Conversational Data Stories

Valentin Grimm
valentin.grimm@th-owl.de
OWL University of Applied Sciences and Arts
Höxter, Germany

Jessica Rubart
jessica.rubart@th-owl.de
OWL University of Applied Sciences and Arts
Höxter, Germany

Patrick Söhlke
ps@nextvision.info
Next Vision GmbH
Hessisch Oldendorf, Germany

ABSTRACT

Data stories are about revealing and communicating insights from complex data. In this paper, we propose *conversational data stories*, which support end users in understanding the key findings of the data analysis at hand by natural language conversation. Creating these stories manually means to put a lot of effort into understanding the data and crafting visuals. With increasingly powerful generative large language models (LLMs), natural language processing as well as automating the creation of data stories is a promising field. We present a concept for a conversational data storytelling system that integrates LLMs as well as explainable AI. We present the collected requirements for our system concept and how the requirements are addressed. To show the potential of our approach, we provide a use case scenario and a discussion in this paper. This is supposed to serve as a basis for future research that will aim at investigating the technical reliability and the user experience of such a system.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; **Natural language interfaces**; **Visual analytics**; **Visualization theory, concepts and paradigms**.

KEYWORDS

Data Storytelling, Conversational Assistant, Conversational Data Storytelling, Explainable AI

ACM Reference Format:
Valentin Grimm, Jessica Rubart, and Patrick Söhlke. 2024. Conversational Data Stories. In *7th Workshop on Human Factors in Hypertext (HUMAN '24), September 10–13, 2024, Poznan, Poland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3679058.3688631

1 INTRODUCTION

Data storytelling transforms numbers and data points into a cohesive narrative [17]. Conventional visualizations using bar charts, pie charts, or scatter plots, for example, are enhanced to present the main findings of the data analysis at hand to a special target group. These enhancements include removing redundant elements, adding explainable annotations, emphasizing important visual elements, and guiding the users through explicit titles.

In our current research and development project with the company Next Vision GmbH we integrate eXplainable Artificial Intelligence (XAI) as well as Generative AI (GenAI) in data stories. With this integration, we turn data stories into conversational systems, which help end users to understand the key findings of the data analysis and which support end users in making inquiries regarding the analysis problem at hand. Resulting visualizations are linked to parts of the conversation. To this end, from the viewpoint of our collaborative project we define a conversational data story with the following characteristics:

- Integration of suitable XAI concepts in order to improve transparency, acceptance, and trust of the underlying ML model or AI approach
- Integration of GenAI, in particular large language models (LLMs), to turn data stories into conversational systems
- Linking the different media, e.g. text and visuals, to achieve a coherent story and improve explainability

We have developed a system concept with our partner and started to implement the system. A prototypical example is shown in fig. 1. It depicts how users make textual requests and receive answers from a conversational assistant. The answers are explicitly linked with data visualizations through annotations. The main technical innovation of the system comes from the interaction between large language models and visualizations to provide meaningful text and visual responses for the users that are in symbiosis.

