Eurographics Conference on Visualization (EuroVis) 2026
N. Andrienko, M. Chen, and B. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 3

Guardrail Selection in Line Charts to Contextualize Persuasive Visualizations

K. A. Nadib[1] M. Kogan[1] A. Lex[2,1] M. Lisnic[3]

[1]University of Utah, USA
[2]Graz University of Technology, Austria
[3]Worcester Polytechnic Institute, USA

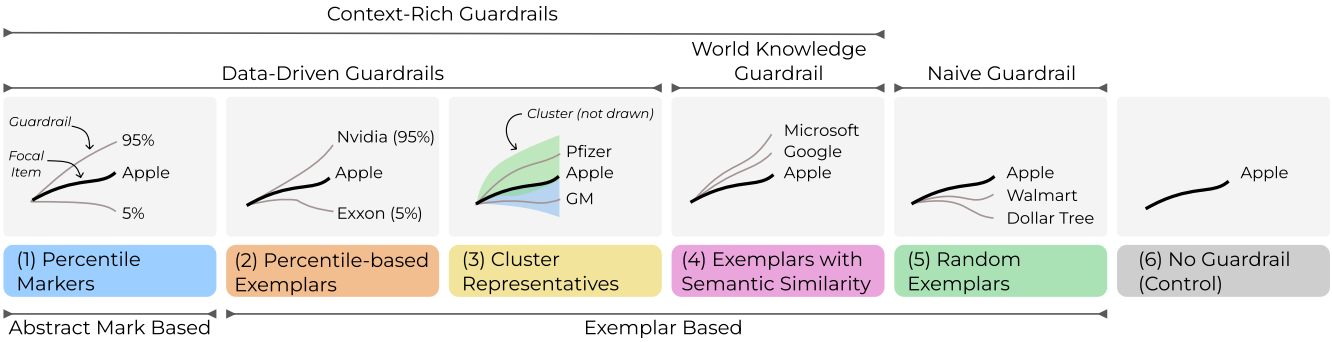


Figure 1: The five guardrails and the control condition evaluated in our experiment. We test different versions of statistical guardrails: (1) percentile markers, i.e., explicitly showing percentiles of the whole dataset over time; (2) percentile-based exemplars, i.e., showing concrete items that are close to a percentile value, (3) cluster representatives, i.e., showing items central to the clusters of the dataset, (4) exemplars with semantic similarities, i.e., based on higher-level knowledge about the data items; (5) randomly drawn exemplars, and (6) no guardrails.

Abstract

Charts used for persuasion can easily veer into being outright misleading when, for instance, cherry-picked data is paired with a deceptive caption, as is commonly encountered on social media. The rise of interactive time-series data explorers for hotly debated topics makes such framing easy to produce and spread. Post-hoc interventions like fact-checking often arrive too late and suffer from persistence of belief. Prior work suggests that guardrails, in the form of contextual comparison lines embedded directly into charts, can reduce these effects. We propose and evaluate a practical set of guardrail sampling strategies for implementing such contextual lines in real systems. In a preregistered mixed-design study with two real-world scenarios (COVID-19 and Stocks), participants viewed persuasive charts with different sets of guardrails and reported trust, estimated rank in the dataset, expressed their perceived completeness of context, as well as subjective preference for different tasks. Across scenarios, guardrails improved trust, accuracy of performance judgments, and perceived completeness of context compared to the control. Taken together, the study offers practical guardrail sampling methods, evidence of their contextual benefits, and insights into participants' preferences.

CCS Concepts

• ***Human-centered computing*** → ***Visualization design and evaluation methods; Empirical studies in visualization; Visualization theory, concepts and paradigms; Empirical studies in HCI; Empirical studies in collaborative and social computing;***

1. Introduction

Modern interactive time-series explorers make it easy for both novice and expert users to produce a factual, persuasive chart in seconds, before captioning it and sending it off over text or social media. Public health-related data exploration platforms, such as the Our World in Data COVID-19 Explorer [Our20] and the Johns Hopkins dashboards [Joh20], offer powerful controls for date windows, smoothing, and metric selection to inform users' health decisions. Finance platforms such as TradingView [Tra25], Yahoo Finance [Yah25], and Google Finance [Goo25] similarly provide range, overlays, and comparators for easy customization of views pertaining to personal finance choices. These platform affordances are essential for domain-specific analyses for both experts and lay

© 2026 The Author(s). Computer Graphics Forum published by Eurographics — The European Association for Computer Graphics and John Wiley & Sons Ltd.
This is an open access article under the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

people alike, but they also make it easy to produce and disseminate misleading charts [LPLK23]. The same affordances that allow users to zoom in and focus on data points of their interest when the interface is used for exploration purposes may also lead to intentional or unintentional cherry-picking that may mislead both the user and their audiences [LCKL25].

Deceptive visualizations and, consequently, visual misinformation are a pressing topic that has been attracting a growing body of work in the visualization community [LYI*21, LPLK23, LGS*22, PRS*15]. There are many tools at visualization authors' disposal that assist in crafting misleading visualizations: framing and narrative choices in visualization expressed via titles, captions, selected time windows, axis ranges, and curated comparators may strongly shift reader interpretation and judgment, even when the underlying data are held constant [KLK18, KSA21, LPLK23]. Visualizations are also persuasive [PMN*14] and are commonly shared on social media. Once they go viral, "belief echoes"—residual attitudinal effects after correction—can persist even after misinformation is clearly retracted [Tho16]. This reveals a pressing need to **design and deploy in-chart preemptive interventions that meet readers at the point of viewing and interpretation rather than relying solely on downstream fact-checking**.

We initiated the work to design such preemptive interventions by previously introducing *visualization guardrails*: auxiliary context embedded directly into data explorer tools so that a *focal item* (the item a user has chosen to display) cannot be read or shared in isolation [LCKL25]. In our original evaluation of guardrail designs, we found that simple guardrails that match the visual language of the main data successfully encourage skepticism while not interfering with the main chart [LCKL25]. However, the question of *what* data to show and how to effectively sample it across data domains remains. For example, when exploring COVID-19 cases, it may matter which other countries are shown as guardrails: when a user is trying to understand Sweden's COVID-19 response, showing similar countries in geographic proximity might make sense. When exploring the stock price of General Motors, showing exemplars at the 10th, 50th, and 90th percentiles may help situate the data against the larger market. This paper aims to fill this gap and take guardrails from design space to practice.

To identify which data is best displayed as a guardrail relative to a single *focal item*, we selected five strategies from a larger set of options we considered, also illustrated in Figure 1: (1) Percentile Markers—markers representing the percentiles of the dataset; (2) Percentile-based Exemplars—example items that are close to statistical percentiles; (3) Cluster Representatives—example items that represent a cluster of the data; (4) Exemplars with Semantic Similarity—items that are similar to the primary item based on external world knowledge about the data; (5) Random Exemplars—a randomly-drawn subset of data, which we use as a naive baseline. In our experiments, we compare these to each other and to a No Guardrail control condition (6). As these guardrails are intended to be implemented by data exploration platforms, they span a spectrum with respect to the development effort and metadata needs: they range from simple statistical options to context-informed options. However, all of these methods can be precomputed as a new dataset is loaded and are feasible to integrate into existing explorers with minimal to no supervision.

To test the effectiveness of the sampling methods, we ran a preregistered, mixed-design crowd-sourced study ($n = 500$) comparing the five methods across two domain scenarios (COVID-19 and Stocks). We organize our evaluation around the following main research question: How do different guardrail sampling methods compare in effectiveness? More specifically, we evaluated four outcomes: **Trust**—perceived chart trustworthiness, **Accuracy**—rank judgment of the focal item in the dataset, **Context**—a judgment of the sufficiency of context, and lastly **Preference**—participants' personal preference of a guardrail method selected from a lineup.

Our findings indicate that, overall, **guardrails consistently outperform the No Guardrail control across all measures**, with each having its own advantages. Methods that surface statistical cues—Percentile-based Exemplars, Percentile Markers and Cluster Representatives work best when the goal is accurately judging the overall performance of the focal item. Exemplars with Semantic Similarity, which shows real-world peers of the focal item, is especially effective when the goal is chart credibility. Meanwhile, Random Exemplars is a surprisingly effective lightweight baseline that outperformed our expectations. In summary, our work provides an empirically validated roadmap for automatically contextualizing time series visualizations, quantifying the design trade-offs between guardrail strategies to reliably optimize user needs for either trust or accuracy.

2. Background & Related Work

In this section, we first discuss interventions aimed at mitigating misleading charts at the times of creation, exploration, and consumption stages. We then explore how interactive explorers may enable cherry-picking. Finally, we review the *guardrails* concept and design space that we build upon in this work.

2.1. Interventions for Mitigating Misinterpretation

There are several key issues that contribute to visualizations misleading their audiences, each requiring its own interventions. One is visual tricks: choices made while constructing charts, such as truncated axes, omitted baselines, and zoomed-in ranges, may lead their viewers to incomplete or biased interpretations [PRS*15]. Researchers developed checkers that help authors mitigate risky specifications at creation time, before charts are published. Draco, for example, encodes design rules as constraints and automatically surfaces specification issues, recommending safer alternatives [MWN*19]. Practical visualization linting prototypes similarly flag command pitfalls programmatically and surface them to the author [MK18, HCS20]. To help the audiences combat visual tricks that the authors did not catch (or intended to spread), literature proposed tools that present a check view or add salient annotations that can help viewers notice exaggerations without leaving the primary figure [FMMM22,RWC19]. More recently, researchers developed AI-assisted chatbots that transform views and describe the tricks upon user demand [DM25].

Another issue that leads to misleading visualizations arises during exploratory visual analyses, when reviewing many subgroups

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

increases the likelihood of spurious patterns, a phenomenon known as the multiple comparisons problem [ZZZK18, ZDSZ*17]. Interventions addressing this problem propose incorporating safeguards that monitor user actions to validate whether the discovered pattern is statistically sound [ZDSZ*17].

All interventions against misleading visualizations, however, share a common limitation in that they are not effective unless the viewer is skeptical in the first place and trusts the intervention. Once a misleading chart has been flagged as such, after it has already spread, it becomes an uphill battle for post-facto interventions, such as issuing warnings, conducting fact checks, and adding appended notes. Moreover, people often continue to rely on initial implications even after a clear correction has been published, known as the continued-influence effect; 'belief echoes' can persist long after a claim has been refuted or retracted [LES*12, Tho16]. Platform dynamics compound this issue, as false or misleading items tend to proliferate faster than corrections, which widens the gap between the initial exposure and any subsequent correction [VRA18]. Additionally, cognitive biases and motivated reasoning may lead audiences to not be convinced even by a corrected chart. Evidence shows that misleading visualizations can attract substantial engagement and rebuttal, yet the first impression remains mostly intact [LLK24].

2.2. Persuasive Framing and Cherry-picking in Visualization

We use the term *persuasive visualization* to refer to charts presented in a communicative context to influence beliefs or decisions [CP08, MRK*23]. One particularly common pathway to create misleading visualizations is cherry-picking items in data explorer platforms, where a user can use the platform to make a data point appear unusually good or bad and then share such a view on social media [LPLK23]. Contemporary data explorers, such as public health dashboards and retail investing platforms, expose various degrees of freedom for the user to adjust the view: selecting a suitable time frame, modifying axes and scales, and curating comparators. This freedom of modification is crucial for exploration but can also enable cherry-picking [LCKL25]. Prior work confirms that visual exploration choices can introduce selection bias, which can be exploited by adversarial framing and persuasive tactics [ZDSZ*17]. Literature also documents concrete deceptive tactics that were used to create charts that spread misinformation about the COVID-19 pandemic and vaccination measures, such as including curated subsets and comparators, as well as captioning and annotation [LPLK23].

