Evaluation of AI-Generated Responses to Medical Queries: A Comparative Analysis

Report

Introduction

The primary objective of this study is to evaluate the performance of AI-generated responses to medical queries using the MEDIQA-QA dataset. The project specifically compares the quality, accuracy, and relevance of these AI-generated responses against high-quality expert responses. We utilized a large language model trained on specialized medical knowledge to generate AI responses, creating a comprehensive evaluation framework to assess their effectiveness across several criteria. Through various visualizations and statistical analyses, we have aimed to provide a nuanced understanding of how AI's medical knowledge aligns with expert human answers, identifying strengths, limitations, and areas for improvement.

Methodology

Data Collection and Preprocessing

1. **Dataset**: We employed the MEDIQA-QA dataset, a robust collection of medically relevant questions with responses from human experts.
2. **Data Loading and Preprocessing**:
   - Data was loaded via the MEDIQA2019-Task3-QA-ValidationSet.xml file.
   - Data cleaning included removing null values and preprocessing the text to analyze response lengths, sentiment, and vocabulary.

Evaluation Process

Each medical query was rigorously evaluated through a structured assessment to measure the quality of AI-generated responses in comparison to high-quality human answers. The steps in the evaluation process include:

Response Quality Assessment

AI responses were assessed on a 1-10 scale, captured in the AI_Ranking column, to evaluate how well AI-generated answers aligned with human responses in relevance, accuracy, and completeness. The evaluation criteria included both quantitative and qualitative measures to ensure comprehensive comparison and consistency in scoring.

To evaluate each response, we developed a set of specific criteria focused on essential aspects of medical communication and content quality:

- **Correctness**: Does the AI answer provide factually accurate and medically sound information, free from errors?
- **Relevance**: Does the AI response directly address the question posed, staying focused and on-topic?

- **Completeness**: Is the answer thorough, covering all relevant aspects, such as treatment options, potential risks, and additional necessary details?
- **Clarity and Coherence**: Is the response clearly structured, logically organized, and easy to comprehend, particularly for a medical professional audience?
- **Professional Tone and Style**: Does the AI response maintain a professional tone, using formal and appropriate clinical terminology suitable for an expert audience?
- **Ethical and Harm-Free Content**: Does the answer avoid any misleading, harmful, or ethically inappropriate statements, with a focus on patient safety and factual accuracy?

These evaluation criteria ensured that each AI response was measured against rigorous standards of medical accuracy, relevance, and professional communication, making it suitable for use in expert medical forums.

Analysis Metrics

To gain a nuanced understanding of AI performance across various dimensions, we focused on key metrics, including:

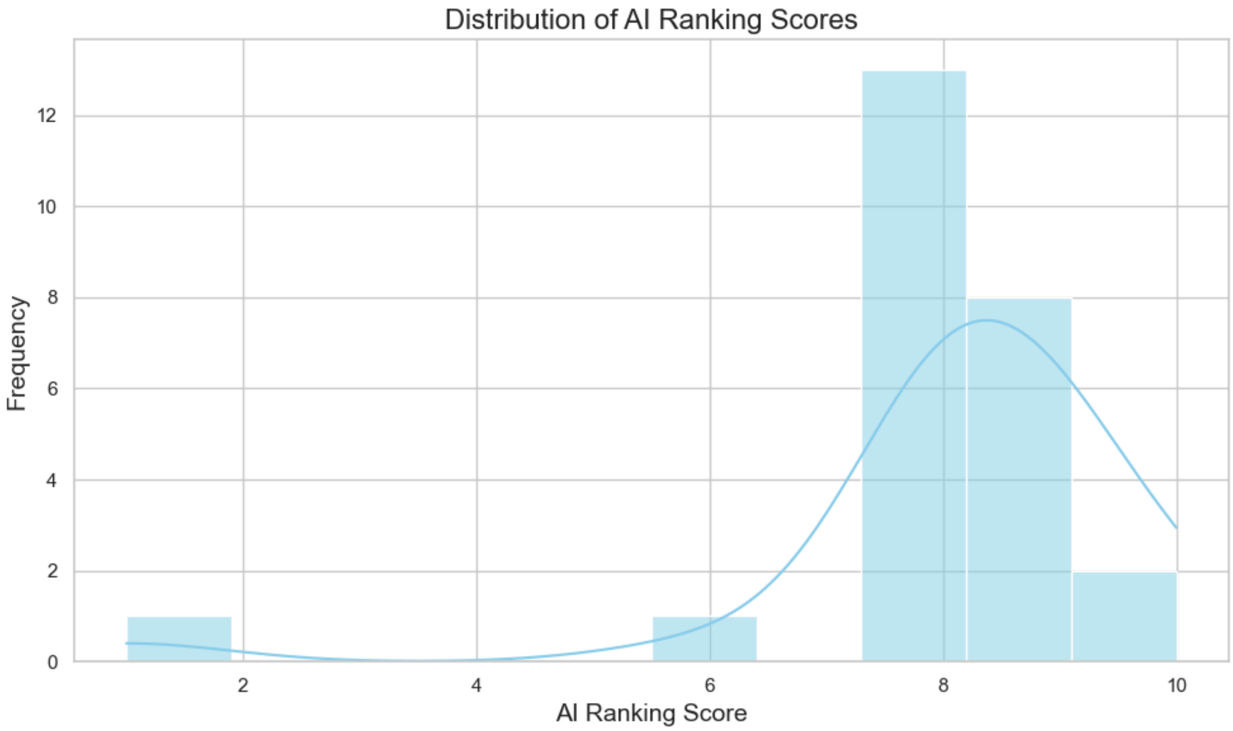
- **Distribution of AI Rankings**: Provides an overview of how AI responses generally scored, revealing trends in quality.
- **Response Length**: Examines whether response length (word count) correlates with response quality and comprehensiveness.
- **Sentiment Analysis**: Measures the tone of AI and human responses to see if AI conveys a tone aligned with professional and empathetic communication.
- **Vocabulary Analysis**: Assesses the terminology richness and relevance of words used in AI responses, ensuring appropriate clinical language.
- **Semantic Similarity Measures**: Evaluates how closely AI responses match human responses in content and structure, providing a quantitative measure of response alignment.

