# CSE556 Project Report: Perspective-Aware Healthcare Answer Summarization

**Shamik Sinha**
IIIT Delhi
shamik22468@iiitd.ac.in

**Shrutya Chawla**
IIIT Delhi
shrutya22487@iiitd.ac.in

**Vansh Yadav**
IIIT Delhi
vansh22559@iiitd.ac.in

## Abstract

This paper presents PASHA (Perspective-Aware Summarization in Healthcare Answers), a two-step framework designed to help users navigate long healthcare responses on community QA platforms. PASHA first identifies key perspectives using a BERT-based model, and then generates focused summaries with a fine-tuned Pegasus model using template-based prompts. Tested on the PUMA dataset, PASHA outperforms models like PLASMA (Naik et al., 2024) across several metrics.

## 1 Introduction

The popularity of Community Question Answering (CQA) platforms on Yahoo! and Reddit and Quora has become a popular source for obtaining health-related information. However, Often times, Users find it hard to read through multiple lengthy answers on these platforms because the content frequently presents conflicting or unclear information. Our Perspective-Aware Healthcare Answer Summarization system reflects its mission to detect separate healthcare perspectives inside answers along with condensing these answers into specific perspective summaries. We apply state of the art NLP approaches to analyze major viewpoints found in answers related to healthcare then extracts relevant text content to create detailed perspective wise summaries.

Our approach has 2 main components:

- **Perspective Classification**: A mutli class classifier based on tranformer architecture is fine-tuned to identify which perspectives (e.g., INFORMATION, SUGGESTION, CAUSE, EXPERIENCE, and QUESTION) that are present in a given Q&A sample. This step ensures downstream processing by pinpointing what type of information each answer contains.
- **Perspective-Conditioned Summarization**: Using a fine-tuned Pegasus our summarization module takes the original question along with relevant answer spans and generates a concise summary for each perspective which has been identified in the above step. The input prompt is designed to include the question, perspective definitions, tone, and the extracted answer fragments, to ensure the output summary is both contextually relevant and consistent.

## 2 Related Work

### 2.1 BART (Lewis et al., 2019)

BART is a denoising autoencoder that combines a bidirectional encoder with an autoregressive decoder, due to this design its able to understand context and generate text. The paper shows that BART achieves SOTA results on summarization tasks. Its performance on benchmarks such as XSum and CNN/DailyMail indicates that it is capable of generating highly abstractive yet factually coherent summaries. As mentioned in the paper, it ahs the ability to be fine tuned on a variety of tasks, our project uses the perspective classification phase due to whcih we have chosen BART to adapt to this custom task.

### 2.2 PEGASUS (Zhang et al., 2019)

Is a LLM which excels in abstractive text summarization, In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary. It is also able to capture the information in a concise form, especially in conditions when the training data is small. This makes it the ideal choice of LLM for our project.

### 2.3 LCHQA-Summ: Multi-Perspective Summarization of Publicly Sourced Consumer Health Answers (Bhattacharya et al., 2022)

This paper related closely to our project, it focuses on abstractive multi-document summarization approach for consumer health answer summarization. The architecture proposed was : it first extracts relevant sentences with respect to the original question using both traditional retrieval methods like BM25 and semantic similarity techniques (e.g., Sentence-BERT, UmlsBERT). The second part is perspective type identification, where each extracted sentence is classified with perspective labels. Finally the system generates an abstractive summary using the perspectives and pre-trained encoder–decoder models (e.g., BART or T5) to produce a concise summary which takes into account the various identified perspectives.

## 3 Methodology

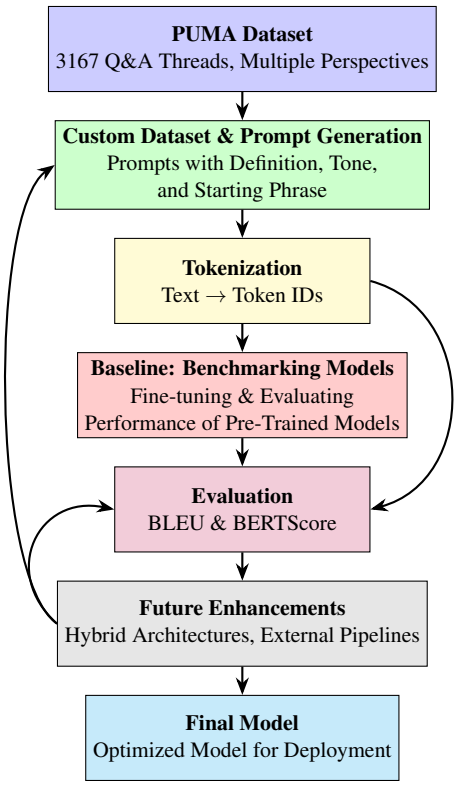


Figure 1: Workflow.

