GROUP-20 ; Efficient Summarization of Healthcare Responses

Dhairya
Roll No: 2022157
dhairya22157@iiitd.ac.in

Harsh Vishwakarma
Roll No: 2022205
harsh22205@iiitd.ac.i

Pandillapelly Harshvardhini
Roll No: 2022345
pandillapelly22345@iiitd.ac.in

Abstract

Today, Natural Language Processing has become very advanced which has enabled the development of large and complex models which are capable of generating high quality summaries from the given text taken from the Internet for its corresponding task. In this work, we explore the effectiveness of FLAN-T5 and PLASMA models for perspective-aware healthcare summarization tasks. Meanwhile the existing methods only focuses on general purpose summarization, they often fail to capture the contextual and perspective based that are critical for real world applications. Our study tells how these models handle multi-perspective summarization by incorporating structured prompt-based learning and energy controlled loss function. We evaluate there performance using the evaluation metrices like BLUE, BERTscore etc. highlighting their effectiveness in generating concise and contextually related summaries.

1 Introduction

With the rise of digital platforms such as Quora, Reddit, and Yahoo!—community question-answering (CQA) forums—how people seek and share information has been revolutionized. These platforms allow users to discuss various topics, such as healthcare, technology, and personal experiences, enabling the collaborative exchange of knowledge.

Online platforms have become a popular space for medical discussions, with communities such as r/AskDocs, r/DiagnoseMe, r/Medical_Advice, and r/Medical serving as hubs for users seeking health-related guidance. These forums cater to a wide range of inquiries, from general health concerns and symptom assessments to questions about medical procedures and professional practices.

While these platforms provide valuable support and collective knowledge, the vast number of responses—ranging from medically accurate insights to personal anecdotes and subjective opinions—can make it difficult for users to extract reliable and meaningful information. This challenge is further amplified by the variation in response quality, the lack of standardized medical verification, and the potential for misinformation, making it crucial for users to critically assess the advice they receive. Summarizing these responses in structured and perspective aware manner is important for improving the accessibility and comprehension. Traditional methods of summarization focuses on extracting the key information, they often overlook the importance of capturing the multiple viewpoints from the text. Recent advancements in transformers based models, particularly, FlanT5 and PLASMA, these models have addressed the problem of traditional model as mentioned above by integrating the prompt based learning. In this paper, we explore the efficacy of these models in generating structured, perspective-driven summaries and assess their performance using datasets and evaluation metrics.

2 Problem Statement

In the digital age, online health forums and medical Q&A platforms serve as vital resources for individuals seeking medical advice. However, the sheer volume of responses—ranging from expert-backed insights to personal experiences—poses a significant challenge in extracting reliable and meaningful information. The lack of standardized verification, varying response quality, and potential for misinformation further complicate the process, leaving users uncertain about which advice to trust.

Our goal is to develop a platform that generates summaries by extracting diverse perspectives—such as causes, suggestions, and personal experiences—from multiple responses to a single medical question. By utilizing advanced recommendation algorithms, our system will intelligently generalize user-posed medical queries and curate the most accurate, relevant, and evidence-based information. It will filter out inconsistencies, prioritize medically verified content, and present users with a well-rounded, credible, and informative summary, ensuring improved access to trustworthy health information.

3 Related Work

3.1 Perspective-aware Summarization

(Gauri, Akhtar et al. 2024) proposed a novel task of perspective-specific summarization in the healthcare domain by annotating answers with five labels – cause, suggestion, experience, question, and information. They introduced a dataset called PUMA and a summarization model named PLASMA which uses prompt-based generation with Flan-T5 and prefix tuning. Additionally, they designed an energy-based loss to enforce alignment with the desired perspective. To handle multiple perspectives, they incorporated perspective-specific prompts and learned separate control signals to generate summaries conditioned on each perspective.