Whether deceptive or correct, harmful or useful, visualizations are often used for persuasion. Design and presentation choices can affect attitudes even when the underlying data remains unchanged. Prior work on the *persuasive phase* of the chart lifecycle argues that once a view exits exploratory analysis, it enters a communicative context where framing and emphasis aim to persuade, rather than to discover [CP08]. Furthermore, charts can be more persuasive than equivalent text summaries [SSC*23]. Presentation details, such as sources and annotations, can modulate perceived credibility and influence [PMN*14]. Persuasive impact also depends on viewers' prior beliefs and goals, which can shape what they extract from a visualization [MRK*23]. In modern data explorers, these persuasive tools are readily available to users through item selection, axis adjustments, and metric choices, among other features.

In short, persuasive use of charts and the widespread availability of cherry-picking affordances create the perfect conditions for the spread of misleading, yet "data-driven" messages, presumably against the intentions of the creators of the data exploration platforms that enable them. This motivates the interventions that add context at the moment of viewing or exploring.

2.3. Guardrails: Concept and Design Space

In our prior work, we introduced *visualization guardrails* to combat cherry-picking in interactive explorers at both exploration and explanation times [LCKL25]. Visualization guardrails are an in-chart intervention that embeds contextual comparisons directly into an interactive data explorer. In our evaluation, we compared several presentation styles (displayed in the appendix, Figure A5) and revealed a clear pattern: contextual items that share the same visual language with the main items (such as simple grayed-out lines) successfully revealed cherry-picking, increased skepticism, and were easy to understand. At the same time, more complex distributional and statistical visual designs were ineffective, presumably because they are more difficult to understand. Our prior work, however, focused solely on presentation styles and did not explore how to best select items to display as the guardrail.

In this paper, we build on this evidence, focusing only on the proven effective type of guardrails, and ask the practical question our prior work left open: Which contextual lines should the system show? We translate the design concept of guardrails into deployable sampling strategies that real explorers can precompute and render quickly. We formalize these sampling methods, analyze their effects across a series of tasks, and provide guidance for selecting a method to match platform and task goals.

3. Guardrail Selection Methods

Next, we first describe our approach for designing a variety of guardrail selection methods and then introduce each method implemented and evaluated in this work.

3.1. Design Process

One of the primary design goals was to generate a small set of implementable and conceptually-distinct guardrail selection strategies that we could evaluate both by logical analysis and with a study. Initially, all of the authors met to brainstorm ideas for selection strategies, prioritizing coming up with as many different ideas as possible. Our session was informed by the results of our prior work [LCKL25], which showed that visually and conceptually simpler guardrails are more understandable and effective. We came up with a set of 11 distinct strategies, which we implemented and iteratively tested out on real-world data. We then conducted multiple rounds of (i) adjusting the conceptual and implementation details to make sure that guardrails result in understandable, effective, and uncluttered views, and (ii) grouping together or further differentiating similar ideas, eventually paring down to five distinct sampling strategies.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

Another key choice we had to make is *how many* comparators (exemplars, statistical markers) to show. Designers have to find a balance between a good coverage of the data space and clutter. Prior research on showing COVID-19 forecasts [PFCB22] suggests that trust hits a plateau when showing 6–9 forecasts, while prediction accuracy plateaus at 5–7 forecasts. We found that when considering historical data associated with different items, as opposed to modeled forecast data, there is a lot of volatility in the data (e.g., prices of different stocks fluctuate independently of each other), leading to more clutter. As such, we believe that displaying $n = 5$ contextual items alongside the focal item is a justifiable choice, yet we note that our methods work with fewer or more items as well.

3.2. Selection Methods

As a result of our design process, we developed five selection methods for evaluation, each with its own advantages and potential pitfalls. Figure 1 summarizes the five methods, spanning data-driven, world-knowledge, distributional, and exemplar-based comparators. One method, Random Exemplars, serves as a naive baseline against which we evaluate the rest, termed Context-rich guardrails. *Context-rich guardrails* use either the dataset or world knowledge to select useful contextual items. Most guardrails are *exemplar-based*, meaning they display other items from the dataset and thus are of the same data type as the focal item. One method is *abstract mark-based*; it shows statistical information and does not correspond to any actual item from the data.

Random Exemplars The Random Exemplars condition shows n items randomly drawn from the same dataset alongside the chosen focal items, without regard for equal distribution or meaning. We primarily implemented this condition as a naive baseline of a guardrail. This method's main strength lies in its simplicity and applicability to any domain and dataset. Additionally, even small random samples can sufficiently convey distributional cues, such as noisiness vs homogeneity [HRA15, ACG14]. At the same time, the randomness may also lead to uninformative or outright misleading comparisons.

Percentile Markers A more data-informed (yet still domain-agnostic) approach would be to show distributional information about the rest of the dataset alongside the focal item. Percentile Markers shows aggregate percentile bands. We calculated five roughly equally spaced percentiles—5th, 25th, 50th, 75th, and 95th—and show them as lines alongside the focal item. We create the percentile lines by calculating the five percentiles at each timestep t across the entire dataset. So, in essence, these bands do not necessarily correspond to "real" data or trends found in the dataset and serve as synthetic distributional markers.

Percentile-based Exemplars In order to still convey the distribution but anchor it in real items and thus potentially increase understandability, we designed the Percentile-based Exemplars strategy. Here, for each previously calculated synthetic percentile line, we find the single item (stock or country) that most closely tracks the percentile line using the minimum least-squares distance across all time steps. This method allows for apples-to-apples comparison with the focal item while still conveying the distribution. Alongside the item labels, we additionally display the percentile that they serve to represent.

Cluster Representatives The Cluster Representatives condition similarly aims to display the variety of trends in the dataset while anchoring it to actual items. In this strategy, we first cluster all time series in the dataset into n clusters using the k-means algorithm. For each cluster, we find the item closest to the cluster centroid by computing the minimum L2 distance. The advantage of this method compared to Percentile-based Exemplars is that it does not prioritize uniformly-trending items that best align with a synthetic percentile curve. In cases where a dataset has a lot of items that show big increases or big drops throughout the period, clustering such shapes would allow us to surface them as context and potentially get a more realistic view of the variation of trajectories in the data. On the other hand, this condition is somewhat less intuitive to explain to the user.

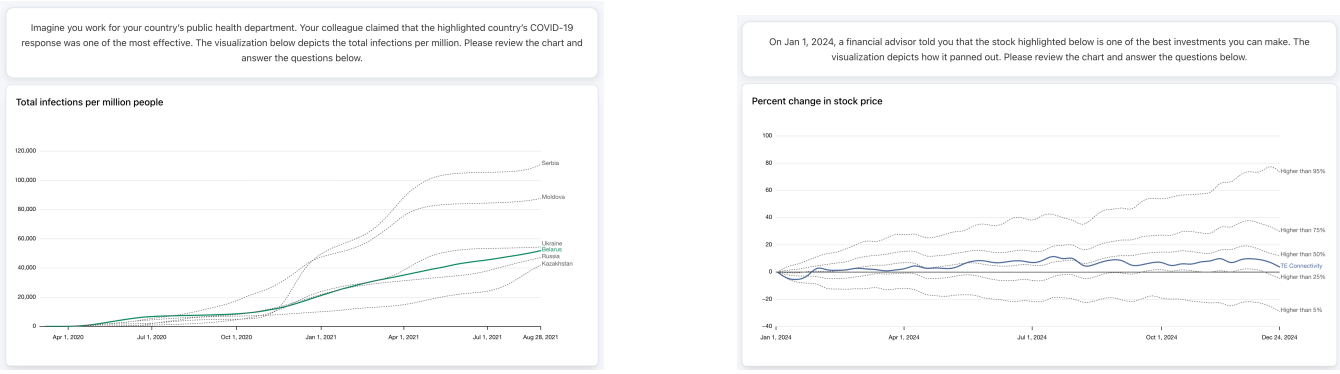
Exemplars with Semantic Similarity One limitation of all of the above methods is that they do not capture any of the potentially meaningful semantics of the data. For instance, when comparing COVID-19 case curves it might be most useful to look at countries from the same region, with similar demographics, alongside other metrics. The goal of this method is to sample a set of n semantically comparable peer items that would make sense to contextualize the focal item in real-world scenarios. This could be the same sector and similar market cap for stocks, or the same region and similar demographics for countries. Our implementation employs a large language model (LLM) ensemble and a high-agreement filtering mechanism to select robust comparators. We generate ten independent samples by querying the model (ChatGPT-5) with a fixed prompt that requests the five most appropriate comparators for the focal item and task. We then apply a majority-vote consensus heuristic, retaining any entity that achieves a high level of inter-response agreement (appearing in seven or more of the ten generated samples). The full prompts are detailed in the appendix.

There are a number of alternative methods that could be employed here and that we considered in the process: collecting domain-specific metadata (such as UN region classes for countries) or collaborating with a subject matter expert to identify an appropriate context for each focus item. The LLM approach we used offers significant flexibility as it works for many domains, especially those of interest to data explorers targeting the general public. It also offers low overhead; however, we urge any implementation to first thoroughly investigate whether the method yields appropriate results.

4. Evaluation Methodology

To identify user preferences across guardrails sampling methods, as proposed in Section 3, we conducted a preregistered mixed-design crowd-sourced study. In this section, we outline the recruitment and assignment of participants, study design, and analysis methodology. All prompts, stimuli, datasets, analysis code, and other materials are provided in the supplemental materials. Our study, developed with reVISit [CWS*26], can be reviewed by following the link. The code for the study is available on GitHub.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.



(a) Exemplars with Semantic Similarity in the COVID scenario

(b) Percentile-based Exemplars in the Stock Scenario

Figure 2: Examples of our guardrails as seen in the Task 1 stimulus. Figure (a) illustrates the Exemplars with Semantic Similarity strategy, showing items semantically similar to the focal entity (here, countries demographically and geographically similar to Belarus). Figure (b) illustrates the Percentile-based Exemplars strategy, which instead selects context based on their percentile bins in the global distribution.

4.1. Study Procedure

In designing our study, we aimed to elicit guardrail preferences across a variety of measures, task framings, and scenarios. As such, our study utilized two datasets, two task types, and several measures of preference, described in more detail below. The factors for data scenario (COVID-19 vs. Stock Performance) and the specific guardrail conditions (only two of the five guardrail types were shown to an individual participant) were manipulated between subjects. The factors for the focal item (three specific items seen by all participants) and the presence of a guardrail versus the control condition were manipulated within subjects. This approach allowed us to compare the effect of having a guardrail versus no guardrail and account for variation across specific focal items at the individual participant level.

In terms of process, each participant was assigned to a scenario—either COVID-19 cases by country or stock performance over time—and completed two tasks, followed by a survey. Due to a combination of randomization and participant dropout, the conditions are not perfectly balanced. Additionally, due to technical issues, we separately followed up with participants to complete the post-study survey for an additional payment of $0.25 and received 468 responses. In the survey, we collected their self-reported familiarity with charts, statistics, the domain of the scenario, as well as their political leaning. We report the results of the post-task survey in the supplemental materials; however, we note that the insights did not meaningfully inform the results of our main analysis, and we thus omit them from the discussions below.

Task 1: Chart & Survey

The first task was **Chart & Survey**, in which we presented the participants with a chart accompanied by a persuasive caption. In the COVID scenario, the chart was shared by a colleague in public health who was advocating for adopting a given country's pandemic protocol, describing it as "one of the most effective" (seen in Figure 2a). In the Stocks scenario, the chart was shared by a financial advisor who had promoted a given investment, by similarly describing it as "one of the best investments" (seen in Figure 2b). Each participant saw three such charts (with different data) in random order, one by one: one Control chart (with no guardrails whatsoever), and two randomly selected guardrail methods. In all cases, however, the focal item was selected to be middle-of-the-pack as to evaluate the usefulness of providing guardrails as context. The specifics of data selection are discussed further in Section 4.3.

