Vision, Modeling, and Visualization (2025)
B. Egger and T. Günther (Eds.)

Val-LLM: A Visual Analytics Approach for the Critical Validation of LLM-Generated Tabular Data

Madhav Sachdeva[1], Christopher Narayanan[1], Marvin Wiedenkeller[1], Jana Sedlakova[1,2], and Jürgen Bernard[1,2]

[1]University of Zurich, Switzerland; [2]Digital Society Initiative, Zürich, Switzerland

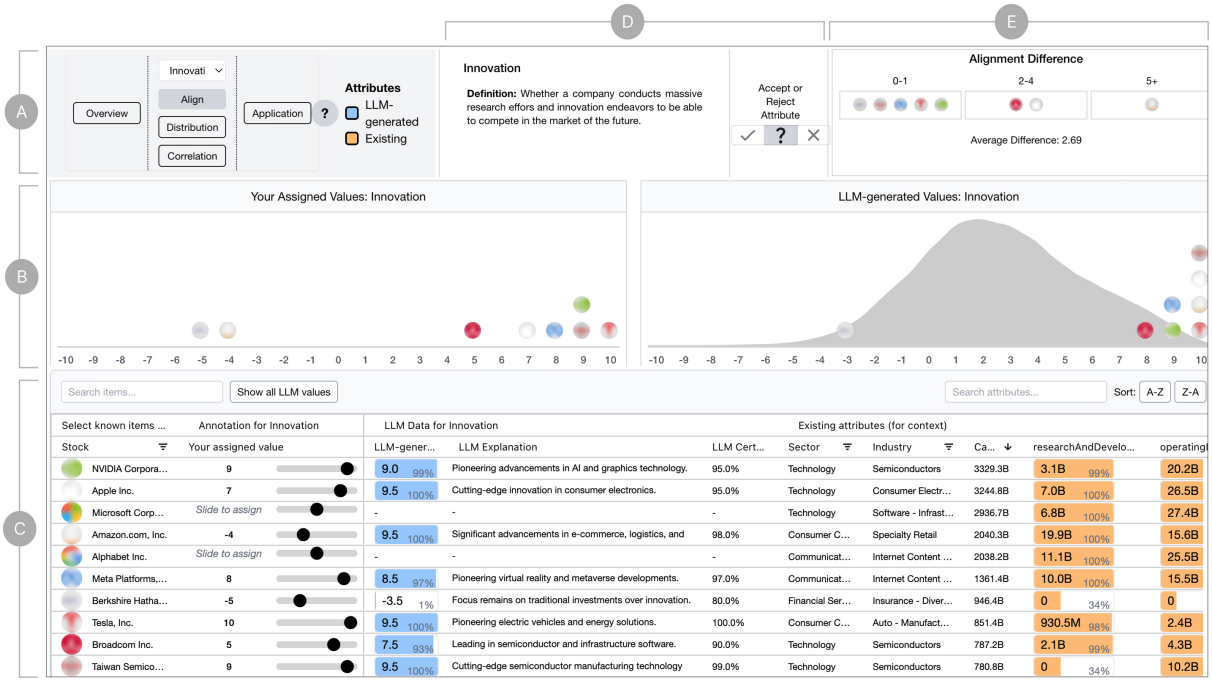

Figure 1: *The Val-LLM interface for the critical of LLM-generated tabular data. (A) Users can navigate the validation workflow from the Overview to three complementary validation views at different analysis granularities (Align, Distribution, and Correlation), and assess the downstream utility of the generated data in the Application. The Align View is active, enabling humans to assign data values through knowledge externalization for a series of items (C), as a basis for the LLM-based data validation through critical comparison (B). Users can review the differences between these aligned values at a glance (E) and make an informed decision on whether to accept or reject the LLM-generated data attribute (D) for downstream analysis.*

Abstract

Large Language Models (LLMs) are emerging as promising approaches for tabular data generation and enrichment, helping to ease constraints related to data availability. However, the reliable use of LLM-generated data remains challenging, e.g., due to hallucinations and inconsistencies. While some validation approaches exist, five key challenges remain: the lack of explanations and transparency in how values are generated, balancing fine-grained accurate with coarse-grained scalable validation, validating generated data without ground truth, and evaluating plausibility, semantic relevance, and downstream utility. To address these challenges, we present Val-LLM, a novel visual analytics approach for the critical validation of LLM-generated tabular data. Val-LLM enables users to contextualize generated data values with explanations, externalize human expert knowledge, relate LLM outputs with existing data, and assess the data utility in an application downstream. We conducted a user study to evaluate Val-LLM. Results highlight the usefulness of supporting multiple levels of granularity and enabling human knowledge externalization for validation. The study also indicates the need to study validation workflows and workflow flexibility, based on user domain experience and user preferences. Our work supports the trustworthy and effective use of LLM-generated tabular data by integrating visual analytics for systematic data validation.

CCS Concepts

• ***Human-centered computing*** → ***Interactive systems and tools; Visual analytics;***

DOI: 10.2312/vmv.20251235

1. Introduction

Large Language Models (LLMs) are a promising approach for tabular data generation [DW24, SZZ*24]. In data-limited domains, practitioners are increasingly exploring the use of LLMs to generate tabular data that reflects real-world properties [JJC*24]. Furthermore, LLM-generated data opens up opportunities for rapid data-centric hypothesis generation, customized dataset creation, and domain-specific data fine-tuning, helping analysts work with data that better reflects real-world contexts [FXT*24]. However, integrating LLM-generated data into analytical workflows introduces major challenges across domains, including concerns about hallucinations [JJC*24], inconsistencies [XLS*24,LZF*25], and different types of biases [BWSG25, ZYJ*24, WA25, Res25], requiring careful assessment. Our goal is to support practitioners in the process of critically validating such data.

