Lightweight LLM Adaptation for Medical Summarisation: Roux-lette at PerAnsSumm Shared Task

Anson Antony and **Peter Vickers** and **Suzanne Wendelken**
Northeastern University, Boston, MA
{a.antony, p.vickers, s.wendelken}@northeastern.edu

Abstract

The PerAnsSumm Shared Task at CL4Health@NAACL 2025 focused on Perspective-Aware Summarization of Healthcare Q/A forums, requiring participants to extract and summarize spans based on predefined perspective categories. Our approach leveraged LLM-based zero-shot prompting enhanced by semantically-similar In-Context Learning (ICL) examples. Using Qwen-Turbo with 20 exemplar samples retrieved through NV-Embed-v2 embeddings, we achieved a mean score of 0.58 on Task A (span identification) and Task B (summarization) mean scores of 0.36 in Relevance and 0.28 in Factuality, finishing 12th on the final leaderboard. Notably, our system achieved higher precision in strict matching (0.20) than the top-performing system, demonstrating the effectiveness of our post-processing techniques. In this paper, we detail our ICL approach for adapting Large Language Models to Perspective-Aware Medical Summarization, analyze the improvements across development iterations, and finally discuss both the limitations of the current evaluation framework and future challenges in modeling this task. We release our code for reproducibility.[1]

1 Background

Healthcare community question-answering (CQA) forums serve as information resources for patients seeking accessible explanations outside clinical settings, caregivers navigating medical decisions, and curious individuals performing health research whilst avoiding stigma or costs tied to formal consultations (Beloborodov et al., 2013). However, the unstructured discussions typical of online forums often bury actionable insights under noise such as anecdotal claims, off-topic debates, or incorrect advice (Naik et al., 2024).

Perspective-aware summarization addresses this by categorizing forum responses into domains like suggestions ("ERCP procedures minimize scarring") or experiences ("Phantom pain persisted post-surgery")—enabling users to contrast evidence-based options with peer-endorsed narratives. Perspective-Aware Summarization [PAS] addresses this challenge by identifying and categorizing diverse viewpoints within healthcare forum responses. Unlike traditional summarization into a single version, PAS structures information into distinct perspective categories: 'Cause' (explanations of medical conditions), 'Suggestion' (recommended treatments or actions), 'Experience' (personal accounts), 'Question' (follow-up inquiries), and 'Information' (factual medical knowledge). The PerAnsSumm Shared Task at CL4Health@NAACL 2025 split this approach into two subtasks: Span Identification: Tagging text segments in answers aligning with five perspectives (Cause, Suggestion, Experience, Question, and Information). Summarization: Generating concise summaries for each of the five perspectives.

Building on the **P**erspective s**UM**marization d**A**taset (**PUMA**) dataset, a corpus of 3,167 annotated CQA threads annotated with 10K Human-authored Perspective-Aware Summarizations, the task encouraged models to move beyond single-view summaries common in traditional methods (Agarwal et al., 2025).

Our approach used the few-shot capabilities of Large Language Models to learn novel tasks with minimal exposure to labeled examples, including in the Medical Domain. Using just 20 exemplar samples from the training set, we are able to obtain a mean score of 0.58 on task A and 0.36 in Relevance and 0.29 in Factuality on Task B.

In this paper, we detail our approach, including releasing the code for all of our attempts. We then outline further approaches to improve performance. Finally, we discuss the difficulties of the task it-

[1] https://github.com/petervickers/Roux-PerAnsSumm

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 389–397

self, including bias and ambiguity in community question-answer forums.

2 Related Work

Healthcare community question-answering (CQA) forums serve as information resources for patients seeking medical information outside clinical settings, though unstructured discussions often bury actionable insights beneath anecdotal claims and incorrect advice (Beloborodov et al., 2013; Naik et al., 2024). Traditional summarization approaches typically condense information into a single narrative, whereas perspective-aware summarization (PAS) addresses this limitation by categorizing content into distinct perspective types (cause, suggestion, experience, question, and information) (Agarwal et al., 2025).