For each chart, participants rated how trustworthy they found the chart on a 7-point Likert scale, estimated the focal item's position in the entire dataset from 0 to 100 (equivalent to percentile rank), and lastly reported how appropriate they found the context in the chart to make an informed decision given the scenario, also on a 7-point Likert scale. After seeing the three charts, participants completed a simple attention check question, which can be seen here.

Task 2: Preference Selection

The second task was **Preference Selection**, in which the participants saw a lineup of four charts. Each chart depicted the same focal country or stock, but with different guardrails: Random Exemplars, Cluster Representatives, Percentile-based Exemplars, and Exemplars with Semantic Similarity. We excluded the Percentile Markers here to limit the conditions to Exemplar-based guardrails.

We asked the participants to select the chart that they find the most useful, and then provide their rationale in an open text response field. To assess the differences in guardrail preferences according to the **specifics of the task**, we randomly assigned the participants into two groups that differed in their task framing. The first group was asked to complete a task in what we called a **Holistic** framing: they were asked to choose a chart that best supports a relative evaluation. For instance, in the COVID condition, the participants were asked to judge the effectiveness of a country's COVID response. Such a task, in theory, is best judged when compared to demographically and geographically similar countries; thus, we hypothesized that most would select Exemplars with Semantic Similarity. The second group saw what we called a **Precise** framing, which tasked the participants to choose a chart that best supports an overall or global performance. For example, in the context of stocks, they were asked to evaluate the return of the stock. This task, in contrast, is best judged on an absolute scale (i.e., when choosing an investment to maximize profit, how peer companies

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

did in the market does not matter as much as all alternative investments), thus we theorized that most participants would select the Percentile-based Exemplars view.

4.2. Procedure & Recruitment

We first conducted an informal pre-pilot with colleagues from our university. As a result, we clarified task wording and introduced the two diverging framings in the Preference Selection task, as feedback indicated that participants would often assume one of these two task framings. We then conducted a pilot study with 85 crowdsourced participants on Prolific. We used the results of the pilot to estimate the payment based on completion time, as well as to size our experimental sample using a power analysis. We aimed to detect moderately-sized effects based on our pilot results with 80% power and at the standard statistical significance level of 5%. As a result, we recruited 500 participants for the main study. The median completion time was just under 7 minutes, and we paid each participant $2.10 (for an hourly rate of $18/hour). Our IRB deemed the study exempt from full review. All participants gave informed consent, and we collected no identifying data.

4.3. Dataset, Stimuli and Framing Materials

We sourced the COVID-19 dataset from Our World in Data [Our20]. The charts in the study showed the participants cumulative infections per million inhabitants from April 2020 to August 2021. We used cumulative infections as opposed to the daily rate because our pre-pilots showed that daily variance made the charts visually noisy and less clear for the judgment tasks. We collected the S&P 500 stock prices via a script using Yahoo Finance API and showed data for the calendar year 2024 [Yah25]. To reduce noise and smooth the trajectories for clearer presentation, similar to the COVID-19 data, we plotted the weekly closing prices and transformed the values into percentage changes over the period, all starting at 0 at the beginning of the chart. This is consistent with common stock market data exploration sites, such as Yahoo Finance and Google Finance.

We designed the charts—seen in Figure 2—to be consistent across conditions and aesthetically minimalist as to not interfere with performance. Our initial work on evaluating guardrail designs showed that it is important to clearly visually distinguish the guardrails from the main data [LCKL25]. Consequently, we designed each chart to highlight the focal item with a saturated color and a slightly thicker stroke, while auxiliary guardrail lines were grayed and dashed, allowing them to recede into the background while still remaining salient.

In choosing the focal items to highlight, we aimed to select the "(lower) middle of the pack" views that could believably be used for persuasive visualization (thus not the worst or blatantly unconvincing), yet underperforming enough that contextual data would provide revealing evidence. In other words, we wanted to ensure that the decision to trust or distrust the chart is not trivial and be able to measure the relative effect of contextual data on persuasiveness. To select such datapoints for Stocks scenario, we filtered the set of S&P 500 companies to those that met the following three criteria: (i) roughly visually smooth weekly trajectories, to control for potential volatility interference; (ii) no significant dips below 0% to avoid significant floor cues; and (iii) approximately 35th percentile performance for the year across the entire dataset, so not the best but also not the worst. We selected three stocks closest to the 35th percentile that satisfied these criteria for task 1: TEL (TE Connectivity), COR (Cencora), and CHD (Church & Dwight), and one for task 2: VZ (Verizon). To select countries to show in the COVID scenario, we followed a similar protocol, but in this case, sampling from the 65th percentile of cases per capita, as with COVID the "goodness" metric is flipped and lower is better. We again selected the three closest items that satisfied the criteria for task 1, yielding Greece, Germany, and Belarus, and one for task 2: Norway.

4.4. Analysis Overview

We adhered to our preregistration, which can be found here and describes the analysis in detail. We excluded responses that: (i) failed attention checks, (ii) open-ended rationales that were non-substantive or gibberish, (iii) incomplete trials, or (iv) data missing because of technical issues. As a result, we excluded 13 participants and analyzed 487 responses. The raw responses, analysis scripts, and participants' anonymized demographic data are available in the supplemental materials.

For Task 1 (Chart & Survey) outcomes (Trust, Accuracy, and Context), we investigated two main hypotheses (see Table 1). First (H1), we tested if adding any guardrail improved responses compared to the baseline No Guardrail condition. Second (H2), we tested if Context-rich guardrails performed better than a naive Random Exemplars guardrail. We employed linear mixed-effects analyses, including participant and stimulus as random intercepts. Fixed effects were evaluated using likelihood-ratio tests, followed by planned pairwise contrasts with Holm correction.

For Task 2 (Preference Selection) responses, we tested which selection method was preferred most often across two task prompts to identify whether users' preferences depend on the question at hand (H3, see Table 1). We first used a chi-square test of independence to determine if the prompt type (Holistic vs. Precise) influenced the distribution of chosen chart types. Within each prompt, we ran preregistered one-sided two-proportion z-tests, comparing the hypothesized winner to the other options. We adjusted p-values using Holm correction, and we report proportions with Wilson 95% Confidence Intervals (CIs) and Cohen's h.

Finally, we performed a thematic analysis with inductive open coding on the Task 2 preference rationales. To code the results, the first author conducted an initial pass over the data, assigning individual codes to participants' preference rationales. The last author then completed their own pass, suggesting code edits, deletions, and combinations. The first and last authors met to further group codes into higher-level categories and organize the final codebook.

5. Study Results

In this section, we report the results of our evaluation across the two tasks. Table 1 presents a concise summary of the study results relative to the preregistered hypotheses. The exact p-values for all preregistered hypothesis tests are reported in Appendix Table A1.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

Table 1: Overall, guardrails improved *Trust*, *Context*, and relative-performance *Accuracy* versus Control (H1a-c supported). Context-rich guardrails did not exceed a Random guardrail on Trust or Context (H2a, H2c not supported), but did improve Accuracy relative to Random (H2b supported). Task framing (Holistic vs. Precise) did not alter guardrail preferences (H3a-c not supported).

ID	Statement	Support	Findings summary
H1: Guardrails vs No Guardrails			
H1a	Guardrails increase *Trust*.	**Supported**	Higher with any guardrail vs No Guardrail with a small but reliable effect, about +0.5 Likert points.
H1b	Guardrails improve *Accuracy*.	**Supported**	Significantly lower error for all guardrails except Exemplars with Semantic Similarity, where the results are highly variable.
H1c	Guardrails increase *Context*.	**Supported**	Significant improvement of about +1.5 Likert points, on average, compared to No Guardrail.
H2: Context-rich Guardrails vs Random Items			
H2a	Context-rich guardrails increase *Trust* over Random Exemplars.	**Not supported**	No improvement over random, with the exception of Exemplars with Semantic Similarity in the COVID scenario.
H2b	Context-rich guardrails improve *Accuracy* over Random Exemplars.	**Supported**	Significantly lower error for all guardrails except Exemplars with Semantic Similarity, where the results are highly variable.
H2c	Context-rich guardrails increase *Context* over Random Exemplars.	**Not supported**	No improvement over random, with the exception of Exemplars with Semantic Similarity in the COVID scenario.
H3: Task Framing Influence on Guardrail Preferences			
H3a	Chosen chart *Type* differs between *Holistic* vs *Precise* prompt.	**Not supported**	No significant effect of prompt on preference choices.
H3b	Under *Holistic*, Exemplars with Semantic Similarity is chosen most frequently.	**Not supported**	Exemplars with Semantic Similarity not reliably preferred more under Holistic vs other types.
H3c	Under *Precise*, Percentile-based Exemplars is chosen most frequently.	**Not supported**	Percentile-based Exemplars not reliably preferred more under Precise vs alternatives.

5.1. Task 1 Results

Figure 3 presents an overview of the results of Task 1 (Chart & Survey). Across scenarios and measures, we see that presenting *any* type of guardrail improved responses, however the specifics vary.

We observe that respondents consistently report **higher trust in charts that do feature guardrails**, confirming hypothesis H1a (COVID: $\chi^2 = 16.46$, $p < 0.001$, stocks: $\chi^2 = 22.33$, $p < 0.001$). The effect, albeit statistically significant, is not large: trust increases by roughly half of a Likert point with guardrails. Notably, while guardrails overall are associated with higher reported trust, using Context-rich guardrails does not yield a difference relative to Random Exemplars (COVID: $\chi^2 = 1.83$, $p = 0.18$, stocks: $\chi^2 = 0.324$, $p = 0.57$), leading us to reject hypothesis H2a. In other words, **any random guardrails are approximately as effective at increasing trust as the more meaningful samples.**

Participants' responses to the context appropriateness question (e.g., "Do you feel that the visualization provided appropriate context to make an informed decision about your investment?" for stocks) follow a similar pattern, however with stronger effects. **Guardrails increase the context substantially compared to No Guardrail** (COVID: $\chi^2 = 100.17$, $p < 0.0001$, stocks: $\chi^2 = 113.08$, $p < 0.0001$), by an average of 1.5 Likert points and moving the results from below to above the midpoint, confirming H1c. Here again, notably, **the Random Exemplars condition performs as effectively** as the more Context-rich guardrail methods (COVID: $\chi^2 = 1.78$, $p = 0.18$, stocks: $\chi^2 = 0.35$, $p = 0.56$), rejecting H2c.

We also confirm that **guardrails significantly decrease the error when estimating where the data point falls in the distribution**, reducing it by an average of 5.3 points in COVID scenario, and 3.25 points in Stocks (COVID: $\chi^2 = 15.7$, $p < 0.0001$, stocks: $\chi^2 = 12.03$, $p < 0.001$) and confirming our hypothesis H1b. The Context-rich results also prevail in the Stocks scenario, decreasing the error on average by 4.2 more points compared to the Random Exemplars guardrail ($\chi^2 = 11.18$, $p < 0.001$). In the COVID scenario, however, the improvement was marginal ($\chi^2(1) = 3.11$, $p = 0.08$). We also note that, as shown in Fig. 3, all methods *except* Exemplars with Semantic Similarity significantly improved accuracy; Exemplars with Semantic Similarity underperforms all other guardrails, including No Guardrail in the COVID scenario. Our findings confirm a potential negative effect of this sampling strategy: while Exemplars with Semantic Similarity might be trustworthy and provide helpful context, it may also misleadingly frame the data within a non-representative subgroup: the semantic subset may be significantly skewed relative to the overall distribution.

