Medical RAG ChatBot Evaluation Report

Emre Büyükyılmaz

September 18, 2025

1 Executive Summary

This report presents a comprehensive performance evaluation of the Medical RAG system, based on a 100-question test set. The analysis combines automated metrics with human judgments to assess answer quality, latency, and system behavior. All figures are generated from the analysis notebook (`src/utils/metrics_Analysis.ipynb`).

Methodological Note: For this analysis, 47 responses where the model abstained (i.e., the `generated_answer` contained the phrase ”I do not have sufficient information”) were excluded. This filtering ensures that the evaluation focuses on the quality of substantive answers, preventing abstentions from skewing the results.

Key Findings:

- **High-Quality Human Ratings:** The system achieves near-perfect scores for fluency and high relevance, as rated by human evaluator. Factual accuracy is more variable, indicating a dependency on the quality of retrieved documents and the complexity of the query.
- **Semantic vs. Lexical Metrics:** Traditional n-gram-based metrics (BLEU, ROUGE) are low, reflecting the system's abstractive nature and its tendency to paraphrase rather than extract. In contrast, semantic similarity metrics like BERTScore show strong alignment with reference answers, confirming contextual correctness.
- **Latency Profile:** System latency is overwhelmingly dominated by the token generation phase. Retrieval is exceptionally fast. As expected, total response time correlates directly with the length of the generated answer.

2 Metrics Under Evaluation

The analysis incorporates a diverse set of metrics to capture different aspects of system performance:

- **Human Evaluation:** `Accuracy`, `Relevance`, `Fluency`, `Source_Citation` (1-5 Likert scale)
- **Latency (ms):** `Retrieval_Time_MS`, `Generation_Time_MS`, `Total_Time_MS`
- **Automated Quality:** `BLEU`, `ROUGE-1/2/L`, `METEOR`, `BERTScore-P/R/F1`, `Perplexity`
- **Input/Output Features:** `Query_Length`, `answer_len_words`, `answer_len_chars`

3 Summary Statistics

Metric	Mean	Notes
Retrieval_Time_MS	*13.183*	Consistently fast and stable.
Generation_Time_MS	*3090.362*	The primary contributor to latency; varies with answer length.
Total_Time_MS	*3111.684*	Distribution is right-skewed due to longer answers.
BLEU	*0.135*	Low score suggests significant lexical divergence from references.
ROUGE-1	*0.436*	Moderate unigram overlap.
ROUGE-2	*0.258*	Low bigram overlap, indicating paraphrasing.
ROUGE-L	*0.358*	Moderate longest-common-subsequence overlap.
METEOR	*0.438*	Moderate alignment, sensitive to synonyms and stemming.
BERTScore-F1	*0.771*	Strong semantic alignment with reference answers.
Perplexity	*38.404*	Low value indicates high model confidence and fluency.
Accuracy (1–5)	*3.811*	Good on average, but with notable variance.
Relevance (1–5)	*4.491*	Consistently high; answers address the query.
Fluency (1–5)	*5.000*	Perfect score; answers are grammatically correct.