Large Language Models (LLMs) have demonstrated strong performance on healthcare tasks in few-shot settings without domain-specific fine-tuning (Brown et al., 2020; Liu et al., 2023). Notably, Nori et al. (2023) showed that general-purpose models like GPT-4, when enhanced with appropriate prompting techniques (termed "MedPrompt"), can match or exceed specialized medical models. MedPrompt combines dynamic few-shot selection using k-nearest neighbors, self-generated chain-of-thought reasoning, and choice shuffling ensembles.

For span identification tasks similar to our work, named entity recognition (NER) approaches have been adapted for more complex extraction tasks. Tools like Spacy-LLM (Honnibal et al., 2020; Explosion AI, 2025) provide structured templates for guiding LLMs in entity extraction, which we adapt for perspective categories. However, perspective identification presents unique challenges compared to traditional NER: perspective spans often cross sentence boundaries, have ambiguous boundaries, and require subjective interpretation based on annotator guidelines.

Current limitations in perspective-aware systems include reliance on domain-specific training that limits generalization, handcrafted prompts requiring medical expertise, difficulties identifying perspective boundaries in conversational text, and challenges maintaining factual accuracy while generating perspective-specific summaries.

3 Methodology

Building on recent advances in LLM-based medical text processing, we introduce a novel approach to the PerAnsSumm Shared Task, which requires a two-stage cascaded pipeline: (1) Perspective-Aware Span Extraction followed by (2) Perspective-Aware Span Summarization (Agarwal et al., 2025). Our system addresses the key limitations identified in the related work through a specialized adaptation of the MedPrompt framework (Nori et al., 2023) for perspective-based tasks.

3.1 Overview of Our Approach

While MedPrompt has demonstrated state-of-the-art performance on medical multiple-choice questions (Nori et al., 2023), adapting it to open-ended perspective identification and summarization tasks presents several unique challenges. We preserve MedPrompt's core strength—dynamic few-shot selection—while modifying its architecture to accommodate span extraction rather than option selection. We term this MedPrompt Adaptation for Perspective Tasks.

Our system leverages semantic similarity to identify relevant examples from the training data. This addresses the scalability limitations of expert-dependent systems while maintaining the flexibility to adapt to diverse healthcare topics.

For both tasks (A) and (B), our system implements a four-component architecture:

1. **Dynamic In-Context Learning Sampling**: We extend MedPrompt's k-nearest neighbors approach to perspective-specific content by encoding samples using NVEmbed-v2. Our ICL strategy differs between the two subtasks:
   - For Task A (span extraction), we generate embeddings with the input question as the query and all questions in the training dataset as documents. For each test instance, we compute cosine similarity between its question embedding and all training question embeddings.
   - For Task B (summary generation), we generate embeddings at the perspective level, using the input spans as the query and retrieving training examples where the spans share the same perspective category. This focuses similarity computation on perspective-specific content rather than general question context.

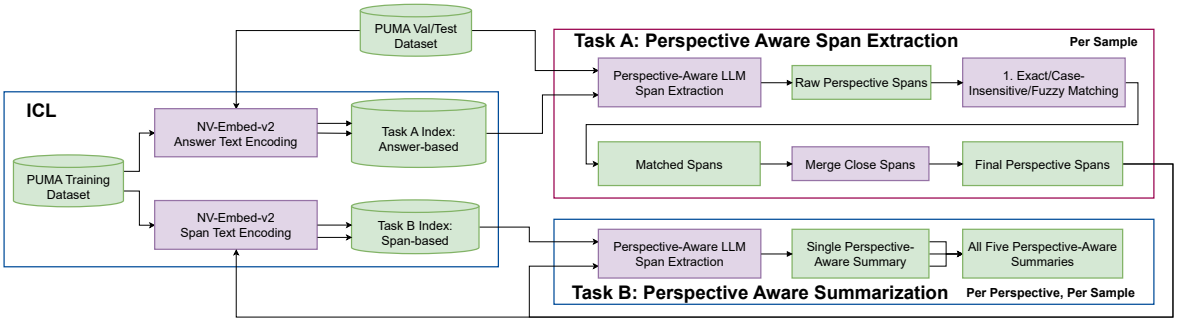

Figure 1: Data Flow Diagram of our Span Extraction and Summarization System

In both cases, we select the k=16 most similar examples based on these similarity scores to serve as in-context examples.

