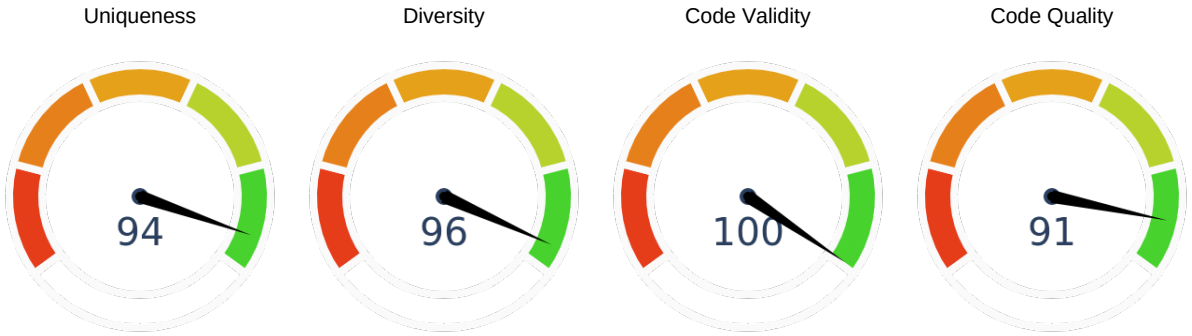
Data Quality Report

Key Metrics



Please see the Appendix for metric definitions, and the Conclusion for AI/ML considerations.

Dataset Overview

Metric	Value
Data Completeness	100.0%
Number of Rows	10000
Number of Columns	12
Categorical Columns	5
Text Columns	3
Numerical Columns	0
Seed Columns	2

Metric	Value
Unique Rows	94.0%
Semantically Unique Rows	-470.4%
Avg Words per Row	2.96
Avg Tokens per Row	60.61
Total Tokens	606052
Avg Text Diversity	0.95
Avg Gini-Simpson Index	N/A