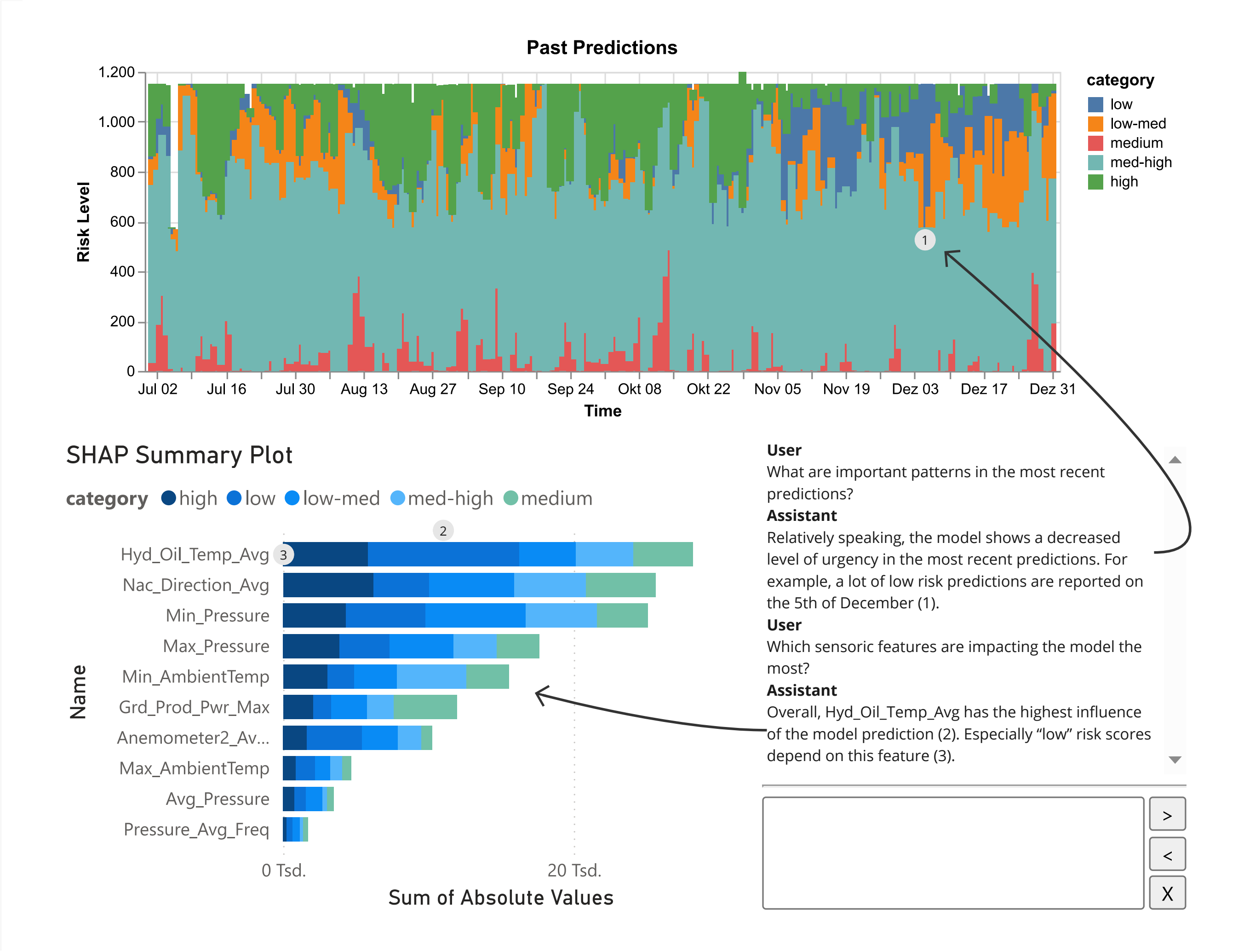
In this work, we present the requirements for our conversational data storytelling system and how we address these requirements in the implementation. Moreover, we showcase a use case scenario as a proof of concept and provide a discussion and critical assessment of our concept.

2 RELATED WORK

Echeverria et al. explore the effectiveness of different data storytelling elements in the learning context [5]. They experience encouraging first results, e.g that teachers perceived prescriptive titles, text labels and shaded areas as data storytelling elements adding contextual information. Shao et al. have shown that data stories improve the efficiency and effectiveness of comprehension tasks compared with conventional visualizations [17]. The data story design solution of Mosconi et al. [11] focuses on curating and sharing heterogeneous data sources; data storytelling is used to facilitate contextualization.

Renda et al. present Melody (*Make mE a Linked Open Data storY*), a platform for data storytelling using Linked Open Data and the SPARQL query language to visualize query results as web-based data stories [13]. Prior knowledge about SPARQL, a query language for data stored in the Resource Description Language (RDF) format, is required. Conversational data stories, the approach presented

HUMAN '24, September 10–13, 2024, Poznan, Poland
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1120-6/24/09
https://doi.org/10.1145/3679058.3688631



User
What are important patterns in the most recent predictions?
Assistant
Relatively speaking, the model shows a decreased level of urgency in the most recent predictions. For example, a lot of low risk predictions are reported on the 5th of December (1).
User
Which sensoric features are impacting the model the most?
Assistant
Overall, Hyd_Oil_Temp_Avg has the highest influence of the model prediction (2). Especially "low" risk scores depend on this feature (3).

Figure 1: An Example of a Conversational Data Story

in this paper, target end users by integrating natural language processing (NLP) rather than a query language.

