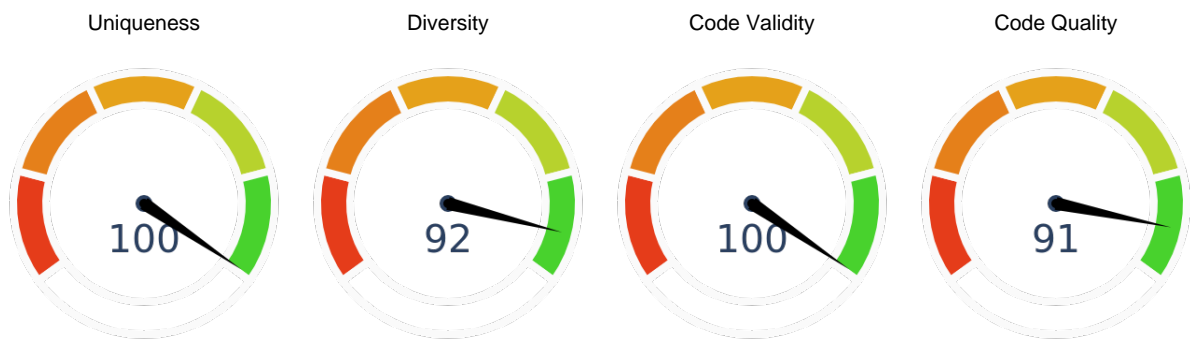


Data Quality Report

Key Metrics



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

Dataset Overview

Metric	Value	Metric	Value
Data Completeness	100.0%	Unique Rows	100.0%
Number of Rows	100	Semantically Unique Rows	92.0%
Number of Columns	12	Avg Words per Row	2.82
Categorical Columns	4	Avg Tokens per Row	57.02
Text Columns	3	Total Tokens	5702
Numerical Columns	0	Avg Text Diversity	0.93
Seed Columns	2	Avg Gini-Simpson Index	N/A

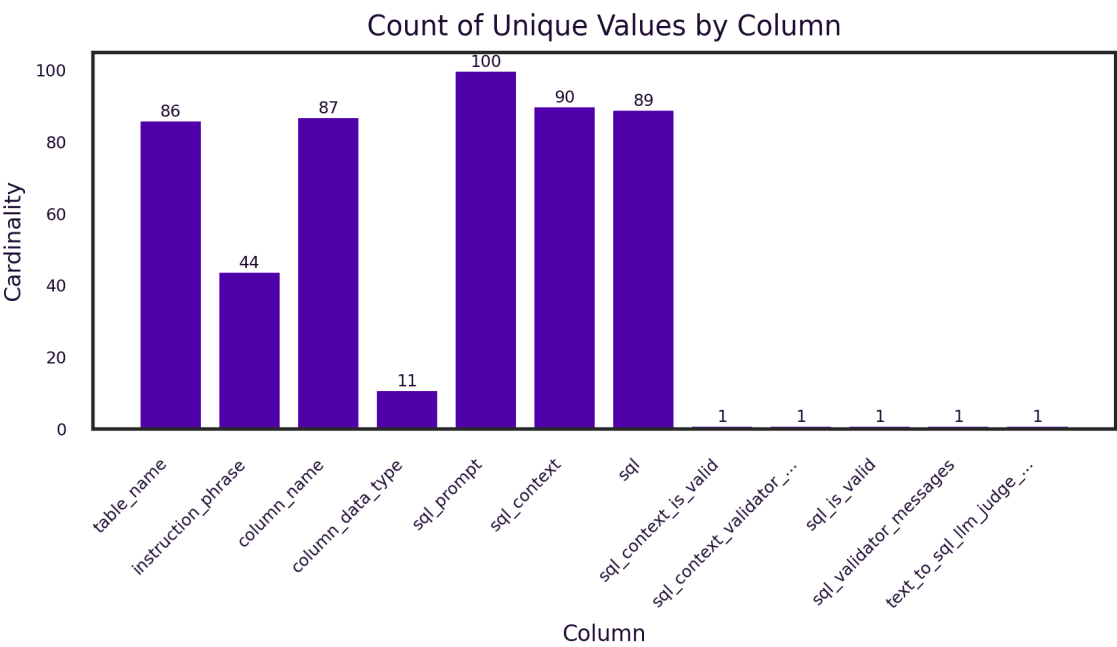
Data Preview

```
table_name: patches
instruction_phrase: Prepare an SQL command to
column_name: patch_version
column_data_type: VARCHAR(255)
sql_prompt: Prepare an SQL command to select exactly 1 patch_version from the patches.
sql_context: CREATE TABLE patches ( patch_version VARCHAR(255) );
sql: SELECT patch_version FROM patches;
sql_context_is_valid: True
sql_context_validator_messages: []
sql_is_valid: True
sql_validator_messages: []
text_to_sql_llm_judge_results: {'readability': {'reasoning': 'The query is well-formatted and easy to understand, but it could bene...
```