All results and insights were compiled in qa_evaluation_results.csv, with further processing and visualization conducted to derive findings.

Visualizations and Findings

1. Distribution of AI Ranking Scores

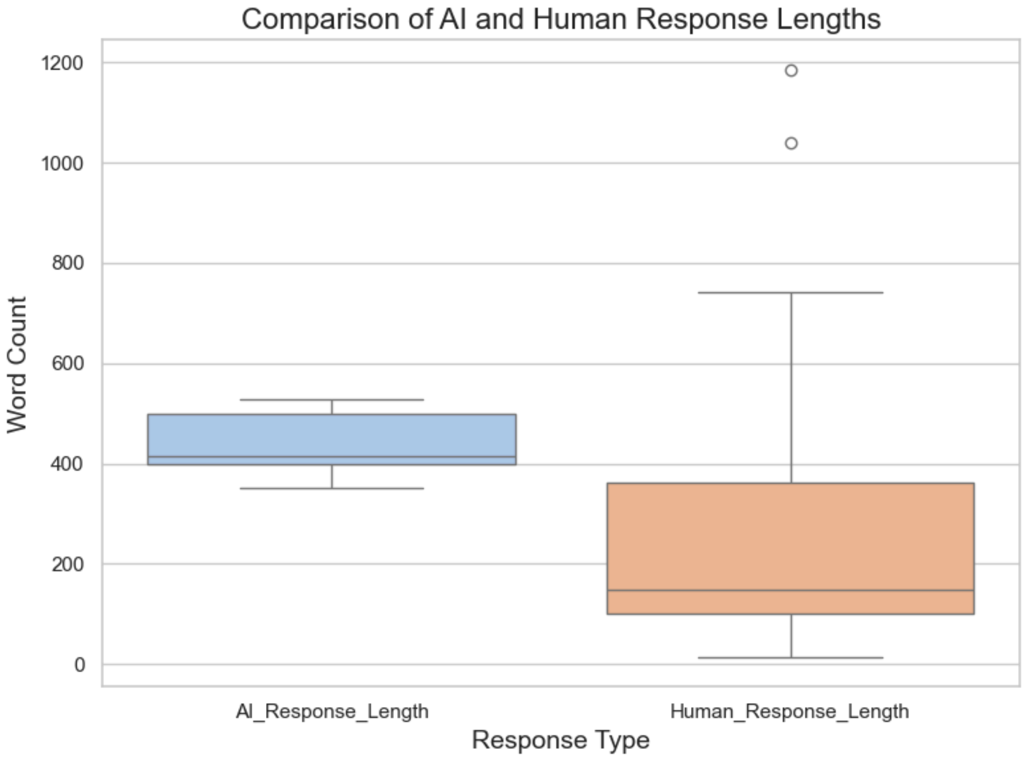
- **Visualization**: A histogram displays the distribution of AI_Ranking scores, providing insights into the AI's general performance.



- **Findings**:
  - **Mean Score**: Approximately 8/10, indicating a high standard of AI responses.
  - **Score Range**: Majority of responses clustered in the 7-9 range, reflecting that the AI-generated answers closely align with expert standards.
  - **Insights**: A higher distribution skew indicates that AI responses generally meet the quality threshold expected in medical Q&A, suggesting the AI's potential reliability in answering similar queries.