2. **Task-Specific Prompt Engineering**: We adapt Spacy-LLM's NER templates (Honnibal et al., 2020; Explosion AI, 2025) to the more complex task of perspective identification. Where traditional NER identifies concrete entities with clear boundaries, perspective identification requires identifying abstract categories that may span multiple sentences.

3. **Annotator Bias Replication**: We include explicit instructions directing the model to mirror the subjective biases present in the In Context Learning annotations.

4. **Span Post-Processing**: We implement a three-stage cascading alignment strategy to overcome LLMs' known limitations in returning precise character indices (Wu et al., 2023). This approach significantly improves upon the exact matching typically used in NER systems, which fails to account for the flexibility needed in perspective boundary identification.

To ensure consistent output formatting across both subsystems, we enforce JSON output structure by constraining the first token of the model's response to be '{', effectively force-decoding the beginning of a JSON object.

As Figure 1 shows, our system consists of three high-level components: Perspective Aware Span Extraction, Perspective Aware Summarization, and In-Context Learning. In-Context Learning (left) leverages the PUMA training dataset through dual NV-Embed-v2 encoding pathways—one optimized for Task A using answer-based text encoding and another for Task B using span-based encoding. This creates semantic indices for efficient retrieval of relevant examples during inference. Task A (upper right) performs perspective-aware span extraction through a three-stage cascading alignment process (exact, case-insensitive, and fuzzy matching), followed by a span merging step to produce cohesive perspective-specific text segments. These extracted spans then feed into Task B (lower right), which generates perspective-aware summaries organized across the five predefined categories (cause, suggestion, experience, question, and information).

For Task A, we found no advantage in perspective-level span extraction. For Task B, performance improved with perspective-specific summarization, so we generate summaries separately for each perspective and merge the results.

This modular design allowed us to conduct controlled experiments, isolating the impact of different embedding models and varying quantities of in-context examples on system performance.

3.2 Evaluation Metrics

The PerAnsSumm Shared Task evaluation framework comprises distinct metrics for both span identification (Task A) and perspective-aware summarization (Task B). Of note, the perspective-aware summarization was dependent on the output of the span identification model. Gold standard span inputs were not provided for Task B. Metrics compared between the generated outputs and the max-voted labels from the test split of the PerAnsSumm/PUMA dataset.

Task A: Span Identification Metrics

1. Macro-averaged F1 score: Evaluates performance across all five perspective categories (cause, suggestion, experience, question, and information), mitigating class imbalance effects.

2. Strict Matching: Measures exact correspondence between predicted and ground-truth

spans, considering both boundaries and classification labels.

3. Proportional Matching: Allows partial credit for spans that overlap with the ground truth, accounting for minor discrepancies in extraction.

"Ground-truth" reference spans were from the task/PUMA dataset annotations, which were manually labeled for each perspective and reflected annotation bias discussed elsewhere.

Task B: Perspective-aware Summarization Metrics

1. ROUGE (R-1, R-2, R-L) (Lin, 2004): Measures unigram overlap (R-1), bigram overlap (R-2), and longest common subsequence (R-L) between generated summaries and reference summaries.

2. BLEU (Papineni et al., 2002): Computes n-gram precision against reference summaries, commonly used in machine translation but adapted here for summarization.

3. METEOR (Banerjee and Lavie, 2005): Extends BLEU by incorporating synonymy and stemming, better capturing semantic equivalence.

4. BERTScore (Zhang et al., 2020): Uses contextualized BERT embeddings to compare generated and reference summaries at the semantic level, overcoming limitations of n-gram-based metrics.

Additionally, "factuality" assessments were included to evaluate the alignment of generated summaries with the source content:

1. AlignScore (Zha et al., 2023): attempts to measures factual consistency using a unified alignment function to compare source text and generated summaries.

2. SummaC (Laban et al., 2022): attempts to detect contradictions and hallucinations in summarization by leveraging natural language inference (NLI) models and sentence-level document-summary pairs.

Reference summaries were the annotator-provided summaries from the task/PUMA dataset, which were written post-hoc based on extracted spans. As with Task A, these summaries inherit the dataset's biases and limitations, influencing how models were evaluated.

4 Experimental Setup

We developed our approach over four system variants (summarized in Table 1), each representing incremental improvements to our initial baseline implementation. All systems were evaluated on the PerAnsSumm Shared Task.

