EuroVis Workshop on Visual Analytics (2025)
H.-J. Schulz and A. Vilanova (Editors)

A Design Space 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

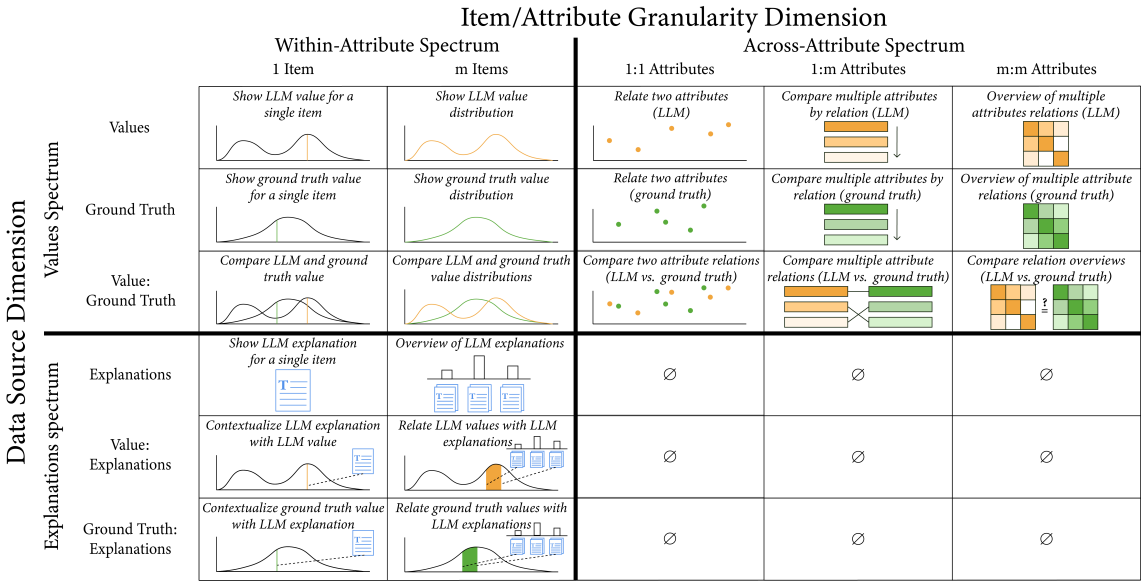

Figure 1: *Design space for the critical validation of LLM-generated tabular data. The x-dimension describes the analysis granularity, ranging from within attributes to across attributes. The y-dimension structures approaches by the data used for validation, including generated values, explanations, ground truth data, and their combinations. Visual representations per cell exemplify numerical data use, while the design space is generalizable for mixed data. Lower right: validations across attributes including explanations are so far uncharted.*

Abstract

LLM-generated tabular data is creating new opportunities for data-driven applications in academia, business, and society. To leverage benefits like missing value imputation, labeling, and enrichment with context-aware attributes, LLM-generated data needs a critical validation process. The number of pioneering approaches is increasing fast, opening a promising validation space that, so far, remains unstructured. We present a design space for the critical validation of LLM-generated tabular data with two dimensions: First, the Analysis Granularity dimension–from within-attribute (single-item and multi-item) to across-attribute perspectives (1×1, $1 \times m$, and $n \times n$). Second, the Data Source dimension–differentiating between LLM-generated values, ground truth values, explanations, and their combinations. We discuss analysis tasks for each dimension cross-cut, map 19 existing validation approaches, and discuss the characteristics of two approaches in detail, demonstrating descriptive power.

