Athena Interactive Analytics of Experiment

Welcome to the analytics report for the experiments. This report includes fully interactive visuals generated with Plotly. You can explore the data by turning visuals on or off by clicking items in the legend.

Exercise Information

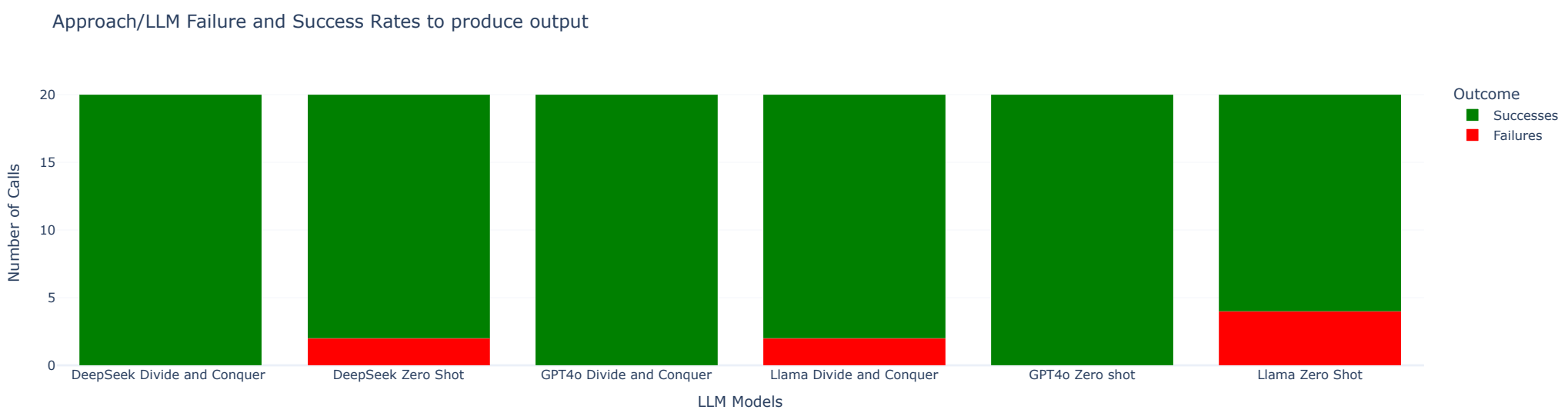
Exercise: System Design Review (SS21)

Maximum Points

9 points

Credits Analaytics

Plot 1



Plot 2

Normalized Absolute Differences Between LLM and the Tutor

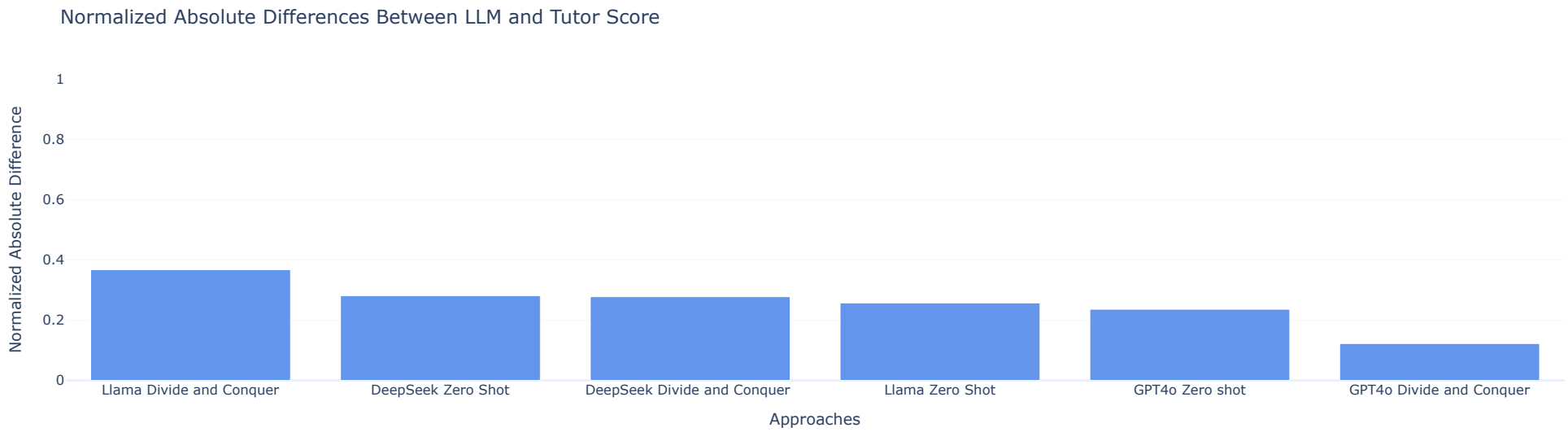
This bar plot visualizes the **normalized absolute differences** in scores between the LLM and other approaches. Each bar represents an approach, sorted from the highest to the lowest difference, and normalized by dividing the average absolute difference by the maximum possible score.