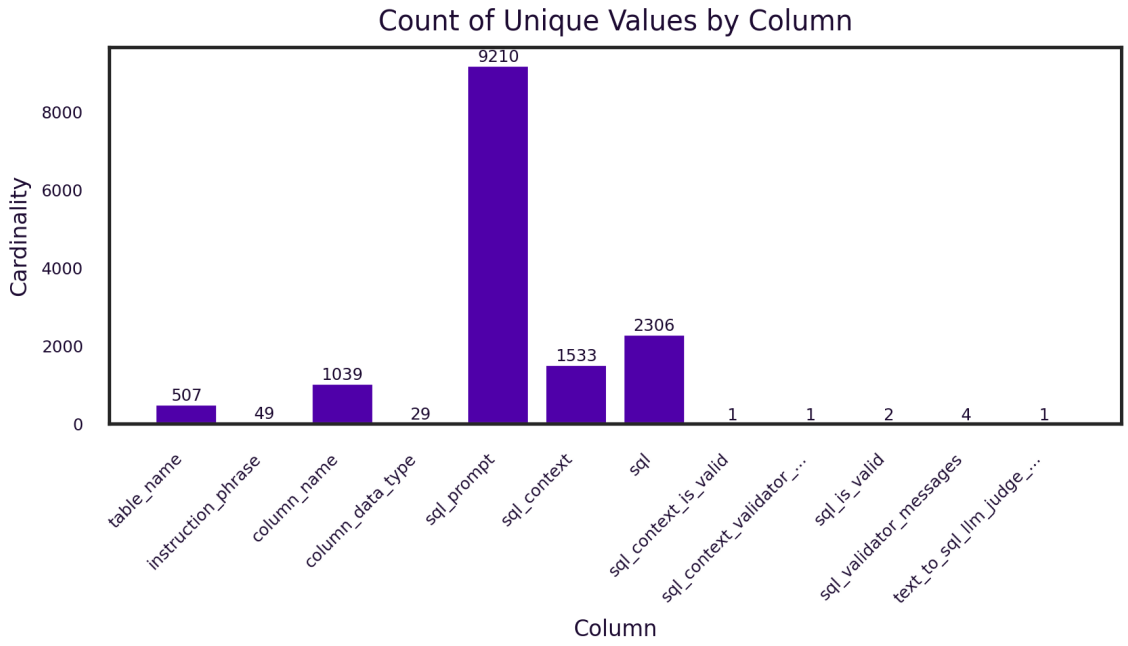
Data Preview

table_name: login_attempts
instruction_phrase: Create a selection of
column_name: user_id
column_data_type: INT
sql_prompt: Create a selection of user_id from login_attempts
sql_context: CREATE TABLE login_attempts (user_id INT);
sql: SELECT user_id FROM login_attempts;
sql_context_is_valid: True
sql_context_validator_messages: []
sql_is_valid: True
sql_validator_messages: []
text_to_sql_llm_judge_results: {'readability': None, 'relevance': None, 'scalability': None, 'standards': None}

