Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3706599.3719832

WORK IN PROGRESS

VisQuestions: Constructing Evaluations for Communicative Visualizations

RUIJIA GUAN, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

ELSIE LEE-ROBBINS, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

XU WANG, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

EYTAN ADAR, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

Open Access Support provided by:

University of Michigan, Ann Arbor

PDF Download
3706599.3719832.pdf
06 April 2026
Total Citations: 0
Total Downloads: 348

Published: 25 April 2025

Citation in BibTeX format

CHI EA '25: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

VisQuestions: Constructing Evaluations for Communicative Visualizations

Ruijia Guan
Weinberg Institute for Cognitive Science
University of Michigan
Ann Arbor, Michigan, USA
rjguan@umich.edu

Elsie Lee-Robbins
School of Information
University of Michigan
Ann Arbor, Michigan, USA
elsielee@umich.edu

Xu Wang
Computer Science and Engineering
University of Michigan
Ann Arbor, Michigan, USA
xwanghci@umich.edu

Eytan Adar
School of Information
University of Michigan
Ann Arbor, Michigan, USA
eadar@umich.edu

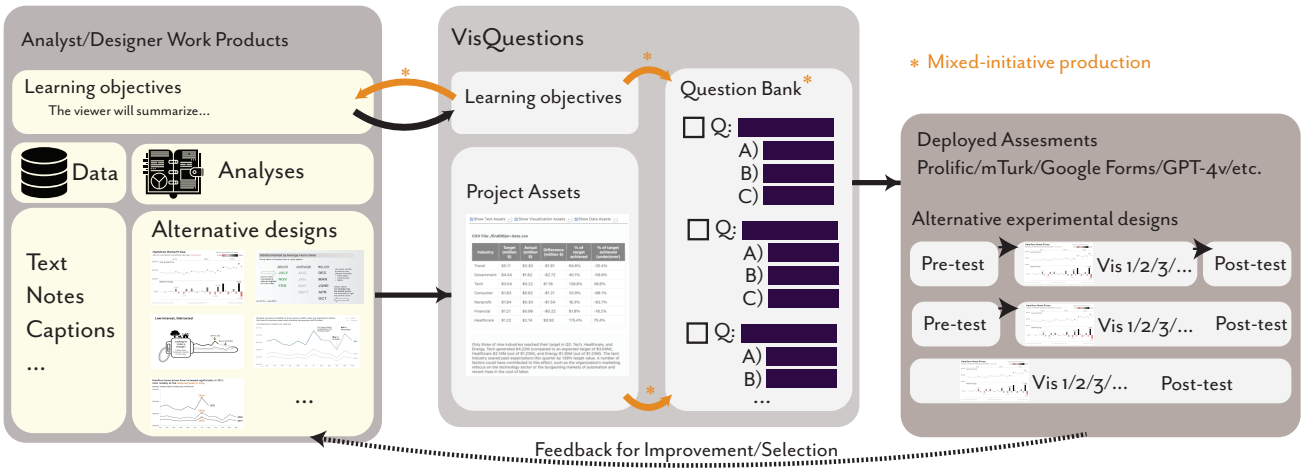


Figure 1: VisQuestions aims to help designers create objective-oriented assessments for their communicative visualizations. The system allows designers with specific communicative intents to evaluate one or more visualizations by generating multiple choice questions in a mixed-initiative interface. LLM-based features help generate effective 'tests' that can then be deployed to different platforms.

Abstract

Communicative visualization designers devote few resources to evaluating their designs. Although general guidelines may guide them, the uniqueness of each communicative visualization makes evaluation challenging. Recent research has suggested that modeling designer intents as learning objectives can guide the construction of evaluative assessments. In this paper, we present *VisQuestions*, a system that streamlines the creation of multiple-choice questions for visualization evaluation. Using source material (e.g., data, notes, prototypes) and contextual data (e.g., captions, annotations, and other text), VisQuestions guides the creation of questions. The system's mixed-initiative features can help create or improve learning objective definitions, questions, and distractors. We report on the performance of the questions by deploying different visualization designs and quizzes through an online experiment. Our findings support the utility of the tool for crafting assessments that can identify how much visualizations support designer intents.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI EA '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3719832

CCS Concepts

• **Human-centered computing → Visualization systems and tools**.

Keywords

Communicative visualization, evaluation, learning objectives

ACM Reference Format:
Ruijia Guan, Elsie Lee-Robbins, Xu Wang, and Eytan Adar. 2025. VisQuestions: Constructing Evaluations for Communicative Visualizations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3706599.3719832

1 Introduction

Communicative visualizations are powerful tools for conveying complex data, enabling designers to present facts, insights, and emotional appeals to diverse audiences. Given the diversity of goals and the one-off nature of many communicative visualizations, designers are not incentivized to evaluate empirically. Rather, they rely on existing guidelines to justify designs [6, 9, 10, 16]. However, these guidelines largely focus on perceptual factors: bar charts are easier to read because we are better at reading lengths; pie charts are better for part-versus-whole comparisons, etc. The goals of communicative visualization often extend beyond perception and readability. Designers do not aim for audiences to only see the data, but also to learn and retain key insights. Recent work suggests that modeling intent as learning objectives (LOs) can better capture designer goals [1, 20], can be more easily translated to assessments, and lead to better visualization selection [21].