Insights:

- **The height of each bar:** Indicates how far, on average, the scores from a specific approach differ from the tutor's feedback after normalization.
- **Lower values:** Suggests a lower divergence on average from the tutor assessment
- **Higher values:** Indicate greater deviation, showing approaches that diverge more from tutor assessments.

Note: **THIS PLOT IS NOT AN ACCURATE REPRESENTATION OF ALIGNMENT WITH TUTOR FEEDBACK**

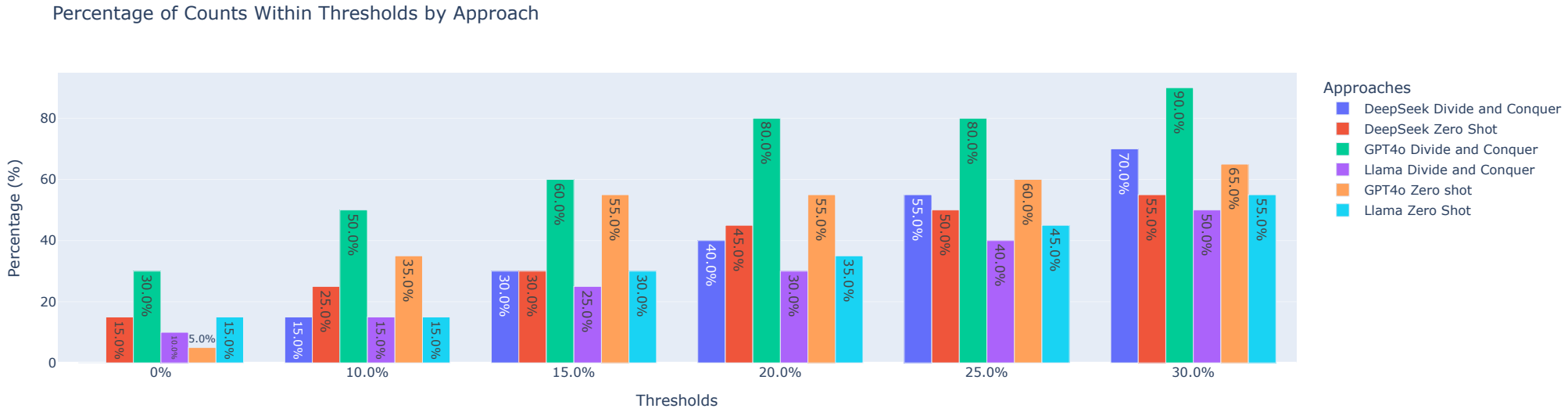
Note: **Refer to the next plot for a better representation of alignment**