1. Introduction

Large language models (LLMs) have evolved from simple text prediction to generating context-aware output across diverse data types, with data-generation qualities comparable to state-of-the-art deep learning methods [MS24]. LLMs are increasingly used to enrich tabular datasets with contextually relevant values for items and attributes, and create new items relevant to analytical needs. Common LLM-based data enrichment methods include imputing missing values, labeling data items, and adding semantically rich attributes [FXT*24]. However, critical validation is essential to ensure the usability, plausibility, and reliability of LLM-generated tabular data. Key problem areas in LLM-generated data include subjective quality [KLS*24], lack of reproducibility [CHM*24],

DOI: 10.2312/eurova.20251101

factuality hallucinations (e.g., inconsistency with real-world facts), faithfulness hallucinations (e.g., deviating from intended prompt), epistemic opacity (black-box nature), biases, and adoption of human misconceptions [MKB24]. Further challenges include inconsistencies in attribute relationships appearing as logical contradictions, or inconsistencies in logical reasoning steps [LGL*24]. Research also identified LLM biases in choice preferences (e.g., favoring first position, favoring style of input data (e.g., preferring longer responses), and egocentric bias (e.g., favoring their own responses) [LPD23]. Without critical validation, integrating such data into analytical workflows risks propagating errors, which misinforms decision-making processes [PPD*24].

Validation methods differ in their degree of human and computational involvement, creating opportunities for visual analytics (VA) research. Fully automated methods, like statistical measures [SWS24] and self-validating techniques like Retrieval-Augmented Generation (RAG) [AWW*23], offer scalability but reduce transparency. Mixed-initiative methods, such as LLM-as-a-Judge [ZCS*23] and human-calibrated evaluations [SZPH*24, KTP*24], combine automation with human feedback. Direct human evaluation maximizes interpretability, but is resource-intensive and less scalable [CL23].

While validation is well-established for unstructured textual data [LWX*24], recent work increasingly targets tabular data, showing high heterogeneity in methods. Validation strategies include the statistical alignment with benchmark data [SWS24], LLM-as-a-Judge [ZCS*23], explanation-based methods [HMJ*23], and ground truth comparisons [WKR*24]. Ground truth may be human-labeled or from external datasets [WKR*24]. Human-labeled data is often used in mixed-initiative approaches [KLS*24, KTP*24, SZPH*24]. Approaches operate at different analysis granularities, trading accuracy for scalability, ranging from fine-grained single-item assessments [KLS*24], multi-item or entire attributes for aggregated metrics [CFB*23, CE23], to coarse assessments across attributes [TWB*20]. Some provide textual explanations for single items [HMJ*23], or multiple items to show systematic biases [KTP*24]. Approaches also differ across workflow phases: LLM-based data generation (prompt engineering) [MS24], core validation (output assessments) [KTP*24], and evaluating LLM outputs in domain-specific downstream applications [CE23]. Contextual factors such as the LLM model, application domain, and user group introduce further variations.

To understand this emerging and multifaceted validation space, several challenges must be addressed. Existing approaches operate in an unstructured space without an overarching perspective, leaving key characteristics implicit. No comprehensive meta-analysis compares validation methods, leaving differences and commonalities unexplored. The lack of structured analysis prevents easy identification of gaps and the guided selection of appropriate strategies. A unified framework is needed to organize methods, identify gaps, and guide future research.

We introduce a design space that systematically structures validation approaches for LLM-generated tabular mixed data. Derived from a systematic literature review, it is defined along two key dimensions: The *Item and Attribute Granularity Dimension* captures the validation granularity, within and across attributes. The *Data Source Dimension* covers core validation data sources: LLM-generated values, ground truth, LLM-explanations, and their combinations. At the intersection of these dimensions, we identify essential validation tasks that guide validation methods for LLM-generated tabular data. Our contributions are:

1. **A design space** for the critical validation of LLM-generated structured tabular data. It structures and describes methods by their commonalities and differences, and guides designers and developers toward novel validation strategies.
2. **A mapping of existing validation approaches** onto this design space, demonstrating its descriptive power. It offers a systematic characterization of approaches and the identification of gaps.
3. **A discussion on key insights** based on critical reflection, including identified patterns, uncharted subspaces, and additional structural characteristics to inform future research.

Our design space maps the validation landscape to support research into robust, scalable, and interpretable validation methods. It aims to enhance the plausibility, reliability, and utility of LLM-generated tabular data in data analysis and decision-making workflows, accounting for different types of human and algorithmic involvement. It supports researchers, system designers, and tool developers to build VA interfaces for critical data validation.