4.1 System Implementation Details

Our implementation leveraged the core MedPrompt architecture with targeted adaptations for perspective-aware tasks:

Model Selection: We initially employed OpenAI's GPT-4o-mini model (et al., 2024) (versions v1-v2) before transitioning to Qwen/Qwen-turbo (Qwen et al., 2025) (versions v3-v4) based on preliminary performance evaluations.

Dynamic In-Context Learning: For version v1, we used zero-shot prompting without in-context learning examples. Version v2 incorporated 5 in-context examples selected using OpenAI's text-embedding-3-small model to match samples, while versions v3-v4 employed NVIDIA's NV-Embed-v2 (Lee et al., 2025) with retrieval sets of 5 and 20 examples, respectively. This progression allowed us to evaluate the impact of both example quantity and embedding quality on performance.

Post-processing Pipeline: All systems employed our three-stage cascading alignment strategy for span reconciliation, with refinements in later versions to address edge cases identified during development:

1. **Exact substring matching**: First attempting verbatim matches using Python's native string.find() function, with extension to word boundaries for cleaner spans

2. **Case-insensitive matching**: If exact matching failed, converting both source and target texts to lowercase before applying the find() function again

3. **Sentence-level fuzzy matching**: For spans still unmatched, breaking the text into sentences and applying thefuzz library's ratio() algorithm to find the best matching sentence, with early termination at 95% similarity

System	Model	K	Embedder	Prompts		Scores	
				Task A (Span Extraction)	Task B (Summarization)	A	B
v1	openai/gpt-4o-mini	None	None	Span-Prompt-V1	Summ-Prompt-V1	0.58	0.30
v2	openai/gpt-4o-mini	5	OpenAI text-embedding-3-small	Span-Prompt-V2	Summ-Prompt-V2	0.58	0.33
v3	qwen/qwen-turbo	5	NVIDIA NV-Embed-v2	Span-Prompt-V3	Summ-Prompt-V3	0.58	0.35
v4	qwen/qwen-turbo	20	NVIDIA NV-Embed-v2	Span-Prompt-V4	Summ-Prompt-V4	0.58	0.36

Table 1: System configurations and performance comparison. K indicates the number of in-context learning examples, Embedder refers to the model used for retrieving similar examples. Scores represent macro-averaged metrics: Task A scores show span alignment accuracy (Avg. column from Table 2), Task B scores show relevance performance (Relevance Avg. from Table 3).

The fuzzy matching threshold ($\theta = 0.7$) served as a quality filter, with spans scoring below this threshold being discarded. Our implementation also included specialized handling for overlapping spans through the, which merged spans of the same perspective category that were within 5 characters of each other.

4.2 Experimental Configurations

Table 1 summarizes our four experimental configurations. The progression from v1 to v4 represents an evolution from simple baseline approaches to sophisticated in-context learning with optimized similarity matching:

Key experimental parameters were:

- **Similar Example Selection:** For ICL-based systems (v2-v4), we selected examples from the training corpus based on cosine similarity between embedding vectors. Version v4's expanded number of ICL samples (K=20) allowed for more diverse exemplars.
- **Fuzzy Matching Threshold:** We empirically determined a similarity threshold of $\theta = 0.7$ for accepting predicted spans, with scores below this threshold triggering rejection during post-processing.

4.3 Evaluation Process

Systems were evaluated using the official PerAnsSumm metrics as described in Section 3.2. We submitted all versions, v1-v4, to the shared task evaluation server, with v4 representing our best-performing configuration. The detailed prompt specifications for all system variants are provided in Appendix A.

5 Results

Tables 2 and 3 present the performance of our four system variants on Tasks A and B, respectively. Our final system (v4) achieved an average score of 0.58 on Task A (span identification) and 0.36 on Task B's relevance metrics with 0.28 on factuality metrics, placing our team 13th out of 23 teams overall in the shared task.

For Task A, all four of our system variants achieved consistent performance with a macro F1 classification score of 0.81, strict matching F1 of 0.22, and proportional matching F1 of 0.64. Our overall Task A average of 0.58 placed us within 3.4% of the top-performing system's score of 0.60.