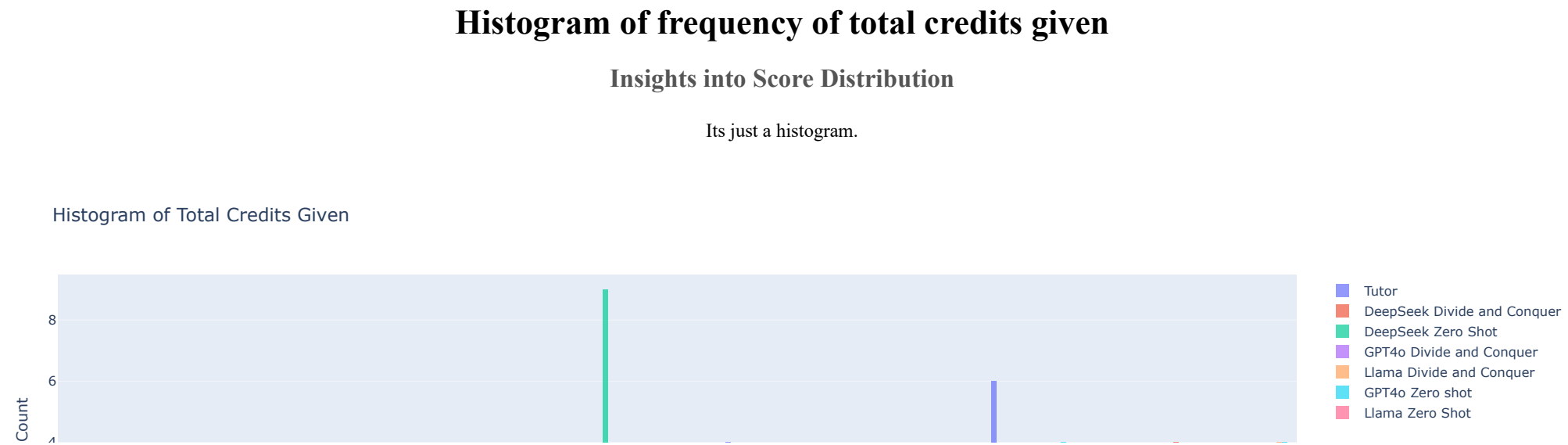
Plot 3

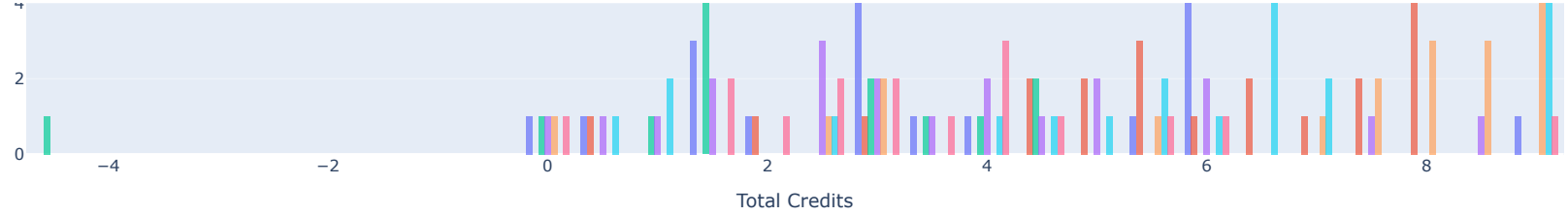
Percentage of LLM results, that fall within a range of maximum points difference from the tutor feedback

For example, for threshold 10 per cent. If the max points are 5, it only includes those llm results that are within 0.5 points of the tutor feedback. A bar of 20 per cent would translate to 20 per cent of the llm results being within 0.5 points of the tutor feedback.



Plot 4





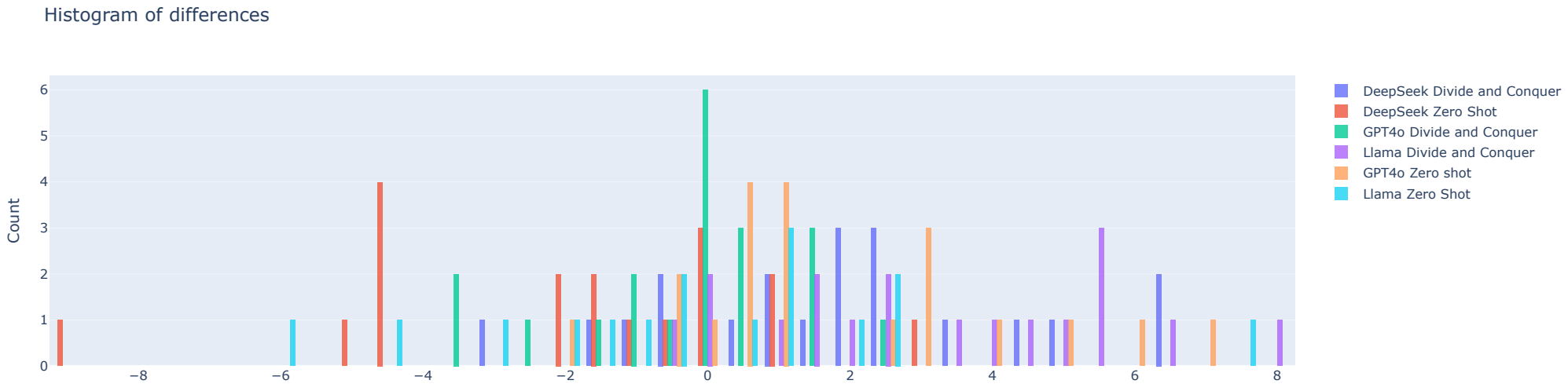
Plot 5

Distribution of Score Disparity Between LLM and Tutor

This graph represents the distribution of score differences between the LLM and the tutor. Negative values indicate that the LLM has scored the submission lower than the tutor, while positive values suggest the opposite.

The chart provides insights into the consistency and bias of the LLM's grading compared to the tutor. Viewers should look for patterns such as a strong concentration of values near zero, which would indicate agreement, or significant skew towards negative or positive values, highlighting systematic under- or over-grading by the LLM.