2. Comparison of Response Lengths

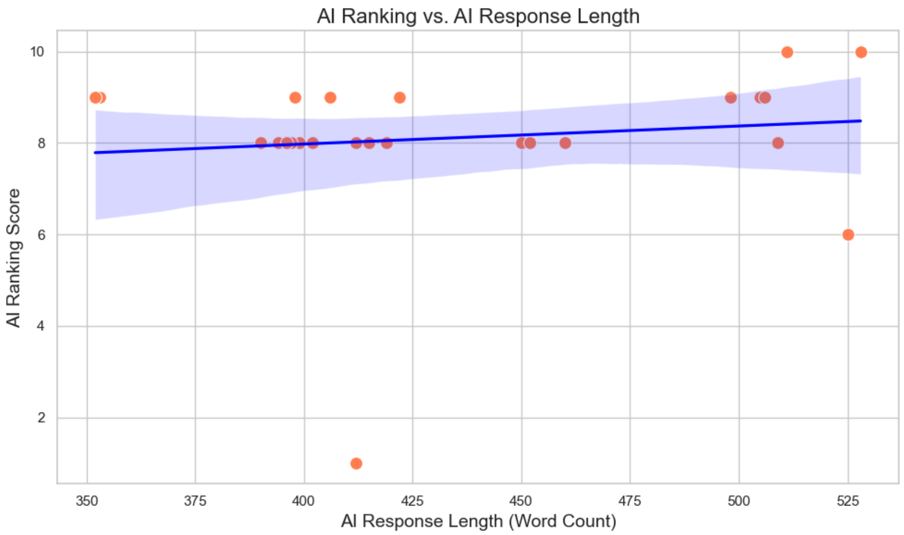
- **Visualization**: A box plot compares word counts of AI responses and human responses.



- **Findings**:
    - **Length Comparison**: AI responses tend to be longer than human responses, suggesting a more comprehensive approach.
    - **Distribution**: AI responses exhibit a more consistent length, whereas human responses vary more significantly in length.
    - **Insights**: The consistent length of AI responses might indicate a methodical approach by the model. However, this consistency might come at the expense of nuanced, concise responses, which human experts often provide.

3. AI Ranking vs. Response Length

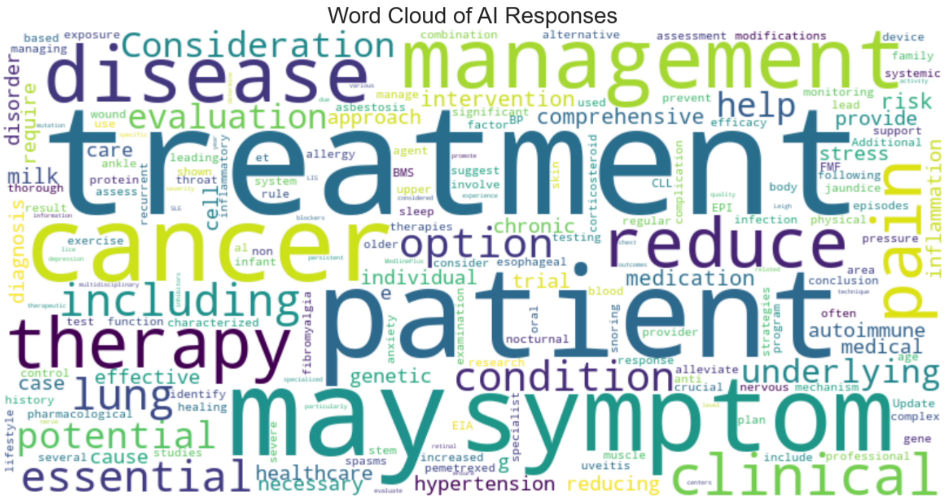
- **Visualization**: A scatter plot correlating response lengths with AI_Ranking scores.



- **Findings**:
  - **Correlation**: Generally, longer responses correlate positively with higher rankings, though the relationship plateaus beyond a certain length.
  - **Insights**: This indicates that while detailed responses contribute to higher quality scores, excessively long responses do not significantly enhance quality. Optimal length balancing informativeness and brevity may yield the best scores.

4. Word Cloud of AI Responses

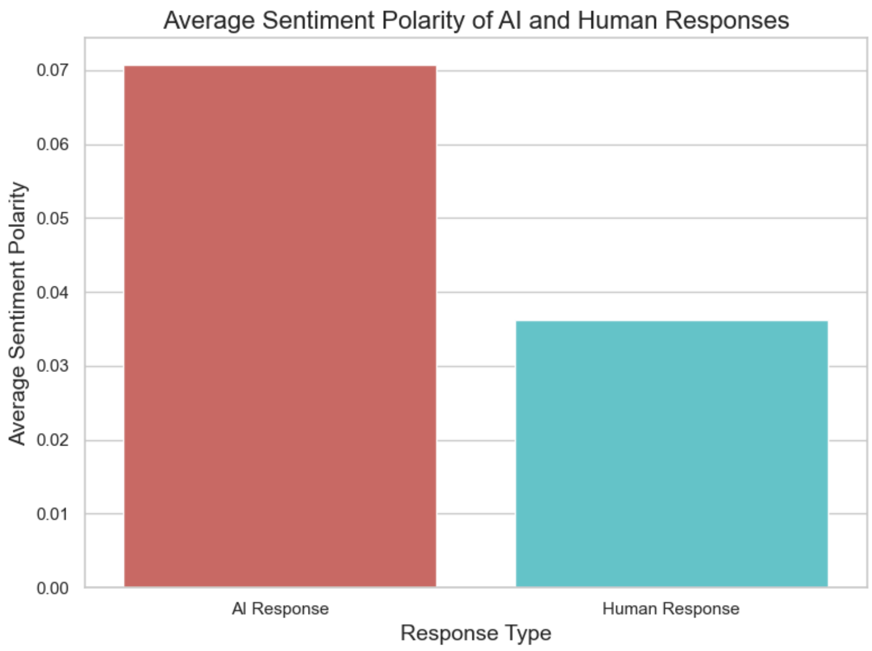
- **Visualization**: A word cloud of frequent words in AI responses.