2. Design Space

2.1. Methodology and Overview

We conducted a systematic literature search, analysed the identified related work, and discussed it among the authors. Three authors identified and categorized validation approaches for LLM-generated tabular data via Google Scholar search with the terms 'LLM tabular data validation', 'LLM human-in-the-loop validation', 'Explanation LLM-generated tabular data', 'Explainable LLM interactive visualizations' and 'LLM scoring', followed by a forward and backward search, with a focus on the papers EvalLM [KLS*24], LLM Comparator [KTP*24], and two surveys by Brasoveanu et al. [BSNA24] and Long et al. [LWX*24]. We excluded works published before 2017 (prior to transformer-based LLM architectures), those not written in English, mentioning validation only superficially, or addressing only textual data (like EvalGen [SZPH*24]). Borderline cases were discussed to reach consensus. In total, we identified 19 relevant works as the basis for design space ideation; we refer to the supplemental material for their characterization. The main visualizations used by existing approaches are bar charts (6), followed by scatter plots (3) and heat maps (3).

We aimed to support the systematic study of existing approaches and the identify gaps and opportunities, in both numerical and categorical tabular data. We chose a 2D structure for spatialization, using expressive, independent dimensions that reveal validation tasks supported by VA. Both dimensions are discrete to allow systematic cross-cuts and clear mapping of approaches in a table.

Our design space structures validation methods for LLM-generated tabular data by two key dimensions: *Data Source*, and *Item and Attribute Granularity*, reflecting *what* and *why* experts analyze [Mun14]. The *Data Source* dimension spans data types serving as inputs essential for validation reasoning, consisting of (combinations of) LLM-generated data, explanations, and (human) ground truth. The *Item and Attribute Granularity* dimension distinguishes approaches by analysis granularity, from fine-grained item-based validations and coarse attribute-based assessments, offering

different trade-offs between accuracy and scalability. Figure 1 illustrates the design space with important analysis tasks for cross-cutting dimensions. Each task reflects the need for specific visual reasoning, depending on Data Source and analysis granularity. For most tasks, existing approaches already indicate the need for interactive human judgment enabled through VA support. We exemplify the design space with visual identifiers for numerical data due to its intuitiveness, while the space is mixed-data capable. Table 1 displays the mapping of existing approaches onto the design space.

2.2. Design Space Dimension: Data Source

The Data Source dimension refers to the input data used for validation tasks. We use visual encodings for the three types of Data Sources: LLM-generated values (V), ground truth values (V_g) and LLM-generated explanations (E), and their combinations.

- **Data Value** V : is the most granular information unit requested from the LLM. Data values belong to an item and have an attribute type defined by its associated attribute schema.
- **Ground truth Value** V_g : represents a data value for an item that is known or assumed to be true. This can be provided by pre-existing data or elicited from human externalized knowledge.
- **Explanations** E : are textual justifications provided by an LLM about its generated data values. This explanation spectrum enables to contextualize data values or assess their semantic meaning with VA relation-discovery, per item or more condensed.

A key observation from many approaches highlight a key distinction: value-based validation (V , V_g) relies on measurable data, statistical reasoning, and alignment checks, while explanation-based (E) validation requires subjective judgments of plausibility, consistency, and bias. This distinction directly impacts the types of visualization tasks and the design of VA solutions.

2.3. Design Space Dimension: Item and Attribute Granularity

This dimension characterizes the validation granularity, in a spectrum from fine-grained *Item-based* validations within an attribute, to more coarse *Attribute-based* validation, across attributes.

- 1 Item: the most atomic granularity of validation is for a single item (and attribute), preferring accuracy over scalability.
- m Items: multiple items are taken into account for a single attribute. This includes the assessment of summaries and aggregates, and the comparative analysis between item values.
- 1×1 Attribute: raising the granularity to an across-attributes perspective, one class of validation approaches is on relation discovery tasks between an attribute pair.
- $1 \times m$ Attributes: for some validation tasks, users take multiple attributes into account, e.g., to assess the strengths of relations of attributes to a focus attribute.
- $m \times m$ Attributes: the coarsest granularity is the brute-force assessment of all relations between all attributes at scale.

