Perspective-aware Summarization for Healthcare QA

Shaikh Mohd Areeb
MT24111
areeb24111@iiitd.ac.in

Abstract

Healthcare Community Question-Answering (CQA) forums help people get information on medical topics in an easy and open way. Users turn to these forums to share personal health issues, seek advice, and understand medical conditions better. However, answers are often varied and go off-topic, making it hard to find useful information. This creates a need for good summaries of the discussions. Most existing work focuses on general topics and ignores the different viewpoints in responses. This project works on a task: summarizing answers by grouping them based on their perspectives. We have used a dataset (PUMA) with over 3,000 threads and over 6,000 such summaries. We have used a model that uses prompts to generate perspective-based summaries. It includes a special loss function and tuning method to capture the healthcare-specific viewpoints.

1 Introduction

Community Question Answering (CQA) platforms like Quora, Reddit, and Yahoo! Answers have become popular sources for users to seek and share medical advice. In forums like Reddit's r/AskDocs, users post health-related questions and receive a wide range of responses—ranging from personal experiences to factual information, suggestions, and follow-up questions.

While this diversity is valuable, it also poses a challenge: users must sift through numerous, often lengthy, responses to find relevant and reliable information. Summarizing these responses can greatly improve accessibility. However, most existing summarization methods in the medical domain focus on structured texts like clinical reports or doctor-patient dialogues, ignoring the varied perspectives found in CQA answers.

To address this, we are doing a task: perspective-specific answer summarization. Given a medical question and its responses, the goal is to generate concise summaries tailored to specific perspectives such as cause, suggestion, experience, question, and information.

We are using a part of PUMA, a dataset containing 3,167 healthcare CQA threads annotated with over 6,000 perspective-specific summaries. Additionally, we are replicating PLASMA, a controllable summarization model and applying fine-tuning on top of that which combines prompt learning with an energy-based loss function to better enforce perspective alignment. Our model significantly outperforms existing baselines across multiple evaluation metrics, demonstrating the effectiveness of our approach.

2 Related Work

2.1 No perspective, no perception!! Perspective-aware Healthcare Answer Summarization

This paper introduces a novel task of perspective-specific answer summarization in healthcare-related Community Question Answering (CQA) forums. The authors present PUMA, a dataset of 3,167 CQA threads annotated across five perspectives—cause, suggestion, experience, question, and information. They also propose PLASMA, a prompt-based summarization model using prefix tuning and an energy-controlled loss function to enforce perspective alignment. PLASMA outperforms several baselines in both automatic and human evaluations, offering more accurate and coherent perspective-specific summaries. [7]

2.2 Detect and Classify – Joint Span Detection and Classification for Health Outcomes

This paper proposes a joint model called LCAM (Label Context-aware Attention Model) for simultaneously detecting health outcome spans and classifying their types in clinical texts. Unlike previous approaches that treated these tasks separately,

LCAM leverages both word-level and sentence-level attention to capture contextual and semantic relationships. The authors also introduce a label alignment method to merge datasets with different annotation standards, improving model performance. Experiments across multiple benchmark datasets show LCAM outperforms standalone models like BioBERT and SciBERT, especially in low-resource settings. [5]

2.3 AnswerSumm: A Manually-Curated Dataset and Pipeline for Answer Summarization

This paper introduces AnswerSumm, a large-scale, manually curated dataset for multi-perspective answer summarization in Community Question Answering (CQA) forums. It provides a four-step annotation pipeline: selecting relevant sentences, clustering them by perspective, summarizing each cluster, and generating a final fused summary. The authors also propose an automatic data augmentation pipeline and reinforcement learning-based training with semantic and factual rewards. Experimental results show that their methods improve both the quality and factual consistency of summaries compared to baselines. [1]