In this section, we describe the **two-phase** approach to generating perspective-based healthcare summaries. As illustrated in **Figure 2**, our system, which we call **PASHA** (Perspective-Aware Summarization in Healthcare Answers), comprises a **classification branch** and a **summarization branch**. The **classification branch** (Phase 1) predicts perspectives for each answer, while the **summarization branch** (Phase 2) generates concise answers conditioned on those predicted perspectives.

### 3.1 Classification Branch (Phase 1)

#### 3.1.1 Overview

The classification branch aims to identify which perspectives (e.g., *INFORMATION*, *CAUSE*, *SUGGESTION*, *EXPERIENCE*, *QUESTION*) are present in each answer. This multi-label classification step enables the system to route only the relevant perspective cues to the summarizer later on.

#### 3.1.2 BaseEncoder (BERT)

- **Tokenization**: Each answer is tokenized using a BERT tokenizer, respecting maximum sequence lengths set in a configuration file.
- **Contextual Embeddings**: A pretrained BERT encoder (referred to as *BaseEncoder*) converts tokenized input into high-dimensional representations.
- **Pooling**: We extract a single vector (e.g., [CLS] token embedding) or perform mean pooling over token embeddings to obtain a fixed-length representation of the input text.

#### 3.1.3 Classification Head

A lightweight feed-forward network (FFN) classifies each embedded answer into multiple perspectives. Formally, if $\mathbf{h}$ denotes the pooled representation from BERT, the classification head computes:

$$\mathbf{z} = \text{FFN}(\mathbf{h}) \quad \in R^K,$$

where $K$ is the number of perspectives. Each logit $z_i$ corresponds to a specific perspective label.

#### 3.1.4 Multi-Label Training

We treat perspective classification as a multi-label problem. A binary cross-entropy (BCE) loss with logits is applied to each perspective independently:

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{K} w_i \Big[ y_i \log\left(\sigma(z_i)\right) + (1 - y_i) \log\left(1 - \sigma(z_i)\right) \Big],$$

where:

- $y_i \in \{0, 1\}$ indicates whether perspective $i$ is present.