The learning objective framework, in particular the cognitive variety [3, 4], describes the desired cognitive impact on the viewer *after* seeing the visualization (e.g., *the viewer will recall the unemployment levels in Texas*). These objectives engage increasingly sophisticated cognitive operations and specific types of knowledge (facts, groupings, procedures, etc.). Critically, design intents stated as LOs are more easily converted into assessments (e.g., *What was the unemployment level in Texas?*). In the same way that student scores help a teacher understand their teaching effectiveness, visualization assessments can capture the 'programmatic' effectiveness of the visualization. Alternative designs can be evaluated in the same way, with higher scoring visualizations preferred. Research has shown that the learning objective framework can help designers select between competing designs [21].

Unfortunately, learning objectives are often difficult to construct [13, 26, 29]. Even if well-formed, objectives themselves may be insufficient. Assessment with an audience is critical, as designer are often unable to intuitively identify the best design [21]. In response, we built VisQuestions, a system to help designers craft learning objectives and assessments in the form of multiple choice questions. Figure 1 illustrates the basic flow of VisQuestions. VisQuestions allows designers to use different work products (e.g., notes, datasets, captions, annotations, design prototypes, etc.) to specify objectives for their visualizations, translate these into assessments, and use these assessments to evaluate one or more visualization alternatives. VisQuestions can optionally utilize AI features to create or improve both objectives and questions (e.g., by creating better distractor choices for questions).

We evaluated VisQuestions using: (1) a small population of design experts (n=7), who constructed a question bank and evaluated system usability; and (2) a large crowd population (n=1096), who answered the questions to identify which visualizations best met the objectives. Our study shows the viability of the LO approach in evaluating visualizations and the limitations of different types of automation. In addition, our work demonstrates the potential of a designer-focused tool for the creation of objectives and assessments. We discuss how non-visualization work products can guide this process and in which situations AI can help. Ultimately, VisQuestions provides a way to translate the theoretical framework of learning objectives into practice.

2 Related Work

2.1 Evaluating Visualizations

Visualization evaluation varies by task [15, 19]. Heuristic evaluations, guided by predefined principles [12, 38, 45], are cost-effective but subjective and limited [11, 41]. Combining heuristics with performance metrics provides richer insights [14]. However, aligning design intent with user outcomes remains a challenge [21]. Automated heuristics tools [44] assist designers, but gaps persist in bridging design intent and audience experience. Recently, the community has developed tools such as EvalBench, ReVisit, and ETK [2, 7, 42]. These aim to lower these barriers for evaluation by simplifying the process of creating studies. However, many tools either emphasize perceptual experiments or are disconnected from the complex intents that underlie communicative visualizations.

The learning objective framework can begin the process of evaluating communication visualization. LOs give a structured approach to defining outcomes across cognitive levels (beyond perception) [1, 25, 39]. The objectives describe what the viewer should be able to *know or do* after viewing the visualization (that is, the *viewer will be able to* < ***the verb***> < ***the verb***>). This knowledge or ability can be directly translated into assessment tasks. Many LO frameworks build upon Bloom's Taxonomy, ranging from simple to complex verbs: Remember, Understand, Apply, Analyze, Evaluate, and Create [3]. Modern versions also specify knowledge descriptors (the noun): facts, concepts, procedures, and meta-cognitive knowledge [3]. This framework supports higher-level evaluation tasks, such as analyzing trends or contrasting patterns, rather than merely identifying values [1, 25, 39]. For communicative visualizations, learning objectives articulate design goals and measure the desired cognitive or emotional impact on audiences [1].

Well-crafted objectives are SMART (Specific, Measurable, Achievable, Relevant, Time-Bound), ensuring consistency and rigor [5, 8, 34]. This approach is particularly suited for communicative visualizations where clarity and impact are critical. VisQuestions is built to help designers translate their intents and goals into learning objectives. From these, the tool supports the creation of questions that satisfy our need for (M)easurability.

2.2 Question Generation

VisQuestions allows a designer to manually craft both objectives and questions. In addition, it provides a number of automated features to help. For this, we integrate a number of AI features, including large language models (LLMs).

Before LLMS, question generation (QG) was often domain specific or highly templatized and constrained [18]. More recent tools facilitate human-AI collaboration, enabling instructors to generate questions in real time [23]. These mixed-initiative approaches offer greater control compared to fully automated systems. Generative AI tools have advanced QG by leveraging LLMs to generate learning

objectives [40], questions [31, 37], and assessments [24, 27]. With VisQuestions, we extend these approaches to support communicative visualization designers. While we model designers as 'teachers' they are not exactly the same. Their goals are not to evaluate students. Additionally, their work products (e.g., data, sketches, etc.) are different than standard teaching materials.

Our initial version of VisQuestions is focused on multiple choice questions (MCQs). Although this imposes some constraints on the assessments we can implement, MCQs are a proven method for effective knowledge assessment [13, 22]. Unlike traditional QG systems reliant on lengthy texts, VisQuestions focuses on shorter inputs, aligning with the needs of communicative visualization assessment. Its mixed-initiative approach allows users to guide model inputs, ensuring relevance and precision. MCQs are well suited for the learning objectives commonly seen in communicative visualizations, especially tasks at lower levels of Bloom's taxonomy, such as recall, classification, and comparison [1, 13]. Standard MCQs, consisting of a question stem, a correct answer, and distractors, need to be well designed [13, 33]. VisQuestions adopts established guidelines [13, 22, 36], ensuring alignment with learning objectives, high quality distractors, and effective generation of questions through LLM prompting and user training.

3 System Details

VisQuestions offers a front-end web application (Javascript and HTML) to generate evaluations and a back-end system (Python) for project management and AI-features. Figure 1 illustrates the conceptual system workflow and connection of components. All AI functionalities in VisQuestions are designed to be mixed-initiative, allowing users to complete tasks independently or override AI suggestions as needed. Key AI features include: automated learning objective suggestions with textual component classifications based on LOs; various forms of automated question and test generation; and distractor improvement.

