DataHacks at PerAnsSumm 2025: LoRA-Driven Prompt Engineering for Perspective Aware Span Identification and Summarization

Vansh Nawander
IIIT Hyderabad
vanshnawander@gmail.com

Chaithra Nerella
IIIT Hyderabad
chaithra.nerella@research.iiit.ac.in

Abstract

This paper presents the approach of the DataHacks team in the PerAnsSumm Shared Task at CL4Health 2025, which focuses on perspective-aware summarization of healthcare community question-answering (CQA) forums. Unlike traditional CQA summarization, which relies on the best-voted answer, this task captures diverse perspectives, including 'cause,' 'suggestion,' 'experience,' 'question,' and 'information.' The task is divided into two subtasks: (1) identifying and classifying perspective-specific spans, and (2) generating perspective-specific summaries. We addressed these tasks using Large Language Models (LLM), fine-tuning it with different low-rank adaptation (LoRA) configurations to balance performance and computational efficiency under resource constraints. In addition, we experimented with various prompt strategies and analyzed their impact on performance. Our approach achieved a combined average score of 0.42, demonstrating the effectiveness of fine-tuned LLMs with adaptive LoRA configurations for perspective-aware summarization.

1 Introduction

Community Question Answering (CQA) forums for healthcare care serve as valuable resources for individuals seeking information on illnesses, treatments, therapies, personal experiences, and medical advice. These communities include a number of varied viewpoints, such as factual information, expert advice, personal anecdotes, causal justifications, recommendations, and follow-up questions. Although these platforms provide diverse perspectives, the large number of responses, often containing conflicting points of view, makes it difficult for users to extract clear and reliable information.

A well-structured summary is crucial for enabling users to quickly access relevant information within this complex content. However, traditional summarization models, like RNN-based encoder-decoder architectures, often fail to handle the complexity of CQA discussions. They struggle with capturing multiple viewpoints, handling contradictions, and preserving key information which is present in CQA threads. (Chowdhury et al., 2020).

Recent advancements in summarization techniques have attempted to address these challenges. Perspective-aware summarization models ensure that critical viewpoints are retained (Naik et al., 2024), while inconsistency detection methods such as SummaC use NLI-based approaches to improve factual reliability and coherence in summaries (Laban et al., 2022). Furthermore, CQA-specific summarization corpora have provided high-quality reference summaries to better adapt models to the unique nature of CQA data (Chowdhury and Chakraborty, 2019). Despite these developments, existing methods still struggle to effectively capture the nuanced and sometimes contradictory perspectives present in CQA discussions.

Large Language Models (LLMs) have emerged as powerful tools for text summarization, excelling at processing lengthy contexts and generating coherent summaries (Minaee et al., 2024). However, adapting these models to domain-specific tasks like healthcare CQA remains a challenge due to the high computational costs associated with full fine-tuning. To overcome this limitation, Low-Rank Adaptation (LoRA) (Hu et al., 2022) has gained prominence as an efficient fine-tuning technique that enables LLMs to specialize in specific tasks with minimal parameter updates. By leveraging LoRA, LLMs can be adapted for perspective-aware summarization while significantly reducing computational costs.

The PerAnsSumm Shared Task at CL4Health 2025 (Agarwal et al., 2025) is designed to advance the development of perspective-aware summarization systems for healthcare CQA forums, focusing on two subtasks: (A) Span Identification and Classi-

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 374–379

fication and (B) Perspective-Aware Summary Generation. This problem highlights the necessity for sophisticated methods that can summarize and distinguish between various points of view while preserving the content's cohesion and factual integrity. In our work, we fine-tuned Mistral-7B(Jiang et al., 2023) and analyzed the impact of LoRA ranks and prompting strategies on the performance of both tasks.

2 Dataset

The task included PUMA dataset (Naik et al., 2024), a perspective-aware corpus specifically annotated for medical question-answer pairs. The dataset consists of 3,167 CQA threads with approximately 10,000 answers sourced from Yahoo! L6 corpus. Each answer is annotated with perspective-specific spans across five categories: *experience, information, cause, suggestion and question*. Each data instance has several key components- *Question, Context, Answers, Labelled Answers Spans, Labelled summaries*.