To that aim, five key validation challenges must be addressed. First, the lack of transparency in LLM decision-making processes hinders trust, leaving questions about why or how specific values were generated open. Providing explanations and model confidence scores can help address this issue ($\mapsto R_1$). Second, the absence of ground truth values impedes users' ability to build confidence in the validity of the generated data, as many real-world scenarios lack reference datasets, making it difficult to assess accuracy beyond manual sanity checks. Human-centered support for ground truth generation for well-known items could support systematic validation workflows, reflecting user-defined expectations and domain expertise. ($\mapsto R_2$). Third, users are tasked to validate both at fine-grained levels (e.g., individual items) for plausibility with high accuracy and at coarse-grained levels (e.g., statistical distributions) for broader error-pattern detection at scale. Multi-granularity support will enable users to validate more systematically ($\mapsto R_3$). Fourth, it remains difficult to determine if LLM-generated tabular data is plausible, semantically relevant, and practically useful for existing datasets and downstream applications. Systems should provide validation mechanisms that relate generated data to the structure and semantics of existing data ($\mapsto R_4$) and support assessments within actual data analysis contexts (e.g., application) $\mapsto R_5$). Together, these challenges define a complex problem space that demands Visual Analytics (VA) solutions, combining automated methods with human expertise and interactive control.

To address these challenges, we propose `Val-LLM`, a VA approach for the critical validation of LLM-generated tabular data. Based on a workflow and task abstraction, we designed and developed the `Val-LLM` VA interface, implementing the validation workflow using the tasks as design targets. `Val-LLM` enables experts to gain an overview of the space of LLM-generated attributes, identify and select an attribute, and critically validate its values along three essential analysis granularities: from fine-grained assessment of the alignment of LLM-generated values with human knowledge, to the critical assessment of value distributions for many/all items, to the scalable validation of many LLM-generated attributes with many pre-existing data attributes. The workflow is completed by applying the LLM-generated data to a real-world application, as financial stock data, used in this work as the primary case, for its practical relevance and task clarity. Our evaluation strategy includes a usage scenario and a mixed-methods user study with a qualitative and quantitative assessment of `Val-LLM`. Results highlight the usefulness of `Val-LLM` in its support for critical validation workflows, and indicate important avenues for future work. Our contributions are as follows.

- Conceptualization of the critical validation of LLM-generated tabular data, integrating item-based (values and value distributions), and attribute-based assessments in a structured workflow aligned with corresponding analysis tasks, to address remaining challenges and validation requirements.
- Implementation of `Val-LLM`, a novel VA approach that facilitates the critical validation of LLM-generated tabular data through multi-granularity support, explanations of LLM values, human knowledge externalization, and a classification-based application case on financial stocks data.
- Demonstration of the perceived usefulness of `Val-LLM` in a usage scenario and identification of generalizable insights from a user study with six participants and a critical discussion, informing the design and development of future approaches.

`Val-LLM` advances VA methods for interactively validating LLM-generated tabular data by enabling human-machine collaboration at multiple data granularities. It enables systematic error detection, fosters responsible use of generative models, and strengthens data-driven decision making in financial stock data and beyond.

2. Related Work

With the promising capabilities of LLMs as embedded knowledge bases [SAT*23], empirical studies have investigated LLMs for their capabilities to enrich data for various applications such as e-commerce for product listing enrichment [JJC*24] and metadata enrichment [MKSvO24, Kus24]. However, LLM-enriched data comes with multifaceted validation challenges [BTKC24, LWX*24]. LLMs can amplify biases present in their training data [GRB*24]. Bias can manifest in topics [BWSG25], ordering of items [ZYJ*24], data representation [WA25], and statistical bias [Res25]. Other issues, such as hallucinations, occur when LLMs generate plausible but factually incorrect information. Several approaches are proposed to detect hallucinations in textual data. These include approaches that use model confidence scores [VYZ*23], external knowledge based approaches [MIT23], multiple prompting [MLG23] or semantic screening by another LLM [JJC*24]. Further, it has been demonstrated that LLMs lack in maintaining consistent relations across attributes [XLS*24, LXB25].

Understanding how and why LLMs generate tabular values is critical for building trust. Tools such as EvalGen [SZPH*24] and EvalLM [KLS*24] enable the evaluation of LLM-generated data by relying on LLM-as-a-Judge [ZCS*23] approaches, in which one LLM is used to validate another. However, these methods offer limited transparency into the generation process, making the underlying value generation difficult to interpret. Validating LLM-generated tabular data requires balancing fine-grained checks with coarse-grained assessments. Tools like EvalGPT [YLL*24], LLM Comparator [KTP*24], iScore [CHM*24], CoEval [LCKB23] support evaluation at item and distribution-level granularities, primarily for text generation tasks. Ferret [LSPL23] highlights the task of structured inspection when the data source is unknown, focusing on detecting manipulations. While these systems inspire our work, many typically lack structured support for tabular data, or target intentionally falsified input.

Fan et al. [FSM24] verify LLM-generated data by referencing values from known data lakes. However, in the absence of ground truth, tools such as EvaluLLM [PAD*24] and EvalGen [SZPH*24]

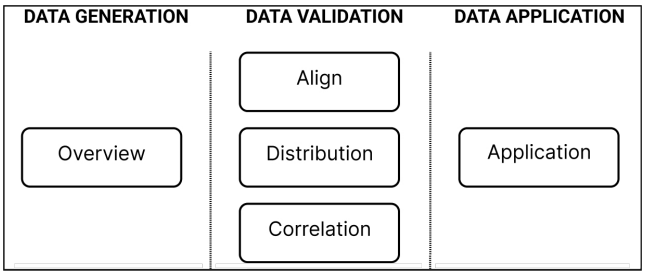

Figure 2: *The Validation Workflow of* `Val-LLM` *consists of three consecutive phases, ranging from support for LLM-generated data, core data validation activities, to the application of enriched data.*

use human knowledge externalization to steer the LLM-as-a-judge evaluation criteria but do not use it to annotate values for items. Evaluating the plausibility of newly generated attributes is largely unexplored. Most works evaluate text output [MDS*25] or model behavior [CFB*23], but not the semantic integration of new attributes. Current approaches focus on validating LLM-generated data with metrics such as win rates [KTP*24, YLL*24], the outputs preferred by evaluation models or humans, but not in evaluating the downstream utility for data analysis tasks.

