

# Multi-Model Benchmark Report

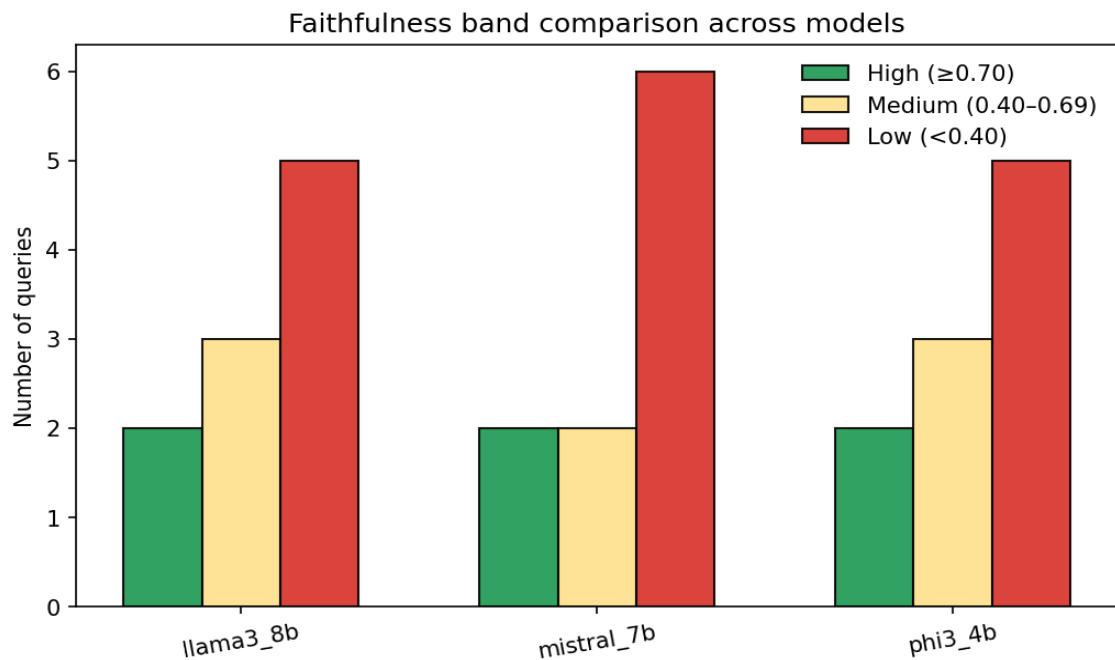
Generated: 2025-11-16 00:58

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	0.502	1.000	0.426	0.001	0.168
mistral_7b	1.000	0.457	1.000	0.386	0.001	0.181
phi3_4b	1.000	0.478	1.000	0.409	0.001	0.156

## 2. Faithfulness Band Comparison



End of Report