Atzenbeck et al. integrate an LLM in the spatial hypertext system *Mother* through suggestion nodes [3]. These are intended to stimulate user creativity in the context of storytelling. Data stories are not in the focus of this work.

He et al. [7] propose a narrative visualization pipeline that showcases the potential of LLMs in the creation of data stories. Most notably, they present the narration step that discusses the relevance of organization of elements such as textual and graphical elements (logic) and the arrangement of the story, e.g. linear or multi-layered (structure). This paper builds upon these ideas by proposing an innovative way to combine the logical arrangement and a multi-layered narrative.

This work builds upon previous work on a digital boardroom [16] and an industrial control room [15], which use visual components and multimodal interaction techniques [14] in collaborative or mixed reality based dashboards. Another previous work presents an approach to provide interactive XAI visualizations by combining different XAI methodologies and incorporating gamification [6]. These approaches can benefit from conversational data stories.

3 SYSTEM SUPPORT FOR CONVERSATIONAL DATA STORIES

We have developed our approach in the context of a specific business use case of analyzing a prediction based on machine learning (ML). For example, if an ML model provides insights into the maintenance status of a wind turbine, a maintenance worker might want to look deeper into the decision process for unusual predictions. This includes, among others, to look into unusual data patterns, to analyze the overall and specific model performance and to observe similar predictions of the model.

In practice, it is usually necessary to define a clear procedure (cf. [10]). We have defined a four step process for building trust towards a specific model prediction. This process is concerned with analyzing the underlying data (1), analyzing the model (2), understanding the model prediction (3) and analyzing the prediction context (4).

Our concept is generalizable for multi-step data analysis procedures (e.g. looking into business data with a macro vs. micro perspective). The system is intended to categorize user requests based on the presented steps to adapt the User Interface (UI) for the specific step, respectively.

One important challenge in the context of data visualization and, more specifically in explainable AI, is to provide information not just mathematically correct, but also in a way that is human-friendly [9]. For example, Abdul et al. [1] have investigated into the trade-off between the reduction of cognitive load and reducing the accuracy of information. They found that the evaluation performance could improve even though the information contained in the visualization was less because the cognitive load was reduced.

Among others, conversational systems are discussed as a promising approach to increase performance in XAI applications [9]. This is grounded in the idea that humans are used to explanations in a conversational fashion as a series of inquiries and responses.

3.1 Requirements

Therefore, we developed requirements for a conversational system that incorporates data storytelling and evaluating trust towards a model prediction:

(1) Provide textual insights and expert knowledge based on the data at hand
(2) Visualize the utilized data in a meaningful way
(3) Explicitly link the aspects of the visuals that are directly mentioned in the textual response
(4) Interact with the visuals
(5) Stay in the defined framework of categories (e.g. data, model, prediction, context)
(6) Provide meaningful guidance based on the request history

3.2 System Design

In the following, we describe how all of the six requirements are addressed. Fig. 2 depicts the internal process of the system and the work of the large language model for an example use case scenario. Currently, we consider the use of foundation models (i.e. general purpose LLMs) that are not specifically fine-tuned for the domain and guided by sophisticated and parameterizable prompts. In an industrial setting with confidential data an on-premise solution with open-source LLMs, such as LLAMA3, can be used.

The system properties are described in the same order as the listed requirements.

(1) The user request is processed by an LLM with a prompt that is enriched by information about the available data and parameters to narrow down the data. The LLM returns a structured response that contains machine-readable instructions to fetch the data. The general knowledge of the LLM, a guiding prompt and the fetched data is combined to create a natural language response from the LLM (steps 1 - 4)
(2) The machine-readable instruction to fetch the data (step 6) is also used to choose and populate a fitting visual by the LLM. For this, a library of visual components is predefined in the system.
(3) The visuals from the custom library also support parametrization for annotations, which also can be used for explicit linking. Based on the textual response (step 7), the parameters are derived with the help of the LLM and fed into the custom visual together with the information from step 6.
(4) The custom visuals are shipped with interactive elements. This may include hover functionality and filters. E.g., enable updating the time frame or get details about a positional element.
(5) The incoming user request is categorized in step 1 and results in the creation or update of a category-specific thread/page for the newly rendered visuals. Also, a guidance LLM is deployed (see next bullet point).
(6) Additionally to the LLMs that are concerned with the individual categories (based on prompts), two specialized LLMs are integrated. One is responsible for answering general questions that are related to ML, e.g. "How do SHAP values work?". The other is concerned with user guidance. If the user is unsure about how to proceed or asks questions that are out of the story context, the guidance LLM will be activated.

