Multi-Model Benchmark Report

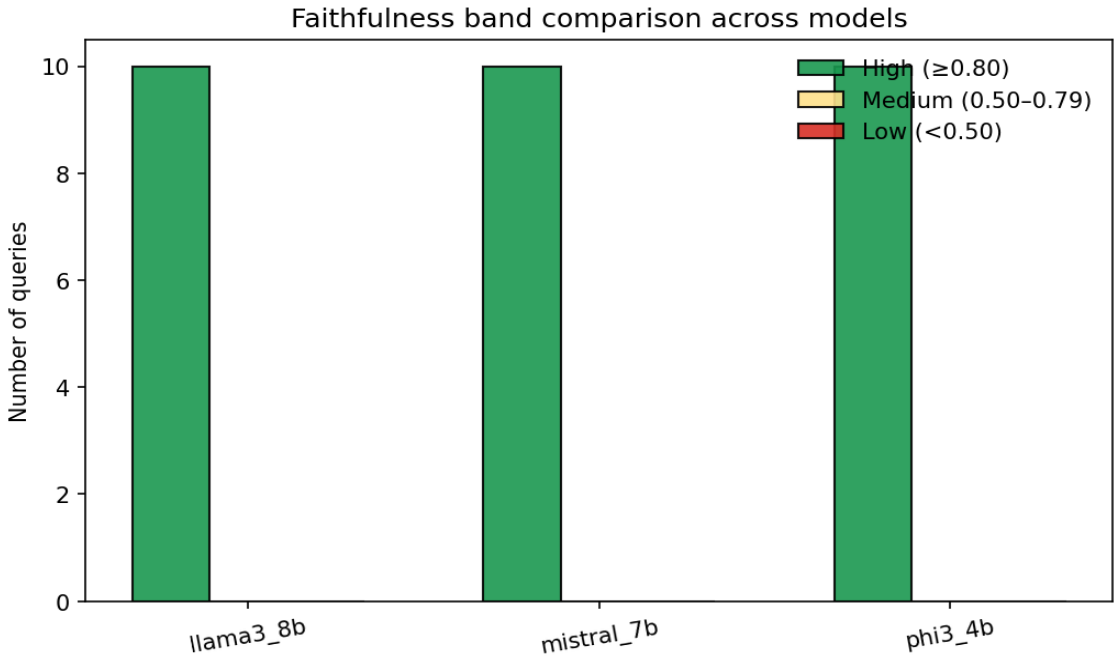
Generated: 2025-11-15 22:48

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	0.992	1.000	1.000	0.001	0.017
mistral_7b	1.000	0.991	1.000	1.000	0.001	0.018
phi3_4b	1.000	0.996	1.000	1.000	0.001	0.013

2. Faithfulness Band Comparison



End of Report