

Washington University in St. Louis

## Washington University Open Scholarship

---

McKelvey School of Engineering Theses & Dissertations

McKelvey School of Engineering

---

Spring 5-13-2024

### Mending Trust in AI: Trust Repair Policy Interventions for Large Language Models in Visual Data Journalism

Hangxiao Zhu

*Washington University – McKelvey School of Engineering*

Follow this and additional works at: [https://openscholarship.wustl.edu/eng\\_etds](https://openscholarship.wustl.edu/eng_etds)



Part of the [Computer Sciences Commons](#), and the [Social and Behavioral Sciences Commons](#)

---

#### Recommended Citation

Zhu, Hangxiao, "Mending Trust in AI: Trust Repair Policy Interventions for Large Language Models in Visual Data Journalism" (2024). *McKelvey School of Engineering Theses & Dissertations*. 1013.  
[https://openscholarship.wustl.edu/eng\\_etds/1013](https://openscholarship.wustl.edu/eng_etds/1013)

This Thesis is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in McKelvey School of Engineering Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact [digital@wumail.wustl.edu](mailto:digital@wumail.wustl.edu).

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering  
Department of Computer Science & Engineering

Thesis Examination Committee:

Alvitta Ottley, Chair

Caitlin Kelleher

Yevgeniy Vorobeychik

Mending Trust in AI: Trust Repair Policy Interventions for Large Language Models in  
Visual Data Journalism  
by  
Hangxiao Zhu

A thesis presented to  
the McKelvey School of Engineering  
of Washington University in  
partial fulfillment of the  
requirements for the degree  
of Master of Science

May 2024  
St. Louis, Missouri

© 2024, Hangxiao Zhu

# Table of Contents

- List of Figures . . . . . iv
- List of Tables . . . . . v
- Acknowledgments . . . . . vi
- Abstract . . . . . vii
- Chapter 1: Introduction . . . . . 1**
- Chapter 2: Related Work . . . . . 5**
  - 2.1 Data Visualization in Journalism . . . . . 5
  - 2.2 Authoring Tools . . . . . 6
  - 2.3 Large Language Models . . . . . 7
  - 2.4 Trust in Intelligent Systems . . . . . 7
- Chapter 3: Methodology . . . . . 10**
  - 3.1 Data Visualization Interpretations Generation . . . . . 10
  - 3.2 Experimental Design . . . . . 11
- Chapter 4: Results . . . . . 15**
  - 4.1 Behavioral Trust Analysis . . . . . 16
    - 4.1.1 Participant Response Trends . . . . . 16
    - 4.1.2 Semantic Similarities . . . . . 18
  - 4.2 Cognitive Trust Analysis . . . . . 20
    - 4.2.1 Participant Rate Trends . . . . . 20
    - 4.2.2 Cognitive Trust Stability . . . . . 22
- Chapter 5: Discussion and Conclusion . . . . . 24**
  - 5.1 Training and Perception in Data Visualization . . . . . 24
  - 5.2 The Adaptability of Trust in AI . . . . . 25
  - 5.3 The Limited Role of Apologies in Trust Repair . . . . . 26
  - 5.4 Limitations and Future Work . . . . . 27
  - 5.5 Conclusion . . . . . 28

<b>References</b> . . . . .	<b>29</b>
<b>Appendix A: Evaluation</b> . . . . .	<b>33</b>
A.1 Behavioral Trust Supplement Graph . . . . .	33
A.2 Semantic Similarities . . . . .	33
A.3 Average Cognitive Trust Ratings . . . . .	34
A.4 Kruskal-Wallis H-test Results . . . . .	38

# List of Figures

Figure 3.1: Experiment Procedure . . . . .	12
Figure 3.2: Data Visualization Analysis Page . . . . .	13
Figure 3.3: Post-round Trust Level Check . . . . .	14
Figure 4.1: Behavioral Trust Trends in Trust Enhanced Group and Trust Eroded Group . . . . .	16
Figure 4.2: Behavioral Trust Trends in Trained Group and Untrained Group . . . . .	16
Figure 4.3: Comparison of Ratings by Apology . . . . .	21
Figure 4.4: Comparison of Ratings by Experience . . . . .	23
Figure A.1: Behavioral Trust Trends in Trained Group and Untrained Group Stacked by Trust Group . . . . .	33

# List of Tables

Table 4.1:	Semantic Similarities Analysis . . . . .	19
Table A.1:	Wilcoxon Signed-rank Test for Semantic Similarities . . . . .	34
Table A.2:	Average Quality Ratings . . . . .	35
Table A.3:	Average Factualness Ratings . . . . .	36
Table A.4:	Average Trustiness Ratings . . . . .	37
Table A.5:	Kruskal-Wallis H-test of Apology Efficacy . . . . .	38
Table A.6:	Kruskal-Wallis H-test of Experience Effect . . . . .	38

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Alvitta Ottley. Her guidance, support, and encouragement have been instrumental throughout my research journey. Professor Ottley's expertise and insights have greatly shaped my work, and her kindness and willingness to support my decisions have been incredibly motivating. I am truly grateful for the opportunity to learn from and work with such an outstanding mentor.

I would also like to extend my sincere thanks to Saugat Pandey, a PhD student in Professor Ottley's group, who has been an invaluable colleague and friend. Saugat's assistance, especially during challenging times, has been crucial to my progress. His knowledge, patience, and willingness to help have made a significant impact on my research experience.

Last but not least, I am grateful to my friends and lab mates who participated in the pilot study of my research. Their constructive feedback and insights have greatly contributed to the quality and success of my work.

This work would not have been possible without the collective support and contributions of all these individuals. I am truly thankful for each and every one of them.

Hangxiao Zhu

*Washington University in St. Louis*

*May 2024*



## ABSTRACT OF THE THESIS

Mending Trust in AI: Trust Repair Policy Interventions for Large Language Models in  
Visual Data Journalism

by

Hangxiao Zhu

Master of Science in Computer Science

Washington University in St. Louis, 2024

Professor Alvitta Ottley, Chair

Trust in Large Language Models (LLMs) emerged as a pivotal concern. This is because, despite the transformative potential of LLMs in enhancing the interpretability and interactivity of complex datasets, the opacity of these models and instances of inaccuracies or biases have led to a significant trust deficit among end-users. Moreover, there is a tendency for people to personify AI tools that utilize these LLMs, attributing abilities and sensibilities that they do not truly possess. This thesis exploits this personification and proposes a comprehensive framework of trust repair policies tailored to address the challenges inherent in LLM annotations within data journalism contexts. Grounded in principles of transparency, accountability, user control, feedback integration, and ethical consideration, our research aims to mend the trust breach and foster a more reliable, user-centric approach to AI-assisted data interpretation. Employing a novel experimental design with 84 participants across diverse demographics, we simulate the dynamics of trust formation, breach, and repair in the context of data visualizations, maps, and other visual journalism from The New York Times Graphics Desk and Washington Post. Our findings reveal that journalists, regardless of data visualization expertise, can identify inaccuracies in AI-generated content. Initial AI accuracy did not significantly influence long-term trust, but journalists with relevant expertise exhibited higher cognitive trust when faced with incorrect summaries. Surprisingly, specific

apology strategies had limited impact on trust repair; instead, accuracy and reliability of AI-generated content played a crucial role in maintaining and restoring trust. These findings emphasize the importance of accuracy and transparency in fostering trust between journalists and AI tools, highlighting the need for AI systems that prioritize real-time accuracy. This research contributes to the discourse on the responsible use of AI in data journalism and underscores the significance of collaborative efforts within newsrooms to ensure the integrity of AI-assisted storytelling.

# Chapter 1

## Introduction

The integration of artificial intelligence (AI) tools into both everyday life and professional domains has become a prominent trend. A key area of this integration is using AI to enhance data analysis and presentation for journalists, essentially aiding in storytelling through data. Data visualization plays a vital role in effective data storytelling [37]. This form of journalism is becoming increasingly common in mass media for sharing data-driven stories with the general public. Therefore, there is a high demand for AI-powered authoring tools that allow journalists to embed visualization into their stories while also utilizing Large Language Models (LLMs) to generate or refine the content. These tools promise to enhance the depth and accessibility of journalistic stories by offering capabilities to generate, refine, and contextualize textual content alongside creating compelling data visualizations. Indeed, previous research has demonstrated that LLMs possess the capability not only to interpret and comprehend data visualizations [23] but also to assist journalists in crafting articles that revolve around these visual narratives [35]. This support is particularly invaluable for journalists in creating or interpreting data visualizations, undoubtedly alleviating their workload and enhancing their storytelling capacity.

However, despite the potential of LLMs and decision support systems to significantly augment human capabilities in processing and analyzing vast datasets, challenges persist. Journalists frequently confront the ‘hallucinations’ and inaccuracies that LLMs might produce. [42]. This introduces problems with identifying when the information is inaccurate and building and repairing appropriate levels of trust. These issues can be more complicated in mixed media settings like visual data journalism. In such settings, the journalist, although skilled in crafting captivating stories, may have a comparatively lower level of visualization literacy. Thus, our research zeroes in on this specific intersection of human-AI collaboration: the interpretation of data visualizations in journalism. We focus on the critical challenges at the juncture of AI assistance and journalistic storytelling, seeking to address the complex

dynamics of trust in AI-generated content. Specifically, we investigate the following research questions:

- **RQ1.** Can journalists who may not be trained in data visualization authoring identify when AI summaries are inaccurate?
- **RQ2.** How does the initial accuracy of the AI summaries affect behavioral trust and cognitive trust?
- **RQ3.** How do trust tendencies—both behavioral and cognitive—vary among journalists with different levels of expertise?
- **RQ4.** Can we exploit factors affecting trust and use apologies to repair trust?

To explore the research questions, we conducted an experiment involving crowd workers from Prolific, all proficient in English. We utilized ChatGPT<sup>1</sup> to generate both *accurate* and *inaccurate* summaries for various data visualizations, with participants exposed to only one type of summary at a time. Participants were tasked with reviewing the data visualizations alongside the AI-generated summaries and were asked to submit a final summary they deemed satisfactory. This final submission could either be the original AI-generated summary, an edited version of the AI-generated summary, or an entirely new summary crafted by the participants themselves.

For RQ1, we surveyed participants’ familiarity and experience with data visualization to identify those potentially lacking formal training in this area. By examining their behavior in response to AI-generated summaries, specifically their capacity to identify inaccuracies, we aimed to gauge their proficiency. The inaccuracies were quantified by converting participants’ submitted summaries into word vectors and performing cosine similarity tests to assess the quality of their responses. We hypothesized that untrained journalists might not be able to identify errors in AI-generated data visualization summaries as effectively as trained journalists. However, our findings did not support this hypothesis, indicating that even journalists without formal training in data visualization possess the ability to detect inaccuracies in AI-generated content.