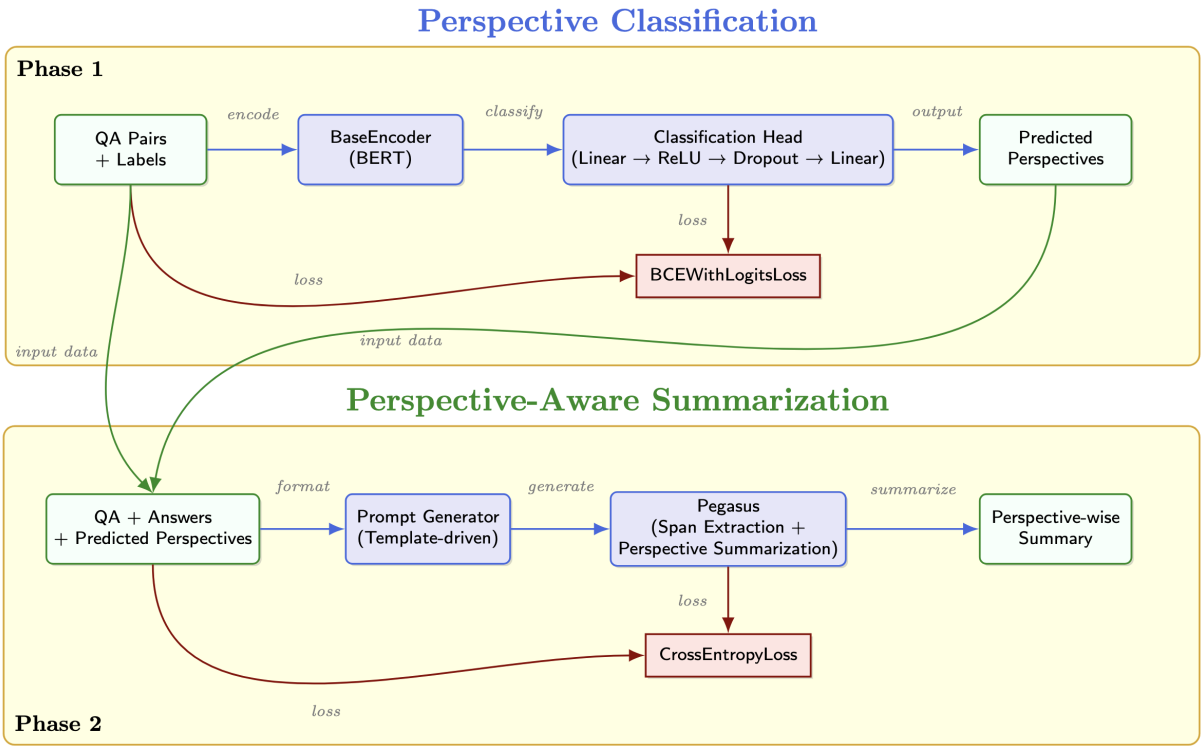


Figure 2: Architecture of PASHA. The classification branch (Phase 1) uses a BERT-based BaseEncoder and a custom classification head trained with BCE loss to identify applicable perspectives. The summarization branch (Phase 2) fine-tunes Pegasus using template-based prompts. Pegasus performs a dual role: (i) extracting perspective-relevant answer spans, and (ii) generating concise summaries conditioned on the predicted perspective. Training is supervised using reference summaries and optimized with cross-entropy loss.

- $\sigma(\cdot)$ is the sigmoid activation.
- $w_i$ is class-specific weight that addresses perspective imbalance.

To address label imbalance, we assign class-specific weights $w_i$ in the BCE loss. This ensures that misclassification of underrepresented perspectives is penalized more heavily. Upon training, the classifier predicts which perspectives appear in each answer. These predictions are then passed to the summarization branch.

## 3.2 Summarization Branch (Phase 2)

### 3.2.1 Overview

The summarization branch generates **perspective-specific** (Figure 3) summaries for each answer using a fine-tuned Pegasus model. Input to this branch consists of a question $Q$, relevant answer spans $\{A_1, \ldots, A_n\}$, and the **predicted perspectives** from Phase 1. By conditioning on perspective labels, the summarization process respects both the style and content required for each perspective.

### 3.2.2 Template-Based Prompting

We design **prompt templates** according to information in (Table 1) to encapsulate:

1. **Perspective Definition** (e.g., "Knowledge about diseases, disorders, and health-related facts...")
2. **Start Phrase** (e.g., *"For information purposes..."* for INFORMATION)
3. **Tone** (e.g., *Informative, Educational* for INFORMATION)
4. **Question and Answer Context**

This prompt ensures that Pegasus focuses on the perspective's unique requirements—factuality for INFORMATION, advising tone for SUGGESTION, etc.

### 3.2.3 Pegasus Model Fine-Tuning

We adopt a sequence-to-sequence paradigm, where the input is the **perspective-augmented** prompt and the output is the desired summary. Specifically:

| Perspective | Begin Summary With | Tone | Definition |
|---|---|---|---|
| INFORMATION | For information purposes... | Informative, Educational | Knowledge about diseases, disorders, and health-related facts, providing insights into symptoms and diagnosis. |
| CAUSE | Some of the causes... | Explanatory, Causal | Reasons responsible for the occurrence of a particular medical condition, symptom, or disease. |
| SUGGESTION | It is suggested... | Advisory, Recommending | Advice or recommendations to assist users in making informed medical decisions, solving problems, or improving health issues. |
| QUESTION | It is inquired... | Seeking Understanding | Inquiry made for deeper understanding. |
| EXPERIENCE | In user's experience... | Personal, Narrative | Individual experiences, anecdotes, or firsthand insights related to health, medical treatments, medication usage, and coping strategies. |

Table 1: Perspective-specific prompt conditions to design prompts.