4 Dataset Description and EDA

The dataset comprises three JSON files—`train.json`, `valid.json`, and `test.json`—structured for the task of perspective-based summarization. Each file contains a list of examples represented as JSON objects with the following fields:

- **uri**: A unique identifier for each question-answer instance.
- **question**: A natural language question posed by a user.
- **context**: Background context associated with the question, often empty.
- **answers**: A list of raw community-generated answers from public forums, varying in tone, accuracy, and content type.
- **labelled_answer_spans**: A dictionary mapping perspective types (e.g., `INFORMATION`, `CAUSE`, `SUGGESTION`, `EXPERIENCE`) to a list of relevant answer spans. Each span includes:
  - `txt`: The actual text snippet extracted from an answer.
- **labelled_summaries**: A dictionary of human-written summaries for each perspective category present in `labelled_answer_spans`.
- **raw_text**: The concatenated textual form of the original question, context, and answers, used for reference or token-level alignment.

This structure supports supervised training of models to generate perspective-based summaries by mapping raw answers to categorized summary targets.

Dataset Split	Number of Samples
Train	2236
Validation	959
Test	640

Table 1: Dataset split for training, validation, and testing

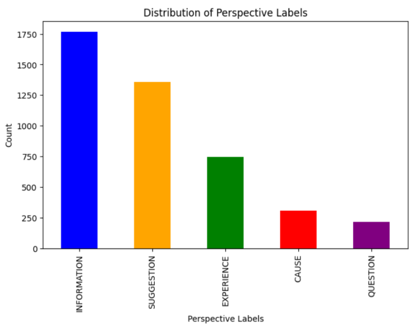

Figure 1: Distribution of Perspective Types in Labelled Answer Spans

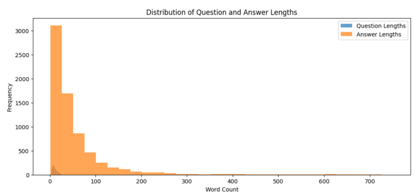

Figure 2: Length Distribution of Answers and Summaries

5 Baseline

We evaluated our approach using two baseline model given in the research paper: Flan-T5 and PLASMA, where Flan-T5 used for summerization and PLASMA is used perspective-specific summarization on the PUMA dataset.
Flan-T5: Flan-T5 is a fine-tuned version of T5 model. A standard fine-tuned Flan-T5 trained on question and answer concatenations with perspective labels. This model provides a strong

starting point for generating summaries using frozen backbone weights.
PLASMA: PLASMA (Perspective-aware heaLthcare Answer SuMmarizAtion) builds on top of Flan-T5 by incorporating prefix tuning and an energy-based loss function. It introduces soft prefixes learned during training to guide the generation toward a target perspective. Additionally, PLASMA includes a perspective-controlled objective that calculates energy values based on perspective alignment, tone, and anchor-text matching. This allows PLASMA to generate more accurate and perspective-specific summaries without modifying the base model parameters.

Epoch	Training Loss	Validation Loss
1	1.364	0.872
2	0.858	0.837
3	0.829	0.821
4	0.812	0.814
5	0.802	0.793

Table 2: Training and Validation Loss per Epoch of Plasma

Metric	PLASMA	FLAN-T5
METEOR	0.0586	-
BLEU-1	0.0332	0.30
BLEU-2	0.0127	0.30
BLEU-3	0.0074	0.30
BLEU-4	0.0046	0.30
BERTScore	0.7598	-

Table 3: Comparison of evaluation metrics for the PLASMA and FLAN-T5

6 Methodology

This project addresses the task of abstractive summarization in a question-answering context using advanced Natural Language Processing (NLP) techniques. The methodology encompasses data preparation, model development, ensemble integration, and comprehensive evaluation.

6.1 Data Preparation

The dataset consists of structured JSON files containing the following fields: question, context, answers, labelled summaries. Separate files were provided for training, validation, and testing: `train.json`, `valid.json`, and `test.json`.