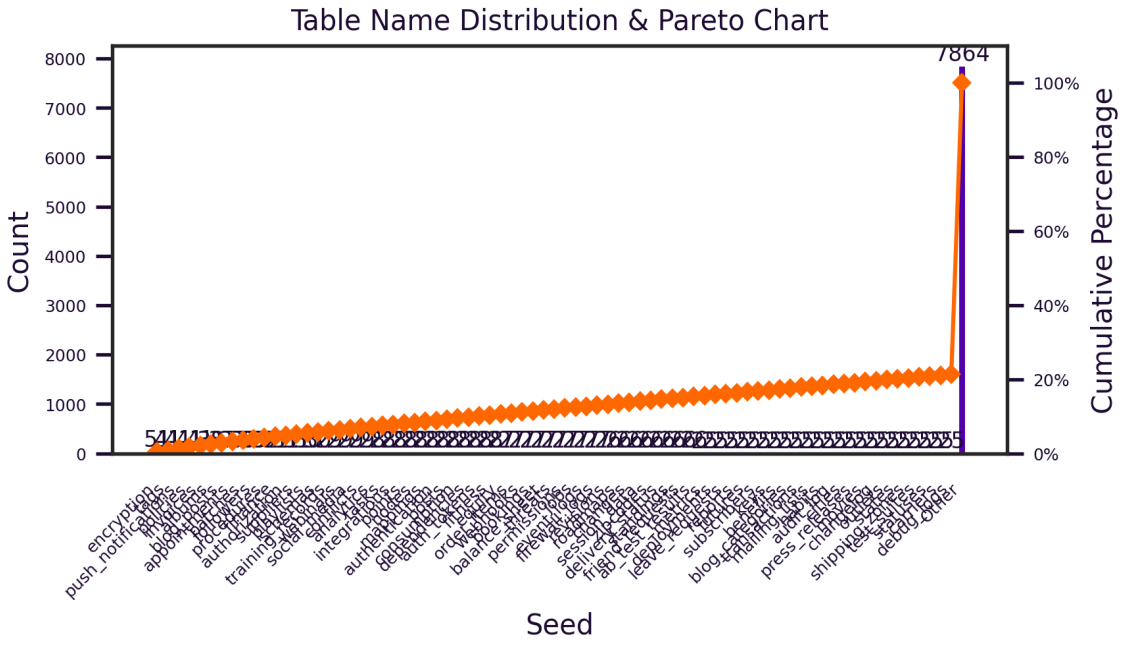
Dataset Schema

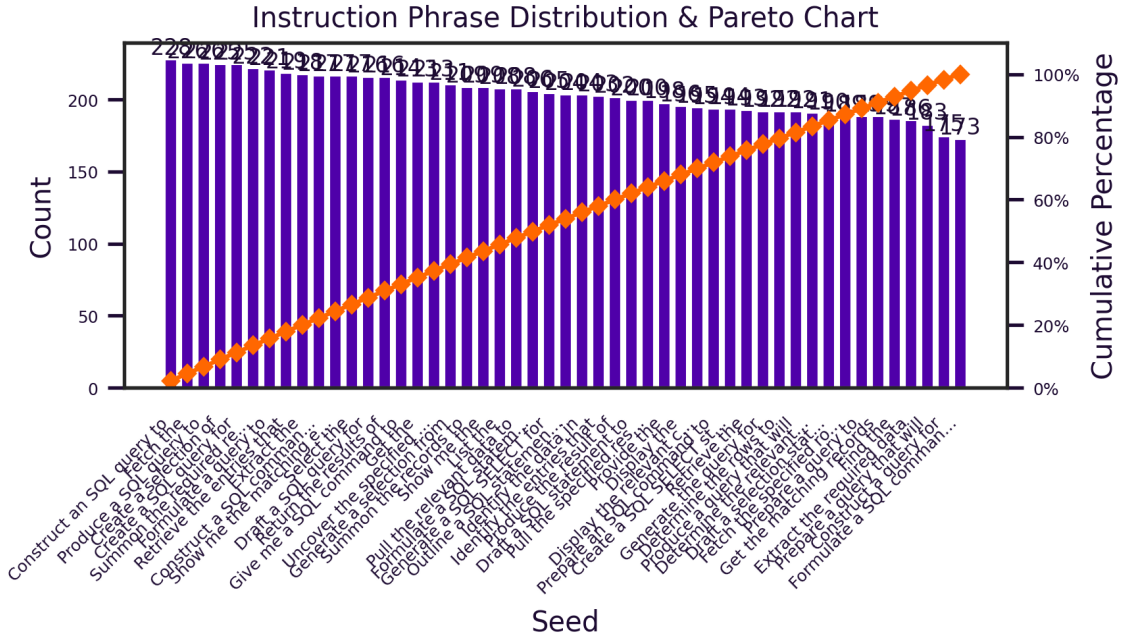
Column Name	Data Type	Total Count	% Null	Avg Length	Avg Tokens	Note
table_name	object	10000	0.00%	1.0	1.6573	Seed Column
instruction_phrase	object	10000	0.00%	4.0775	4.2998	Seed Column
column_name	object	10000	0.00%	N/A	2.2345	Requested Column
column_data_type	object	10000	0.00%	1.0	2.6428	Requested Column
sql_prompt	object	10000	0.00%	9.1515	11.5342	Requested Column
sql_context	object	10000	0.00%	7.0	11.6687	Requested Column
sql	object	10000	0.00%	4.3417	6.9695	Post-Processing Column
sql_context_is_valid	bool	10000	0.00%	0.0	1.0	Post-Processing Column
sql_context_validator_messages	object	10000	0.00%	1.0	1.0	Post-Processing Column
sql_is_valid	bool	10000	0.00%	0.0	1.0	Post-Processing Column
sql_validator_messages	object	10000	0.00%	1.003	1.0091	Post-Processing Column
text_to_sql_llm_judge_results	object	10000	0.00%	4.0	16.0	Post-Processing Column