3.1 User Interface

Figure 2 illustrates the main components of VisQuestions. The system supports multiple simultaneous projects, each allowing the upload of various assets (**A1**), such as textual notes, datasets, and prototype visualizations. The interface is divided into three main columns: the left column for creating and editing learning objectives, the center for accessing project assets, and the right (**A3**) for creating MCQs. A widget at the top (**B1**) highlights selected datasets or text as input for the MCQ generator or to guide the designer's focus. When selecting text in the assets column, a modal dialogue provides a summary to refine inputs for question generation (**A2**).

The question creation form lets the designer associate questions with objectives (**B2**), specify target answers (**B3**), and generate one or more questions (**B4**). Users can modify generated content (e.g. stems and distractors) and regenerate elements as needed. The finalized questions are added to the database (**B5**) and linked to specific visualizations (**B6**), ensuring alignment between objectives and visualizations. A deployment window (not shown) facilitates the distribution of the assessment, ensuring that all objectives are covered with targeted questions.

To clarify the use of these components, we introduce Sasha, a visualization designer, who has been hired to create a visualization for an article on housing affordability. Since there is space for only one visualization, she needs to take care that it satisfies the client's key objectives. To do this, she must identify and formulate these as learning objectives, generate assessments, prototype visualizations, and evaluate their effectiveness. Figure 2 represents this overall process in the VisQuestions interface.

She uses VisQuestions to create a new project (**C**). Sasha uploads the client's notes and allows the system to generate and classify learning objectives. She accepts some objectives and adds her own. Next, she selects a target objective, such as recalling which month had the highest sale prices. She identifies the key insight in her notes and selects the relevant sentence. VisQuestions summarizes the insight, and she inputs the correct answer ('October') into the interface (**D**). The system then generates assessment questions, including distractors, by analyzing both the dataset and project notes. Sasha can edit or regenerate these as needed. After creating at least one assessment per objective, she uploads prototype visualizations and selects which questions to associate with each. VisQuestions supports different study deployment environments, such as Prolific and Google Forms. Participants view the visualization, then answer the questions. With the collected responses, Sasha analyzes the findings, discusses them with the editor, and iterates on or selects the final visualization design.

3.2 Objective Generation

Designers can manually add their own objectives to their projects. VisQuestions helps by guiding them through examples and training. Additionally, VisQuestions automatically proposes LOs from project work products (e.g., visualization text, captions, annotations, notes, etc.) using tailored LLM prompts. The pipeline prompts the LLM to behave as a professional course designer tasked with generating up to seven unique objectives, following the format: *the viewer will* <***verb***> <***noun***>. Verbs like apply or recall are encouraged, while vague ones like understand are avoided [1]. The generated objectives can be manually edited. When text is provided as work products, VisQuestions classifies sentences likely related to visualization content under relevant learning objectives. This is achieved by tokenizing at the sentence level, and using an embedding model [30] to check for similarity. Sentences that do not seem to fit, are placed in an 'other' category (they can be moved or can inspire a new LO). This process helps to iteratively create and refine the LOs that reflect that designer's visualization intent.

3.3 Test Generation

MCQs in VisQuestions consist of a question (stem), a correct answer, and distractors. Users can create all parts manually, or rely on VisQuestions and provide text 'snippets' to guide question generation. Snippets can be anything: short descriptions of important insights found in analysis (e.g., 'unemployment in TX was 5%'); the potential title or subtitle of visualization the designer is building; an annotation from the visualization; or even the intended target answer (e.g., '5%'). VisQuestions uses these, in conjunction with LOs and other work products, to create LO-aligned assessments.

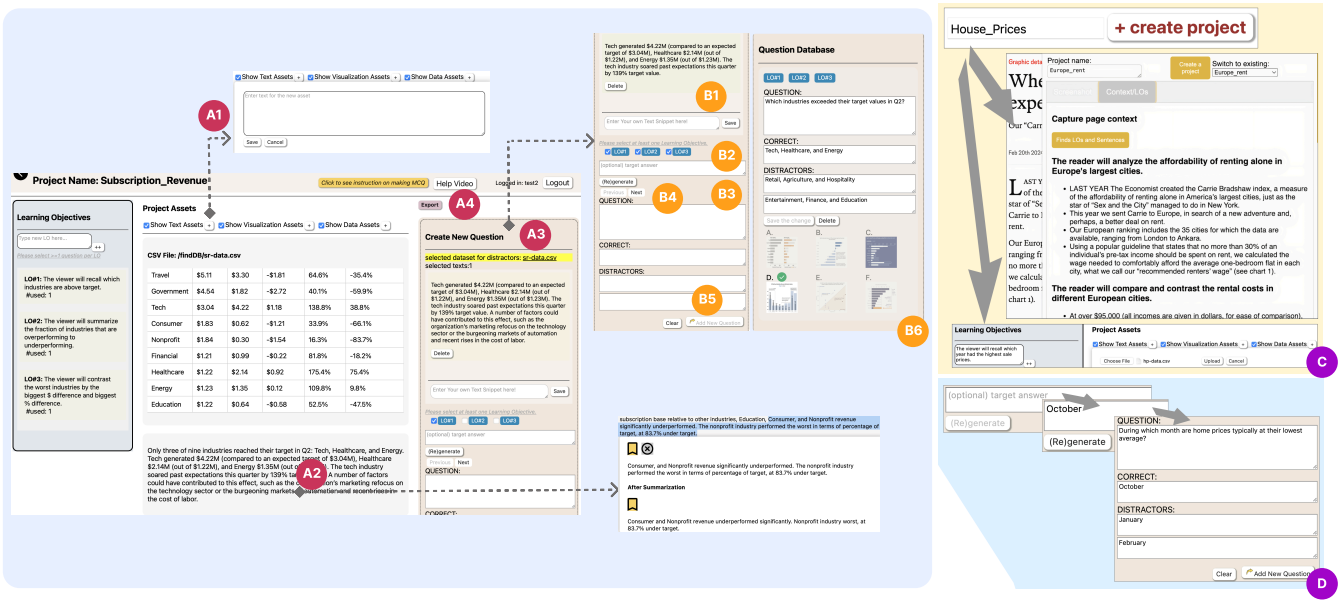


Figure 2: An overview of VisQuestions' interface and example walkthrough.