2.4 LCHQA-Summ: Multi-perspective Summarization of Publicly Sourced Consumer Health Answers

The paper presents a novel approach to summarizing health-related answers from community question answering forums like Yahoo! Answers. The authors introduce a dataset and summarization pipeline that captures multiple perspectives (e.g., cause, treatment, experience) from layperson-contributed answers. Their method includes filtering relevant sentences, identifying the perspective, and generating summaries using models like BART or T5. This work fills a gap by focusing on multi-perspective summarization in the consumer healthcare domain, which is largely underexplored. [6]

2.5 Aspect-oriented Consumer Health Answer Summarization

The paper proposes a novel approach for summarizing health-related responses from community Q&A platforms like Yahoo! Answers. It introduces CHA-Summ, a dataset with human-written summaries categorized into four aspects: Suggestion, Experience, Information, and Question. The authors build a multi-step pipeline involving relevant sentence selection, aspect classification using linguistic and transformer-based models, and aspect-based abstractive summarization using models like BART and Pegasus. Their pipeline significantly improves both the relevance and coverage of summaries, as confirmed by human evaluations. [2]

2.6 CQASumm: Building References for Community Question Answering Summarization Corpora

The paper introduces a large-scale annotated dataset for summarizing threads from Yahoo! Answers. It constructs 100-word reference summaries using filtered best answers, enabling evaluation of summarization algorithms. The authors benchmark traditional multi-document summarization techniques and propose OpinioSumm, a novel method that separates factual and opinion-based content for improved summaries. OpinioSumm outperforms baselines like TextRank by up to 4.6% in ROUGE-1, addressing the unique challenges of noisy, user-generated content in CQA platforms. [4]

2.7 Community Answer Summarization for Multi-Sentence Question with Group L1 Regularization

The paper proposes a novel method to generate complete and non-redundant summaries for community question answering (cQA) threads. It addresses the problem of "incomplete best answers" by segmenting complex questions into subquestions and using a Conditional Random Field (CRF) model with group L1 regularization to identify important answer sentences. The model leverages both textual and non-textual features, along with contextual interactions between sentences. Experiments on Yahoo! Answers show significant improvements over baselines like SVM and standard CRF in ROUGE and F1 scores. [3]

3 Dataset

We utilize a domain-specific dataset tailored for community healthcare answer summarization. The dataset consists of a total of **3,835 samples**, split into three subsets:

- **train.json**: 2,236 samples
- **valid.json**: 959 samples
- **test.json**: 640 samples

Each sample in the dataset is represented as a JSON object containing the following key components:

- `uri`: A unique identifier for the question thread.
- `question`: The original question posted by the user.
- `answers`: A list of community-generated answers.
- `labelled_answer_spans`: Manually annotated answer fragments grouped by summary aspects (e.g., INFORMATION, SUGGESTION).
- `labelled_summaries`: A human-written abstractive summary for each available perspective.
- `raw_text`: The raw concatenated version of all fields for textual reference.

Below is an example illustrating a single annotated data entry:

```
{ 'question': 'what is parkinesonism?',
  'answers': [
    "Parkinson's disease is one of the
most common neurologic disorders...",
    "Parkinsonism describes the common
symptoms of Parkinson's disease..."
],
     'labelled_answer_spans':        {
'INFORMATION':  [{  'txt':   "..."  }]
},
      'labelled_summaries':          {
'INFORMATION_SUMMARY': "..." }
}
```

The dataset is rich in diversity, covering various perspectives such as **INFORMATION**, **SUGGESTION**, **EXPERIENCE**, **CAUSE**, and **QUESTION**, making it suitable for training and evaluating perspective-aware summarization models in the healthcare domain.

4 Methodology

The objective is to generate multi-aspect summaries for complex community health-related questions by using transformer-based models trained on annotated data. The key components of the methodology are: The first line of the file must be

4.1 Problem Definition

In Community Question Answering (CQA), questions are often long, multi-sentence, and expect diverse answers from different users. The "best answer" may miss some perspectives. This project aims to automatically generate comprehensive summaries from all community answers to a given question.

4.2 Data Preparation & Analysis

Three datasets are used: `train.json`, `valid.json`, and `test.json`, sourced from the **PUMA** dataset.

