

Multi-Model Benchmark Report

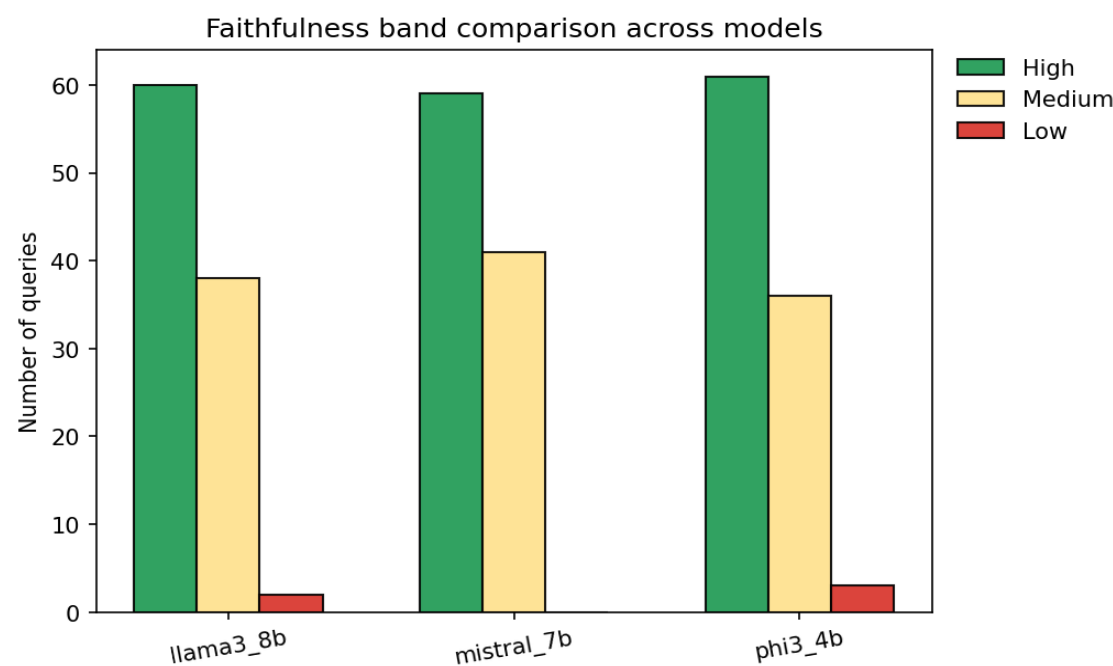
Generated: 2025-11-16 21:36

This report compares multiple LLM profiles in terms of retrieval relevance (NDCG@k) and factual grounding (faithfulness). Thresholds for band classification are applied internally but deliberately hidden in the visualization.

1. Summary Statistics per Model

| Model | Mean NDCG | Mean Faith | Median NDCG | Median Faith | Std NDCG | Std Faith |
|------------|-----------|------------|-------------|--------------|----------|-----------|
| llama3_8b | 1.000 | 0.585 | 1.000 | 0.629 | 0.000 | 0.188 |
| mistral_7b | 1.000 | 0.569 | 1.000 | 0.583 | 0.000 | 0.187 |
| phi3_4b | 1.000 | 0.561 | 1.000 | 0.607 | 0.000 | 0.183 |

2. Faithfulness Band Comparison



End of Report