VisQuestions also supports a more complete automation by generating multiple questions simultaneously given a set of learning objectives. In this case, we utilize a different prompt to generate a cohesive 'exam' as a test. The prompt explicitly has the LLM output the reasoning for selecting a specific question in relation to the learning objective (leveraging chain-of-thought prompts [43]). For example, for an article about online dating[1], VisQuestions suggested a learning objective of: *The reader will examine demographic patterns in filtering choices among online daters.* It also identified the sentence, *"Women block 70% of potential matches, compared with 55% for men, mostly because they tend to exclude users who are shorter or younger"* as being related to this objective. A question it generated based on this information was: *What is the approximate percentage of potential matches blocked by women and men, respectively?* with the choices: *A) Women: 55%, Men: 70% ; B) Women: 70%, Men: 55%; C) Women: 86%, Men: 48%.* Additional details and prompt designs are provided in the Appendix A.

3.4 Distractor Improvement

VisQuestions addresses issues like overly simplistic or irrelevant distractors. For example, given the snippet "Colombia had the highest Gini coefficient at 54.2 in 2021," a question might be: "Which country had the highest Gini coefficient in 2021?" Ideally, the LLM will generate South American distractors such as Brazil or Peru. However, it could also include irrelevant options like France, making the question trivial. Adding context to snippets (e.g., specifying 'South America') or using user-provided titles (e.g., "Most unequal Latin American countries by Gini coefficient") biases the LLM to generate more relevant distractors.

When a visualization's source data is available, VisQuestions uses a SQL-based distractor builder. The builder works in four steps. First, the database or CSV file is indexed in an embedded database (we use a SQL embedder from Langchain). Second, we utilize the language model to transform the question (stem) text to a SQL query. For example, our test question might become: `SELECT _percentage FROM tempdb_1 WHERE _country = 'Brazil' AND _response = 'Better off' LIMIT 5;`. When the correct answer is returned (e.g., 72), we proceed with a third step: modifying the SQL to produce the wrong answer. Here, each SQL condition is transformed into its inverse, one at a time (e.g., each = is transformed to a != and vice versa, each >= to <, and so on). Additional details on this process are provided in Appendix D.

4 Evaluation and Analysis

To evaluate VisQuestions, we tested whether designers were able to generate questions that allowed them to assess how well visualization designs achieve learning objectives. Our main study was conducted in two steps: a design study with professional data visualization designers (n = 7) to evaluate the usability of the system; and a large-scale study (n = 996) in Prolific to evaluate the effectiveness of MCQ generated by designers in the first stage. A good test should help designers differentiate visualizations and measure their contribution to the envisioned learning objectives. We achieved this goal by comparing subjects' performance in MCQs for different visualization-LO pairs.

4.1 Materials

We modeled two visualization projects based on two design challenges from Storytelling With Data (SWD)[2], a community platform

[1]https://www.economist.com/graphic-detail/2023/03/22/online-daters-are-less-open-minded-than-their-filters-suggest

[2]https://www.storytellingwithdata.com/

focused on creating effective communicative visualizations [16]. The challenges involved redesigning visualizations based on dataset: (1) House Prices (HP) [17]–Home sales data over three years for a fictional town. (2) Subscription Revenue (SR) [32]–Quarterly target vs. actual subscription data across industries. Our research team and the SWD team selected six representative submissions with diverse design decisions per topic based on the SWD votes and the diversity of style. We developed three matched learning objectives for each topic and ensured equivalent complexity between projects [1, 3]. We also created descriptive text summaries of the data findings.

4.2 Analysis of System Usability

We recruited nine members of the Data Visualization Society, requiring at least one year of visualization experience and a minimum age of 18. Two participants were excluded as pilot participants, leaving seven participants with an average of six years of experience. The details of the position and experience are in Appendix B. Each participant was compensated with a $60 gift card. Participants were provided with the learning objectives, text, visualizations, and datasets via a modified VisQuestions interface.

In a 90-minute individual Zoom session, participants received an instruction sheet, two practice problems on MCQ design, and two brief videos on the use of VisQuestions. They then created questions for one project using the non-AI version of VisQuestions, followed by a second project with AI features enabled. Project order was counterbalanced to minimize bias. The task was to create a set of multiple-choice questions (i.e., a test) given a set of learning objectives to contrast six different visualization designs. After creating MCQs, participants linked visualizations to their questions. Each session ended with an ease of use rating [35], a usability survey [28], and a semi-structured interview on their experience, opinions on AI integration, and suggestions for improvement.

