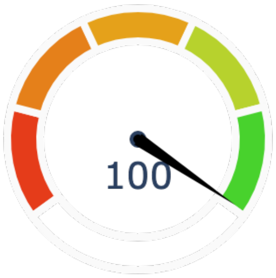
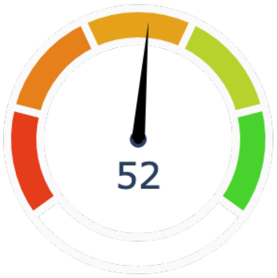
Data Quality Report

Key Metrics

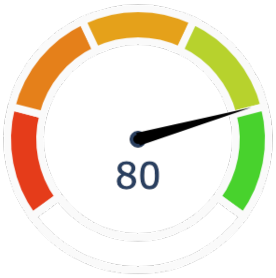
Unique Rows



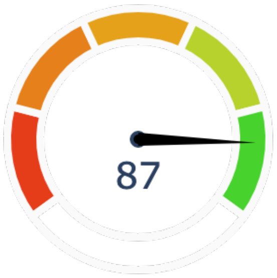
Data Diversity



Code Validity



Code Quality



Dataset Overview

This section provides key metrics on the structure, uniqueness, complexity, and quality of the data.

Metric	Value
Data Completeness	100.0%
Number of Rows	10
Number of Columns	12
Categorical Columns	3
Text Columns	8
Numerical Columns	0
Seed Columns	4

Metric	Value
Unique Rows	100.0%
Semantically Unique Rows	100.0%
Avg Words per Row	44.55
Avg Tokens per Row	920.60
Total Tokens	9206
Avg Text Diversity	0.5208
Avg Gini-Simpson Index	N/A

Single Row Preview

id: 482

domain: Financial Services

topic: Portfolio Management

complexity: Beginner: Basic syntax, data types, and control structures

prompt: Write a Python function named 'calculate_portfolio_value' that takes two arguments: 'stocks' and 'cr...

dependency_list: ['pandas', 'matplotlib', 'numpy', 'scikit-learn', 'seaborn']

code: import numpy as np stocks = {'AAPL': 10, 'GOOG': 5, 'AMZN': 2} crypto = {'BTC': 0.5, 'ETH': 2.0} s...

code_is_valid: True

code_score: 7.5

code_severity: warning

code_messages: [{'symbol': 'redefined-outer-name', 'msg': "Redefining name 'stocks' from outer scope (line 3)", 'ca...

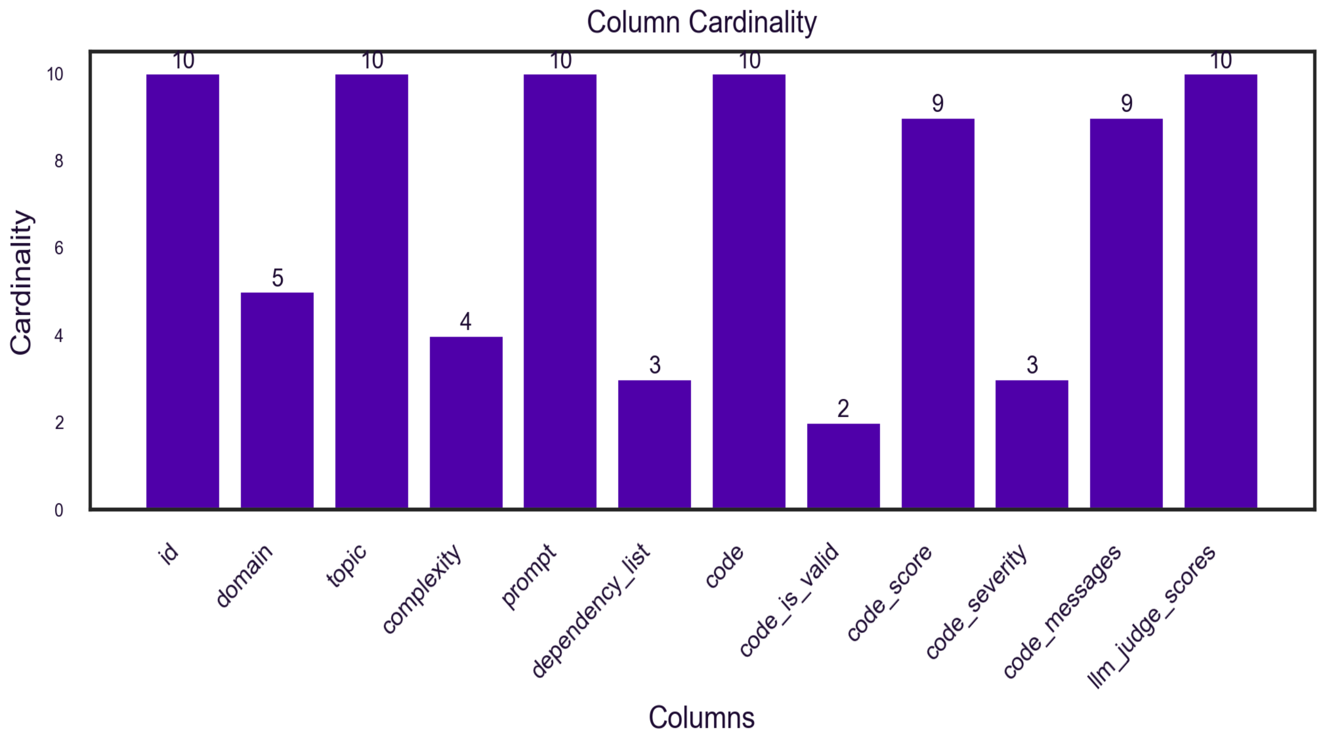
llm_judge_scores: { "relevance": {"score": 4, "reasoning": "The code perfectly meets all specified requirements, i...

