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

Supplemental materials for `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

1. Data Generation Prompt example

Listing 1: *A Python function to create a prompt for the LLM for generating tabular data. Here,* `company` *and* `isin` *refer to an item (stock), aspect refers the name of the attribute,* `description` *is the definition of the attribute, and* `context` *refers to description of the item, in this case, company. We expect the output in a valid JSON format which we later parse into our tabular database.*

```
def create_prompt(context, company, ISIN, aspect,
     description):
    prompt = f"""
    For company "{company}" with ISIN {ISIN},
        evaluate '{aspect}', defined as: '{
        description}'.
    Return: {{
        "name": '{company}',
        "isin": '{ISIN}',
        "aspect_name": '{aspect}',
        "score": '{aspect}' score on a
            quantitative scale from 1 (worst) to
            10 (best) (this is an integer value,
            also don't have the value in quotes),
        "score_explanation": short explanation of
             the assigned score (up to 15 words),
        "score_confidence": the confidence of the
             score estimation in % (0 for lowest,
             100 for highest),
    }}

    Respond in JSON without any markdown notation
        . Make sure to provide a valid JSON
        object.
    Base your evaluation on the following context
        :
    <context>
    {context}
    </context>
    """
    return prompt.strip()
```

2. `Val-LLM` Interface

In this section we describe the `Val-LLM` interface in detail. Icons for stocks are blurred in the referenced figures.

