Healthcare Answer Summarization

Avikalp Rewatkar
MT24022

Sai Krishna Kota
MT24078

Shubhankar Tiwary
MT24139

Abstract

In this project, we focus on summarizing healthcare-related answers according to a given perspective. This task enables healthcare information seekers to view different perspectives. With this, users can quickly assess all the answers, making them feel less overwhelmed and leading to a more informed decision-making process. The paper "No perspective, no perception!! Perspective-aware Healthcare Answer Summarization" got us to do experiments, which included fine-tuning several language models, like Flan-T5-Base and BART-Large-CNN, to generate clear and relevant summaries. To improve results further, we also implemented a Mixture of Experts (MoE) model based on Flan-T5-Base. We aim to compare how these models and architectures perform and explore better ways to summarize healthcare answers.

1 Introduction

In recent years, the advancement of Natural Language Processing (NLP) has revolutionized various applications, notably in tasks that require explicit reasoning and logical deduction. While much research has focused on solving problems through traditional, vertical approaches—emphasizing systematic, step-by-step reasoning in tasks like question-answering and text classification—the domain of lateral thinking has remained relatively underexplored.

Healthcare community question-answering platforms present a wealth of information from diverse user-generated responses. These responses range from factual medical information to personal experiences and recommendations. However, the abundance and heterogeneity of these insights often overwhelm users, making it difficult to distill the most relevant information. In response, our project introduces a perspective-aware summarization framework that fine-tunes a Flan-T5 model using Low-Rank Adaptation (LoRA) techniques. This approach enables the model to generate concise, relevant summaries that capture distinct perspectives.

The motivation behind this project originates from the need to structure the existing overwhelming amounts of unstructured healthcare information into concise, actionable summaries. In real-world scenarios, patients and doctors alike are faced with diverse perspectives on medical issues—ranging from factual explanations to personal experiences—that can complicate decision-making and often lead to misinformation. Moreover, the project is an excellent opportunity to refine our skills in the summarization fine-tuning domain, bridging cutting-edge research with practical applications and fostering advancements in how NLP systems handle complex, multifaceted tasks.

2 Related Work

Recent research in perspective-aware summarization has highlighted the importance of capturing diverse viewpoints in healthcare QA platforms. In particular, the paper "No perspective, no perception!! Perspective-aware Healthcare Answer Summarization" introduces the PUMA dataset and the PLASMA model, which utilizes prompt-based control mechanisms, prefix tuning, and an energy-controlled loss function to generate perspective-specific summaries. This work demonstrates significant improvements over traditional summarization models by explicitly modeling varied perspectives such as information, suggestion, cause, and experience. Our project builds upon these insights, aiming to extend the approach through efficient fine-tuning techniques and enhanced sum-

marization strategies in real-world healthcare settings.

3 Dataset

The dataset used in this project is structured and annotated to capture the multiple perspectives of healthcare question-answer threads. It consists of seven key columns: a unique identifier ("uri"), the primary healthcare query ("question"), and additional context ("context") that may supplement the query. The "answers" column contains a list of responses from various users, reflecting a range of viewpoints. Crucially, two additional columns—"labelled_answer_spans" and "labelled_summaries"—provide detailed annotations for different perspectives. The "labelled_answer_spans" column marks specific portions within the answers according to perspectives, delineating where each type of content appears. In parallel, the "labelled_summaries" column contains manually crafted summaries corresponding to these perspectives, with keys like "INFORMATION", "SUGGESTION", "CAUSE", and "EXPERIENCE" (and few "QUESTION"). The final column, "raw_text", combines the entire dialogue unprocessed.

The training dataset contains 2,236 entries, with each entry including multiple summaries from different perspectives. The validation dataset comprises 959 instances, and the test dataset has 640 entries. To enable perspective-specific modeling, we expanded the training dataset by duplicating each entry for every available perspective, assigning a unique summary per instance. This resulted in a total of 1,252 entries in the expanded training set. This transformation provides a robust foundation for developing a perspective-aware summarization model.

Lastly, we analyzed the word and token counts for each perspective in the training set to better understand the summary length characteristics and guide our generation settings accordingly.

INFORMATION_SUMMARY: Average word count: 66.95, Average token count: 96.58
SUGGESTION_SUMMARY: Average word count: 50.74, Average token count: 70.04
CAUSE_SUMMARY: Average word count: 30.66, Average token count: 44.30
EXPERIENCE_SUMMARY: Average word count: 51.70, Average token count: 71.08
In the train dataset, the count of instances of each perspective is shown below:
INFORMATION: 1742
CAUSE: 305
SUGGESTION: 1363
EXPERIENCE: 745
QUESTION: 213

Since length is highly relevant in summarization tasks, we use this information to determine the appropriate generation length for the summaries.