---

<sup>1</sup>We used the ChatGPT-4 web client to complete the data visualization analysis, without using the official OpenAI API.

For RQ2, we randomly assigned participants to two groups, with one group initially receiving accurate AI summaries and the other group encountering inaccurate summaries. This setup allowed us to observe their subsequent interactions with AI-generated summaries and measure their trust levels. By analyzing whether participants chose to utilize, edit, or disregard AI-generated summaries in their final submissions and by evaluating their confidence ratings in AI, we could infer the impact of initial AI summary accuracy on both behavioral and cognitive trust. We hypothesized that participants who are initially presented with accurate AI-generated data visualization summaries would exhibit greater ease of trust repair compared to those who are initially presented with inaccurate summaries when subjected to subsequent trust repair strategies. However, our findings did not support this hypothesis, suggesting that the initial accuracy of AI summaries does not significantly influence the ease of trust repair.

For RQ3, employing the aforementioned methodology, we compared data across journalists of varying expertise levels to examine how professional experience influenced their interactions with AI-generated summaries and their trust evaluations. We hypothesized that behavioral and cognitive trust might vary more strongly in experienced journalists compared to inexperienced journalists, as they might trust themselves more and identify even minor mistakes made by the AI. However, our findings did not align with this hypothesis, indicating that the level of professional experience does not significantly impact the strength of trust variations.

To address RQ4, after deliberately eroding trust by presenting inaccurate AI summaries in the middle of the experiment, we implemented different apology strategies aiming to mend participants' trust in AI. The effectiveness of these apologies was then assessed by analyzing changes in participants' subsequent actions and trust ratings. We hypothesized that participants who receive an apology following inaccurate AI advice would find it easier to repair their trust in the AI system compared to those who do not receive any form of apology. Among participants who receive an apology, we expected that those who receive an apology framed in terms of the AI's ability (e.g., acknowledging a mistake or a failure in processing) would find it more difficult to repair their trust compared to those who receive apologies framed in terms of integrity (e.g., commitment to accuracy) or benevolence (e.g., concern for the user's well-being). However, our findings did not support these hypotheses. Instead, we observed that participants' trust is restored once they are presented with correct AI responses, regardless of the presence or type of apology.

Our research offers significant contributions to the evolving field of AI-assisted journalism, particularly in understanding how journalists interact with AI-generated data visualizations and the trust dynamics involved.

- By examining the ability of journalists to identify inaccuracies in AI-generated content (RQ1), our study reveals that journalists, regardless of their expertise in data visualization, possess an intrinsic ability to detect errors in AI-generated summaries.
- Our investigation into the impact of initial AI accuracy on trust (RQ2) shows that, under the test conditions, journalists' trust in AI is primarily shaped by the current performance of the system rather than their prior experiences or preconceptions. This adaptability of trust emphasizes the need for AI systems to consistently deliver accurate and reliable content to maintain journalists' trust.
- Exploring the variations in trust tendencies among journalists with different levels of expertise (RQ3), we found that while overall trust tendencies were similar, trained journalists exhibited higher cognitive trust when faced with incorrect summaries.
- Our investigation into apology strategies as a means of trust repair (RQ4) reveals that different types of apologies had a limited impact on trust restoration. Instead, the accuracy of AI-generated content played a more crucial role in maintaining and repairing trust. This insight suggests that investing in robust error detection and correction mechanisms may be more effective than developing elaborate apology strategies.

However, it's important to acknowledge the limitations of our study, including its ecological validity. Our experimental design, focused primarily on summary evaluation, may not fully replicate the complex realities of journalism practice. Furthermore, by intentionally presenting incorrect (rather than merely misleading) summaries to test trust erosion, we recognize the need for subsequent research that more accurately mirrors the nuances of real-world AI usage in journalism, including studies that present a mix of correct, misleading, and incorrect summaries to better gauge journalists' ability to rely on AI.

# Chapter 2

## Related Work

### 2.1 Data Visualization in Journalism

Data visualization has a long and diverse history, evolving from simple data tables in the 2nd century to sophisticated interactive visualizations seen in today's digital media [10]. The strength of data visualization resides in its profound storytelling capabilities, which not only engage readers but also simplify complex data, making it accessible and comprehensible [31]. This characteristic has contributed to the extensive adoption of data visualization in journalism, a field where storytelling is paramount and continually evolving [26].

In contemporary journalism, data visualization is indispensable for distilling complex datasets into digestible and engaging visual formats that enhance storytelling. Visual elements like charts, maps, and infographics enable journalists to convey stories more effectively, making abstract data tangible and accessible to a broader audience [18]. These visual stories often play crucial roles in shaping public understanding of critical issues, such as electoral results, pandemic trends, or economic changes, demonstrating the power of visuals in driving public discourse.

It's important to recognize that the efficacy of data visualization in journalism hinges on the collaborative efforts within the newsroom, where every team member, be it a programmer, designer, or statistician, embraces the ethos of journalism [40]. This interdisciplinary approach presents a challenge for journalists who may lack expertise in data visualization and depend on technological tools for assistance. With the advent of advanced Large Language Models (LLMs), there is a newfound capacity to enhance how stories are visualized and presented, potentially empowering journalists who are less skilled in traditional data visualization techniques to produce high-quality visual narratives [28].

Nevertheless, this technological empowerment comes with its challenges, which are central to our research: assessing whether journalists can critically evaluate the accuracy of content generated by AI and understanding their trust dynamics with AI-powered tools. As AI becomes more entrenched in data journalism, understanding its impact on both the content created and the confidence journalists place in these tools is imperative. Our work connects the historical foundations of data visualization with contemporary advancements, investigating the integration of AI in journalism and its implications for trust and content verification.

## 2.2 Authoring Tools

Authoring Tools (ATs) are designed to produce professional, engaging, and interactive training content [13]. While traditionally utilized for creating digital courses, their functionality is increasingly recognized in broader applications due to their ability to minimize technical challenges and leverage WYSIWYG (“what you see is what you get”) interfaces [3]. These features simplify the authoring process and lower the barrier to entry in terms of required skill sets. Consequently, the use of authoring tools has expanded into other domains, including the creation of data documents [19] and the development of self-explanatory visualizations for journalists [34].

Thus, the application of ATs across various domains is gaining momentum, with their use in creating data documents emerging as a particularly promising direction. The process of producing data documents involves several recurring challenges, such as ensuring consistency in formatting, aligning textual descriptions with corresponding charts, and maintaining clarity and readability for diverse audiences [36]. Recent studies have shown that these challenges can be effectively addressed through the intelligent design of ATs. Features like automatic suggestions for text-chart references [19] and the integration of onboarding concepts into data narratives [34] can significantly enhance the functionality of these tools, making complex data more accessible and engaging for users.

Despite the effectiveness of ATs in the data visualization arena, integrating advanced LLMs to enhance their performance has yet to be fully explored. The Kori paper highlights the key concerns, including accuracy and reliability, in current methods for automatic text-chart linking and points to the potential benefits of a more robust, data-driven approach that



leverages advanced AI models [19]. Our study, while not directly focused on developing such an AT, anticipates the possibilities that integrating advanced LLMs into ATs could offer. This exploration provides a preliminary insight into how enhanced ATs could potentially improve the effectiveness of data visualization and storytelling.

## 2.3 Large Language Models

Drawing on the Transformer architecture [39], LLMs have driven significant advancements in the field of natural language processing. The introduction of bidirectional training in the BERT model [7] significantly enhanced performance across numerous tasks. Following this, OpenAI’s GPT series—including GPT-2 [30], GPT-3 [4], and the latest iteration, GPT-4 [1]—have continually expanded the capabilities and applications of LLMs. These models have demonstrated remarkable proficiency across various fields, offering vast potential and innovation to everyday applications and professional domains alike [38, 5, 12]. One of the most notable applications is in the field of data analysis, where these models contribute substantially, including through synthetic data generation [23], enriching dataset exploration with profound insights [25], and facilitating automated, high-precision data analyses [24].

Our research specifically explores one notable capability of LLMs in data analysis: interpreting data visualizations. Prior studies have investigated various aspects of this application, ranging from enhancing chart accessibility [11] and advising on visualization design [14] to refining color palettes for charts [33] and generating textual narratives to accompany charts [35]. In our study, we focus primarily on using ChatGPT to generate analysis and interpretations for provided charts, aiming to assess how its performance influences user trust in the model. This exploration is critical as it helps understand the reliability and effectiveness of LLMs in presenting data in a way that is both insightful and trustworthy to users.

## 2.4 Trust in Intelligent Systems

The concept of trust has been extensively explored within interpersonal contexts, particularly highlighted in seminal research that delineates trust as a multidimensional construct

encompassing ability, benevolence, and integrity [27]. The first dimension, ability, encompasses the skills, competencies, and characteristics that empower an entity to exert influence within a specific domain. Benevolence, the second dimension, refers to the degree to which a trustee is perceived to have intentions of goodwill towards the trustor, beyond any self-serving motives. Lastly, the concept of integrity involves the trustor’s perception that the trustee adheres to a set of principles deemed acceptable by the trustor. These dimensions collectively form the bedrock of trust dynamics, offering a lens through which the complexity of trust, particularly in the context of human-to-human interactions, can be dissected and understood.

Given the crucial role and inherent complexities of trust in societal dynamics, much scholarly attention has also been devoted to the scenarios in which trust is breached and the subsequent repair mechanisms. Historical research highlights that trust restoration is both possible and challenging [21], entailing a range of strategies categorized into short-term and long-term approaches [20]. Short-term strategies often include verbal statements, apologies, compensation, and denial, while long-term strategies might involve creating structural arrangements to monitor and control future interactions, reframing the incidents to alter perceptions or emotional responses, promoting forgiveness, and recognizing the strategic use of silence in the wake of trust violations.