For Task B, we observed progressive improvements across our system versions. The relevance metrics improved from 0.30 in v1 to 0.36 in v4 (+20%), while factuality scores declined slightly from 0.29 to 0.28. The gap between our system and the top-performing system was more pronounced in Task B, with our relevance average trailing the leader by 14% relative.

Each system iteration brought incremental improvements: v1 (zero-shot GPT-4o-mini) achieved 0.30 on Task B relevance, v2 (GPT-4o-mini with ICL) improved to 0.33 (+10%), v3 (Qwen-Turbo with NV-Embed-v2) reached 0.35 (+6%), and v4 (expanded to 20 examples) achieved our best score of 0.36 (+3%).

Analysis of the overall leaderboard reveals that the top 13 teams were tightly clustered, with scores ranging from 0.457 to 0.400, indicating that minor implementation differences had significant impact on final rankings.

6 Discussion

6.1 Task A Performance Analysis

Despite transitions from GPT-4o-mini (v1-v2) to Qwen-Turbo (v3-v4) as our base LLM, our Task A performance remained remarkably consistent. This stability suggests that our model effectively learned to distinguish between the five perspective categories regardless of the specific implementation details or embedding model used for in-context

	CLASSIFICATION		STRICT MATCHING			PROPORTIONAL MATCHING			
Submission	**Macro F1**	**Weighted F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Avg.**
Roux-lette 1	0.81	0.87	0.20	0.22	0.21	0.59	0.73	0.64	0.58
Roux-lette 2	0.81	0.87	0.20	0.23	0.22	0.57	0.72	0.64	0.58
Roux-lette 3	0.81	0.87	0.20	0.23	0.22	0.57	0.72	0.64	0.58
Roux-lette 4	0.81	0.87	0.20	0.23	0.22	0.57	0.72	0.64	0.58
Mean Gradient	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall Improvement	0.00	0.00	0.00	0.00	0.00	-0.02	-0.01	0.00	0.00
Leader (WisPerMed)	0.88	0.92	0.17	0.23	0.20	0.62	0.74	0.68	0.60
delta (Roux - WisPerMed)	-0.07	-0.05	0.03	0.00	0.02	-0.05	-0.02	-0.04	-0.02

Table 2: Task A Results

	RELEVANCE							FACTUALITY		
Submission	**ROUGE1**	**ROUGE2**	**ROUGEL**	**BERTScore**	**METEOR**	**BLEU**	**Avg.**	**AlignScore**	**SummaC**	**Avg.**
Roux-lette 1	0.31	0.09	0.27	0.80	0.27	0.07	0.30	0.37	0.22	0.29
Roux-lette 2	0.34	0.12	0.30	0.82	0.31	0.08	0.33	0.36	0.22	0.29
Roux-lette 3	0.37	0.15	0.33	0.83	0.33	0.11	0.35	0.31	0.23	0.27
Roux-lette 4	0.38	0.17	0.34	0.83	0.33	0.12	0.36	0.32	0.23	0.28
Mean Gradient	0.02	0.03	0.02	0.01	0.02	0.02	0.02	-0.02	0.00	-0.01
Overall Improvement	0.07	0.07	0.07	0.03	0.06	0.06	0.06	-0.05	0.01	-0.02
Leader (WisPerMed)	0.45	0.22	0.41	0.90	0.41	0.13	0.42	0.41	0.30	0.35
delta (Roux - WisPerMed)	-0.07	-0.05	-0.07	-0.07	-0.08	-0.01	-0.06	-0.09	-0.07	-0.08

Table 3: Task B Results

example retrieval.

Notably, our precision in strict matching (0.20) exceeded the top-performing system (WisPerMed's 0.17), indicating that our cascading alignment strategy with fuzzy matching was particularly effective at identifying precise span boundaries. While our recall matched the leader (0.23), our overall strict matching F1 (0.22) slightly outperformed the leader's 0.20, demonstrating the effectiveness of our three-stage cascading alignment strategy with fuzzy matching threshold ($\theta = 0.7$).

The small performance gap between participating teams in Task A is striking, with the top 13 systems achieving scores within a narrow range (0.58-0.62). This clustering suggests that the task may have reached a performance ceiling with current LLM-based methods, possibly due to inherent ambiguities in perspective boundary identification.