Lastly, we note that the results are largely consistent across the two data scenarios, with several deviations. Firstly, the results primarily vary in the Exemplars with Semantic Similarity condition, which follows from the fact that the guardrail items are highly dependent on the specifics of the domain and the focal item and thus result in higher variance. Secondly, the absolute values across all three measures are notably different: participants find Stock charts more trustworthy and complete than COVID, and make more accurate performance judgments about them. This result likely follows from the fact that participants are both more familiar with (and more biased by) the ubiquitous COVID-19 visualizations.

5.2. Task 2 Results

Figure 4 shows participants' responses in Task 2: Preference Selection. We observed no significant framing effect on preferences, rejecting our hypothesis H3a. The distribution of the selected meth-

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

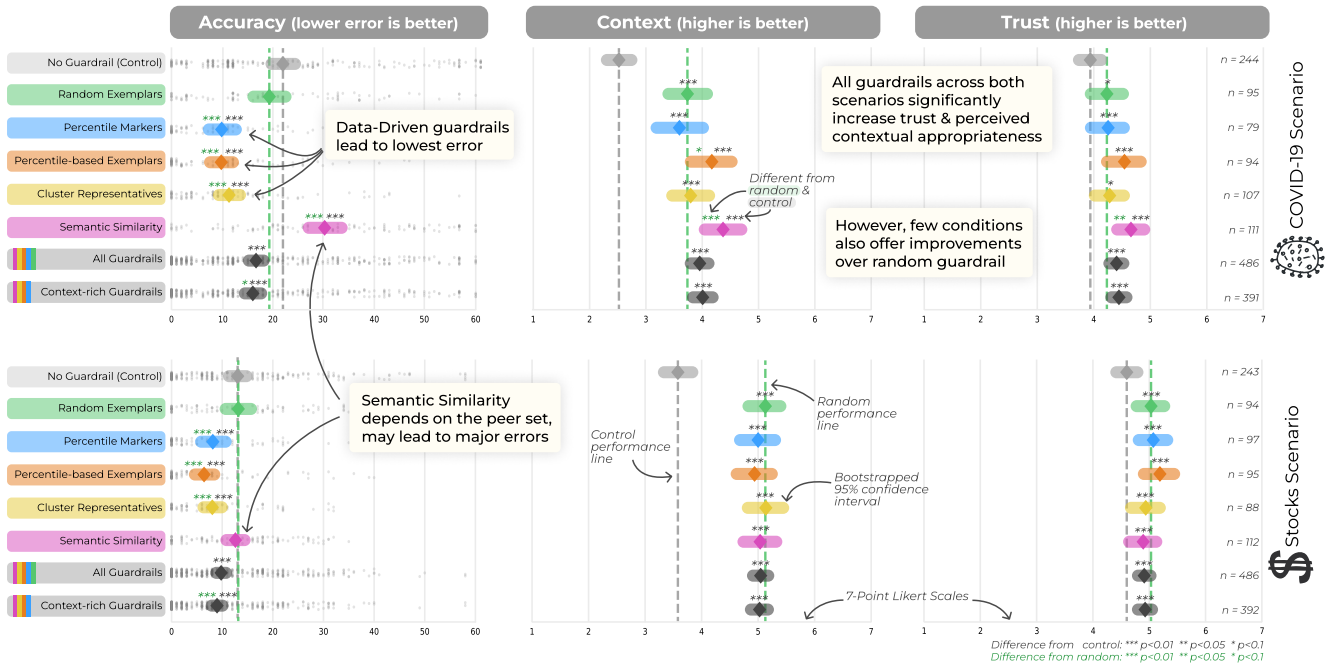


Figure 3: Across all tasks and scenarios, guardrails improve accuracy, trust, and context. For trust and context, Random Exemplars performs as well as data-driven guardrails, and Exemplars with Semantic Similarity performs better in the COVID scenario. Data-driven guardrails—Percentile-based Exemplars, Percentile Markers, and Cluster Representatives— however, offer a significant improvement in accuracy. Exemplars with Semantic Similarity method greatly depends on the distribution of the semantically-similar peers, thus may lead to major accuracy errors. Note that the statistical significance markers follow from mixed-effects models that also account for participant and item effects and may not directly reflect visible confidence interval differences.

ods did not differ between Holistic and Precise prompts in either scenario (COVID: $\chi^2 = 0.50$, $p = 0.92$, stocks: $\chi^2 = 0.55$, $p = 0.91$). Consequently, participants also did not choose the Exemplars with Semantic Similarity guardrails more often in the Holistic prompt or the Percentile-based Exemplars in the Precise prompt, leading us to reject hypotheses H3b and H3c, respectively.

Interestingly, the results are consistent across both scenarios. Although not a tested hypothesis, we suspected that Exemplars with Semantic Similarity would be preferred for COVID-19 data, as pandemic data was typically compared against peers, while Percentile-based Exemplars would be more preferable to evaluate stock data, since investors are typically not limited to investing among a group of stocks. Taken together, the **selection results primarily reflect participants' general preferences and do not vary with the specifics of the task prompt or the domain**. By and large, participants prefer either Exemplars with Semantic Similarity or Percentile-based Exemplars guardrails to complete their tasks. In summary, although both sampling methods performed similarly in Task 1, Task 2 demonstrates a significant user inclination toward specific guardrail types.

6. Qualitative Rationale

To understand *why* participants selected certain charts, we conducted an *exploratory* analysis of open-ended text rationales. Figure 5 shows our codebook used to annotate the responses. Below, we present three themes that describe the codebook.

6.1. Theme 1: Preference for a Specific Context

One way in which participants expressed their preference for a guardrail was by describing their desire for the specific property that the guardrail displayed. Codes 1.1–1.4 in the codebook reflect the preferences for contextual strategies that fairly directly map onto the specific guardrail selection methods we tested. For instance, some participants described preferring peers that make sense together in the real world (same sector for stocks; neighboring regions, similar demography for countries). As such, they chose the Exemplars with Semantic Similarity method (code 1.2) and cited seeking "apples-to-apples" comparisons. One participant described preferring the chart that showed Norway's immediate neighbors: "*I was influenced by how close the other countries' data was geographically to Norway, because this closeness/similarity meant that you might be comparing similar countries, so getting rid of some of the factors influencing COVID-19 infections.*"

At the same time, other participants had different expectations of best context. Participants who selected Percentile-based Exemplars charts (code 1.1.2) cited the preference for the ability to make global comparisons and glean the distribution: "*This chart has percentages that seem to indicate how a labeled nation fares among the global COVID-19 outbreaks. It showcases nations with low outbreaks versus nations with high outbreaks in a simple manner.*" Notably, these preferences reflect personal expectations, rather than a selection that best matches the task or scenario—as previously discussed in the quantitative results.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

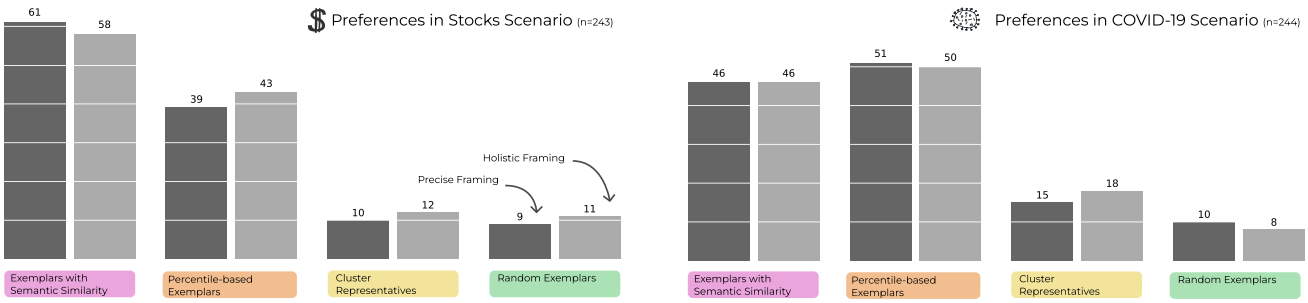


Figure 4: When asked to choose, participants prefer Exemplars with Semantic Similarity and Percentile-based Exemplars at roughly the same rate, independent of framing, even though we considered Exemplars with Semantic Similarity the better choice for the Holistic framing (the question targeted judgment relative to peers), and percentiles the better choice for Precise framing (the question targeted estimating absolute performance). Cluster Representatives and Random Exemplars are infrequently chosen for these tasks.

1. **Context strategies**
   - 1.1. **Selection coverage**
     - 1.1.1. `Broader contextual coverage`
     - 1.1.2. `Broader coverage by distribution`
   - 1.2. **Similar contextual peers**
   - 1.3. **Specific entities included**
   - 1.4. **Information specificity**
     - 1.4.1. `More information preferred`
     - 1.4.2. `Objective numbers preferred`
2. **Chart aesthetics preferences**
   - 2.1. **Declutter and clarify**
     - 2.1.1. `Clean chart`
     - 2.1.2. `Fewer overlapping lines`
   - 2.2. **Scale preferences**
     - 2.2.1. `More granular axis`
     - 2.2.2. `Wider axis range`
3. **Perceived accuracy & trust**
4. **Focal series take-away**

Figure 5: Final codebook used for open-ended responses. We organize the themes based on the highest level of codes. Each text response may be described by one or more codes. We excluded codes that indicated no preference or a clearly mistaken justification.

6.2. Theme 2: Preference for Chart Aesthetics

Moreover, the actual contextual content was not the only reason for participants' preferences. Many of the responses cited clean layouts, reduced overlap (codes 2.1), and axis scaling (codes 2.2) as rationales for their guardrail preferences: "*The highest total of infections per million people in chart B only goes up to 80,000 and increments of 10,000 vs 20,000 in the other choices, showing slightly more detail in the changes over time.*" While chart aesthetics and clarity of guardrails were not central to our guardrail design, they are often the primary reason for participants' preferences. This highlights an important challenge of competing demands: when attempting to show, for instance, semantically-relevant contextual data, one may end up with a more cluttered chart as a side effect. For instance, one respondent simply reported their preference being due to "*there [being] less overlapping elements for the other countries*", emphasizing chart clarity above all.

6.3. Theme 3: Preference for Specific Takeaways

In another set of preferences, participants chose their preferred chart not because of the context itself, but rather based on the light in which it presented the focal item (codes 3, 4). Text responses sometimes directly referenced the performance of the focal country or stock, using terms such as "*middle of the pack*", "*no big declines*", or "*stable*". Participants often prefer a chart that reveals a specific quality of the target data that a condition shows, such as: "*It shows it is middle of the pack, whereas others showed it was less so*". Alternatively, others simply prefer seeing the focal country performing better or worse, whichever aligns with their expectation: "*It showed in comparison to others, there was no period of significant decline, and the overall performance was positive [...]*"

7. Discussion

In this section, we interpret the findings and articulate their implications for visualization practice. We highlight the key trade-offs and limits that guide the implementations of guardrails in real systems.

7.1. Guardrails help, but there's no one-size-fits-all

Across both the COVID and Stock scenarios, the presented guardrails consistently improved context and generally increased trust. Gains in accuracy, however, were concentrated in data-driven methods that explicitly encode distributional structure. This pattern implies that there is no *single best* guardrail; instead, the suggested guardrail selection methods form a **toolbox** where each tool is optimized to serve different goals.

When the goal is to make a persuasive chart feel **credible and complete**, we found that, surprisingly, nearly any guardrail was effective. Exemplars with Semantic Similarity and Percentile-based Exemplars offered the largest gain, but were sometimes statistically indistinguishable from Random Exemplars. Our results indicate that randomly-drawn items might be a "false friend," offering credibility without providing any semantically-meaningful context.

On the other hand, when the goal is to help readers **accurately estimate the global rank** of a focal item, the best-performing methods are those that signal the data spread and typical trajectories. These methods include Cluster Representatives, Percentile

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

Markers, and Percentile-based Exemplars. The Exemplars with Semantic Similarity method, conversely, led to increased error. These results confirm that participants leverage data-driven methods that surface distributional information but are biased by semantic context, which is entirely dependent on how the peer items perform relative to the rest of the data. An important insight is that Exemplars with Semantic Similarity has the potential to mislead: while significantly improving credibility, it substantially biases global judgments, potentially exacerbating cherry-picking.