As the capabilities of intelligent tools have advanced, the discussion around trust and its repair has similarly extended to encompass human-technology interactions. While there are fundamental similarities between trust dynamics in human-human and human-intelligent system relationships, previous studies have shown the unique aspects of various intelligent systems and the diverse contexts of their interactions with humans necessitate a nuanced application of traditional trust repair principles [9, 32, 16]. The factors influencing trust in AI, as identified in prior research, encapsulate a spectrum of elements including knowledge, transparency, explainability, certification, as well as self-imposed standards and guidelines [2]. Furthermore, the literature suggests that the efficacy of different apology types can vary significantly based on the nature of the trust breach in intelligent systems [15]. Building on these insights, our research specifically investigates trust dynamics within the context of AI-assisted data visualization—a relatively underexplored area within the broader discussion on trust in intelligent systems. This focus allows us to examine how traditional trust repair mechanisms might be adapted or rethought to effectively address and mend trust breaches in scenarios where AI tools play a critical role in processing and presenting complex data,

thereby contributing novel perspectives to the existing body of knowledge on trust repair in the era of advanced intelligent systems.

# Chapter 3

## Methodology

In this section, we outline the procedure we have followed when leveraging the Large Language Model (LLM) to generate analysis for provided charts, as well as the design of our experimental system.

### 3.1 Data Visualization Interpretations Generation

To initiate our study, we curated a collection of 15 diverse data visualizations from The New York Times Graphics Desk, including geographical maps, line charts, area charts, treemaps, histograms, scatter plots, and more. This selection was made to guarantee a diverse range of data visualizations in our experiment, providing a foundation for the generalization of our experimental findings.

In order to obtain high-quality responses from ChatGPT-4, structured prompts were designed and can be divided into three main components:

- **Background Introduction and Role-Playing Setting:** This initial part introduces the task to ChatGPT 4, framing it as a data visualization analyst. This role-playing scenario is designed to contextualize the AI’s function, guiding it to adopt the mindset of analyzing and interpreting data visualizations accurately.
- **Chain-of-Thought Prompting:** Following the setup, we employ the Chain-of-Thought prompting technique [41], which involves presenting a structured example of the desired analysis format. This method helps in steering the AI’s responses by clearly demonstrating how outputs should be logically organized and articulated.

- **Disclaimer and Emotional Stimuli:** The final segment of the prompt includes a disclaimer to overcome ChatGPT’s programming to reject tasks that involve generating intentionally incorrect data analysis, specifying that the request is for academic purposes. Additionally, we incorporated Emotional Stimuli techniques [22] to potentially enhance the quality and engagement of the AI’s output.

This structured approach to prompting aims to optimize the LLM’s output, ensuring that the generated interpretations are both contextually relevant and of high quality, suitable for academic analysis and discussion.

## 3.2 Experimental Design

Our experiment was conducted using a  $2 \times 4$  between-subjects design, with the Initial Trust State (Trust states: Enhanced Trust vs. Eroded Trust) and attribute of the apology (Apology type: None vs. Ability vs. Integrity vs. Benevolence) as independent variables. The participants were randomly assigned to one of the trust groups (Enhanced Trust-None:  $n = 11$ , Enhanced Trust-Ability:  $n = 8$ , Enhanced Trust-Integrity:  $n = 13$ , Enhanced Trust-Benevolence:  $n = 12$ , Eroded Trust-None:  $n = 14$ , Eroded Trust-Ability:  $n = 9$ , Eroded Trust-Integrity:  $n = 7$ , Eroded Trust-Benevolence:  $n = 7$ ). Trust was the primary dependent variable, evaluated through both behavioral and cognitive measures.

We recruited 84 participants through the Prolific platform, specifically targeting individuals with experience in journalism. Of these, 3 participants did not complete all questions, and 4 failed to pass the attention checks, leaving 77 valid profiles for analysis. All these 77 participants completed preliminary screening, including demographic surveys, initial trust assessments, and a color blindness test. The demographic breakdown of these participants included 38 females, 35 males, and 4 others. Education levels ranged with 21 holding a High School Diploma/GED, 7 with an Associate Degree, and 49 with a Bachelor’s Degree or higher. Age distribution was as follows: 2 participants were under 20, 18 were aged 20-29, 18 were aged 30-39, 25 were aged 40-49, and 14 were over 50.

Upon passing these criteria, participants were randomly assigned to either the *Trust Enhanced Group* or the *Trust Eroded Group*. Additionally, to examine how the journalistic

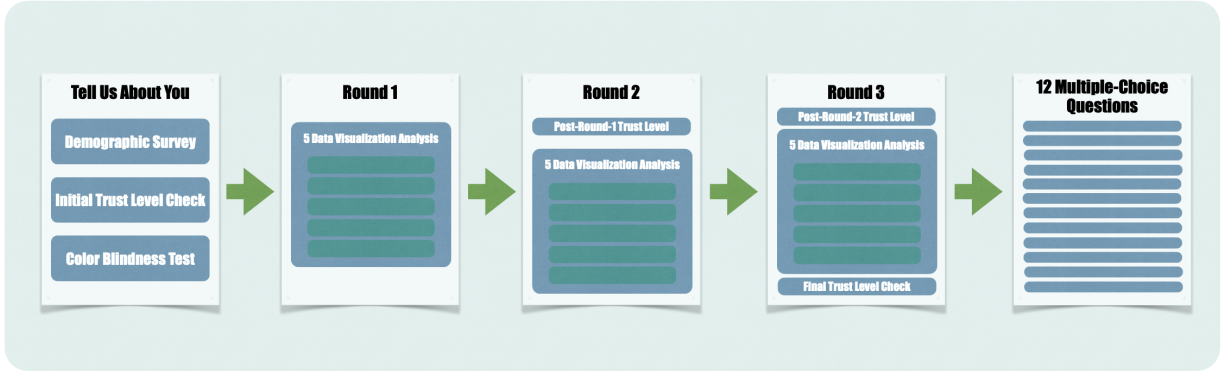


Figure 3.1: Experiment Procedure

experience with data visualization impacts participant performance, we categorized participants based on their familiarity with and professional exposure to data visualizations. Those with no relevant industry experience or unfamiliarity with data visualizations were classified as the *Untrained Group* ( $n = 23$ ), while those with relevant experience were designated as the *Trained Group* ( $n = 54$ ). This distinction allowed for a nuanced analysis of how expertise in data visualization influences participant responses within the experimental framework.

Each participant engaged in a series of 15 tasks divided into three rounds, with each round consisting of 5 data visualization analysis tasks followed by a break. Each task involved presenting participants with a data visualization sourced from The New York Times Graphics Desk along with an accompanying AI-generated description provided by ChatGPT-4. In the Trust Enhanced Group, participants initially received accurate AI descriptions to foster trust in Round 1, followed by incorrect descriptions to challenge this trust in Round 2, and finally, apologies paired with correct descriptions to repair trust in Round 3. Conversely, the Trust Eroded Group started with incorrect descriptions to undermine trust, followed by further incorrect descriptions, and ended with apologies and correct descriptions to attempt trust repair. The experiment procedure is shown in Figure 3.1.

Inspired by measurements of trust in previous study [17], participants were required to evaluate the AI-generated description based on its *reliability*, *accuracy*, and *trustworthiness* using a 5-point Likert scale for each aspect. In addition to these criteria, an attention check was implemented to ensure the quality of the data by filtering out participants not adequately engaged in the task. The specific dimensions assessed were as follows:

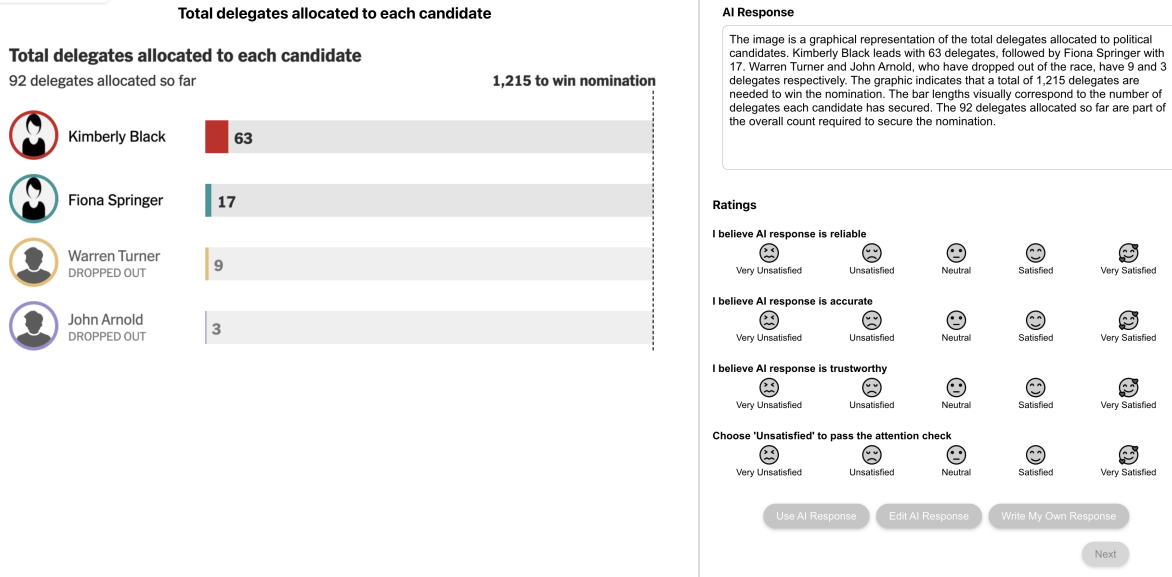
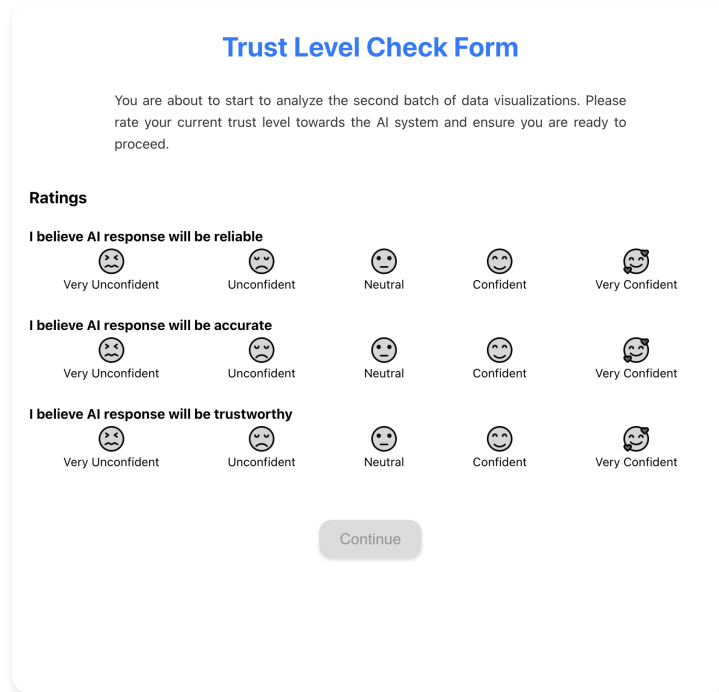


Figure 3.2: Data Visualization Analysis Page

- **Reliability:** How dependable do you find the AI's initial response?
- **Accuracy:** How precisely does the AI's response capture the essence of the data visualization?
- **Trustworthiness:** Can you trust the AI's initial response?
- **Attention Check:** Are you attentively participating in this experiment? 4 participants failed 7 or more out of 15 attention checks and were considered to be guessing randomly. These participants were excluded from further analysis, based on a critical value derived from a binomial distribution.