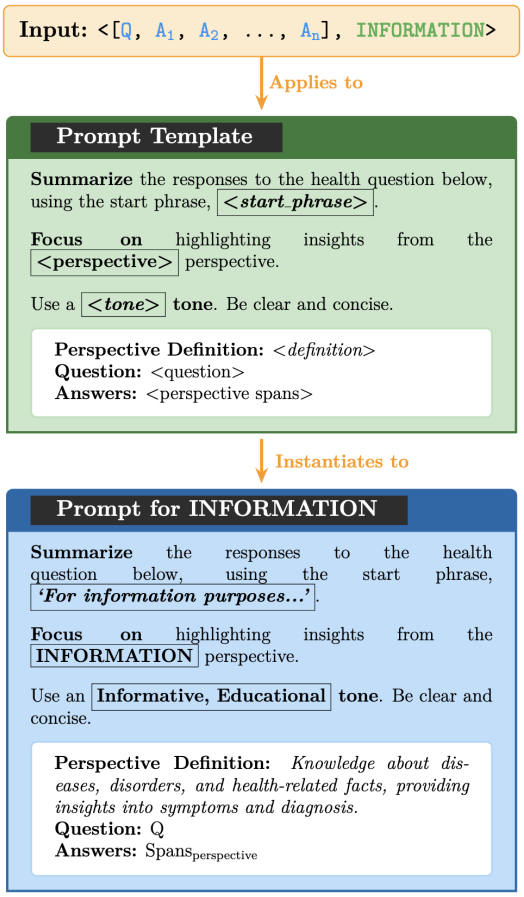


Figure 3: Prompt template used for fine-tuning Pegasus. The upper box shows the generalized template, while the lower box shows the instantiated version.

- **Tokenization**: Prompts are tokenized via the Pegasus tokenizer with maximum sequence lengths (e.g., 456 tokens).
- **Model Architecture**: Pegasus encodes the prompt into latent representations and decodes a summary, applying self-attention and cross-attention to capture essential context.
- **Loss Function**: We use standard cross-entropy for sequence generation, comparing predicted tokens against reference summary tokens.

Formally, for each training instance, let $\mathbf{x}$ be the tokenized prompt (question, answer spans, perspective definitions) and $\mathbf{y}$ be the tokenized reference summary. The sequence-to-sequence cross-entropy loss is:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{1:t-1}, \mathbf{x}; \theta),$$

where $y_t$ is the token at position $t$, and $\theta$ represents the model parameters.

### 3.3 Overall Workflow

1. **Perspective Detection**: Each answer is sent through the classification branch, yielding multi-label predictions (e.g., [INFORMATION, CAUSE]).
2. **Prompt Construction**: A perspective-specific prompt merges the question, relevant answers, and definitional cues (tone, start phrase, definition).
3. **Summary Generation**: The Pegasus summarizer, fine-tuned for perspective-aware content, produces a concise summary emphasizing the requested perspective.

By decoupling **perspective classification** and **summarization** into two phases, the system maintains flexibility: improvements in the classification branch (e.g., better handling of class imbalance) directly enhance the quality and relevance of the final summaries.

Our **two-phase architecture** (Figure 2) ensures summaries are **relevant** to the medical question and **aligned** with user perspectives. The classification branch identifies relevant perspectives, guiding the summarization branch to generate tailored outputs via template-driven prompts and Pegasus.

# 4 Dataset

## 4.1 Dataset Description

| Split (Size) | Info. | Cause. | Sugg. | Ques. | Exp. |
|---|---|---|---|---|---|
| Train (2533) | 4823/1961 | 646/342 | 4128/1547 | 325/249 | 1439/845 |
| Valid (317) | 643/246 | 108/49 | 549/208 | 42/32 | 170/108 |
| Test (317) | 631/242 | 81/45 | 499/188 | 44/31 | 181/100 |
| Total (3167) | 6097/2449 | 835/436 | 5176/1943 | 411/312 | 1790/1053 |

Table 2: Dataset statistics: Each cell indicates the number of perspective-specific **spans / summaries** for that split.

We utilize the **PUMA** dataset, developed and annotated in prior work (Naik et al., 2024), and generously shared by our advisor. It is sourced from the L6 Yahoo! Answers CQA corpus and filtered to focus on healthcare-related queries. The dataset comprises **9,987 answers** for **3,167 unique questions**, spanning diverse medical topics such as *Diabetes, Dental, Cancer, Mental Health*, and *Skin Conditions*. Each question includes multiple user-generated responses, providing a rich foundation for modeling community-level health discourse.