3.3 Use Case Scenario

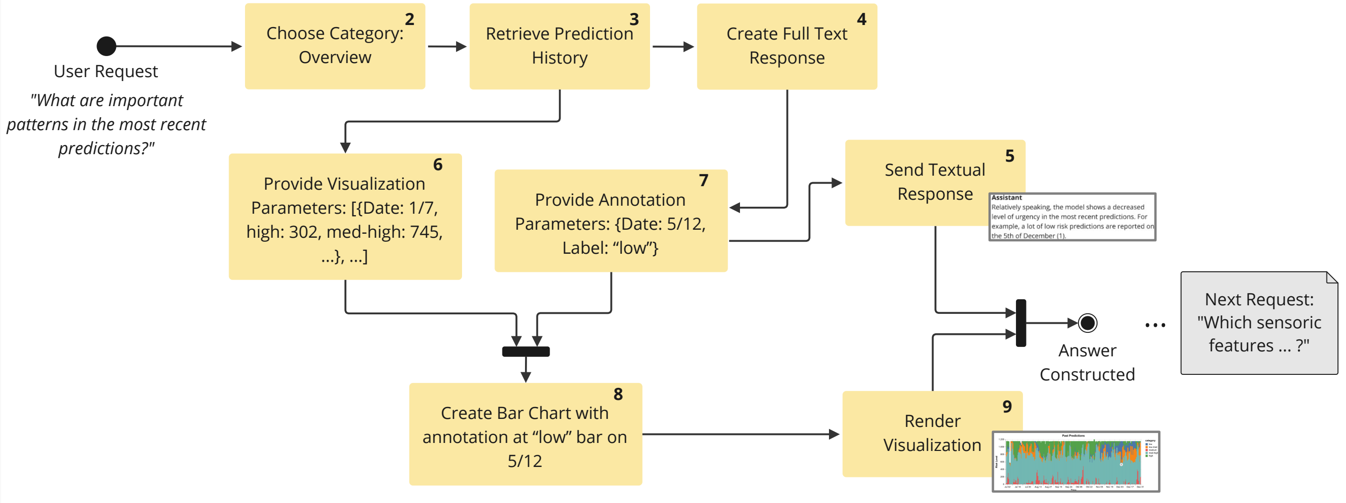
The UI with an example highlighting the implementation of requirements 1, 2, 3 and 4 is shown in fig. 1.

The user is interacting with the assistance through the chat window on the bottom right. Based on the first user request, the textual answer in the chat window is generated and the "Past Predictions" timeline is generated. The visual includes an annotation at the "low" bar on the 5th of December because it is mentioned in the textual answer of the LLM.

The second request dives deeper into the decision process of the model and asks for the most impactful features. The assistance's textual response is again accompanied by a fitting visual that shows the ordered importance of features based on SHAP values. The assistance mentions that the highest importance is given to the Hyd_Oil_Temp_Avg and especially "low" predictions are impacted. These insights are also highlighted in the SHAP importance visual.

If the user asks a guiding question, such as "What else should I explore from the high level perspective of the model?", the response will not trigger the creation of new visual components, but provide a recommendation based on the available data. E.g. "Explore the model performance based on the test data and compare them with the training data performance to explore overfitting of the model."

After exploring the high level perspective, the user can make a request about the specific prediction from the 5th of December and ask for information about this, such as "How did feature values impact that specific prediction?". This will trigger a recategorization and, as there is no thread/page that contains visuals about that category ("prediction"), a new thread/page will be created with the textual answer and first visuals.


User Request
"What are important patterns in the most recent predictions?"
Choose Category: Overview 2
Retrieve Prediction History 3
Create Full Text Response 4
Send Textual Response 5
Provide Visualization Parameters: [{Date: 1/7, high: 302, med-high: 745, ...}, ...] 6
Provide Annotation Parameters: {Date: 5/12, Label: "low"} 7
Create Bar Chart with annotation at "low" bar on 5/12 8
Render Visualization 9
Answer Constructed
...
Next Request: "Which sensoric features ... ?"