2.4. Validation Tasks at the Crosscut of Dimensions

The systematic combination of the *Data Source* and the *Item and Attribute Granularity* dimension opens the perspective of discrete cells of the design space. For each cell, we present the dominating validation task, according to existing approaches (see Table 1). Blank cells at the lower right indicate non-allocated regions (e.g., aggregating explanations across (multiple) attributes simultaneously.) Figure 1 provides detailed task descriptions and visual exemplifications, to align validation tasks with abstract analysis/visualization tasks, e.g., "Show LLM value for a single item" supports *Lookup* whereas "Compare multiple attributes by relation" supports *Compare*. For each cross-cut, we include expected analysis outcomes and example analysis questions.

2.4.1. 1 Item Focus

Expected analysis outcome: identify and clarify discrepancies between human-expected and LLM-generated values for a single item. Example analysis question: “Is the LLM-generated happiness index for the city of Luxembourg plausible?”

V × 1 Item: **Show LLM value:** The isolated inspection of a single LLM-generated value enables users to assess its plausibility and relevance. Single-value assessment is often supported in the context of value distributions, e.g., to enable comparative analysis.
V_g × 1 Item: **Show ground truth value:** Allows users to create, inspect, or refine a ground truth value, which can be later used for comparative assessment with LLM-generated values.
V × V_g × 1 Item: **Compare LLM value and ground truth value:** Users can compare the LLM-generated value with the ground truth value, enabling the assessment of value alignment/deviation.
E × 1 Item: **Show LLM explanation:** Users can engage with explanations for a single LLM-generated value. This can potentially show flaws in the LLM’s reasoning or justification.
E × V × 1 Item: **Contextualize LLM explanation with LLM value:** Displaying the LLM-generated value and its explanation helps assess its justification and reveal inconsistencies. This pairing offers insight into the LLM’s logic, indicating whether it considers relevant factors. It also highlights weaknesses when explanations are vague, uncertain, or misaligned with generated values.
V_g × E × 1 Item: **Contextualize ground truth value with LLM explanation:** This task enables the assessment of the LLM-generated explanation by using the ground truth value as a trusted frame of reference, e.g., to shed light on the reasoning of the LLM.

2.4.2. m Items Focus

Expected analysis outcome: identify and clarify discrepancies between human-expected and LLM-generated value distributions for multiple items. Example analysis question: “Are LLM-generated unemployment rates biased toward a specific range?”

V × m Items: **Show LLM value distribution:** Displaying the distribution of LLM-generated values, possibly in comparison to the entire item value distribution, can reveal subset characteristics, like potential biases or a lack of diversity.
V_g × m Items: **Show ground truth value distribution:** Show ground truth value distribution to reveal subset patterns and potential shifts from the overall distribution, helping to assess biases.
E × m Items: **Overview of LLM explanations:** Understanding explanation patterns occurring across multiple items helps to cross-validate explanations, understand rationales that drive LLMs when generating values, and helps to assess the LLM’s consistency.

