

KEYWORDS

Large Language Models, Artificial Intelligence, Explainable AI, Human-Computer Interaction

ACM Reference Format:
Philipp Spitzer, Sebastian Celis, Dominik Martin, Niklas Kühl, and Gerhard Satzger. 2024. Looking Through the Deep Glasses: How Large Language Models Enhance Explainability of Deep Learning Models. In *Proceedings of Mensch und Computer 2024 (MuC '24), September 01–04, 2024, Karlsruhe, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3670653.3677488

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License.

MuC '24, September 01–04, 2024, Karlsruhe, Germany
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0998-2/24/09
https://doi.org/10.1145/3670653.3677488

1 INTRODUCTION

The recent rise of artificial intelligence (AI) has revolutionized the way data is used and interpreted to make decisions, profoundly impacting industries from various sectors [5, 28]. Despite these advancements, one of the major challenges of using AI models is that they are often difficult to interpret. This makes it difficult for decision-makers to appropriately rely on their output [23] and integrate them into everyday processes.

Past research shows that making AI models more interpretable and transparent can help their adoption [9, 18, 26]. As highlighted by Samek et al. [21] and Felzmann et al. [6], eXplainable AI (XAI) is crucial for providing decision-makers with the necessary tools and techniques required to understand the underlying mechanisms of the systems. However, interpreting these techniques correctly remains a challenge [24]. Past works reveal that explanations can sometimes deceive decision-makers [17], resulting in automation bias and blind reliance on AI even though their output is incorrect [24].

A persistent issue within the field of AI is the trade-off between model performance and interpretability [4]. Oftentimes, AI development prioritizes performance over interpretability. However, recent research suggests that this trade-off is not necessarily inevitable. By leveraging large language models (LLMs), it is possible to enhance interpretability without significantly compromising performance. These studies explore using natural language explanations for model predictions to make AI more transparent and user-friendly. For instance, Werner [31] attempts to provide a rule-based interactive chatbot drawing from information gained with common XAI methods. Slack et al. [27] present an interactive chat system employing a language model to analyze user input and parse it into a predefined set of actions.

Despite this paradigm shift towards using LLMs to improve AI interpretability, little is known about how LLMs support the interaction between humans and AI from the decision-maker's perspective. The few studies that exist like Slack et al. [27] focus on inherently interpretable "white-box" models. To expand our understanding of how LLMs can improve the interpretability of white-box models and "black-box" models (e.g., deep learning models) in the field of human-computer interaction (HCI), we propose the following research question:

RQ: How can LLMs facilitate the interaction between decision-makers and deep learning models within an interactive environment?

To address our research question and advance the understanding in HCI on the use of LLMs as facilitators for decision-makers, we implement state-of-the-art augmentation strategies for LLMs: retrieval-augmented and context-augmented strategies. Retrieval-augmented strategies involve fetching relevant information in real-time to generate explanations whereas context-augmented strategies integrate relevant information directly into the model's context before generating explanations. We explore the use of these strategies through a user study involving two user groups: domain experts and AI experts. In our study, we compare their preference and satisfaction during interactions with the LLMs, and we also analyze the computational efficiency of each strategy by measuring the token usage during explanation generation. Our results indicate that context-augmented generation (CAG) was preferred and deemed more satisfactory. However, retrieval-augmented generation (RAG) was computationally cheaper, presenting an alternative to CAG when costs are a factor to be considered.

Overall, our contributions to the HCI field are two-fold: First, we outline the design of how LLMs can be utilized as interactive explanation methods for deep learning models. Second, we extend the knowledge in HCI on the use of augmentation strategies by comparing state-of-the-art methods and evaluating them in an empirical study. Our findings provide guidelines for designers of AI systems on how to use LLMs as a means of explanation.

