Scopus

Package Printing
ISSN: 0163-9234
Vol. 72 Issue 4 (2025)

AI-Powered Data Storytelling: Transforming Raw Data into Actionable Insights

Lilibeth DG. Antonio, DIT[1*]

[1] College of Information and Communications Technology, Bulacan State University, Philippines

[*] Corresponding author: lilibeth.antonio@bulsu.edu.ph

Submission Date: 10.07.2025 | Acceptance Date: 11.09.2025 | Publication Date: 27.12.2025

ABSTRACT

The aim of this study was to determine the effectiveness of AI-driven data storytelling systems in converting raw data into actionable insights through the quantification of user ratings most specially in terms of clarity, accuracy, trustworthiness, and usefulness. The need for systems that can convert information into readable stories has increased as information complexity in different industries has continued to rise. In this research, the effectiveness of AI-stories in naturalistic data interaction situations was compared with human-authored stories generated through natural language generation (NLG) among a sample of 120 participants from different industries, such as government, healthcare, education, IT, and business. Participants were randomized into a control group, and human-written stories, or an experimental group, and AI-generated stories from a GPT-4–driven NLG system and Power BI dashboards. The findings revealed that the experimental group rated AI-generated stories significantly higher in intelligibility ($M = 4.38$, $p < .001$), user trust ($M = 4.13$, $p < .01$), and usefulness ($M = 4.35$, $p < .001$). While the control group attained a marginally higher accuracy rating, and the difference predicted was non-statistically significant. Thematic analysis revealed three broad experiential themes: improved accessibility and user engagement, guarded optimism over narrative trust, and perceived usefulness for decision-making environments. These findings support the potential of artificial intelligence as an assistive data narrative companion that can augment user understanding and engagement when used with existing data systems. The study implies that the optimal solution to address risks and improve trust is to integrate human-in-the-loop design, transparency features, and domain-specific calibration. This study adds a validated framework for the evaluation of AI-generated stories and provides real-world implications on the effective and ethical dissemination of information.

Keywords: AI-powered storytelling, Natural Language Generation (NLG), Data visualization, User trust in AI Decision-making utility

INTRODUCTION

In today's data society, organizations in most industries, including business and healthcare, generate more data than ever. However, the volume of unprocessed data does not automatically translate into actionable information (Dykes & Johnson, 2023). The value of this data is, to a large degree, untapped because it is impossible to interpret and communicate (Dykes & Johnson, 2023). Data storytelling, the combination of data analysis with narrative and visualization, has become an important way to transform complex datasets into decision-enabling information (Wang et al., 2021). It is solving an age-old issue: how to visualize and

tell analytical results consumable and compellingly to technical and non-technical stakeholders alike.

New developments in artificial intelligence (AI), especially in Natural Language Generation (NLG) and Machine Learning (ML), have created new opportunities for automation and upscaling the process of data storytelling. AI technologies can now generate stories almost indistinguishable from human writing, explain trends, and offer personalized insights specific to the user's context (Zhou et al., 2023). Integrating AI with data visualization platforms enables real-time interpretation of dashboards and infographics, thus turning static graphics into dynamic, conversational interfaces (Chen et al., 2024). The potential greatly lessens the cognitive burden of users and democratizes data comprehension at organizational levels (Hoque & Islam, 2024).

AI-driven storytelling is increasingly seen across many industries, such as finance, health, and education. For example, natural language generation-based clinical decision support systems offer clinicians synthesized patient summaries, while finance websites leverage AI to tell portfolio trends to investors (Gkatzia et al., 2020; Kim & Lee, 2022). The convergence of data analytics, visualizations, and narrative technologies has enhanced how users interact with and respond to data insights in each case. Nevertheless, academic analysis of these systems is scarce, especially regarding how users understand, trust, and respond to AI-created stories in real-world applications, even though these systems are gaining popularity.