Dataset Schema & Preview

The schema table provides an overview of each column in the dataset.

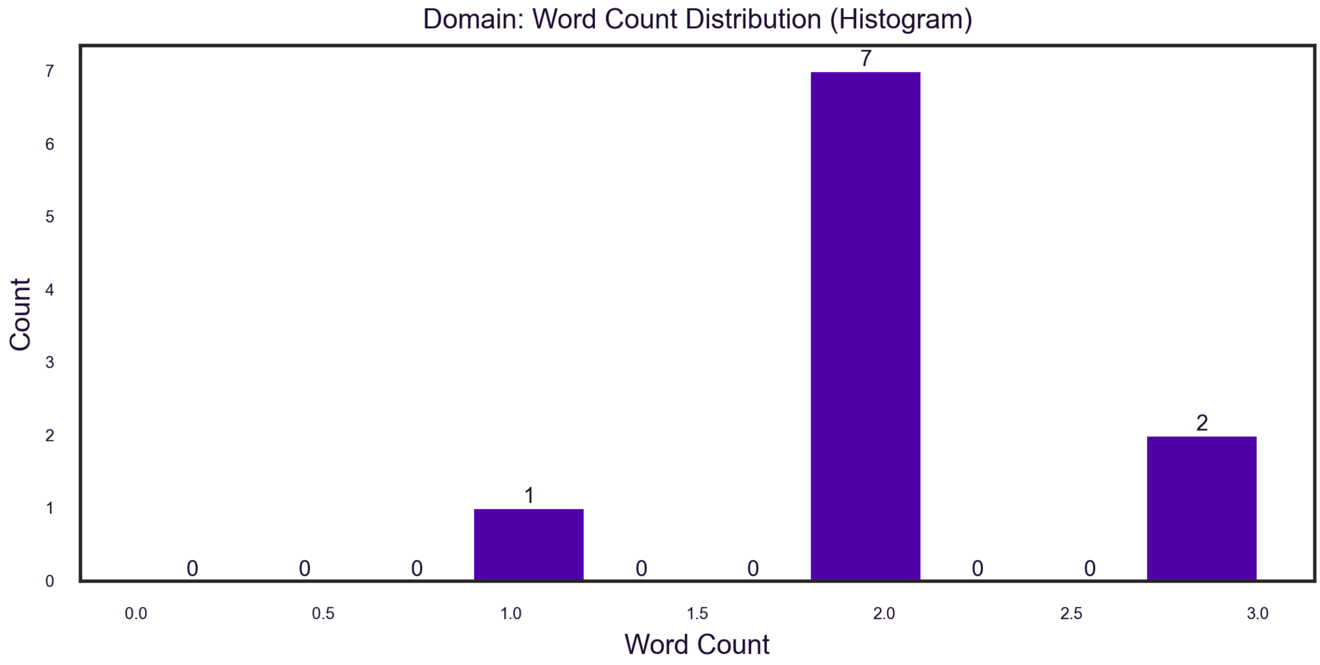
Column Name	Type	Total Count	% Null	Average Length	Avg Tokens	Note
id	int64	10	0.00%%	N/A	1.0	Unique ID
domain	object	10	0.00%%	2.1	2.5	Seed Column
topic	object	10	0.00%%	2.2	2.6	Seed Column
complexity	object	10	0.00%%	6.6	10.9	Seed Column
prompt	object	10	0.00%%	174.1	245.7	
dependency_list	object	10	0.00%%	5.0	19.8	Seed Column
code	object	10	0.00%%	119.6	281.2	
code_is_valid	bool	10	0.00%%	0.0	1.0	Post Processing Column
code_score	float64	10	0.00%%	0.0	5.1	Post Processing Column
code_severity	object	10	0.00%%	1.0	1.0	Post Processing Column
code_messages	object	10	0.00%%	38.1	119.3	Post Processing Column
llm_judge_scores	object	10	0.00%%	141.3	230.4	Post Processing Column