7.2. Context preferences are micro-level, not macro-level

We found **no framing effect** on guardrail preference distributions in Task 2, as shown in Figure 4. Participants did not adjust their preferred chart based on the instructions or domain specifics. This may reflect the limitations of using prompts in an experimental setup to guide tasks, but also the fact that users form their preferences based on lower-level factors that are external to the task.

While we designed the guardrail sampling methods from a **macro-level** point of view (i.e., *what data would make sense as context?* or *how to best surface the distributional patterns in the dataset?*), the responses in Task 2 reveal that participants primarily base their selections on **micro-level** preferences. These include chart aesthetics, specific comparisons or insights they can make from the charts, or preferences for seeing a specific stock or country they expected to observe. Although many participants did cite broader reasons aligning their selections with the task they were performing, the quantitative results being consistent across scenarios indicate that this was primarily not the determining factor.

Overall, both the results of Task 2 suggest that participants prioritized **relevance and clarity** over the exact prompt. Taken together with the results of Task 1, we can conclude that when a chart surfaces the benchmarks people expect to see, provides insights people find interesting and non-contradictory, and is easy to read and aesthetically pleasing, people prefer it and trust it more.

8. System Design Implications

Based on our studies, we distill the following design implications:

1. **Prioritize Distributional Context for Accuracy:** When the objective is to help users accurately estimate the **global rank** of a focal item, only **data-driven guardrails** that encode distributional structure (Percentile-based Exemplars, Percentile Markers, Cluster Representatives) offer significant improvement.
2. **Use Semantic Context for Trust and Relevance:** When the goal is to maximize **user trust and perceived context completeness**, any guardrail is effective. Use Exemplars with Semantic Similarity to satisfy user preference for **relevance**, but be aware: it can reduce accuracy for global judgments by misleadingly framing the data within a non-representative subgroup.
3. **Explain the Contextual Choice:** Since user preferences are often pre-determined and independent of the task or domain, the system should **explicitly surface and explain** why the specific guardrail items were chosen. This is necessary to bridge the gap between user-preferred micro-level cues and the macro-level goal of the guardrail (task relevance, distributional context).
4. **Adopt a Goal-Oriented Guardrail Toolbox and Mitigate Risk:** Implement a suite of context selection methods and deploy the method that best serves the system's use cases. Crucially, recognize that seemingly successful and easiest to implement methods like Random Exemplars can be a "false friend," offering credibility without meaningful context and potentially masking important data patterns.

9. Limitations and Future Work

Our study focuses on two real-world domains (COVID-19 cases and stocks) and examines a small set of non-extreme items. Future work could extend guardrail evaluations to other issues such as climate or election data, vary focal item rank to contrast high- and low-performing groups, and test whether effects depend on how consequential or familiar a domain is to participants.

We also fixed the number of contextual lines to five, following prior evidence that this strikes a balance between trust and performance [PFCB22]. Subsequent research can map how trust, context, and accuracy shift as the number and saliency of contextual lines change, identifying task-specific sweet spots. Our study also kept the number of focal items fixed, and future work could explore methods to accommodate multiple selections. Most methods are independent of the focal item and should thus scale easily to more focal points, whereas Exemplars with Semantic Similarity may require selecting guardrails that are similar to each focal item. As the number of focal items grows, it may also become advantageous to reduce the number of guardrails to address visual clutter.

Our design used concise, persuasive captions, but alternative framings—such as more adversarial narratives or neutral descriptions—may shift perceptions and outcomes. Future work may explore this narrative space, including how guardrails interact with varying framing strengths. Finally, because our US and UK crowd-sourced sample reflects a specific cultural lens, generalizability may differ across contexts, as notions of "appropriate" or useful comparators can vary by culture and experience.

10. Conclusion

Guardrails embed a set of contextual comparators into data explorers, better preparing basic charts for the various tasks people use them for. We proposed five guardrail sampling strategies and confirmed that all significantly improve trust and context, while data-driven methods improve accuracy in rank judgments. Through a crowd-sourced evaluation across multiple framings, tasks, and scenarios, our paper presents actionable recommendations for data explorer platform governance. Specifically, platforms should surface semantically-comparable peers for maximizing credibility, and select percentile-based exemplars for achieving precise judgment. In summary, our strategies allow data explorer platforms to move beyond simply displaying data to actively contextualizing decisions and fostering reliable interpretation for the general public.

11. Acknowledgments

This work is supported by the National Science Foundation (CNS 2213756).

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

References

[ACG14] ALBERS D., CORRELL M., GLEICHER M.: Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2014), CHI '14, Association for Computing Machinery, pp. 551–560. doi:10.1145/2556288.2557200. 4

[CP08] CHIH C. H., PARKER D. S.: The persuasive phase of visualization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas Nevada USA, Aug. 2008), ACM, pp. 884–892. doi:10.1145/1401890.1401996. 3

[CWS*26] CUTLER Z., WILBURN J., SHRESTHA H., DING Y., BOLLEN B., NADIB K. A., HE T., MCNUTT A., HARRISON L., LEX A.: ReVISit 2: A Full Experiment Life Cycle User Study Framework. *IEEE Transactions on Visualization and Computer Graphics (VIS) 32* (2026). arXiv:2508.03876, doi:10.48550/arXiv.2508.03876. 4

[DM25] DAS A. K., MUELLER K.: MisVisFix: An Interactive Dashboard for Detecting, Explaining, and Correcting Misleading Visualizations using Large Language Models, Aug. 2025. arXiv:2508.04679, doi:10.48550/arXiv.2508.04679. 2

[FMMM22] FAN A., MA Y., MANCENIDO M., MACIEJEWSKI R.: Annotating Line Charts for Addressing Deception. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2022), CHI '22, Association for Computing Machinery, pp. 1–12. doi:10.1145/3491102.3502138. 2

[Goo25] GOOGLE LLC: Google Finance, 2025. 1

[HCS20] HOPKINS A. K., CORRELL M., SATYANARAYAN A.: VisuaLint: Sketchy In Situ Annotations of Chart Construction Errors. *Computer Graphics Forum 39*, 3 (2020), 219–228. doi:10.1111/cgf.13975. 2

[HRA15] HULLMAN J., RESNICK P., ADAR E.: Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE 10*, 11 (Nov. 2015), e0142444. doi:10.1371/journal.pone.0142444. 4

[Joh20] JOHNS HOPKINS UNIVERSITY CENTER FOR SYSTEMS SCIENCE AND ENGINEERING: COVID-19 Dashboard, 2020. 1

[KLK18] KONG H.-K., LIU Z., KARAHALIOS K.: Frames and Slants in Titles of Visualizations on Controversial Topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC Canada, 2018), ACM, pp. 1–12. doi:10.1145/3173574.3174012. 2

[KSA21] KIM D. H., SETLUR V., AGRAWALA M.: Towards Understanding How Readers Integrate Charts and Captions: A Case Study with Line Charts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, May 2021), CHI '21, Association for Computing Machinery, pp. 1–11. doi:10.1145/3411764.3445443. 2

[LCKL25] LISNIC M., CUTLER Z., KOGAN M., LEX A.: Visualization Guardrails: Designing Interventions Against Cherry-Picking in Interactive Data Explorers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, Apr. 2025), ACM, pp. 1–19. doi:10.1145/3706598.3713385. 2, 3, 6, 4

[LES*12] LEWANDOWSKY S., ECKER U. K., SEIFERT C. M., SCHWARZ N., COOK J.: Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest 13*, 3 (2012), 106–131. 3

[LGS*22] LO L. Y.-H., GUPTA A., SHIGYO K., WU A., BERTINI E., QU H.: Misinformed by Visualization: What Do We Learn From Misinformative Visualizations? *Computer Graphics Forum 41*, 3 (2022), 515–525. doi:10.1111/cgf.14559. 2

[LLK24] LISNIC M., LEX A., KOGAN M.: 'Yeah, this graph doesn't show that': Analysis of online engagement with misleading data visualizations. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (Honolulu, HI USA, 2024), CHI '24, ACM, pp. 1–14. doi:10.1145/3613904.3642448. 3

[LPLK23] LISNIC M., POLYCHRONIS C., LEX A., KOGAN M.: Misleading Beyond Visual Tricks: How People Actually Lie with Charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg Germany, 2023), ACM, pp. 1–21. doi:10.1145/3544548.3580910. 2, 3

[LYI*21] LEE C., YANG T., INCHOCO G. D., JONES G. M., SATYANARAYAN A.: Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, 2021), CHI '21, Association for Computing Machinery, pp. 607:1–607:18. 2

[MK18] MCNUTT A., KINDLMANN G.: Linting for visualization: Towards a practical automated visualization guidance system. In *VisGuides: 2nd Workshop on the Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization* (2018), vol. 1, p. 9. 2

[MRK*23] MARKANT D., ROGHA M., KARDUNI A., WESSLEN R., DOU W.: When do data visualizations persuade? The impact of prior attitudes on learning about correlations from scatterplot visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2023), CHI '23, Association for Computing Machinery, pp. 1–16. doi:10.1145/3544548.3581330. 3

[MWN*19] MORITZ D., WANG C., NELSON G. L., LIN H., SMITH A. M., HOWE B., HEER J.: Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 438–448. doi:10.1109/TVCG.2018.2865240. 2

[Our20] OUR WORLD IN DATA: COVID-19 Data Explorer, 2020. 1, 6

[PFCB22] PADILLA L., FYGENSON R., CASTRO S. C., BERTINI E.: Multiple Forecast Visualizations (MFVs): Trade-offs in Trust and Performance in Multiple COVID-19 Forecast Visualizations. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–11. doi:10.1109/TVCG.2022.3209457. 4, 10

[PMN*14] PANDEY A. V., MANIVANNAN A., NOV O., SATTERTHWAITE M., BERTINI E.: The Persuasive Power of Data Visualization. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec. 2014), 2211–2220. doi:10.1109/TVCG.2014.2346419. 2, 3

[PRS*15] PANDEY A. V., RALL K., SATTERTHWAITE M. L., NOV O., BERTINI E.: How Deceptive are Deceptive Visualizations? An Empirical Analysis of Common Distortion Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea, 2015), CHI '15, Association for Computing Machinery, pp. 1469–1478. doi:10.1145/2702123.2702608. 2

[RWC19] RITCHIE J., WIGDOR D., CHEVALIER F.: A Lie Reveals the Truth: Quasimodes for Task-Aligned Data Presentation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland Uk, May 2019), ACM, pp. 1–13. doi:10.1145/3290605.3300423. 2

[SSC*23] STOKES C., SETLUR V., COGLEY B., SATYANARAYAN A., HEARST M. A.: Striking a Balance: Reader Takeaways and Preferences when Integrating Text and Charts. *IEEE Transactions on Visualization and Computer Graphics 29*, 1 (Jan. 2023), 1233–1243. doi:10.1109/TVCG.2022.3209383. 3

[Tho16] THORSON E.: Belief echoes: The persistent effects of corrected misinformation. *Political Communication 33*, 3 (2016), 460–480. 2, 3

[Tra25] TRADINGVIEW INC.: TradingView, 2025. 1

[VRA18] VOSOUGHI S., ROY D., ARAL S.: The spread of true and false news online. *Science 359*, 6380 (Mar. 2018), 1146–1151. doi:10.1126/science.aap9559. 3

[Yah25] YAHOO INC.: Yahoo Finance, 2025. 1, 6

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

[ZDSZ*17] ZHAO Z., DE STEFANI L., ZGRAGGEN E., BINNIG C., UPFAL E., KRASKA T.: Controlling False Discoveries During Interactive Data Exploration. In *Proc. SIGMOD* (New York, 2017), ACM, pp. 527–540. doi:10.1145/3035918.3064019. 3

[ZZZK18] ZGRAGGEN E., ZHAO Z., ZELEZNIK R., KRASKA T.: Investigating the effect of the multiple comparisons problem in visual analysis. In *Proc. CHI* (Montreal QC Canada, 2018), ACM, p. 479. doi:10.1145/3173574.3174053. 3

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

Appendix A: Stimuli, Datasets, and Materials

To facilitate review and replication, we provide direct links to the deployed study:

- COVID-19 scenario: https://vdl.sci.utah.edu/guardrail-samples-study/stage-1-covid
- Stocks scenario: https://vdl.sci.utah.edu/guardrail-samples-study/stage-1
- Condition explorer (all guardrail variants used in-study, plus prototypes): https://vdl.sci.utah.edu/guardrail-samples-study/sandbox

Appendix B: LLM Prompts and Final Selection Sets for Semantic Similarity (LLM) Condition

COVID-19 prompt and selections

Prompt.