The dataset Structure is shown in Table 1

3.1 Data Split

The dataset used for this task was initially split into 2,236 instances for training, 959 for validation, and 640 for testing. To better train the model for each perspective, we expanded the training set by treating each perspective-specific summary as a separate instance. This involved duplicating the original questions while associating them with individual perspective summaries. After this expansion, the training dataset grew to 1,252 instances, with each entry focusing on a single, distinct perspective.

4 Methodology

In this section, we describe the approaches and techniques used to tackle the healthcare summarization task. Our objective was to generate high-quality summaries from question-answer pairs annotated with spans like information, cause, suggestion, experience, and question. We experimented with both standard fine-tuning and a Mixture of Experts (MoE) architecture to assess their effectiveness.

4.1 Baseline 1

We benchmarked various open-source LLMs using a zero-shot setup to evaluate their ability to generate perspective-based summaries without task-specific fine-tuning. In particular, we explored DeepSeek-R1-Distill-Qwen-1.5B and BART-Large models.

4.2 Baseline 2

We tried to Reproduce the PLASMA Model from the attached reference paper. The authors made the source code publicly available. We tried to keep the training process close to the original implementation.

Column Name	Example
uri	4367393
question	“what is parkinesonism?”
context	“I’ve been feeling tired and my joints hurt.”
answers	[”Parkinson’s disease is one of the most comm...]
labelled_answer_spans	{INFORMATION: [(0, 34)], SUGGESTION: [(35, 60)]}
labelled_summaries	{INFORMATION_SUMMARY : “Parkinson’s disease is one...”, SUGGESTION_SUMMARY: “...”}
raw_text	“uri: 4367393 question: what is parkinesonism?contex...”

Table 1: Structure and sample entries of the puma dataset

4.3 Standard Fine-Tuning Architecture

We applied standard fine-tuning techniques on pre-trained sequence-to-sequence models using our healthcare dataset.

For standard fine-tuning, we introduced the structured prompt technique. In this method, we combined the perspective and its definition, the question being asked, multiple answers to that question, and the summary for that specific perspective. This approach gives the model a clear and organized view of the information, helping it understand the context more effectively. Here’s an example of the structured prompt:
”Summarize the following content according to Perspective: Suggestion; Suggestion Definition: Defined as advice or recommendations to assist users in making informed medical decisions, solving problems, or improving health issues; Begin Summary with: ‘It is suggested’ ; Tone of summary : Advisory, Recommending Content to summarize: A1, A2...An”

4.3.1 BART

For this summarization task, we fine-tuned BART-large-cnn. The input to the model included the concatenated structured prompt template as mentioned above, while the target output was the perspective summary. The model was trained using a cross-entropy loss function. We used early stopping and validation ROUGE scores to prevent overfitting.

We chose BART-Large-CNN because it is a pre-trained model on BART architecture with the CNN dataset specifically designed for summarization tasks, hence it performs well to understand context, handle long input documents, and generate summaries.

4.3.2 Flan-T5

We also fine-tuned the Flan-T5-small model for this task. Similar to BART, the input consisted of the question and answer span texts, and the model was trained to generate concise summaries. These models were trained using fp16 precision for faster training and better resource utilization.

4.4 Mixture of Experts

We employed a Mixture of Experts architecture to train specialized models for each summary perspective. Specifically, we fine-tuned four separate models, each on a subset of the dataset corresponding to a particular summary perspective. During inference, we designed a pipeline that routes each input to its respective expert model, allowing us to generate and evaluate summaries tailored to each perspective.

4.4.1 Flan-T5

We used the Flan-T5 model as the base for our Mixture of Experts architecture for its instruction-tuned tasks. Flan-T5 is pre-trained on a diverse set of tasks and follows an encoder-decoder structure, making it well-suited for summarization. In our MoE setup, we fine-tuned separate instances of Flan-T5 for each summary perspective—INFORMATION, SUGGESTION, EXPERIENCE, and CAUSE—allowing each expert model to specialize in generating summaries aligned with its target viewpoint. This modular approach helped capture the semantic patterns unique to each perspective, improving generation quality and interpretability without needing a single monolithic model.

For Flan-T5 fine-tuning, we experimented with two approaches concerning summary lengths. Ini-

tially, we fine-tuned the expert models and observed that the generated summaries were too short, which is reflected in the evaluation scores presented in Table 3. To address this, we explicitly set a minimum generation length of around 60–70 words based on the average word count observed in the training dataset. The improved results following this adjustment are shown in Table 4.

4.4.2 Flan-T5 with LoRA

We applied the Low-Rank Adaptation (LoRA) technique using Flan-T5 as the base model to fine-tune it for this task.