Column Cardinality

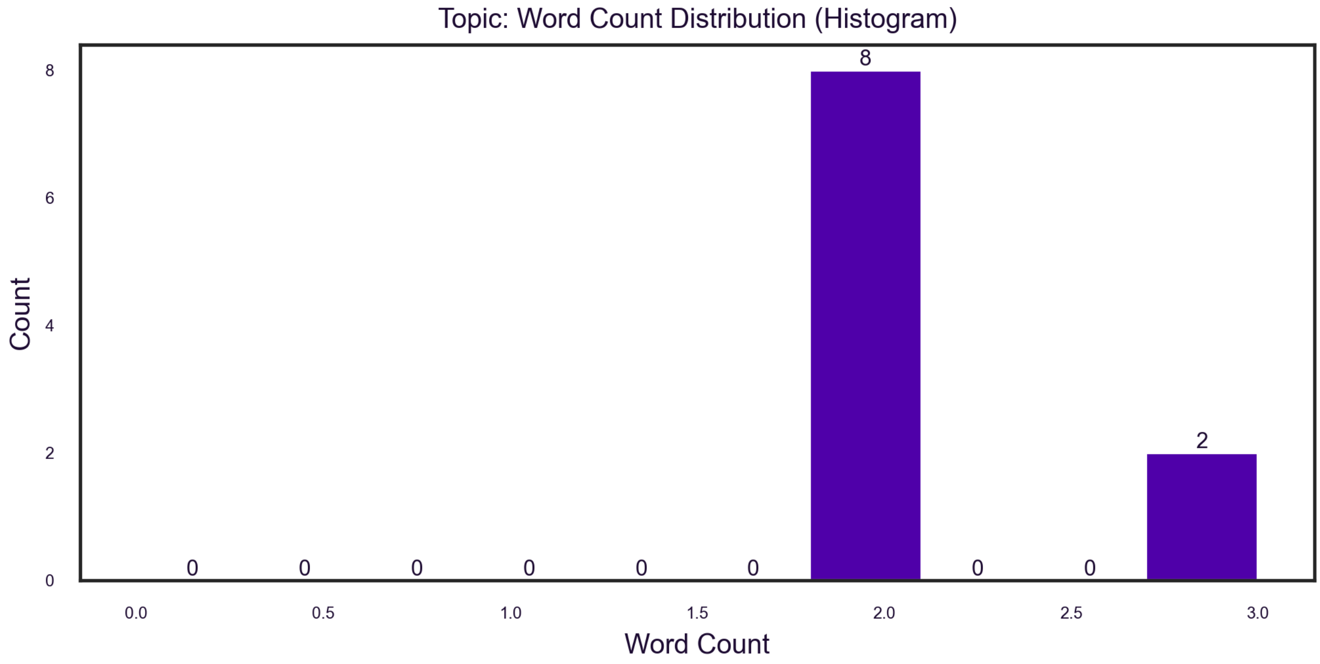


Seed Column Distributions

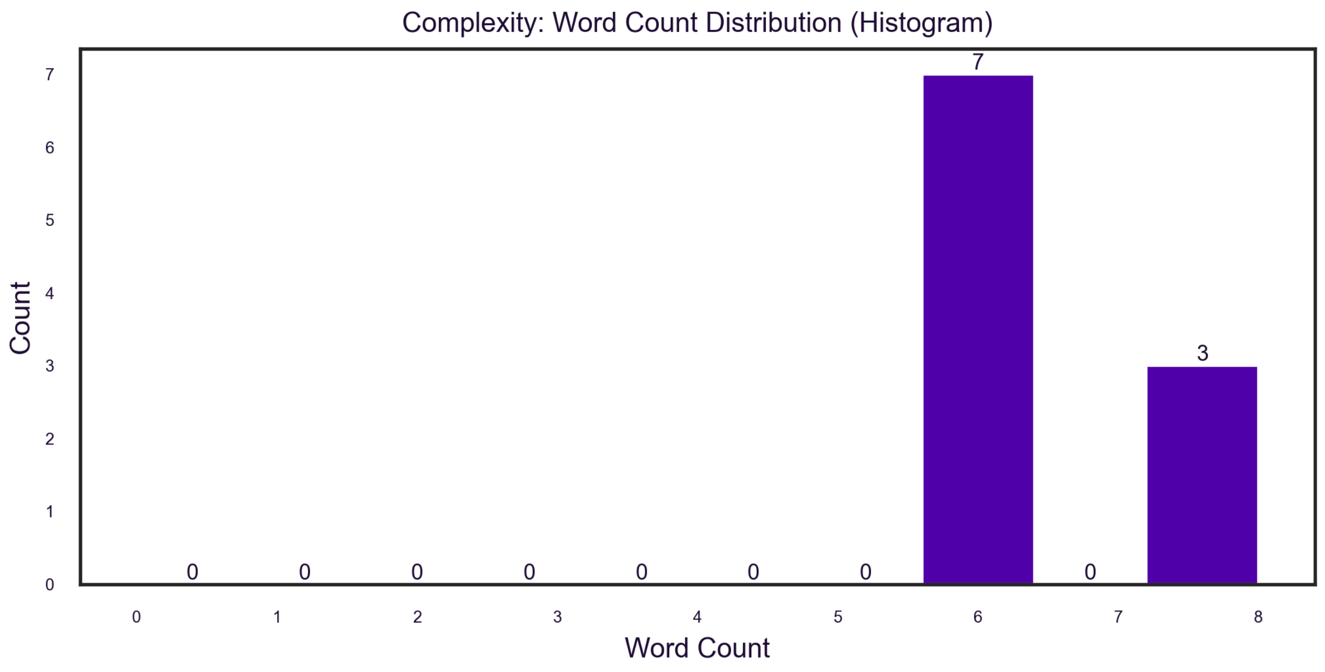