2 RELATED WORK

The field of XAI has gained considerable traction and has been extensively researched, specifically addressing the need to explain, interpret, and make AI models more transparent [21]. XAI can significantly enhance the collaboration between humans and AI by improving interpretability, trust, and usability through various types of explanations [6]. A key challenge in the use of XAI revolves around the trade-off between model performance and interpretability. The utilization of more interpretable models often comes at the cost of accuracy [4, 16]. For this reason, it is crucial to ensure high explainability within accurate AI models. This provides optimal performance and establishes a high level of trust and reliability for end-users [10]. Given the significance of providing transparency for AI models, this section delineates the potential of LLMs as a means of explanation and the extent to which previous research has explored their utility in this context.

In the domain of XAI, the incorporation of LLMs is an emerging strategy to enhance the understanding and interpretation of decisions made by AI models [3, 13, 30]. One focal point of research is generating natural language to provide specific explanations on model predictions, often highlighting what components of the inputs lead to the decision. By focusing more on the model outputs for the explanation, both Monje et al. [15] and Alonso et al. [2] utilize fuzzy linguistic modules for natural language interpretations of model outputs. The former extracts linguistic rules from the data, while the latter integrates rule-based language generation for more coherent textual explanations. Instead of using the outputs as the explanation source, Kaczmarek-Majer et al. [12] generates template-based fuzzy linguistic summaries of the Shapley analysis run on the model to provide more detailed insights on the model's decision.

In the context of interactive explanations, in which the system is capable of adapting to multiple back-and-forth questions, several works have been able to leverage the generalization capabilities of LLMs to provide continuous explanations in a chat-like format. Werner [31] attempts to provide a rule-based interactive chatbot drawing from information gained with common XAI methods. In a different approach, Yang et al. [32] merge LLMs with an extracted tabular analysis of advertisements from brand campaigns to provide interpretations on model decisions. Similarly, Gao et al. [7] extend a recommendation tool by integrating ChatGPT for personalized and history-based suggestions, while leveraging the recommendation history and in-context learning from LLMs to explain the recommendations. Slack et al. [27] present an interactive chat system employing a LLM to analyze user input and parse it into a predefined set of actions. The actions are executed to gather additional information, trying to improve the interpretability of the model.

While several works analyze the techniques that explain AI models, evaluating the implementation of these approaches empirically and in the interaction with actual users is a crucial task to facilitate the adoption of this approach. Similar to Aechtner et al. [1] who focus on comparing different XAI techniques against each other or Rong et al. [20], Silva et al. [25], Tekkesinoglu [29] which explore the objective and subjective understanding of various methods, rigorous research is necessary to explore the use of LLMs to explain deep learning models. Despite the contribution LLMs make to the XAI field through natural language generation, the recent development leap of LLMs has significantly expanded research potential to enhance explainability. Existing research mostly concentrates on providing summarized model predictions to the user in a non-interactive format. Only a few studies (e.g., Slack et al. [27]) investigate the use of LLMs to explain white-box models in an interactive environment. Deep learning models have not been investigated yet, presenting a gap in research. To address this research gap, this work investigates different augmentation-generation strategies for LLMs to explain deep learning models in a first empirical study.

3 METHODOLOGY

To explore different augmentation strategies for LLMs, we conducted a user study in which we had participants interact with LLMs that were used to improve the interpretability of a deep learning model. The deep learning model predicted cost trends over time.

The online study was divided into four parts. In the first part, participants gave their consent for participation, were introduced to the study, and answered demographic questions. Part two comprised a tutorial for the task and the usage of the system. Participants were explained that they would interact with a deep learning model and were shown the basic functionality based on an example. After the tutorial, participants started the actual task in the third part of the study. They were given a prediction of the deep learning model and could interact with the LLM to retrieve further information, such as the domain of the task or to enhance their understanding of how the deep learning model worked. For each prediction, the participants had to interact with the LLMs for at least five different questions. For each question, both LLMs provided an answer. One LLM used a CAG method and one LLM used a RAG method. The participants could either select pre-defined questions or type in their own questions. After each question, both LLMs generated a response side-by-side. To prevent bias in the selection of the first generated response, the display location of the LLM's output was randomized for each new question. In the final part of the study, participants provided general feedback on their experience and the study overall.