3. Workflow and Task Abstraction

Based on the literature review, the analysis of remaining validation challenges, and guidance received from a design space for the critical validation of LLM-generated tabular data [SNW*25], we outline five key requirements for systems enabling the critical validation of LLM-generated tabular data. We present the corresponding conceptual validation workflow, followed by abstracted tasks.

3.1. Validation Workflow

We present a workflow for systematic validation of LLM-generated tabular data, presented in Figure 2. It consists of three consecutive phases: *Data Generation*, *Data Validation*, and *Data Application*. The Data Validation phase is further subdivided to account for validation tasks at different levels of analysis granularity. Our `Val-LLM` interface, presented in Section 4, utilizes and implements this workflow as an interactive navigation compass for process control; in the VA interface, always present at the upper left.

Data Generation: As an upstream phase, tabular data is generated by prompting an LLM. This includes defining a new attribute with a name and a precise description of its intended semantics. The prompt also specifies the output format, textual explanations, and model confidences for interpretability (R_1) [SNW*25] (see supplemental materials). The set of LLM-generated attributes results in an overview analysis step, as a basis for targeted validation activities. Our workflow does not make recommendations for the type of LLM or its form of interaction. In turn, our emphasis is on the LLM output, as direct input for critical validation workflow activities.

Data Validation: This core phase in the workflow includes activities of users for the data validation at different levels of analysis granularity, ranging from fine-grained assessments with high accuracy to coarse analysis to optimize for scalability.

- *Align:* One workflow step is on validating individual LLM-generated values for small item sets, using sanity checking methods (R_2). Human knowledge can be externalized when no ground truth exists, followed by comparing it to LLM outputs (R_3).
- *Distribution:* For a coarser and more scalable validation (R_2), the workflow supports analyzing entire value distributions and comparing distributions of selected item subsets. Subsets can also be created to assess LLM confidences, explanation topics, or known attribute properties (R_1).
- *Correlate:* Validation reaches the most scalable level (R_2) by relating LLM-generated attributes to other dataset attributes to detect plausibility errors and semantic alignment (R_4).

Data Application: LLM-generated data is combined with existing data in downstream analyses. The analysis goal with this enriched tabular data is agnostic to the upstream validation activities and can reflect any usage context of tabular data. In this work, we focus on financial stocks data, applying the LLM-enriched data in a buy-hold-sell classification scenario, as a representative supervised machine learning example in the finance domain. In general, the effect and impact of LLM-generated attributes can be validated by comparing models and outcomes between the original dataset and the enriched dataset, including the LLM-generated attributes (R_5).

3.2. Task Abstraction

Our validation workflow breaks down complex validation challenges into tangible phases and steps, easing the abstraction of tasks. This process was further informed by a design space for the critical validation of LLM-generated tabular data [SNW*25].

T1 **Gain Overview of Attributes**: Users can explore LLM-generated attribute output through a structured overview. This overview includes the validation state (accept/reject), where reject may promote re-prompting attribute values in the upstream LLM process. Other important attribute information includes their description, value distribution, and correlations to other attributes (LLM-generated or pre-existing). (R1)

T2 **Externalize Values**: Users can externalize their knowledge through known-item annotation for well-known items [SNW*25], identified through search and filtering. Human-created values can serve as ground truth for comparisons with LLM-generated values and their assessment through fine-grained sanity and plausibility checking. (R2)

T3 **Inspect Value Distributions**: Users inspect the value distribution of an LLM-generated attribute and compare it with distributions of item subsets, enabling scalable attribute-level analysis to reveal overarching patterns [SNW*25]. (R3)

T4 **Correlate Attributes**: Users relate LLM-generated attributes with multiple other attributes (LLM-generated or pre-existing) [SNW*25] to assess relation consistency, assess attribute plausibility, and identify potential mismatches. (R4)

T5 **Analyze Data Utility**: Users can validate the utility of LLM-generated data directly in analysis applications. The focus is on revealing the impact of the newly introduced LLM-generated data attributes by comparing analysis outcomes before and after their integration. This can be done by comparing output values or by applying quantitative quality measures. (R5)

4. `Val-LLM` Interface

`Val-LLM` enables validating LLM-generated tabular data through five views, each supporting one task abstracted in Section 3.2: Attribute Overview ($\mathbf{T}_1$, Figure 3), Align View ($\mathbf{T}_2$, Figure 1), Distribution View ($\mathbf{T}_3$, Figure 4), Correlation View ($\mathbf{T}_4$, Figure 5), and

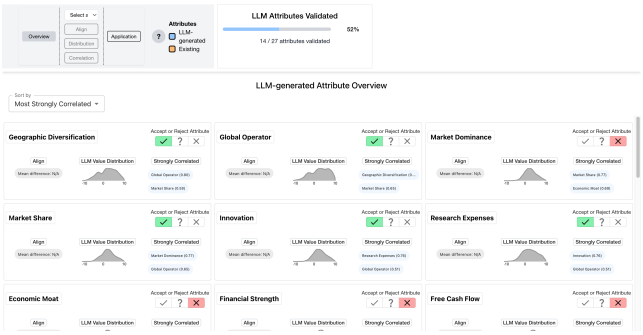

Figure 3: *Attribute Overview interface of* `Val-LLM`*. Each attribute is displayed as a card that summarizes the name, validation status, strongly correlated attributes, and interactive links to key validation views (Align, Distribution, and Correlation).*

