Performance Comparison of llama3.3 and llama_d3 Models

Analysis Report

May 19, 2025

1 Introduction

This document compares the performance of two language models, `llama3.3` and `llama_d3`, based on a set of evaluation metrics: hallucination score, ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L), METEOR, and cosine similarity (ground truth to prediction and context to prediction). The analysis aims to determine which model performs better and under what conditions.

2 Metric Definitions

The following metrics were used to evaluate the models:

- **Hallucination Score**: Measures the extent of incorrect or unsupported information in the output (lower is better).
- **ROUGE (1, 2, L)**: Measures lexical overlap between generated and reference cape text, with ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence) (higher is better).
- **METEOR**: Evaluates semantic similarity, accounting for synonyms and word order (higher is better).
- **Cosine Similarity (`gt_to_pred`)**: Measures similarity between generated text and ground truth (higher is better).
- **Cosine Similarity (`context_to_pred`)**: Measures alignment between generated text and input context (higher is better).

3 Metric Comparison

The table below summarizes the performance metrics for both models.

Table 1: Performance Metrics for `llama3.3` and `llama_d3`

Metric	`llama3.3`	`llama_d3`
Hallucination Score	0.0707	**0.0108**
ROUGE-1	**0.2075**	0.1631
ROUGE-2	**0.1427**	0.1188
ROUGE-L	**0.2075**	0.1631
METEOR	**0.6894**	0.6133
Cosine Similarity (`gt_to_pred`)	**0.8983**	0.8440
Cosine Similarity (`context_to_pred`)	0.7014	**0.7029**

4 Analysis

- **Hallucination Score**: `llama_d3` (0.0108) significantly outperforms `llama3.3` (0.0707), indicating fewer incorrect or unsupported outputs, making it more reliable for factual accuracy.

- **ROUGE Scores**: `llama3.3` outperforms `llama_d3` in ROUGE-1 (0.2075 vs. 0.1631), ROUGE-2 (0.1427 vs. 0.1188), and ROUGE-L (0.2075 vs. 0.1631), showing greater lexical similarity to the ground truth.
- **METEOR**: `llama3.3` (0.6894) outperforms `llama_d3` (0.6133), indicating better semantic alignment with the ground truth.
- **Cosine Similarity (`gt_to_pred`)**: `llama3.3` (0.8983) outperforms `llama_d3` (0.8440), confirming its stronger alignment with the ground truth.
- **Cosine Similarity (`context_to_pred`)**: `llama_d3` (0.7029) slightly outperforms `llama3.3` (0.7014), but the difference is minimal, suggesting comparable context relevance.

5 Visualization

A bar chart comparing the metrics was generated separately using Chart.js. The chart visualizes the performance differences across all metrics, with `llama3.3` showing higher scores in most categories except hallucination score and context-to-prediction cosine similarity.

6 Conclusion

`llama3.3` is the better model for most use cases, excelling in ROUGE, METEOR, and ground truth cosine similarity, which are critical for tasks requiring high fidelity to reference answers (e.g., question-answering or summarization). However, **`llama_d3`** is superior in minimizing hallucinations, making it preferable for applications where factual accuracy is paramount (e.g., medical or legal text generation). The choice depends on the specific use case: prioritize **`llama3.3`** for general accuracy and semantic alignment, or **`llama_d3`** for minimal hallucinations.

7 Recommendations

- Use **`llama3.3`** for tasks prioritizing lexical and semantic similarity to ground truth.
- Use **`llama_d3`** for tasks where minimizing hallucinations is critical.
- Consider ensemble methods or fine-tuning to combine the strengths of both models if both accuracy and low hallucination are priorities.