LoRA is a parameter-efficient fine-tuning technique that adapts large pre-trained language models to specific downstream tasks with significantly fewer trainable parameters. Instead of updating all weights in the model, LoRA injects low-rank matrices into certain layers (typically attention weights), allowing effective task-specific learning while keeping most of the original model frozen. This drastically reduces computational and memory costs.

In our project, we used LoRA with Flan-T5 to fine-tune models for different summary perspectives, enabling faster training and better generalization with limited data and resources.

5 Experimental Setup

- **Hardware:**
  - **GPU:** NVIDIA Tesla T4 (16 GB) available on both Google Colab and Kaggle Notebooks, ensuring efficient training and fine-tuning.
  - **RAM:** 16 GB, which is sufficient for smooth model training and testing without memory issues.
  - **CPU:** Standard CPUs provided by both platforms, suitable for data preprocessing and evaluation tasks.
- **Software:**
  - **Framework:** Hugging Face Transformers library for accessing pre-trained models (e.g., BART, FLAN-T5) and for fine-tuning and evaluation.
  - **Deep Learning Library:** PyTorch, chosen for its flexibility and compatibility with the Transformers library.
  - **Operating System:** Linux-based environments provided by Google Colab and Kaggle Notebooks.
  - **Python Version:** Python 3.8, ensuring compatibility with the latest versions of necessary libraries.

6 Evaluation Metrics

6.1 BLEU

BLEU-4 measures the n-gram overlap between generated summaries and reference texts, focusing on matching sequences of four consecutive words. A high BLEU-4 score indicates that the generated text closely mirrors the phrasing and structure of the reference, suggesting good lexical similarity, Whereas, a low BLEU-4 score means there is little direct overlap in word sequences, which could indicate that the model is missing key details, or producing less fluent outputs which are not similar to the referenced outputs.BLEU-4 is typically on a scale of 0 to 100.

6.2 BERT

BERT-F1, on the other hand, assesses semantic similarity by comparing contextual embeddings of the generated and reference texts. A high BERT-F1 score demonstrates that the model has successfully captured the underlying meaning and nuance of the reference summary, even if the exact wording differs. A low BERT-F1 score signals that the generated output diverges semantically, indicating that the summary might be off-topic, lacking details, or not in the context intended.BERTScore ranges from 0 to 1.

7 Results

7.1 Baseline 1

Table 2 presents the performance metrics for both the BART and DeepSeek-R1-Distill-Qwen-1.5B models in a zero-shot setup.

Model	BERT-F1	BLEU-4
BART	0.828	0.109
DeepSeek	0.852	0.073

Table 2: Benchmarking Results: BART DeepSeek-R1-Distill-Qwen-1.5B

7.2 Baseline 2

The PLASMA model was replicated following the guidelines in the paper. Table 3 compares the performance of our trained PLASMA model with that reported in the paper, highlighting differences in BLEU-4 and BERTScore metrics. Model from the paper and our training.

PLASMA Model	BERT-F1	BLEU-4
Paper	0.869	0.040
Our Model	0.810	0.007

Table 3: Comparision Results: Replication of PLASMA model comparision with metric in the paper

7.3 Standard Fine-Tuning Architecture

7.3.1 BART-Large-CNN

The resulted BLEU-4 score is 0.038 and BERT-F1 score is 0.684. This score shows that this model, although it captures the key points, may not capture the relevant information/details from the answer and may miss the important details from a given perspective.

The BLEU score of 0.0380 suggests that the generated summaries may have limited overlap with reference summaries in terms of word choice and structure.

7.3.2 Flan-T5

For Flan-T5-small, we achieved a comparable high score for both BERT-F1 score and BLEU-4 score. The resulting BERT-F1 score is 0.870. This score suggests that FLAN-T5-small is effective at capturing essential details and context, providing a more reliable extraction of key points.

BLEU-4 score for this model is 0.051 which signifies stronger performance in generating summaries that closely matches with given reference summaries in terms of selection of words and overall structure.

Model	BERT-F1	BLEU-4
Flan-T5-Small	0.870	0.051
BART-Large-CNN	0.684	0.038

Table 4: Evaluation metrics for fine-tuned models

7.4 Mixture of Experts

7.4.1 FlanT5

For the minimum-length-constrained fine-tuned model, the BLEU-4 score is 4.28 and the BERT-F1 score is 0.085 (as shown in Table 6). Both scores are quite low, indicating that the model is likely generating overly short or irrelevant summaries that fail to capture the semantic meaning of the reference summaries.