To assess the capability of LLMs to explain deep learning models, we utilized a cost prediction model. This model predicts the costs of a specific product from the polymer industry, given its history of costs. It outputs the predicted cost per unit for a given polymer-based product. We implemented CAG by following Li et al. [14], Yue et al. [33] and RAG by following Slack et al. [27]. To evaluate the use of LLMs as a means of explainability, we recruited participants from two different backgrounds applying a purposeful sampling approach [19]: domain experts and AI experts. Overall, there were five participants with a background in AI and six domain experts. The participants were recruited within an industrial company that had deployed the AI system.

ChatGPT The AI model selected in this work was a long short-term memory model. This model is a variant of a recurrent neural network, which is specialized in sequence and temporal data processing due to its ability to maintain information over extended time intervals. The choice of this particular model was motivated by the task domain to predict the costs of a given polymer over time. On top of the AI model, we implemented two LLMs with different augmentation strategies: CAG and RAG. CAG is a technique where a prepared prompt based on a template is created and injected into the LLM as prior context [14]. RAG is an alternative method that retrieves additional context depending on the user's query [8]. As the LLM, we selected *"gpt-3.5-turbo-1106"*[1]. It contains both a sufficiently large context window (roughly 16,000 tokens) and provides a function-calling interface. This foundation model was augmented with additional context based on the two content retrieval strategies. The CAG model received data from a prior XAI analysis injected as context. The RAG model took advantage of function-calling to access the data from the XAI analysis. The foundation model had been trained to detect when a function should be called and what inputs to provide to it while adhering to a predefined schema. Based on the XAI analysis generated for the black box model, function definitions were outlined for the RAG model to call.

After each question for which participants interacted with the LLM, they had to give their preference for which of the two answers provided by both LLMs they preferred and rate how satisfied they were with the answer. The response preference was designed to understand which of the two methods yielded better and preferred explanations to the user. Additionally, we measured participants' satisfaction with each output they received. The satisfaction levels of the user for each response provided a more detailed insight into the perceived quality of explanations. Satisfaction is particularly beneficial when both explanations are very similar, in which case the user does not have a direct preference but chooses one. In addition, satisfaction reveals explanations of poor quality, which could not directly be discovered through preference selection alone. We measured the satisfaction for each response of the LLM through a slider that ranged from "very unsatisfied" in five steps to "very satisfied" (see Appendix A Figure 3.

4 RESULTS

The study was conducted in February 2024. In total, eleven participants took part in the study. On average, the participants completed the study in about 28 minutes. To evaluate both augmentation strategies, we first analyze the preference scores of both models. Comparing both augmentation strategies with each other, the CAG model was preferred slightly more than the RAG model: 5% of the answers generated by the CAG model was preferred over 46% of the answers generated by the RAG model. We illustrate the findings in Figure 1. In terms of satisfaction, however, both models appear to be roughly equal, with the average satisfaction at 66.6% for the RAG model and 66.4% for the CAG model (we transformed the used Likert scale into numeric values: 1 - very unsatisfied, 5 - very satisfied). Notably, the CAG model had relatively polarized satisfaction—more very positive and very negative satisfaction selections—whereas the RAG model was more centered around neutral satisfaction results.

