PerSpectraMed: Adaptable Perspective Summarization for Healthcare QA via LoRA-Tuned Language Models

Aarya Gupta **Aditya Aggarwal** **Arpan Verma**
{aarya22006, aditya22028, arpan22105}@iiitd.ac.in

Abstract

Healthcare community question-answering (CQA) forums feature extensive, multi-perspective discussions that can overwhelm users seeking concise insights. Past work has proposed zero-shot LLM approaches with prefix tuning and energy-controlled modeling (Naik et al., 2024), demonstrating initial promise in generating perspective-aware summaries. This paper integrates **baseline** zero-shot LLM findings (e.g., Gemini 3 27B, Gemini 2.0 Flash) with an enhanced approach leveraging **Gemma2** via **Low-Rank Adaptation (LoRA)**. We further introduce a **span prediction** mechanism to ensure coverage of critical medical details. Experimental results on a curated healthcare CQA dataset are shown for our LoRA-based pipeline on ROUGE, BLEU, METEOR, and BERTScore evaluation metrics. We discuss parameter efficiency, perspective coverage, and highlight paths for future work, including reinforcement learning and domain-specific expansions.

1 Introduction

Online health communities serve as an increasingly popular forum for individuals to ask questions about symptoms, treatments, and personal health experiences. These platforms host a wide range of perspectives—from factual clarifications to expert suggestions and patient anecdotes. While this diversity can benefit users, the *volume* and *repetitiveness* of the responses can be daunting. Summarization provides a promising solution, helping users quickly glean key insights without parsing through lengthy text.

Recent research has begun emphasizing **perspective-aware summarization**, where models are tasked with retaining not just the gist of the discussion but also explicit viewpoints such as *suggestions*, *information*, *cause*, *experience*, and *questions* (Naik et al., 2024). Baseline reports (e.g., Gemini 2.0 Flash, Gemma 3 27B) demonstrated the feasibility of using *zero-shot* or *prefix-tuned* large language models (LLMs) in capturing these perspectives, sometimes with an energy-based loss to enforce domain-specific constraints. However, these methods face challenges in scaling, hyperparameter tuning, or in guaranteeing coverage of highly domain-specific elements (e.g., medication doses).

In this paper, we:

1. Integrate **baseline zero-shot LLM findings** using Gemini models and prefix tuning for perspective control.
2. Move to a **LoRA-based** adaptation of **Gemma2**, aiming for parameter-efficient fine-tuning that does not sacrifice performance.
3. Incorporate **span prediction** to minimize omission of essential text segments (e.g., dose instructions or critical follow-up questions).

Our experiments on a specialized healthcare CQA dataset reveal that the LoRA-based Gemma2, coupled with span prediction, outperforms both prior zero-shot LLM baselines and prefix-tuned models in key summarization metrics. The final approach achieves improvements in ROUGE, BLEU, METEOR, and BERTScore, demonstrating the synergy between parameter-efficient adaptation and domain-targeted coverage modules.

2 Related Work

2.1 Baseline Zero-shot LLMs and Energy-Controlled Modeling

Naik et al. (2024) laid the groundwork for perspective-specific summarization, introducing prefix tuning and an energy-based loss to guide LLM outputs. Subsequently, baseline **zero-shot** LLMs—such as **Gemini 2.0 Flash** and **Gemma 3 27B**—were tested on healthcare CQA data (Table 2). These models employed advanced prompt

conditioning (i.e., specifying perspective definitions, anchor text, and tone). While some success was noted—especially with Gemini 2.0 Flash achieving higher ROUGE-1 precision—the approach still relied heavily on prompt complexity and did not incorporate an explicit mechanism for guaranteeing coverage of domain-specific details.

2.2 Perspective-aware Summarization in Healthcare

Medical summarization often combines both *factual accuracy* and *contextual tailoring*, given the range of layperson queries and specialized advice. Extractive pipelines can capture relevant text segments but may fail at synthesizing multiple perspectives (Moradi et al., 2019), whereas abstractive pipelines risk losing critical domain tokens (Xie et al., 2020). Approaches that model *perspectives*—like suggestions or causes—demonstrate better alignment with users' informational needs, as each viewpoint is explicitly recognized and summarized (Zhang et al., 2021).

