

Manual Evaluation Report: RAG vs Base Model

Evaluation Criteria

1. Relevance - Does the answer directly address the question?
2. Factual Accuracy - Are the statements medically accurate?
3. Completeness - Is the answer thorough and informative?
4. Clarity - Is the response well-structured and easy to understand?
5. Fluency - Is the grammar and language usage natural and correct?
6. Hallucinations - Are there false or nonsensical claims? (1 = severe, 5 = none)

Evaluation Summary Table

| Question | Model | Rel | Acc | Comp | Clar | Flu | Halluc. |
|---------------------------|------------|-----|-----|------|------|-----|---------|
| What is Knee Replacement? | RAG | 5 | 5 | 5 | 5 | 5 | 5 |
| | Base Model | 1 | 1 | 1 | 2 | 3 | 1 |
| What is Glaucoma? | RAG | 5 | 5 | 5 | 5 | 5 | 5 |
| | Base Model | 1 | 1 | 1 | 2 | 2 | 1 |
| What is Anxiety Disorder? | RAG | 5 | 5 | 4 | 5 | 5 | 5 |
| | Base Model | 3 | 2 | 1 | 3 | 3 | 2 |

Average Scores

- RAG Model: 29.6 / 30
- Base Model: 11.2 / 30

Conclusion

Yes RAG Model provides factually accurate, clear, and complete responses tailored to the query, with no hallucinations.

No Base Model often provides incomplete, irrelevant, or incorrect answers, with frequent hallucinations and poor structure.