After evaluating the AI's response, participants chose one of three actions:

- **Use AI Response:** Opt to submit the AI-generated response without modifications.
- **Edit AI Response:** Modify or refine the AI-generated response before submission.
- **Write My Own Response:** Create and submit a response based on personal interpretation, ideally within 2 to 3 sentences.



The image shows a 'Trust Level Check Form' with a blue title. Below the title is an introductory paragraph: 'You are about to start to analyze the second batch of data visualizations. Please rate your current trust level towards the AI system and ensure you are ready to proceed.' The form is divided into three sections, each with a bold heading and five rating options represented by smiley faces. The first section is 'I believe AI response will be reliable', the second is 'I believe AI response will be accurate', and the third is 'I believe AI response will be trustworthy'. Each section has five options: 'Very Unconfident' (frowny face), 'Unconfident' (neutral face), 'Neutral' (neutral face), 'Confident' (smiling face), and 'Very Confident' (happy face). At the bottom center of the form is a grey 'Continue' button.

Figure 3.3: Post-round Trust Level Check

The page for the above data visualization analysis tasks is shown in Figure 3.2.

At the start of the second and third analysis phases, participants reassessed their trust toward the system using the same rating criteria, as shown in Figure 3.3. This step encouraged participants to reflect on their previous experiences with the system and anticipate their future interactions. After Round 3, participants provided their overall trust ratings for the system.

These tasks were structured to mimic real-world uses of AI in journalism and explore the complexities of trust in AI-generated content through firsthand interaction with the technology. After completing the visualization tasks, participants undertook a series of 12 multiple-choice questions [29] aimed at evaluating their ability to analyze and interpret different data visualizations. This section of the experiment was designed to assess the participants' skills in understanding and analyzing visual data representations, further highlighting their engagement and trust in AI-supported analysis.



# Chapter 4

## Results

In the multiple-choice question segment, participants demonstrated a high level of proficiency in analyzing and understanding data visualizations. The average score for all participants was 8.21 out of 12, indicating that they answered over 8 questions correctly on average. Interestingly, there was no significant difference in performance between the Untrained Group and the Trained Group, with untrained journalists scoring an average of 8.13 and trained journalists scoring 8.24. Out of the 77 participants, only 11 scored 6 or less out of 12 points, where 6 was determined as the critical value. However, it is important to note that these participants still passed the attention checks with scores of at least 4 and were subject to a 25-second time limit per question. This suggests that even those with lower scores were actively engaged in the task and possessed a sufficient level of data visualization literacy. These findings highlight the overall competence of our participants in interpreting and analyzing data visualizations, regardless of their prior training or experience in the field. This sets the stage for our subsequent evaluations, where we delve into the dynamics of behavioral and cognitive trust among participants across various experimental conditions.

In the following sections, we investigate how trust levels evolve throughout the experiment and examine the impact of participants' backgrounds, particularly their experience with data visualization, on their responses and decision-making processes. By analyzing these trust dynamics, we aim to gain insights into how participants navigate the challenges of working with AI-generated content in the context of data journalism.

# 4.1 Behavioral Trust Analysis

This section evaluates the behavioral trust exhibited by participants in different groups during the experiment. Behavioral trust was primarily gauged by observing the tendency of participants to accept AI-generated responses across a series of surveys.

## 4.1.1 Participant Response Trends

In our analysis, we operate under the assumption that when a participant opts to ‘Use AI Response’, it indicates they did not detect errors in the AI-generated content, demonstrating a measure of behavioral trust in the AI system. This assumption forms the basis of our investigation into the trends of participants’ choices throughout the experiment.

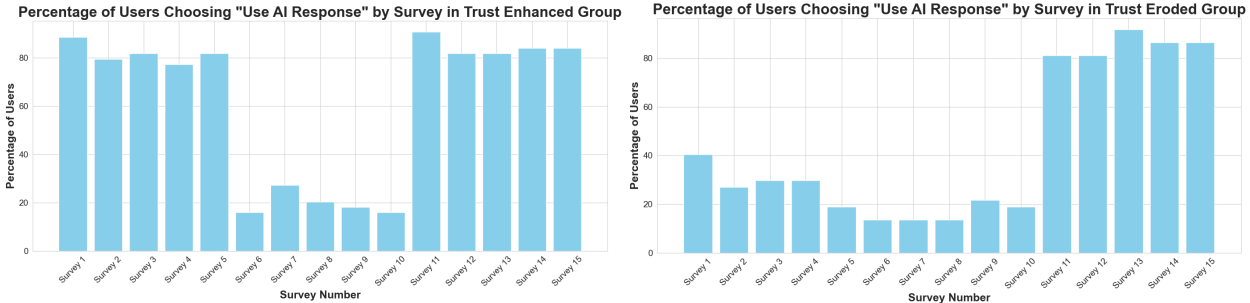


Figure 4.1: Behavioral Trust Trends in Trust Enhanced Group and Trust Eroded Group

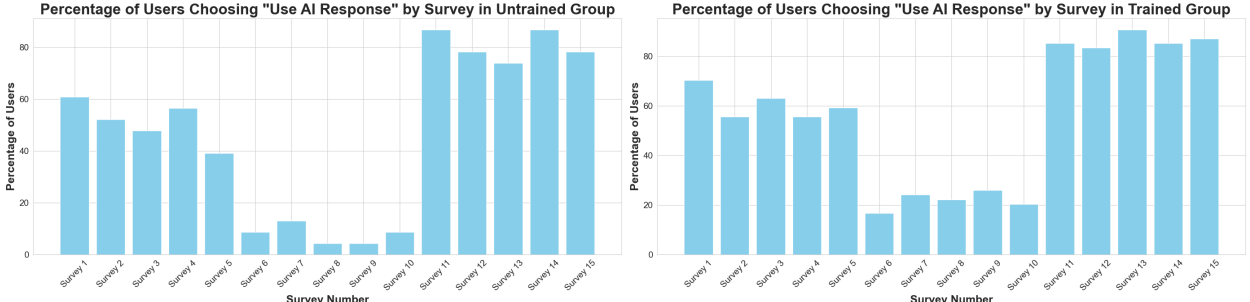


Figure 4.2: Behavioral Trust Trends in Trained Group and Untrained Group

Figure 4.1 illustrates participants’ trends in accepting AI responses. The Trust Eroded Group group exhibited high levels of initial trust with 88.64% of participants accepting the AI response in the first survey. Trust levels remained high through the initial correct

responses but significantly declined when incorrect responses were introduced in surveys 6-10. The acceptance rate dropped to as low as 15.91%, indicating a sharp decline in trust when faced with errors. In the Trust Eroded Group, initially, participants in this group showed a moderate level of trust with 40.54% opting to use the AI response in the first survey. However, as they continued to receive incorrect AI responses, their trust visibly eroded, reflected by a steady decline in accepting the AI's responses, reaching a low of 13.51% by the fifth survey. Notably, when the AI responses were corrected in the final rounds (surveys 11-15), the percentage of acceptance surged dramatically, suggesting a partial restoration of trust. However, similar to the Trust Enhanced Group, trust levels rebounded when correct responses resumed in the later surveys.

A Chi-squared test was conducted to compare the behavior of the two groups during the surveys where they were presented with incorrect AI responses, specifically focusing on surveys 6-10 for both groups. This period is critical as it represents the phase where both groups were exposed to similar conditions of incorrect information, providing a direct comparison of trust behavior under equivalent conditions of AI error. The results showed a Chi-squared statistic of 0.20 and a p-value of 0.66, suggesting no significant difference between the groups in their ability to recognize and react to the erroneous AI outputs. These results indicate that the initial trust conditioning—whether aimed at enhancing or eroding trust—did not significantly influence participants' behavioral responses to AI inaccuracies. Instead, both groups demonstrated a similar capacity to adjust their trust levels dynamically, based on the veracity of the AI's outputs, rather than the initial trust conditions set by the experiment.

To further understand the impact of professional training and experience in data visualization on trust dynamics, we examined how both the Untrained and Trained Groups responded to AI-generated descriptions during the surveys, particularly when they encountered incorrect responses, as shown in Figure 4.2. The Untrained Group, comprising participants with little to no formal training in data visualization, initially showed relatively high trust levels with 60.87% using the AI response in the first survey. However, this trust eroded significantly when faced with continued inaccuracies, plummeting to as low as 4.35% by the ninth survey. Despite this sharp decline, trust rebounded to 86.96% once accurate responses were reintroduced, indicating a conditional restoration of trust based on the correctness of AI responses. The Trained Group, consisting of participants experienced in data visualization, began with even higher initial trust levels at 70.37%. Although their trust also declined upon encountering errors, with the lowest of 16.67% participants accepting AI responses, the decrease

was less stark compared to the Untrained Group, and they maintained a somewhat higher baseline of trust even during the phase of inaccuracies. Trust levels similarly rebounded once the AI provided correct responses again. The different percentages of participants accepting AI responses in the first five surveys of Figure 4.2 compared to Figure 4.1 can be attributed to the fact that participants in the Trained and Untrained Groups could also belong to the Trust Eroded Group. Participants in the Trust Eroded Group faced inaccurate AI responses in the first five surveys, leading to a higher tendency to reject AI responses. This overlap in group composition contributes to the lower average rate of accepting AI responses in the Trained and Untrained Groups during these initial surveys. The stacked bar graphs in Appendix A.1 provide a visual representation of the composition of participants in the Trained and Untrained Groups, highlighting the proportion of participants from the Trust Eroded Group and Trust Enhanced Group within each group.

The Chi-squared test performed on the responses from surveys 6 through 10 showed a Chi-squared statistic of 1.66 and a p-value of 0.20, indicating that there was no statistically significant difference in the likelihood of participants from the Untrained versus Trained Groups choosing to ‘Use AI Response’ during the phase of incorrect information. This suggests that despite their differing backgrounds, both trained and untrained participants displayed a similar level of ability in identifying inaccurate AI-generated content. These findings reveal that while professional training in data visualization does not significantly alter the immediate behavioral trust responses to AI inaccuracies, it may influence the degree of trust erosion experienced. Both groups showed an ability to recover trust once the accuracy of AI responses was restored, but trained participants displayed a slightly more robust trust during the period of errors.