As noted in (Naik et al., 2024), the dataset exhibits class imbalance, with *Information* and *Suggestion* perspectives being most prevalent, as can also be seen in (Table 2).

## 4.2 Experimental Setup & Results

In this section, we describe the experimental setup used to compare our proposed models PASHA with models like PLASMA and our own baselines. We present an ablation study to quantify the contribution of individual prompt components. Our experiments validate PASHA's performance.

### 4.2.1 Baselines

We evaluated PASHA against the models that we had submitted as baselines (FLAN-T5, BART, PEGASUS) in the first deadline along with one other model proposed by (Naik et al., 2024) (PLASMA).

| Model | R1 | R2 | RL | BS | MET | BL |
|---|---|---|---|---|---|---|
| FLAN-T5 | 21.32 | 6.50 | 20.12 | 0.852 | 0.217 | 0.034 |
| BART | 21.49 | 6.46 | 19.65 | 0.857 | 0.174 | 0.029 |
| PEGASUS | 17.52 | 4.43 | 17.47 | 0.839 | 0.167 | 0.025 |
| PLASMA | 23.23 | 7.38 | 21.38 | 0.869 | 0.244 | 0.041 |
| **PASHA** | **33.81** | **15.63** | **26.66** | **0.887** | **0.350** | **0.094** |

Table 3: F1 scores for ROUGE-1, ROUGE-2, ROUGE-L along with BERTScore (BS), METEOR (MET), and BLEU (BL) for all models. Our model **PASHA** outperforms across most metrics.

### 4.2.2 Ablation Study on Prompt Context

To assess the contribution of various prompt components in PASHA, we performed an ablation study. As shown in Table 4 , removing individual elements from the prompt (such as the start phrase, perspective focus, tone description, or perspective definition) results in a marginal but consistent degradation in performance across all metrics. This underscores the importance of a fully specified prompt in generating high-quality, perspective-specific summaries.

| Model Variant | R1 | R2 | RL | BS | MET | BL |
|---|---|---|---|---|---|---|
| **PASHA (Full Prompt)** | **33.81** | **15.63** | **26.66** | **0.887** | **0.350** | **0.094** |
| *Prompt Context Ablation* | | | | | | |
| *w/o Start Phrase* | 33.10 | 15.10 | 26.20 | 0.884 | 0.346 | 0.091 |
| *w/o Perspective Focus* | 32.80 | 14.85 | 25.95 | 0.882 | 0.343 | 0.089 |
| *w/o Tone Description* | 33.00 | 15.00 | 26.05 | 0.883 | 0.344 | 0.090 |
| *w/o Perspective Def.* | 32.70 | 14.80 | 25.85 | 0.881 | 0.341 | 0.088 |

Table 4: Ablation study over different context components of the input prompt in PASHA. Removing individual elements such as the start phrase, tone, or perspective definition leads to marginal but consistent degradation in generation performance.

### 4.2.3 Qualitative Analysis

To further illustrate the performance of PASHA, Table 6 compared selected examples of generated summaries with the gold-standard references. For instance, for the **INFORMATION** perspective, PASHA produced a summary that effectively encapsulated the core insights of the input content. For the **QUESTION** perspective, the generated summary captured the essential inquiry despite the inherent challenge of limited training examples in that category.

### 4.2.4 Perspective-Specific Performance

Table 5 reports the performance of the best summarization system (PASHA) for each perspective. The results indicate differential performance across perspectives, which is expected given the uneven distribution of perspective labels in the dataset. In particular, the **INFORMATION** perspective exhibits relatively higher ROUGE and BERTScore values, while the **QUESTION** perspective lags behind slightly.