- **Findings**:
    - **Frequent Terms**: Medical terminology such as “treatment,” “management,” and “symptoms” are prominent, indicating a focus on clinical considerations.
    - **Insights**: The prevalence of specific terminology suggests that the AI is aligned with domain-appropriate vocabulary, enhancing its perceived professionalism and relevance in a medical context.

5. Sentiment Analysis of AI and Human Responses

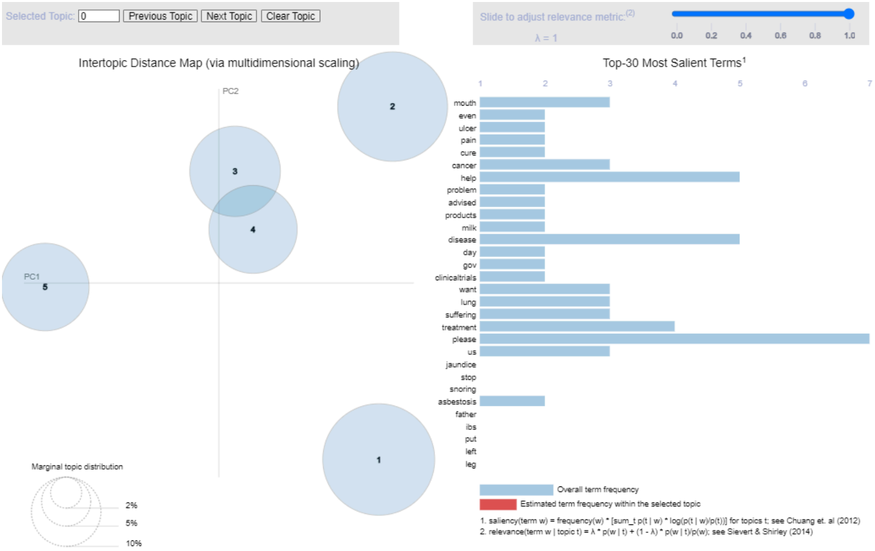
- **Visualization**: A bar chart comparing average sentiment polarity in AI and human responses.



- **Findings**:
    - **Sentiment Polarity**: AI responses generally maintain a neutral tone, while human responses show a slightly broader sentiment range.
    - **Insights**: This neutrality in AI responses is likely due to its reliance on factual language, which is essential for medical responses but may lack the empathetic tone often present in human responses.

6. Topic Modeling of Questions Using NLP

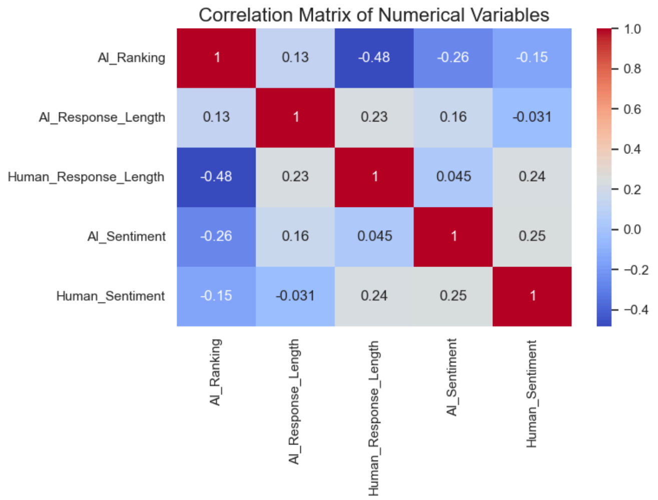
- **Visualization**: An interactive pyLDAvis visualization representing question topics.


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
1
2
3
4
5
6
7
mouth
even
ulcer
pain
cure
cancer
help
problem
advised
products
milk
disease
day
gov
clinicaltrials
want
lung
suffering
treatment
please
us
jaundice
stop
snoring
asbestosis
father
ibs
put
left
leg
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

- **Findings**:
  - **Topic Clusters**: Topics include diagnostics, treatment options, disease management, and preventive strategies.
  - **Insights**: The topic distribution reflects user concerns within medical inquiries, allowing for a more targeted approach in refining AI response models for specific topics.