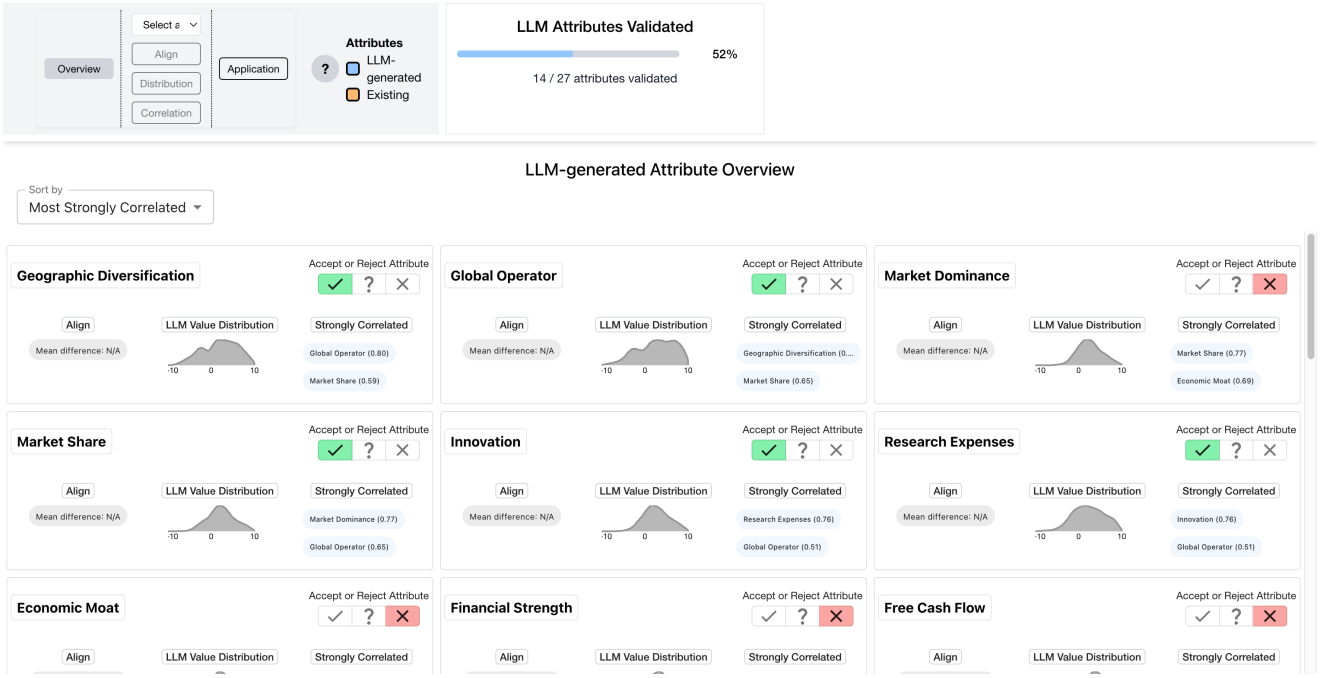

Figure 1: *Attribute Overview interface of* `Val-LLM`*. On the top left, the navigation compass provides a global navigation control through the entire interface. The top center shows a progress bar with the number of LLM-generated attributes validated. An attribute is validate once its validation status changes to accept or reject. Users can sort the LLM-generated attributes by criteria (most strongly correlated, name, standard deviation of LLM-generated scores, alignment difference. The LLM-generated attributes are shown in a card-like format, encapsulating the attribute name, validation status and three summary components for the Align, Distribution and Correlation Views - to help the user overview the three granularities (when applicable).*