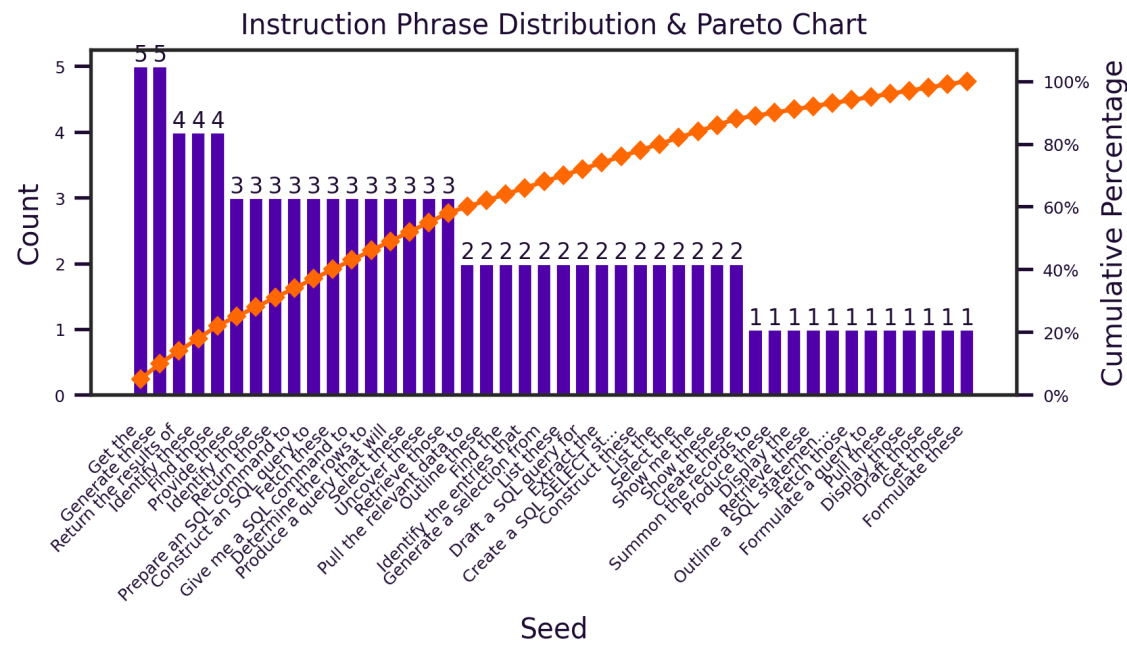
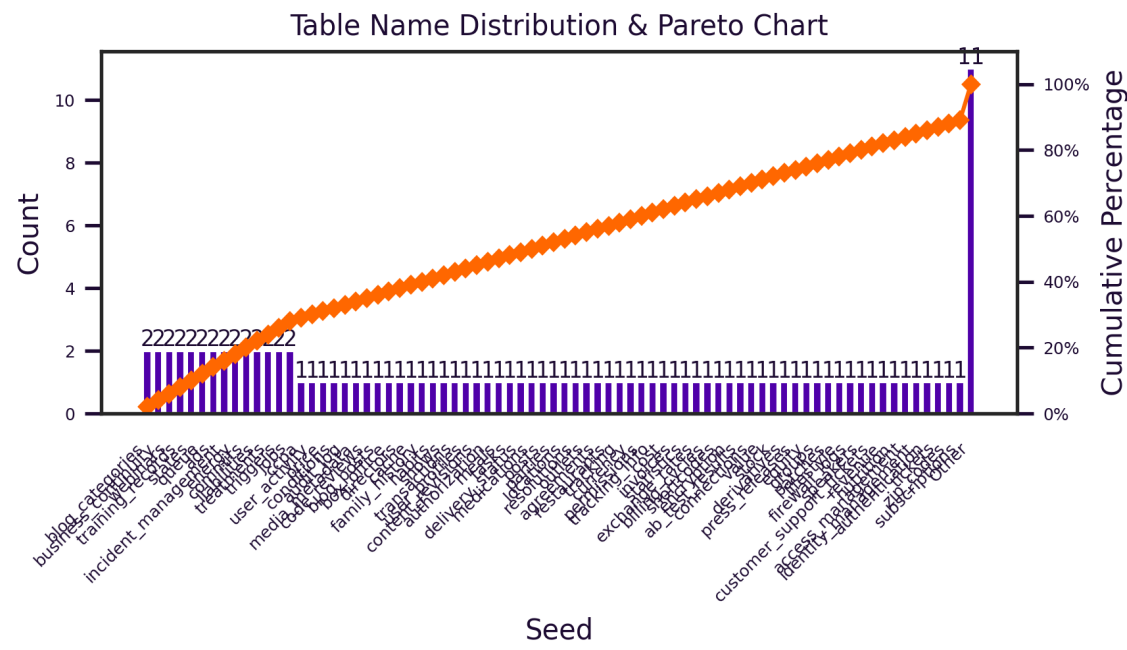
Dataset Schema

