Medifact at PerAnsSumm 2025: Leveraging Lightweight Models for Perspective-Specific Summarization of Clinical Q&A Forums

Nadia Saeed
Computational Biology Research Lab
Department of Computer Science
National University of Computer and Emerging Sciences (NUCES-FAST)
Islamabad, Pakistan
i181606@nu.edu.pk

Abstract

The PerAnsSumm 2025 challenge focuses on perspective-aware healthcare answer summarization (Agarwal et al., 2025). This work proposes a few-shot learning framework using a Snorkel-BART-SVM pipeline for classifying and summarizing open-ended healthcare community question-answering (CQA). An SVM model is trained with weak supervision via Snorkel, enhancing zero-shot learning. Extractive classification identifies perspective-relevant sentences, which are then summarized using a pretrained BART-CNN model. The approach achieved 12th place among 100 teams in the shared task, demonstrating computational efficiency and contextual accuracy. By leveraging pretrained summarization models, this work advances medical CQA research and contributes to clinical decision support systems.[1]

1 Introduction

Healthcare Community Question-Answering (CQA) forums have become a vital source of medical information to seek advice and share experiences (Jiang, 2024; Zhang et al., 2024). These platforms generate diverse responses, ranging from factual knowledge to personal opinions like PUMA dataset (Naik et al., 2024). Traditional CQA summarization methods focus on selecting a single best-voted answer as a reference summary (Tsatsaronis et al., 2015; Kell et al., 2024). However, a single answer often fails to capture the broad range of perspectives available across multiple responses. To better serve users, it is essential to generate structured summaries that represent various viewpoints effectively.

To address this, we introduce a hybrid framework that combines perspective classification and summarization, as shown in Figure 1. The first step involves classifying user responses into predefined perspectives using a multi-step learning pipeline. This pipeline integrates Snorkel-based weak supervision (Ratner et al., 2017), support vector machine (SVM) classification with sentence embeddings (Rueping, 2010), and zero-shot learning (ZSL) using transformer models (Lewis, 2019). The goal is to enhance classification accuracy, especially when labeled data is scarce.

Once classified, responses undergo a two-step summarization process. We employ extractive summarization using BART to select key sentences from classified perspectives (Lewis, 2019). Then, we refine these summaries using abstractive summarization with Pegasus to improve fluency and coherence (Zhang et al., 2020). The composed model is evaluated on the **PerAnsSumm Shared Task - CL4Health@NAACL 2025**, which focuses on analyzing multi-perspective responses in Community Question Answering (CQA) (Agarwal et al., 2025). Given a user-generated question Q and a set of responses A, the task is divided into two key objectives:

(1)Perspective Classification, where response spans are categorized into predefined perspectives such as *cause*, *suggestion*, *experience*, *question*, and *information*;

(2)Perspective Summarization, which generates structured summaries that condense key insights while preserving essential details. Our approach integrates both tasks into a single pipeline, ensuring efficient classification and summarization of CQA responses.

By leveraging weak supervision and fine-tuning pre-trained models, we balance computational efficiency with adaptability, making the solution practical for real-world applications. This hybrid approach ensures that summaries retain critical information while being concise and easily understandable. This study makes the following key contributions:

[1]Models Code available: https://github.com/NadiaSaeed/PerAnsSumm2025/tree/main

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 331–339

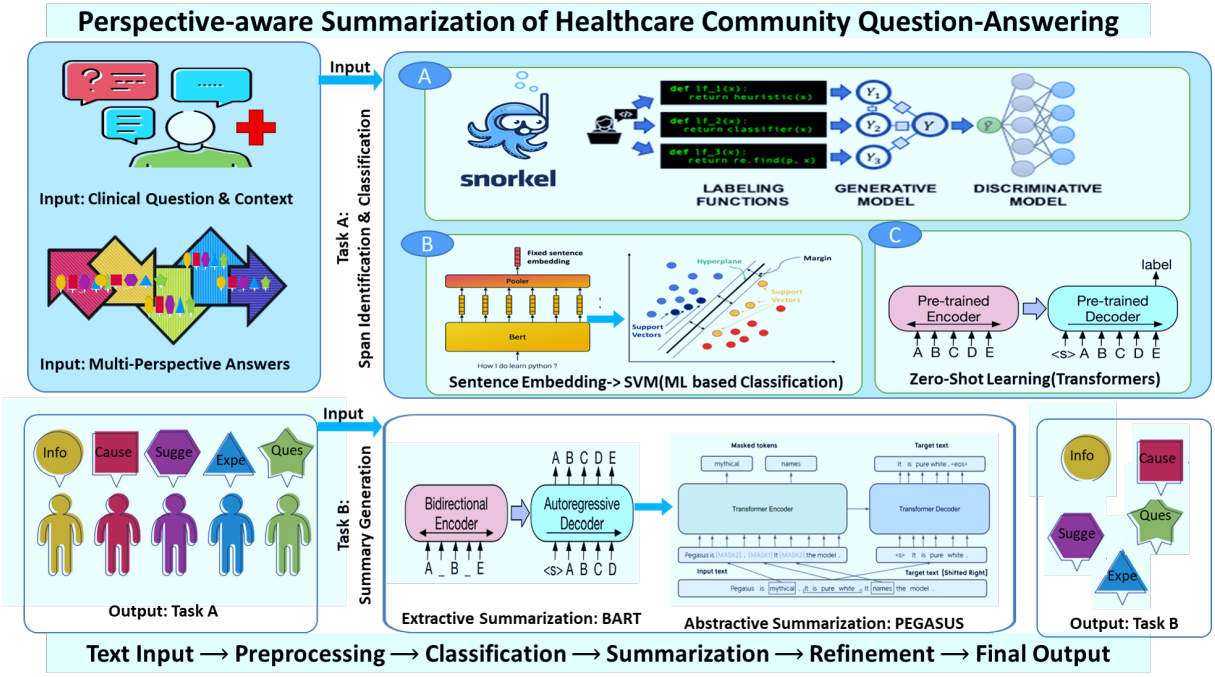

Figure 1: Hybrid workflow for perspective classification and summarization. Perspectives are classified using heuristic labeling (Snorkel), SVM-based classification, and a zero-shot model fallback. Summarization is performed in two stages: extractive (BART) and abstractive (Pegasus), integrating the context for a refined output.

1. A hybrid classification framework combining weak supervision, machine learning, and deep learning techniques for robust perspective identification.

2. A rule-based weak supervision method using Snorkel's labeling functions to generate high-quality probabilistic labels.

3. Feature extraction via sentence embeddings, leveraging transformer-based models to enhance classification.

4. A zero-shot learning (ZSL) classifier to handle unseen data without additional labeled examples.

5. A two-stage summarization pipeline that integrates extractive (BART) and abstractive (Pegasus) techniques for structured summaries.

6. A thorough evaluation demonstrating the effectiveness of our approach on real-world CQA datasets.

By combining classification with summarization, our method ensures that user-generated responses are structured, informative, and accessible. This enhances the usability of healthcare CQA forums and facilitates better decision-making for users.

2 Methodology

2.1 Task A: Perspective Classification

2.1.1 Problem Definition

Given a dataset of textual responses, our goal is to classify each response x_i into one of the predefined perspective categories (Naik et al., 2024):

$$\mathcal{P} = \{EXPE, INFO, CAUS, SUGG, QUES\} \quad (1)$$

Each response consists of multiple sentences, and our objective is to determine the category y_i by maximizing the conditional probability:

$$y_i = \arg\max_{p \in \mathcal{P}} P(p \mid x_i) \quad (2)$$

2.1.2 Hybrid Classification Pipeline

To achieve robust classification, we employ a three-stage hybrid pipeline:

1. Weak Supervision with Snorkel: Rule-based labeling functions assign probabilistic labels (Ratner et al., 2017; Fries et al., 2020; Rühling Cachay et al., 2021).