Column Cardinality

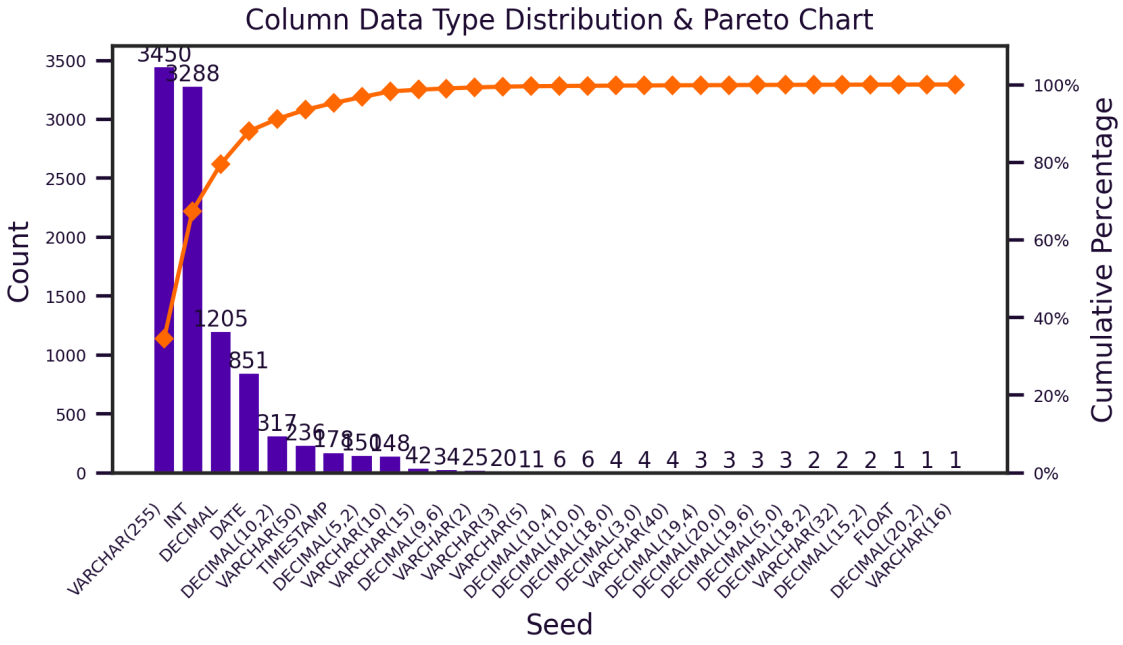


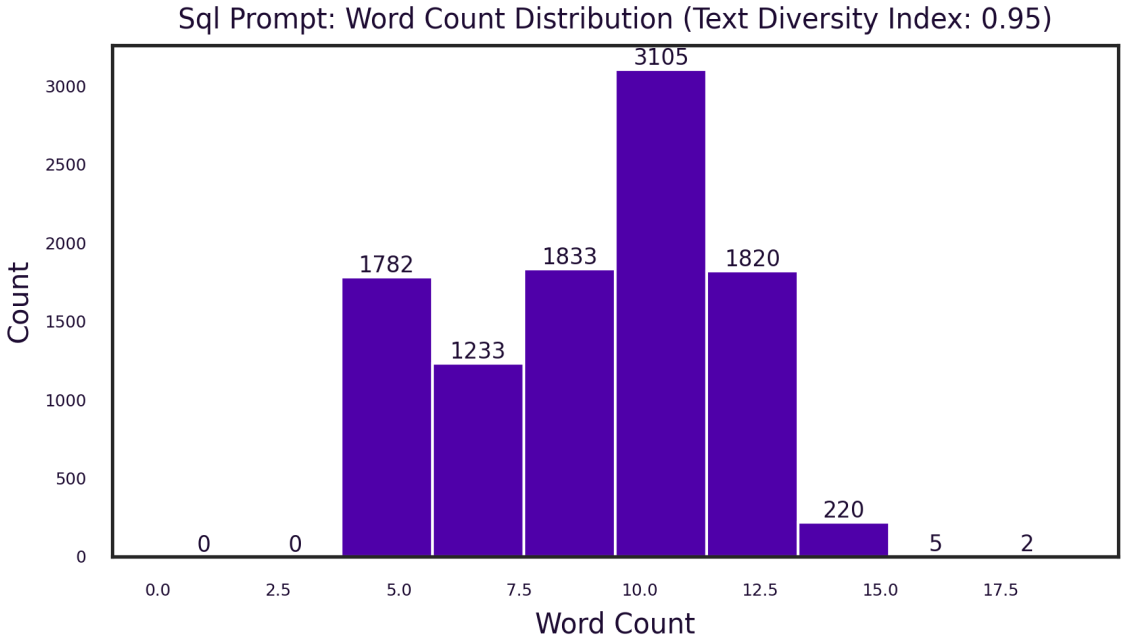
Seed Column Distributions

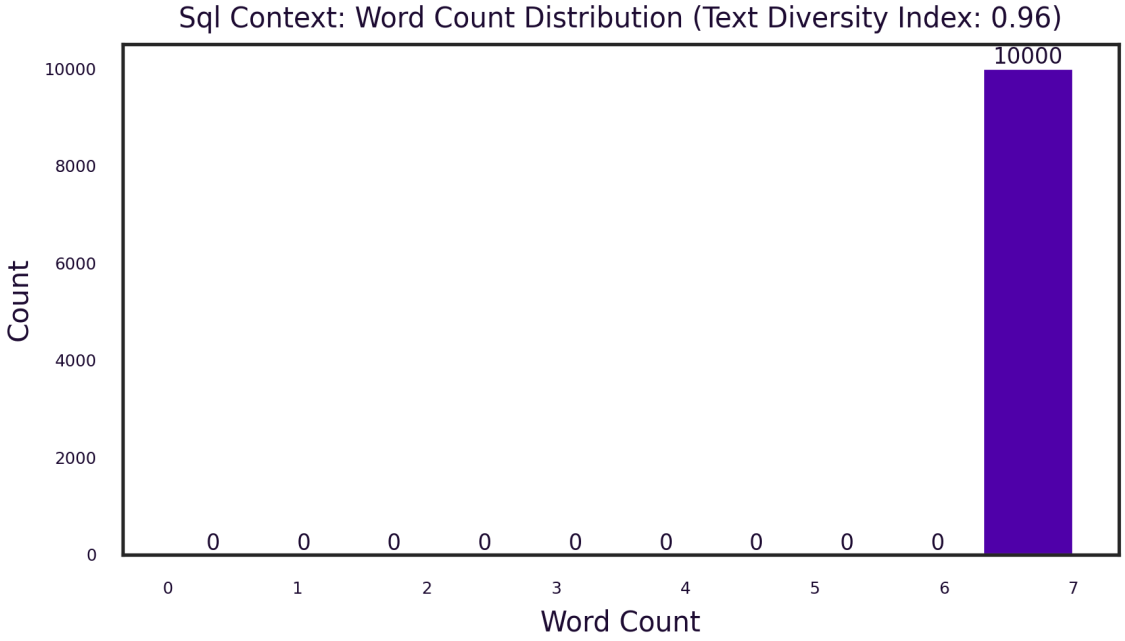




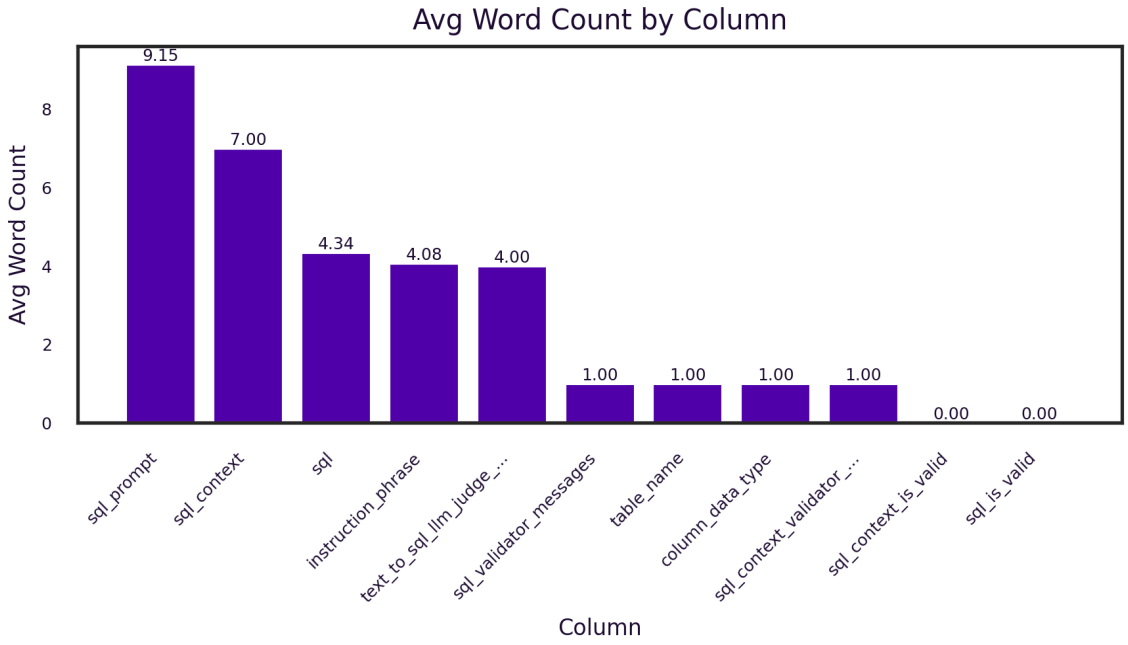
Generated Column Distributions