Domain Distribution (Text Diversity Index: 0.47)



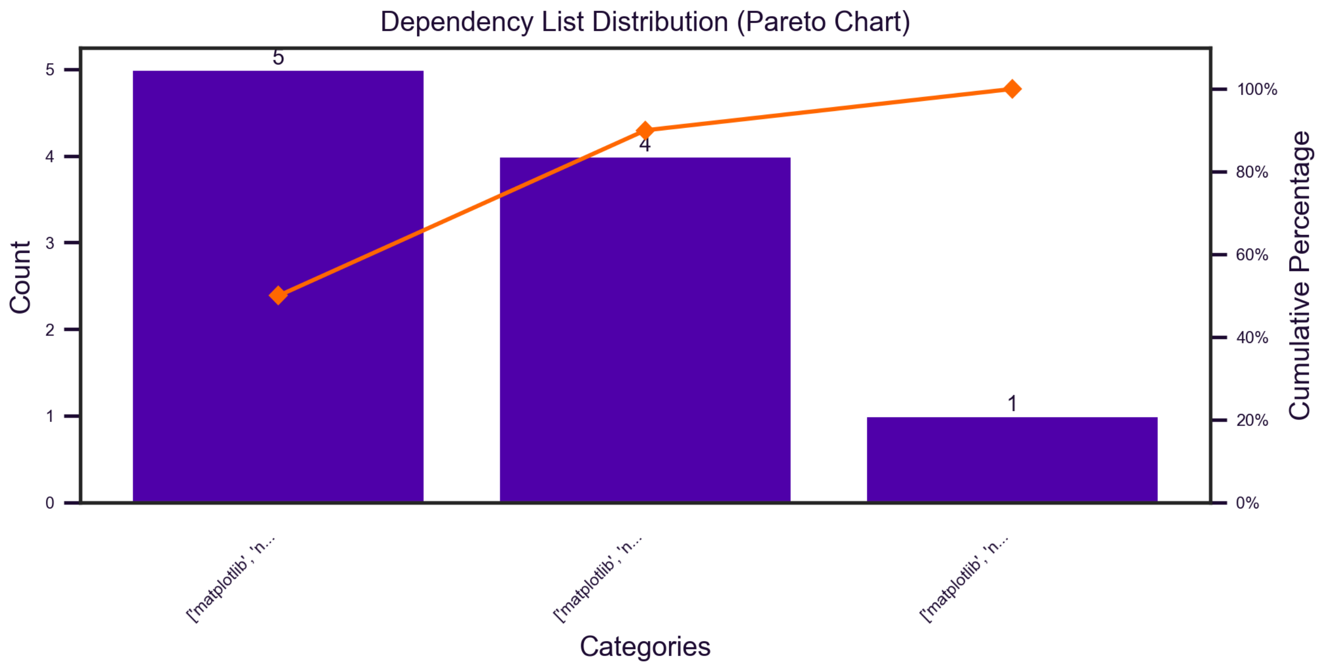
Topic Distribution (Text Diversity Index: 0.68)