> You are curating contextual comparisons for a public data-exploration platform that visualizes COVID-19 cumulative cases per million. The goal is to help people make better sense of charts, surface missing context, and ultimately support better decisions and a more holistic understanding of the metric. For each highlighted country, select five other countries as meaningful comparisons to co-plot as auxiliary lines. Consider geographic proximity (e.g., immediate neighbors), similar stages of economic development, comparable demographics and urbanization, and health-system capacity. Avoid random picks, duplicates, and microstates unless the anchor is one; relax constraints only as needed, and break ties by geographic proximity, then alphabetically.

Final selection sets used in-study.

- **Greece**: Italy, Spain, Portugal, Cyprus, Croatia
- **Germany**: France, Netherlands, Austria, Sweden, Denmark
- **Belarus**: Russia, Ukraine, Kazakhstan, Moldova, Serbia
- **Norway**: Sweden, Denmark, Finland, Iceland, Netherlands

Stocks prompt and selections

Prompt.

> You are curating contextual comparisons for a public data-exploration platform that visualizes S&P 500 stock price performance (percentage change). The goal is to help people make better sense of charts, surface missing context, and ultimately support better decisions and a more holistic understanding of the metric. For each highlighted stock, select five other stocks as meaningful comparisons to co-plot as auxiliary lines. Consider industry proximity (e.g., same or adjacent GICS industry or sub-industry), similar market capitalization, comparable growth/volatility profiles, and operating capacity. Avoid random picks, duplicates, and non-S&P 500 stocks; relax constraints only as needed, and break ties by industry proximity, then alphabetically by ticker.

Final selection sets used in-study.

- **COR**: MCK, CAH, CVS, WBA, HSIC
- **CHD**: PG, CL, KMB, CLX, EL
- **TEL**: APH, KEYS, GLW, HUBB, ETN
- **VZ**: T, TMUS, CMCSA, CHTR, AMT

Appendix C: Participant Demographics

We recruited a total of $N = 487$ participants after preregistered exclusions. Participants had a mean age of 41.74 years (SD = 12.44; median = 39.50).

Figures A1–A4 provide detailed visual summaries of the demographic distributions.

Appendix D: Analysis Scripts

Full analysis notebooks (model fits, bootstrap CIs, and figure generation) are available here: https://colab.research.google.com/drive/1YyBcm-NKntIQLPzOI-K49cCf6YAfwKqg?usp=sharing

Appendix E: Artifacts

- Study code repository: https://github.com/visdesignlab/guardrail-samples-study
- COVID-19 dataset: https://github.com/visdesignlab/guardrail-samples-study/blob/data-v1.0/public/stage-1-covid/data/clean_data.csv
- Stocks dataset: https://github.com/visdesignlab/guardrail-samples-study/blob/data-v1.0/public/stage-1/data/sp500_stocks.csv

Appendix F: Additional Tables, Figures and Screenshots

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

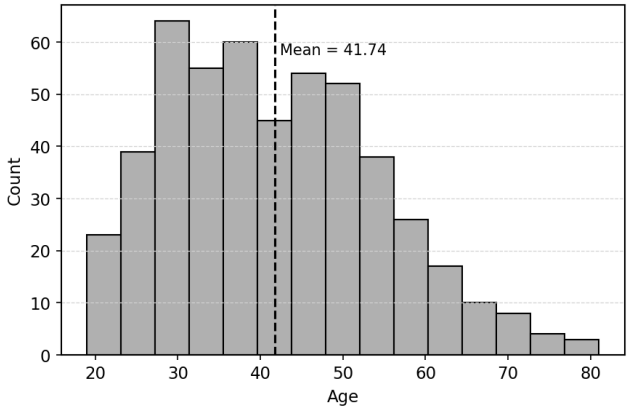


Figure A1: Age distribution of participants.

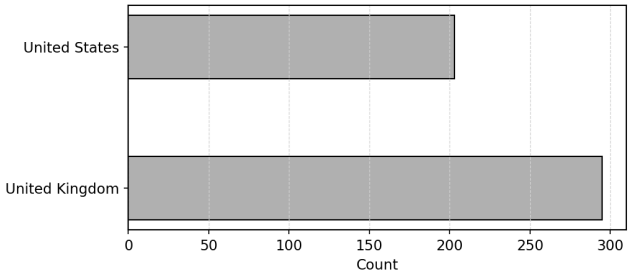


Figure A2: Distribution of participants by country of residence.

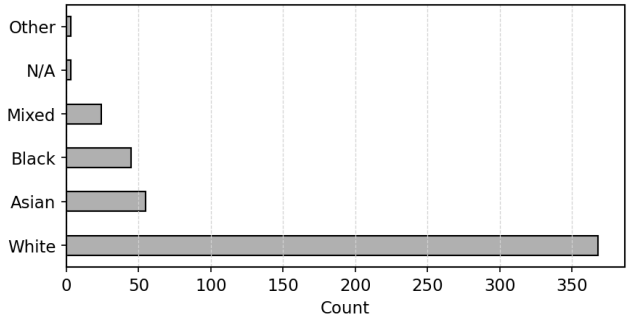


Figure A3: Distribution of participants by self-reported ethnicity.

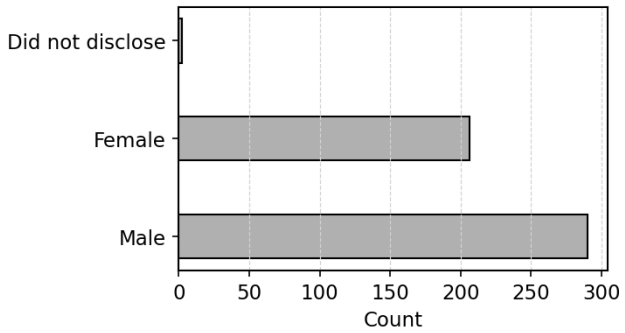


Figure A4: Distribution of participants by self-reported sex.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

Table A1: Exact p-values for preregistered hypothesis tests.

Hypothesis	Measure	Scenario	χ^2 (df)	Exact p
H1: Guardrails vs. No Guardrail				
H1	Trust	COVID	16.457 (1)	4.9771×10^{-05}
H1	Trust	Stocks	22.327 (1)	$2.2998e \times 10^{-06}$
H1	Accuracy	COVID	15.704 (1)	7.4049×10^{-05}
H1	Accuracy	Stocks	12.026 (1)	5.2451×10^{-04}
H1	Context	COVID	100.171 (1)	1.3978×10^{-23}
H1	Context	Stocks	113.079 (1)	2.0734×10^{-26}
H2: Context-rich vs. Random				
H2	Trust	COVID	1.8130 (1)	1.7813×10^{-01}
H2	Trust	Stocks	0.3240 (1)	5.6916×10^{-01}
H2	Accuracy	COVID	3.1100 (1)	7.7793×10^{-02}
H2	Accuracy	Stocks	11.176 (1)	8.2882×10^{-04}
H2	Context	COVID	1.7780 (1)	1.8237×10^{-01}
H2	Context	Stocks	0.3470 (1)	5.5556×10^{-01}
H3: Task Framing				
H3	Framing	COVID	0.5050 (3)	9.1782×10^{-01}
H3	Framing	Stocks	0.5500 (3)	9.0779×10^{-01}

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

Figure A5: Example of the guardrail representation types identified by previous work [LCKL25]. We build upon these results, identifying sampling strategies for the most effective Superimposed Primary Data condition.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

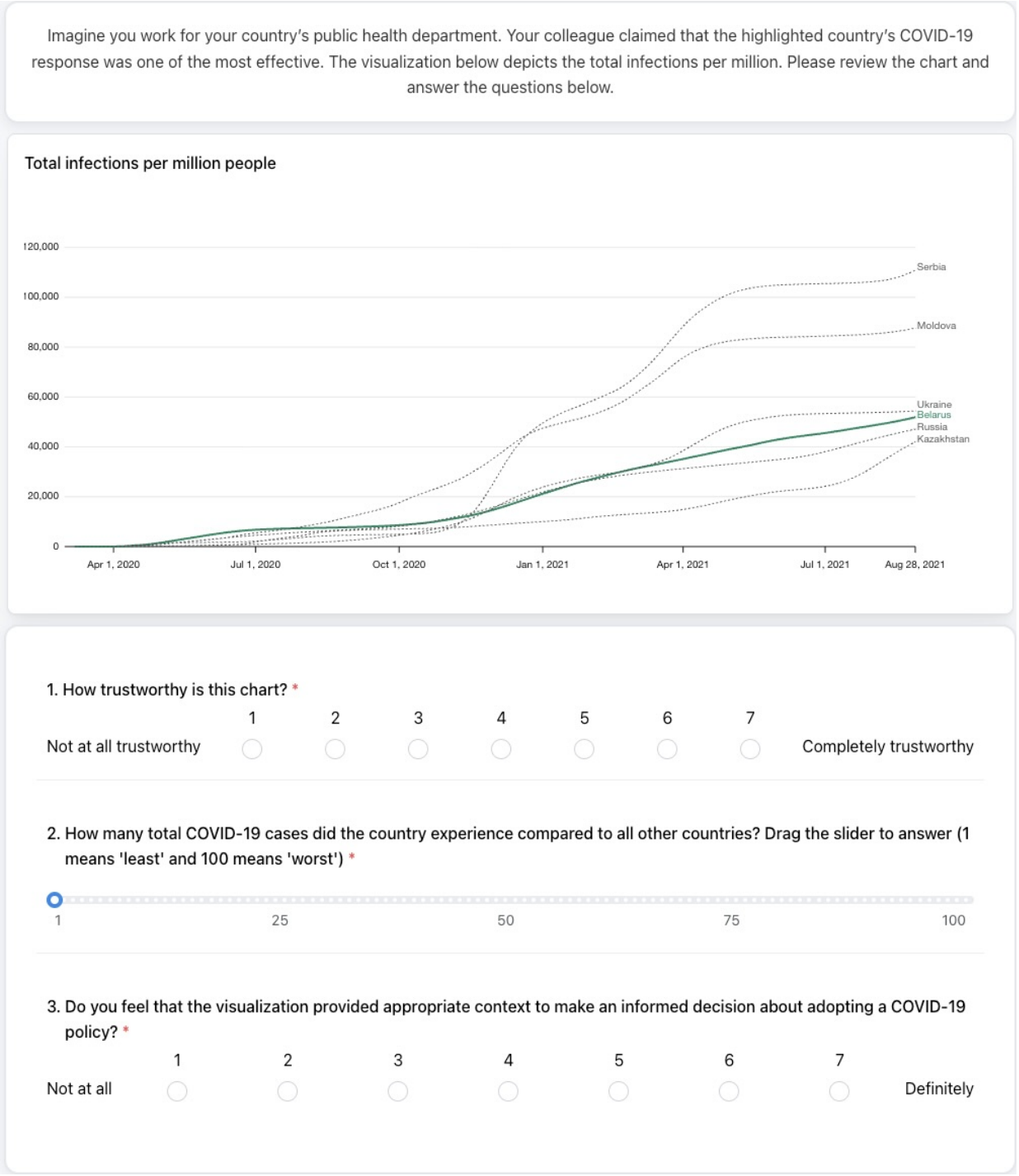


Figure A6: Screenshot of the Task 1 stimulus and survey in the COVID-19 scenario.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