6.2 Task B Performance Analysis

The clear progression in our Task B performance correlates directly with improvements in our LLM and embedding models. The significant gains in ROUGE-2 (0.09 to 0.17, +89%) and BLEU (0.07 to 0.12, +71%) indicate better capture of n-gram sequences and improved alignment with reference summaries as we enhanced our embedding model quality and expanded ICL example counts.

The inverse relationship between relevance and factuality scores raises important questions about evaluation metrics in perspective-aware summarization. As our systems better matched reference summaries (higher relevance), they simultaneously drifted from factual alignment with source content (lower factuality). This trade-off, particularly evident in the drop in AlignScore (0.37 to 0.32, -13.5%), suggests that human annotators may have introduced interpretations or simplifications in their summaries that deviated from the original forum content.

The leaderboard reveals a significant gap in Task B performance between the top 5 teams (relevance scores of 0.40-0.42) and the remainder of the field (scores below 0.39), suggesting that certain architectural approaches may have offered substantial advantages in summarization quality.

6.3 Effectiveness of In-Context Learning Approaches

The most substantial improvements in our systems came from the transition from zero-shot to in-context learning with semantically similar examples. The progression from v1 to v4 underscores the importance of both the quality of embedding models for finding related samples and the quantity of in-context examples in achieving optimal

performance for perspective-aware summarization.

The diminishing returns observed when increasing from 5 to 20 examples (+6% vs. +3% improvement) suggests that example quality may be more important than quantity beyond a certain threshold. This finding aligns with recent research showing that carefully selected few-shot examples often outperform larger random samples in in-context learning scenarios.

6.4 Bias Learning vs. Medical Understanding

We speculate that the structure of Task A encouraged models to imitate annotator biases rather than developing genuine understanding of medical discourse. Our experiment with explicit bias instruction did not significantly improve results, suggesting that the bias patterns were either inconsistent or difficult for the model to internalize.

This observation is supported by our Task A performance remaining stable across different LLMs and embedding models, indicating that the task primarily measures how effectively systems can approximate existing annotation patterns rather than demonstrating true innovation in perspective identification. The tight clustering of team performances on the leaderboard further supports this hypothesis.

Examining the leaderboard, we observe that the top-performing systems achieved their advantage primarily through Task B (summarization) rather than Task A (span identification), where scores were more tightly clustered. This suggests that while span identification may have reached a performance ceiling, summarization quality remains an area where significant improvements are possible.

7 Conclusion

In this work, we explored an LLM-driven approach to perspective-aware summarization in the PerAnsSumm shared task, leveraging a **lightweight, zero-shot ICL methodology that requires no fine-tuning** and can be readily applied to any LLM. Our approach used semantic similarity-guided in-context learning with minimal example retrieval, demonstrating the efficacy of model-agnostic techniques for structured medical text understanding.

For Task A, we used Qwen-Turbo guided by 20 semantically similar training samples retrieved using NV-Embed-v2 embeddings, achieving a mean score of 0.58 and notably exceeding the top-performing system in strict matching precision (0.20 vs. 0.17). Our three-stage cascading alignment strategy (exact, case-insensitive, and fuzzy matching with $\theta = 0.7$) proved effective for capturing perspective boundaries without the need for task-specific training.

For Task B, we extended this model-agnostic methodology to summarization, incrementally improving relevance metrics from 0.30 (zero-shot) to 0.36 (20 examples), while maintaining factuality scores around 0.28. Our experimental progression showed embeddings quality and example selection significantly impact performance, with the transition from zero-shot to ICL (v1→v2: +10%) yielding greater improvements than embedding upgrades (v2→v3: +6%) or increasing example count (v3→v4: +3%).

Our results suggest potential limitations in the current task framework. The narrow performance range across teams in Task A (0.58-0.62) may indicate a ceiling effect possibly attributable to inherent ambiguities in perspective boundary identification. The observed inverse relationship between relevance and factuality metrics raises questions about potential annotation biases or simplifications in reference summaries. Additionally, the patterns we observed suggest the task design may encourage models to replicate annotation patterns rather than develop genuine medical understanding.

The primary advantage of our approach lies in its **simplicity and transferability across models**, requiring only basic API access to any capable LLM rather than expensive fine-tuning or domain-specific architectures. Future perspective-aware summarization tasks would benefit from more clinically relevant, open-ended evaluation frameworks that foster methodological innovation with real-world impact rather than alignment with pre-existing annotation patterns.