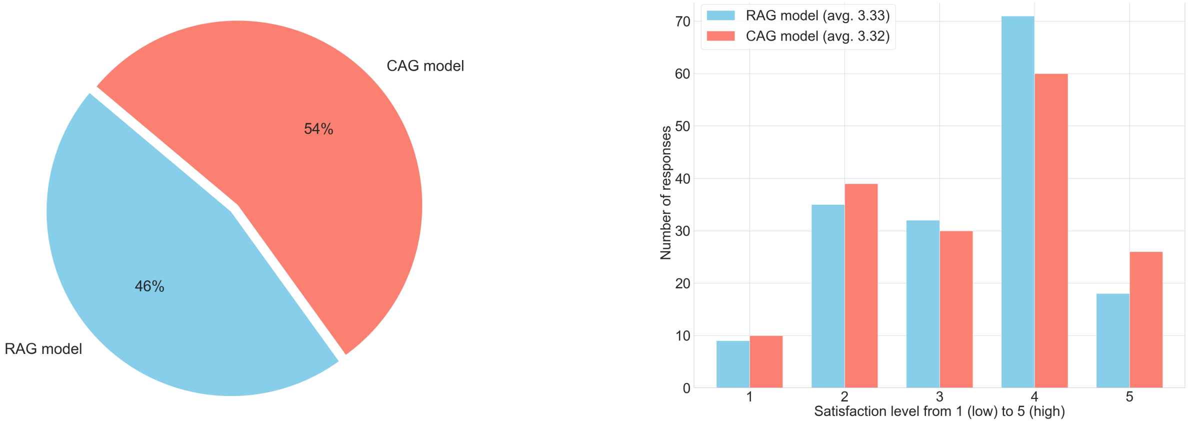
Based on the two different user groups in the study, we derived further insights. The user group of AI experts is largely related to departments in the domains of AI, machine learning, and data science (five participants). The other user group, domain experts, work mostly in sales, management, and other business areas (six participants). Overall, there is a discrepancy in satisfaction preference between the two groups. On the one hand, the AI experts slightly preferred the RAG model (70%) over the CAG model (66%). On the other hand, for the domain experts, the data was reversed: domain experts preferred the CAG model (68%) over the RAG model (64%).

Both LLMs convert text into tokens through a tokenization process. These tokens only include the input context each language model uses to generate the responses. Higher token usage corresponds to higher usage costs. The CAG model used roughly 4,200 input tokens on average, whereas the RAG model used approximately 1,700 tokens on average. This discrepancy is due to the fact that the CAG model had XAI data available prior to each response, while the RAG model only retrieved partial extracts of it. The CAG model has less deviation around the average than the RAG model. Overall, all responses of the CAG model require higher token usage than the responses of the RAG model.

5 DISCUSSION

This work explores the use of LLMs to explain deep learning models to decision-makers. By investigating different augmentation strategies for LLMs to facilitate decision-making, we take the first steps towards answering our RQ and revealing LLMs' capability to

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo



(a) Preference results.

(b) Satisfaction results.

Figure 1: Results of the overall user preference and satisfaction level between both LLMs with retrieval-augmentation generation in red and context-augmentation generation in blue.

facilitate the interaction between decision-makers and deep learning models. While more research is necessary to answer the research question entirely, we find first empirical evidence for the support LLMs can provide. Moreover, we compare CAG against RAG throughout a user study with two different user groups: domain experts and AI experts. The data of our study suggests that the CAG model is overall slightly more preferred and satisfactory, while domain experts prefer the CAG model and AI experts the RAG model. This could mean that explanations by the RAG model reflect the underlying rationale of the AI model in a more detailed way which is easier to interpret for AI experts, whereas the explanations provided by the CAG model are more extensive and complete, giving the domain experts a better overview of all the information. The findings suggest that the explainability of AI models can be enhanced through the use of interactive natural language explanations for different user groups. While past work focuses on utilizing LLMs to enhance the explainability of white-box AI models [27], our work expands these findings by providing an interactive explanation system for black-box AI models. The application potential of these findings is therefore not constrained by the architecture of the model, making the proposed methodology an important tool to improve the transparency of AI models in practice.

Building on existing evidence that knowledge-augmented models perform better in classification and reasoning tasks [11], this work provides novel insights into the strengths and weaknesses of both RAG and CAG models. The effectiveness of each strategy is empirically evaluated in a user study and the resulting data aligns with the theory that knowledge-augmented models also provide more contextually relevant and accurate explanations. To further explore how LLMs can support decision-makers in interacting with LLMs and to gain deeper insights, future work could verify satisfaction levels and our findings in field experiments with users through interviews or think-aloud studies. This could reveal further insights and reasoning.