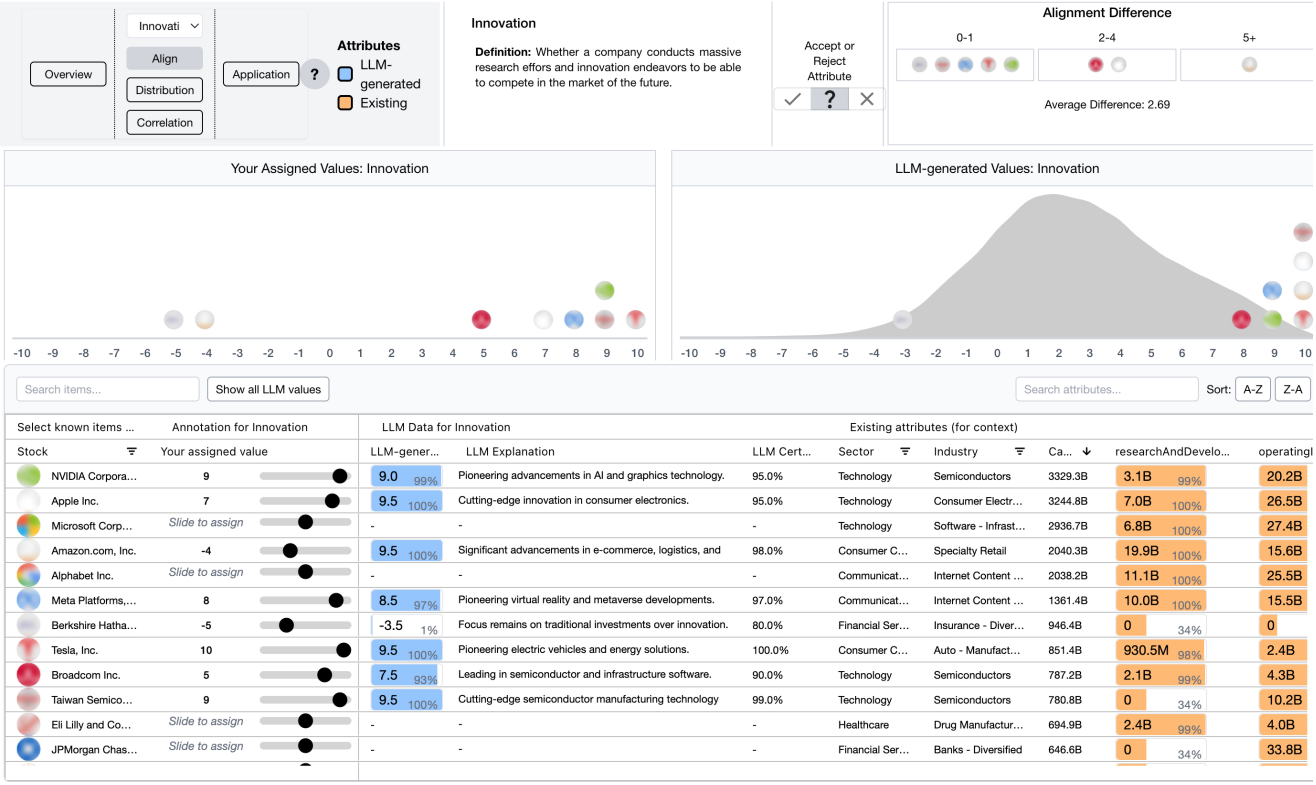

Figure 2: *Align View interface of* `Val-LLM`*. On the top left, the navigation compass provides a global navigation control through the entire interface. In the top center, the attribute being validated and its definition (defined by the user in an upstream process) are stated. On the top right, the validation status is displayed to allow the user to accept or reject the LLM-generated data for the attribute. In the center of the interface, the left side shows the assigned values externalized by the human and the right side shows the overall LLM-generated value distribution (in gray) and the LLM-generated values for items that were annotated by the human. In the top right, the Alignment Difference provides a summary of the difference in values between the human externalized values for items compared with the LLM-generated values for that respective item. In the bottom half, the table lists all the items. Users can move the slider to assign a value for an item. Once a value is assigned, the LLM-generated values, LLM-generated explanation and LLM-generated certainty is shown to the user in the table. To show all the values, without moving the slider, the user can select the "Show all LLM values" button. To help the user contextualize, they can browse the table columns for existing attributes. For the LLM-generated values and existing values, we also encode the percentile of the value, both with text and bar length. Users can search for known-items using the search bar. They can also search for existing attributes and sort them with the controls on the right side.*

