Abdelmalak at PerAnsSumm 2025: Leveraging a Domain-Specific BERT and LLaMA for Perspective-Aware Healthcare Answer Summarization

Abanoub Abdelmalak
University of Bonn, Bonn, Germany
ZB MED - Information Centre for Life Sciences, Cologne, Germany
abdelmalak@zbmed.de

Abstract

The PerAnsSumm Shared Task - CL4Health@NAACL 2025 aims to enhance healthcare community question-answering (CQA) by summarizing diverse user perspectives. It consists of two tasks: identifying and classifying perspective-specific spans (Task A) and generating structured, perspective-specific summaries from question-answer threads (Task B). The dataset used for this task is the PUMA dataset. For Task A, a COVID-Twitter-BERT model pre-trained on COVID-related text from Twitter was employed, improving the model's understanding of relevant vocabulary and context. For Task B, LLaMA was utilized in a prompt-based fashion. The proposed approach achieved 9th place in Task A and 16th place overall, with the best proportional classification F1-score of 0.74.

1 Introduction

Perspective-aware summarization of multiple text sources has recently been studied and used in different applications. One application is the reviews summarization on online shopping websites, where the summarization model can generate a summary that reflects the different perspectives of the reviewers or summarization of different news articles based on the news domain (Liu et al., 2021). Another application is the summarization of question-answering threads in healthcare communities, where the summarization model should be able to generate a summary that reflects the different perspectives of the users. The PerAnsSumm Shared Task - CL4Health@ NAACL 2025 (Agarwal et al., 2025) aims to improve healthcare community question-answering (CQA) by summarizing diverse user perspectives. The goal is to transform the enormous amount of knowledge that is available on these forums into structured information that could be beneficial to others.

The shared task is structured into multiple subtasks to systematically process community question-answering (CQA) threads. The first subtask involves identifying relevant answers and extracting specific spans that convey meaningful information. The second subtask focuses on categorizing these spans into the appropriate perspective classes. Finally, the third subtask entails generating concise summaries for each perspective class, ensuring that the diverse viewpoints present in the discussions are effectively captured. Task A emphasizes the identification and classification of perspective-specific spans, while Task B is dedicated to generating structured summaries. For further details and examples of the dataset, refer to the Appendix A.

The dataset used for this task is PUMA (Naik et al., 2024). It comprises 3,167 CQA threads with around 10,000 answers filtered from the Yahoo! L6 corpus. Each answer in PUMA is annotated with five perspective spans: 'cause', 'suggestion', 'experience', 'question', and 'information'. Based on these perspective and span annotations, summaries are crafted for each identified perspective. These summaries provide concise representations of the underlying perspectives contained within the spans across all answers. Each CQA thread includes up to five perspective-specific summaries.

2 Methodology

The proposed approach frames the first task as a sequence classification problem, where each sentence within an answer is assigned to one of the five predefined perspective classes. For the second task, a summary is generated for each perspective class, utilizing relevant information from the classified sentences. A pre-trained BERT model (Devlin et al., 2019) was fine-tuned on the training dataset to accurately classify sentences based on their perspective labels. For summarization, the LLaMA model (Dubey et al., 2024) was em-

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 428–436

ployed in a zero-shot manner, generating concise summaries for each perspective class without additional fine-tuning.

2.1 Dataset Preparation

The challenge organizers divided the dataset into training, validation, and testing sets. The training and development datasets included additional fields, such as ground truth perspective spans and perspective-specific summaries. However, these fields were not present in the testing dataset.

The training dataset was used to fine-tune the models for Task A. To accomplish this, answers needed to be broken into sub-sequences (sentences) before being fed into the BERT model. A comparison between the spans in the dataset and the actual text revealed inconsistencies. Some spans were incomplete, often missing letters at the beginning or end. This happened because the dataset was annotated based on the exact locations where the perspective appeared in the text. Therefore, Spacy (Honnibal et al., 2020) was not only used to split the text into sentences but the text was also tokenized using the Spacy tokenizer and the tokens were then compared with the tokens in the spans. Out of 22361 sentences, 15027 were found exactly in the labeled spans and 7334 were partially found (partially means that 45% of the larger span matches with the span in question) This criterion was used to filter the sentences that were used in the fine-tuning of the BERT model. Any sentences that didn't match were labeled as negative non-relevant sentences. The reason to restructure the training data is to make sure that the model is trained on the right data that will be used in the testing phase.

