Medical Summaries Evaluation

Evaluating Model Outputs with NLP Metrics and Human Feedback

Shreeti Shrestha
Northeastern University
CS 6983: Research in Human-Centered NLP
Instructor: Hye Sun Yun
Date Submitted: February 15, 2025

Overview

The task involved exploring the performance of an NLP model on summarization of randomly selected 150 abstracts of medical articles. The purpose of this evaluation was to make medical knowledge more accessible by generating summaries that can be understood by non-medical personnel.

Methods

Model Used: falconsai/medical_summarization

- The model is a variant of the pre-trained T5 transformer model, fine-tuned for the task of summarizing medical texts. It is designed to generate concise and coherent summaries of medical documents, research papers, clinical notes and other healthcare related texts.

Prompt Used:

- "Generate a plain language summary highlighting key points and removing unnecessary details that can be easily read and understood by non-medical people"
- The prompt is added as a prefix to each "abstract_text" when processing the inputs for the model

Additional Details:

- Used hugging face pipeline for summarization
- Parameters set: min_length = 30, max_length = 150 to generate brief summaries
- Generated output: A list summarized_outputs[] that stores 150 summaries for each record
- Output file located at summarized_outputs.csv:
    - abstract_text: original input
    - target_text: original generated summary
    - generated_text: model generated summary

Evaluation Methods and User Criteria:

- Automated Metrics:
    - ROUGE (the average of ROUGE-1, ROUGE-2, ROUGE-L scores) (0 – 1)
    - BLEU Score (0 – 1)
    - BERT Score (0 – 1 each for precision, recall, f1 scores)
    - Flesch-Kincaid Grade Level (where a score of 9.7 means that a ninth grader would be able to read the document)
    - Flesch Reading Ease Score (0 – 100, where readability increases with the score, 100 being the easiest to read)
- Human Evaluation:
    - A small-scale human evaluation study of 3 participants was conducted. Each participant read 15 summaries and answered 4 follow-up questions for each of the readings.
    - Criteria:
        - Persona: A layperson seeking to understand health advice quickly
        - Needs: Simplicity, absence of jargon
    - Among the 15 summaries allocated to each participant, 5 of them were common for all participants. The remaining 10 were selected at random.
    - Participant Demographics:

	Participant 1	**Participant 2**	**Participant 3**
Academic Background in Medicine	No	No	No
Grade Level	College	Graduate	Professional

Results

Automated Metrics

Rouge Scores (Aggregated scores for 150 records)

Metric Name	**Score (0 – 1)**
Rouge 1	0.38
Rouge 2	0.12
Rouge L	0.21
Rouge L Sum	0.21
Average of ROUGE-1, ROUGE-2, and ROUGE-L	0.24

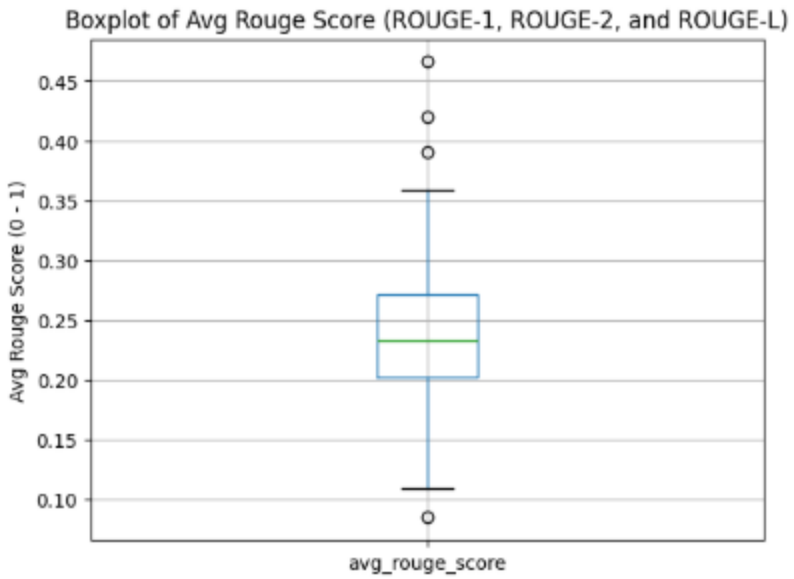

Figure 1: Box-plot for average rouge score (mean of Rouge1, Rouge2 and RougeL)

The average Rouge Score of 0.24 indicates that the model has some overlap, but lacks strong alignment between the generated and the reference texts.

BLEU Score

BLEU Score: 0.475

Similar to the rouge score, a BLEU score of 0.48 indicates about a moderate (48%) match to the reference texts on average, although BLEU score's sensitivity to exact phrasing may underestimate readability.