Complexity Distribution (Text Diversity Index: 0.45)

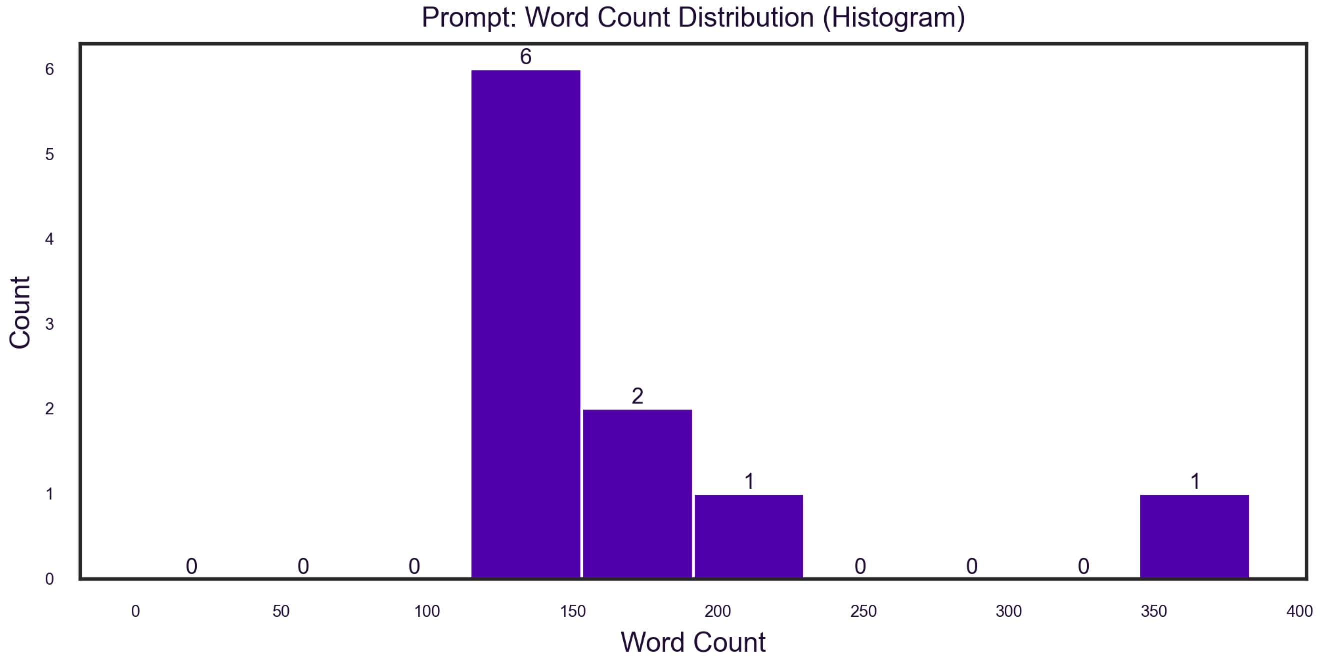


Dependency List Distribution (Gini-Simpson Index: 0.58)