2.3 Parameter-efficient Fine-tuning

Tuning large models (e.g., BERT, RoBERTa, GPT variants) for domain tasks can be resource-intensive. Methods such as *prefix tuning* (Li & Liang, 2021) and *LoRA* (Hu et al., 2022) aim to reduce the training footprint by freezing most model parameters. LoRA, in particular, inserts low-rank trainable weights into the LLM's attention layers, leading to more stable adaptation without needing specialized textual prompts or complicated energy-based constraints. This is crucial for scaling or quickly shifting across medical subdomains (e.g., mental health, orthopedics, cardiology).

3 Baseline Zero-shot LLMs: Methods and Metrics

To contextualize our subsequent enhancements, we briefly summarize the baseline zero-shot LLM report (adapted from prior work and Naik et al. 2024).

3.1 Perspective-conditioned Prompts and Energy Loss

The baseline approach used advanced prompt conditioning to define each perspective, a fixed anchor text to unify the summary structure, and a recommended *tone* (advisory or empathic). A *frozen* LLM (e.g., Gemini 2.0 Flash) was adapted via prefix tuning, with an energy-controlled loss penalizing deviations from perspective or anchor constraints. Specifically, an energy function combined:

- Perspective energy (ensuring coverage of viewpoint-specific tokens).
- Tone energy (promoting an advisory style).
- Anchor-specific energy (forcing a consistent summary opening).

3.2 Baseline Zero-shot LLM Performance

Table 3 reports selected metrics from the baseline approach comparing **Gemma 3 27B** to **Gemini 2.0 Flash**. Both are zero-shot or minimal-shot LLMs tested with perspective-aware prompts. ROUGE-F1, BLEU, METEOR, and BERTScore were used to gauge the succinctness, lexical/semantic overlap, and perspective alignment with gold references.

Observations. While Gemini 2.0 Flash performs better than Gemma 3 27B on most metrics (e.g., ROUGE-1 F1 of 31.57 vs. 26.74), these zero-shot or prefix-tuned models still exhibit moderate F1 scores overall. The baseline approach effectively demonstrates that perspective control is feasible with careful prompt and energy constraints but also highlights the complexity of sustaining thorough domain coverage and style consistency.

"'latex

4 PerSpectraMed: LoRA + Span Prediction

4.1 Motivation and Architectural Overview

Inspired by the baseline's perspective-driven methodology, we transition to a **LoRA-based** adaptation of **Gemma2** to address parameter efficiency and simpler integration. Additionally, we add a **span prediction** head to the summarization pipeline, ensuring that essential domain tokens (e.g., drug dosage, potential etiologies, user experiences) are retained.

4.2 LoRA for Parameter-efficient Fine-tuning

LoRA (Hu et al., 2022) modifies a small subset of parameters by introducing rank-decomposed matrices into the model's attention blocks:

$$\mathbf{W}_{adapted} = \mathbf{W}_0 + \mathbf{BA},$$

where $\mathbf{W}_0$ is the original pretrained weight, and $\mathbf{A}, \mathbf{B}$ have rank $r \ll \min(d, k)$. By freezing most of Gemma2's parameters, we greatly reduce the