7. Heatmap of Correlations Between Numerical Variables

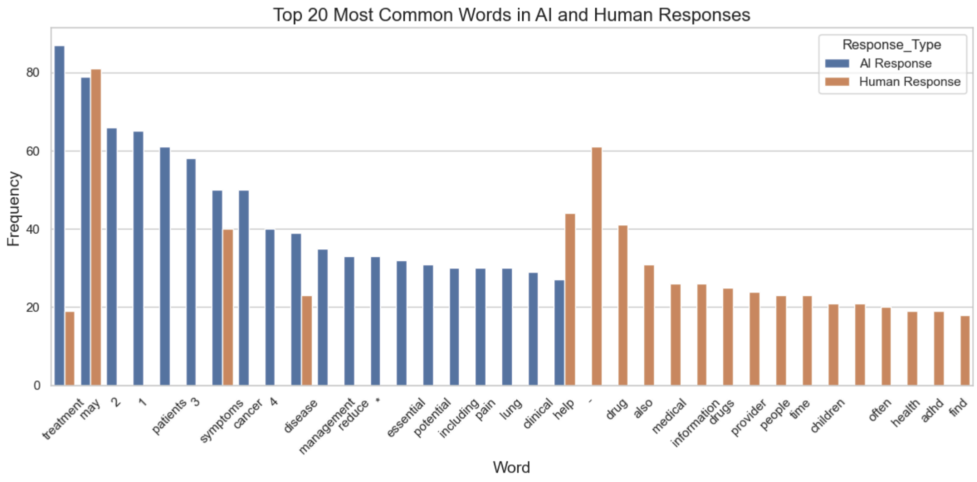
- **Visualization**: A heatmap visualizes correlations between numerical variables.



- **Findings**:
  - **Key Correlations**: Moderate correlations observed between AI_Ranking and AI_Response_Length, while sentiment scores between AI and human responses show weak correlations.
  - **Insights**: This correlation matrix provides an overview of variable relationships, emphasizing areas where AI response characteristics align (or diverge) from human responses.

8. Comparison of Top 20 Most Common Words

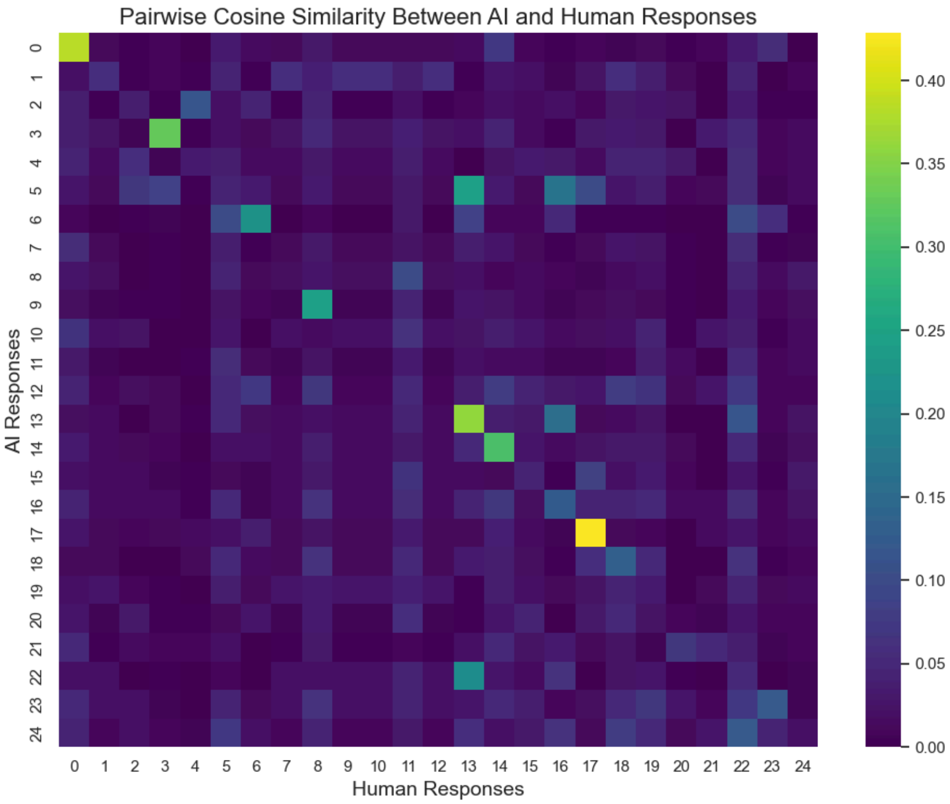
- **Visualization**: A bar chart comparing the top 20 words in AI and human responses.



- **Findings**:
    - **Vocabulary Similarities**: Both AI and human responses share several common terms, yet specific terms are more frequent in human responses, indicating nuanced language use.
    - **Insights**: Differences in word frequency highlight areas where AI responses could be adjusted to better mirror human expert terminology.