The Question represents the user's inquiry related to a healthcare topic. The Context provides additional background information, which may be empty or contain relevant details to aid in understanding the question. The Answers consist of a list of user-provided responses related to the question. These answers are further enriched with Labelled Answer Spans, which are annotated text segments categorized under the perspective labels. Each span includes the text itself along with its character-level position, enabling precise identification of the perspective within the answer. Additionally, the dataset includes Labelled Summaries, which are perspective-specific summaries that aggregate relevant spans across all answers in a thread. These summaries serve as concise representations of the underlying perspectives, facilitating a comprehensive understanding of the various viewpoints expressed in the data set.

3 Methodology

Our goal was to enhance perspective-aware answer summarization by fine-tuning Mistral-7B using Low-Rank Adaptation (LoRA).We experimented with different LoRA ranks and prompting strategies to assess their impact on performance. Mistral-7B was chosen for its strong language understanding capabilities, efficiency, and ability to generate coherent and contextually rich summaries. Instead of full fine-tuning, we opted for LoRA to preserve model generalization while optimizing computational efficiency making it feasible under resource constraints.

3.1 Data Preprocessing

The dataset provided contained perspective-specific spans and summaries annotated across five categories: *experience, information, cause, suggestion, and question*. To prepare the data for training, we systematically extracted these segments from the original JSON annotations and reformatted them into a structured format.

Each instance in the dataset was converted into a standardized dictionary structure where every category was explicitly represented. For example, even if a response contained only *information* spans, the format ensured that placeholders for other perspectives were included like: {information:[.....], suggestion:[ ], experience:[ ], cause:[ ], question:[ ]}. This transformation allowed uniform processing across all data instances and ensured that the model learned to differentiate between perspectives effectively.

3.2 Prompt engineering

We experimented with various prompt strategies and documented the results of two key variations. For Task A, the model was instructed to generate spans for each perspective label. For Task B, the same prompt structure was used, but the model was asked to generate summaries instead of extracting spans. In the first approach, the prompt presented the question, context, and answer as a single block of text without explicitly differentiating them. While this approach produced reasonable outputs, it often resulted in vague or incomplete summaries, as the model struggled to clearly distinguish between different components. Additionally, the absence of clear section markers sometimes led to misclassification in span extraction and inconsistencies in summaries.

To address these issues, we refined the prompt by explicitly separating the question, context, and answer into distinct sections. This structured approach improved the model's ability to identify relationships between different components, leading to more accurate perspective classification. It also minimized errors caused by misinterpretation and ensured greater consistency in the generated outputs. A comparative analysis of both strategies revealed that the structured prompt method signifi-

cantly improved both the accuracy and coherence of the summaries, making it the preferred choice for our experiments. The prompts are detailed in the Appendix section.

3.3 Evaluation Metrics

The submissions were evaluated across different metrics for each task. Task A (span identification and classification) was evaluated across 3 main metrics *F1 score (Macro F1, Weighted F1), Strict Matching, Proportional Matching*. Macro and weighted F1 scores can assess the classification performance, ensuring a balanced evaluation across all the classes including minority ones. Strict Matching and Proportional Matching metrics for precision, recall and F1 score were used to evaluate span identification accuracy. Strict Matching checks if the span boundaries match exactly, while Proportional Matching allows for partial overlaps, making the evaluation more flexible.

Task B (Perspective-Specific Summarization) was evaluated across two metrics- *Relevance and Factuality*. Relevance was assessed using *ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) (Lin, 2004), BERTScore (Zhang et al., 2020), METEOR (Banerjee and Lavie, 2005), and BLEU (Papineni et al., 2002). ROUGE* measures lexical overlap by comparing n-grams between the generated and reference summaries. *BERTScore* goes beyond surface-level overlap by using contextual embeddings to evaluate semantic similarity. *METEOR* considers synonymy and stemming to better capture meaning, while *BLEU* focuses on matching n-grams but is more sensitive to exact word choice. Factuality was assessed using *AlignScore(Zha et al., 2023), SummaC* ensuring that summaries remained factually consistent and aligned with source content. This multifaceted evaluation approach allowed us to thoroughly analyze the effectiveness and reliability of our models in capturing diverse perspectives and generating high-quality summaries.

4 Experiments and Results

4.1 Experimental Setup

We fine-tuned the Mistral-7B model using Low-Rank Adaptation (LoRA) to optimize computational efficiency while preserving model generalization. LoRA enables efficient adaptation by injecting low-rank matrices into key transformer layers, significantly reducing the number of trainable parameters while maintaining model performance. To systematically analyze the impact of LoRA configurations, we experimented with different LoRA ranks—64, 128, and 256—while keeping the LoRA scaling factor (lora_alpha) fixed at 128.

