Multi-Model Benchmark Report

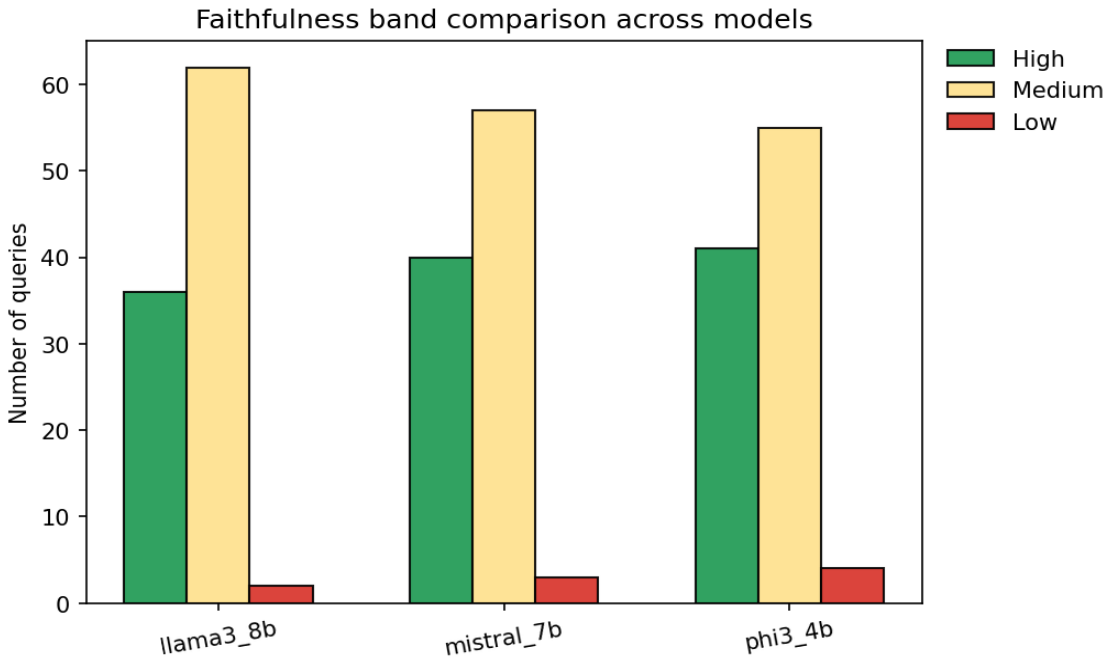
Generated: 2025-11-18 14:39

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	0.998	0.461	1.000	0.452	0.011	0.134
mistral_7b	0.998	0.466	1.000	0.429	0.011	0.146
phi3_4b	0.998	0.469	1.000	0.460	0.011	0.139

2. Faithfulness Band Comparison



End of Report