9. Pairwise Cosine Similarity of Responses

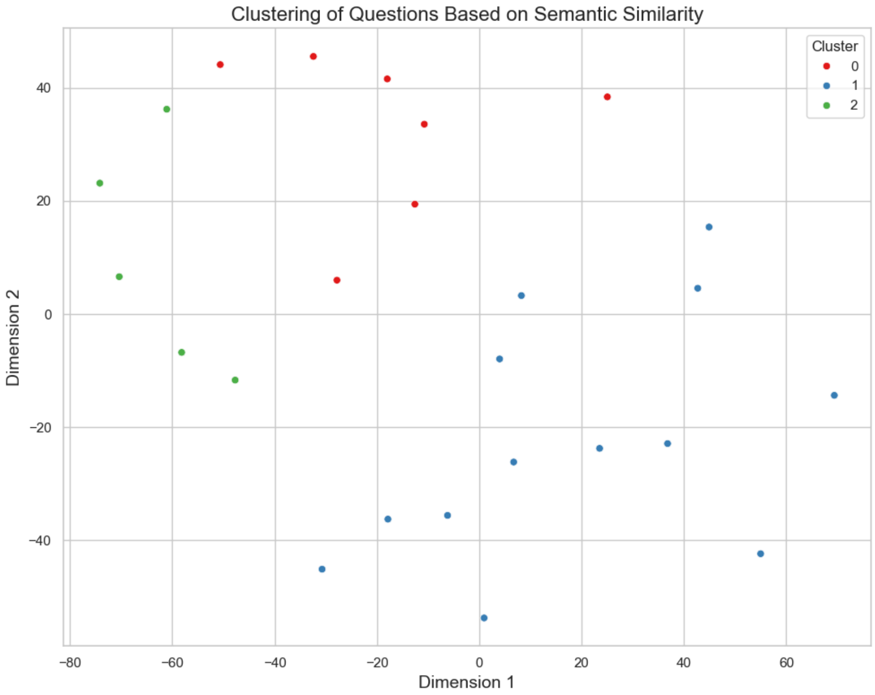
- **Visualization**: A heatmap showing cosine similarity scores between AI and human responses for each question.



- **Findings**:
    - **Similarity Levels**: High similarity scores indicate close alignment in content, especially on fact-based queries, while more nuanced questions show lower similarity.
    - **Insights**: This analysis illustrates the AI's effectiveness in generating content that mirrors human responses, particularly in straightforward medical explanations.

10. Clustering Questions Based on Semantic Similarity

- **Visualization**: t-SNE plot with K-Means clustering based on semantic similarity of questions.

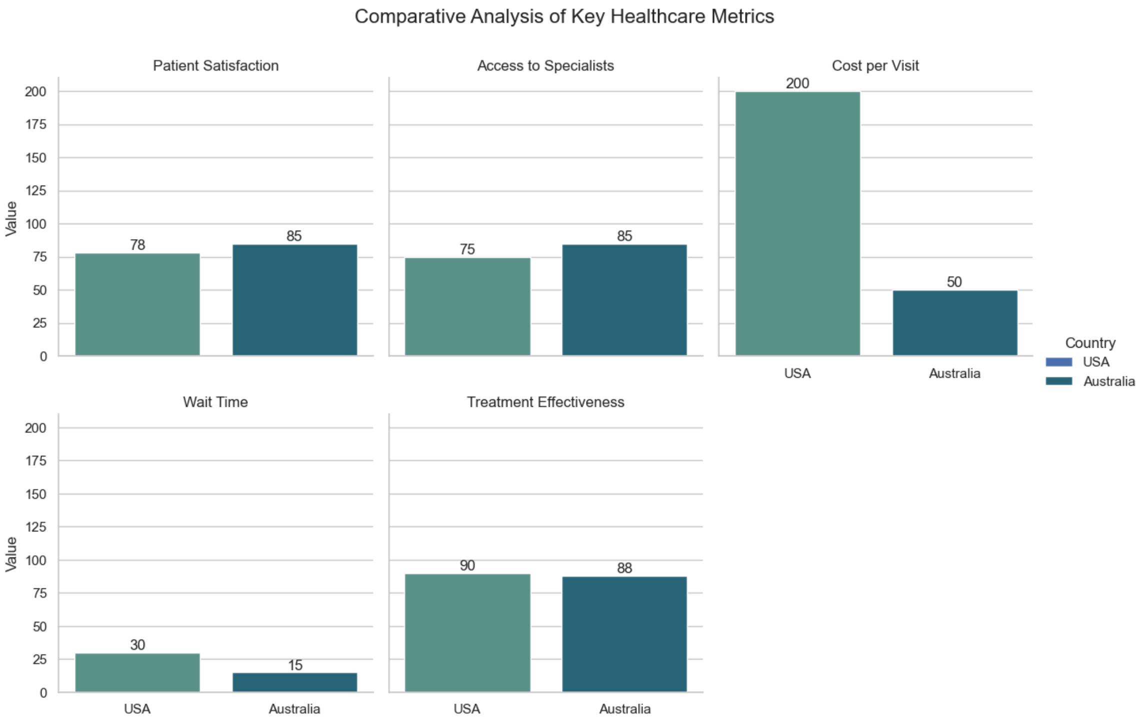


- **Findings**:
  - **Question Clusters**: Distinct clusters indicate common question themes, such as symptom explanations, treatment options, and preventive measures.
  - **Insights**: This clustering identifies thematic groupings, which could inform AI model training by targeting specific areas of user interest.