Column Name	Data Type	Total Count	% Null	Avg Length	Avg Tokens	Note
table_name	object	100	0.00%	1.0	1.71	Seed Column
instruction_phrase	object	100	0.00%	2.92	3.13	Seed Column
column_name	object	100	0.00%	N/A	2.36	Requested Column
column_data_type	object	100	0.00%	N/A	2.54	Requested Column
sql_prompt	object	100	0.00%	7.18	9.4	Requested Column
sql_context	object	100	0.00%	7.0	11.68	Requested Column
sql	object	100	0.00%	4.12	6.71	Post-Processing Column
sql_context_is_valid	bool	100	0.00%	0.0	1.0	Post-Processing Column
sql_context_validator_messages	object	100	0.00%	1.0	1.0	Post-Processing Column
sql_is_valid	bool	100	0.00%	0.0	1.0	Post-Processing Column
sql_validator_messages	object	100	0.00%	1.0	1.0	Post-Processing Column
text_to_sql_llm_judge_results	object	100	0.00%	4.0	16.0	Post-Processing Column

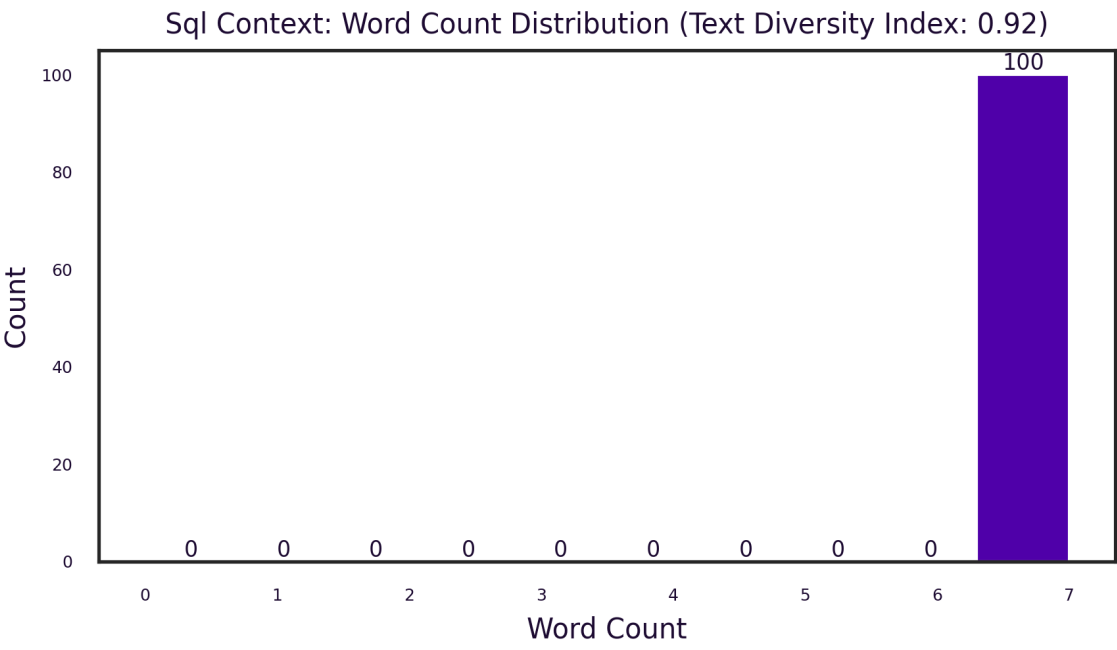
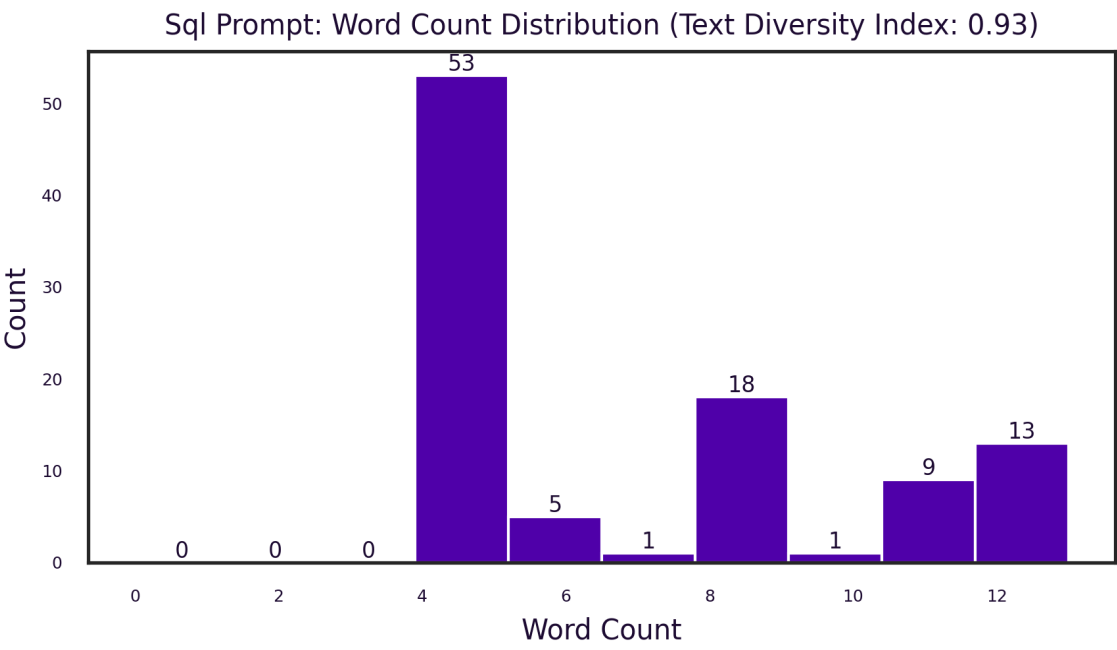
Column Cardinality