The model was fine-tuned for five epochs, with a per-device batch size of four and gradient accumulation set to two, resulting in an effective batch size of eight. We used the AdamW optimizer with fused updates, a learning rate of 2e-4, and a linear scheduler without warm-up. Mixed precision training was enabled, utilizing FP16 or BF16 (based on hardware support) to further optimize memory usage and training speed. Training was monitored using epoch-wise evaluation, with key metrics tracked via Weights & Biases (W&B). The best-performing model was selected based on evaluation results, with a checkpoint limit of six to manage storage efficiently.

4.2 Results

Table 1(a) presents the performance metrics for Task A. Among the different LoRA configurations, the refined prompt (RP) with a LoRA rank of 256 achieved the highest overall performance, with a Task A score of 0.5441, outperforming initial prompt (IP) configurations with a small margin. The RP (256) setting also led to the best Strict F1 and Proposition F1, indicating improved precision and recall in structured prediction. Among the IP configurations, LoRA rank 256 performed best, followed by rank 128 and the lowest performance was observed in IP (64).

Table 1(b) reports the evaluation metrics for Task B. Similar to Task A, RP (256) achieved the highest scores, particularly in TASK B Factuality (0.3663) and TASK B Relevance (0.3504). While IP (256) demonstrated competitive performance (Factuality = 0.3521), RP (256) still outperformed all other configurations. The improvements in factuality and relevance suggest that refined prompts help generate more accurate and contextually appropriate responses, making them particularly effective for knowledge-based tasks like summarization.

Table 2 consolidates the performance across both tasks. The highest combined average score of 0.4203 was obtained using RP (256). The results indicate that increasing the LoRA rank improves performance by a small margin, with LoRA rank 256 yielding the best results. The refined prompt (RP) strategy outperformed initial prompts (IP) for the combined average. However, their effect across individual metrics was not consistent.

(a) Performance Metrics for Task A

LoRA rank	Macro F1	Weighted F1	Strict P	Strict R	Strict F1	Prop P	Prop R	Prop F1	Task_A
IP (64)	0.8382	0.8778	0.1148	0.0857	0.0981	0.4542	0.6318	0.5285	0.5015
IP (128)	**0.8787**	**0.9181**	0.0156	0.0495	0.0238	**0.5422**	0.6301	**0.5829**	0.5082
IP (256)	0.8689	0.9009	0.131	0.1048	0.1164	0.4546	0.6659	0.5403	0.5192
RP (256)	0.8635	0.9044	**0.1599**	**0.1352**	**0.1465**	0.5149	**0.6678**	0.5815	**0.5441**

(b) Performance Metrics for Task B

LoRA rank	ROUGE1	ROUGE2	ROUGEL	BERT	METEOR	BLEU	TASK B Relevance	Align	SummaC	TASK B Factuality
IP (64)	0.3671	0.1607	0.3345	0.7849	0.3386	0.1052	0.3485	0.4002	0.2661	0.3331
IP (128)	**0.3787**	0.1679	**0.3428**	**0.8041**	0.3406	0.1072	0.3569	0.2846	**0.4302**	0.3574
IP (256)	0.3778	**0.1747**	0.343	0.7927	**0.3452**	0.1092	**0.3571**	0.4211	0.2831	0.3521
RP (256)	0.3708	0.1683	0.3365	0.7762	0.3391	**0.1116**	0.3504	**0.4427**	0.2899	**0.3663**

Table 1: Performance Metrics with different LoRA ranks (in bracket) and IP - Initial Prompt, RP - Refined Prompt

LoRA rank	Combined Average
IP (64)	0.3944
IP (128)	0.4075
IP (256)	0.4095
RP (256)	**0.4203**

Table 2: Task A + B Combined Average Scores

5 Conclusion

This study shows that combining Low-Rank Adaptation (LoRA) with well-structured prompts can significantly improve perspective-aware summarization in healthcare Q&A forums. By fine-tuning the Mistral-7B model, we captured different perspectives—cause, suggestion, experience, question, and information—while keeping the approach efficient. LoRA rank played a key role, with higher ranks generally improving precision and recall, though the gains leveled off at a certain point. The refined prompt strategy also boosted classification accuracy, proving that clear guidance helps models generate better responses. These results highlight the importance of both efficient fine-tuning and good prompt design in building accurate and context-aware summarization systems for healthcare applications.