4 Human Evaluation of Answer Quality

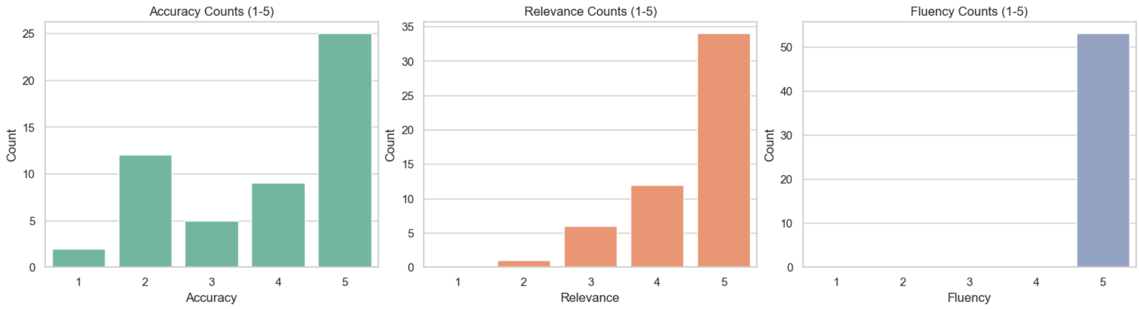


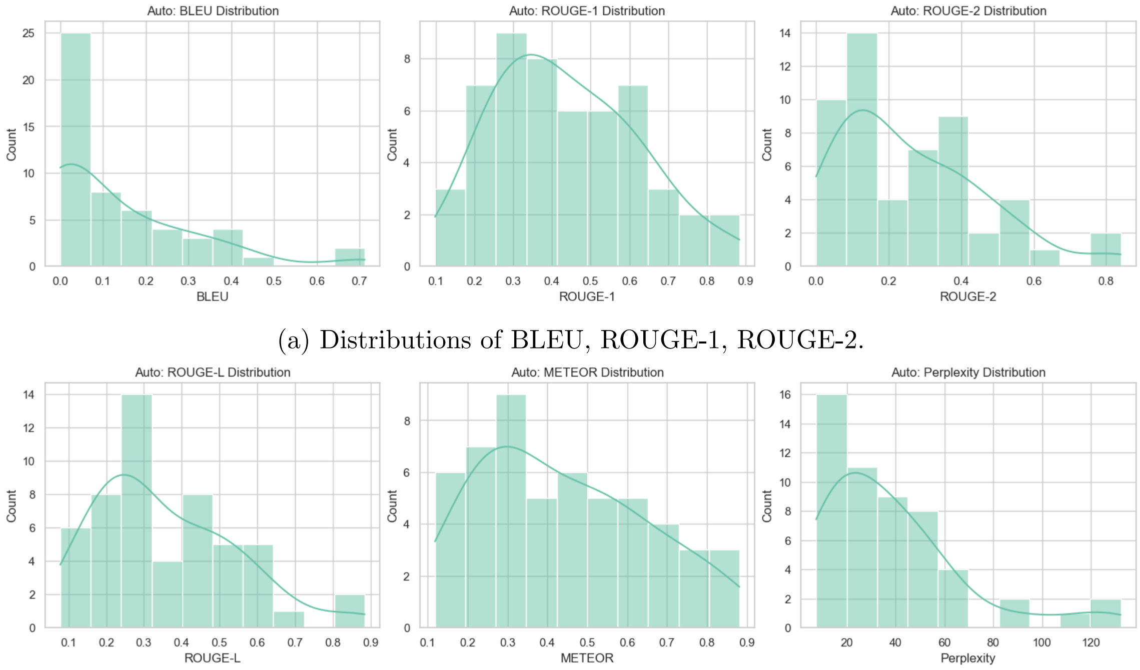
Figure 1: Distribution of human ratings for Accuracy, Relevance, and Fluency (1-5 scale).

Analysis: Human evaluation highlights the system's strengths in language generation and its challenges in maintaining factual consistency.

- **Fluency and Relevance:** The system achieves perfect and near-perfect scores for Fluency and Relevance. This indicates that the generated responses are consistently well-formed, coherent, and directly address the user's question.
- **Accuracy:** Factual Accuracy shows a bimodal-like distribution, with a large cluster of perfect scores (5) alongside a significant number of less accurate responses (scores of 2 and 3). This variance suggests that while the system is often correct, its accuracy is highly dependent on the quality of the retrieved context and the inherent difficulty of the question. Failures in the retrieval step are the likely cause of lower accuracy scores.

5 Automated Metrics Analysis

5.1 N-gram Overlap Metrics (BLEU, ROUGE, METEOR)



(a) Distributions of BLEU, ROUGE-1, ROUGE-2.

(b) Distributions of ROUGE-L, METEOR, Perplexity.

Figure 2: Distributions of n-gram-based metrics and perplexity.

Analysis: The low scores across all n-gram-based metrics (BLEU, ROUGE, METEOR) are characteristic of abstractive, generative systems. These metrics penalize lexical and syntactic variations, even when semantic meaning is preserved. The results confirm that the RAG system generates novel, paraphrased responses rather than simply extracting text. While useful for tracking lexical overlap, these metrics are insufficient for capturing the semantic quality of the system's output. The low Perplexity aligns with the perfect human Fluency scores, indicating the model is highly confident in its generations.