For more information on the data set, refer to the original paper by Naik et al. (2024).

2.2 Task A: Sentence Classification

This task is approached as two subtasks: eliminating non-relevant sentences and assigning relevant sentences to their corresponding perspectives. These tasks are modeled as a sequence classification problem, specifically, sentence pair classification, where the model takes the question and sentence as input, separated by the special token [SEP], and the first token [CLS] is used for classification. A set of experiments was conducted using only sentences without the question, resulting in lower training and validation F1-scores. Previous work by Chaturvedi et al. (2024) demonstrated through experimentation that encoder-based models (e.g., BERT, RoBERTa (Liu et al., 2019)) perform better in identifying the relationship between two sentences (in this case, the question and sentence). As a result, the basic BERT model with single sentences as input was used as a baseline.

Considering the nature of the dataset and the target of the models, the first choice model is COVID-Twitter-BERT (Müller et al., 2023) which is published on the Hugging Face model hub (Wolf et al., 2020). The model was originally pre-trained on COVID-related text from Twitter, which matches the same language used in the question-answering forums where people use informal language and also matches the use of health-related symptoms in that case which means it should have a richer dictionary of tokens.

2.2.1 Irrelevant sentences elimination

To achieve this, a COVID-Twitter-BERT model was fine-tuned on both question-sentence pairs and single sentences to classify sentences as relevant or not. The model was fine-tuned on the training dataset, with a sample of relevant sentences selected to balance the dataset (50%). The dataset only contained 2 labels (relevant and irrelevant). It was fine-tuned for 5 epochs with a batch size of 16 and a learning rate of 2e-5. The model was then tested on question-sentence pairs and single sentences, predicting the relevance of sentences in the validation dataset. Table 1 shows that the model achieved an F1-score of 0.74 in the development dataset.

2.2.2 Perspective Classification

A new instance of COVID-Twitter-BERT model was employed for classifying relevant sentences into their corresponding perspective classes. The model was fine-tuned on the training dataset for 5 epochs, using a batch size of 16 and a learning rate of 2e-5. Table 4 showed that it achieved an F1-score of 0.68 on the validation dataset. Additionally, Table 5 shows the performance of the best model on the different classes. Table 2 shows that the distribution of sentences across the perspective categories is imbalanced, which is a common issue in many datasets. To address this, a weighted cross-entropy loss function was utilized to assign more weight to the minority classes, helping to balance the model's sensitivity to different perspectives. The class weights were calculated based on the number of sentences in each class. The weights were calculated using inverse frequency as shown

Model	Precision	Recall	F1 (Macro)
COVID-Twitter-BERT (Single sentences)	0.74	0.73	0.74
BERT-base (Single sentences)	0.74	0.72	0.73
COVID-Twitter-BERT (Pairs)	0.75	0.73	0.74
BERT-base (Pairs)	0.75	0.72	0.73

Table 1: Performance comparison of models on precision, recall, and F1-score for identification of relevant sentences on the validation set to identify irrelevant sentences.

Perspective	No. Sentences
EXPERIENCE	2933
QUESTION	311
CAUSE	677
SUGGESTION	6695
INFORMATION	10723
O	4916

Table 2: Sentence count for each perspective category in the training set after using Spacy's en_core_web_sm model to tokenize each answer into sentences

Class	Weight
EXPERIENCE	7.28
QUESTION	68.61
CAUSE	31.52
SUGGESTION	3.19
INFORMATION	1.99

Table 3: Computed class weights for cross-entropy loss.

in Table 3.

2.3 Task B: Perspective-specific Summarization

The summarization process utilizes the **Meta-LLaMA-3.1-8B-Instruct** (Dubey et al., 2024) model to generate concise, perspective-specific summaries. The model runs with `bfloat16` precision and a maximum of 500 new tokens using the `transformers` pipeline.

A structured prompt ensures that the summary answers a given question while adhering to a predefined category and writing style. The model is instructed to avoid repeating the question or context and to generate a clear, one-line summary that explicitly references the subject. Each category follows a distinct tone: **EXPERIENCE** and **QUESTION** use a third-person perspective, **CAUSE** emphasizes causal reasoning, **SUGGESTION** adopts an advisory tone, and **INFORMATION** maintains a scientific style. The prompt structure ensures high-quality, structured outputs suitable for downstream analysis. A sample prompt can be found in the Appendix A.

