

# Multi-Model Benchmark Report

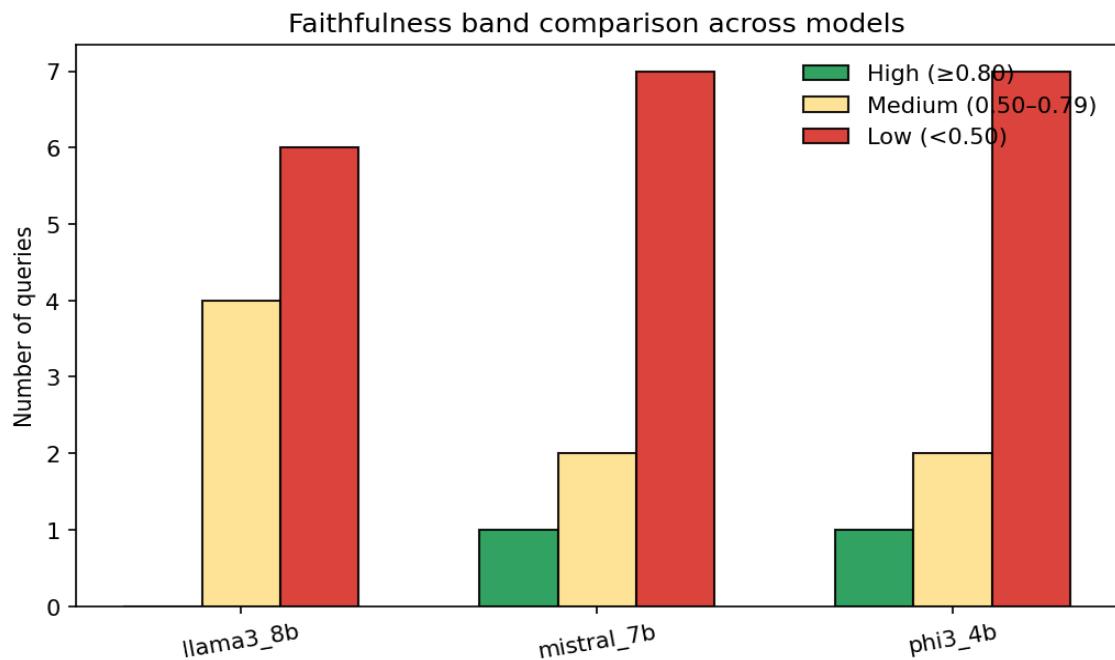
Generated: 2025-11-15 23:21

This report compares multiple LLM profiles in terms of retrieval relevance (NDCG@k) and factual grounding (Faithfulness). All evaluations were performed using an identical prompt set, identical retrieval stack, and identical parameters.

## 1. Summary Statistics per Model

Model	Mean NDCG	Mean Faith	Median NDCG	Median Faith	Std NDCG	Std Faith
llama3_8b	1.000	0.472	1.000	0.410	0.001	0.155
mistral_7b	1.000	0.491	1.000	0.428	0.001	0.216
phi3_4b	1.000	0.503	1.000	0.433	0.001	0.210

## 2. Faithfulness Band Comparison



End of Report