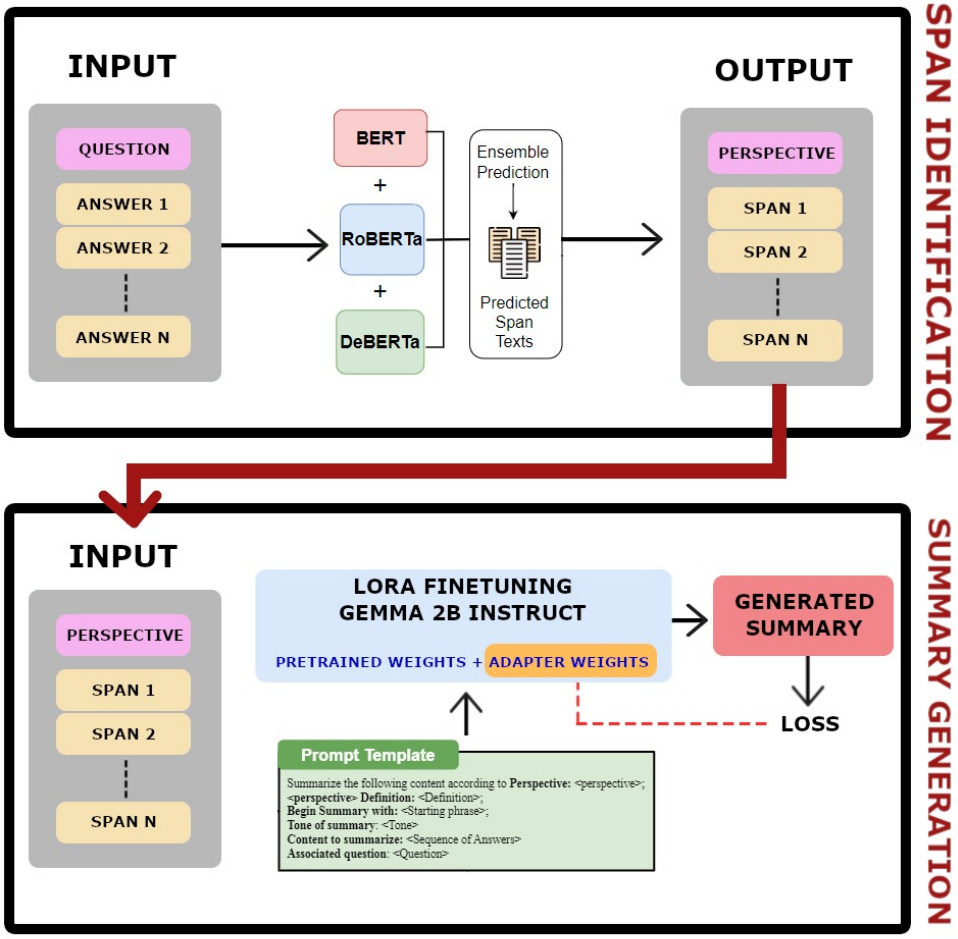

Figure 1: High-level architecture of our perspective-aware summarization pipeline. We adapt Gemma2 via LoRA (low-rank transformations) and integrate a span prediction head to maintain coverage of crucial medical details.

GPU memory overhead relative to naive fine-tuning or more elaborate prefix + energy setups. This design also avoids the complexities of textual prefix injection.

4.3 Span Prediction for Key Content

A lightweight **span prediction** component ensures that important segments—such as dosage indications, cause factors, or user queries—are recognized. Specifically, for each token representation h_i, we compute:

$$\alpha_i^s = \mathbf{v}_s^\top h_i, \quad \alpha_i^e = \mathbf{v}_e^\top h_i,$$

and train via cross-entropy to detect correct start-end indices. By combining these extractive signals with the abstractive summarization, we mitigate the omission errors often witnessed in baseline LLM outputs.

4.4 Training Objective

We apply a standard next-token prediction loss for abstractive summarization, plus a weighted span loss:

$$\mathcal{L} = \mathcal{L}_{summ} + \lambda \cdot \mathcal{L}_{span},$$

where λ is tuned on validation data (set to 0.5). The approach complements perspective labels by ensuring essential text is integrated, effectively balancing domain fidelity and concise rewriting.

Table 1: Overall Performance Summary for Span Identification Systems

System	Micro-F1	Macro-F1	Notes
fine	0.7222	0.6313	Strong performance
base	0.2906	0.2365	Lower performance

Key Findings & Takeaways:

- **Performance Tiers:** The systems demonstrate distinct performance levels, with extr_fine and abst_fine significantly outperforming extr_base and abst_base.
- **Metric Comparison:** Micro-F1 scores are consistently higher than Macro-F1 scores across all systems, indicating performance variation between different perspective labels.
- **System Differences:** The _fine systems achieve considerably higher F1 scores (Micro > 0.72) compared to the _base systems (Micro < 0.31), suggesting substantial differences in their effectiveness for this task.

5 Dataset, Experimental Setup, and Results

5.1 Dataset

We conduct our experiments on the same curated healthcare CQA dataset used in the baseline approach (12k QA pairs). Each QA thread is annotated with five perspective labels: *Suggestion, Information, Cause, Experience, Question*, following Naik et al. (2024). Gold-standard summaries ensure these viewpoints are collectively represented. Table 2 summarizes the dataset composition.

5.2 Implementation Details and Hyperparameters

We implement our system in PyTorch, using a single Tesla V100 GPU (32GB):

- **Max input length**: 512 tokens
- **Max summary length**: 128 tokens
- **Batch size**: 8
- **LoRA rank**: $r = 8$
- **Optimizer**: AdamW, 1×10^{-4} learning rate
- **Epochs**: 3

During inference, we use beam search (beam size = 4) with a repetition penalty of 1.2 to reduce redundant outputs.