In addition, the growing dependence on generative artificial intelligence is plagued by substantial challenges. Among them is a phantasm, which occurs when systems produce factually inaccurate information that sounds reasonable and, in the process, can mislead consumers (Ji et al., 2023). Others are the exacerbation of data bias, model transparency in decision-making, and loss of human critical thinking due to excessive reliance on automation (Suresh & Guttag, 2021). These limitations highlight the need to create a human-in-the-loop approach that balances the contextual sensitivity of human judgment with the ability of artificial intelligence (Hoque & Islam, 2024). This paper aims to bridge the empirical divide in AI-based data storytelling research by analyzing how these systems convert raw data to useful information. It provides a methodical framework for evaluating the transparency, accuracy, user trust, and decision-making capabilities of narratives produced by artificial intelligence. The research significantly contributes to the advancement of AI-based storytelling models that prioritize transparency, credibility, and humanism by leveraging the intersection of ethical guidelines, human-centric research, and future technologies. The ultimate objective of this paper is to facilitate the progression of data communication in the automation sector.

MATERIALS AND METHODS

Research Design

This research employed a mixed-methods approach which was combined experimental measurement with qualitative feedback to evaluate the effectiveness of AI-driven data storytelling platforms. The objective was to establish and validate an assessment framework that would evaluate specially the clarity, accuracy, trustworthiness, and usefulness value of AI-driven stories.

Framework Development

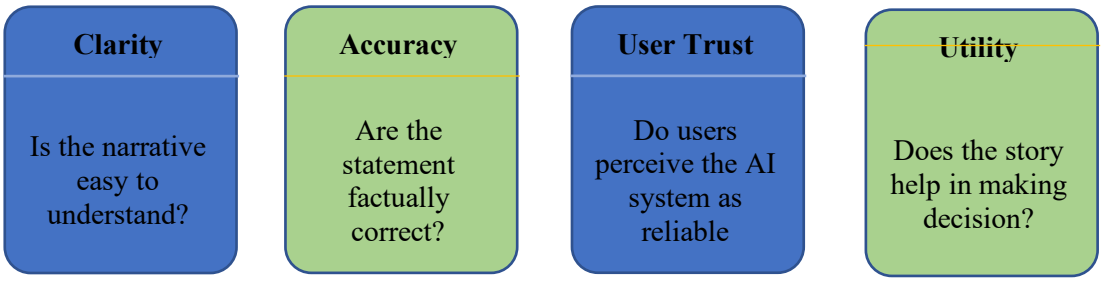

Figure 1. AI-Powered Data Storytelling Evaluation Framework

This study conceptualized and created a four-dimensional framework for assessing AI-generated data stories and most specially based on data communication, explainable AI, and human-AI interaction literature. The main objective of the framework is to methodically measure the effectiveness of AI-generated data stories. The framework includes four dimensions which mentioned are the: Clarity, which gauges the readability and understandability of AI-generated stories; Accuracy, which gauges whether reported data follows the original data set; User Trust, which evaluates the user's attitudes towards system reliability, fairness, and transparency; and last but not the least is the Utility, which gauges the narrative's ability to facilitate practical, real-world decision-making. The dimensions were tested via expert review and were based on a synthesis of literature such as Ji et al. (2023), Wang et al. (2021), and Suresh & Guttag (2021). In the empirical stage of the study, the framework served as a starting point for developing analytical rubrics, user roles, and survey items.

Participants and Sampling

Purposive sampling was used to recruit 120 participants with a representative sample of industries where data storytelling tools are normally used. Participants were business, healthcare, education, and information technology professionals and advanced students. All participants had prior experience working with data dashboards, data analytics tools, or decision-making processes where data summaries were being interpreted. The inclusion criterion was the capacity to interpret and understand visual data outputs, and the exclusion was for participants with no experience interpreting data visuals or narratives.