Assistant
Relatively speaking, the model shows a decreased level of urgency in the most recent predictions. For example, a lot of low risk predictions are reported on the 5th of December (1).

Figure 2: Examplary process of the LLM-driven architecture to handle a user request.

4 DISCUSSION

Conversational data stories fit well into the understanding of *augmented analytics*, a term coined by Gartner [12] that points out the integration of ML and NLP in data analytics environments for business users to improve the quality and transparency of decisions. Our conversational data stories assist end users in interacting with the story and understanding the key findings.

To address digital responsibility [18] additional aspects should be considered in analytics use cases, such as consideration of fairness principles to make sure that the underlying training data is fairly biased or such as alignment with special target groups to make the data stories effective.

The presented system has been developed in the context of the presented use case. The system can be generalized to other data storytelling use cases. Categories, custom visuals and the underlying data are variation points in the system and can be adapted for each use case.

For now, we made the design decision to develop visualizations manually, because it decreases the load on the LLM, which decreases the likelihood of errors and, at the same time, increases the reliability. Although there are efforts to automatically create visuals (e.g. [4]), we believe that the automatic creation of visuals is currently too error-prone, let alone the inclusion of annotations.

The idea of categorization is intended to improve the flow of the data storytelling and decrease information overload due to intersecting information from different categories. It also potentially decreases the load on the LLM. This being said, it is a clear design trade-off and combining visuals from different categories has its own set of advantages.

4.1 Retrieval Augmented Architecture

One of the most important limitation, which is a challenge for all systems that integrate large language models, is its unpredictability. Despite providing a clear and well-defined structure to describe visuals and their parameters, the LLMs will occasionally make up information and include them inside the outputs. While our system is still in development, we conducted informal tests and made the experience that GPT-4o (The most current LLM from openAI as of writing this paper) mostly does a good job at outputting the correct structure and information most of the time. Still, errors occur and a way to handle them efficiently has yet to be explored.

To kick-off the discussion on improving the reliability of a system described in this paper, we present an architecture that integrates RAG, and a variation called Self-RAG [2] (see fig. 3). Self-RAG is specialized in self-reflection and aims to handle hallucinations, as well as evaluate the correctness of answers.

After the routing step, where an LLM decides between a category, guidance or general route, documents are retrieved from a vector store. In the case of XAI, this can be a store that may contain detailed information about XAI and machine learning as well as other business knowledge that is important for the use case. As presented before, in case of routing through a category, there is a path to retrieve data and build visuals. After generating the textual answer, Self-RAG is deployed, which utilizes LLMs to check the correctness of the output based on the general knowledge and vector store. Lastly, it is evaluated if the answer is actually answering the question. In our test system without RAG capabilities, answers that were not properly addressing the question were a prominent problem.

4.2 Evaluation Strategies

In this paper, we presented an approach to create conversational data stories. In future work, it is necessary to evaluate this approach in a structured way. Here, we want to outline potential starting points for the evaluation.

4.2.1 Data Literacy of Large Language Models. While there are many large language models that are specialized on code generation and the processing of structured data, we do not know about scientific work that focuses on the understanding of data structures by LLMs. When we talk about an architecture that is primarily