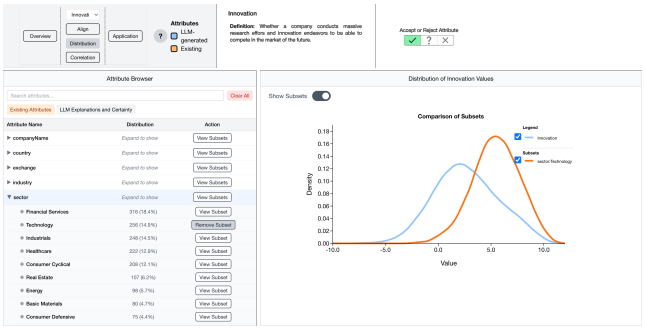

Figure 4: *Distribution View interface showing the value distribution of a focus attribute (blue). Users can browse other attributes (left) to identify subsets of interest (like Technology sector), to add subset distributions for comparison (right, orange distribution).*

Application View ($\mathbf{T}_5$, Figure 6). Users navigate views using the global navigation control always at the top-left, implementing the validation workflow (Figure 2) as an interactive compass.

We study `Val-LLM` in the context of financial stock data. In our approach, the prompt requests numeric output scores for attributes ranging between -10 (worst) and 10 (best), a textual explanation for the assigned value, and the LLM's confidence (0-100). This score range was chosen as a synthetic interpretable scale to facilitate relative assessments of company performance across multiple stocks. It allows for straightforward attribute comparisons and helps to keep the focus on the study of novel validation techniques, rather than distractions through data preprocessing operations like attribute normalizations. The full prompt is provided in the supplemental materials. Despite the running example with financial stock data, the workflow, validation tasks, and interface designs are completely domain-agnostic.

4.1. Attribute Overview

The Attribute Overview presents LLM-generated attributes with their value distribution, alignment with ground truth (if applicable), and most correlated existing attributes (see Figure 3). Each LLM-generated attribute has a validation status (accept, undecided, reject). A progress bar (center top) tracks how many attributes users have already accepted/rejected through validation activities. Selecting an attribute starts the validation process in the Align View.

4.2. Align View

The Align View is named after the alignment of human-generated values with LLM-generated values, enabling critical validation of single values through visual comparison (see Figure 1). Users inspect a table of items (bottom) with columns for the existing attribute context. They can search, filter, and sort items by name or attributes. For known items of interest, users can externalize their knowledge by assigning a value using a range slider. Two value distributions at the top of the Align View enable the comparison of all human-externalized values (Your Assigned Values, left) with aligned LLM-generated values (LLM-generated Values, right). The Alignment Difference panel (top right) summarizes the differences, grouped into near matches (left), minor differences (center), and large differences (right). These visualizations help users decide whether to accept or reject the LLM-generated value.

4.3. Distribution View

The Distribution View enables distribution-based validation for multiple/all items for a selected attribute (see Figure 4). It shows the overall value distribution of a selected LLM-generated attribute, enabling the assessment of distributional skew and the identification of potential distributional divergences. In addition, users can dynamically add the distributions of user-created item subsets (by different criteria), enabling the visual comparison of different subset distributions, e.g., to identify biases. Criteria for subset creation are i) any category of a categorical attribute, ii) ranges (low/mid/high) of any numerical attribute, iii) confidence levels (low/mid/high) of the LLM, and ultimately, topics from textual explanations of the LLM derived through topic modeling. For example, users can examine whether distributions of LLM scores differ across regions like Europe, the USA, or Asia.

4.4. Correlation View

The Correlation View supports the scalable validation at the between-attribute granularity, by examining the relationship between the focused LLM-generated attribute and other attributes (see Figure 5). It presents a list of attributes on the left, sorted by correlation strength. A toggle allows correlation analysis also for other LLM-generated attributes. Focusing on a second attribute of interest reveals their relation in a scatterplot on the right (here: the LLM-generated Innovation attribute and the dividend Yield). Patterns selected in the scatterplot reveal corresponding items in a table below, including their attribute details.

4.5. Application View

The Application View enables assessing the impact of LLM-generated data on a focused data analysis scenario. Our implementation showcases a buy-hold-sell classification task on stock data (see Figure 6), i.e., a supervised machine learning scenario. The view uses a matrix idiom to compare the agreement of two classification runs, based on two different, user-selectable attribute sets: Validated, Existing (default), Existing + Validated, LLM-generated, Existing + LLM-generated (default). Statistical information about the agreement between results is shown at the top. Users can conduct detailed on-demand reasoning about classified item subsets by clicking matrix cells, e.g., where results disagree (like for the selected cell with 303 items). A table on the right

Figure 5: *Correlation View interface of* `Val-LLM`*. An attribute browser (left) indicates the correlation between the focused attribute and all other attributes. The scatterplot displays its correlation to a second user-selected attribute. Users can select items through lasso selection, which are then shown in the lower table.*

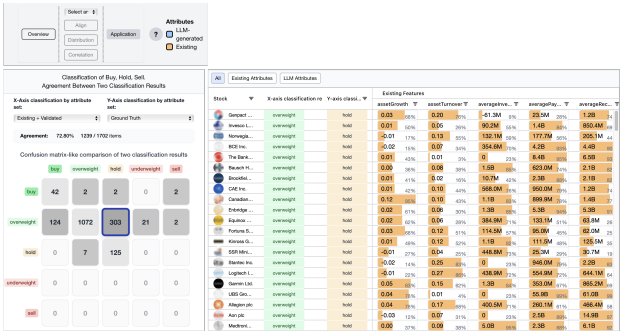

Figure 6: *Application View interface of* `Val-LLM`*. Users can compare the classification performance of validated or non-validated attributes with the ground-truth through a matrix (left). Darker cells indicate higher misclassifications. Users can inspect the misclassified items in the table (right).*

represents selected items, their attributes, predicted classification, ground truth labels, and existing attributes for contextualization.