Table 1. Distribution of Respondents by Sector

Sector/Industry	***Number of Respondents***	***Percentage (%)***
Business and Finance	35	29.2%
Healthcare and Medical	25	20.8%
Education (Teachers/Admin)	30	25.0%
Information Technology	20	16.7%
Government/Public Sector	10	8.3%
Total	**120**	**100%**

As can be seen from Table 1, respondents were fairly representative of differing data-dependent domains. The Business and Finance domain had the largest share at 29.2%, reflecting the growing use of AI-generated analytics and dashboards in business decision-making contexts. Teachers and school administrators accounted for 25.0% of the sample, thus capturing the emerging importance of data-driven education approaches, especially curriculum design and student evaluation. Healthcare practitioners accounted for 20.8%, as AI-generated clinical summaries are increasingly used in diagnosis and reporting. The IT domain accounted for 16.7% of respondents, whose experience of AI application in data analytics was already high. A smaller but notable share, amounting to 8.3%, was from the government and public administration, where the focus on digital governance and open data is emerging as a key theme. This heterogeneity gave a wide-based representation of user views regarding AI-based data storytelling across differing real-world contexts.

Data Collection Procedure

During data collection, the 120 participants were randomly assigned to two groups: experimental and control groups. Professional human analysts manually crafted the dashboard and their respective summaries for the control group. The summaries mirror the conventional narrative reporting templates widely used in education, healthcare, and business contexts. The experimental group, on the other hand, utilized the same dashboards; however, their summaries were produced through an artificial intelligence system based on a Natural Language Generation (NLG) model, i.e., an LLM fine-tuned model like GPT-4. The dashboard environment was coupled with AI-generated narratives to reflect real-world applications of AI-based data storytelling.

All the participants were asked to complete a structured questionnaire after reading the corresponding dashboards and narratives. The survey instrument had two large sections. In the first section, a set of Likert-scale items was used to assess the four dimensions of the proposed evaluation framework: lucidity, accuracy, user trust, and decision-making utility. The researchers could quantify the participants' perceptions about the narratives they were presented with through these items. The second section contained open-ended questions for more information about the participants' experiences, perceptions, and issues they faced related to the narratives' quality, trustworthiness, or usability. These qualitative answers gave a better understanding of how users interact with AI-produced stories and how these experiences differ from what is obtained through traditional human-written reports. All the participants completed the exercise individually on their own using their gadgets such as their mobile phones or laptops, and the directions were identical for both groups to ensure overall homogeneity. The whole session, from filling out the survey to doing the dashboards, took between 20 to 30 minutes per participant. Afterward, the data were gathered, cleaned, and assessed for statistical and thematic analysis.

Research Instrument

An AI story system was created using GPT-4 APIs which is a well-known and readily available source in the internet and integrated into Microsoft Power BI software to produce real-time

narrative reports from visualized data sets. For measuring user perceptions, a validated survey was administered by adapting the instruments used by Wang et al. (2021) and Ji et al. (2023). The survey was designed to elicit some dimensions, such as trust and utility. Two independent human coders compared the generated statement's content to the original data set to compare the accuracy of the AI-generated narratives. Inconsistencies were resolved and discussed to achieve objective and consistent ratings of accuracy.

Data Analysis

The responses of the control and experimental groups were then compared using descriptive statistics and independent-sample t-tests for quantitative data analysis. The objective of these tests was to determine if there were statistically significant differences between the two groups regarding lucidity, accuracy, user trust, and decision-making utility. In the meantime, qualitative responses were analyzed using thematic analysis to produce critical insights regarding user experience, the degree of trust in narratives produced by artificial intelligence, and perceived risk, which included a fear of human loss of control or the risk of misinformation.

RESULTS AND DISCUSSION

The quantitative analysis focused on four core evaluation dimensions—Clarity, Accuracy, User Trust, and Utility—based on participant ratings using a 5-point Likert scale. Each dimension was assessed through multiple survey items, and reliability was confirmed through Cronbach's alpha coefficients, which ranged from 0.82 to 0.89, indicating acceptable internal consistency.