Each question thread includes the following components:

- **question**: a multi-sentence user query.
- **answers**: a list of user-contributed answers.
- **labelled_answer_spans**: spans of answer text annotated as different types (e.g., *fact*, *opinion*).
- **labelled_summaries**: reference summaries labeled by aspect, including:
  - Suggestion
  - Experience
  - Information
  - Question

The dataset is analyzed to compute:

- The frequency of each label type in the answer spans.
- The presence and distribution of different summary types across questions.

This analysis guides the development of a **multi-perspective summarization model**, allowing the system to learn summaries that reflect diverse answer viewpoints.

4.3 Question Segmentation

In community question answering (CQA) platforms, questions are often multi-sentence and encompass multiple sub-questions or intents. Summarizing answers to such questions effectively requires understanding and addressing each of these sub-components. Although explicit question segmentation is not performed in the code, the summarization framework inherently models this aspect by leveraging labeled summaries categorized by

perspective types (e.g., *Suggestion*, *Experience*, *Information*, *Question*).

To emulate question segmentation, each multi-sentence question is treated as a complex query comprising several latent sub-questions. The summarization model is trained to generate a comprehensive summary that covers various subtopics embedded within the question. This is achieved by fine-tuning the BART model on summaries that capture different perspectives of community answers.

This design enables the model to:

- Attend to distinct parts of the input question and answers.
- Generate summaries that reflect multiple viewpoints.
- Avoid redundant or narrowly focused outputs by covering multiple aspects in a unified summary.

By utilizing aspect-labeled summaries and full question context during training, the model implicitly learns to segment and address sub-questions without the need for an explicit segmentation module.

4.4 Model Architectures

In this work, we explore two methodologies for generating abstractive summaries of healthcare-related community Q&A threads using transformer-based models. The first approach utilizes well-established pretrained models like BART, GPT-2, and FLAN-T5. The second approach involves a multi-task learning framework that integrates DeBERTa and BART to improve perspective alignment in summaries.

4.4.1 Method 1: Transformer-Based Summarization Using BART, GPT-2, and FLAN-T5

Objective: The goal of this approach is to generate abstractive summaries using various transformer architectures, each offering a different degree of control over the structure, tone, and content of the generated output.

Architecture Overview: The models used in this approach are BART-base, BART-large, GPT-2, and FLAN-T5 base. BART and FLAN-T5 are encoder-decoder architectures, making them highly effective for sequence-to-sequence tasks. GPT-2, on the other hand, is a decoder-only autoregressive model that generates summaries by continuing a given prompt. BART models use prefix-based control mechanisms, GPT-2 utilizes continuation prompts, and FLAN-T5 leverages instruction-style inputs to guide the summarization.

Input Format: The input consists of concatenated answers from a community thread and the associated question. For BART and FLAN-T5, the input is enhanced with structured phrases or instructions that define the desired summary perspective (e.g., *INFORMATION*, *SUGGESTION*). In the case of GPT-2, the input is a flat prompt that the model learns to continue.

Output: The output is a reference summary extracted from the `labelled_summaries` field in the dataset. For FLAN-T5, the output is expected to start with a specific phrase aligned with the requested perspective and tone.

Loss Function: BART and GPT-2 are optimized using standard cross-entropy loss. FLAN-T5 employs a combination of cross-entropy loss and a custom auxiliary loss that enforces alignment with the expected summary perspective. This includes:

- A perspective classifier loss using RoBERTa.
- A phrase-matching score using ROUGE to compare starter phrases.
- A tone similarity loss computed using cosine similarity between BERT embeddings.

Benefits: This method provides strong flexibility and control:

- BART supports robust, structured generation suitable for noisy, real-world text.
- GPT-2 offers a prompt-based approach to free-form generation.
- FLAN-T5 excels at instruction-following and can produce summaries that reflect both tone and content control.

4.4.2 Method 2: Perspective-Controlled Summarization using Multi-Task DeBERTa-BART Framework

Objective: This method is designed to generate summaries that not only convey key information from healthcare Q&A threads but are also explicitly aligned with a target perspective label such as *SUPPORTIVE*, *NEUTRAL*, or *DOUBTFUL*.