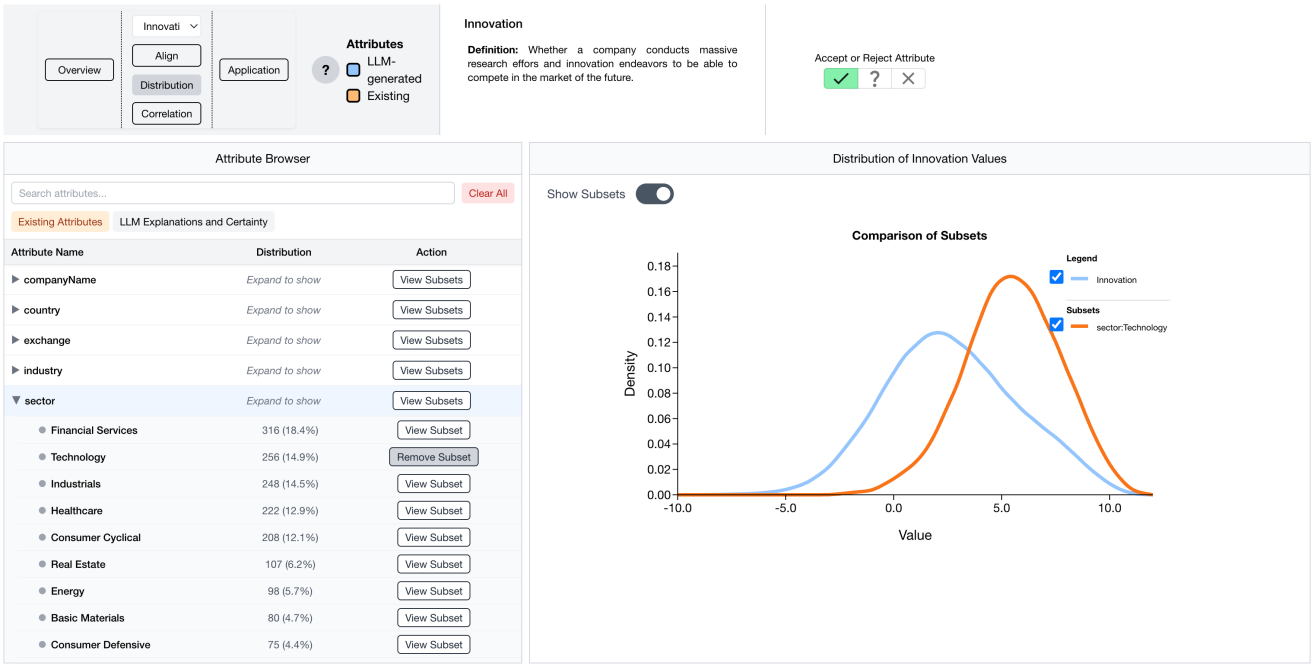

Figure 3: *Distribution View interface of* `Val-LLM`*. On the top left, the navigation compass provides a global navigation control through the entire interface. In the top center, the attribute being validated and its definition (defined by the user in an upstream process) are stated. On the top right, the validation status is displayed to allow the user to accept or reject the LLM-generated data for the attribute. In the main interface, users can browse existing attributes and LLM-generated explanations and certainty using the Attribute Browser. Users can also search for attributes using the search bar. For hierarchical categorical attributes (e.g. sector), users can expand them. Users see the attribute name, distribution (categorical distributions show the number of items and how much percent that from the overall) and numerical values show a small sparklines plot (not illustrated in this example). Users can click on the View Subset to visualize the subset together with the LLM-generated attribute. Users can toggle which lines to visualize on the line chart. This plot shows the density of the LLM-generated values. Users can overlay multiple subsets.*

Figure 4: *Correlation View interface of* `Val-LLM`*. On the top left, the navigation compass provides a global navigation control through the entire interface. In the top center, the attribute being validated and its definition (defined by the user in an upstream process) are stated. On the top right, the validation status is displayed to allow the user to accept or reject the LLM-generated data for the attribute. In the center of the interface, the users have the a table on the left to browse the correlation of attributes with the target LLM-generated attribute. Users can browse the attributes by toggling existing attributes or LLM-generated attributes. Users can also search for an attribute. In the table, the attribute name (feature name), correlation strength using pearson correlation is displayed. Users can select an attribute using the View Correlation button to show it on the Correlation View on the right side in the scatterplot. In the scatterplot, users can hover over items to see their names. Users can also lasso select multiple items. By doing so, the selected items will appear in the top rows of the table. The table consists of the name of the item (i.e. stock), and LLM-generated values, LLM-generated explanations and LLM-generated certainty. Users can also contextualize these items by oviewview with the existing attributes in table.*