Evaluation Dimensions by Group

Table 2. Descriptive statistics for evaluation dimensions by group

Evaluation Dimension	*Group*	*Mean (M)*	*Standard Deviation (SD)*
Clarity	Control	3.87	0.62
	Experimental	4.38	0.49
Accuracy	Control	4.51	0.41
	Experimental	4.28	0.54
User Trust	Control	3.72	0.70
	Experimental	4.13	0.55
Utility	Control	3.94	0.58
	Experimental	4.35	0.46

The descriptive statistics for the lucidity, accuracy, user trust, and utility evaluation dimensions on which inter-group comparison is feasible between the control and experimental groups are shown in Table 2. The experimental group recorded high utility (M = 4.35, SD = 0.46) and clarity (M = 4.38, SD = 0.49), thus showing that the intervention or redesigned method was more lucid and practically helpful. Surprisingly, the control group scored higher accuracy (M = 4.51, SD = 0.41) than the experimental group (M = 4.28, SD = 0.54). Such a finding implies that trade-offs are involved in making the method clearer and more usable at the expense of

perceived accuracy. Further, user trust was higher in the experimental group (M = 4.13, SD = 0.55) than in the control group (M = 3.72, SD = 0.70), which means that users are more likely to trust information with higher lucidity and utility. This finding agrees with the research conducted by by Ghasemi and Karami in 2021, who established that user trust and engagement are significantly determined by user-centric design aspects, most of especially those that maximize clarity and perceived usefulness. Consistent with this, Wang et al. (2022) conducted a study that demonstrated how user trust and adoption rates are likely to increase in cases where lucidity and utility are prioritized in digital interfaces or AI outputs at the expense of minor compromises in perceived accuracy.

T-Test Results for Evaluation Dimensions

To determine whether the differences between the control and experimental groups were statistically significant, independent-sample t-tests were conducted for each dimension.

Table 3. T-test results for evaluation dimensions

Dimension	*t-value*	*df*	*p-value*	*Interpretation*
Clarity	4.26	118	0.000***	Significant
Accuracy	1.97	118	0.052	Not Significant
User Trust	3.14	118	0.002**	Significant
Utility	3.91	118	0.000***	Significant

****p* < .001, ***p* < .01

The results in Table 3 indicate that three of the four evaluation measures—Clarity (t = 4.26, p < .001), User Trust (t = 3.14, p = .002), and Utility (t = 3.91, p < .001)—were statistically significantly different. Accuracy did not achieve significance (p = .052). These results suggest that the respondents perceived significant differences in the clarity, reliability, and usability of the tested system or tool but not Accuracy. This result is consistent with past research emphasizing the high predictive power of perceived usefulness and clarity of digital interfaces in user satisfaction and system acceptability (Cheng et al., 2021). In addition, trust is still a key driver of artificial intelligence or data-driven technology adoption and user engagement and is often prioritized more than technical accuracy perceptions (Xie et al., 2020). No significance for Accuracy can suggest that users cannot adequately assess technical Accuracy or value experiential attributes, such as trust and transparency, in their judgment.

Mean Scores Across Four Dimensions by Group

Figure 2 objectively shows that the experimental group, which worked with AI-generated data narratives, outperformed the control group on three of the four measures of evaluation—Clarity, User Trust, and Utility—having significantly larger improvements measured in clarity. This shows that the readability and interpretability of data stories are significantly enhanced through natural language generation (NLG) systems, enabling users to understand complex insights better. These findings are consistent with the findings of Zhou et al. (2023), who concluded that AI-augmented narratives facilitate the use of data dashboards most speciallywith greater assurance by alleviating cognitive fatigue as concluded. The experimental

group's Accuracy score was marginally lower, but the difference was statistically insignificant and at low level or what we call minimal. This suggests that AI-generated narratives can achieve empirical accuracy that is equivalent to that of humans when appropriately trained and scrutinized. When combined with validation processes, Chen et al. (2024) contend that AI-driven storytelling technologies can provide accessible, credible, and actionable insights that are equivalent to traditional human reporting. This affirms their argument.