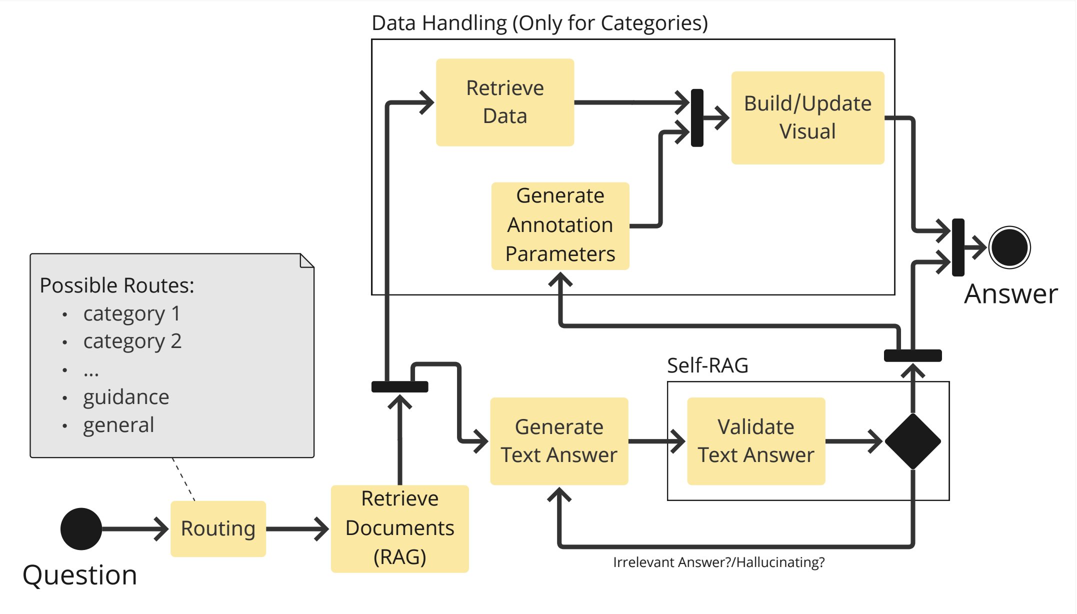


Figure 3: LLM Architecture enhanced by RAG and Self-RAG

concerned with understanding data and representations of visualizations, it is necessary to understand the data literacy of large language models and their abilities of understanding different structures.

One promising approach is to create a "data literacy test bench", which we feed with data structures (e.g. a confusion matrix) and questions about it. These questions should address different levels of complexity. Shao et al. did user studies where they categorized the question difficulty into *information retrieval* (level 1) and *comprehension* (level 2) [17]. Such a test bench could compare the capabilities of different LLMs with respect to data literacy. At the same time it could evaluate familiarity with different data structures as well as different ways to represent data structures.

4.2.2 Fixed Workflow. We already conducted unstructured tests with our initial implementation of the system. As a next step we want to define a fixed workflow with questions to build trust towards the application from understanding the basic data to the specific model prediction. For each consecutive step in the workflow, we will define one question and a target response. If the model does not respond in the desired way, we will prompt it until the target response is achieved. To handle the risk that the LLM is not able to provide the desired response, we can introduce a maximum number of trials n. As it is usually desirable to use a non-deterministic model (temperature > 0), we will rerun the workflow several times. In this way, we will build a structured record of the model behaviour and assess it based on the number of failures and output quality.

4.2.3 Usage Studies. To evaluate the system from users' end perspective, we are planning a comparative study where we compare the feedback for a non-conversational system that is similar to a dashboard experience and the conversational system that we presented in this paper. The non-conversational system will borrow from the same custom visualization library and categorization, providing comparability.

A central aspect of our system are the annotations that link the visual elements with the textual elements. We plan to evaluate different forms of linking from a user perspective and explore different ways to annotate and combine annotations. This may include the use of different colors, high-level annotations between text and the full visuals and approaches to annotate the specifically mentioned part of the visuals (grey numbers in fig. 1).

5 CONCLUSION

In this paper, we have presented the idea of *conversational data stories*, which integrate XAI as well as Generative AI. In addition, we have presented a system concept for implementing conversational data stories. It combines conversational agents and interfaces with visual components that are driven by data storytelling elements.

Up to now, we mostly conceptualized the system and built a first prototype that we tested in an unstructured way. To guide our next steps, we also presented an LLM architecture that is improved by RAG and outlined ways to evaluate such a system.

Future work will be focused on further refining the technical reliability of the system, i.e. decreasing and handling undesired behaviour by the LLMs. This includes the evaluation of relevancy of chosen data for the request (e.g. using CRAG [19]) as well as handling hallucinations and recognizing if the question was actually answered (e.g. Self-RAG [2]). In general, we want to further investigate how to utilize RAG to address domain knowledge-intensive tasks [8].

Also, we aim to test the system with users and evaluate how users rate such a user interface compared to more traditional user interfaces that are not driven by textual inputs.