### 4.1.2 Semantic Similarities

In the evaluation of semantic similarities within the context of behavioral trust, our analysis was centered around the participants’ responses and their alignment with AI-generated content. By leveraging word vector transformations and computing cosine similarities, we were able to quantitatively assess how closely participants’ manual inputs—either generated independently or through edits—matched the AI’s responses. Table 4.1 summarized the comparative performances across different groups.

	Trust Enhanced Group		Trust Eroded Group	
	Untrained	Trained	Untrained	Trained
Write My Own to Correct	0.93 (+0.03)	0.94 (+0.03)	0.92 (+0.02)	0.92 (+0.02)
Write My Own to Incorrect	0.90	0.91	0.90	0.90
Edit AI Response to Correct	0.96 (+0.01)	0.96 (+0.00)	0.96 (+0.00)	0.96 (+0.01)
Incorrect to Correct	0.95	0.96	0.96	0.95

Table 4.1: Semantic Similarities Analysis

The semantic similarity analysis was specifically conducted on data where participants had interacted with AI responses deemed inaccurate. This subset included instances where participants chose to either ‘Write My Own’ response or ‘Edit AI Response’.

For participants who chose to ‘Write My Own’, the analysis compared the semantic similarity of their responses to both the correct and incorrect AI-generated answers. The results showed that participants’ self-written responses exhibited higher semantic similarities to the correct answers than to the incorrect ones they were initially shown. This finding suggests that even without being exposed to the correct responses, participants were able to recognize errors in the AI-generated content and provide more accurate answers. The Wilcoxon signed-rank test results (Table A.1) support this conclusion, with significant p-values across all groups, indicating that participants’ ability to generate responses closer to the correct answers is statistically significant.

For participants who opted to ‘Edit AI Response’, the analysis focused on comparing the semantic similarity of their edited responses to the correct answers, as well as the similarity of the original incorrect responses to the correct answers. The results revealed that the edited responses showed higher semantic similarity to the correct answers compared to the original incorrect responses. This suggests that through the editing process, participants were able to identify and rectify the inaccuracies in the AI-generated content, effectively transforming the responses to be more aligned with the correct information. The Wilcoxon signed-rank test results (Table A.1) further support this finding, with low p-values across all groups, indicating that the improvement in similarity to the correct answers after editing is statistically significant.

The analysis demonstrates that participants, regardless of their group or experience in relevant fields, possess the ability to detect and correct errors in AI-generated content. By choosing to write their own responses or edit the provided ones, participants consistently

produced answers that were more semantically similar to the correct information than the incorrect AI-generated responses.

## 4.2 Cognitive Trust Analysis

This section explores the cognitive trust demonstrated by participants across various groups during the experiment. Cognitive trust is assessed by analyzing the trust ratings provided by participants at different stages of the experiment, reflecting their confidence in the AI's performance.

### 4.2.1 Participant Rate Trends

Figure 4.3 displays the trends in trust ratings for both the Trust Enhanced and Trust Eroded Groups, segmented by ratings toward Reliability, Accuracy, and Trustworthiness. Each group consists of subgroups differentiated by their respective apology strategies, which include a control group without any apology, an Ability apology addressing perceived deficiencies in capability, an Integrity apology focusing on breaches of honesty, and a Benevolence apology concerning failures to prioritize user well-being and interests. These metrics shed light on the participants' perceptions of the AI's outputs, offering insights into how different apology approaches affect the perceived accuracy, reliability, and overall trustworthiness of the AI under varying conditions. Details of participants' ratings can be found in Tables A.2, A.3, and A.4.

In the Trust Enhanced Group, the initial trust ratings in survey 1 were slightly higher than their initial levels, indicating an initial optimism toward the AI system. Both groups demonstrated a rebound in cognitive trust with higher ratings observed when correct AI responses were reintroduced from survey 11 onwards. This trend highlights the participants' responsiveness to the accuracy of the information provided. Additionally, while trust levels significantly rebounded in the later surveys, there was a discernible decline in trust ratings post-experiment compared to those observed from surveys 11 to 15, returning to levels similar to those seen pre-experiment. This pattern suggests a complex long-term effect on trust, potentially influenced by participants' repeated exposure to both accurate and inaccurate

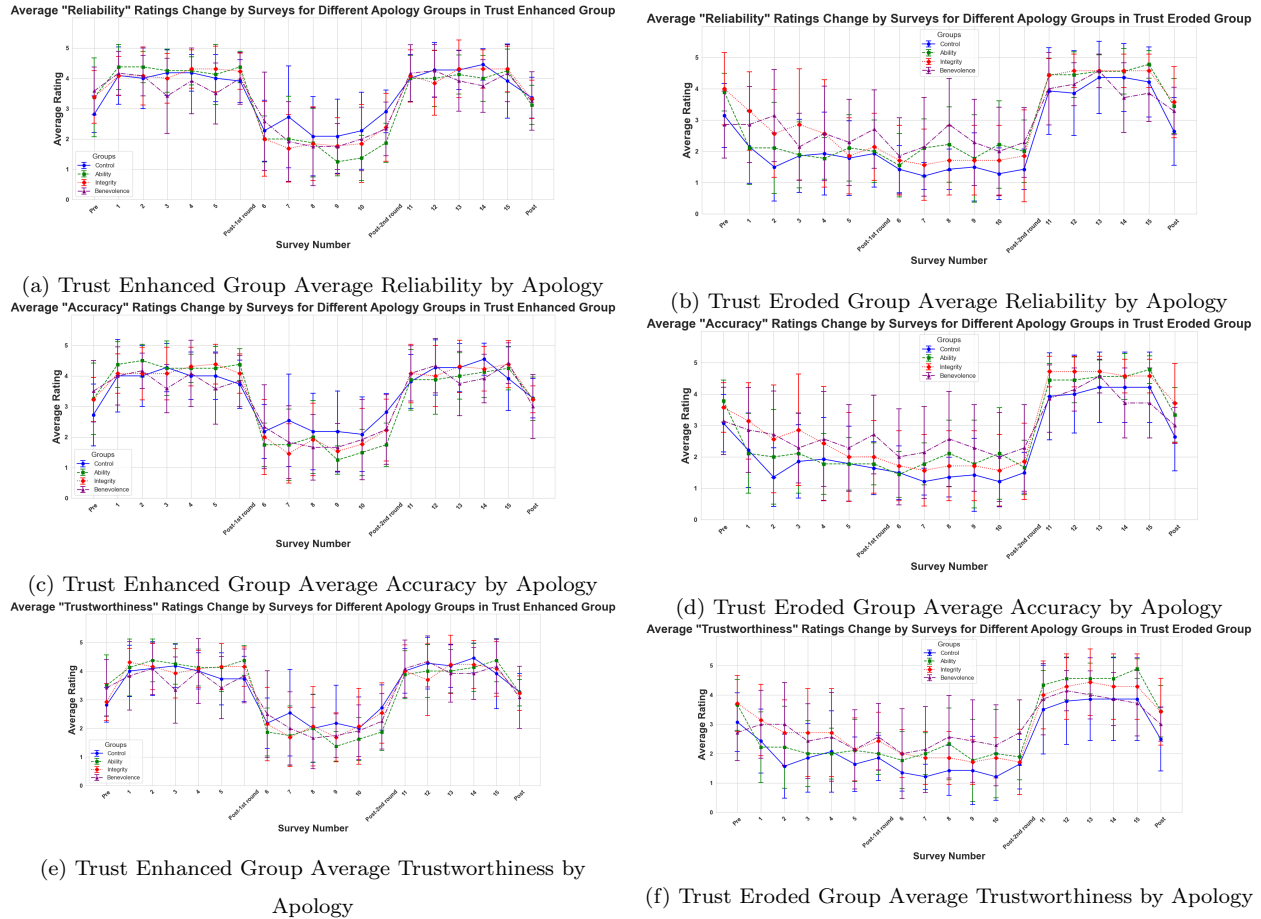


Figure 4.3: Comparison of Trust Enhanced Group and Trust Eroded Group on Reliability, Accuracy, and Trustworthiness Metrics by Apology

AI responses throughout the experiment. It also implies that users' preconceived notions or stereotypes about AI may not be easily or significantly altered over the short term.

Kruskal-Wallis H-tests were performed to evaluate the differences among the subgroups treated by different apology policies concerning their ratings across different surveys, specifically focusing on surveys 11 and 15. These tests aimed to determine if the different apology strategies (Control, Ability, Integrity, Benevolence) significantly influenced the cognitive trust ratings. Across both Trust Enhanced and Trust Eroded groups, the statistical tests revealed no significant differences in ratings between the different apology subgroups A.5. This indicates that the type of apology did not have a discernible impact on participants' cognitive trust. The lack of significant differences in the cognitive trust ratings among the different apology conditions suggests that the apologies, irrespective of their nature, did not

significantly affect the participants' trust restoration. This finding implies that the correction of information and the subsequent accuracy of AI outputs may be more influential in rebuilding trust than the specific content of apologies. The observation that control group participants also experienced restored trust supports the notion that improving the quality and accuracy of AI responses is crucial for effective trust management in AI systems.

### 4.2.2 Cognitive Trust Stability

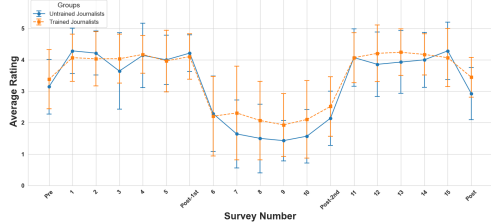
Figure 4.4 displays the trends in trust ratings for both the Trust Enhanced and Trust Eroded Groups, segmented by ratings toward Reliability, Accuracy, and Trustworthiness. The analysis segregates each group into Untrained and Trained subgroups based on their data visualization experience. This segmentation allows for a deeper understanding of cognitive trust behaviors among participants with varying levels of expertise.

The figures reveal that both Untrained and Trained journalists' trust levels are reinstated once correct AI responses are presented, indicating the resiliency of cognitive trust in response to accurate information. However, during periods when inaccurate AI responses were provided, the Trained subgroup consistently showed higher trust ratings than the Untrained subgroup. This suggests that trained individuals maintain a higher degree of trust stability even when faced with misinformation.