Participants rated AI-assisted question generation as easier than manual generation (6 of 7 participants), with a reported decrease in difficulty from manual ($mean = 4$) to AI-assisted ($mean = 3.5$). Usability scores indicated that the system was relatively usable ($mean = 68.93$, $SD = 24.15$) on a scale of 0 to 100. AI features improved efficiency, reducing the average time per question ($mean =$ 7.19 to 6.53 minutes) and the final question time ($mean = 5.48$ to 3.80 minutes). Due to the small sample size, these results are indicative, but not statistically significant. Two participants used AI-generated outputs as a foundation. Two highlighted the value of distractor generation, and four appreciated AI's ability to provide inspiration and alternative perspectives. However, five participants found the interface somewhat cluttered.

4.3 Analysis of Generated Questions

4.3.1 Study Procedure. To validate the constructed tests, we ran a large-scale study of 1096 participants on Prolific to evaluate how well each question differentiated various visualization designs. Prolific participants answered six random questions (one for each of the 3 objectives × 2 projects) in a pre-test and post-test. Pre-test questions were presented individually, followed by the same questions in the post-test paired with randomly selected visualizations (see Appendix C for an example). This setup ensured consistent comparisons while minimizing priming effects. Each House Prices (HP) LO had three AI-generated questions, and each Subscription Revenue (SR) LO had four. To enhance test robustness and improve generalizability, in addition to the six visualizations per project from the first-stage evaluation, we included one additional "optimal" visualization from SWD staff. This ensured that the test also runs on high-quality design unseen by MCQ creators, who might be biased by seeing the initial set of six visualizations. In addition to the question sets from the first stage study, we had the LLM construct a complete set of multiple-choice questions without human intervention. These questions were created using an 'exam generation' prompt based on learning objectives (LOs) and textual content, with raw data refining distractors. This design gives us 3 conditions, *human, AI, and mixed-initiative*.

4.3.2 MCQ Reliability. To evaluate the quality of the questions, we began by looking at the quality of the answers and distractors. In an ideal scenario, selecting the correct answer and distractors in the pre-test should all be equally likely (i.e., a probability of 1/3 here). Comparing the correct ratio in the pre- and post- tests allows us to assess the learning effect brought by the visualization. We find that human-generated and AI-generated questions have the best distractors, with the correct answer chosen 41% of the time in the pre-test, which is slightly over random chance. A χ^2 test of independence comparing choices of AI-generated questions to human-generated questions is not significant, $\chi^2(2, N = 4196) = 4.409, p = 0.11$. This suggests that AI-generated distractors are as good as human ones. The mixed-initiative generated questions are more guessable, with the correct answer chosen 47% of the time in the pre-test. We find that there is a significant difference between the human and mixed-initiative distribution of choices $\chi^2(2, N = 4704) = 19.098, p < .001$.

We next measured the internal consistency across question items within a test with Cronbach's alpha. For each learning objective of a task, there are 10-11 multiple-choice questions created of three types: human, AI, and mixed-initiative. Since they are all designed for one specific learning objective, they should yield a consistent ranking of the seven visualizations. We first compute an overall Cronbach's alpha for the average score of the seven visualizations on these 11 questions, and then re-compute Cronbach's alpha when this question is dropped from the set. If the Cronbach's alpha increases when the item is dropped, the question is indicated to be inconsistent with the rest of the questions. In total, 4 out of 25 human generated questions (16%), 9 out of 24 mixed initiative generated questions (38%), and 8 out of 21 AI-generated questions (37%) were identified as inconsistent question items. This finding highlights the importance of human involvement in communicative visualization. Although every designer manually modified the MCQs before deploying them, it seems that the AI would introduce bias and negatively affect the result.

4.3.3 Visualization Performance. Figure 3 displays a heatmap of test results for each condition. Rows represent the visualizations and columns are the learning objectives. Each cell is colored by the average score for questions assigned to that objective. The rows are ordered by the total score across the three objectives. Thus, the most successful visualization are at the top (visualization ids, which are anonymized, are numbered by this rank). The designer can quickly

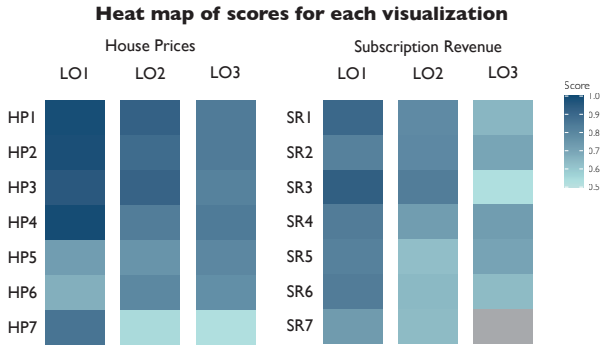


Figure 3: A heatmap of assessment scores for the two experimental conditions. Each row represents a visualization (e.g., HP1, HP2, SR1, SR2, etc.) and each column a learning objective (e.g., LO1). Cells aggregate the scores for all questions in that objective. Visualizations are ordered by total score across all objectives. For example, visualization HP1 does well across the three objectives, while HP7 does well for LO1 but not LO2 and LO3.

determine which visualization performs the best (assuming that all LOs are equally important). More critically, the designer would be able to determine areas of failure and implement fixes. For example, SR3 would seem to perform well (even better than SR1) with the exception of LO3. The designer may implement changes to that specific visualization to improve performance on LO3. Similarly, the designer can rapidly assess tradeoffs. SR3 may be preferable to SR1 if LO3 is not particularly important. The designer may also be able to identify 'bundles' of visualizations to optimize performance. For example, displaying HP7 *and* HP6 together may yield a similar or better performance to HP3.