5. Usage Scenario

We present a usage scenario demonstrating the usefulness of `Val-LLM`. The scenario accompanies Linda, a financial expert analyzing stock data. Through various APIs, she already accesses many financial attributes. However, she lacks specific types of information that are not easily available through these paid services. This includes the innovation, sustainability, leadership, customer satisfaction, supply chain dependencies, and geopolitical risks of companies/stocks. Given the volume of stocks she analyzes, manually creating, labeling, and curating such attributes is too costly and time-intensive. She uses an LLM to generate a series of missing data attributes. She is now concerned about the reliability of these LLM-generated tabular data, concerned about issues like hallucinations, potential distribution divergences, and biases. She decides to use `Val-LLM` to solve her data validation problem.

Her validation process starts with an overview of the LLM-generated attributes (Figure 3). Along with the attributes, Linda also identifies the functionality to determine whether to accept or reject the LLM-generated attribute for her analyses, after critical validation. As a next step, Linda will decide which attribute to validate systematically. She applies attribute filtering and sorts the LLM-generated attributes by their correlation strengths. One attribute with high correlations that she considers essential in evaluating stocks is the "*Innovation*" of stocks. By clicking on the *Innovation* attribute, she is directed to the **Align View**.

The **Align View** supports Linda in value-based assessments of LLM-generated data through informed sanity checks (Figure 1). To do so, she introduces her own knowledge to the analysis by externalizing *Innovation* values for stocks of interest. To find stocks for value externalization, she uses the search bar and applies filters, and uses the slider toggle to annotate stocks of interest. For each annotated stock, she receives instant feedback, unveiling the LLM's value, confidence, and explanation, and how she and the LLM-generated value align. Using the NVIDIA stock as an example, her assigned value is 9.0, representing a high *Innovation* score (ranging from -10.0 to 10.0). The comparison with the LLM's output reveals an exact match, so her value expectations match the LLM-generated output, helping her in the trust-building process. After a series of annotations, she identifies in the Alignment Difference panel (top right) that the model broadly aligns with her knowledge, even if not always exactly. It stands out that there exists an anomaly (mockup-anomaly, for demonstration purposes) for the Amazon stock: a value difference of -4.0 (Linda) versus 9.5 (LLM). It is now up to Linda to draw meaningful conclusions from this deviation in human-machine collaboration: either to reconsider her own judgment, or to inspect the textual LLM explanation ("Significant advancements in e-commerce (...)") and possibly make changes to the prompt. Beyond this stock, she does not notice any other anomalies while annotating and decides to proceed with the validation at a coarser granularity to increase validation speed. Using the workflow compass interaction, she navigates to the **Distribution View**.

In the **Distribution View** (Figure 4), Linda inspects the distribution of all Innovation values (blue). Linda wants to see if subsets of interest yield a different distribution, to check for plausibility, or for biases. Here, she creates an item subset of all tech stocks (attribute subset browser on the left), assuming that this generates a higher distribution of *Innovation* values. She added the *Technology* subset to the view on the right (orange): The technology sector does indeed score higher than the average of all stocks, which she finds plausible, contributing positively towards accepting the LLM-generated attribute. She now proceeds to the **Correlation View**, aiming for validations at scale.

When analyzing the correlations of the *Innovation* attribute with all other attributes (Figure 5), Linda identifies that *dividend yield* and *research and development* expenses are correlated, albeit correlation strengths are moderate to low. In the scatter plot for *Innovation* versus *dividendYield*, she performs several lasso selections on extreme ends, to identify which stocks had low *innovation* but *high research and development expenses*, and vice versa. She studies the details of selected items in the table below, providing additional item context for her selection-based validation activities. Having detected and validated expected correlations of the *Innovation* attribute, she decides to accept the attribute. Along similar lines, Linda proceeds with other LLM-generated attributes until she reaches a set of accepted LLM-generated attributes for application.

Linda now uses the **Application View** (Figure 6) to see the LLM-generated attributes in action, i.e., in her actual analysis scenario on buy-hold-sell classification. She decides to compare the

classification results of buy-hold-sell decisions for two attribute sets: Existing + Validated (x-axis) and her Ground Truth, obtained from her reliable reference dataset (y-axis). She identifies an agreement of 73% of classification results. Using the matrix visualization, she explores disagreement in the overweight- hold cell, with 303 stocks. She selects the cell and inspects these items on the table (right), revealing important context information about attributes (features) involved in the classification process.

Overall, Linda was able to validate the *Innovation* attribute (and others) through multiple granularities. Along the process, she externalized her knowledge, received immediate feedback and LLM-explanations, followed by structured visual comparisons and correlations that helped her build confidence in the LLM-generated data.

6. Evaluation

We conducted a user study to assess the usefulness and usability of `Val-LLM`, using a mixed-methods design.

6.1. Experiment Design

Participants We recruited (n=6) participants through the authors' academic network, including colleagues, students, and researchers. The inclusion criteria were familiarity with LLMs to ensure they could provide relevant feedback, and a personal interest in stocks and stock characteristics to maintain a decent degree of expertise.

Procedure The evaluation followed a mixed-methods design. First, we introduced the background and aim of `Val-LLM` to the participants. Next, we presented the `Val-LLM` VA interface through a short tutorial case. Participants were then invited to use the system freely without instructions to get familiar with it. Next, participants performed tasks in a task-based assessment structured by the `Val-LLM` views. Finally, participants evaluated the system for the following usability and perceived usefulness measures:

- System Usability Scale (SUS) [B*96]: A standardized questionnaire measuring usability on a 5-point Likert scale.
- Technology Acceptance Model (TAM) [DBW89]: SUS extension, including 4 questions about perceived usefulness on a 5-point Likert scale.
- Questions that collect quantitative and qualitative data on the perceived usefulness of the validation views and workflow.
- Semi-structured interviews to elicit qualitative insights on the impact of addressing validation challenges.