5.2 Semantic Similarity (BERTScore)

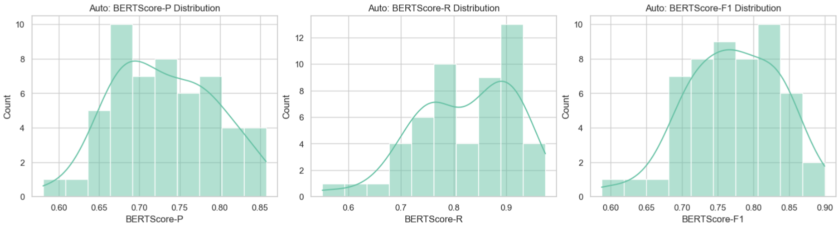


Figure 3: Distributions of BERTScore Precision (P), Recall (R), and F1-score.

Analysis: BERTScore, which measures semantic similarity, provides a more meaningful assessment of quality. The high Recall (R) indicates that the generated answers successfully capture most of the semantic content from the reference answers. The slightly lower Precision (P) suggests the model may include additional, correct information not present in the reference. The strong F1 score (mean 0.77) confirms a high degree of semantic alignment, making it a reliable automated metric for this task.

5.3 Correlation of Automated vs. Human Metrics

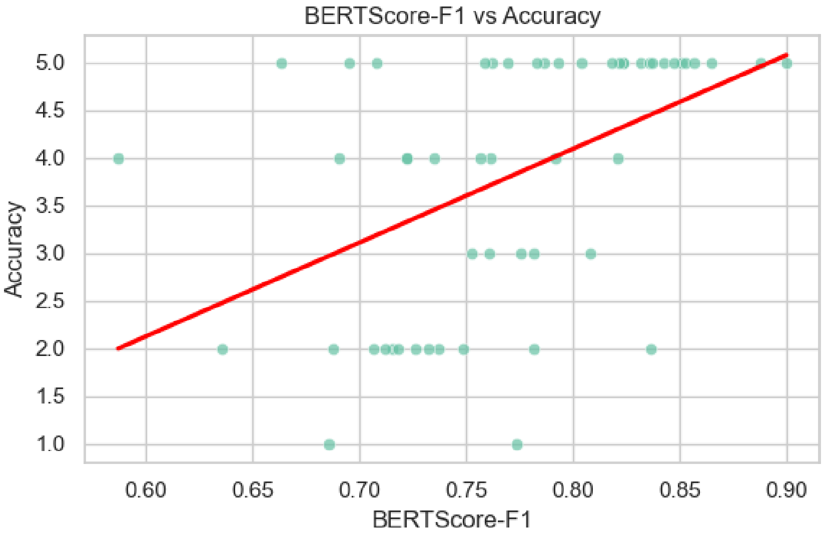


Figure 4: Relationship between BERTScore-F1 and human-rated Accuracy.

Analysis: A clear positive correlation is observed between human-rated Accuracy and BERTScore-F1. This alignment is crucial, as it validates BERTScore as a reliable proxy for human judgment of semantic correctness. Higher BERTScore values are strongly associated with higher accuracy ratings, reinforcing its utility for automated monitoring and regression testing of the system.

6 Latency and Performance Analysis

6.1 Total Response Time Distribution

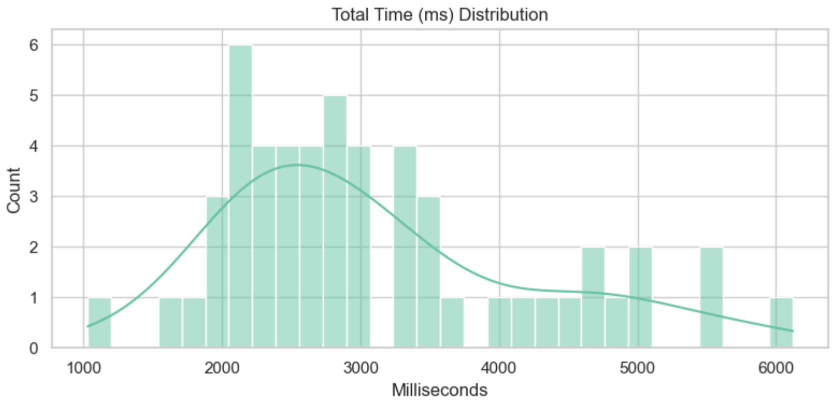


Figure 5: Distribution of Total_Time_MS across all queries.