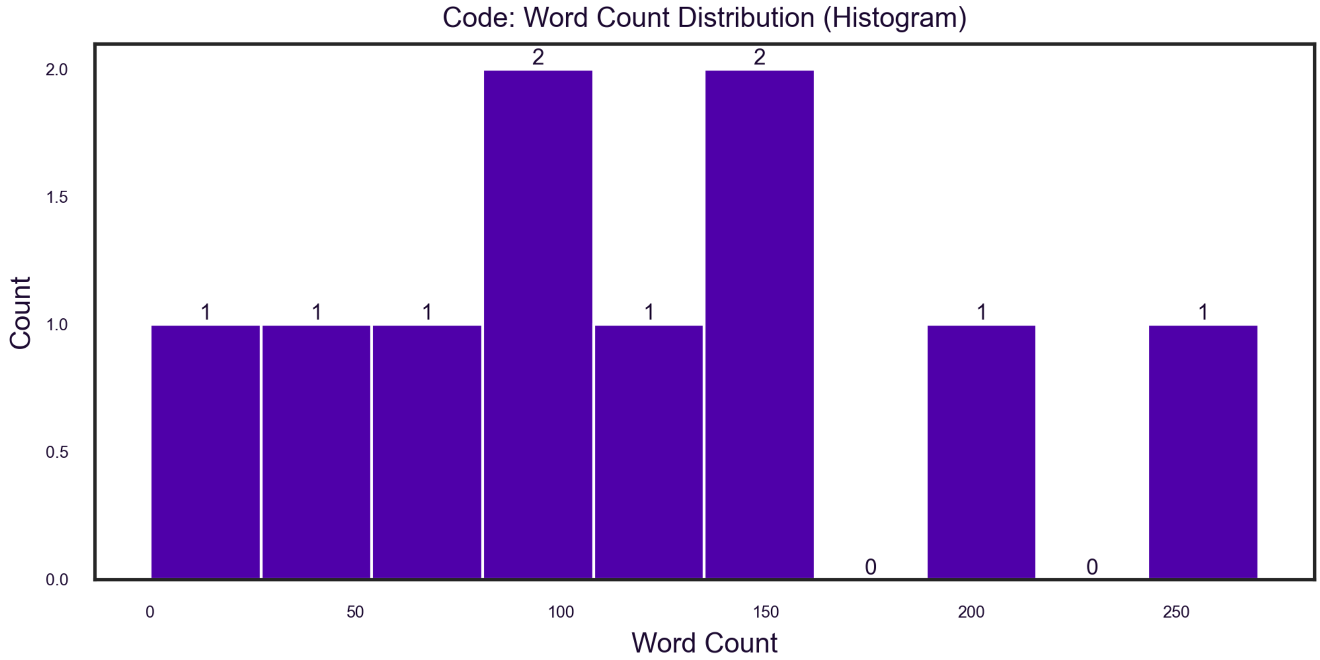


Generated Column Distributions

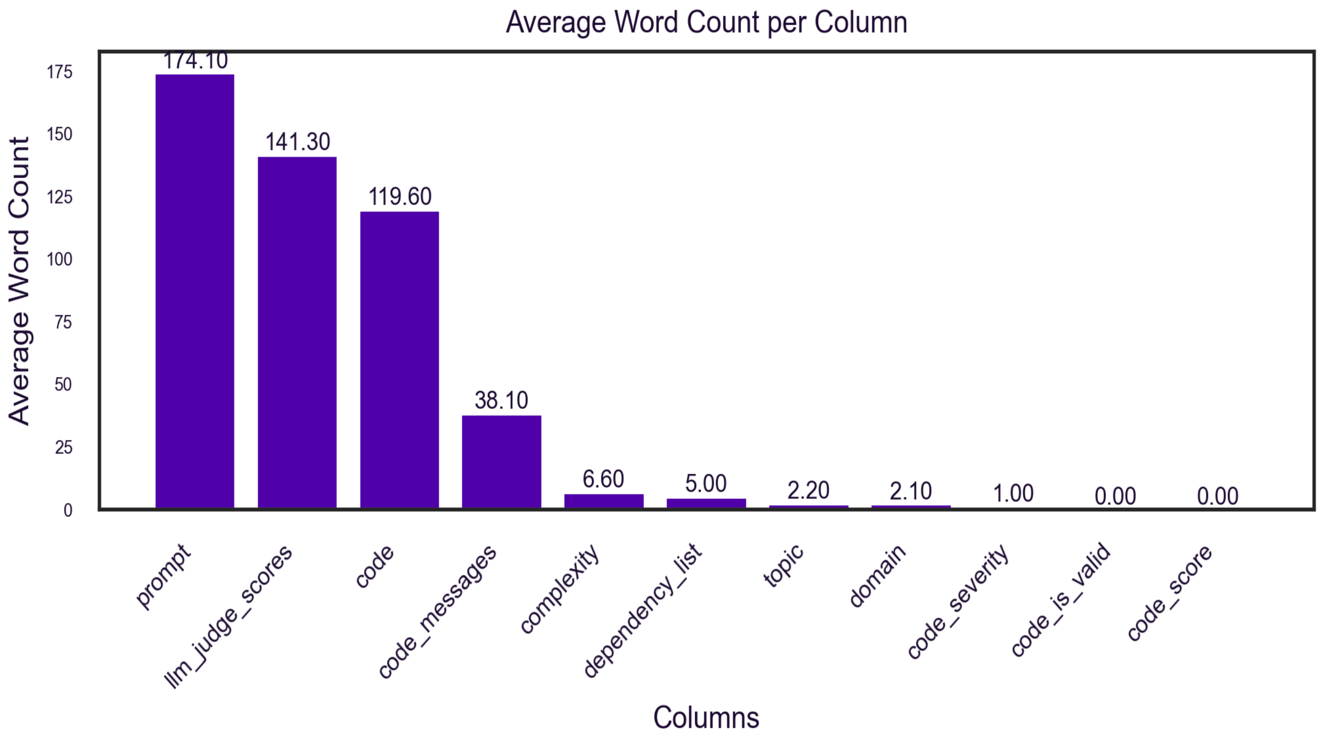
Prompt Distribution (Text Diversity Index: 0.42)