Architecture: This architecture integrates DeBERTa as the encoder and BART as the decoder. DeBERTa provides enhanced semantic understanding, particularly beneficial in clinical contexts. In addition to the summarization head, a classification head is added on top of the encoder for perspective classification, enabling multi-task learning.

Multi-Task Setup: The model jointly learns two tasks:

1. **Summarization:** Sequence-to-sequence generation based on DeBERTa-BART.
2. **Perspective Classification:** Predicting the intended perspective label.

Loss Function: The total training loss is a weighted sum of summarization and classification losses:

$$L_{\text{total}} = L_{\text{summarization}} + \lambda \cdot L_{\text{classification}}$$

where λ is a hyperparameter that controls the trade-off between the two tasks.

Input and Output: The input includes the full QA thread content along with a perspective label. During training, the model learns to produce summaries that are consistent with this label. The output is a fluent, perspective-aligned summary.

Benefits: This method enhances summary factuality and alignment with intended perspectives. Multi-task learning improves generalization, while DeBERTa provides strong contextual encoding tailored to clinical text. The inclusion of explicit supervision for perspective classification enables better control over summary stance.

4.5 Training Approach

4.5.1 Training of BART, GPT-2, and FLAN-T5 Models

The BART, GPT-2, and FLAN-T5 models are fine-tuned using standard supervised learning, where the model parameters are optimized to minimize the cross-entropy loss between the generated tokens and the target summaries.

For FLAN-T5, an additional auxiliary loss is incorporated to enforce stylistic and semantic alignment with the expected summary perspective. This auxiliary component includes a classifier-guided perspective loss, a ROUGE-based phrase match loss for the summary starter, and a cosine similarity loss using BERT embeddings to ensure tonal correctness.

Hyperparameters:

- Optimizer: AdamW
- Learning Rate: 5×10^{-5}
- Epochs: 3–5
- Loss Function:
  - BART and GPT-2: Cross-entropy loss
  - FLAN-T5: Cross-entropy loss + custom auxiliary loss
- Sequence Lengths: Truncation applied up to 1024 tokens for both input and output

4.5.2 Training of Multi-Task DeBERTa-BART Framework

The DeBERTa-BART model is trained in a multi-task setup where both summarization and perspective classification are optimized jointly. A combined loss function is used that integrates the summarization loss and the classification loss. The summarization component uses cross-entropy loss, while the classification head on top of the DeBERTa encoder is trained using categorical cross-entropy.

The total loss function is defined as:

$$L_{\text{total}} = L_{\text{summarization}} + \lambda \cdot L_{\text{classification}}$$

where λ is a weighting factor that balances the two objectives during training.

Hyperparameters:

- Optimizer: AdamW
- Learning Rate: 5×10^{-5}
- Epochs: 3–5
- Loss Function: Combined summarization and classification loss
- Sequence Lengths: Input and output sequences truncated to a maximum of 1024 tokens
- Multi-Task Weight: λ empirically tuned based on validation performance

4.6 Inference & Evaluation

Once training is complete, the fine-tuned model is used to generate summaries on the test set. These predictions are then compared against human-written reference summaries to evaluate the model's effectiveness.

Evaluation Metrics

We employ a comprehensive set of automatic evaluation metrics to assess the quality of the generated summaries:

- **ROUGE (ROUGE-1, ROUGE-2, ROUGE-L):**
  - **ROUGE-1:** Measures unigram (word-level) overlap.
  - **ROUGE-2:** Measures bigram (2-word sequence) overlap.
  - **ROUGE-L:** Captures the longest common subsequence, reflecting fluency and structure.
- **BERTScore:** Evaluates semantic similarity by computing contextual embeddings using a pre-trained BERT model. It captures meaning even when exact words differ.
- **METEOR:** Considers synonym matching, stemming, and paraphrasing in addition to exact word overlap. It provides a more linguistically informed evaluation.
- **BLEU:** A precision-based metric originally developed for machine translation that measures n-gram overlap between the candidate and reference summaries.