5 Discussion, Limitations, and Future Work

In this paper, we explored how a mixed-initiative system could help designers evaluate their visualizations by creating MCQs based on the learning objective framework. Our results support the usability of the system and demonstrate how the MCQs result can help evaluate visualization performance given designers' intents. The current work would benefit from further evaluation and system improvements. First, the evaluation has mainly focused on question generation, while features such as learning objective generation need further study. An end-to-end evaluation, where designers create visualizations, could offer valuable insights. Second, VisQuestions struggles with complex inputs, particularly in generating appropriate distractors for parallel components in lengthy text. For example, given the input: "Google LLC is an American multinational technology company focusing on online advertising, search engine technology, cloud computing, computer software, artificial intelligence, and consumer electronics," the system may generate unrelated distractors like "Network Security, Video Conferencing." Addressing this may require enhanced NLP techniques, deterministic checks, or advances in LLM models.

Based on the two-stage evaluation, we also detected several limitations of our current work. Firstly, in our usability evaluation, we detected that the AI-assisted version only brings a small improvements in the difficulty (*4 vs. 3.5*) and time-taking scores (*5.48 vs. 3.80*). Secondly, we were somewhat surprised by the worse MCQ performance in the mixed-initiative case. One interesting future direction is to investigate how AI reliance might negatively affect performance. These results may also be due to a learning curve to effectively use automation. It is also worth noting that many participants reported that the interface is too crowded and hard to use. Participants' feedback on needing more support during the AI-enabled sessions suggests the necessity of onboarding and workflow refinement.

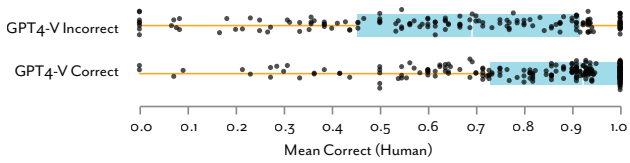


Figure 4: A visualization of GPT-4v model correctness versus average human score. Each point represents the average score for a unique visualization/question pair. The underlying boxplot captures the overall distribution.

Inspired by the feedback above, we are developing a new version of the system, VisQuestions 2.0. The new version fixes various issues with automation, provides a simpler user-friendly interface, and also improves the adaptability of the system. Firstly, the updated interface introduces a streamlined "wizard" for uploading assets, specifying the number and type of generated questions, and exporting projects as ReVISit experiments [7] to support more diverse applications. Secondly, we are aiming to support non-MCQ question generation and evaluation of affective (rather than cognitive) LOs. This would expand the range of projects that VisQuestions could help with.

In addition to interface updates, VisQuestions 2.0 introduces AI test-taking to overcome the challenges of conducting user experiments. This feature leverages advanced LLMs (e.g., OpenAI's GPT-4V, Google's Gemini, Anthropic's Opus) to simulate human responses to questions paired with visualizations. This enables rapid iterative testing without requiring full-scale experiments. While a comprehensive evaluation of this feature is still needed, early results are encouraging. In initial trials, GPT-4V achieved a 61% accuracy rate (out of 420 answers), with human-graded scores significantly higher when the model's answers were correct ($p < 0.001$, two-tailed t-test, see Figure 4).

6 Conclusion

In this paper, we present VisQuestions, a mixed-initiative system developed to create evaluations for communicative visualizations. VisQuestions supports designers in creating custom assessments by which they can contrast, select and improve their visualizations. Using LLMs-based features, VisQuestions allows users to easily create multiple-choice questions. Generated questions can be used

in deployed surveys and evaluations that validate a design's effectiveness relative to the designer's intent (expressed as learning objectives). By collecting text, visualizations, data and other information, VisQuestions can construct well-formed multiple-choice questions and distractors. Our system offer a way for designers to leverage the learning objective and assessment 'model' in communicative visualizations.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant No. IIS-1815760. We sincerely appreciate the support of the Storytelling With Data team, whose feedback greatly contributed to the preparation and evaluation of the visualizations.

References

[1] Eytan Adar and Elsie Lee. 2020. Communicative visualizations as a learning problem. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 946–956.

[2] Wolfgang Aigner, Stephan Hoffmann, and Alexander Rind. 2013. Evalbench: A software library for visualization evaluation. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 41–50.

[3] Lorin W. Anderson and David R. Krathwohl (Eds.). 2001. *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives (Complete Edition).* Longman.

[4] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1956. *Taxonomy of educational objectives, handbook I: The cognitive domain.* Vol. 19. New York: David McKay Co Inc.

[5] Alyxander Burns, Cindy Xiong, Steven Franconeri, Alberto Cairo, and Narges Mahyar. 2021. Designing with pictographs: Envision topics without sacrificing understanding. *IEEE transactions on visualization and computer graphics* 28, 12 (2021), 4515–4530.

[6] Alberto Cairo. 2012. *The Functional Art: An introduction to information graphics and visualization.* Pearson Education.

[7] Yiren Ding, Jack Wilburn, Hilson Shrestha, Akim Ndlovu, Kiran Gadhave, Carolina Nobre, Alexander Lex, and Lane Harrison. 2023. reVISit: Supporting Scalable Evaluation of Interactive Visualizations. In *2023 IEEE Visualization and Visual Analytics (VIS).* 31–35. https://doi.org/10.1109/VIS54172.2023.00015

[8] George T Doran. 1981. There's a SMART way to write managements's goals and objectives. *Management review* 70, 11 (1981).

[9] Stephanie D. H. Evergreen. 2019. *Effective data visualization: The right chart for the right data.* SAGE Publications. 1–11 pages.

[10] S. Few. 2004. *Show Me the Numbers: Designing Tables and Graphs to Enlighten.* Analytics Press.

[11] Camilla Forsell. 2012. Evaluation in information visualization: Heuristic evaluation. In *2012 16th International Conference on Information Visualisation.* IEEE, 136–142.

