

Evaluation Metrics

Retrieval Metrics

Precision@K

relevant in top K / K

Recall@K

relevant in top K / total relevant

NDCG

Normalized Discounted Cumulative Gain

MRR

Mean Reciprocal Rank

Generation Metrics

ROUGE-L

Longest common subsequence

BLEU

N-gram overlap with reference

BERTScore

Semantic similarity

Medical-Specific

Clinical Relevance

Expert judgment (1-5 scale)

Safety Score

Harm potential assessment

Citation Accuracy

Correct source attribution %

Recommended Thresholds

Precision@10: > **90%**

NDCG@10: > **0.85**

Clinical Relevance: > **4.0/5.0**

Safety Score: **100%**