The structure of the prompt is as follows:

- **Text Input**: The relevant text and the guiding question are provided to the model.
- **Category**: The prompt specifies the category under which the summary should fall (e.g., EXPERIENCE, QUESTION, CAUSE, SUGGESTION, or INFORMATION).
- **Writing Style**: The summary is generated according to the tone associated with the chosen category:
  - **EXPERIENCE** and **QUESTION**: Use third-person perspective and discuss the subject as users.
  - **CAUSE**: Focuses on causal reasoning and logical connections between events.
  - **SUGGESTION**: Uses an advisory tone, often starting with "It is suggested" when applicable.
  - **INFORMATION**: Presents information in a scientific and informative style.
- **Constraints**: The model is instructed to provide a clear, concise, one-line summary that explicitly references the subject of the question. The summary must not repeat the question or context and should follow the specified writing style.

2.4 Experimental Setup

For fine-tuning the BERT-based models, 2 A40 GPUs and AMD EPYC "Milan" 64-core/128-thread 2.00GHz CPUs were used. More details about it can be found in Appendix A. For the use of the LLaMA model in inference mode, 2 Nvidia A40 48GB GPUs were used.

For each of the BERT-based models, different settings and different datasets were experimented as follows:

Model	Precision	Recall	F1 (Macro)
COVID-Twitter-BERT (Single)	0.63	0.67	0.65
COVID-Twitter-BERT (Pairs)	0.67	0.69	0.68

Table 4: Performance of COVID-Twitter-BERT on precision, recall, and F1-score on the validation set to identify the different perspectives (EXPERIENCE and QUESTION, CAUSE, SUGGESTION and INFORMATION).

Class	F1-Score	Instances
CAUSE	0.42	274
EXPERIENCE	0.70	1248
SUGGESTION	0.76	3044
QUESTION	0.72	175
INFORMATION	0.79	4581
Macro Avg	**0.68**	9322

Table 5: F1-scores and instance count for each class, along with the macro average F1-score for the best-performing model on the validation set.

1. **BERT-base (Single sentences)**: BERT base-uncased fine-tuned on only the sentences from the answers.

2. **COVID-Twitter-BERT (Single sentences)**: COVID-Twitter-BERT fine-tuned on just the sentences from the answers.

3. **BERT-base (Pairs)**: BERT-base-uncased fine-tuned on the question-sentence pairs.

4. **COVID-Twitter-BERT (Pairs)**: COVID-Twitter-BERT fine-tuned on the question-sentence pairs.

This was applied to the irrelevant and relevant models and then was applied to the test data through the evaluation platform. For the model selection criteria and the model loss functions, the macro F1-score was used as the main evaluation metric. The models were fine-tuned using the Adam optimizer with a learning rate of 2e-5 and a batch size of 16. The models were trained for 5 epochs.

There were different data representations used for the fine-tuning as seen from the model names and also for the sentence processing after the classification. The consecutive sentences that were from the same class were merged to form a single sentence. This was done to see if it affects the exact matching results or not. Also, one experiment discarded the part where the sentences were classified as relevant or not to see if it affected the results or not. The results of the experiments are shown in Table 4. For the summarization part, the LLaMA model was used in inference mode. The model was run on the labeled sentences to generate the summaries for each perspective class.

3 Results and Discussion

This section presents the results of the experiments conducted on the PUMA dataset. The results are presented in two parts: the first part is the results of the sentence classification task and the second part is the results of the summarization task.

3.1 Evaluation

The BERT-based models were evaluated on the validation dataset after each epoch. The model with the highest F1-score was selected as the final model. For the sentence classification task, to get over the low classification scores for the minority classes, the weighted cross entropy loss function was used. The weights were calculated based on the number of sentences in each class. The weights were calculated using inverse frequency as shown in Table 3. The test phase For Task A (Span Identification and Classification), evaluation is conducted using the macro-averaged F1-score for classification. Additionally, span identification is assessed through Strict-matching and Proportional-matching methods to measure the accuracy of detected spans.