ACKNOWLEDGMENTS

The German Federal Ministry for Economic Affairs and Climate Action supports this research under the grant KK5438301ER1. We thank our project collaborators for fruitful discussions as well as the anonymous reviewers for their valuable comments to a previous version of this paper.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. https://doi.org/10.48550/arXiv.2310.11511 arXiv:2310.11511 [cs.CL]

[3] Claus Atzenbeck, Sam Brooker, and Daniel Roßner. 2023. Storytelling Machines. In *Proceedings of the 6th Workshop on Human Factors in Hypertext* (Rome, Italy) *(HUMAN '23)*. Association for Computing Machinery, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3603607.3613481

[4] Victor Dibia. 2023. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. https://doi.org/10.48550/arXiv.2303.02927 arXiv:2303.02927 [cs.AI]

[5] Vanessa Echeverria, Roberto Martinez-Maldonado, Roger Granda, Katherine Chiluiza, Cristina Conati, and Simon Buckingham Shum. 2018. Driving data storytelling from learning design. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (Sydney, New South Wales, Australia) *(LAK '18)*. Association for Computing Machinery, New York, NY, USA, 131–140. https://doi.org/10.1145/3170358.3170380

[6] Valentin Grimm, Jonas Potthast, and Jessica Rubart. 2023. Motivational Exploration of Explanations in Industrial Analytics*. In *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*. IEEE, New York, NY, USA, 1–6. https://doi.org/10.1109/INDIN51400.2023.10217864

[7] Yi He, Shixiong Cao, Yang Shi, Qing Chen, Ke Xu, and Nan Cao. 2024. Leveraging Large Models for Crafting Narrative Visualization: A Survey. https://doi.org/10.48550/arXiv.2401.14010 arXiv:2401.14010 [cs.HC]

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474. https://doi.org/10.5555/3495724.3496517

[9] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. https://doi.org/10.48550/arXiv.2110.10790 arXiv:2110.10790 [cs.AI]

[10] Dongyu Liu, Sarah Alnegheimish, Alexandra Zytek, and Kalyan Veeramachaneni. 2022. MTV: Visual analytics for detecting, investigating, and annotating anomalies in multivariate time series. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30. https://doi.org/10.1145/3512950

[11] Gaia Mosconi, Dave Randall, Helena Karasti, Saja Aljuneidi, Tong Yu, Peter Tolmie, and Volkmar Pipek. 2022. Designing a data story: A storytelling approach to curation, sharing and data reuse in support of ethnographically-driven research. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23. https://doi.org/10.1145/3555180

[12] Nicolas Prat. 2019. Augmented analytics. *Business & Information Systems Engineering* 61 (2019), 375–380. https://doi.org/10.1007/s12599-019-00589-0

[13] Giulia Renda, Marilena Daquino, and Valentina Presutti. 2023. Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media* (Rome, Italy) *(HT '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 8 pages. https://doi.org/10.1145/3603163.3609035

[14] Jessica Rubart. 2018. Multimodal Interaction with Hypermedia Structures. In *Proceedings of the 1st Workshop on Human Factors in Hypertext*. ACM, New York, NY, USA, 17–21. https://doi.org/10.1145/3215611.3215613

[15] Jessica Rubart, Valentin Grimm, and Jonas Potthast. 2022. Augmenting Industrial Control Rooms with Multimodal Collaborative Interaction Techniques. *Future Internet* 14, 8 (2022), 224. https://doi.org/10.3390/fi14080224

[16] Jessica Rubart, Benjamin Lietzau, Patrick Söehlke, Bastian Alex, Stephan Becker, and Tim Wienböeker. 2017. Semantic navigation and discussion in a digital boardroom. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, IEEE, New York, NY, USA, 290–296. https://doi.org/10.1109/ICSC.2017.39

[17] Hongbo Shao, Roberto Martinez-Maldonado, Vanessa Echeverria, Lixiang Yan, and Dragan Gasevic. 2024. Data Storytelling in Data Visualisation: Does it Enhance the Efficiency and Effectiveness of Information Retrieval and Insights Comprehension?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 195, 21 pages. https://doi.org/10.1145/3613904.3643022

[18] Matthias Trier, Dennis Kundisch, Daniel Beverungen, Oliver Müller, Guido Schryen, Milad Mirbabaie, and Simon Trang. 2023. Digital Responsibility: A Multilevel Framework for Responsible Digitalization. *Business & Information Systems Engineering* 65, 4 (2023), 463–474. https://doi.org/10.1007/s12599-023-00822-x

[19] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective Retrieval Augmented Generation. https://doi.org/10.48550/arXiv.2401.15884 arXiv:2401.15884 [cs.CL]