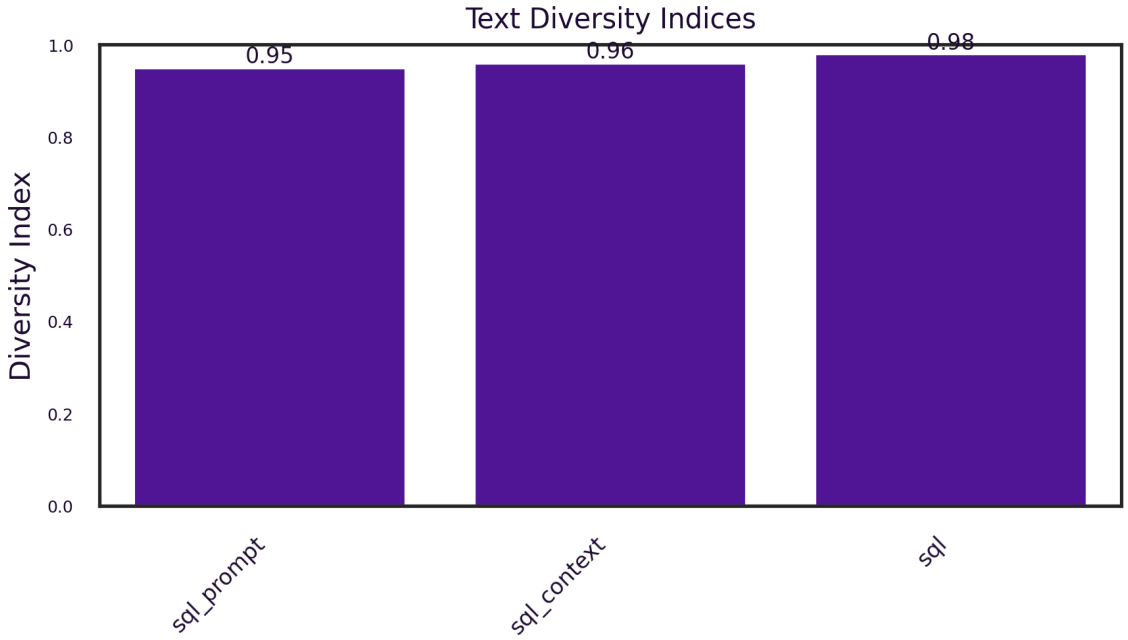



Sql Context: Word Count Distribution (Text Diversity Index: 0.96)
Count
Word Count
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
0
0
0
0
0
0
0
0
0
10000

Average Word Count per Column



Text Diversity Indices



Conclusion

Main Takeaways

Data Uniqueness: The dataset has 94.0% unique rows and -470.4% semantically unique rows. This high uniqueness suggests excellent variety in the generated data.