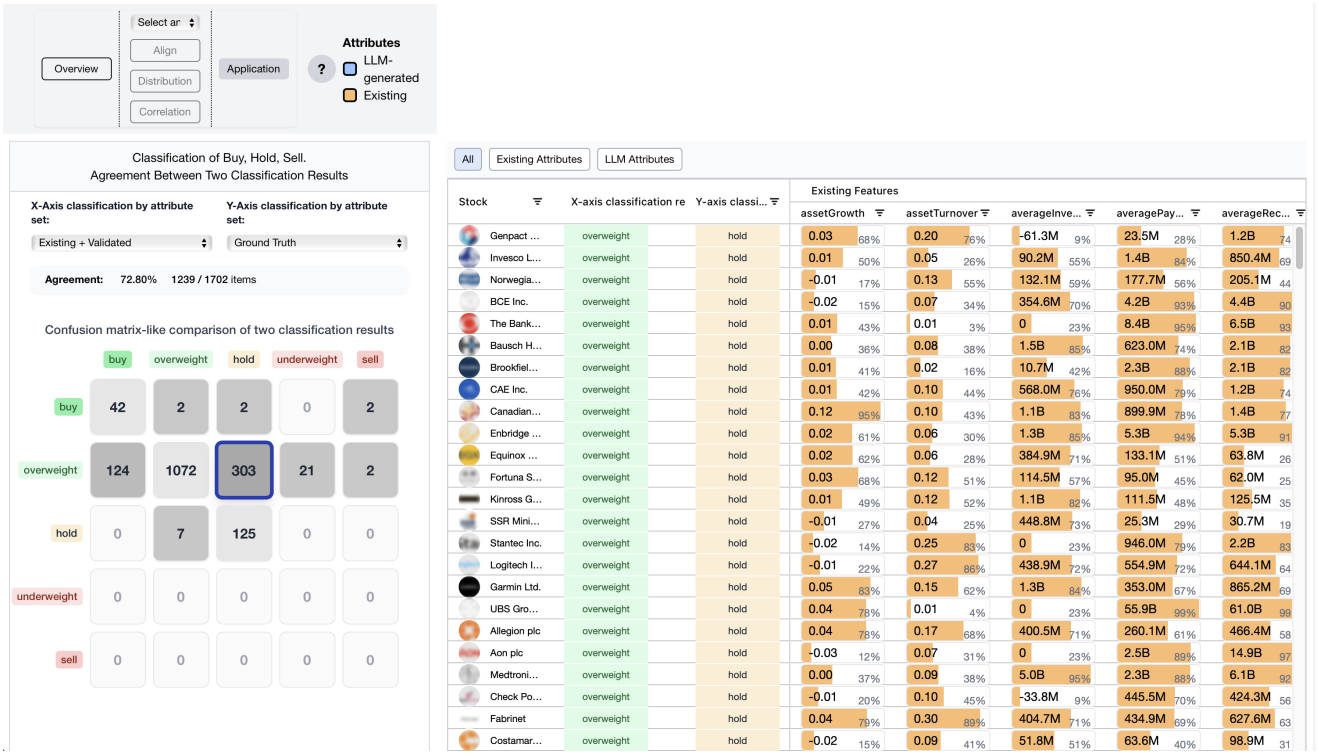

Figure 5: *Application View interface of* `Val-LLM`*. On the top left, the navigation compass provides a global navigation control through the entire interface. On the left, users can explore the classification results of attribute sets. The two dropdown menus enable users to compare attribute sets and their classification performances. The agreement (i.e, match) between the classification results are shown in the confusion matrix idiom. Users can select the attribute sets:(Validated, Existing, Existing + Validated, LLM-generated, Existing + LLM-generated). In the confusion matrix, cells are color-coded based on their values to help highlight areas with high counts. However, we left the diagonal cells uncolored, as they represent correct matches. Our focus is to highlight the off-diagonal cells, which refer to misclassifications. To inspect the mislassiciations on an item-level, users can click on the cell to show the items in the table on the right. Users can view the item (i.e, stock), the resulting classifications and the existing and LLM-generated attributes for these items. Considering the list of exisiting and LLM-generated attributes is long, users can toggle which attribute sets to view in the table (when applicable).*

3. User Study

3.1. Task-based Assessment

3.2. System Usability Scale (SUS)

3.3. Technology Adoption Model (TAM)

3.4. Usefulness assessment

Task	P2	P3	P4	P5	P6	P7
How many LLM-generated attributes are there?	1	1	1	1	1	1
Identify which LLM-generated attribute has the strongest correlation?	1	1	1	1	0	1
Find and assign a known value for the NVIDIA corporation (based on your knowledge) for the Innovation attribute	1	1	1	1	1	1
Compare this known value with the LLM-generated value	1	1	1	1	1	1
How does the "Financial Services" sector subset compare to the LLM-generated values for Innovation?	1	1	1	1	1	1
Examine whether observed differences in distribution remain consistent or significantly change with other sectors?	1	1	1	0	0	1
Identify the existing attribute most correlated with the Innovation attribute	0	1	0	1	0	1
Despite high overall correlation, which items have a high LLM Innovation value but lower research and development expenses?	1	0	1	1	0	1
Identify where LLM-generated classifications differ from ground truth the most.	0	1	0	1	1	1
Determine the percentage of agreement between predicted classifications and ground truth when LLM-generated attributes are integrated.	0	1	0	1	1	1