In contrast, the standard fine-tuned model, trained for 3 epochs over the dataset, achieves a BERT-F1 score of 0.197 and a BLEU score of 0.79 (as shown in Table 5). While these scores are moderate and improvable, they represent a clear performance gain over the length-constrained model, suggesting better semantic alignment and content overlap.Still they suggest that the summaries generated are too short or quality of data is lacking.

Model	BERT-F1	BLEU-4
Overall	0.197	0.79
Information_Summary	0.166	0.50
Suggestion_Summary	0.199	0.38
Cause_Summary	0.256	1.68
Experience_Summary	0.234	3.38

Table 5: Evaluation metrics for different models of Flan-T5 with MoE Architecture

Model	BERT-F1	BLEU-4
Overall	0.085	4.28
Information_Summary	0.756	3.94
Suggestion_Summary	0.086	3.18
Cause_Summary	0.110	2.52
Experience_Summary	0.091	6.01

Table 6: Evaluation metrics for different models of Flan-T5 with MoE Architecture and specified minimum length

7.4.2 Flan-T5 with LoRA

The BERT-F1 score of 0.192 is relatively low, indicating that the generated summaries capture only a limited portion of the semantic content from the reference summaries.

The BLEU-4 score, which measures the overlap of 4-gram sequences between the generated and reference summaries, is 0.50. This is a moderate score, suggesting that while the generated summaries match some phrases from the reference texts, they may fail to fully convey the intended meaning.

Model	BERT-F1	BLEU-4
Overall	0.192	0.50
Information_Summary	0.150	0.24
Suggestion_Summary	0.210	1.60
Cause_Summary	0.248	0.23
Experience_Summary	0.228	2.73

Table 7: Evaluation metrics for Flan-T5 with LoRA on MoE Architecture

8 Discussion of Results

8.1 Standard Fine-Tuning Architecture

The final results shows that Flan-T5-small fine-tuned for this task gives relevant summary as compared to BART-Large-CNN. The results of our experiments show that FLAN-T5-Small outperforms BART-Large-CNN in healthcare answer summarization. With a higher BERT F1 score (0.8709 vs. 0.6841), FLAN-T5-small demonstrated better accuracy in extracting relevant information. Additionally, FLAN-T5-small achieved a higher BLEU score (0.0512 vs. 0.0380). This indicates more fluent and coherent summaries. These findings suggest that FLAN-T5 is more effective at understanding healthcare content and generating high-quality summaries. While both models show potential, FLAN-T5-small's superior performance highlights its suitability for healthcare summarization tasks, though further improvements are still possible, particularly in generating more fluent summaries.

8.2 Mixture of Experts

Some metrics for individual "expert" models are quite poor, indicating that the primary issue with the Mixture of Experts architecture is the lack of uniform data across each expert model. Specifically, for perspectives like SUGGESTION or EXPERIENCE, where the number of instances is significantly lower, the models struggle to generate high-quality summaries due to the insufficient amount of training data.

8.3 Comparision

On comparing the two architectures based on current metrics, the Flan-T5-Small model with standard fine-tuning produces the most semantically accurate summaries. In contrast, the Mixture of Experts (MoE) architecture shows several drawbacks with its current approach, likely due to insufficient and unevenly distributed data and limited computational resources during training. While the MoE approach is currently underperforming, it could potentially yield better results if a more balanced dataset and additional training resources are provided.

9 Conclusion

In this study, we concluded that that standard fine-tuning of Flan-T5-Small performs better than BART-Large-CNN with comparatively high scores.

For the Mixture of Experts architecture, it shows potential for perspective-aware summarization, but its effectiveness is highly dependent on the availability of balanced and sufficient training data. With the current dataset, the performance of certain expert models—especially for underrepresented perspectives such as SUGGESTION and EXPERIENCE—remains suboptimal. This shortfall suggests that the existing data volume is inadequate for achieving high-quality outputs using this architecture.

10 Future Work

Future work should focus primarily on improving the training data, particularly for perspectives with limited instances, to create a more uniform dataset. Additionally, exploring alternative fine-tuning methods, hyperparameter tuning, and allocating more computational resources could potentially enhance model performance. Improving these aspects may enable the MoE architecture to reach its full potential in generating better summaries.

Also, we will experiment on different baselines and their fine-tuning for each perspectives based on which specific model will generate summaries for that particular perspective. This may improve performance as different model will be trained for different perspectives.

11 Links

Github Repo: Link

Baseline Models: Google Drive

Standard Fine-Tuning Models: Google Drive

MoE Saved Models: Google Drive

12 References

[1] G. Naik, S. Chandakacherla, S. Yadav, and M. S. Akhtar, "No perspective, no perception!! Perspective-aware Healthcare Answer Summarization," *arXiv preprint arXiv:2406.08881*, Jun. 2024.