The full set of questions and responses for each assessment is provided in the supplemental materials. In this main paper, we report the key findings. We conducted a pilot study (n=1) before the user study to refine the study design.

6.2. Results

Task-Based Assessment: In the Attribute Overview, all participants succeeded except one who struggled with identifying the strongest correlated attribute. In the Align View, all participants completed both tasks successfully. In the Distribution View, participants had minor struggles regarding the interpretation of distribution comparisons, with P5 and P6 failing in both tasks. In the Correlation View, participants found it difficult to find strong correlations (P2, P4, and P6) and locate items in the scatterplot (P3, P6). In the Application View, performances were mixed. Two participants (P2, P4) failed tasks on error analysis and classification agreement, suggesting these aspects were particularly challenging.

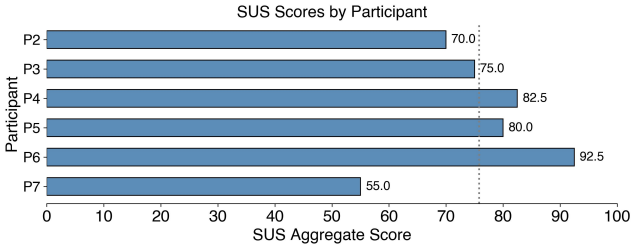

Figure 7: *System Usability Scale [B*96] aggregate scores by participant (P1 was piloting). The gray dotted line indicates the mean.*

System Usability Scale: The SUS scores ranged from 55 (P7) to 92.5 (P6) (Figure 7). The average score was 75.8, indicating moderate usability according to standard benchmarks. The variation reflects differences in perceived ease of use among participants.

Technology Adoption Model (TAM) Ratings: Participants rated the TAM statements positively, with an average score of 4.67 out of 5. The highest-rated aspect was the system's ability to improve validation against domain knowledge (mean = 5). The perceived usefulness for systematic validation and integration into workflows was rated slightly lower. The results suggest participants found `Val-LLM` useful for validation tasks.

Questions Collecting Qualitative and Quantitative Data: Participants found the tool effective in inspecting individual LLM-generated values (mean = 4.33) and finding trends across attributes (mean = 4.30). These findings were supported by the distribution plots present in the Overview, Align View, and Distribution View. Participants expressed difficulty in externalizing their reasoning (mean = 3.50), indicating a need for further support. P3 noted that only the Align View supported knowledge externalization through direct item annotation, but this support could be extended to other views. P3 also remarked that the LLM explanations seemed "high-level" and were not actively used, a sentiment partially echoed by P5 and P6. P5 and P6 highlighted that the system effectively provided context for assessing the reliability of LLM-generated data.

Semi-Structured Interview *Most Useful View*: The Align View was considered the most useful (3/6). Participants valued its support for direct, blind comparison between known values and LLM-generated values, and P3 valued its capability to reduce bias. P4 and P5 reported that externalizing knowledge through item annotation and comparing it with LLM-generated values increased confidence in the validation. The Distribution View was the second most useful (2/6), by participants preferring to start with a high-level overview before undertaking fine-grained validation. P2 appreciated its support for trend recognition, and P7 found it useful for identifying focus areas for inspection. The Correlation View was most valued by P6 for understanding relations between attributes.

Least Useful View: The Application View (2/6) and Distribution View (2/6) were mentioned twice, while the Attribute Overview (1/6) and Correlation View (1/6) were less common answers. P2 and P7 found the matrix idiom in the Application View difficult to understand. In the Distribution View, P5 did not understand the

impact of subsets, and P6 noted having limited insights. P3 found the Overview overloaded with information, reducing its effectiveness. P4 stated that the Correlation View required extensive domain knowledge and was less intuitive.

Re-prompting LLM: There is a high consensus that the Align View (5/6) motivates the re-prompting of the LLM to generate data, while P7 noted that the Distribution View motivates them. Participants emphasized that the Align View enabled direct comparison between item values, without being influenced by prior expectations. P2-6 noted that item-level alignment was crucial and provides actionable feedback to refine the generation process.

Workflow Preferences: Most participants (4/6) preferred a workflow from fine-grained (item-level) to coarse-grained (overview) validation. Participants argued that item-level validation builds confidence before generalization, and three participants noted that this workflow helps the process of validation. In contrast, P2 and P7 preferred starting from a coarse overview (Distribution View) and then drilling down for specific validations. In summary, while many participants preferred fine-to-coarse workflows, workflow flexibility can accommodate different user strategies and needs.

7. Discussion and Reflection

Using Scores as Attribute Values During the design of `Val-LLM`, we determined that LLM-generated values will be scores, in the range -10 to 10. This enables comparability across attributes and eases the study of validation methods. To be generalizable to other scales like market capitalization, temperatures, or velocities, the interface design would require slight modifications in its legends. Future work includes studying full user flexibility in defining attribute values, ideally across different application domains.

Workflow Granularity and Flexibility Participants particularly valued the validation functionalities at different levels of granularity: item-based (Align View), item distribution-based (Distribution View), and multi-attribute-based (Correlation View). Most participants preferred to start validation with fine-grained item-level checks, building confidence before moving to coarser levels. This preference may have been influenced by the layout of the Validation Workflow (Figure 2). Nonetheless, some participants preferred to start with coarse-grained overviews first. The diversity of observed validation behavior highlights the need for flexible workflows that support multiple entry points and allow iterative use across granularity levels. Future work could investigate controlled studies that isolate layout effects and investigate how the displayed workflow order could affect the preference in granularity assessment. A second factor influencing workflow preference was users' data literacy and domain expertise. These aspects appear to affect how much control users expect over the validation process. Interfaces should therefore support both bottom-up (drill-down) and top-down (aggregation) strategies to accommodate diverse validation behavior.