Table 1: *Task success by participants.*

System Usability Scale Question	P2	P3	P4	P5	P6	P7
I think that I would like to use this system frequently.	3	3	4	3	4	3
I found the system unnecessarily complex.	2	2	2	2	1	3
I thought the system was easy to use.	3	4	4	4	5	3
I think that I would need the support of a technical person to be able to use this system.	2	1	1	2	1	4
I found the various functions in this system were well integrated.	4	4	5	4	5	4
I thought there was too much inconsistency in this system.	2	1	1	1	1	1
I would imagine that most people would learn to use this system very quickly.	5	4	4	5	5	2
I found the system very cumbersome to use.	1	2	2	1	2	3
I felt very confident using the system.	3	4	4	4	5	4
I needed to learn a lot of things before I could get going with this system.	3	3	2	2	2	3

Table 2: *System Usability Scale (SUS) responses by participants*

TAM Question	P2	P3	P4	P5	P6	P7
The use of this visual analytics tool could help me with validating and understanding LLM-generated tabular attributes.	4	5	5	5	5	4
The use of this tool could improve how I validate and compare LLM-generated tabular data with my own domain knowledge.	5	5	5	5	5	5
The use of this tool could improve my performance in reviewing, assessing, and integrating LLM-generated insights into my workflow.	4	5	4	5	5	4
The use of this tool could facilitate the systematic validation of LLM-generated tabular data in an interpretable way.	5	4	5	4	5	4

Table 3: *Technology Acceptance Model (TAM) on perceived usefulness –responses by participants.*

Usefulness Question	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
How effective was the tool in inspecting individual model-generated values?	4	4	4	5	5	4
Please elaborate on the previous question. For example, what made the inspection easier or more difficult?		I was able to look at the values for specific companies in the table of the 'Align' view. Combined with the sorting of the table, this allowed me to inspect them.	It was a straightforward process of choosing the attribute and then adding the own judgment of it and finally view the correct data, so that I am not influenced by the prediction of the LLM.	The ability to immediately compare my own suggested value against the model-generated one made the inspection easier.	It was easier to mark your knowledge and compare with LLM-generated values and also able to see where you and LLM stand in the space and the difference score.	The tool seems very centered around comparison of LLM data vs actual/human generated data – which is close to what I would want for this task.
How effective was the tool in letting you spot overall trends across entire attributes?	4	5	4	4	5	4
Please elaborate on the previous question. For example, what helped or prevented your ability to see these patterns?		The distributions were visible on various views (Overview, Align, Distribution), so I was always aware of the overall distribution.	It was intuitively clear what to do to see the desired trends, but I lacked the domain knowledge to correctly interpret them.	The plotted distributions of the selected variables were helpful.	The trend graph helps you understand the overall distribution and where the LLM-generated distribution and actual data stand.	I liked the distribution section which gave me trends for different slices of the data. The correlation section seemed useful in theory, but the scatterplot was offset by outliers which made it difficult to use.
When comparing your externalized value with the LLM's generated value, how did your confidence in the results change?	4	4	3	4	5	4
Please elaborate on the previous question. For example: Why do you think your confidence changed (or stayed the same)	Because I could compare the values and they were comparable to how I perceived it.	It was very nice to see the instant feedback after labeling. Being close to the predicted values boosted my confidence. But I feel that the [-10, +10] range is for some attributes a bit confusing – or at least does not align with my internal value range.	I had the intuition that the values of the LLM were pretty good, but not perfect, and this could be observed and validated later on.	If I was certain about my choice and the generated result was close or matched it, my confidence increased. If the values mismatched, the confidence didn't decrease due to lack of familiarity with parts of the data. However, could imagine the opposite situation while working with a well-familiar dataset.	By seeing if the generated data aligned with my intuition, I felt more confident.	Most of my guesses in the Align view were in line with what the LLM generated.
How helpful was the tool to let you think through your reasoning while validating the data?	4	2	4	3	5	3
Please elaborate on the previous question. For example: what features, interactions, visualizations helped or prevented you from externalizing your reasoning.		Only the 'Align' view with the instant feedback helped me in externalizing my knowledge/reasoning. An improvement suggestion for the other views would be to make a 'note' about a generated insight for later provenance.	It gave me a structure and a framework to use this structure. So it guided me through the process and supported me on the journey.		Very useful features like understanding the distribution and alignment with human value, and in all aspects like attribute level, correlation level, and application level.	
How helpful was the Application View in assessing the impact of adding the LLM-generated attributes versus only using the existing attributes?	4	4	5	4	5	5
Please elaborate on the previous question. For example, what aspects of the view helped or prevented you.		This is a bit of a hypothetical answer because in the user study I only had one validated feature. But in general, the Buy-Hold-Sell matrix gives a good indication of the generated attributes.	It was simple to compare them.	The confusion matrices were helpful.	It is good to ensure and trust the application altogether and to see how useful the data you validated improved the performance of an ML model. Also, among classes where misclassification happened, and if there is a feature to re-generate data, it's easy to improve the model training and validation with application view.	
How effectively did the tool provide context that helped you assess the reliability of the LLM-generated information?	4	3	4	5	5	3
Please elaborate on the previous question. For example, what kind of context you found helpful or lacking.		The LLM explanations seemed a bit high-level, but I also didn't make use of them.		The Application view was effective at providing the context for the impact of LLM-generated attributes on classification. Additionally, the Align view was helpful for assessing the reliability in the item-based context.	Align item level and the correlation level analysis helped me a lot to assess the reliability.	