		Item/Attribute Granularity Dimension				
		1 Item	m Items	1×1 Attribute	$1 \times m$ Attributes	$m \times m$ Attributes
Data Source Dimension	Values V	Show LLM value for a single item [KTP*24, KLS*24, CZC*24, TWB*20, CE23]	Show LLM value distribution [KTP*24, PAD*24, KLS*24, CFB*23, CZC*24, TWB*20, CE23]	Relate two attributes (LLM) [KTP*24, KLS*24, TWB*20]	Compare multiple attributes by relation (LLM)	Overview of multiple attribute relations (LLM) [TWB*20]
	Ground Truth V_g	Show ground truth value for a single item [PAD*24, TWB*20]	Show ground truth value distribution [CHM*24, TWB*20]	Relate two attributes (ground truth) [CHM*24, TWB*20]	Compare multiple attributes by relation (ground truth)	Overview of multiple attribute relations [TWB*20]
	Value : Ground Truth V : V_g	Compare LLM and ground truth value [PAD*24, SWS*22, MDS*25, CFB*23, ZZH*23, HBL*23, FSM24, SLA*24, GCM*25, TWB*20, CHM*24]	Compare LLM and ground truth value distributions [CHM*24, PAD*24, SWS*22, MDS*25, CFB*23, ZQH*23, BSL*22, TWB*20]	Compare two attribute relations (LLM vs. ground truth) [FSM24, BSL*22, TWB*20]	Compare multiple attribute relations (LLM vs. ground truth) [FSM24, BSL*22]	Compare relation overviews (LLM vs. ground truth)
	Explanations E	Show LLM explanation for a single item	Overview of LLM explanations			
	Value : Explanations V : E	Contextualize LLM explanation with LLM value [KTP*24, CHM*24, PAD*24, KLS*24, SWS*22, MDS*25, CZC*24, BS24, SJZ25, TWB*20]	Relate LLM values with LLM explanations [KTP*24, BS24]			
	Ground Tr. : Explanations V_g : E	Contextualize ground truth value with LLM explanation [BS24, TWB*20]	Relate ground truth values with LLM explanations [BS24]			

Table 1: *Design space tasks across Item and Attribute Granularity and Data Source Dimension. While the design space accommodates values across all item-attribute granularities, explanations are primarily effective when applied to an item granularity.*

V × E × m Items: **Relate LLM values with LLM explanations:** Relating multiple LLM-generated values with explanations assesses justification consistency, reveals discrepancies across items, and identifies potential pitfalls or biases in the reasoning.

V × V_g × m Items: **Compare LLM and ground truth value distributions:** Assessing the alignment of these two value distributions provides statistical insights, revealing discrepancies, potential biases, or failures in LLM's understanding of key aspects.

V_g × E × m Items: **Relate ground truth values with LLM explanations:** This relation-discovery task allows users to assess the general alignment between ground truth and LLM explanations for the identification of potential pitfalls or biases in the reasoning.

2.4.3. 1×1 Attribute Focus

Expected analysis outcome: clarity and plausibility of a LLM-generated attribute, in relation to an existing attribute. Example analysis question: "Are happiness scores generated by the LLM positively correlated with my income attribute?"

V × 1×1 Attribute: **Relate two attributes (LLM-generated):** Relation-discovery between two attributes, e.g., a LLM-generated and an existing attribute to show interesting dependencies.

V_g × 1×1 Attribute: **Relate two attributes (ground truth):** Relation-discovery between two ground truth attributes, e.g., to identify dependencies or redundancies for feature selection.

V_g × V × 1×1 Attribute: **Compare two attribute relations (ground truth with LLM-generated):** Relate ground truth and LLM-generated attributes to assess alignment and agreement. This helps to gain trust in whether LLMs can reproduce what is already known.

2.4.4. $1 \times m$ Attributes Focus

Expected analysis outcome: clarity about the relation between the LLM-generated attribute and other (existing) attributes. Example analysis question: "Which attributes are most positively/negatively correlated with the LLM-generated happiness index?"

V × $1 \times m$ Attributes: **Compare multiple LLM-generated attributes by relation:** Relation-discovery between LLM-generated attributes can, e.g., put one LLM-generated attribute in focus, allowing for ranking other attributes by their relation strength.

V_g × $1 \times m$ Attributes: **Compare multiple ground truth attributes by relation:** Relation-discovery between ground truth attributes can, e.g., put one attribute in focus, allowing for ranking other attributes by their relation strength.

V × V_g × $1 \times m$ Attributes: **Compare multiple attribute relations (LLM vs ground truth):** Users compare the relation-discovery outcomes of LLM-generated attributes with those for ground truth attributes to identify discrepancies and biases.

2.4.5. $m \times m$ Attributes Focus

Expected analysis outcome: clarity on the relations between all LLM-generated attributes and all existing attributes. Example analysis question: "Do relations in LLM-generated data align with the ground truth and do they enrich existing data for downstream use?"