Code Quality Metrics: The code shows 100.0% validity (excellent validity) and a quality score of 90/100 (excellent quality).

Column Cardinality: The columns show an average of 1223.5 unique values, with the highest cardinality being 9210. Some columns have very high cardinality, suggesting rich, detailed data.

Text Complexity: Average of 3.0 words per row (60.6 tokens), with 606,052 total tokens in the dataset. The text content is relatively concise.

Diversity Metrics: Text columns show an average diversity index of 0.963. This indicates excellent overall diversity in the dataset.

AI/ML Considerations

Pre-training: The dataset's uniqueness and diversity can provide a rich foundation for pre-training language models or other AI systems. High-cardinality columns may help in learning broad representations, while low cardinality columns could aid in learning important categorical distinctions. If text diversity is high, it could be particularly valuable for building robust language models that can handle a wide range of contexts and styles.

Fine-tuning: The distribution patterns in data should guide follow-ups and the fine-tuning process. Columns with high semantic uniqueness could be especially useful for fine-tuning models on specific domains or tasks, as they likely contain a wide range of relevant examples. Consider the average token count per row when deciding on sequence length for transformer-based models during fine-tuning.

Iterating on Data to Fill Data Gaps: Distribution charts should help identify underrepresented categories and opportunities for iterating on data to improve model performance. If certain text diversity scores are low, consider ways to introduce more variety, either through different seeding or better prompting. For columns with very high cardinality, consider if grouping or categorization might be beneficial to prevent overfitting on rare categories. If semantic uniqueness is low in certain areas, it might indicate a need for more diverse examples in those categories to improve model generalization.

General Considerations: Care should be taken to address possible imbalances in data. Monitor for potential biases that could be propagated or amplified by machine learning models. The text complexity (average tokens per row) should inform decisions about model architecture and preprocessing steps.

Appendix

Metric Definitions

Key Metrics

Metric	Definition
Uniqueness	Percentage of rows that are unique in the dataset, based on exact matching
Semantic Uniqueness	Percentage of rows that are semantically unique, based on TF-IDF cosine similarity
Diversity	Average diversity across all text columns. Text Diversity Index/Gini-Simpson Index are used for text/category columns. Higher values indicate more diverse content
Code Validity (when requested)	Average percentage of valid code across all columns for which validation was requested
Code Quality (when requested)	Average code quality across all columns for which evaluation was requested
	*** NOTE *** All key metrics are presented on a scale from 0-100 for ease of interpretation. Only columns requested by the user are included in the calculation of Key Metrics: ['column_data_type', 'sql_prompt', 'sql_context']. Seed columns as well as validation/evaluation and other post-processing columns are excluded.

Dataset Overview Metrics

Metric	Definition
Data Completeness	Overall percentage of non-null values across all columns
Seed Columns	Number of columns used to seed the data generation process
Unique Rows	Number of rows that are that are unique, based on exact matching
Semantically Unique Rows	Number of rows that are semantically unique, based on TF-IDF cosine similarity
Avg Words per Row	Average number of words for a text column
Avg Tokens per Row	Average number of LLM tokens, as determined by Tiktoken
Avg Text Diversity	Average Text Diversity Index (see below) across all text columns
Avg Gini-Simpson Index	Average Gini-Simpson Index (see below) across all categorical columns
Text Diversity Index	A diversity index for text columns. It is defined as the average correlation between each row's TF-IDF vector and the dataset's TF-IDF matrix. Higher values indicate greater diversity
Gini-Simpson Index	A diversity index for categorical columns. It quantifies the probability that two values taken at random from the column (with replacement) are different. Higher values indicate greater diversity

Dataset Schema & Preview Metrics

Metric	Definition
Data Type	Data type of the values in a specific column
Total Count	Total number of values in a column
% Null	Percentage of null values in a column
Avg Length	Average length of column values (text columns only)
Avg Tokens	Average number of LLM tokens in column values, as determined by Tiktoken (text columns only)