Usability and Usefulness Our results indicate moderate usability but high perceived usefulness of `Val-LLM`. Regarding the perceived usefulness, users appreciated the structured support for validating LLM-generated tabular data. However, usability was compromised as some views were perceived as complex. In particular, the correlation-based validation and matrix idiom required higher mental effort. Future work should explore designing validation systems tailored to specific user groups, ideally including non-experts, possibly through targeted design studies.

Human Knowledge Externalization: The Align View's blind comparison design (without revealing LLM-generated values) was valued for reducing confirmation bias and encouraging independent reasoning. A limitation is its reliance on users knowing specific items, which restricts its broader applicability. This could be addressed by extending the workflow with data exploration. Another extension is supporting uncertain expert knowledge, comparable to LLM confidence in value outputs. This opens a new dimension with promising validation possibilities, though it also introduces greater complexity in corresponding visualization designs.

Comparative Evaluation Our study does not include a direct comparison with other tools like LLM Comparator [KTP*24], EvaluLLM [PAD*24], or EvalGen [SZPH*24], these tools focus on different parts of LLM evaluation. Their approach evaluates different prompts, scores output, or tunes parameters instead of tabular data validation. Our approach complements these tools by enabling a visual approach to inspecting issues with LLM-generated tabular data. Future work in this direction could evaluate how `Val-LLM` integrates into LLM evaluation workflows, and compare which approaches potentially offer failure discovery in the workflow.

When LLM-Based Value Generation Fails Throughout this work, we gained experience with the output of LLMs and their quality, especially through the validation features of `Val-LLM`. We observed that value generation with LLMs is highly sensitive to the definition of target attributes. If upstream specifications are imprecise, the model reverts to linguistic heuristics, producing inconsistent and non-comparable values. Small changes in prompts or schema design can shift results and create spurious precision, which risks misleading analyses and decisions. We see potential for visual analytics in supporting precise attribute specification, scale design, and ensemble-based value creation with ensemble ranking.

8. Conclusion

We presented `Val-LLM`, a visual analytics approach for the critical validation of LLM-generated tabular data. Our approach addresses five key challenges by enabling scalable and fine-grained validation, supporting validation through human knowledge externalization, assessing plausibility, and evaluating practical utility through a downstream classification task. In the user study, we found that `Val-LLM` supports the process of critical validation of LLM-generated tabular data. Our findings highlight the importance of supporting multiple levels of granularity in validation systems, allowing users to form informed judgments incrementally. These insights inform the design of future visual analytics tools for validating LLM-generated tabular data. `Val-LLM` demonstrates the feasibility of combining visual analytics with validation strategies to enable reliable and effective use of LLM-generated data.

Supplemental Materials

Supplemental materials covering prompt for data generation, interface views, and user study data are available on OSF.

Acknowledgments

This work was supported by funding from UZH Candoc, grant no. [FK-25-014].

References

[B*96] BROOKE J., ET AL.: Sus-a quick and dirty usability scale. *Usability evaluation in industry 189*, 194 (1996), 4–7. 6

[BTKC24] BHARGAVA U., TERESHA Y., KOUL N., CHAVAN C. P.: Overcoming the challenges of large language models: Introducing a novel proposition for synthetic data validation. In *Big Data and Artificial Intelligence (BDAI)* (2024), IEEE, pp. 290–295. doi:10.1109/BDAI62182.2024.10692968. 2

[BWSG25] BAI X., WANG A., SUCHOLUTSKY I., GRIFFITHS T. L.: Explicitly unbiased large language models still form biased associations. *National Academy of Sciences 122*, 8 (2025), e2416228122. doi:10.1073/pnas.2416228122. 2

[CFB*23] CABRERA Á. A., FU E., BERTUCCI D., HOLSTEIN K., TALWALKAR A., HONG J. I., PERER A.: Zeno: An interactive framework for behavioral evaluation of machine learning. In *CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–14. doi:10.1145/3544548.3581268. 3

[CHM*24] COSCIA A., HOLMES L., MORRIS W., CHOI J. S., CROSSLEY S., ENDERT A.: iscore: Visual analytics for interpreting how language models automatically score summaries. In *Intelligent User Interfaces* (2024), pp. 787–802. doi:10.1145/3640543.3645142. 2

[DBW89] DAVIS F. D., BAGOZZI R. P., WARSHAW P. R.: User acceptance of computer technology: A comparison of two theoretical models. *Management science 35*, 8 (1989), 982–1003. 6

[DW24] DONG H., WANG Z.: Large language models for tabular data: Progresses and future directions. In *ACM SIGIR Conference on Research and Development in Information Retrieval* (2024), pp. 2997–3000. doi:10.1145/3626772.3661384. 2

[FSM24] FAN G., SHRAGA R., MILLER R. J.: Finding support for tabular llm outputs. *VLDB Endowment. ISSN 2150* (2024), 8097. 2

[FXT*24] FANG X., XU W., TAN F. A., ZHANG J., HU Z., QI Y., NICKLEACH S., SOCOLINSKY D., SENGAMEDU S., FALOUTSOS C.: Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey. *arXiv preprint* (2024). doi:10.48550/arXiv.2402.17944. 2

[GRB*24] GALLEGOS I. O., ROSSI R. A., BARROW J., TANJIM M. M., KIM S., DERNONCOURT F., YU T., ZHANG R., AHMED N. K.: Bias and fairness in large language models: A survey. *Computational Ling. 50*, 3 (2024), 1097–1179. doi:10.1162/coli_a_00524. 2

[JJC*24] JIANG L., JIANG K., CHU X., GULATI S., GARG P.: Hallucination detection in llm-enriched product listings. In *Workshop on e-Commerce and NLP@ LREC-COLING* (2024), pp. 29–39. 2