These metrics collectively assess both surface-level overlap and deeper semantic correspondence between the generated and reference summaries.

5 Results

Perspective-wise Evaluation Metrics

To analyze the effectiveness of different models in perspective-controlled summarization, we report results across five distinct perspectives: *QUESTION*, *SUGGESTION*, *CAUSE*, *INFORMATION*, and *EXPERIENCE*. The evaluation metrics include ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BERTScore, METEOR, and BLEU.

5.1 BART-base Results

BART-base performs consistently well across all perspectives, with the strongest results on the **QUESTION** and **EXPERIENCE** categories. High BERTScores and ROUGE values indicate that the model effectively captures both semantic and lexical content, particularly in structured or narrative summaries.

5.2 BART-large Results

BART-large performs best on **INFORMATION** and **CAUSE**, showing strength in handling fact-based summaries. It underperforms on **EXPERIENCE** and **QUESTION**, indicating potential difficulties with handling informal or variable narrative structures.

5.3 GPT-2 Results

GPT-2 performs poorly across all perspectives, particularly on ROUGE and BLEU metrics, due to its decoder-only architecture. While BERTScore remains moderate, the low overlap with reference summaries suggests limited usefulness for this task without encoder-side conditioning.

5.4 FLAN-T5 Base Results

Despite low ROUGE scores, FLAN-T5 achieves high BERTScore and METEOR values, especially for **QUESTION**, indicating strong semantic alignment. Its instruction-based design enables better perspective and tone adherence compared to other models.

5.5 DeBERTa-BART Results

The DeBERTa-BART model achieves balanced results, with especially strong performance on the **INFORMATION** and **CAUSE** perspectives. Its multi-task setup helps align summary content with the intended perspective, improving both semantic fidelity and classification-aware generation.

6 Observation and Interpretation

Based on the evaluation metrics and architectural choices, several key observations emerge when comparing the five models used for perspective-controlled healthcare summarization.

6.1 BART Models (BART-base and BART-large):

The BART-base model outperforms BART-large in most metrics across all perspectives. BART-base achieves high ROUGE and BERTScores, especially on the **QUESTION** and **EXPERIENCE** perspectives. Its ability to generalize well on unstructured, community-driven data is attributed to its balanced depth and strong pretraining on sequence-to-sequence tasks. In contrast, BART-large shows a significant drop in performance for narrative and question-driven content, possibly due to overfitting or increased complexity not fully utilized in this domain.

Perspective	R1	R2	RL	BERTScore	METEOR	BLEU
QUESTION	47.52	32.17	47.29	0.908	0.427	0.199
SUGGESTION	35.54	18.33	33.00	0.890	0.266	0.090
CAUSE	40.79	24.52	38.75	0.895	0.357	0.162
INFORMATION	38.51	18.40	36.20	0.892	0.310	0.099
EXPERIENCE	41.11	27.55	39.01	0.905	0.383	0.182

Table 1: BART-base: Perspective-wise evaluation metrics

Perspective	R1	R2	RL	BERTScore	METEOR	BLEU
EXPERIENCE	16.95	5.18	15.29	0.853	0.186	0.037
QUESTION	12.32	2.60	11.48	0.847	0.159	0.021
INFORMATION	33.20	14.80	30.71	0.885	0.291	0.097
SUGGESTION	24.56	7.72	22.17	0.876	0.223	0.048
CAUSE	31.02	12.78	27.98	0.886	0.306	0.088

Table 2: BART-large: Perspective-wise evaluation metrics

Perspective	R1	R2	RL	BERTScore	METEOR	BLEU
EXPERIENCE	3.77	0.00	3.18	0.795	0.081	0.002
QUESTION	2.77	0.09	2.64	0.817	0.091	0.003
INFORMATION	4.73	0.10	4.27	0.795	0.083	0.003
SUGGESTION	4.87	0.13	4.47	0.812	0.085	0.003
CAUSE	3.89	0.22	3.50	0.810	0.111	0.003