Various modeling approaches were used for summarization:

6.2 Weakly Supervised Perspective-Aware Summarization

6.2.1 Snorkel + LR

From the initial data set we have extracted text spans with its corresponding perspectives into train-df and valid-df. Then each entry is tagged with numerical label 0-4
Now for Embeddings and Classification, we used sentense embedding model all-mpnet-base-v2 to convert text into vectors. Then trained Logistic Regression classifier on the embeddings.
Now this model can predict the perspective of new unseen by seeing their embeddings.
We have created labeling function which assign a label based on presence of keywords. These are used by Snorkel's LabelModel which learns a probabilistic ensemble of all LFs to give a more accurate label, even if individual LFs are weak.
After this we have done perspective prediction on Testset. Each test.json entry is processed in which Answer text is split into sentences. Each sentence is assigned a label using: Snorkel if confident Logistic Regression otherwise. Then we have saved it to classified-test-output.json
Now summarize each group of sentences per perspective.
Extractive Summarization (BART): Join all sentences for a perspective, Use BART to summarize them, Output is informative but factual.
Abstractive Refinement (Pegasus): Take BART summary and use Pegasus, Pegasus adds abstraction, paraphrasing, more fluent summarization.
For each perspective with a reference summary compare the Pegasus summary to the human-written reference using BLEU and BERTScore.

6.2.2 Snorkel + SVM + Zero-Shot

For this we have trained two same models first a simple bart then a bart with few shorts learning. Used SentenceTransformer (MiniLM) to encode each span into a dense vector (embedding) and then train a Support Vector Machine (SVM) classifier to predict the perspective label from the embeddings.
Create Snorkel Labeling Functions (LFs) using regex-based keyword patterns derived from the training data. Each LF tries to identify a specific perspective based on the presence of certain keywords. Then trained Snorkel LabelModel.
If Snorkel LFs abstain (-1) and SVM is unsure or its

prediction isn't mapped to a known label. Then apply a Zero-Shot Classifier (based on facebook/bart-large-mnli) to predict the label using natural language inference. This ensures every sentence gets a perspective, even if other models abstain. Now For testing each test entry's answers are split into sentences. Then each sentence is labeled using the combined logic.

- For simple Bart training, first we loaded train.json and test.json from the PUMAAA dataset then you extract labeled sentence spans and their corresponding perspective labels into a Pandas DataFrame.
  For each perspective group of sentences (e.g., all INFORMATION sen), you generate a summary using BART. The input is concatenated list of all sentences in that perspective.

- we generate perspective-specific summaries using a few-shot learning approach combined with BART. From the train.json of the PUMA dataset, we extract pairs of user questions concatenated with their answers and corresponding human-written summaries, which are then embedded using Sentence-Transformer (MiniLM) to enable semantic similarity search. During inference, for each set of perspective-labeled sentences (predicted using a previous classification pipeline), we compute the embedding of the combined text and retrieve the top-3 most similar examples from the training set. These examples are formatted as few-shot prompts and prepended to the input before being passed to the facebook/bart-large-cnn model for summarization. The model dynamically adjusts summary length based on input size.

6.3 Using stacking ensemble

6.3.1 Fine-tuning Flan-T5 with LoRA + Pegasus

We aimed to generate effective summaries based on multiple perspective-tagged answers (CAUSE, INFORMATION, SUGGESTION, QUESTION, EXPERIENCE). The goal was to generate a single, coherent summary that captures the core idea of the answers. Generating such summaries is complex because different users expect different types of answers: some want background information, others want causes, and some expect suggestions or shared experiences. To address this, we started by fine-tuning the Flan-T5 model using LoRA (Low-Rank Adaptation). LoRA allows us to update only a small subset of parameters during training while keeping the rest of the model frozen, making the process memory-efficient and faster to train. We formatted each input like an instruction: "summarize for PERSPECTIVE: QUESTION [CONTEXT]", where PERSPECTIVE was either CAUSE, INFORMATION, SUGGESTION, Experience , QUESTION. Alongside Flan-T5, we also used the Pegasus model, known for its strength in abstractive summarization. We created a stacking ensemble using the outputs of both models. For each input, we generated summaries from both Flan-T5 (with LoRA) and Pegasus and compared the results.

However, we observed that this combination sometimes lacked relevance — the generated summaries were either too general or did not capture the core points well.

6.3.2 Fine-Tuning BART + Fine-tuning Flan-T5 with LoRA

To improve the quality of the summaries, we fine-tuned another model: BART, a powerful encoder-decoder transformer architecture. BART was trained on the same dataset, with the same input format and perspective-wise summaries.