Table 4: *Quantitative and Qualitative Questions on Usefulness – responses by participants.*

3.5. Semi-structured interview

Semi-Structured Interview Question	P2	P3	P4	P5	P6	P7
What was the most useful feature of the tool?	Distribution view	Align view	Align view	Align view	Correlation view	Distribution view
Why?	Helps to see many item values at once.	Blind testing approach prevents you from being biased.	I could validate that each of the items were correct according to my intuition and this helps have confidence that the distributions would be correct as well.	Allows you to compare value to LLM-generated.	Understand in a holistic view and see what goes wrong without scrolling through one by one attribute.	Useful to have the option to slice the data and see whether it aligns. It was easier to do it on a broad scale. It was a nice starting point.
Least useful feature of the tool	Application view	Overview	Correlation view	Distribution view	Distribution view	Application view
Why?	It was difficult to understand the classifications and interpret the matrix.	Sorting does not help, correlations may not help, there is extra information. I would like to have seen more attributes at an overview without scrolling.	Not sure how to validate based on correlations. It would require more experience with the data. Least intuitive.	Did not understand what the subsets meant.	I could not get many insights from it. I am fine with validation even without looking at it.	It's difficult to understand the matrix and difficult to interpret the classification agreement.
Presumed that prompt engineering would be part of the process: which of the features of the tool motivated you the most to revise the prompt for the LLM?	Align view	Align view	Align view	Align view	Align view	Distribution view
Why?	It provides item values and explanations.	It helps me to get an impression if the LLM values meet my expectations. I like coming from concrete examples for which I know facts.	If I see many values being different from my expectations then it would tell me to reprompt. Correlation view may also help.		You can see item-wise comparison to the model and if you strongly believe you are correct then you can reprompt and ensure in the later steps data is more trustworthy.	I will first see distribution and then if I see something wrong I will go to the align view and regenerate data that were misaligned.
Would you have the workflow go from coarse to fine-grained or from fine-grained to coarse?	Distribution (between fine-to-coarse)	Fine-to-coarse	Fine-to-coarse	Fine-to-coarse	Fine-to-coarse	Coarse-to-fine
Why?	Distribution gives oversight and based on that I decide whether to go fine-grained or coarse.	Helps me to see examples.	Maybe correlations would be correct but not the underlying data. From fine-grained I can be sure since I inspect the data and be sure it is showing what is correct.	It's easier to get familiar with the data from fine-to-coarse. If I knew the data then from coarse-to-fine.	Initially you understand item-wise and then it makes sense to go to more coarse.	I like to get a broad overview first and then go to fine-grained.
Is there something you would like to share?		How is the ground-truth defined? I did not make much use of explanations and certainty.		Perhaps good to have a definition on the overview screen as well.	Would be good to have automatic uploading of dataset and have more control of LLM data generation.	It's useful for simulations and generating data.

Table 5: *Semi-structured interview responses by participants.*