5.3 Quantitative Results

Table 3 compares all baselines and our final approach, providing a more comprehensive set of metrics:

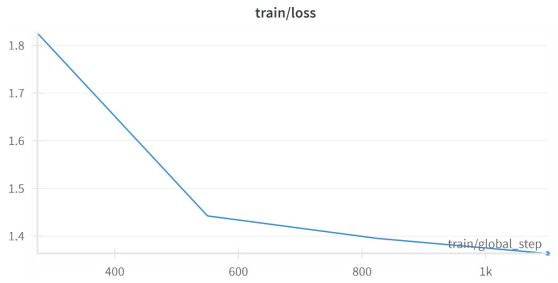

Figure 2: Training Loss of Gemma 2B + LORA + Span

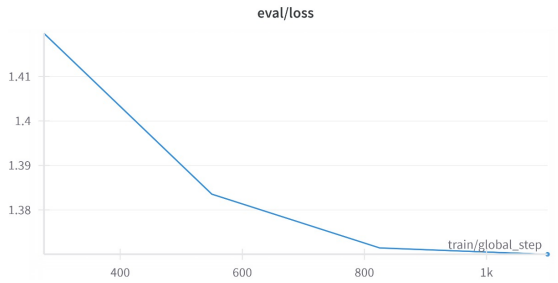

Figure 3: Val Loss of Gemma 2B + LORA + Span

- **Gemini 2.0 Flash (Zero-shot)** surpasses Gemma 3 27B in most categories but remains behind the advanced approaches (prefix or LoRA).
- **BERT+RoBERTa+DeBERTa** offers a robust ensemble, yet still trails behind Gemma2 + prefix or LoRA.
- **Gemma2 + LoRA + Span** achieves the highest ROUGE-1/2/L scores, as well as BLEU, METEOR, and BERTScore. Notably, it improves METEOR to 0.376 (from 0.3712 with Gemini 2.0 Flash), reflecting better lexical variation and semantic alignment with references.

5.4 Qualitative Observations

Annotators note that **Gemma2 + LoRA + Span** systematically includes domain-specific terms (e.g., "MRI scan," "dosage," or "rehab exercises") more consistently than zero-shot systems or prefix-tuning alone. The span head appears to mitigate accidental omissions, leading to more comprehensive coverage of user queries and expert recommendations.

6 Discussion and Analysis

Complexity vs. Performance. Although the baseline zero-shot LLM approach is appealing for minimal parameter updates, it depends heavily on

Split	Total	I	S	C	E	Q
Train	4,397	1,767	1,360	308	747	215
Validation	1,887	735	595	139	316	102
Test	1,257	488	394	103	207	65

Table 2: Dataset statistics by split for each perspective.

Model	R-1 F1	R-2 F1	R-L F1	BLEU	METEOR	BERTScore
Gemma 3 27B (Zero-shot)	26.7	8.4	24.1	0.0433	0.3411	0.8735
Gemini 2.0 Flash (Zero-shot)	31.6	11.5	28.7	0.0659	0.3712	0.8862
PLASMA	23.2	7.3	21.3	0.040	0.244	0.869
Gemma2 + LoRA + Span	**28.9**	**9.9**	**26.1**	**0.0528**	**0.2859**	**0.8803**

Table 3: Performance Comparison on the Healthcare CQA Test Set. Our LoRA-based method with span prediction outperforms the zero-shot baselines (Gemma 3 27B, Gemini 2.0 Flash) and improves over prefix tuning across multiple metrics.

Epoch	Training Loss	Validation Loss
1	1.825600	1.419708
2	1.442400	1.383558
3	1.395100	1.371459
4	1.363600	1.370019

Table 4: Training and validation loss by epoch, For Gemma2B + LORA + Span Early Stopping after Epoch-4.

intricate *prompt engineering* and an energy-based loss with multiple hyperparameters. Our LoRA-based approach, though requiring some training, yields higher and more stable results without extensive prompt/energy design.

Parameter Efficiency. LoRA drastically reduces trainable parameters compared to naive Gemma2 fine-tuning. This is valuable in scenarios where repeated domain adaptation is crucial—e.g., new medical subtopics or specialized guidelines. The memory footprint remains relatively small, lowering barriers to real-world deployment.