BERT Scores (Aggregated scores for 150 records)

Metric Name	**Score (0 – 100)**
Average Precision	85.72
Average Recall	85.05
Average F1 Scores	85.37

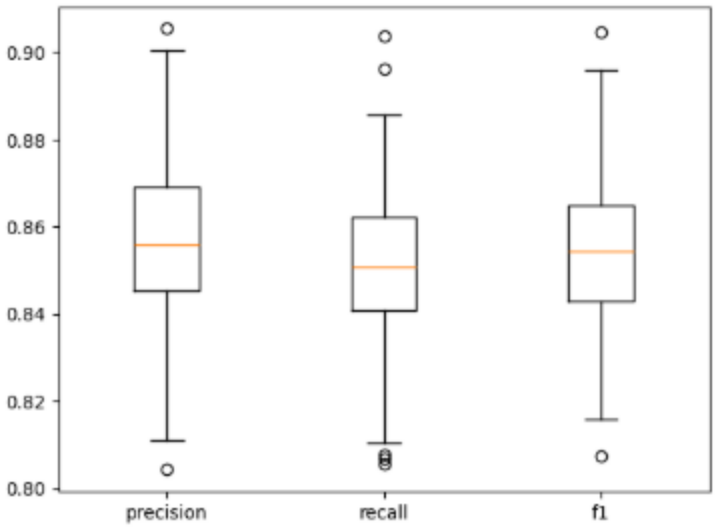

Figure 2: Box-plot of BERT Scores (precision, recall and f1)

Semantically, the generated texts are strongly similar to the reference summaries provided, meaning it conveys similar meanings even if wordings differ.

Flesch-Kincaid Grades (Readability Metric)

- Average Flesch-Kincaid Grade Level (FKGL) for 150 records: 16.51
- Average of Flesch Reading Ease (FRE): 16.16

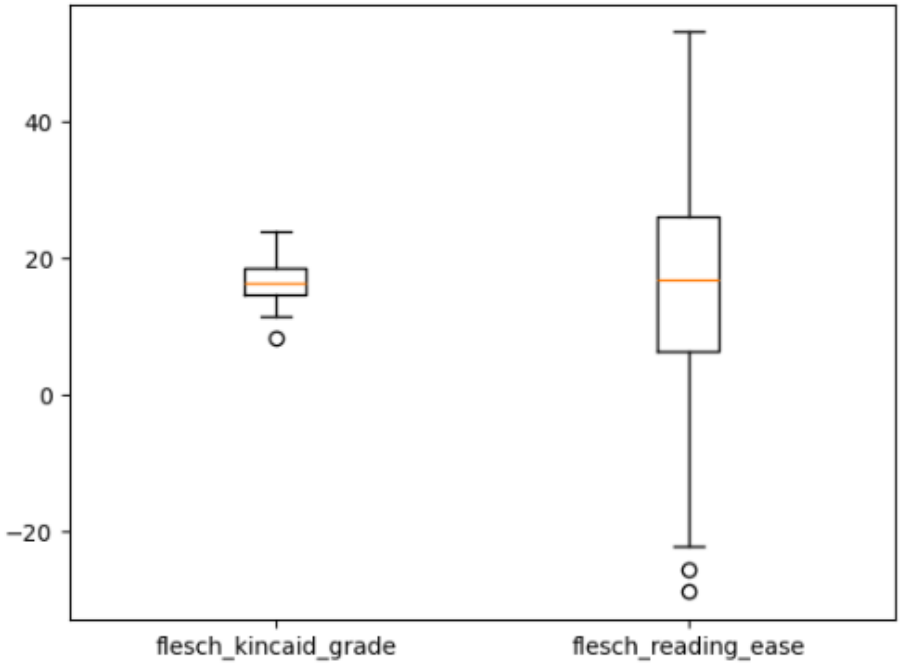

Figure 3: Box-plot for ease of reading and reading grade level

The average reading score is not that high (~16%), which indicates that the summaries generated by the model are generally harder to comprehend. Similarly, the FK grade level is also ~16% which suggests that it requires college-level comprehension.