The evaluation of the summarization component focused on two key aspects: **relevance** and **factuality**. Relevance was assessed using **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** (R1, R2, RL) (Lin, 2004), **Bilingual Evaluation Understudy (BLEU)** (Papineni et al., 2002), **Metric for Evaluation of Translation with Explicit Ordering (METEOR)** (Banerjee and Lavie, 2005), and **BERTScore** (Zhang et al., 2020), measuring lexical and semantic overlap with reference summaries. Factuality was evaluated using **AlignScore** (Zha et al., 2023) and **Summary Consistency (SummaC)** (Laban et al., 2022), ensuring the generated summaries remained faithful to the original content.

Model	Macro F1	Strict F1	Prop. F1	Task A
Covid-Twitter-BERT	0.8859	0.1108	**0.7554**	**0.5931**
BERT	0.8584	0.1068	0.7518	0.5861
Covid-Twitter-BERT + Bingfire[1]	0.8859	0.1108	**0.7554**	**0.5931**
Covid-Twitter-BERT + Merge Sentences	0.8859	**0.1118**	0.7368	0.5872
Skip Irrelevance Step + Covid-Twitter-BERT + Merge Sentences	**0.8931**	0.1081	0.7437	0.5898

Table 6: F1-scores for classification, span matching, and Task A performance. The best values are in **bold**, and the second-best values are underlined.

Model	Task B Relevance	Task B Factuality
Covid-Twitter-BERT	0.2963	**0.2827**
BERT	0.2909	0.2691
Covid-Twitter-BERT + Bingfire	0.2774	0.2403
Covid-Twitter-BERT + Merge Consecutive Sentences	0.2963	**0.2827**
Skip Irrelevance Step + Covid-Twitter-BERT + Merge Consecutive Sentences	**0.3019**	0.2508

Table 7: Evaluation results for Task B: Relevance and Factuality. The best values are in **bold**, and the second-best values are underlined.

3.2 Results

This section reports the results from the challenge's evaluation platform. The results are presented in two parts: the first part is the results of the sentence classification task and the second part is the results of the summarization task.

Table 6 presents the F1-scores for classification, span matching, and Task A performance across several model configurations. The classification macro F1-score evaluates the overall classification performance across all classes, while the strict matching F1 and proportional matching F1 assess the model's ability to correctly identify and match spans at different levels of granularity. The Task A score provides an overall evaluation of the model's performance on the span identification and classification task. In the table, the best values are highlighted in bold, and the second-best values are underlined for easy reference.

The results indicate that the model pre-trained on data more similar to the task's dataset achieved the best overall performance. Additionally, switching the sentence tokenizer from Spacy did not impact the results, as it was only used during testing, not in the fine-tuning phase. Merging sentences did not affect the classification task but slightly improved the overall classification performance. Finally, skipping the relevance task did not enhance the results; in fact, it led to worse overall performance in the test phase.

While only one model and a single prompt were used in the summary generation task, the input text that comes from the first task was the factor that affected the results. The results show that a better classification contributed to a better result overall as shown in Table 7. While adding more sentences through skipping the irrelevance step did not affect the relevance of the summary, it affected the factuality of the summary.

4 Conclusions

The approach presented in this paper, as part of the PerAnsSumm Shared Task - CL4Health@NAACL 2025, aimed to enhance healthcare community question-answering (CQA) by summarizing diverse user perspectives.

A key aspect of the approach was the focus on accurately classifying sentences (parts of answers) into the correct perspectives while eliminating irrelevant text. To achieve this, a specialized BERT model (COVID-Twitter-BERT) was fine-tuned on the training data for each subtask separately. The results demonstrated that the model pre-trained on data more similar to the task's dataset achieved the best overall performance in Task A. The classification model achieved the best results in terms of proportional matching across the challenge, indicating that the data preprocessing for fine-tuning the model to classify the correct perspectives was highly effective. However, the identification of the correct spans was less accurate, even when merging sentences. This suggests that identifying the right sentence boundaries, in line with the dataset's standards, is notably different from the default boundaries applied in common libraries (e.g., SpaCy and Bingfire).

For the second task, which involves using sentences to generate summaries, only one model was tested in inference mode without any fine-tuning. The results demonstrated how the quality of the data from the first task can impact the results of the second. Specifically, better classification contributed to overall better performance in summary generation.