Certainly, this work also comes with limitations. Since there are no ground-truth explanations available, the evaluation scope of this work is limited to subjective feedback through a user study. Without ground-truth explanations, it is difficult to confirm the factual correctness of the provided explanations. Future research could advance this line of research and compare the factual correctness of generated explanations. Furthermore, the outputs of both LLMs were displayed next to each other. To mitigate any biases through this design choice, future research could make use of an in-between study design.

Another focal point for future research is to investigate the trust and information coverage provided by the LLMs to make deep learning models more transparent. As our study builds the foundation for further exploration, future research can focus on measuring the decision-making behavior in terms of appropriate reliance [22, 23].

Through the comparison of two augmentation strategies, this work provides a deeper understanding of generating understandable explanations with LLMs. Further research is necessary to broaden the knowledge of using LLMs as a means of explanation system to support decision-makers. We invite other researchers to follow up on this line of research and actively participate in this discourse.

REFERENCES

[1] Jonathan Aechtner, Lena Cabrera, Dennis Katwal, Pierre Onghena, Diego Penroz Valenzuela, and Anna Wilbik. 2022. Comparing User Perception of Explanations Developed with XAI Methods. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–7. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882743

[2] Jose M. Alonso, A. Ramos-Soto, Ehud Reiter, and Kees van Deemter. 2017. An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015489

[3] Brendan Alvey, Derek Anderson, James Keller, and Andrew Buck. 2023. Linguistic Explanations of Black Box Deep Learning Detectors on Simulated Aerial Drone Imagery. *Sensors* 23, 15 (Aug 2023), 6879. https://doi.org/10.3390/s23156879

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] D. Carter. 2018. How real is the impact of artificial intelligence? The business information survey 2018. *Business Information Review* 35, 3 (2018), 99–115. https:

//doi.org/10.1177/0266382118790150

[6] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics* 26 (2020), 3333–3361. Issue 6. https://doi.org/10.1007/s11948-020-00276-4

[7] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. arXiv:2303.14524 [cs.IR]

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL]

[9] Johannes Hangl, Viktoria Joy Behrens, and Simon Krause. 2022. Barriers, Drivers, and Social Considerations for AI Adoption in Supply Chain Management: A Tertiary Study. *Logistics* 6, 3 (2022). https://www.mdpi.com/2305-6290/6/3/63

[10] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation* 16, 1 (January 2024), 45–74. https://doi.org/10.1007/s12559-023-10179-8

[11] Pedram Hosseini, David A. Broniatowski, and Mona Diab. 2022. Knowledge-Augmented Language Models for Cause-Effect Relation Classification. arXiv:2112.08615 [cs.CL]

[12] Katarzyna Kaczmarek-Majer, Gabriella Casalino, Giovanna Castellano, Monika Dominiak, Olgierd Hryniewicz, Olga Kamińska, Gennaro Vessio, and Natalia Díaz-Rodríguez. 2022. PLENARY: Explaining Black-Box Models in Natural Language through Fuzzy Linguistic Summaries. *Inf. Sci.* 614, C (oct 2022), 374–399. https://doi.org/10.1016/j.ins.2022.10.010

[13] Benjamin J. Lengerich, Sebastian Bordt, Harsha Nori, Mark E. Nunnally, Yin Aphinyanaphongs, Manolis Kellis, and Rich Caruana. 2023. LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs. arXiv:2308.01157 [stat.ML]

[14] Zonglin Li, Ruiqi Guo, and Sanjiv Kumar. 2022. Decoupled context processing for context augmented language modeling. *Advances in Neural Information Processing Systems* 35 (2022), 21698–21710.