After training, we combined the fine-tuned BART model with the Flan-T5 (LoRA) model to create another stacked ensemble. For each question, we generated summaries from both models and selected the more relevant output based on fluency and perspective alignment. This final ensemble helped us combine the strengths of both models:

BART was more fluent and natural in language generation.

Flan-T5 with LoRA was better at following the instruction and producing summaries more aligned to the perspective prompt.

By comparing the outputs from both models and picking the best one, we were able to generate higher-quality summaries that retained the most important aspects of user responses while maintaining clarity and relevance

- Averaging semantic similarity scores (e.g., BERTScore) to select the most contextually aligned summary.

- Using hybrid voting mechanisms based on model confidence or agreement.

This approach aimed to exploit the complementary strengths of individual models to generate higher-quality summaries.

7 Experimental Setup

7.1 Stacking ensemble

We conducted experiments using the PUMA dataset, split into `train.json`, `valid.json`, and `test.json`. The models were trained on the training set, validated on the validation set, and evaluated on the test set.

We fine-tuned three models:

- **Flan-T5 with LoRA**: Efficient tuning using Low-Rank Adaptation.
- **Pegasus**: Pretrained for abstractive summarization.
- **BART**: Fine-tuned using the same perspective-based format.

Inputs were formatted as:

> summarize for {PERSPECTIVE}: {QUESTION} [CONTEXT]

and outputs were the corresponding summaries.

All models were trained using Hugging Face Transformers with batch size 8, 5 epochs, and learning rate 5×10^{-5} using AdamW optimizer. Bert,Bleu scores were used for evaluation.

We also created a stacked ensemble using BART and Flan-T5 outputs,also pegasus and Flan-T5 selecting the better summary per input.

7.2 Weakly Supervised Perspective-Aware Summarization

In addition to supervised fine-tuning, we implemented a weakly supervised pipeline to classify and summarize answer spans by perspective without relying on manually labeled data.

- **Snorkel + Logistic Regression (LR)**: Sentence embeddings were generated using `all-mpnet-base-v2` and a Logistic Regression model was trained to predict perspective labels. Snorkel LabelModel was used to generate weak labels via keyword-based labeling functions.
- **Perspective Labeling on Test Set**: Each answer was split into sentences and labeled using Snorkel (if confident) or the LR classifier otherwise. The output was saved as `classified-test-output.json`.
- **Two-Stage Summarization Pipeline**:
  1. **Extractive Summarization (BART)**: Perspective-grouped sentences were summarized using BART.
  2. **Abstractive Refinement (Pegasus)**: BART summaries were refined by Pegasus to improve fluency and abstraction.
- **Snorkel + SVM + Zero-Shot**: Sentence embeddings from `MiniLM` were used to train an SVM classifier. Snorkel LFs generated weak labels, and zero-shot classification (`facebook/bart-large-mnli`) was used as a fallback when both Snorkel and SVM abstained.

8 Results and Evaluation

8.1 Perspective based summaries for fine-tune flan-t5 with LoRA + Pegasus

Perspective	BERTScore F1	Examples
Information	0.8662	484
Cause	0.8666	102
Suggestion	0.8582	392
Experience	0.8487	205
Question	0.8395	64

Table 4: BERTScore F1 per Perspective Category for FlanT5+Pegasus model

8.2 Metrices for fine-tuned flan-t5 with LoRA + Fine-tuned BART

Metric	BART	T5	Stacked
BERTScore(prcn)	0.9024	0.8645	0.8881
BERTScore(rcl)	0.8800	0.8628	0.8758
BERTScore F1	0.8907	0.8632	0.8815
BLEU	0.0883	0.0363	0.0747
METEOR	0.2544	0.1856	0.2386

Table 5: Performance comparison of BART, T5, and Stacked models across evaluation metrics.