Table 3: GPT-2: Perspective-wise evaluation metrics

Perspective	R1	R2	RL	BERTScore	METEOR	BLEU
SUGGESTION	0.32	0.15	0.30	0.875	0.239	0.079
INFORMATION	0.36	0.17	0.34	0.885	0.304	0.100
EXPERIENCE	0.35	0.18	0.33	0.882	0.293	0.104
CAUSE	0.32	0.14	0.28	0.879	0.298	0.086
QUESTION	0.44	0.28	0.42	0.894	0.382	0.172

Table 4: FLAN-T5 Base: Perspective-wise evaluation metrics

Perspective	R1	R2	RL	BERTScore	METEOR	BLEU
INFORMATION	0.44	0.21	0.33	0.893	0.297	0.084
QUESTION	0.16	0.03	0.12	0.849	0.146	0.007
EXPERIENCE	0.20	0.05	0.14	0.848	0.131	0.007
SUGGESTION	0.31	0.13	0.24	0.877	0.191	0.042
CAUSE	0.31	0.15	0.25	0.875	0.264	0.070

Table 5: DeBERTa-BART: Perspective-wise evaluation metrics

6.2 GPT-2 (Decoder-only Architecture):

GPT-2 yields the weakest results across all evaluation metrics because of the absence of the encoder

The absence of an encoder and its reliance on prompt continuation result in poor factual alignment and coherence, especially for summarization tasks requiring structured generation. While it shows acceptable BERTScores for certain perspectives, the extremely low ROUGE and BLEU values indicate poor lexical and grammatical fidelity.

6.3 FLAN-T5 Base (Instruction-tuned Model):

FLAN-T5 base demonstrates strong semantic alignment, as seen in consistently high BERTScores and METEOR scores across perspectives. It excels in the **QUESTION** perspective, suggesting that instruction-tuned prompting enables better control over the tone and structure of summaries. Although ROUGE scores remain low, the model generates summaries that closely reflect the intended meaning and perspective, even when surface-level overlap is limited.

6.4 DeBERTa-BART (Multi-task Framework):

The DeBERTa-BART model effectively balances generation quality and perspective control. Its highest performance is observed on the **INFORMATION** and **CAUSE** perspectives, likely due to the auxiliary classification task guiding the generation toward perspective alignment. The use of a DeBERTa encoder further enhances semantic understanding, which is particularly useful in the medical domain. However, the model struggles with narrative perspectives like **EXPERIENCE** and low-context questions, similar to other models.

6.5 Impact of Perspective Type:

Across all models, the **INFORMATION** and **CAUSE** perspectives consistently yield better performance. These types of summaries are more structured and fact-based, making them easier for models to learn. On the other hand, **SUGGESTION**, **EXPERIENCE**, and **QUESTION** perspectives exhibit greater variability and often rely on subtle cues, making them harder to model accurately.

Models that incorporate either architectural supervision (DeBERTa-BART) or semantic prompts (FLAN-T5) outperform traditional generation models in aligning outputs with target perspectives. While BART-base remains a strong baseline with robust lexical overlap, FLAN-T5 and DeBERTa-BART show promise in perspective-conditioned summarization, particularly when semantic accuracy is more critical than surface similarity. GPT-2, due to its generative nature without structured input understanding, is not well-suited for this task.

7 Novelty

1. **Model Diversity:** We evaluate a range of models—BART-base, BART-large, GPT-2, FLAN-T5, and a custom DeBERTa-BART—offering a broad view of transformer capabilities in perspective-controlled summarization.

2. **DeBERTa-BART Multi-Task Framework:** We introduce a multi-task model that jointly performs summarization and perspective classification, enhancing alignment and control.

3. **Perspective-Aware Learning:** An auxiliary classifier is integrated to guide generation using the target perspective, improving relevance and stance adherence.

4. **Simplified Control:** Our use of instruction prompts and classification-based control provides effective, interpretable, and low-complexity summarization strategies.