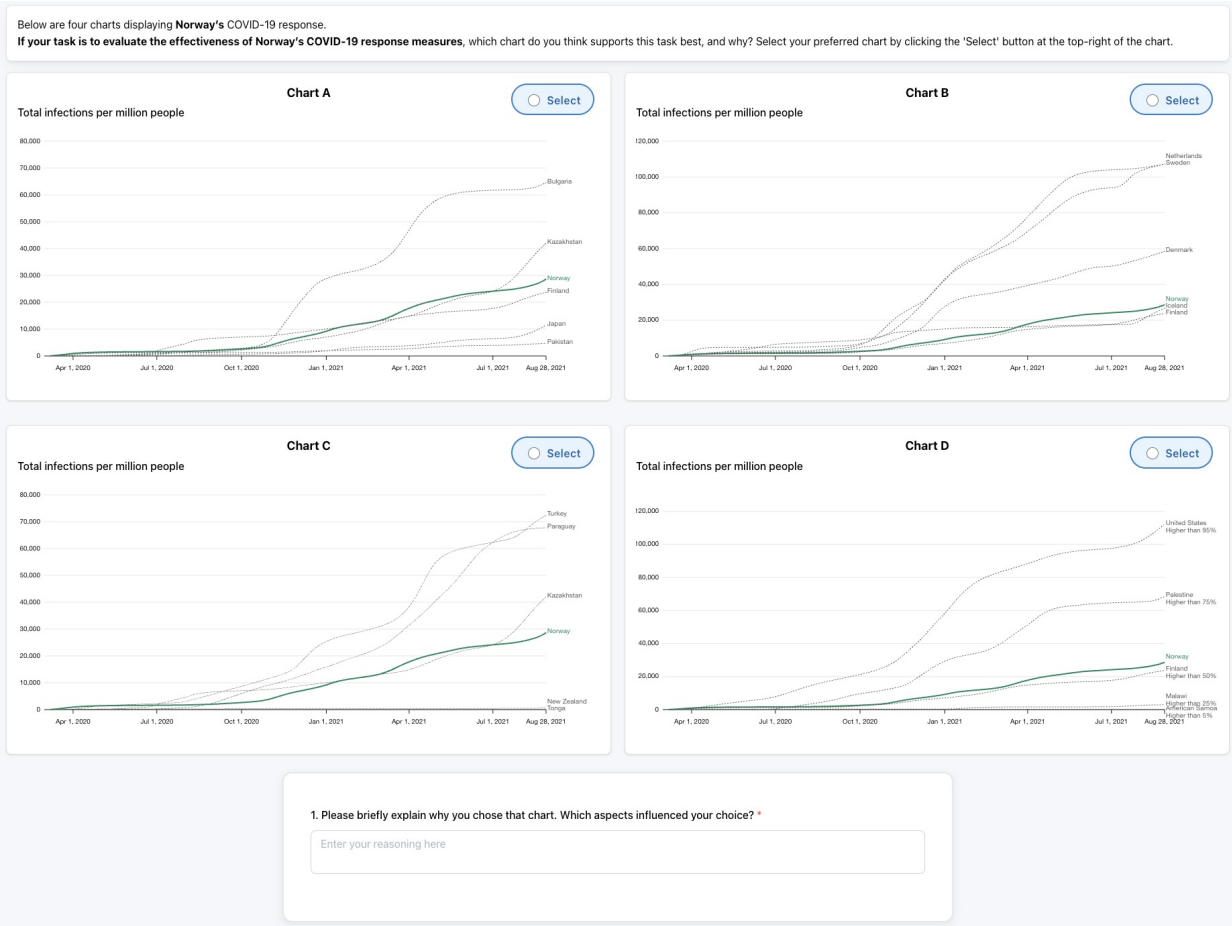


Figure A7: Screenshot of the Task 2 stimulus and survey in the COVID-19 scenario.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

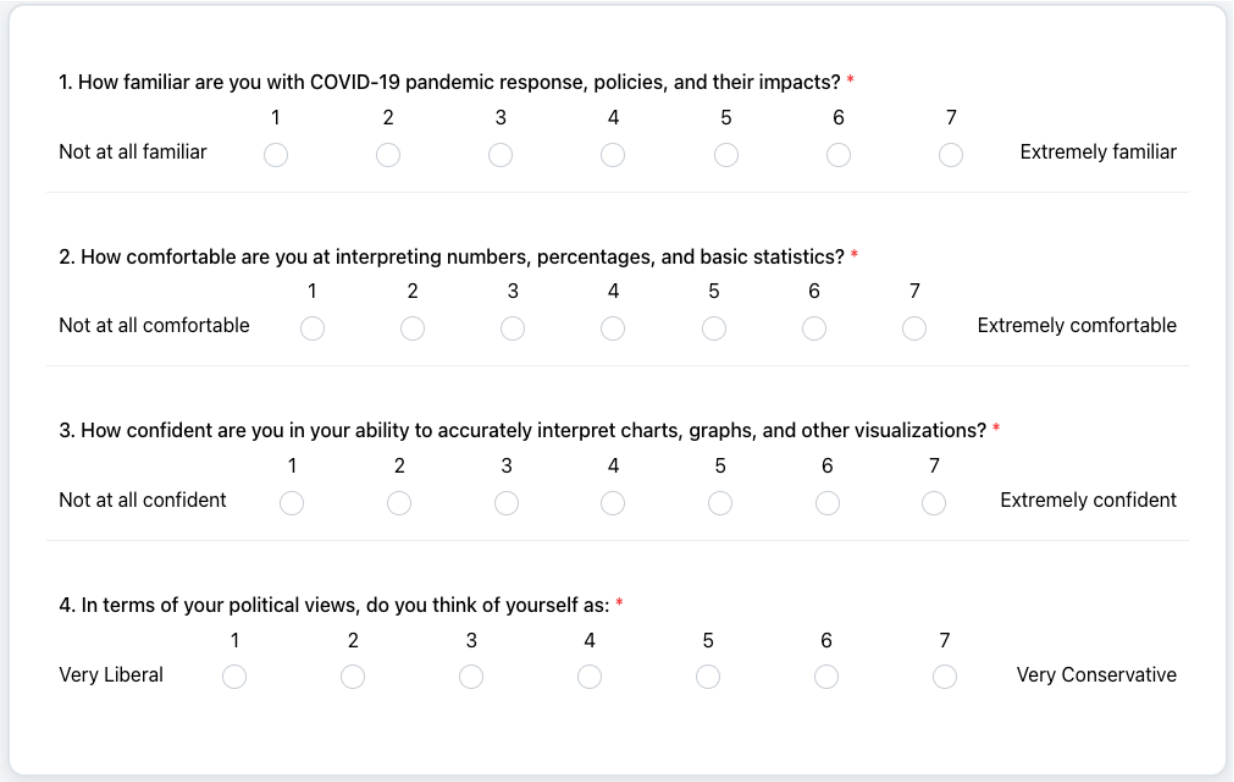


Figure A8: Screenshot of the post-study survey in the COVID-19 scenario.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

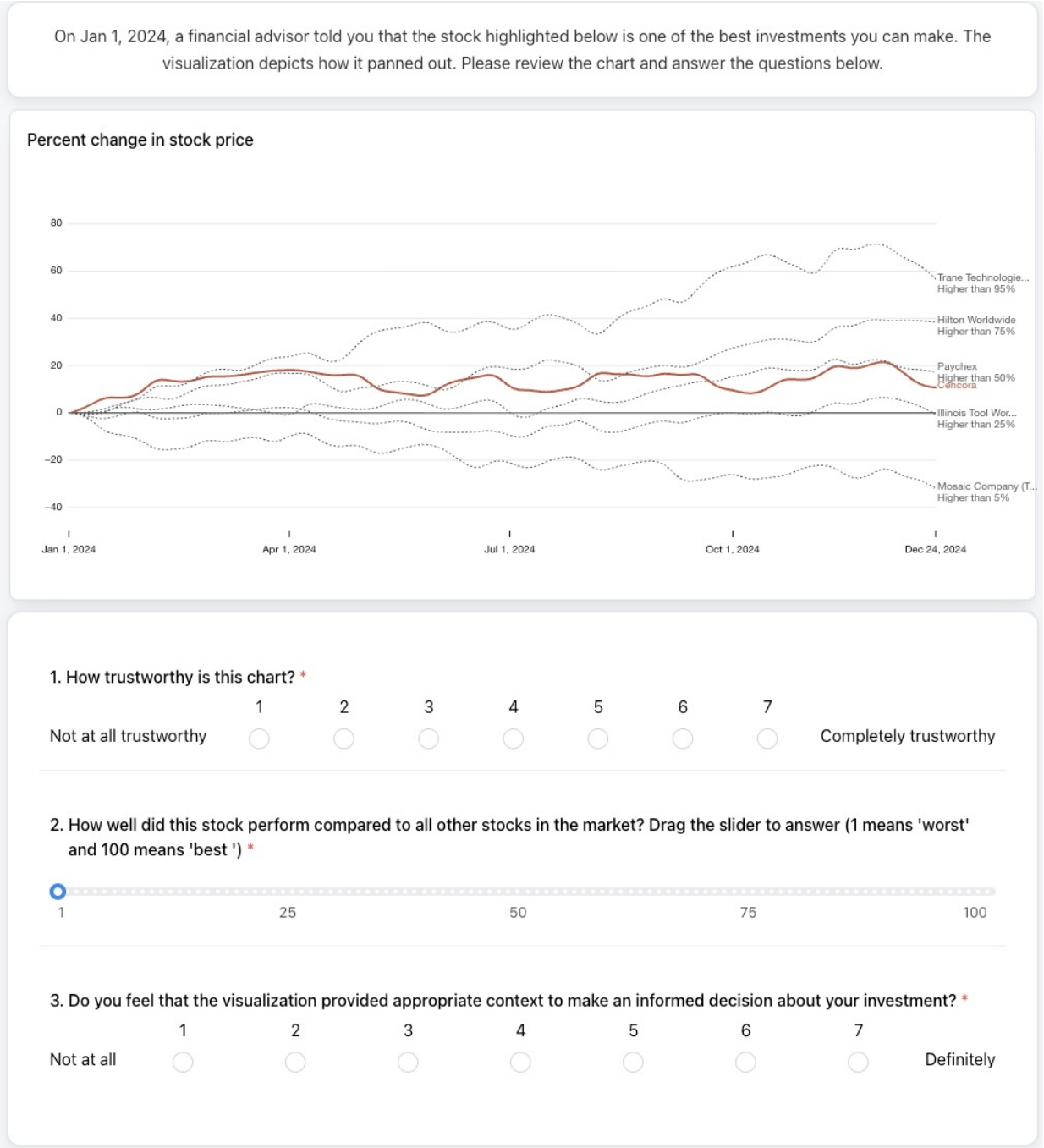


Figure A9: Screenshot of the Task 1 stimulus and survey in the Stocks scenario.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