[12] Camilla Forsell and Jimmy Johansson. 2010. An Heuristic Set for Evaluation in Information Visualization. In *Advanced Visual Interfaces* (Roma, Italy) *(AVI '10).* ACM, New York, NY, USA, 199–206. https://doi.org/10.1145/1842993.1843029

[13] Thomas M Haladyna. 2004. *Developing and validating multiple-choice test items.* Routledge. 3–148 pages.

[14] Marti A Hearst, Paul Laskowski, and Luis Silva. 2016. Evaluating information visualization via the interplay of heuristic evaluation and question-based scoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 5028–5033.

[15] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. 2013. A Systematic Review on the Practice of Evaluating Visualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2818–2827. https://doi.org/10.1109/TVCG.2013.126

[16] Cole N. Knaflic. 2015. *Storytelling with Data: A Data Visualization Guide for Business Professionals.* Wiley.

[17] Cole Nussbaumer Knaflic. 2021. Oct 2021 | Makeover Challenge. https://community.storytellingwithdata.com/challenges/oct-2021-makeover-challenge Accessed: 2025-01-01.

[18] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education* 30, 1 (2020), 121–204.

[19] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536. https://doi.org/10.1109/TVCG.2011.279

[20] Elsie Lee-Robbins and Eytan Adar. 2022. Affective Learning Objectives for Communicative Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1–11.

[21] Elsie Lee-Robbins, Shiqing He, and Eytan Adar. 2021. Learning objectives, insights, and assessments: How specification formats impact design. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 676–685.

[22] Jeri L Little, Elizabeth Ligon Bjork, Robert A Bjork, and Genna Angello. 2012. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological science* 23, 11 (2012), 1337–1344.

[23] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. Readingquizmaker: A human-nlp collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–18.

[24] Xinyi Lu and Xu Wang. 2024. Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. (2024), 16–27. https://doi.org/10.1145/3657604.3662031

[25] Narges Mahyar, Sung-Hee Kim, and Bum Chul Kwon. 2015. Towards a taxonomy for evaluating user engagement in information visualization. In *Workshop on Personal Visualization: Exploring Everyday Life*, Vol. 3.

[26] M. David Miller, Robert L. Linn, and Norman E. Gronlund. 2012. *Measurement and assessment in teaching.* Pearson Education.

[27] Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. 2023. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European Conference on Technology Enhanced Learning.* Springer, 229–245.

[28] Nielsen Norman Group. 2024. Beyond the NPS: Measuring Perceived Usability with the SUS, NASA-TLX, and the Single Ease Question After Tasks and Usability Tests. https://www.nngroup.com/articles/measuring-perceived-usability/. Accessed: 2024-03-31.

[29] Jay Parkes and Dawn Zimmaro. 2016. *Learning and assessing with multiple-choice questions in college classrooms.* Routledge. 45–57 pages.

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[31] Liam Richards Maldonado, Azza Abouzied, and Nancy W Gleason. 2023. ReaderQuizzer: Augmenting Research Papers with Just-In-Time Learning Questions to Facilitate Deeper Understanding. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing.* 391–394.

[32] Elizabeth Ricks. 2020. DEC 2020 | critique then (re)create. https://community.storytellingwithdata.com/challenges/dec-2020-critique-then-recreate

[33] Michael C. Rodriguez. 2005. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice* 24, 2 (2005), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.2005.00006.x

[34] Michael Saenz, Ali Baigelenov, Ya-Hsin Hung, and Paul Parsons. 2017. Reexamining the cognitive utility of 3D visualizations using augmented reality holograms. In *Proc. 2017 IEEE VIS workshop on immersive analytics: exploring future visualization and interaction technologies for data analytics.* 5.

[35] Jeff Sauro and Joseph S Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 1599–1608.

[36] Daniel L Schwartz, Jessica M Tsang, and Kristen P Blair. 2016. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them.* WW Norton & Company.

[37] Machi Shimmei, Norman Bier, and Noboru Matsuda. 2023. Machine-Generated Questions Attract Instructors When Acquainted with Learning Objectives. In *International Conference on Artificial Intelligence in Education.* Springer, 3–15.

[38] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization.* Elsevier, 364–371.

[39] Fabian Sperrle, Astrik Jeitler, Jürgen Bernard, Daniel Keim, and Mennatallah El-Assady. 2021. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics* 100 (2021), 93–105.

[40] Pragnya Sridhar, Aidan Doyle, Arav Agarwal, Christopher Bogart, Jaromir Savelka, and Majd Sakr. 2023. Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives. *arXiv preprint arXiv:2306.17459* (2023).

[41] Melanie Tory and Torsten Moller. 2005. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications* 25, 5 (2005), 8–11.

[42] Terece L Turton, Anne S Berres, David H Rogers, and James P Ahrens. 2017. ETK: An Evaluation Toolkit for Visualization User Studies.. In *EuroVis (Short Papers).* 43–47.

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837.

[44] Ying Zhu and Julia A Gumieniak. 2021. Computer-Assisted Heuristic Evaluation of Data Visualization. In *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II*. Springer, 408–420.

[45] Torre Zuk and Sheelagh Carpendale. 2006. Theoretical analysis of uncertainty visualizations. In *Visualization and data analysis 2006*, Vol. 6060. SPIE, 66–79.