6 Limitations

While RP (256) achieves the highest combined score, no single configuration is best across all metrics. For instance, IP (128) performs better in factuality compared to RP(256) (SummaC: 0.4302 vs. 0.2899), indicating trade-offs between factuality and summarization quality. Although higher ranks (256) generally yield better combined results, IP (128) achieves comparable or better results in some areas (e.g., ROUGE1, SummaC, BERT), indicating that simply increasing LoRA rank does not guarantee uniform improvement.Despite using LoRA to reduce computational costs, fine-tuning large models like Mistral-7B is still computationally intensive, which may not be accessible to all researchers. Since the model is fine-tuned specifically on healthcare CQA data, this might limit its generalizability to other domains or even different types of healthcare texts outside the utilized dataset.

7 Future Work

Using Mistral with prompt variations and LoRA ranks for the tasks shows promised results. Future research could focus on creating more robust prompt templates that generalize across tasks and developing adaptive methods to adjust LoRA ranks based on task complexity.Further, ablation studies comparing different fine-tuning methods, including other parameter-efficient techniques, could provide deeper insights. Expanding prompt strategies for diverse domains, integrating multi-modal data, and analyzing the trade-offs between prompt refinement and model performance are also promising directions. Analyzing the model through a mechanistic interpretability lens might provide more insights into its decision-making process, clarifying things that remain unclear in our analysis.

Acknowledgments

The authors thank Dr Manish Shrivastsava, Datazoic, and LTRC lab IIIT Hyderabad for their constant support and encouragement.

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.

Tanya Chowdhury and Tanmoy Chakraborty. 2019. *CQASumm: Building References for Community Question Answering Summarization Corpora*. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 18–26.

Tanya Chowdhury, Sachin Kumar, and Tanmoy Chakraborty. 2020. *Neural Abstractive Summarization with Structural Attention*. *arXiv preprint arXiv:2004.09739*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Philippe Laban, Tobias Schnabel, Paul 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.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

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 and Virtual Meeting. 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.

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.

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*.

Appendix

Prompts

The prompts used in our experiments are shown in Figures 1,2,3,4

Initial Prompt for Spans

Below is the given input text. Extract the spans for each of the following labels: EXPERIENCE, INFORMATION, CAUSE, SUGGESTION, QUESTION.
Input:
{input_text}
Response:
{{"EXPERIENCE": [], "INFORMATION": [], "CAUSE": [], "SUGGESTION": [], "QUESTION": []}}

Figure 1: The initial prompt used for extracting spans from input text for different categories.

Initial Prompt for Summary

Below is the given input text. Summarize the input text for each of the following labels: EXPERIENCE, INFORMATION, CAUSE, SUGGESTION, QUESTION.
Input:
{input_text}
Response:
{{"EXPERIENCE": "", "INFORMATION": "", "CAUSE": "", "SUGGESTION": "", "QUESTION": ""}}

Figure 2: The initial prompt used for generating summaries from input text for different categories.

Refined Prompt for Spans

Below is the given Question, Context, and Answer. Identify the spans in the user answers that reflect a particular perspective and classify each span to the correct perspective among: EXPERIENCE, INFORMATION, CAUSE, SUGGESTION, QUESTION. Output the results in JSON format.
Question:
{question}
Context:
{context}
Answer:
{answer}
Spans:
{{"EXPERIENCE": [], "INFORMATION": [], "CAUSE": [], "SUGGESTION": [], "QUESTION": []}}

Figure 3: The refined prompt used for identifying and classifying perspective spans in user answers.

Refined Prompt for Summary

Below is the given Question, Context, and Answer. Generate a summary that represents the underlying perspective for each of the following perspectives: EXPERIENCE, INFORMATION, CAUSE, SUGGESTION, QUESTION. Output the results in JSON format.
Question:
{question}
Context:
{context}
Answer:
{answer}
Summaries:
{{"EXPERIENCE": "", "INFORMATION": "", "CAUSE": "", "SUGGESTION": "", "QUESTION": ""}}

Figure 4: The refined prompt used for generating perspective-based summaries from user answers.