V × $m \times m$ Attributes: **Overview of multiple LLM-generated attributes and their relations:** This relation-discovery task applies a brute-force approach to relate all attributes to each other, typically leading to a matrix display of relations.

V_g × $m \times m$ Attributes: **Overview of multiple ground truth attributes and their relations:** This task provides users with the full overview of all relations between ground truth attributes.

V × V_g × $m \times m$ Attributes: **Compare relation overviews (LLM with ground truth):** This task compares all relations of LLM-generated attribute pairs with all relations of ground truth attribute pairs to gain an understanding of their agreement.

3. Validation Examples

We demonstrate the descriptive power of the design space by mapping tasks from two existing approaches and show how their workflows map into the design space. Arrows between tasks illustrate the workflows described in these approaches (see Figure 3).

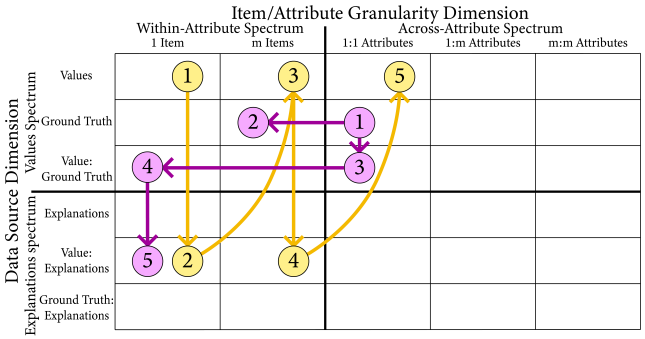

Figure 2: *Mapping the supported tasks and user workflows of two existing validation approaches to our design space. Purple: iScore [CHM*24], Yellow: LLM Comparator [KTP*24]*

Figure 3: *Screenshot with marked validation steps of iScore reproduced from Coscia et al. [CHM*24] with permission (left) and LLM Comparator [KTP*24] with permission (right)*

iScore [CHM*24] is an LLM-based tool for scoring summaries on attributes, such as objectivity and coherence. A scatter plot displays ground truth data for two scoring attributes, providing an overview of expected trends and correlations (1). A bar chart alongside the scatter plot further clarifies value distributions (2). Users provide summaries which the LLM scores and visualizes, on the same scatter plot in a different color. Overlaying these plots allows direct comparison of trends between LLM scores and ground truth values (3). To explore individual relationships, users can select a ground truth value, load the corresponding summary, and have it scored by the LLM for direct attribute comparison (4). For deeper insight into how a score was derived, users can visualize the LLM's attention and receive a non-textual explanation (5).

LLM Comparator [KTP*24] compares the output quality of different LLMs and identifies trends across models. Despite its primary focus on LLM-generated text, LLM Comparator uses an LLM-as-a-judge approach, based on pairwise comparisons, which generates a categorical gradation. Users select an item, view its prompt and two generated outputs, and examine individual comparison results (1). To understand why one output is preferred, textual explanations are collected during the LLM-as-a-judge process (2). For a broader perspective, users can analyze the distribution of aggregated scores to detect patterns in LLM preferences (3). Performance contextualization is possible through bar graphs linking explanations–such as "is more engaging"–to output preferences (4). In the final step, users explore how LLM preferences relate to collected output metrics, such as word count, moving beyond within-attribute analysis to uncover relations and trends (5). This structured comparison allows users to assess LLM performance, understand decision rationales, and detect potential biases or inconsistencies in model outputs.

4. Limitations and Discussion

Data Source Mapping: When validation approaches use two Data Source Dimension elements (values, ground truth, explanations), we currently map them only to the combined cell rather than separately, leading to empty explanation-only cells never occurring in isolation. This reduces redundancy, but affects the descriptive nature of our mapping strategy, which we plan to investigate further.

Item/Attribute Granularity Mapping: Some validation approaches support both fine-grained and coarse analysis in one view. For example, scatter plots enable per-item, multi-item, and 1:1 attribute validation [CHM*24]. Our mapping prioritizes the coarser cell to reduce redundancy, but this may impact the descriptiveness of the design space, requiring further investigation.