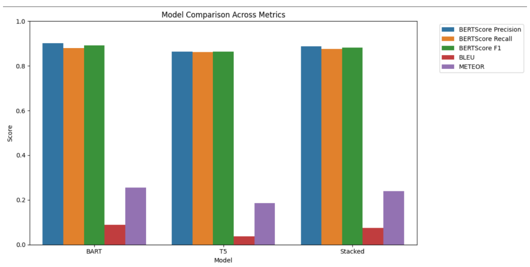

Figure 3: Model performance across Metrices

8.3 Weakly Supervised Perspective-Aware Summarization

Metric	BART+Pegasus
BERTScore (prcn)	0.5618
BERTScore (rcl)	0.5529
BERTScore F1	0.5572
BLEU	0.0045

Table 6: Scores for Snorkel + LR classification

Perspective	BLEU	BERTScore F1
Information	0.0066	0.8348
Cause	0.0192	0.8472
Suggestion	0.0072	0.8415
Experience	0.0131	0.8414
Question	0.01	0.8407

Table 7: Perspective wise scores for Bart for SVM+Snorkel+ZeroShot classification

Perspective	BLEU	BERTScore F1
Information	0.0048	0.8317
Cause	0.0053	0.8430
Suggestion	0.0056	0.8397
Experience	0.0042	0.8331
Question	0.01	0.8354

Table 8: Perspective wise scores for Bart fewshot for SVM+Snorkel+ZeroShot classification

9 Observations and Analysis

9.1 Model Performance Comparison

The comparative analysis across models BART, T5, and the stacked ensemble reveals that:

- **BART** outperforms the other models in BERTScore Precision (0.9024), Recall (0.8800), and F1-score (0.8907), indicating stronger semantic alignment between generated and reference summaries.
- **T5** performs the lowest across all BERTScore metrics, with F1-score of 0.8632, BLEU of 0.0363, and METEOR of 0.1856.
- The **stacked model** shows improved performance over T5 and is close to BART. It acts as a balance between precision and recall, with a BERTScore F1 of 0.8815.
- BLEU and METEOR scores for all models are relatively low, which may be attributed to the abstractive nature of the generated summaries that vary in surface form but preserve meaning.

9.2 Perspective-wise Performance

From the analysis across different perspectives for FlanT5+Pegasus model, the model shows high performance in:

- **Cause** (F1 = 0.8666) and **Information** (F1 = 0.8662), likely due to more consistent and structured language in these categories.
- **Suggestion** and **Experience** show slightly lower performance (F1 = 0.8582 and 0.8487 respectively), which may result from subjective variation in language.
- The **Question** perspective has the lowest F1-score (0.8395), possibly due to fewer training examples (only 64) and higher linguistic variation.

9.3 General Insights

- BERTScore metrics consistently favor models like BART due to their semantic accuracy, highlighting the importance of meaning over exact phrasing.
- Data imbalance (as seen in the perspective distribution) significantly impacts performance, especially for underrepresented categories like *Question*.
- The ensemble approach (stacking) offers a robust compromise in performance, suggesting model complementarities can improve generalization.

9.4 Weakly Supervised Perspective-Aware Summarization

- The performance results show that both the classification strategy and the summarization approach significantly affect output quality.
- **Overall Performance (Table 6):** The BART+Pegasus pipeline using Snorkel + Logistic Regression (LR) achieves a BERTScore F1 of 0.5572 and BLEU of 0.0045, reflecting limited overlap with human-written summaries due to weak perspective separation from LR.

- **Perspective-Wise Performance – BART (Table 7):** With SVM + Snorkel + Zero-Shot classification and BART summarization, Cause (0.8472) and Suggestion (0.8415) achieved the highest BERTScore F1. The Information perspective scored lowest in BLEU (0.0066), indicating difficulty in maintaining factual content, while Question and Experience showed moderate performance.

- **Perspective-Wise Performance – BART Few-Shot (Table 8):** Few-shot BART guided by semantically similar examples maintains strong BERTScores for Suggestion (0.8397) and Cause (0.8430), though slightly lower than standard BART. BLEU scores remain low (max 0.01), suggesting few-shot improves semantic meaning (BERTScore) but not lexical overlap (BLEU).

- **Key Takeaways:**
  - Weak supervision (Snorkel + ZSL + SVM) improves classification and enhances summary quality.
  - Few-shot prompting adds semantic richness, especially in perspectives like Suggestion and Cause.
  - Low BLEU scores highlight the need for improved surface realization or hybrid summarization techniques.