5 Limitations

In this work, the focus was on the classification of the sentences to the correct perspective classes. The results showed that the identification of the correct spans is low which can be highlighted as a limitation of this work. Additionally, due to time and human resources limitations, only one model (LLaMA) was used to generate the summary with few tweaks in the prompt and limited post-processing of the output text.

6 Future work

For future works, it is recommended to add more rules to identify the spans of the text. Or to only fine-tune a model to identify irrelevant parts of the text as a Named-Entity-Recognition task because in some cases it is only one or two words that are discarded which makes it costly in terms of the exact matching. Also, it is recommended to use more models to generate the summaries and to use more prompts to generate the summaries. Additionally, evaluating how fine-tuned summarization models can affect the results.

Acknowledgments

The author gratefully acknowledges the access to the Marvin cluster of the University of Bonn. The author would like to thank the organizers of the PerAnsSumm Shared Task - CL4Health@ NAACL 2025 for providing the dataset and the evaluation platform.

References

Siddhant Agarwal, Md Shad Akhtar, and Shweta Yadav. 2025. Overview of the peranssumm 2025 shared task on perspective-aware healthcare answer summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rochana Chaturvedi, Abari Bhattacharya, and Shweta Yadav. 2024. Aspect-oriented consumer health answer summarization. *arXiv preprint arXiv:2405.06295*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. MultiOpEd: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in artificial intelligence*, 6:1023281.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Dataset Insights

Here are some statistics of the dataset taken from the dataset publication (Naik et al., 2024). Figure 1 shows the examples from the dataset. While Table 8 shows the distribution of the dataset over the different perspectives.

	Information	Cause	Suggestion	Question	Experience
Train (2533)	4823/1961	646/342	4128/1547	325/249	1439/845
Validation (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 8: Statistics of the original dataset (Naik et al., 2024)

A.2 Experiments of the different models combinations

There were different model combinations for Task A that were used but were not worth mentioning in the main body of the paper:

- Single step for Task A: The first experiments used one single model to identify all perspectives and also the irrelevant sentences as an extra class. However, this approach faced many issues due to the imbalance in data. The models are:
  - BERT model to identify all classes and irrelevant classes on pairs of questions and sentences
  - COVID-Twitter-BERT to identify all classes and irrelevant classes on pairs of questions and sentences
- 2-steps for Task A: The adopted approach in this paper was to approach the problem in 2 steps (identifying irrelevant sentences and then classifying the relevant ones from the correct perspectives) For that different formats of the dataset were used:
  - Pairs: The input data instances consist of pairs of questions and sentences.
  - Singles: The input data is only the sentences that should be classified.

  Different instances and models were tested:
  - BERT-base model.
  - COVID-Twitter-BERT
  - Biomed RoBERTa (Gururangan et al., 2020)

  Where the COVID Twitter BERT proved to be superior in terms of results.

A.2.1 Training Parameters

The models were fine-tuned on the training dataset for 5 epochs, using a batch size of 16 and a learning rate of 2e-5.

A.2.2 Hardware

The fine-tuning process was conducted using two Nvidia A40 GPUs and AMD EPYC "Milan" 64-core/128-thread 2.00GHz CPUs. Two Nvidia A40 48GB GPUs were utilized for inference with the LLaMA model.