8 Limitations

Our approach faces several limitations.

First, our models learn to replicate annotator biases rather than develop true medical understanding, evidenced by the tight clustering of Task A scores (0.58-0.62) across teams. Second, the diminishing returns when scaling from 5 to 20 examples (10% → 6% → 3% improvement) suggests fundamental constraints in example-based learning without domain-specific training. Third, the inverse relationship between relevance and factuality scores indicates that optimizing for reference

similarity may reduce source content faithfulness.

Due to time and computational constraints, we were unable to exhaustively test all possible values for the fuzzy matching threshold (θ), optimal number of ICL samples, or evaluate across a broad spectrum of available LLM models.

Finally, and most importantly, our system lacks mechanisms to verify medical accuracy or distinguish between credible and non-credible information in healthcare forums. We highlight broader concerns about using AI for medical applications, which carries documented risks and should never replace physician guidance.

Future work should focus on integrating domain-specific medical knowledge, developing evaluation frameworks better aligned with clinical utility, and establishing robust fact-verification mechanisms for healthcare content.

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.

Alexander Beloborodov, Artem Kuznetsov, and Pavel Braslavski. 2013. Characterizing health-related community question answering. In *Advances in Information Retrieval*, pages 680–683, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

OpenAI et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Explosion AI. 2025. spacy-llm: Structured NLP with LLMs.

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

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.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

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.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *Preprint*, arXiv:2305.15525.

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.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Preprint*, arXiv:2311.16452.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru

Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for Chinese spelling correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10743–10756, Toronto, Canada. 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. *Preprint*, arXiv:1904.09675.

A Appendix A: Prompts

A.1 Task A: Span Extraction Prompts

- **Baseline Prompt (v1):**
  - Core instruction: *Analyze text and identify spans expressing different perspectives (CAUSE, SUGGESTION, EXPERIENCE, QUESTION, INFORMATION)*
  - Added sliding window matching in v2 for phrase boundary detection
  - Integrated overlap handling in v3 with span merging logic
  - JSON structure requirements:
    - Extract complete phrases (under 100 characters)
    - Prefer full sentences where possible
    - Mandatory "text" field in JSON objects
- **In-Context Learning Prompt (v2):** *Enhanced version with example-based guidance:*
  - Detailed perspective definitions:
    - EXPERIENCE: First-hand accounts
    - INFORMATION: Factual data
    - CAUSE: Explanatory reasoning
    - SUGGESTION: Recommendations
    - QUESTION: Information requests
  - Example JSON format:

```
{
    "EXPERIENCE": [{"text": "..."}],
    "INFORMATION": [{"text": "..."}]
}
```

  - Includes 5 retrieved examples using OpenAI embeddings
- **NV-Embed-v2 Prompt (v3):** *Optimized version with:*
  - NVIDIA NV-Embed-v2 for example retrieval
  - OpenRouter API integration
  - Upgraded LLM backend
  - Maintains 5-example context (K=5)
- **Scaled ICL Prompt (v4):** *Enhanced capacity version:*
  - Expands context window to 20 examples (K=20)
  - Retains NV-Embed-v2 retrieval system
  - Optimized for long-context processing

A.2 Task B: Summarization Prompts

- **Merged Baseline Prompt (v1):**
  - Core template: *Summarize {perspective} points about "question"*
  - Requirements:
    - 2-3 sentence summaries
    - Maintain factual accuracy
    - Direct answer alignment
- **ICL Summarization (v2):** *Example-enhanced version:*
  - Incorporates retrieved examples
  - Structured template: *"Analyze text and extract perspective summaries for {perspective}"*
  - Processes span inputs:
    - `{span 1 text}`
    - `{span 2 text}`
- **NV-Embed-v2 Summarization (v3):** *Optimized architecture:*
  - NV-Embed-v2 retrieval system
  - Human-aligned prompt structure
  - Maintains K=5 examples
- **Scaled Summarization (v4):** *Expanded context version:*
  - Processes 20 examples (K=20)
  - Enhanced coherence through extended context
  - Maintains NV-Embed-v2 retrieval