Human Knowledge Externalization: A key validation strategy compares LLM-generated data with human-provided ground truth, leveraging domain knowledge for representative subsets. However, this approach faces scalability challenges, as human knowledge externalization is labor-intensive. Addressing these limitations will be crucial for making human-in-the-loop validation feasible at scale.

Generation-Validation-Application as a Process: We identify a structured process-oriented perspective for validating LLM-generated tabular data, with three key phases: data generation (including prompt engineering), validation (as studied), and downstream applications. Expanding the design space could distinguish validation for upstream generation and downstream application, forming a unified framework adaptable to specific domains.

Numeric vs. Categorical Values: While our design space is attribute-type agnostic, our illustrations focus on numerical representations of LLM-generated data values, and their relations [SBKB23]. Future work may explore illustrative support for categorical relation discovery [BSW*14] and develop idioms to an equal status, to ease the ideation of solutions for categorical data.

Role of Visual Analytics: Most pioneering validation approaches included VA building blocks to cope with the complexity of validation challenges. This motivates a deeper study of forms of human involvement in combination with possibilities for computational support. Future work includes the extension of the design space with a systematic focus on human-model collaboration patterns, studied in relation to the two existing dimensions.

5. Conclusion

We introduced a structured design space for the critical validation of LLM-generated tabular data. The design space systematically characterizes validation approaches, enabling their comparison and guiding developers in identifying validation strategies. We mapped existing approaches onto the design space to identify their characteristics and discussed two approaches in detail to validate the descriptive power. Our work makes a case for VA as a critical layer in the LLM data generation. While automated solutions are valuable, interactive visual validation tools are indispensable for ensuring trust and reliability in validating synthetic tabular data. This design space provides a potential foundation for building such tools. Future work may explore extensions of the design space with further dimensions that could enrich validation approaches. The design space can also be expanded by formulating a unified process framework, distinguishing validation methods for data generation and application.

References

[AWW*23] ASAI A., WU Z., WANG Y., SIL A., HAJISHIRZI H.: Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Conference on Learning Representations* (2023). 2

[BS24] BECKER E., SOATTO S.: Cycles of thought: Measuring llm confidence through stable explanations. *arXiv:2406.03441* (2024). 4

[BSL*22] BORISOV V., SESSLER K., LEEMANN T., PAWELCZYK M., KASNECI G.: Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280* (2022). 4

[BSNA24] BRASOVEANU A. M., SCHARL A., NIXON L. J., ANDONIE R.: Visualizing large language models: A brief survey. In *Conference Information Visualisation (IV)* (2024), IEEE, pp. 236–245. 2

[BSW*14] BERNARD J., STEIGER M., WIDMER S., LÜCKE-TIEKE H., MAY T., KOHLHAMMER J.: Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum (CGF) 33*, 3 (2014), 291–300. 5

[CE23] COSCIA A., ENDERT A.: Knowledgevis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Transactions on Visualization and Computer Graphics 30*, 9 (2023), 6520–6532. 2, 4

[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. 2, 4

[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 *International Conference on Intelligent User Interfaces* (2024), pp. 787–802. 1, 4, 5

[CL23] CHIANG C.-H., LEE H.-Y.: Can large language models be an alternative to human evaluations? In *Association for Computational Linguistics* (2023), Rogers A., Boyd-Graber J., Okazaki N., (Eds.), Association for Computational Linguistics, pp. 15607–15631. 2

[CZC*24] CHENG F., ZOUHAR V., CHAN R. S. M., FÜRST D., STROBELT H., EL-ASSADY M.: Interactive analysis of llms using meaningful counterfactuals. *arXiv preprint arXiv:2405.00708* (2024). 4

[FSM24] FAN G., SHRAGA R., MILLER R. J.: Finding support for tabular llm outputs. *VLDB Endowment. ISSN 2150* (2024), 8097. 4

[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 arXiv:2402.17944* (2024). 1