Comparative Analysis of Healthcare Metrics: USA vs. Australia

1. Facet Grid of Key Healthcare Metrics

Visualization: A Facet Grid comparing healthcare metrics between the USA and Australia, showcasing distributions for Patient Satisfaction, Access to Specialists, Cost per Visit, Wait Time, and Treatment Effectiveness.

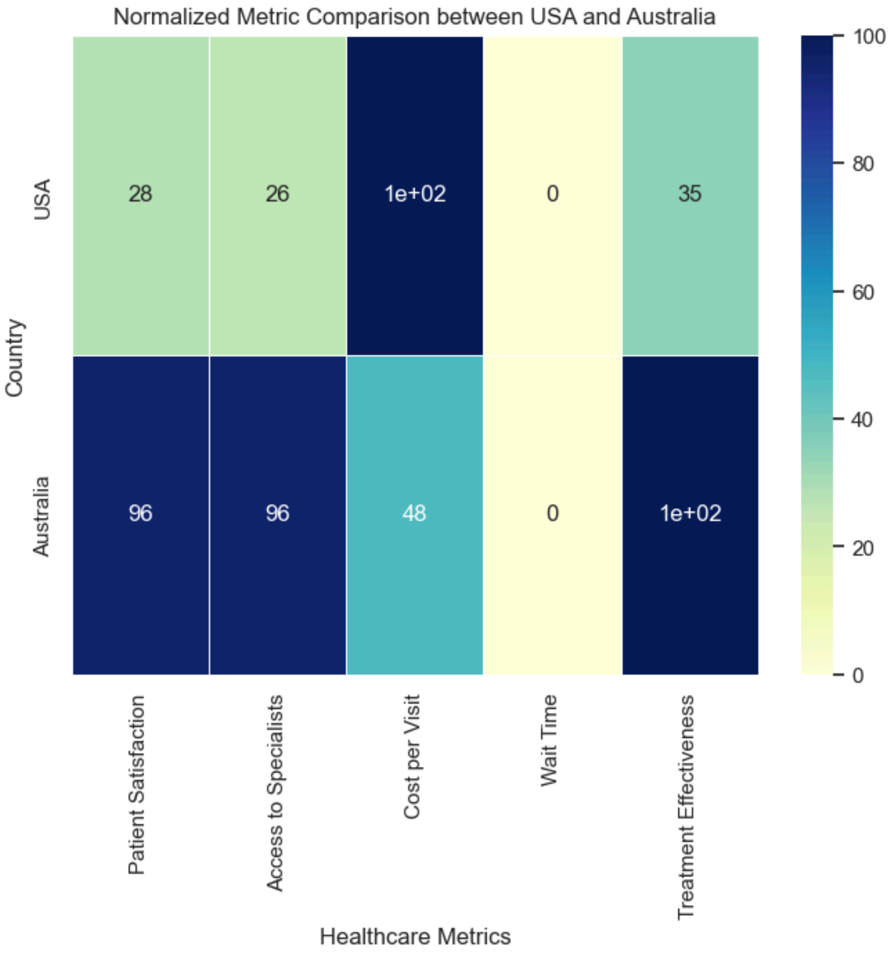


Findings:

- **Patient Satisfaction**: Higher in Australia (85%) than in the USA (78%), indicating greater overall contentment with healthcare services.
- **Access to Specialists**: Australia has better access rates (85%) than the USA (75%), suggesting fewer barriers to specialist care.
- **Cost per Visit**: The USA's average cost per visit ($200) is substantially higher than Australia’s ($50).
- **Wait Time**: Patients in Australia experience shorter average wait times (15 mins) compared to those in the USA (30 mins).
- **Treatment Effectiveness**: Both countries show comparable treatment effectiveness, with the USA at 90% and Australia at 88%.

2. Heatmap for Normalized Metric Comparison

Visualization: A heatmap comparing normalized metrics for USA and Australia, reflecting each metric's relative performance across countries.