Analysis: The system's total response time averages 3.1 seconds, with a right-skewed distribution indicating a tail of longer response times for more complex generations. The overall latency is overwhelmingly dominated by the generation step (mean ˜3.09s), while the retrieval step is extremely efficient (mean ˜13ms).

6.2 Answer Length vs. Time and Accuracy

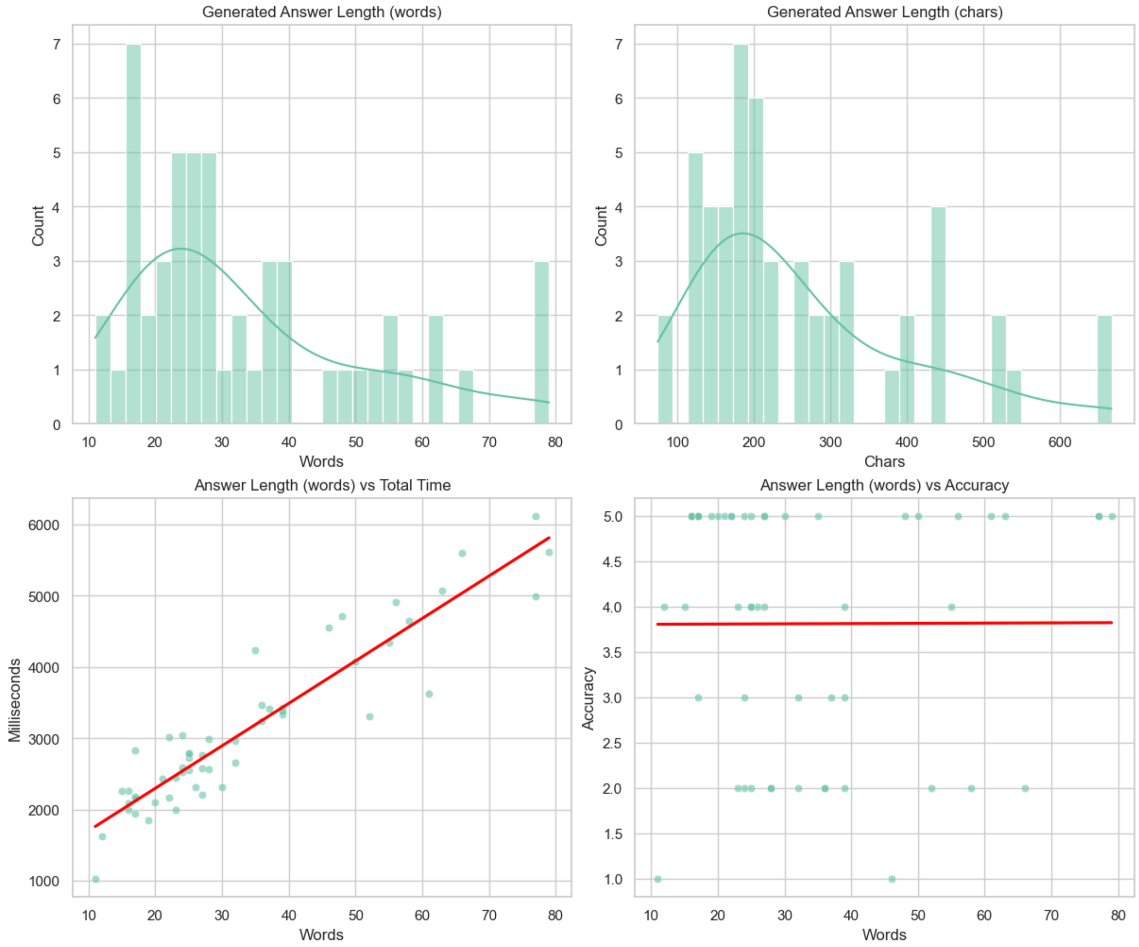


Figure 6: Top: Distributions of answer length in words and characters. Bottom: Relationships between answer length (words) and Total Time (left) and Accuracy (right).