[15] Leticia Monje, Ramón Alberto Carrasco, Carlos Rosado Moral, and Manuel Sánchez-Montañés. 2022. Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain. *Mathematics* (2022). https://api.semanticscholar.org/CorpusID:248383794

[16] W. R. Monteiro and G. Reynoso-Meza. 2022. A review of the convergence between explainable artificial intelligence and multi-objective optimization. (2022). https://doi.org/10.36227/techrxiv.21707261

[17] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–39.

[18] Claire Nicodeme. 2020. Build confidence and acceptance of AI-based decision support systems - Explainable and liable AI. In *2020 13th International Conference on Human System Interaction (HSI)*. 20–23. https://doi.org/10.1109/HSI49210.2020.9142668

[19] Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health Research* 42, 5 (2015), 533–544.

[20] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), 1–20. https://doi.org/10.1109/TPAMI.2023.3331846

[21] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller. 2019. Explainable ai: interpreting, explaining and visualizing deep learning. *Lecture Notes in Computer Science* (2019). https://doi.org/10.1007/978-3-030-28954-6

[22] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. 2023. Towards effective human-ai decision-making: The role of human learning in appropriate reliance on ai advice. (2023).

[23] Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[24] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the influence of explainable AI on automation bias. *arXiv preprint arXiv:2204.08859* (2022).

[25] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2023. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1390–1404.

[26] Rishi P. Singh, Grant L. Hom, Michael D. Abramoff, J. Peter Campbell, Michael F. Chiang, and on behalf of the AAO Task Force on Artificial Intelligence. 2020. Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. *Translational Vision Science & Technology* 9, 2 (08 2020), 45–45. https://doi.org/10.1167/tvst.9.2.45 arXiv:https://arvojournals.org/arvo/content_public/journal/tvst/938366/i2164-2591-9-2-45_1597145394.90976.pdf

[27] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations. arXiv:2207.04154 [cs.LG]

[28] Philipp Spitzer, Niklas Kühl, Marc Goutier, Manuel Kaschura, and Gerhard Satzger. 2024. Transferring Domain Knowledge with (X) AI-Based Learning Systems. (2024).

[29] Sule Tekkesinoglu. 2023. Exploring Evaluation Methodologies for Explainable AI: Guidelines for Objective and Subjective Assessment. (17 12 2023). https://doi.org/10.2139/ssrn.4667052 arXiv:4667052

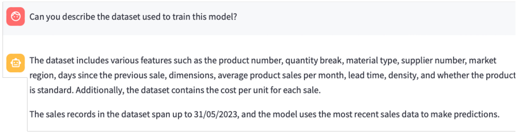
[30] Osman Tursun, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2023. Towards Self-Explainability of Deep Neural Networks with Heatmap Captioning and Large-Language Models. arXiv:2304.02202 [cs.CV]

[31] Christiane Werner. 2020. Explainable AI through Rule-based Interactive Conversation. In *EDBT/ICDT Workshops*. https://api.semanticscholar.org/CorpusID:214765218

[32] Qi Yang, Marlo Ongpin, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. 2023. Against Opacity: Explainable AI and Large Language Models for Effective Digital Advertising. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 9299–9305. https://doi.org/10.1145/3581783.3612817

[33] Thomas Yue, David Au, Chi Chung Au, and Kwan Yuen Iu. 2023. Democratizing financial knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology. *Available at SSRN 4346152* (2023).

A APPENDIX

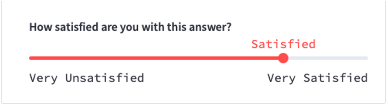


Can you describe the dataset used to train this model?

The dataset includes various features such as the product number, quantity break, material type, supplier number, market region, days since the previous sale, dimensions, average product sales per month, lead time, density, and whether the product is standard. Additionally, the dataset contains the cost per unit for each sale.

The sales records in the dataset span up to 31/05/2023, and the model uses the most recent sales data to make predictions.

Figure 2: An example of the interaction between the AI and the user.



How satisfied are you with this answer?

Satisfied

Figure 3: The questionnaire used to measure satisfaction on a five-point scale.