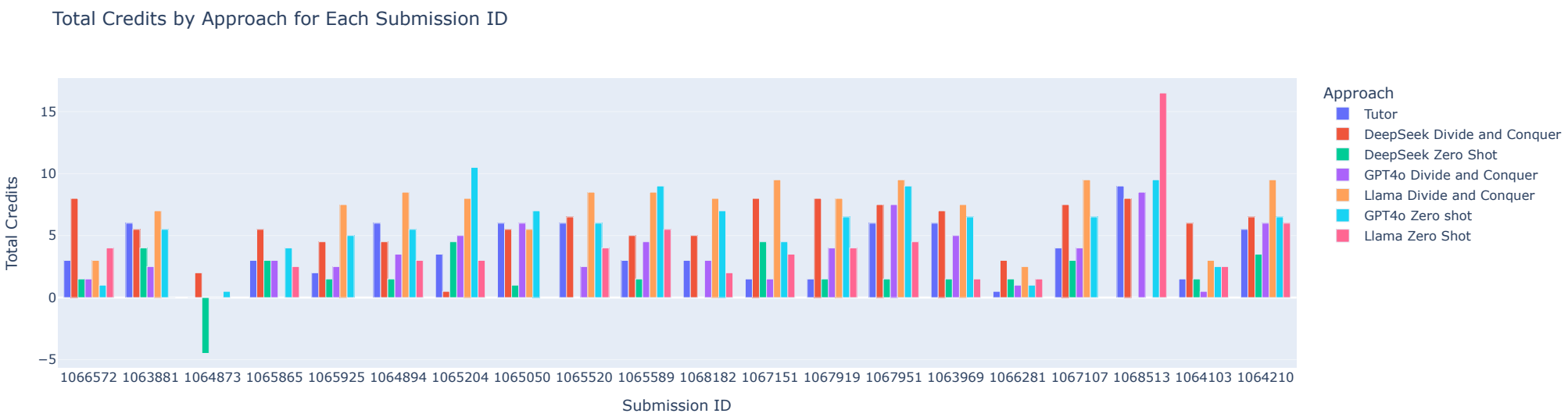
This visualization can help identify discrepancies and areas where the LLM may need calibration or adjustment to align more closely with tutor assessments.



Difference LLM - Tutor

Plot 6

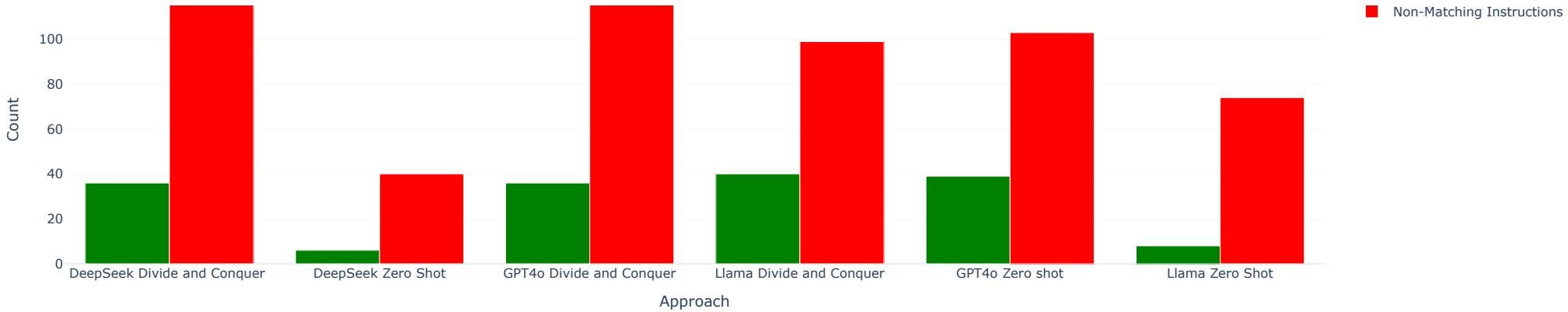
Total Credits awarded by each model on each submission



Structured Grading Instruction IDs Analytics

Grading Instruction Usage Analysis

This visualization compares the grading instructions used by different approaches against the "Tutor" approach. The green bars represent the count of grading instructions that match those of the Tutor approach, while the red bars show the count of non-matching instructions. This analysis highlights alignment and deviations between approaches.


Non-Matching Instructions
Count
0
20
40
60
80
100
DeepSeek Divide and Conquer
DeepSeek Zero Shot
GPT4o Divide and Conquer
Llama Divide and Conquer
GPT4o Zero shot
Llama Zero Shot
Approach