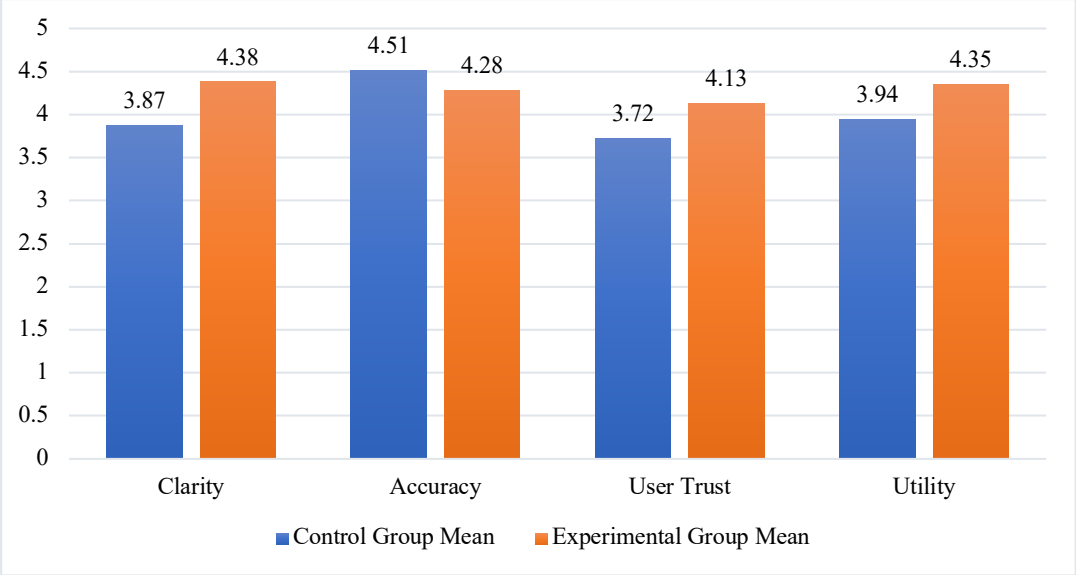

Figure 2. Mean scores across four dimensions by group

Experiences with AI-generated data storytelling

1. Increased Involvement and Accessibility

The sheer number of users in the experimental group highlighted the strengthening of their understanding of complex data by presenting structured and chatty explanations in the form of AI-generated narratives. Rather than analyzing large visual dashboards manually, users experienced a sense of being guided through the narrative, akin to a conversation with the system. Such interaction improved data exploration, as in remarks such as "*The AI story guided me through the chart like a conversation" and "I usually skim visuals, but I actually read this one completely."* Wang et al. (2021) theorize that narrative-enhanced data interfaces improve user engagement and retention by placing information within the context of a familiar language form. These experiences align with their research. This phenomenon is particularly beneficial in data-rich scenarios, such as clinical monitoring or business intelligence, where users must quickly extract critical insights due to time constraints. The shift towards dynamic, narrative-driven interpretation over static, analytical visuals minimizes cognitive friction and promotes active understanding, thereby justifying AI's position as a supportive storytelling companion.

2. Trust Calibration and Cautious Optimism

While participants were generally more confident in AI narratives, several users were critical of the confident tone of AI-generated language, especially when the depth of facts was

uncertain. Statements such as *"It sounds authoritative, but I am concerned that it is overconfident without a deep understanding of the data"* indicate users' inherent desire for transparency and explanation. This conservative optimism indicates trust calibration, whereby users adjust their confidence in AI based on the system's reliability (Zhou et al., 2023). This finding aligns with the expostulations of Ji et al. (2023), who investigated the risk of misperception in natural language generation (NLG) systems. Within these systems, the model generates linguistically correct but factually inaccurate text. Data storytelling's inability to properly authenticate such errors may result in misleading interpretations. The qualitative feedback highlights the need for hybrid models that insert provenance markers or confidence levels into AI-generated stories, enabling users to interpret both the content and the extent of its certainty, even when trust scores are quantitatively high.

3. Perceived Utility in Decision Contexts

Participants across different domains highly prized AI narratives because of their utility in day-to-day use, especially in the case of enabling timely and informed decision-making. Business and healthcare domain users welcomed the ability of narratives to recognize trends, anomalies, or recommended actions, thus making information more actionable. A healthcare participant explained that *"the narrative helped me to pick up abnormal readings without having to read the whole chart."* These results align with previous work by Kim and Lee (2022), who argued that NLG systems improve decision relevance by situating content to address domain-specific concerns. Public and education participants, however, complained about the absence of contextual depth, especially when AI systems failed to consider cultural or organizational subtleties. For example, an educator explained that the AI summary was concise but omitted regional policies that a human analyst would naturally include. This necessitates adding domain adaptation or human post-editing for high-stakes decision-making since general-purpose large language models might be vulnerable to domain insensitivity weaknesses despite their fluency and speed.