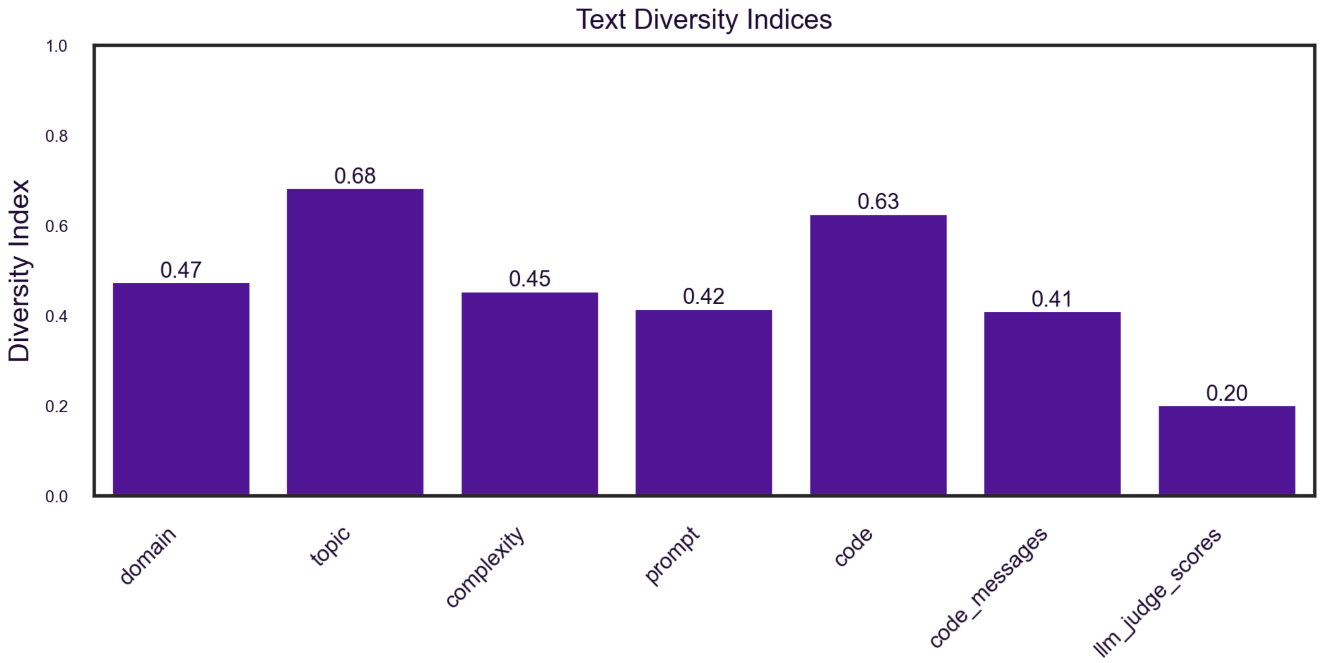
Code Distribution (Text Diversity Index: 0.63)



Average Word Count per Column



Text Diversity Indices



Conclusion

This report provides a comprehensive view of the dataset's structure, content diversity, and the nature of the data it contains. Key takeaways include:
1. Data Uniqueness: With 100.0% unique rows and 100.0% semantically unique rows, the dataset shows a high degree of individuality in its rows. This suggests a rich and varied dataset.

2. Column Cardinality: The dataset contains columns with varying cardinalities. This diversity in column types allows for both granular analysis and higher-level pattern recognition.