A Generating Questions Details

VisQuestions will construct a prompt (using a few-shot learning approach) to generate the question (with or without distractors). Additional information (e.g., the learning objective or other text items) can be fed in as context to better steer the generation. A complete response might look like the following text block. Depending on which elements the user wants generated, different parts are masked out and the LLM is tasked with completing the missing information (e.g., the `CORRECT` and `DISTRACTOR` parts might be omitted).

```
SNIPPET: Unemployment in the euro area reached
a record high in early 2013.
QUESTION: What was the record high unemployment
rate in the euro area in 2013?
CORRECT: 12%
DISTRACTORS: 15%; 19%
```

To better direct the LLM to produce good questions, we manually constructed 69 examples using articles and visualizations from the Economist's Graphic Detail site[3]. Articles on the site are short, targeted pieces that are anchored on one or more visualizations. The authors manually selected sample snippets from the article, generated questions, answers, and distractors for each. The questions were a combination of standard multiple stem (e.g., *When did X happen?*), True/False (e.g., *True or False, X happened in 2019.*), and fill in the blank (e.g., *X happened in* _____). All questions were designed to be likely answered with a visualization. Article snippets were selected based on what we believed the article and visualizations focused on. After generating and reviewing 36 questions, we also began to use LLM to generate new questions for us using the original questions as examples. These were manually filtered and corrected as needed and integrated into the example dataset.

When providing a target target answer, the text does not need to match a word or phrase in the original snippet. We have found that modern LLMs (GPT-3 and above) do a reasonable job working outside of these constraints. For example, taking the following snippet: *Over the last decade we have seen Team A dramatically improve. In contrast, Team B has remained unchanged*, we can specify the target answer as *Team A went up, Team B stayed the same.* The target answer is neither a direct quote nor a specific phrase in the initial snippet. Nevertheless, the LLM can extrapolate from this, and the resulting question is: *How did Team A and Team B change over the last decade?*. Generated distractors include: *Team A stayed the same and Team B went up*; *Both teams stayed the same.*

When providing examples to LLM through the prompt, we do not provide the complete set of 69. Instead, our current implementation uses a *maximum marginal relevance example selector* to select the training examples for a given snippet input. This selects examples based on similarity of the given input snippet to the example (based on cosine similarity of the embeddings), but also penalizes new examples that are too similar to those already selected. We select three examples in this way to become part of the prompt.

[3]https://www.economist.com/graphic-detail

B System Usability Evaluation Participant Profiles

Index	Age	Gender	Position	Year of Visualization Experience
1	25-34	Female	Senior Software Developer	14
2	55-64	Female	Lecturer	8
3	18-24	Female	PhD Student	3
4	25-34	Female	Senior Research Assistant	8
5	35-44	Female	Startup Co-founder	14
6	18-24	Female	Data Visualization Developer Fellow	2
7	25-34	Female	Graphic, Information and Web Designer	4
8	25-34	Male	Senior Business Intelligence Engineer	6
9	35-44	Female	Freelance Designer	5

Table 1: Participant Demographics and GenAI Experience

C Analysis of Generated Question Example

An example test on Prolific:

Pre-test:

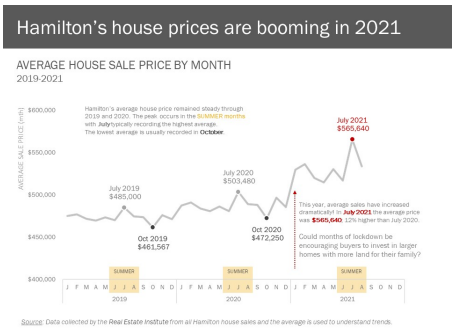
What year had the highest average monthly home sale prices?

A. 2019
B. 2020
C. 2021

Post-test

What year had the highest average monthly home sale prices?

A. 2019
B. 2020
C. 2021



D SQL Fuzzing

Our SQL-based distractor builder works in four steps. First, the database or CSV file is indexed in an embedded database (we use a SQL embedder from Langchain). The index contains both a method to represent the database and a set of prompts that can transform natural-language queries into SQL queries (which are subsequently executed). Every database file in a project is indexed by the system. In our Brazil example, the dataset contains three columns: country, percent, and a categorical survey response (better off; about the same; worse off). In a second step, we ask the language model to answer the generated question using the database. The question above is transformed to: `SELECT _percentage FROM tempdb_1 WHERE _country = 'Brazil' AND _response = 'Better off' LIMIT 5;`. The correct answer (72) is returned when this query is executed. This tells us that the generated query can target the correct 'cell' in the database.

Our next step is to transform this 'correct' SQL statement into an 'incorrect' one to generate alternative answers. Specifically, we replace each condition in the SQL statement with its inverse, one at a time (i.e., each = is transformed to a != and vice versa). In our example above, this leads to two SQL statements: `SELECT _percentage FROM tempdb_1 WHERE _country != 'Brazil' AND _response = 'Better off' LIMIT 5;` (i.e., find the response percent for countries *other than Brazil* for 'Better off') and `SELECT _percentage FROM tempdb_1 WHERE _country = 'Brazil' AND _response != 'Better off' LIMIT 5;` (i.e., find the response percent for the other answers–about the same, and worse off–from Brazil). The same procedure is applied to <=, <, >= and > operators. The result of executing these modified SQL queries provides us a range of possible distractors, all coming from real data and are answers to a subtly modified version of the original question.

In some situations, we are unable to find a correct SQL query in the second step. Here, we have devised a fall-back prompt that looks for other values in the column where the correct answer might appear. For example, VisQuestions generates the following query for the example above: *What other unique values appear in the same column as value '72'?*. The resulting SQL is `SELECT DISTINCT _percentage FROM tempdb_1 WHERE _percentage != 72 LIMIT 5;`. The result are up to five plausible distractors based on real data (from which a subset is selected).