To quantitatively assess this observation, Kruskal-Wallis H-tests A.6 were performed on the trust levels of Untrained and Trained users from surveys 1 to 10 for the Trust Eroded Group and from surveys 6 to 10 for the Trust Enhanced Group—periods during which participants were exposed to inaccurate AI responses. The tests revealed significant p-values across all surveys, confirming that trained individuals consistently rated their trust higher compared to their untrained counterparts when encountering inaccurate answers. The results from the Kruskal-Wallis H-tests emphasize the impact of training and experience on the stability of cognitive trust, particularly in contexts involving the reliability of AI-generated information. The higher trust ratings among trained individuals may imply a deeper or more discerning appreciation of AI capabilities, which permits a sustained trust even when inaccuracies are present. However, this observation also prompts consideration of whether professionals' familiarity with established processes could lead to a potential underestimation of AI inaccuracies, resonating with findings that suggest seasoned professionals sometimes display a

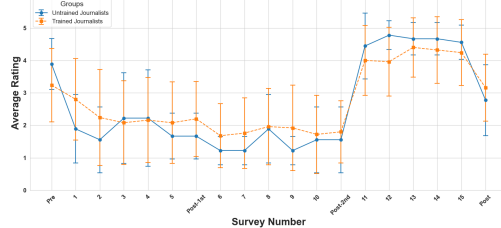


Average "Reliability" Ratings Change by Surveys for Different Experience Groups for Trust Enhanced Group



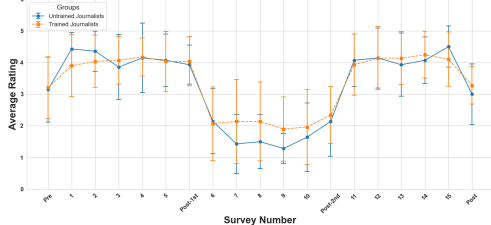
(a) Trust Enhanced Group Average Reliability by Experience

Average "Reliability" Ratings Change by Surveys for Different Experience Groups for Trust Eroded Group



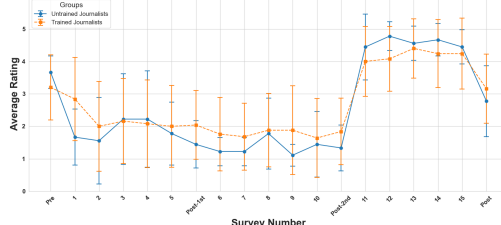
(b) Trust Eroded Group Average Reliability by Experience

Average "Accuracy" Ratings Change by Surveys for Different Experience Groups for Trust Enhanced Group



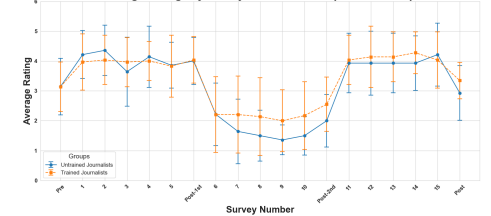
(c) Trust Enhanced Group Average Accuracy by Experience

Average "Accuracy" Ratings Change by Surveys for Different Experience Groups for Trust Eroded Group



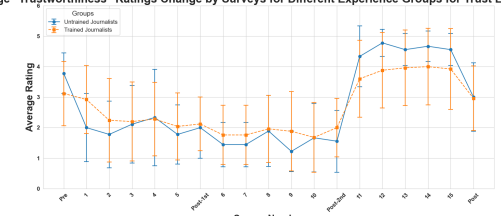
(d) Trust Eroded Group Average Accuracy by Experience

Average "Trustworthiness" Ratings Change by Surveys for Different Experience Groups for Trust Enhanced Group



(e) Trust Enhanced Group Average Trustworthiness by Experience

Average "Trustworthiness" Ratings Change by Surveys for Different Experience Groups for Trust Eroded Group



(f) Trust Eroded Group Average Trustworthiness by Experience

Figure 4.4: Comparative analysis of Trust Enhanced and Trust Eroded Groups across Reliability, Accuracy, and Trustworthiness Metrics by Experience

less meticulous approach in environments that deviate from their routine professional activities [8]. As AI gains a firmer foothold in the realm of data analysis, these insights emphasize the necessity of nurturing a profound understanding of AI among all users to promote a durable and adaptive human-AI alliance.

# Chapter 5

## Discussion and Conclusion

In this paper, we present an initial exploration into the dynamics of trust between journalists and artificial intelligence as it relates to the interpretation of data visualizations—a pivotal component of modern data journalism. Our experiment, designed to mimic the evolving interface of human-AI collaboration, reveals that journalists, irrespective of their expertise in data visualization, possess the intrinsic ability to identify inaccuracies in AI-generated content. These findings challenge preconceived notions about the necessity of formal training in data visualization for effective utilization of AI tools in journalistic settings. Through a series of carefully structured tasks, we have observed the malleability of trust in AI, noting that it fluctuates with the AI’s performance rather than being anchored to initial impressions. This study contributes to the body of knowledge in human-computer interaction by examining the subtleties of trust, its breach, and potential avenues for repair in the context of intelligent systems. Our discussion not only synthesizes these insights but also reflects on the implications for future integration of AI into journalism and the broader media landscape.

### 5.1 Training and Perception in Data Visualization

Our study addresses two critical research questions related to the role of training and expertise in journalists’ interactions with AI-generated content. RQ1 asks whether journalists who may not be trained in data visualization authoring can identify when AI summaries are inaccurate. The multifaceted evidence from our experiment, which encompasses participants’ decisions to accept AI responses, the semantic similarities between their edited or independently written responses and the correct summaries, as well as the observed trends in trust ratings, collectively substantiate the capability of untrained journalists to detect errors in AI-generated content. This finding highlights the inherent critical thinking skills

that journalists possess, which enable them to effectively evaluate the accuracy of AI-assisted content creation, even without extensive formal training in data visualization.

RQ3 delves deeper into the variations in trust tendencies among journalists with different levels of expertise. While both trained and untrained journalists demonstrated similar overall trust tendencies, our analysis revealed that trained journalists exhibited significantly higher cognitive trust when faced with incorrect summaries. This suggests that journalists with expertise in data visualization may have a more nuanced understanding of the capabilities and limitations of AI systems, allowing them to maintain a higher level of trust even when encountering inaccuracies. However, the similarity in overall trust tendencies underscores the importance of fostering a collaborative environment within newsrooms, where journalists with diverse expertise levels can work together to leverage the benefits of AI tools while ensuring the accuracy and integrity of the final output.

These findings have significant implications for the integration of AI tools in journalism practice. They suggest that while formal training in data visualization may enhance journalists' ability to work with AI-generated content, it is not a prerequisite for effective human-AI collaboration. Instead, the focus should be on nurturing the innate critical thinking skills of journalists and creating an environment that encourages open communication and collaboration among professionals with diverse expertise levels.

## **5.2 The Adaptability of Trust in AI**

RQ2 investigates the impact of initial accuracy on behavioral and cognitive trust in AI-generated content. Our experimental results revealed that the initial accuracy of AI summaries did not significantly influence journalists' trust in the long run. Instead, participants consistently reacted to the real-time accuracy of the information provided, adjusting their trust accordingly. This finding highlights the adaptability of trust in AI, suggesting that journalists' trust is primarily shaped by the current performance of the AI system rather than their prior experiences or preconceptions.

The malleability of trust observed in our study has important implications for the use of AI tools in journalism. It emphasizes the need for AI systems to consistently deliver accurate and reliable content to maintain the trust of journalists. Moreover, it suggests that trust in

AI is not a static construct but rather a dynamic one that evolves based on the system’s performance. This understanding can guide the development of AI tools that prioritize real-time accuracy and transparency, enabling journalists to make informed decisions about when to rely on AI-generated content and when to exercise their own judgment.

### 5.3 The Limited Role of Apologies in Trust Repair

Delving into RQ4, our investigation highlights the limited impact of apology strategies on repairing trust within the context of AI errors. Surprisingly, the type of apology—be it based on ability, integrity, or benevolence—did not result in significant differences in participants’ trust repair. Even in the absence of apologies, participants were able to restore trust when provided with accurate AI summaries following a period of inaccuracies.

These observations echo findings from the affective and behavioral forecasting literature [6], where individuals tend to overvalue the impact of apologies, anticipating greater positive effects than what manifests in reality. Just as people are prone to overestimating the value and trust-reinstating power of an apology following interpersonal transgressions, our participants might have expected apologies to have a more pronounced effect on their perceptions of the AI. Yet, the empirical data indicate that the realignment of trust hinges more on the substantive improvement in AI performance rather than the articulation of remorse or acknowledgment of fault.

The limited role of apologies in trust repair, as observed in our experiment, has implications for the design of AI systems in journalism. It suggests that investing in robust error detection and correction mechanisms may be more effective in maintaining trust than developing elaborate apology strategies. However, it is important to note that the effectiveness of apologies may vary depending on the context and the nature of the trust breach. Further research is needed to explore the nuances of trust repair in different scenarios and to identify the most appropriate strategies for addressing trust violations in AI-assisted journalism.

## 5.4 Limitations and Future Work

While our study provides valuable insights into the trust dynamics between journalists and AI-generated content, it is important to acknowledge its limitations and identify avenues for future research. One key limitation is the ecological validity of the experimental setting. In real-world journalism, professionals rarely rely solely on AI-generated summaries but instead have access to the original data and context. Future studies should aim to create more realistic scenarios that closely mirror the complex dynamics of journalism workflows to better understand how journalists interact with AI tools in practice.

Another limitation of our study is the intentional presentation of incorrect summaries to break trust, rather than using misleading or ambiguous summaries that may be more representative of real-world AI inaccuracies. In our experiment, the inaccurate AI responses were generated by ChatGPT-4 on purpose. Thus, its behavior may not accurately reflect the real mistakes it could make. Additionally, the inaccuracies in our study were detectable if participants paid enough attention. Consequently, the conclusion about the influence of participants' training experience on their behavior might not be entirely accurate. To fully address RQ1 and assess journalists' ability to recognize when to use AI-generated content, future research should include a mix of correct, misleading, and incorrect summaries that more closely resemble real-world AI inaccuracies. This approach would provide a more nuanced understanding of journalists' discernment skills in evaluating AI-generated content.

Furthermore, our study focused on a single language model (ChatGPT-4) and did not explore potential variations in trust dynamics across different AI models. While we expect the observed behavioral patterns to remain consistent across models, future research should investigate the use of multiple language models to validate this assumption and provide a more comprehensive understanding of trust in AI-assisted journalism.