Analysis: A strong, positive linear relationship exists between the length of the generated answer and the total response time, as expected. However, there is no discernible correlation between answer length and factual accuracy. This suggests that verbosity is not an indicator of quality; longer answers are not inherently more or less accurate than shorter ones.

6.3 Query Length vs. Total Time

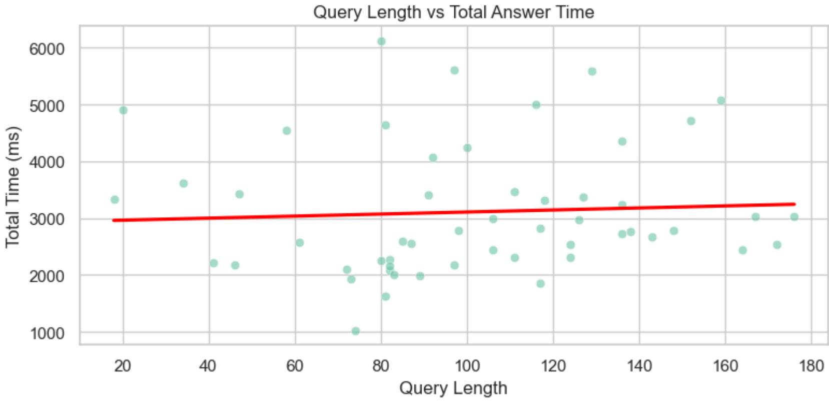


Figure 7: Relationship between Query_Length (characters) and Total_Time_MS.

Analysis: A weak positive correlation exists between input query length and total response time. The effect is minimal because the primary driver of latency is the generation of the answer, which depends more on the complexity of the topic and the retrieved context length than on the length of the initial query.