[1] https://github.com/microsoft/BlingFire

Question: I was just diagnosed with gallstones in my gall bladder I really don't want to have surgery and have been told that there are other ways to get rid of the stones. Suggestions?	
Answer 1	Most gallstones are made of pure cholesterol. You might try a diet with low fat and very low saturated fats. Reducing the saturated fats will lower blood cholesterol and may (I'm not promising anything!) make the stones smaller. Lowering your total fat intake may also help reduce or prevent pain. Gallstones hurt because when you eat fat, the stomach senses it and tells the gallbladder to release some of the gall into the intestines to help digest the fat. If you have stones, that squeezing is about like squeezing a hand full of pointy rocks. However, I've had the surgery, and it really isn't a big deal. There is minimal scarring, ... If you leave the gallstones there, they can get large enough to damage the gallbladder, with the result of a bad infection and that can cause death, which is a very Bad Thing!"
Answer 2	Have you seen a gastroenterologist? They can do a minimally invasive procedure called an ERCP (Endoscopic Retrograde Cholangiopancreatography. An ERCP won't get rid of the stones that are in the gallbladder...just the stones that are stuck in the duct .... They can make a tiny cut in the duct and pull gallstones out with a small balloon. If the stone is too large, they have equipment that will crush the stone so it will pass freely. I had the surgery myself about 10 years ago. It's not as bad as you'd imagine, and you feel much better after it's over. You might still have phantom pain for a while, but it's nowhere near as bad as the pain you started with. A diet high in fat will make gallbladder disease worse, but you can't really get rid of the stones unless they pass naturally or you have them removed, either in surgery or with an ERCP.
Answer 3	The best remedy is surgery. I had surgery to have kidney stones removed. The surgery isn't as bad as you think it may be.
Perspective-based summaries	
Information	Reducing saturated fats may shrink gallstones as they're mostly made of cholesterol. Gallstone pain occurs when the gallbladder squeezes to aid digestion on fat consumption. An ERCP procedure by a gastroenterologist can remove stones stuck in the duct leading to the intestine. This minimally invasive technique involves extracting stones or crushing larger ones for easier passage, but it doesn't eliminate stones within the gallbladder itself.
Cause	Gallstones left untreated can harm the gallbladder, causing severe infection and potentially death.
Suggestion	To eliminate gallstones without surgery, a low-fat diet, particularly low in saturated fats, as it may help reduce pain associated with gallbladder disease. Ultimately, surgical or medical intervention like ERCP may be necessary for complete removal if stones don't pass naturally.
Experience	Multiple people shared their experience of undergoing surgery to remove kidney stones, assuring that the procedure wasn't as daunting as expected. Despite the possibility of post-operative discomfort, the relief from the original pain was significant.
Question	It was asked if the person had seen a gastroenterologist

Figure 1: Example from the dataset to show how the different perspectives are identified (Naik et al., 2024)

A.3 LLaMa Prompt

Figure 3 shows the prompt that was used to generate the perspective-oriented summaries. The prompt follows a structured format, where different placeholders represent key components of the input. Specifically:

- `text`: This refers to the list of sentences associated with a particular perspective. These sentences serve as the content from which the summary is generated.
- `question`: This represents the question that the summary is expected to address. It guides the summarization process by ensuring the generated output remains relevant to the intended query.
- `key`: This corresponds to the perspective class name, which helps differentiate between different viewpoints present in the dataset. By explicitly defining the perspective, the summarization model can tailor its output accordingly.
- `catch_phrase`: This is a perspective-specific command designed to shape the style or focus of the summary. It acts as a guiding phrase that reinforces the perspective's stance or emphasis. Figure 2 shows the different commands according to the corresponding key.

By structuring the prompt in this manner, the model is provided with clear instructions on how to generate summaries that are not only coherent but also aligned with the given perspective. This approach ensures that the summarization process remains consistent and interpretable across different perspectives, ultimately improving the quality of the generated outputs.

```
{"EXPERIENCE": "Use third-person
    perspective and talk about the
    people as users",
 "QUESTION" :"Use third-person
    perspective and talk about the
    people as users",
 "CAUSE" :"Use causality and chain of
    thoughts",
 "SUGGESTION" :"Use Advisory,
    Recommending tone and start by
    **It is suggested** when possible",
 "INFORMATION" :"Use scientific and
    informative tone"}
```

Figure 2: Custom commands to be entered in the summarization generation prompt to adapt the style to the required perspective

```
You are an expert in text analysis. Your task is to summarize the following text
    according to the given category.

### Text:
{text}

### Constraints:
The summary should answer a question regarding: {question}.

###Important: Do NOT repeat the question or the context. Only generate the summary.


### Category:
The summary should follow the {key} category.
### Writing Style:
{catch_phrase}.

### Instructions:
- Ensure that the summary is **one line**
- The summary **must explicitly reference the subject of the question**.
- The summary must not include the question.
- Follow the writing style specified for the given category.
- Ensure the summary is clear, concise, and relevant.
- Generate the summary as a **continuous paragraph** without bullet points.

### Summary:
```

Figure 3: The prompt used to generate perspective-oriented summaries where {text} refers to the list of sentences of one perspective, {question} is the question that the summary should answer, {key} is the perspective class name, and {catch_phrase} is a perspective-specific command.