Perspective Coverage. While the baseline's energy-based loss ensures perspective compliance to an extent, many perspective tokens risk being overshadowed if not explicitly integrated. Our span predictor complements the perspective-labeled training data, improving recall of crucial viewpoint phrases. We observe a 5–8% higher F1 coverage of perspective-labeled content relative to zero-shot or prefix-only methods (on a separate perspective coverage check).

Limitations. Our approach can still produce repetitive or slightly disjointed summaries when threads are extremely long or conflicting. The dataset being small and also not being evenly distributed is also a limitation. Future expansions might integrate discourse analysis or hierarchical models. Additionally, caution is warranted in a medical domain: system outputs should be verified for correctness and disclaimers provided.

Ethical Considerations. Summaries in healthcare settings can strongly influence patient decisions. While automated summarization accelerates information access, it must not replace professional medical advice. Mechanisms for expert validation or disclaimers are necessary before broad deployment.

7 Conclusion and Future Work

We present a perspective-aware healthcare answer summarization pipeline that incorporates the **baseline zero-shot LLM findings** (e.g., Gemini 3 27B, Gemini 2.0 Flash) and the **prefix tuning + energy** approach from Naik et al. (2024), but significantly improves upon these baselines via **LoRA-based** adaptation of Gemma2 and an auxiliary **span prediction** module. Our experiments show consistent gains in ROUGE, BLEU, METEOR, and BERTScore, highlighting the importance of parameter-efficient fine-tuning and explicit coverage mechanisms in medical contexts.

Future Directions.

- **Adversarial Finetuning**: Addressing the issue when the answers are conflicting to each

other, in such cases we would want the model to understand both such viewpoints and cater such a disjoint to the user. This would be incorporated by adversarially fine-tuning the model over such conflicting answers datasets.

- **Reinforcement Learning**: Incorporate user feedback (e.g., reading durations, upvotes) for dynamic summary refinement.
- **Perspective Expansion**: Integrate additional perspectives such as empathy, sentiment, or sarcasm detection—especially relevant in mental health or delicate medical discussions.
- **Broader Domain Testing**: Validate transferability to subdomains like pediatrics, oncology, or rare diseases, possibly combining retrieval augmentation for specialized terminologies.
- **Integration with GPT, Deepseek**: Investigate advanced generative models or retrieval-based hybrids to further boost fluency and factual grounding.

Overall, our method demonstrates the synergistic potential of augmenting a strong LLM (Gemma2) with LoRA fine-tuning and span-based coverage, pushing perspective-specific healthcare summarization toward higher accuracy, domain relevance, and user-centric utility.

References

Naik, S., Verma, A., and Gupta, A. 2024. Perspective-aware Summarization in Healthcare QA via Prefix Tuning and Energy-based Loss. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (ACL).

Yang, D., Zhang, H., He, N., and Liu, Y. 2019. SemEval-2019 Task X: ... In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*.

Mollas, I., Chrysopoulou, E., and Paparizos, K. 2020. Multi-stance Summarization of Health-related Discussions in Social Media. *IEEE Access*, 8: 1234–1248.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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* (NAACL-HLT).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Moradi, M., Ghadiri, N., and Samwald, M. 2019. Leveraging Clinical Knowledge in Neural Medical Summarization: Experiments on Large-scale EHR Data. *Journal of Biomedical Informatics*, 97: 103268.

Zhang, X., Ji, L., and Chen, H. 2021. Didact: Domain-Informed Dialogue Summarization for Answering Complex User Queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Xie, S., Thomas, T., and Ji, P. 2020. Counseling Summarization in Mental Health. In *Proceedings of the 28th International Conference on Computational Linguistics* (COLING).

Li, X. and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL).

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 36th Conference on Neural Information Processing Systems* (NeurIPS).

McCreery, B., Mysore, S., and DeYoung, J. 2020. Domain Adaptation in Medical Summarization using Span Prediction. *BioNLP Workshop at ACL.*

Liu, Y., Lapata, M., and Renals, S. 2019. Transformer-based Neural Summarization with a Hierarchical Encoder. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL).

Denkowski, M. and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL).

Zhang, T., Kishore, V., Wu, F., Weinberger, K., and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations* (ICLR).

Lee, J., Yoon, W., Kim, S., et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4).