7 Overall Metric Correlations

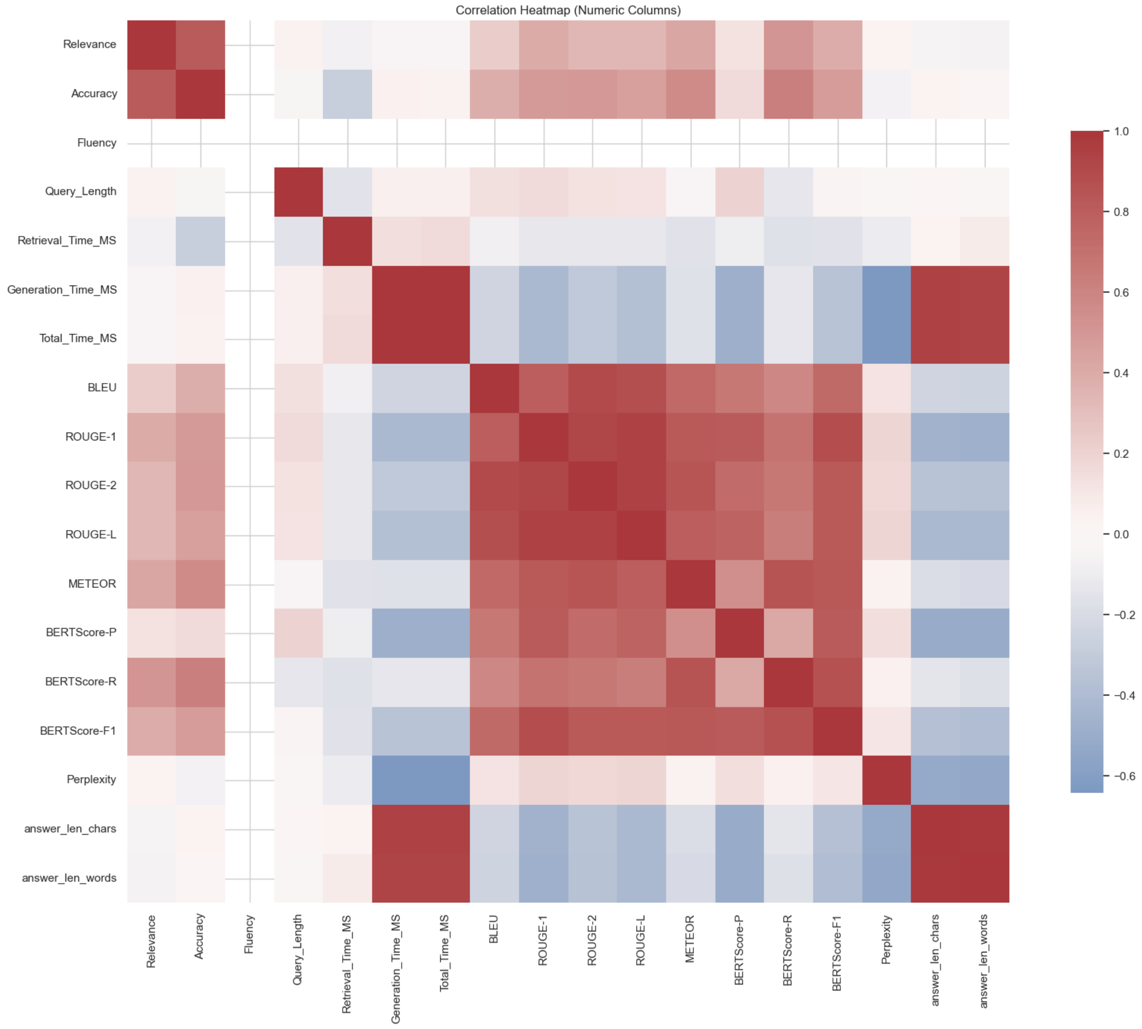


Figure 8: Correlation heatmap of all numeric metrics.

Analysis: The correlation heatmap provides a consolidated view of the inter-relationships between all numeric metrics, revealing several key patterns in the system's behavior and the evaluation framework itself.

- **Distinct Metric Clusters:** Two strong, internally-correlated clusters are evident. The first consists of n-gram-based metrics (`BLEU`, `ROUGE-1/2/L`, `METEOR`), confirming they measure the similar phenomenon of lexical overlap.
- **Latency and Length Correlation:** A very strong positive correlation exists between answer length (`answer_len_words` and `_chars`) and both `Generation_Time_MS` and `Total_Time_MS`. This confirms that output verbosity is the primary driver of system latency.
- **Relationship Between Semantic and Lexical Metrics:** A significant moderate positive correlation exists between `BERTScore-F1` and the n-gram overlap metrics (especially `ROUGE-1` and `ROUGE-L`). This is an important finding: it indicates that while the system is highly abstractive, answers that are semantically correct still tend to share essential keywords and phrases with the reference. However, the fact that this correlation is moderate—not perfect—is crucial. It demonstrates that the system can achieve high semantic similarity without high lexical overlap, confirming its ability to paraphrase and synthesize

information effectively. This reinforces the conclusion that semantic metrics are more suitable for evaluating this system's core capabilities.

- **Validation of Automated Metrics:** Human-rated `Accuracy` shows a moderate positive correlation with `BERTScore-F1`, validating it as a reliable automated proxy for semantic correctness. A similar, though slightly weaker, positive correlation exists between `Accuracy` and the `ROUGE` scores, suggesting that some degree of lexical overlap is still indicative of a correct answer in this dataset.
- **Perplexity as a Quality Indicator:** `Perplexity` exhibits a notable negative correlation with several quality and length metrics. Specifically, it is negatively correlated with `BERTScore-F1` and `answer_len_words`. This indicates that lower perplexity (higher model confidence) is associated with answers that are both longer and more semantically aligned with the reference.
- **Independent Factors:** `Query_Length` and `Retrieval_Time_MS` show negligible correlation with most other metrics, confirming they are not significant drivers of answer quality or overall generation time in this system. Similarly, `Fluency` shows no correlation due to its lack of variance (all scores were 5).