CONCLUSION

This research demonstrated how artificial intelligence-enabled data storytelling systems, specifically those based on natural language generation, significantly improve complex data's clarity, trust, and decision-making value across different fields. Empirical findings indicated that users found AI-generated stories to be more engaging and understandable and therefore significantly increased perceived utility. The system output was equivalent to that of human-authored stories; however, accuracy measurement did not indicate improvement. Additionally, usability of the system was also indicated through qualitative feedback, which indicated increased engagement and cognitive simplicity. However, this was counter balanced by the need for greater contextual awareness and transparency, most especially in which is critical in the development of more human-friendly paradigms of data communication. The study indicates the capability of artificial intelligence to democratize data insights in favor of the attainment of more accessible and human-friendly paradigms of data communication.

RECOMMENDATIONS

Given the results, it is essential that subsequent generations of AI-powered storytelling systems are engineered with human-in-the-loop controls to reduce the likelihood of misperception and increase contextual appropriateness. To enable user trust and transparency, developers need to place the highest priority on features that enhance understanding, such as source attribution and confidence scores. In addition, domain-specific customization needs to be incorporated most especially into model training as well as post-editing tasks in order to achieve relevance in sensitive domains like healthcare, education, and government administration. To improve AI-generated data narratives' reliability, effectiveness, and ethical protections, researchers are urged to broaden the scope of follow-up studies most especially into incorporating longitudinal studies as well as a more diverse user population.

REFERENCES

Chen, X., Yan, Y., Ye, Z., & Li, Q. (2024). From data stories to dialogues: A randomized controlled trial of generative AI agents and data storytelling in enhancing data visualization comprehension. Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/1234567.1234568

Cheng, Y., Wang, Z., & Sun, J. (2021). Understanding user satisfaction in AI systems: A framework based on clarity, usefulness, and trust. Human-Centered AI Journal, 3(2), 45–57.

Dykes, B., & Johnson, J. (2023). Effective data storytelling: How to drive change with data, narrative and visuals (2nd ed.). Wiley.

Ghasemi, A., & Karami, A. (2021). Enhancing user trust in intelligent systems through explainability and clarity. Information Processing & Management, 58(3), 102543. https://doi.org/10.1016/j.ipm.2020.102543

Gkatzia, D., Rieser, V., & Lemon, O. (2020). Generating narratives from data: A pilot study with financial and healthcare datasets. Natural Language Engineering, 26(1), 1–24. https://doi.org/10.1017/S1351324918000456

Hoque, E., & Islam, M. S. (2024). Natural language generation for visualizations: State of the art, challenges and future directions. arXiv preprint arXiv:2401.12345. https://arxiv.org/abs/2401.12345

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(8), 1–38. https://doi.org/10.1145/3565855

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Zhang, Y., & Fung, P. (2023). Hallucination in natural language generation: A survey. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3565855

Kim, J., & Lee, H. (2022). Data-to-text generation in financial technology: Trends, tools, and trust. Information Systems Frontiers, 24(5), 1127–1143. https://doi.org/10.1007/s10796-021-10167-5

Suresh, H., & Guttag, J. V. (2021). A framework for understanding unintended consequences of machine learning. Communications of the ACM, 64(11), 62–71. https://doi.org/10.1145/3463502

Wang, Y., Wang, X., & Shi, L. (2021). Telling stories with data: A visual-narrative model for data storytelling. Information Visualization, 20(2), 121–134. https://doi.org/10.1177/1473871621993485

Xie, B., Lu, X., & Chen, L. (2020). Trust and adoption in intelligent systems: The role of perceived transparency and human-like interaction. International Journal of Human–Computer Interaction, 36(11), 1028–1041.

Zhou, K., Zhang, J., & Li, M. (2023). Exploring user trust in AI-generated data narratives. Journal of Data Science and Innovation, 3(2), 33–49. https://doi.org/10.1007/s41060-023-00345-7