3. Distribution Patterns: The charts reveal the distribution patterns within each column, highlighting potential focus areas or biases in the data. Understanding these distributions is crucial for balanced analysis and identifying underrepresented categories.

4. Text Complexity: With an average of 44.55 words per row, the dataset shows a moderate level of complexity. This gives an indication of the depth of information contained in each row.

5. Text Diversity: The text diversity indices provide insight into the variety of content within text columns. Higher diversity can be beneficial for tasks requiring a broad range of examples, while lower diversity might indicate more standardized content.

Implications for Machine Learning:

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 revealed in the charts should guide the fine-tuning process. Imbalanced categories may require techniques like weighted sampling or loss adjustment to ensure equal representation during fine-tuning.
- 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 word count per row when deciding on sequence length for transformer-based models during fine-tuning.

Designing/Iterating on Data to Fill Data Gaps:
- Analyze the distribution charts to identify underrepresented categories. These areas may require additional data collection or augmentation to ensure comprehensive model performance.
- If certain text diversity scores are low, consider ways to introduce more variety in those columns, either through data augmentation techniques or targeted data collection.
- 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:
- The overall uniqueness of the dataset impacts models that require diverse examples. However, care

should be taken to address any imbalances revealed in the distribution charts.
- Monitor for potential biases in the data that could be propagated or amplified by machine learning models.
- Consider privacy implications, especially for high-cardinality columns that might contain identifiable information.
- The text complexity (average words per row) should inform decisions about model architecture and preprocessing steps.

Metric Definitions

This section provides definitions for the metrics used in the report.

Key Metrics
Only generated columns requested by the user are included in the calculation of Key Metrics: ['prompt', 'code']. Helper columns like ID columns, seed columns or informational columns like code validation columns, data quality evaluation columns are excluded from Key Metrics calculation.
• **Unique Rows:** Percentage of rows that are unique in the dataset.
• **Semantically Unique Rows:** Percentage of rows that are semantically unique, based on TF-IDF.
• **Text Diversity:** Average Text Diversity Index (defined below) across all text columns, with higher values indicating more diverse content.
• **Gini-Simpson Diversity:** Average Gini-Simpson Index (defined below) across all categorical columns. Higher values indicating greater diversity.

Dataset Overview
The enhanced dataset overview provides key metrics about the structure, uniqueness, complexity, and quality of the data:
• **Number of Rows and Columns:** Indicates the size and dimensionality of the dataset.
• **Categorical and Numerical Columns:** Gives insight into the types of data present, helping to guide appropriate analysis techniques.
• **Data Completeness:** Shows the overall percentage of non-null values across all columns, indicating the dataset's overall quality and potential need for imputation.
• **Unique and Semantically Unique Rows:** Demonstrates the level of data diversity and potential redundancy in the dataset.
• **Average Words per Row:** Provides an indication of the typical complexity or detail level of each entry.
• **Average Tokens per Row and Total Tokens:** These metrics correspond to tokens used in Large Language Models (LLMs), giving an estimate of the dataset's complexity from an LLM processing perspective.
• **Average Text Diversity:** Average Text Diversity Index (defined below) across all text columns, with higher values indicating more diverse content.
• **Average Gini-Simpson Index:** Average Gini-Simpson Index (defined below) across all categorical columns. Higher values indicating greater diversity.

Dataset Schema & Preview
The schema table provides an overview of each column in the dataset, including the data type, the count of non-null and null values, and the average length (where applicable). This information is crucial for understanding the structure of the data and identifying potential data quality issues such as missing values or unexpected data types.
• **Data Type:** Categorical, Numeric, Text or Other. Categorical columns are those whose percentage of unique values are low; Text columns are non-Categorical columns with at least 2 spaces per Row, on average.
• **Total Count:** Total number of values in the Column.

- **% Null:** Percentage of null values in the Column.
- **Average Length:** Average character count of the values (for each text column).
- **Avg Tokens:** Average number of tokens in the values (for each text column).

Column Cardinality

- **Column cardinality:** Represents the number of unique values for each column in the dataset. Higher cardinality indicates more diverse values within a Column.

Column Distributions

Column distributions show the frequency of different values within each column. These visualizations help identify common patterns, imbalances, or biases in the data.

- **Pareto Chart:** A Pareto chart illustrates the distribution of domain in the dataset. The bars represent the count for each category, while the line shows the cumulative percentage. Only the top 75 categories are shown individually. The remaining categories are grouped as 'Other'. This visualization helps identify the most significant categories and their relative importance.
- **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.
- **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.