[KLS*24] KIM T. S., LEE Y., SHIN J., KIM Y.-H., KIM J.: Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–21. doi:10.1145/3613904.3642216. 2

[KTP*24] KAHNG M., TENNEY I., PUSHKARNA M., LIU M. X., WEXLER J., REIF E., KALLARACKAL K., CHANG M., TERRY M., DIXON L.: Llm comparator: Visual analytics for side-by-side evaluation of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–7. doi:10.1109/TVCG.2024.3456354. 2, 3, 7

[Kus24] KUSANO G.: Ga-tag: Data enrichment with an automatic tagging system utilizing large language models. In *IEEE International Conference on Data Engineering (ICDE)* (2024), IEEE, pp. 5397–5400. doi:10.1109/ICDE60146.2024.00412. 2

[LCKB23] LI Q., CUI L., KONG L., BI W.: Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *arXiv e-prints* (2023), arXiv–2310. doi:10.48550/arXiv.2310.19740. 2

[LSPL23] LANGE D., SAHAI S., PHILLIPS J. M., LEX A.: Ferret: Reviewing tabular datasets for manipulation. *Comput. Graph. Forum 42*, 3 (2023), 187–198. doi:10.1111/CGF.14822. 2

[LWX*24] LONG L., WANG R., XIAO R., ZHAO J., DING X., CHEN G., WANG H.: On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126* (2024). 2

[LXB25] LONG Y., XU L., BRINTRUP A.: Llm-tabflow: Synthetic tabular data generation with inter-column logical relationship preservation. *arXiv preprint* (2025). doi:10.48550/arXiv.2503.02161. 2

[LZF*25] LU W., ZHANG J., FAN J., FU Z., CHEN Y., DU X.: Large language model for table processing: A survey. *Frontiers of Computer Science 19*, 2 (2025), 192350. doi:10.1007/s11704-024-40763-6. 2

[MDS*25] MISHRA A., DANZY B., SONI U., ARUNKUMAR A., HUANG J., KWON B. C., BRYAN C.: Promptaid: Visual prompt exploration, perturbation, testing and iteration for large language models. *IEEE Transactions on Visualization and Computer Graphics* (2025). doi:10.1109/TVCG.2025.3535332. 3

[MIT23] MARTINO A., IANNELLI M., TRUONG C.: Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference* (2023), Springer, pp. 182–185. doi:10.1007/978-3-031-43458-7_34. 2

[MKSvO24] MARTORANA M., KUHN T., STORK L., VAN OSSENBRUGGEN J.: Zero-shot topic classification of column headers: Leveraging llms for metadata enrichment. In *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI*. IOS Press, 2024, pp. 52–66. doi:10.3233/SSW240006. 2

[MLG23] MANAKUL P., LIUSIE A., GALES M. J.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint* (2023). 2

[PAD*24] PAN Q., ASHKTORAB Z., DESMOND M., SANTILLÁN COOPER M., JOHNSON J., NAIR R., DALY E., GEYER W.: Human-centered design recommendations for LLM-as-a-judge. 16–29. doi:10.18653/v1/2024.hucllm-1.2. 2, 7

[Res25] RESNIK P.: Large language models are biased because they are large language models. *Computational Ling.* (2025), 1–21. doi:10.1162/coli_a_00558. 2

[SAT*23] SINGHAL K., AZIZI S., TU T., MAHDAVI S. S., WEI J., CHUNG H. W., SCALES N., TANWANI A., COLE-LEWIS H., PFOHL S., ET AL.: Large language models encode clinical knowledge. *Nature 620*, 7972 (2023), 172–180. doi:10.1038/s41586-023-06291-2. 2

[SNW*25] SACHDEVA M., NARAYANAN C., WIEDENKELLER M., SEDLAKOVA J., BERNARD J.: A Design Space for the Critical Validation of LLM-Generated Tabular Data. In *EuroVis Workshop on Visual Analytics* (2025), Eurographics. URL: https://www.eurova.org/eurova-2025, doi:10.2312/eurova.20251101. 3

[SZPH*24] SHANKAR S., ZAMFIRESCU-PEREIRA J., HARTMANN B., PARAMESWARAN A., ARAWJO I.: Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *ACM User Interface Software and Technology* (2024), pp. 1–14. doi:10.1145/3654777.3676450. 2, 7

[SZZ*24] SUI Y., ZHOU M., ZHOU M., HAN S., ZHANG D.: Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *ACM International Conference on Web Search and Data Mining* (2024), pp. 645–654. doi:10.48550/arXiv.2305.13062. 2

[VYZ*23] VARSHNEY N., YAO W., ZHANG H., CHEN J., YU D.: A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint* (2023). doi:10.48550/arXiv.2307.03987. 2

[WA25] WU M., AJI A. F.: Style over substance: Evaluation biases for large language models. 297–312. URL: https://aclanthology.org/2025.coling-main.21/. 2

[XLS*24] XU S., LEE C.-T., SHARMA M., YOUSUF R. B., MURALIDHAR N., RAMAKRISHNAN N.: Are llms naturally good at synthetic tabular data generation? *arXiv preprint* (2024). doi:10.48550/arXiv.2406.14541. 2

[YLL*24] YANG X., LIANG Y., LIU L., WU L., YUAN X.: Evalgpt: A visual analytic framework for enhancing trust in large language models. *VLDB Endowment. ISSN 2150* (2024), 8097. 2, 3

[ZCS*23] ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E., ET AL.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems 36* (2023), 46595–46623. 2

[ZYJ*24] ZHANG Z., YANG F., JIANG Z., CHEN Z., ZHAO Z., MA C., ZHAO L., LIU Y.: Position-aware parameter efficient fine-tuning approach for reducing positional bias in llms. *arXiv preprint* (2024). doi:10.48550/arXiv.2404.01430. 2