Another limitation of our experiment design is the lack of measurement of "future" trust. As trust mainly affects users' future behavior, while our experiment measures their current trust levels and reactions, our conclusions might not accurately reflect users' trust if taken out of this experimental setting. In other words, we cannot precisely measure users' trust levels using this experiment design; instead, we are more likely monitoring users' satisfaction ratings towards AI. To gain a more comprehensive understanding of trust dynamics, future

research should include long-term experiments to observe how trust evolves over time as journalists interact with AI tools in real-world settings.

In addition to addressing these limitations, future research could explore the long-term effects of AI-assisted journalism on trust dynamics. Longitudinal studies that track journalists' trust in AI over extended periods could provide valuable insights into how trust evolves as journalists become more familiar with AI tools and as the capabilities of these tools advance. Furthermore, we could redesign the experiment to incorporate elements that simulate real-world consequences of trust, such as introducing monetary incentives or penalties based on the accuracy of the AI-generated content that participants choose to trust. This type of experimental design, known as a "trust game" or "investment game," could provide a more realistic measure of trust by capturing the potential risks and rewards associated with relying on AI tools in journalism. By incorporating these elements, future research can better assess how trust influences journalists' decision-making processes and behaviors when working with AI-assisted tools in high-stakes situations.

## 5.5 Conclusion

In conclusion, our study offers a foundation for understanding the complex trust dynamics between journalists and AI in the context of data visualization interpretation. By addressing the research questions and discussing the implications of our findings, we contribute to the growing body of knowledge on human-AI collaboration in journalism. As AI continues to transform the media landscape, it is crucial to build upon this research and explore ways to foster trust, transparency, and accuracy in AI-assisted journalism. Through continued investigation and the development of robust AI systems that prioritize these values, we can unlock the full potential of AI in enhancing data-driven storytelling and supporting the evolving roles of journalists in the digital age.

# References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] P. Bedué and A. Fritzsche. Can we trust ai? an empirical investigation of trust requirements and guide to successful ai adoption. *Journal of Enterprise Information Management*, 35(2):530–549, 2022.
- [3] P. Berking. Choosing authoring tools. *Attribution-Noncommercial-Share Alike*, 2016.
- [4] 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.
- [5] C.-W. Chiang, Z. Lu, Z. Li, and M. Yin. Enhancing ai-assisted group decision making through llm-powered devil’s advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI ’24*, page 103–119, New York, NY, USA, 2024. Association for Computing Machinery.
- [6] D. De Cremer, M. M. Pillutla, and C. R. Folmer. How important is an apology to you? forecasting errors in evaluating the value of apologies. *Psychological Science*, 22(1):45–48, 2011.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [8] O. Dieste, E. R. Fonseca C., G. Raura, and P. Rodríguez. Professionals are not superman: Failures beyond motivation in software experiments. In *2017 IEEE/ACM 5th International Workshop on Conducting Empirical Studies in Industry (CESI)*, pages 27–32, 2017.
- [9] C. Esterwood and L. P. Robert. Do you still trust me? human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, pages 183–188, 2021.
- [10] S. Few and P. Edge. Data visualization: past, present, and future. *IBM Cognos Innovation Center*, pages 1–12, 2007.
- [11] J. Gorniak, Y. Kim, D. Wei, and N. W. Kim. Vizability: Enhancing chart accessibility with llm-based conversational interaction, 2024.

- [12] E. Jo, D. A. Epstein, H. Jung, and Y.-H. Kim. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [13] M. Khademi, M. Haghshenas, and H. Kabir. A review on authoring tools. In *Proceedings of the 5th International Conference on Distance Learning and Education, IPCSIT*, volume 12, pages 40–44, 2011.
- [14] N. W. Kim, G. Myers, and B. Bach. How good is chatgpt in giving advice on your visualization design?, 2024.
- [15] P. H. Kim, K. T. Dirks, C. D. Cooper, and D. L. Ferrin. When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1):49–65, 2006.
- [16] T. Kim and H. Song. How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61:101595, 2021.
- [17] S. C. Kohn, E. J. De Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12:604977, 2021.
- [18] J. Lankow, J. Ritchie, and R. Crooks. *Infographics: The power of visual storytelling*. John Wiley & Sons, 2012.
- [19] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2022.
- [20] R. J. Lewicki and C. Brinsfield. Trust repair. *Annual review of organizational psychology and organizational behavior*, 4:287–313, 2017.
- [21] R. J. Lewicki and C. Wiethoff. Trust, trust development, and trust repair. *The handbook of conflict resolution: Theory and practice*, 1(1):86–107, 2000.
- [22] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie. Large language models understand and can be enhanced by emotional stimuli, 2023.
- [23] Z. Li, H. Zhu, Z. Lu, and M. Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore, Dec. 2023. Association for Computational Linguistics.



- [24] S.-C. Liu, S. Wang, T. Chang, W. Lin, C.-W. Hsiung, Y.-C. Hsieh, Y.-P. Cheng, S.-H. Luo, and J. Zhang. JarviX: A LLM no code platform for tabular data analysis and optimization. In M. Wang and I. Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 622–630, Singapore, Dec. 2023. Association for Computational Linguistics.
- [25] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. InsightPilot: An LLM-empowered automated data exploration system. In Y. Feng and E. Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 346–352, Singapore, Dec. 2023. Association for Computational Linguistics.
- [26] V. Markova and O. Sukhoviya. Storytelling as a communication tool in journalism: Main stages of development. *Journal of History Culture and Art Research*, 9(2):355–366, 2020.
- [27] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995.
- [28] A. L. Opdahl, B. Tessem, D.-T. Dang-Nguyen, E. Motta, V. Setty, E. Throndsen, A. Tverberg, and C. Trattner. Trustworthy journalism through ai. *Data Knowledge Engineering*, 146:102182, 2023.
- [29] S. Pandey and A. Ottley. Mini-vlat: A short and effective measure of visualization literacy. *Computer Graphics Forum*, 42(3):1–11, June 2023.
- [30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [31] M. T. Rodríguez, S. Nunes, and T. Devezas. Telling stories with data visualization. In *Proceedings of the 2015 Workshop on Narrative & Hypertext*, NHT ’15, page 7–11, New York, NY, USA, 2015. Association for Computing Machinery.
- [32] B. G. Schelble, J. Lopez, C. Textor, R. Zhang, N. J. McNeese, R. Pak, and G. Freeman. Towards ethical ai: Empirically investigating dimensions of ai ethics, trust repair, and performance in human-ai teaming. *Human Factors*, 66(4):1037–1055, 2024.
- [33] C. Shi, W. Cui, C. Liu, C. Zheng, H. Zhang, Q. Luo, and X. Ma. Nl2color: Refining color palettes for charts with natural language. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):814–824, 2024.
- [34] C. Stoiber, S. Radkohl, F. Grassinger, D. Moitzi, H. Stitz, E. Goldgruber, D. Girardi, and W. Aigner. Authoring tool for data journalists integrating self-explanatory visualization onboarding concept for a treemap visualization. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, CHIItaly ’23, New York, NY, USA, 2023. Association for Computing Machinery.

- [35] N. Sultanum and A. Srinivasan. Datatales: Investigating the use of large language models for authoring data-driven articles. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pages 231–235, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.
- [36] C. L. Teles De Oliveira, A. T. D. A. Silva, E. M. Campos, T. D. O. Araújo, M. P. Mota, B. S. Meiguins, and J. M. D. Morais. Proposal and evaluation of textual description templates for bar charts vocalization. In *2019 23rd International Conference Information Visualisation (IV)*, pages 163–169, 2019.
- [37] A. Unwin. Why is data visualization important? what is important in data visualization? *Harvard Data Science Review*, 2(1):1, 2020.
- [38] P. Vaithilingam, T. Zhang, and E. L. Glassman. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [40] W. Weber and H. Rall. Data visualization in online journalism and its implications for the production process. In *2012 16th International Conference on Information Visualization*, pages 349–356, 2012.
- [41] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [42] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples, 2023.

# Appendix A

## Evaluation

### A.1 Behavioral Trust Supplement Graph

Figure A.1 provides additional insights into the behavioral trust trends among trained and untrained users, with a focus on the composition of participants from the Trust Eroded and Trust Enhanced Groups within each experience level. The stacked bar graphs illustrate the percentage of users choosing "Use AI Response" for each survey, with the bars divided into two segments representing the proportion of users from the Trust Eroded Group and the Trust Enhanced Group.

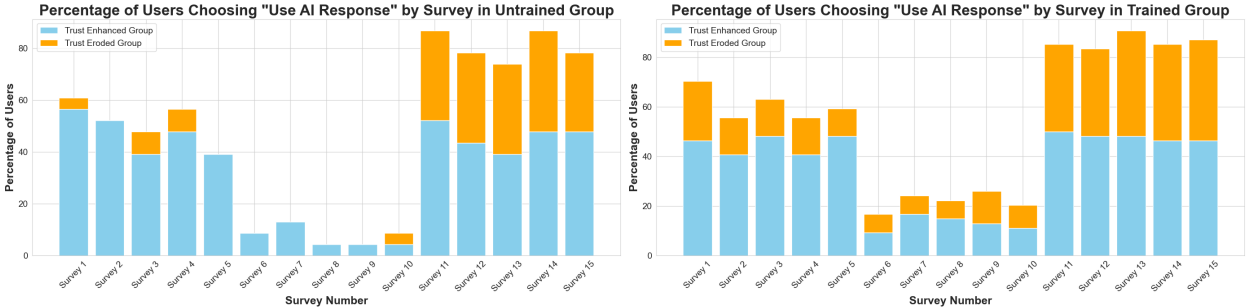


Figure A.1: Behavioral Trust Trends in Trained Group and Untrained Group, stacked by Trust Group

### A.2 Semantic Similarities

Table A.1 displays the results of the Wilcoxon Signed-rank Test, which evaluates the semantic similarities between participants' responses and the AI's correct and incorrect summaries. The 'p-value for Write My Own Group' involves comparing the semantic similarities of

participants’ self-generated responses to the correct AI responses against the similarities of the initially shown incorrect AI responses to the correct AI responses. This test determines if participants’ own responses align more closely with the correct content than the incorrect content they were initially provided. For the ‘p-value for Edit AI Response Group,’ the analysis focuses on participants who chose to edit the inaccurate AI responses they were given. It compares the semantic similarity between participants’ edited responses and the correct AI responses against the similarity between the original, unedited inaccurate AI responses and the correct AI responses. This assesses whether edits made by participants significantly enhance the alignment of the summaries with the correct AI-generated content. The statistically significant p-values across these tests confirm that journalists, regardless of their training background, are capable of identifying and amending inaccuracies in AI-generated summaries, thereby reinforcing the notion of AI as a supportive, rather than infallible, tool in data visualization tasks.

