Analysis Report for Train_with_outliers.jsonl

1. Executive Summary

This report presents a comprehensive analysis of a question-answer dataset containing **144137** pairs. The dataset exhibits **high** complexity with a diversity score of **0.843** (95% CI: 0.841-0.845). Based on content and structural analysis, we recommend fine-tuning a **meta-llama/Llama-3-70b-instruct** model with the optimized hyperparameters detailed in Section 5.

2. Dataset Statistics

Metric	Questions	Explanations
Average Length (tokens)	17.2	255.2
Median Length (tokens)	16.0	189.0
Maximum Length (tokens)	129	982
90th Percentile Length	27.0	532.0
Contains Code	48.7%	98.2%

The dataset contains a total of approximately **39,266,954** tokens. The distribution of tokens suggests a high-complexity dataset that requires a sequence length of at least **2048** tokens to accommodate the longest samples.

3. Content Analysis

3.1 Topic Distribution

The semantic analysis identified **2** distinct topics in the questions:

- **Topic_1**: file (19585.419), using (14514.607), files (14161.968), seq (13726.935), data (12439.254)
- **Topic_2**: using (12160.530), data (12017.339), gene (11786.735), genome (8931.855), sequences (7328.517)

3.2 Complexity Assessment

The dataset complexity was assessed as **High** with a complexity score of **12.61**. This assessment is based on multiple factors:

- Semantic diversity: 0.843
- Average explanation length: 255.2 tokens
- Code presence in explanations: 98.2%

- Average question length: 17.2 tokens
- Topic model perplexity: 471.34

4. Training Data Recommendations

Based on the dataset characteristics, we recommend the following data split:

- Training samples: 115309 (79% of dataset)
- Testing samples: 28827 (19% of dataset)

5. Hyperparameter Recommendations

For optimal fine-tuning results, we recommend the following hyperparameters:

Parameter	Recommended Value	Rationale
Base Model	meta-llama/Llama-3-70b-instruct	Selected based on dataset complexity and size
Sequence Length	2048	Accommodates 184% of the maximum required length
Epochs	10	Optimized for dataset size of 144137 samples
Learning Rate	1e-05	Adjusted for high complexity content
Batch Size	2 with 4 gradient accumulation steps	Optimized for model size and memory efficiency
LoRA Rank (r)	32	Selected based on dataset diversity and complexity
LoRA Alpha	64	Set to 2x LoRA rank for optimal adaptation
LoRA Dropout	0.05	Higher value for robustness with complex data

For generation during evaluation and inference, we recommend:

- Maximum generation length: 798 tokens
- Temperature: 0.7
- Top-p (nucleus sampling): 0.92
- Minimum-p: 0.1

6. Visualization Summary

This analysis includes the following visualizations (available in the 'plots' directory):

1. Token length distributions for questions and explanations
2. Code presence analysis
3. Topic model visualization
4. Key terms wordcloud

5. Summary dashboard of dataset characteristics

7. Conclusion

This dataset demonstrates high complexity with 2 distinct topics. The recommended fine-tuning approach with a Llama-3-70b-instruct model and optimized parameters should yield strong results in capturing the question-answer patterns present in the data.