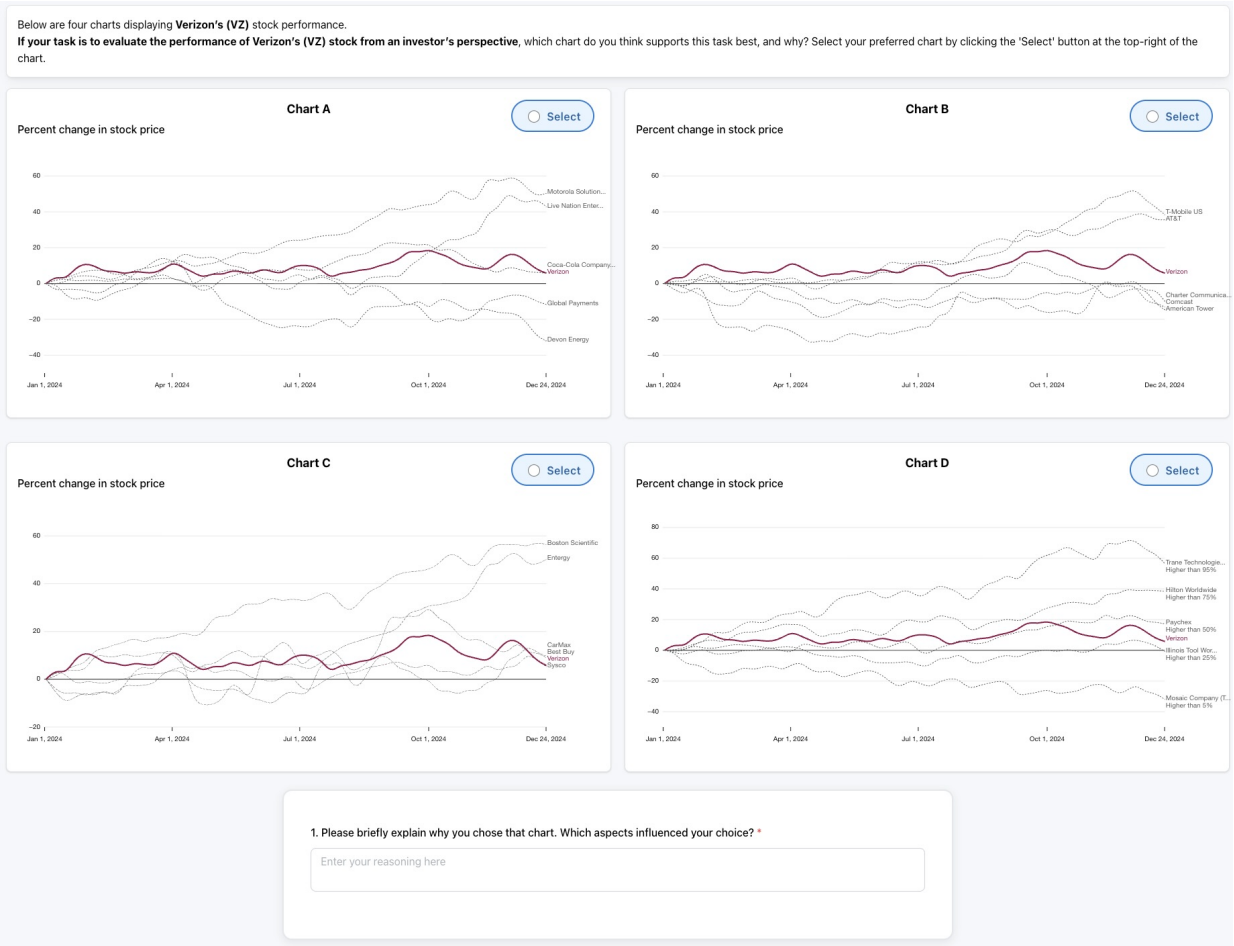


Figure A10: Screenshot of the Task 2 stimulus and survey in the Stocks scenario.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

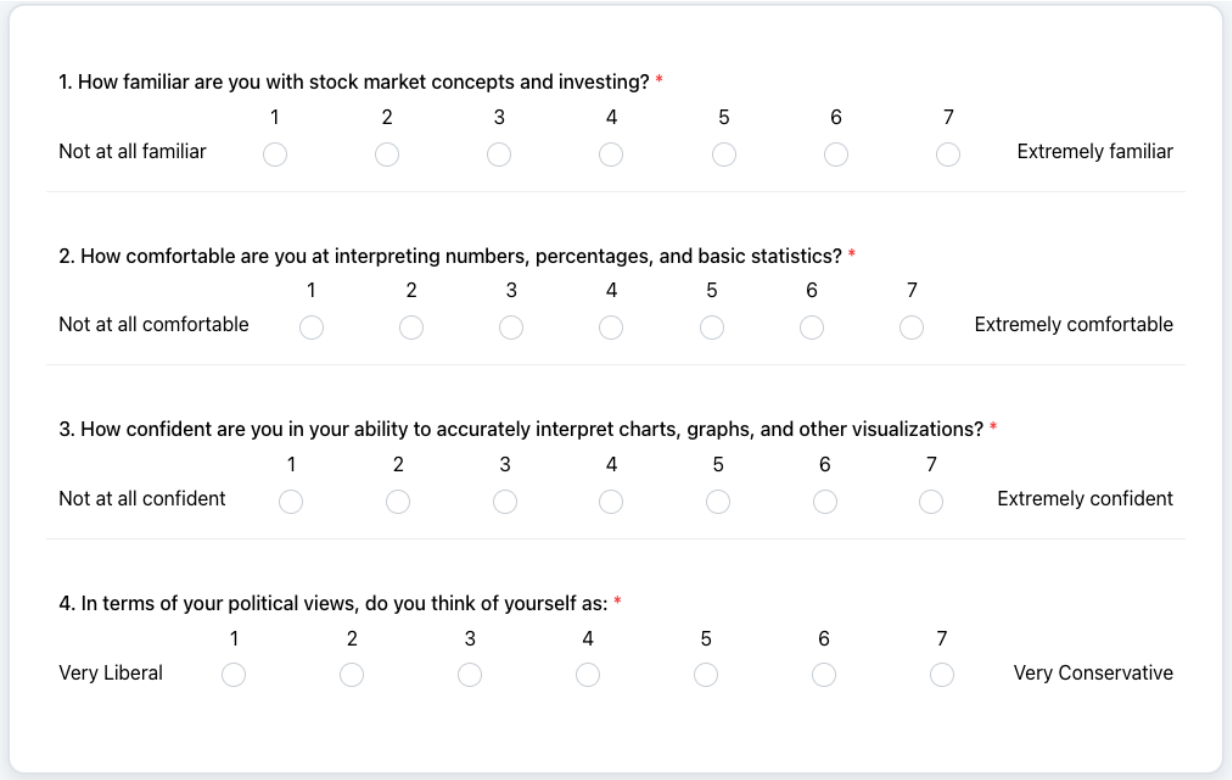


Figure A11: Screenshot of the post-study survey in the Stocks scenario.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

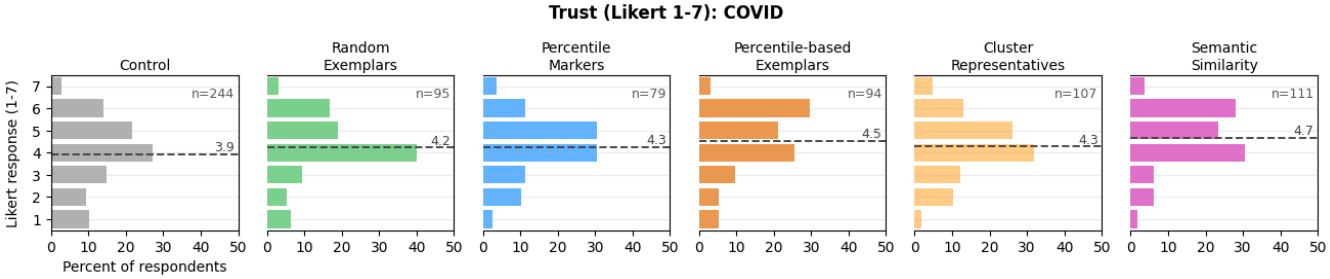


Figure A12: Distributions of Trust responses by guardrail selection method in the COVID-19 scenario in Task 1.

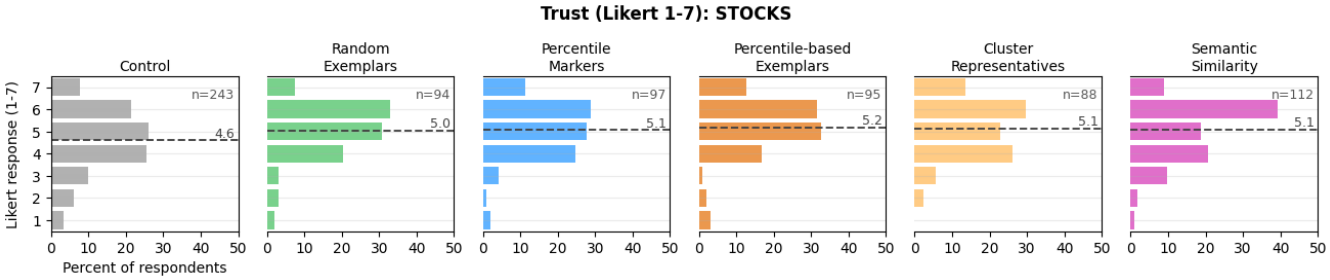


Figure A13: Distributions of Trust responses by guardrail selection method in the Stocks scenario in Task 1.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

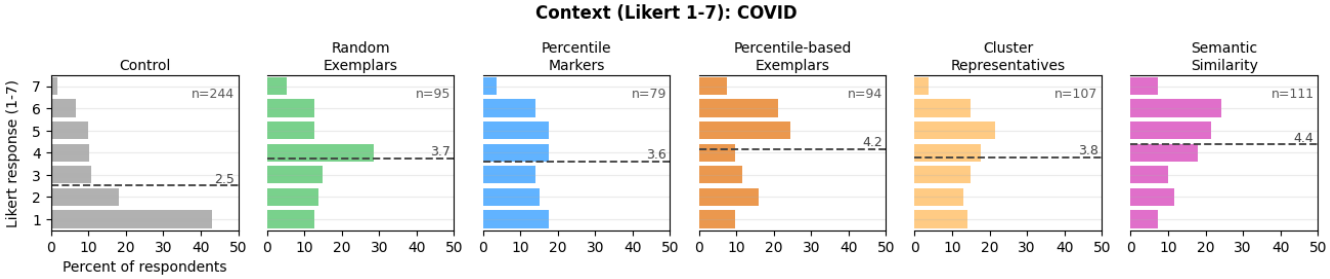


Figure A14: Distributions of Context responses by guardrail selection method in the COVID scenario in Task 1.

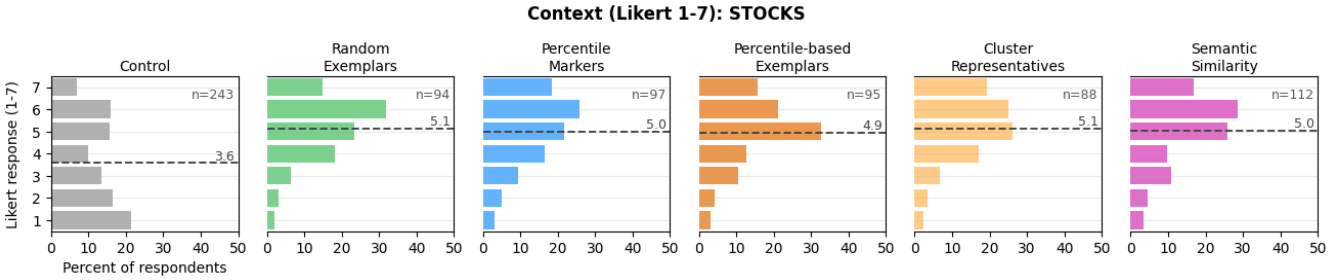


Figure A15: Distributions of Trust responses by guardrail selection method in the Stocks scenario in Task 1.

© 2026 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.