	Trust Enhanced Group		Trust Eroded Group	
	Untrained	Trained	Untrained	Trained
p-value for Write My Own Group	0.03770	0.00001	0.03245	0.00540
p-value for Edit AI Response Group	0.00046	0.00152	0.00913	0.01305

Table A.1: Wilcoxon Signed-rank Test for Semantic Similarities, comparing the differences in what participants wrote compared to what they saw

### A.3 Average Cognitive Trust Ratings

Table A.2, Table A.3, and Table A.4 depict the average ratings for Quality, Factualness, and Trustiness that participants assigned at different stages of the experiment. These tables elucidate the evolution of cognitive trust as participants navigated through accurate and inaccurate AI summaries. The ratings reflect the contingent nature of trust on AI performance, illustrating the criticality of content accuracy and the secondary role of apology in re-establishing trust.

	Trust Enhanced Group				Trust Eroded Group			
	Control	Ability	Integrity	Benevolence	Control	Ability	Integrity	Benevolence
<b>Pre</b>	<b>2.82</b>	<b>3.38</b>	<b>3.38</b>	<b>3.58</b>	<b>3.14</b>	<b>3.89</b>	<b>4</b>	<b>2.86</b>
Survey 1	4.09	4.38	4.08	4.17	2.14	2.11	3.29	2.86
Survey 2	4	4.38	4	4.08	1.5	2.11	2.57	3.14
Survey 3	4.18	4.25	4	3.42	1.86	1.89	2.86	2.14
Survey 4	4.18	4.25	4.31	3.92	1.93	1.78	2.57	2.57
Survey 5	4	4.12	4.31	3.5	1.79	2.11	1.86	2.29
<b>Post-1st</b>	<b>3.91</b>	<b>4.38</b>	<b>4.23</b>	<b>4</b>	<b>1.93</b>	<b>2</b>	<b>2.14</b>	<b>2.71</b>
Survey 6	2.27	2	2	2.58	1.43	1.56	1.71	1.86
Survey 7	2.73	2	1.69	1.92	1.21	2.11	1.57	2.14
Survey 8	2.09	1.88	1.85	1.75	1.43	2.22	1.71	2.86
Survey 9	2.09	1.25	1.77	1.75	1.5	1.78	1.71	2.29
Survey 10	2.27	1.38	1.85	2	1.29	2.22	1.71	2
<b>Post-2nd</b>	<b>2.91</b>	<b>1.88</b>	<b>2.38</b>	<b>2.33</b>	<b>1.43</b>	<b>2</b>	<b>1.86</b>	<b>2.29</b>
Survey 11	4	4	4.08	4.17	3.93	4.44	4.43	4
Survey 12	4.27	4	3.85	4.25	3.86	4.44	4.57	4.14
Survey 13	4.27	4.12	4.31	3.92	4.36	4.56	4.57	4.57
Survey 14	4.45	4	4.31	3.75	4.36	4.56	4.57	3.71
Survey 15	3.91	4.25	4.31	4.17	4.21	4.78	4.57	3.86
<b>Post</b>	<b>3.36</b>	<b>3.12</b>	<b>3.31</b>	<b>3.25</b>	<b>2.64</b>	<b>3.44</b>	<b>3.57</b>	<b>3.29</b>

Table A.2: Average Quality Ratings Across Different Survey Phases

	Trust Enhanced Group				Trust Eroded Group			
	Control	Ability	Integrity	Benevolence	Control	Ability	Integrity	Benevolence
<b>Pre</b>	<b>2.73</b>	<b>3.25</b>	<b>3.23</b>	<b>3.5</b>	<b>3.07</b>	<b>3.78</b>	<b>3.57</b>	<b>3.14</b>
Survey 1	4	4.38	4.08	4	2.21	2.11	3.14	2.86
Survey 2	4	4.5	4.08	4.17	1.36	2	2.57	2.71
Survey 3	4.27	4.25	4.08	3.58	1.86	2.11	2.86	2.29
Survey 4	4	4.25	4.31	4.08	1.93	1.78	2.43	2.57
Survey 5	4	4.25	4.38	3.58	1.79	1.78	2	2.29
<b>Post-1st</b>	<b>3.73</b>	<b>4.38</b>	<b>4.08</b>	<b>3.83</b>	<b>1.64</b>	<b>1.78</b>	<b>2</b>	<b>2.71</b>
Survey 6	2.18	1.75	2	2.33	1.5	1.44	1.71	2
Survey 7	2.55	1.75	1.46	1.83	1.21	1.78	1.57	2.14
Survey 8	2.18	2	1.92	1.67	1.36	2.11	1.71	2.57
Survey 9	2.18	1.25	1.54	1.67	1.43	1.78	1.71	2.29
Survey 10	2.09	1.5	1.77	1.92	1.21	2.11	1.57	2
<b>Post-2nd</b>	<b>2.82</b>	<b>1.75</b>	<b>2.23</b>	<b>2.25</b>	<b>1.5</b>	<b>1.67</b>	<b>1.86</b>	<b>2.29</b>
Survey 11	3.82	3.88	4.08	4.08	3.93	4.44	4.71	3.86
Survey 12	4.27	3.88	4	4.33	4	4.44	4.71	4.14
Survey 13	4.27	4	4.31	3.75	4.21	4.56	4.71	4.57
Survey 14	4.55	4.12	4.23	3.92	4.21	4.56	4.57	3.71
Survey 15	3.91	4.25	4.38	4.42	4.21	4.78	4.57	3.71
<b>Post</b>	<b>3.27</b>	<b>3.25</b>	<b>3.23</b>	<b>3</b>	<b>2.64</b>	<b>3.33</b>	<b>3.71</b>	<b>3</b>

Table A.3: Average Factualness Ratings Across Different Survey Phases

	Trust Enhanced Group				Trust Eroded Group			
	Control	Ability	Integrity	Benevolence	Control	Ability	Integrity	Benevolence
<b>Pre</b>	<b>2.82</b>	<b>3.5</b>	<b>2.92</b>	<b>3.42</b>	<b>3.07</b>	<b>3.67</b>	<b>3.71</b>	<b>2.71</b>
Survey 1	4	4.12	4.31	3.83	2.43	2.22	3.14	3
Survey 2	4.09	4.38	4.15	4.08	1.57	2.22	2.71	3
Survey 3	4.18	4.25	3.92	3.33	1.86	2	2.71	2.43
Survey 4	4	4.12	4.08	4	2.07	2	2.71	2.57
Survey 5	3.73	4.12	4.15	3.42	1.64	2.11	2.14	2.14
<b>Post-1st</b>	<b>3.73</b>	<b>4.38</b>	<b>4.15</b>	<b>3.83</b>	<b>1.86</b>	<b>2</b>	<b>2.43</b>	<b>2.57</b>
Survey 6	2.18	1.88	2.15	2.5	1.36	1.78	2	2
Survey 7	2.55	1.75	1.69	2	1.21	2	1.86	2.14
Survey 8	2	2	2.08	1.67	1.43	2.33	1.86	2.57
Survey 9	2.18	1.38	1.69	1.75	1.43	1.78	1.71	2.43
Survey 10	2	1.62	2.08	1.92	1.21	2	1.86	2.29
<b>Post-2nd</b>	<b>2.73</b>	<b>1.88</b>	<b>2.54</b>	<b>2.25</b>	<b>1.64</b>	<b>1.89</b>	<b>1.71</b>	<b>2.71</b>
Survey 11	4	3.88	4	4.08	3.5	4.33	4	3.86
Survey 12	4.27	4	3.69	4.33	3.79	4.56	4.29	4.14
Survey 13	4.18	4	4.23	3.92	3.86	4.56	4.43	4
Survey 14	4.45	4.12	4.23	3.92	3.86	4.56	4.29	3.86
Survey 15	3.91	4.38	4.08	4.17	3.86	4.89	4.29	3.71
<b>Post</b>	<b>3.27</b>	<b>3.25</b>	<b>3.23</b>	<b>3.08</b>	<b>2.5</b>	<b>3.44</b>	<b>3.43</b>	<b>3</b>

Table A.4: Average Trustiness Ratings Across Different Survey Phases

## A.4 Kruskal-Wallis H-test Results

Table A.5 outlines the results of Kruskal-Wallis H-tests performed to examine the efficacy of various apology strategies on the cognitive trust repair across different groups. The test results reveal no significant differences in the effectiveness of the apology strategies, emphasizing that the rectification of information accuracy outweighs the role of apologies in rebuilding cognitive trust among users.

Table A.6 presents Kruskal-Wallis H-test results for the effect of participants’ training experience on trust levels. Significant p-values indicate that training in data visualization contributes to higher trust stability in the face of inaccuracies in AI-generated content. This table confirms the influence of professional expertise on the resilience of cognitive trust during interactions with intelligent systems.

	Trust Enhanced Group			Trust Eroded Group		
	Reliability	Accuracy	Trustworthiness	Reliability	Accuracy	Trustworthiness
Survey 11	p = 0.90	p = 0.85	p = 0.90	p = 0.92	p = 0.36	p = 0.71
Survey 15	p = 0.91	p = 0.58	p = 0.86	p = 0.13	p = 0.13	p = 0.07

Table A.5: Results of Kruskal-Wallis H-tests assessing the efficacy of different apology subgroups (Control, Ability, Integrity, Benevolence) in repairing trust within Trust Enhanced and Trust Eroded Groups. The tests evaluate the impact on three metrics: Reliability, Accuracy, and Trustworthiness, at two survey checkpoints (Survey 11 and Survey 15). P-values are reported to indicate the statistical significance of variations in trust repair across subgroups.

	Trust Enhanced Group			Trust Eroded Group		
	Reliability	Accuracy	Trustworthiness	Reliability	Accuracy	Trustworthiness
p-value	0.02	0.01	0.01	0.04	0.03	0.02
H-statistic	5.63	6.31	7.81	4.41	4.80	5.12

Table A.6: Results of Kruskal-Wallis H-tests assessing the effect of training experience (Untrained and Trained) in data visualization within Trust Enhanced and Trust Eroded Groups. The tests evaluate the impact on three metrics: Reliability, Accuracy, and Trustworthiness, at survey checkpoints where participants were facing inaccurate AI responses. P-values are reported to indicate the statistical significance of variations in trust levels across subgroups.