Human Evaluation

Each participant rated 15 summaries, using a 5-point Likert scale

Questions	Median	InterQuartile Range	Mean	Standard Deviation
How easy was it to read the text? (1: Easy, 5: Difficult)	3.0	1.0	2.83	0.87
The text was filled with technical and/or medical jargon you could not understand. (1: Strongly Disagree, 5: Strongly Agree)	3.0	1.75	2.37	0.99
The text was brief and engaging while still conveying essential information. (1: Strongly Disagree, 5: Strongly Agree)	2.0	1.0	2.37	0.67
How useful did you find the text in understanding the key idea being discussed? (1: Not Useful at all, 5: Very Useful)	2.0	1.0	2.3	0.88
How much prior knowledge did you have about the concept being discussed in the text? (1: Never heard of it, 5: I'm an expert)	1.0	0.0	1.07	0.25

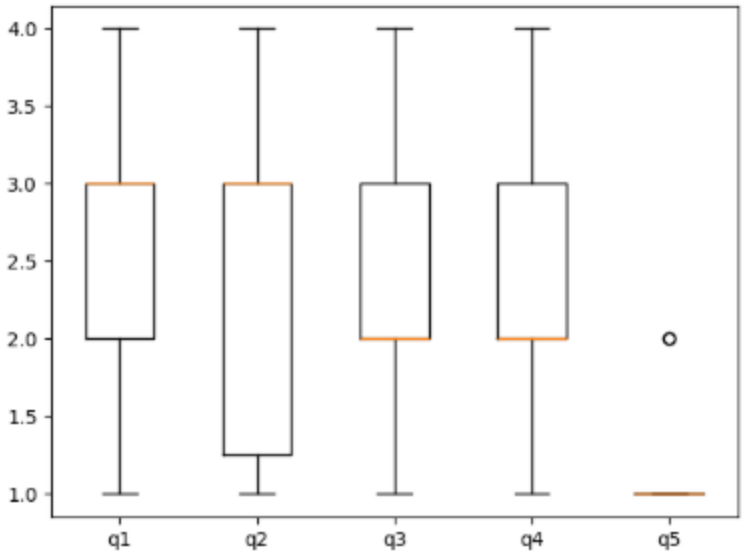

Figure 4: Box-plot for quantitative questions

The questions were meant to assess the following key areas:

- Readability: The relatively average score for Q1 indicates that most summaries lacked clarity, with a standard deviation of 0.87 to suggest some variation in difficulty perception.
- Jargon and Accessibility: Participants, on average, encountered some technical and/or medical jargon they struggled to understand, with high variability in evaluation where some summaries were significantly harder to grasp than others
- Conciseness and Engagement: An average of 2.37 suggests that participants found that some summaries were too dense and lacked a natural flow.
- Usefulness: 2/3 participants also indicated that some sentences towards the end were cut off for relatively longer summaries, and hence, it did not prove useful in conveying the idea.
- Prior Knowledge: A low score and low variation for the last question confirms that all participants were indeed non-experts with no prior knowledge of the medical concepts being discussed, reinforcing the need for simpler and jargon-free explanation.

Discussion

Gaps Identified /Limitations of Standard Metrics

While standard metrics such as ROUGE, BLEU or BERT scores assess textual similarity, and the Flesch-Kincaid grade level helped assess readability, there is no metric to assess whether the information is useful, and actionable. A high-scoring summary result might still fail to provide practical information/purpose or recommendations to the reader.

Another limitation of using standard metrics is that they could show lower values for summaries in plain language that are more natural and readable, as these metrics penalize paraphrasing, and focus more on similarity based on overlap. This also runs the risk of having medical jargon or technical terms that distract a layperson reading the texts.

While the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) scores measure readability, they do not indicate whether the summary remains informative and relevant after simplification, and so, these scores often neglect the balance between readability and relevance.

Improvement Ideas

The current model struggles with jargon removal and engagement possibly due to being pre-trained on medical texts. One potential improvement could be to preprocess and train the summarization model on datasets specifically designed for laypersons. Another approach could be to use word embeddings or predefined medical vocabulary mappings to find simpler synonyms for complex medical terms. We can also make more nuanced use of prompting in prompt tuning the model to encourage action-driven outputs so that the summaries provide useful and actionable steps that vague information.

For human based evaluation, for each summary, we could add questions based on the context to better assess understandability and usefulness. It could also be helpful to ask users if the summaries they read felt unbiased to assess readers' trust in the generated texts.

Conclusion

Overall, this task demonstrated that designing plain language summarization, specifically in the context of medical articles, requires a user-first approach. The evaluation of the model tested in this study highlighted critical gaps in readability, engagement and actionability. While the standard metrics helped measure word overlap and readability, human evaluation revealed the lack of engagement and the lack of actionable outcome for a layperson reading these texts. Hence, to truly make medical summaries accessible to a general reader, our evaluation needs to put end readers at the forefront of assessing these models.

References

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know Your Audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11--16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA 26 Pages. https://doi.org/10.1145/3613904.3642289

Guo, Yue, et al. "APPLS: Evaluating Evaluation Metrics for Plain Language Summarization." arXiv preprint arXiv:2305.14341 (2023).