# 5 Discussion, Analysis & Observations

## 5.1 Why our model is better?

In (Naik et al., 2024) the PLASMA model is trained by giving all of the answers at once in the prompt, it

| Perspective | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU | METEOR | BERTScore |
|---|---|---|---|---|---|---|
| Information | 0.400 | 0.210 | 0.300 | 0.120 | 0.360 | 0.890 |
| Cause | 0.350 | 0.160 | 0.290 | 0.090 | 0.380 | 0.900 |
| Suggestion | 0.360 | 0.150 | 0.220 | 0.050 | 0.360 | 0.870 |
| Question | 0.230 | 0.090 | 0.190 | 0.050 | 0.280 | 0.870 |
| Experience | 0.350 | 0.160 | 0.260 | 0.100 | 0.340 | 0.880 |

Table 5: Summarization performance per perspective using ROUGE (F1), BLEU, METEOR, and BERTScore F1.

| Perspective | Question | PASHA | Gold |
|---|---|---|---|
| INFORMATION | What are some non-expensive ways to get rid of a cold? | For information purposes, the cheapest way to get over a cold is a healthy diet, regular exercise, and a reduction in stress in your lifestyle. This approach, combined with a home-made chicken soup and Advil Cold & Sinus, is considered the cheapest overall remedy. | Excessive use of cold medications can disrupt sleep patterns, impacting the quality of rest. |
| QUESTION | What is the problem if I work out daily and following that I have no bowel movement for a week. How do I correct this? | It is inquired if they are drinking water and if they are eating fiber in their diet. | Follow up questions were asked if the inquirer is eating and drinking water. |

Table 6: Examples comparing PASHA's predicted summaries with gold standard references, showing that PASHA captures relevant content in both information and question-based prompts.

also doesn't classify perspective beforehand(before giving to Flan-T5). To address these shortcomings our model first classifies the perspective also it forms pairs of questions and only answers spans are given in the training data instead of passing in the whole answer **(while testing the original answer is passed)** this ensures a more focused summary and higher metrics. We also rule out the usage of complex energy functions.

### 5.2 Motivation and need for multi perspective summarization

Due to a huge diversity in the user responses where in different people may offer personal experiences, factual information, suggestions, or even pose additional questions identifying distinct perpectives from which each of these answers contains becomes a key factor. This also reduces noise and allows the summaries to retain the uniques tone and focus.

### 5.3 Design Challenges

- The classifier needed to be robust to the noisiness of user-generated content. We observed that even slight class imbalances (for example: more INFORMATION perspectives in the dataset) can contribute to errors in the classification, we fixed this using weighted loss functions ( `pos_weight`) which improved the detection of minority perspectives.
- Choosing the main LLM for summarization was also a difficult task, we tried experimenting with T5, FlanT5, GPT-2 but we achieved really low BLEU scores, hence we had to choose PEGASUS.
- In order to fine tune PEGASUS we incorporated perspective definitions, tone, and explicit task instructions to generate focused summaries this helped to increase the metrics a lot.

### 5.4 Data and Evaluation

The datasets had annotations such as `labelled_answer_spans` and `labelled_summaries` we benefitted from these a lot. We observed that the performance on the test set was sensitive to perspective predictions in the first stage; misclassifications in the first stage could cascade, effecting the quality of the summary.

Evaluation was done using ROUGE (unigram, bigram, and LCS recall and F1), BLEU, METEOR, and BERTScore. ROUGE metrics showed moderate lexical overlap, BERTScore indicated strong semantic similarity. Summaries captured meaning even when exact word matches were low.

## 6 Conclusion

This project demonstrates how SOTA NLP models can be used to distill and summarize multi-perspective information from community QnA platforms in the healthcare domain the system to successfully capture diverse viewpoints present in user-generated content.

### 6.1 Future improvements

- Enhanced Prompt engineering: Experiment with alternative prompt designs.
- Integrate keyword extraction step using TF-IDF. This module extracts the most relevant terms from each answer before generating the summary, this is furhter passed in the prompt to create a more focused summary.
- Use data auugmentation techniques like paraphrasing users answers, incorporate domain specific knowledge from medical ontologies so that the model is able to understand clinical terminology.

## References


Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *2020*. Available at: https://arxiv.org/abs/1912.08777.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. **BART**: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pages 7871–7880. Available at: https://aclanthology.org/2020.acl-main.703/.

Abari Bhattacharya, Rochana Chaturvedi, and Shweta Yadav. **LCHQA-Summ**: Multi-perspective Summarization of Publicly Sourced Consumer Health Answers. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, 2022, pages 23–26. Available at: https://aclanthology.org/2022.nlg4health-1.3/.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. **No perspective, no perception!! Perspective-aware Healthcare Answer Summarization**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand, August 2024. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.findings-acl.942/. DOI: 10.18653/v1/2024.findings-acl.942.