8 Limitations

Despite promising results, this study has several limitations. First, there is a mismatch between semantic quality and lexical evaluation—models like FLAN-T5 achieve high BERTScores but low ROUGE scores, revealing shortcomings in traditional evaluation metrics. Additionally, the dataset suffers from perspective imbalance, particularly in underrepresented classes like *EXPERIENCE* and *QUESTION*, which may hinder generalization. All evaluations are automatic; the absence of human judgment limits insights into fluency and factual correctness. Instruction-based models such as FLAN-T5 are also sensitive to prompt phrasing, which can affect consistency. Lastly, although DeBERTa aids clinical understanding, none of the models were pre-trained on large-scale medical corpora, which may constrain domain-specific accuracy.

9 Conclusion

This study evaluated five transformer-based models for perspective-controlled summarization in the healthcare domain. BART-base emerged as a strong baseline with balanced performance across most perspectives. FLAN-T5 base showed high semantic accuracy through instruction tuning, while the DeBERTa-BART framework benefited from multi-task learning for perspective alignment.

In contrast, GPT-2 underperformed due to its lack of encoder conditioning, confirming that decoder-only models are less suitable for structured summarization. Across models, *INFORMATION* and *CAUSE* perspectives were easier to model than narrative or inquiry-based summaries like *EXPERIENCE* and *QUESTION*.

10 Future Work

Future work can focus on enhancing control and generalization through:

- **Perspective-aware pretraining** to improve alignment with stance labels.
- **Adaptive prompting** for models like FLAN-T5 to better handle varying styles.
- **Reinforcement learning** for optimizing generation with semantic feedback.
- **Data augmentation** and larger domain-specific corpora for improved coverage.
- **Human evaluation** to assess real-world usefulness in medical communication.

Model Checkpoints

The fine-tuned model checkpoints used in this study are publicly available and can be accessed via the following Google Drive link: https://drive.google.com/models

References

[1] A. Fabbri, X. Wu, S. Iyer, H. Li, and M. Diab, "AnswerSumm: A Manually-Curated Dataset and Pipeline for Answer Summarization," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2508–2520. doi: 10.18653/v1/2022.naacl-main.180.

[2] R. Chaturvedi, A. Bhattacharya, and S. Yadav, "Aspect-oriented Consumer Health Answer Summarization," May 10, 2024, arXiv: arXiv:2405.06295. doi: 10.48550/arXiv.2405.06295.

[3] W. Chan, X. Zhou, W. Wang, and T.-S. Chua, "Community Answer Summarization for Multi-Sentence Question with Group L1 Regularization," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, and J. C. Park, Eds., Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 582–591. Accessed: Apr. 15, 2025. [Online]. Available: https://aclanthology.org/P12-1061/

[4] T. Chowdhury and T. Chakraborty, "CQA-SUMM: Building References for Community Question Answering Summarization Corpora," in Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, in CODS-COMAD '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 18–26. doi: 10.1145/3297001.3297004.

[5] M. Abaho, D. Bollegala, P. Williamson, and S. Dodd, "Detect and Classify – Joint Span Detection and Classification for Health Outcomes," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8709–8721. doi: 10.18653/v1/2021.emnlp-main.686.

[6] A. Bhattacharya, R. Chaturvedi, and S. Yadav, "LCHQA-Summ: Multi-perspective Summarization of Publicly Sourced Consumer Health Answers," in Proceedings of the First Workshop on Natural Language Generation in Healthcare, E. Krahmer, K. McCoy, and E. Reiter, Eds., Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics, Jul. 2022, pp. 23–26. Accessed: Apr. 15, 2025. [Online]. Available: https://aclanthology.org/2022.nlg4health-1.3/

[7] G. Naik, S. Chandakacherla, S. Yadav, and M. S. Akhtar, "No perspective, no perception!! Perspective-aware Healthcare Answer Summarization," Jun. 13, 2024, arXiv: arXiv:2406.08881. doi: 10.48550/arXiv.2406.08881.