[GCM*25] GENG S., COOPER H., MOSKAL M., JENKINS S., BERMAN J., RANCHIN N., WEST R., HORVITZ E., NORI H.: Generating structured outputs from language models: Benchmark and studies. *arXiv preprint arXiv:2501.10868* (2025). 4

[HBL*23] HEGSELMANN S., BUENDIA A., LANG H., AGRAWAL M., JIANG X., SONTAG D.: Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics* (2023), PMLR, pp. 5549–5581. 4

[HMJ*23] HUANG S., MAMIDANNA S., JANGAM S., ZHOU Y., GILPIN L. H.: Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv:2310.11207* (2023). 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* (2024), pp. 1–21. 1, 2, 4

[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 *CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–7. 2, 4, 5

[LGL*24] LIU Y., GUO Z., LIANG T., SHAREGHI E., VULIĆ I., COLLIER N.: Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205* (2024). 2

[LPD23] LI R., PATEL T., DU X.: Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762* (2023). 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

[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). 4

[MKB24] MALIN B., KALGANOVA T., BOULGOURIS N.: A review of faithfulness metrics for hallucination assessment in large language models. *arXiv preprint arXiv:2501.00269* (2024). 2

[MS24] MILETIC M., SARIYAR M.: Assessing the potentials of llms and gans as state-of-the-art tabular synthetic data generation methods. In *Privacy in Statistical Databases* (2024), Springer, pp. 374–389. 1, 2

[Mun14] MUNZNER T.: *Visualization analysis and design*. CRC press, 2014. 2

[PAD*24] PAN Q., ASHKTORAB Z., DESMOND M., COOPER M. S., JOHNSON J., NAIR R., DALY E., GEYER W.: Human-centered design recommendations for llm-as-a-judge. *arXiv:2407.03479* (2024). 4

[PPD*24] PARK Y.-J., PILLAI A., DENG J., GUO E., GUPTA M., PAGET M., NAUGLER C.: Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making 24*, 1 (2024), 72. 2

[SBKB23] SACHDEVA M., BURMEISTER J., KOHLHAMMER J., BERNARD J.: LFPeers: Temporal similarity search and result exploration. *Computers & Graphics* (2023). 5

[SJZ25] SCHROEDER N. L., JALDI C. D., ZHANG S.: Large language models with human-in-the-loop validation for systematic review data extraction. *arXiv preprint arXiv:2501.11840* (2025). 4

[SLA*24] SHANKAR S., LI H., ASAWA P., HULSEBOS M., LIN Y., ZAMFIRESCU-PEREIRA J., CHASE H., FU-HINTHORN W., PARAMESWARAN A. G., WU E.: Spade: Synthesizing data quality assertions for large language model pipelines. *arXiv preprint arXiv:2401.03038* (2024). 4

[SWS*22] STROBELT H., WEBSON A., SANH V., HOOVER B., BEYER J., PFISTER H., RUSH A. M.: Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE trans. on visualization and computer graphics 29*, 1 (2022), 1146–1156. 4

[SWS24] SUN K., WANG R., SØGAARD A.: Comprehensive reassessment of large-scale evaluation outcomes in llms: A multifaceted statistical approach. *arXiv preprint arXiv:2403.15250* (2024). 2

[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 *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (2024), pp. 1–14. 2

[TWB*20] TENNEY I., WEXLER J., BASTINGS J., BOLUKBASI T., COENEN A., GEHRMANN S., JIANG E., PUSHKARNA M., RADEBAUGH C., REIF E., ET AL.: The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122* (2020). 2, 4

[WKR*24] WANG X., KIM H., RAHMAN S., MITRA K., MIAO Z.: Human-llm collaborative annotation through effective verification of llm labels. In *CHI Conference* (2024), pp. 1–21. 2

[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

[ZQH*23] ZHUANG H., QIN Z., HUI K., WU J., YAN L., WANG X., BENDERSKY M.: Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122* (2023). 4

[ZZH*23] ZHU Y., ZHANG P., HAQ E.-U., HUI P., TYSON G.: Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145* (2023). 4