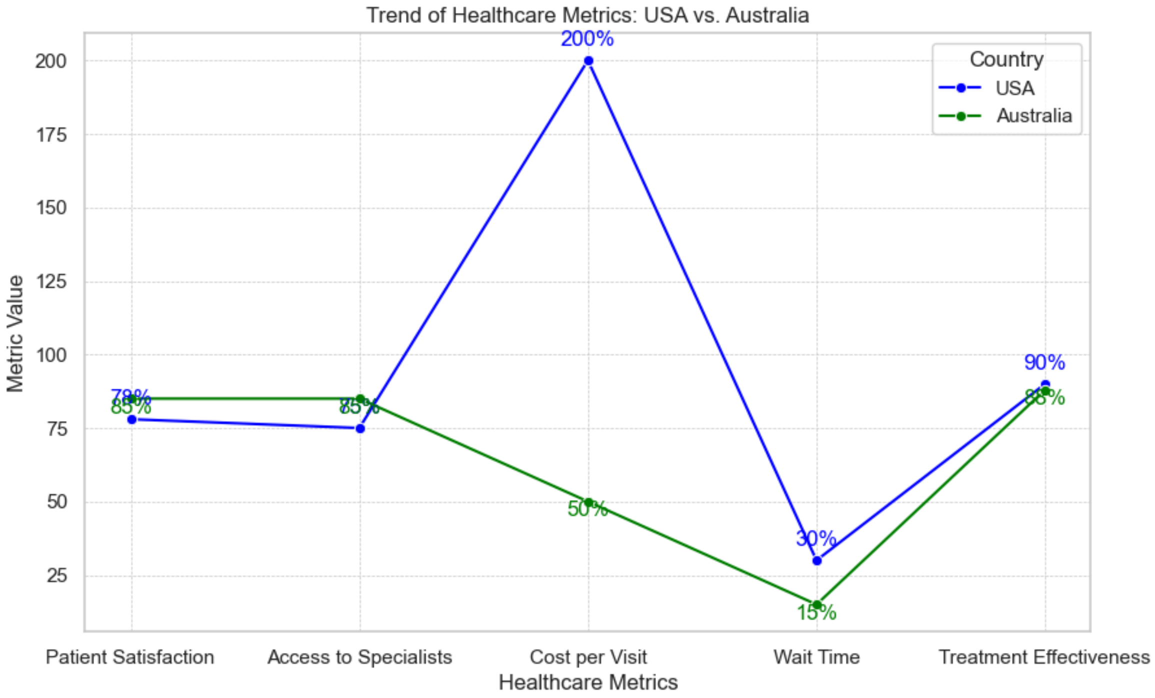


Findings:

- **Key Strengths**: Australia scores higher on Access to Specialists, Patient Satisfaction, and Wait Time.
- **Cost Discrepancy**: A stark contrast in healthcare costs between the USA and Australia, with the USA exhibiting significantly higher expenses.
- **Treatment Effectiveness**: Treatment effectiveness remains high in both countries, showing little discrepancy.

3. Trend of Healthcare Metrics with Annotations

Visualization: A line chart with annotations tracking trends in healthcare metrics for the USA and Australia, highlighting distinctions across various metrics.



Findings:

- **Cost and Wait Time**: The USA exhibits higher costs and longer wait times.
- **Specialist Access and Patient Satisfaction**: Trends reveal that Australia maintains an edge in both accessibility and patient satisfaction.
- **Treatment Effectiveness**: Treatment outcomes are closely aligned between both countries, indicating similar quality of care.

Conclusions and Recommendations

This study analysis evaluates AI-generated responses in healthcare contexts, contrasting USA and Australia across dimensions such as relevance, accessibility, cost, and patient satisfaction. Utilizing the MEDIQA-QA dataset, we have analyzed how these responses align with expert answers, examining the capabilities of AI in producing medically accurate, relevant, and empathetic content. The findings underscore the strengths and current limitations of AI as a tool for medical question-answering, with insights into areas where additional refinement could elevate its quality to expert levels.

Key Findings

- **High Quality Scores**: AI responses achieve an average rating of 8/10, suggesting effective alignment with expert answers in clarity, completeness, and medical accuracy.
- **Appropriate Medical Terminology**: AI-generated responses effectively utilize relevant medical vocabulary, reinforcing the credibility of the information provided.
- **Consistent Tone and Neutrality**: The tone remains factual and neutral, suitable for providing straightforward medical information, though it often lacks empathetic language that might enhance patient comfort.

Recommendations

1. **Enhanced Validation**: Implement additional validation checks for sensitive medical queries to further ensure response accuracy.
2. **Citation Support**: Integrate supporting citations within AI responses to substantiate medical claims and bolster trustworthiness.
3. **Balanced Response Length**: Strive for concise responses that enhance readability without compromising critical details, fostering improved comprehension.

Future Work

1. **Expand Dataset Scope**: Broaden the dataset to include specialized medical fields, enabling AI to develop expertise in diverse healthcare sub-domains.
2. **Real-Time Updates**: Establish continuous updates to incorporate the latest medical knowledge, ensuring relevance and accuracy in fast-evolving fields.
3. **Automated Quality Assessment**: Develop real-time, domain-specific quality metrics to assess AI response quality dynamically, facilitating on-the-go quality control.