10 Additional Files

Dataset Files Used in the Project

The following dataset files were used during this project:

- `train.json` – Contains the training data with questions, multiple answers, and corresponding perspective labels (CAUSE, INFORMATION, SUGGESTION, etc.).

- `valid.json` – Validation set used to evaluate model performance during training.

- `test.json` – Test set used for generating final predictions and analyzing model behavior.

Trained Model Files

The following pre-trained model files are available for download:

- **Model Fine-tuned of Flan-t5 with LoRA**: `https://drive.google.com/file/d/1B7Y0v7PilShiwwZpYC9gqfX-c6dW5LeK/view?usp=drive_link`

- **Model Fine-tuned of BART**: `https://drive.google.com/file/d/1gcOZbf_eemWJDFbYhMnZTkGXUgcB2cNu/view?usp=drive_link`

Plasma: `https://drive.google.com/drive/folders/1fSkgWWQRqOLh9H4O-YfxwzM3rs7baTNH`

11 Discussions and Future Work

Our experiments demonstrated that combining models like fine-tuned Flan-T5 with LoRA and BART in a stacked ensemble leads to more fluent and relevant summaries for perspective-based healthcare questions. The use of LoRA significantly reduced training cost while maintaining competitive performance. Additionally, the weakly supervised approach using Snorkel and sentence-level classification helped in handling unlabeled data and supported summarization even when reference summaries were not available.

However, there are still challenges. The model sometimes struggles to fully separate perspectives when multiple types of information are present in answers. Pegasus, while strong at abstraction, occasionally generated less factually grounded content. The ensemble method, though effective, relies on manual selection between outputs and lacks an automated scoring mechanism. In the future, we aim to enhance our model by exploring reinforcement learning or reward-based fine-tuning strategies that align the generated summaries more closely with human preferences. This could improve the overall quality and perspective alignment of the outputs. Additionally, collecting or annotating more high-quality perspective-labeled data would significantly benefit model training and generalization, especially for underrepresented perspectives like Experience or Question.

References

[1] No perspective, no perception!! Perspective-aware Healthcare Answer Summarization Available at: arxiv.org

[2] FLAN-T5, HuggingFace Available at: Link

[3] SpanBERT Link

[4] Medifact at PerAnsSumm 2025: Leveraging Lightweight Models Available at: Link

Input Text
Summarize from the INFORMATION perspective: Question: what is orgasm? Answers: An orgasm, also known as a sexual climax, is a pleasurable physical, psychological or emotional response to prolonged sexual stimulation...
True Summary
For information purposes, an orgasm is a pleasurable response to sexual stimulation, involving physical, emotional, and sometimes spiritual reactions. Both genders can experience it, but the specific response varies. It's the third stage in the human sexual response cycle, influenced by physical touch, emotional connection, mental intimacy, and complete involvement in the moment. Various factors like thoughts of procreation, power dynamics, and genuine love can contribute to the experience. While having orgasms with someone you love can be great, it doesn't guarantee the long-term success of a relationship.
BART Prediction
For information purposes, an orgasm, also known as a sexual climax, is a pleasurable physical, psychological or emotional response to prolonged sexual stimulation, often accompanied by a notable physiological reaction such as ejaculation, blushing or spasm and may be followed by aftershocks. However, the term "angust" has become obscure in older literature. It is often associated with a similar intensity of emotional excitement, but the exact response varies across genders. Generally speaking, orgasm is the third stage of the human sexual response cycle, which is the currently accepted model of the physiological process of sexual stimulation.
T5 Prediction
The information perspective provides information on the meaning of orgasm.
Stacked Prediction
For information purposes, an orgasm, also known as a sexual climax, is a pleasurable physical, psychological or emotional response to prolonged sexual stimulation, often accompanied by a notable physiological reaction such as ejaculation, blushing or spasm and may be followed by aftershocks. However, the term "angust" has become obscure in older literature. It is often associated with a similar intensity of emotional excitement, but the exact response varies across genders. Generally speaking, orgasm is the third stage of the human sexual response cycle, which is the currently accepted model of the physiological process of sexual stimulation.