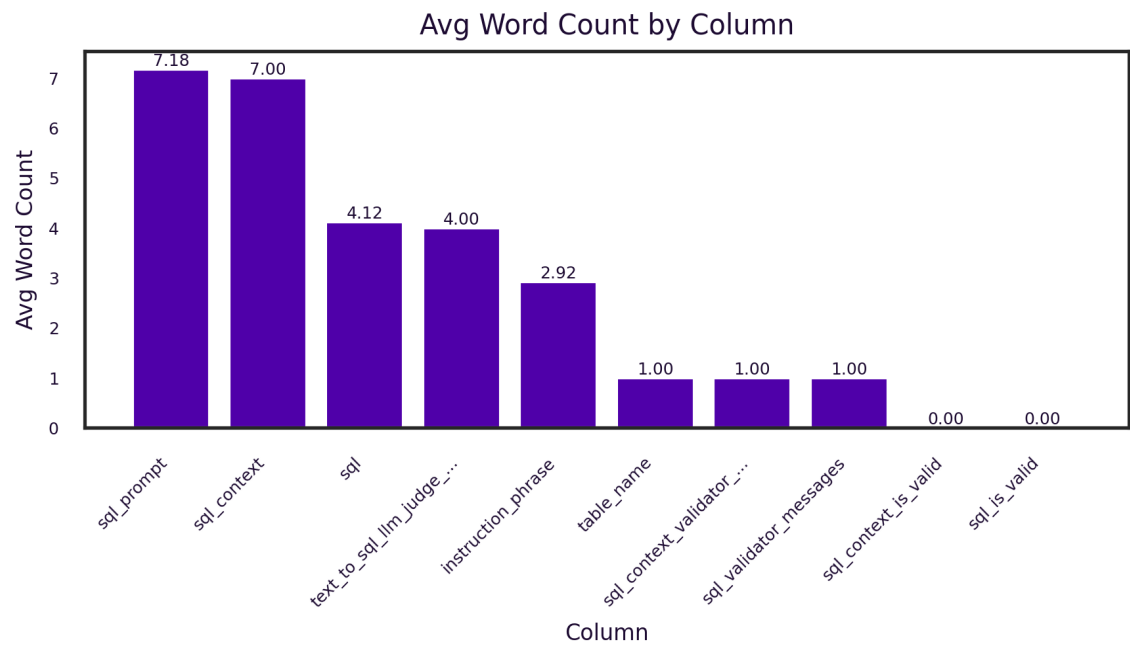
Seed Column Distributions



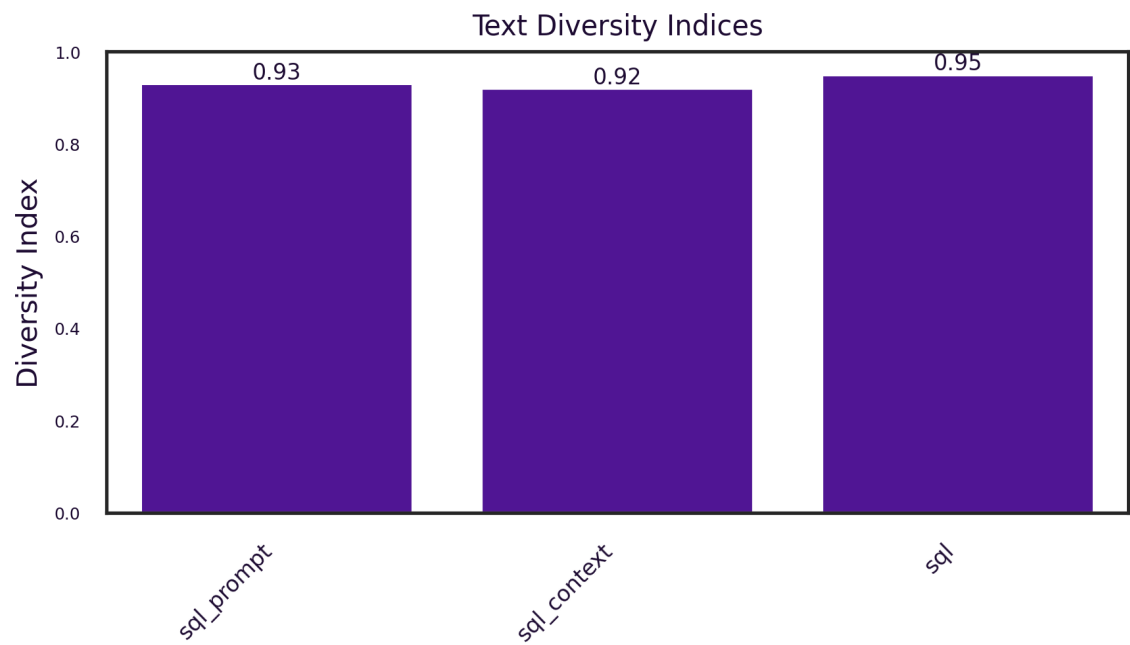
Generated Column Distributions



Average Word Count per Column



Text Diversity Indices



Conclusion

Main Takeaways

Data Uniqueness: The dataset has 100.0% unique rows and 92.0% 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 91/100 (excellent quality).

Column Cardinality: The columns show an average of 42.7 unique values, with the highest cardinality being 100. The cardinality levels suggest more categorical or constrained data.

Text Complexity: Average of 2.8 words per row (57.0 tokens), with 5,702 total tokens in the dataset. The text content is relatively concise.

Diversity Metrics: Text columns show an average diversity index of 0.933. 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: ['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)