2. Supervised Learning with SVM: A Support Vector Machine (SVM) refines classification using sentence embeddings (Ala'M et al., 2023).

3. Zero-Shot Classification: A transformer model is applied when previous methods yield uncertain labels (Gera et al., 2022; Schopf et al., 2022).

2.1.3 Weak Supervision Using Snorkel

Manual annotation is time-intensive, so we use Snorkel's labeling functions (LFs) to generate weak labels based on pattern recognition:

$$LF(x) = \begin{cases} l_p, & \text{if pattern } p \text{ is found in } x \\ -1, & \text{otherwise} \end{cases} \quad (3)$$

where l_p is the assigned label, and -1 indicates abstention. To aggregate multiple weak labels, Snorkel's Label Model M estimates the true label distribution:

$$\hat{Y} = M(L) \quad (4)$$

where L represents the label matrix from different LFs. To efficiently label textual data, LFs based on regex patterns extracted from frequent words in the dataset. Each LF detects specific linguistic cues for perspective categories like EXPERIENCE or SUGGESTION. If a match is found, a label is assigned; otherwise, it abstains (as shown in Figure 1). The PandasLFApplier applies these LFs to generate a label matrix (Tok et al., 2021), which is then refined using Snorkel's Label Model to resolve conflicts and improve accuracy. This approach speeds up annotation while ensuring consistency through statistical aggregation.

2.1.4 Sentence Embeddings and SVM Classification

We convert textual responses into sentence embed

$$E(x) = \text{SentenceTransformer}(x) \quad (5)$$

These embeddings are used by an SVM classifier to enhance prediction accuracy:

$$\hat{y} = \text{SVM}(E(x)) \quad (6)$$

SVM is trained on sentence embeddings from a labeled dataset to classify text into perspective categories. Using a linear kernel, it learns decision boundaries in high-dimensional space. During inference, new sentences are embedded and classified based on their positions in the learned feature space.

2.1.5 Few-Shot Learning with Zero-Shot Classification

If Snorkel and SVM fail to provide a confident classification, we apply zero-shot learning (ZSL) using a transformer-based model:

$$P(p \mid x) = f_{ZSL}(x, \mathcal{P}) \quad (7)$$

where f_{ZSL} is a BART-based ZSL classifier, selecting the category with the highest probability. The ZSL model (facebook/bart-large-mnli) is applied using Hugging Face's pipeline (Lewis, 2019). When a sentence remains unclassified, the ZSL model evaluates the text without prior training on specific labeled data by comparing it to predefined perspective categories (*P*). It then assigns the most probable label by ranking all categories based on their semantic similarity to the input sentence. This ensures that even unseen or ambiguous responses can still be categorized effectively.

2.1.6 Final Classification Decision

The classification decision follows a hierarchical approach (as Shown in Figure 1 A, B and C):

$$y_i = \begin{cases} \hat{Y}_i, & \text{if } \hat{Y}_i \neq -1 \\ \text{SVM}(E(x_i)), & \text{if Snorkel abstains} \\ f_{ZSL}(x_i, \mathcal{P}), & \text{otherwise} \end{cases} \quad (8)$$

2.2 Task B: Hybrid Summarization

2.2.1 Overview

To generate high-quality summaries, we integrate extractive and abstractive techniques as shown in Figure 1 and 2:

2.2.2 Extractive Summarization Using BART

We use the facebook/bart-large-cnn model to extract salient content (Lewis, 2019):

$$S = \text{BART}(X) \quad (9)$$

where X is the concatenated input text and S is the generated extractive summary. The process involve following steps as shown in Figure 1 and 2:

1. Tokenizing input text with BART's tokenizer.
2. Using a task-specific prefix (summarize:).
3. Truncating text to 1024 tokens.
4. Applying beam search with: *max_length = 150, min_length = 50, length_penalty = 2.0, num_beams = 4*

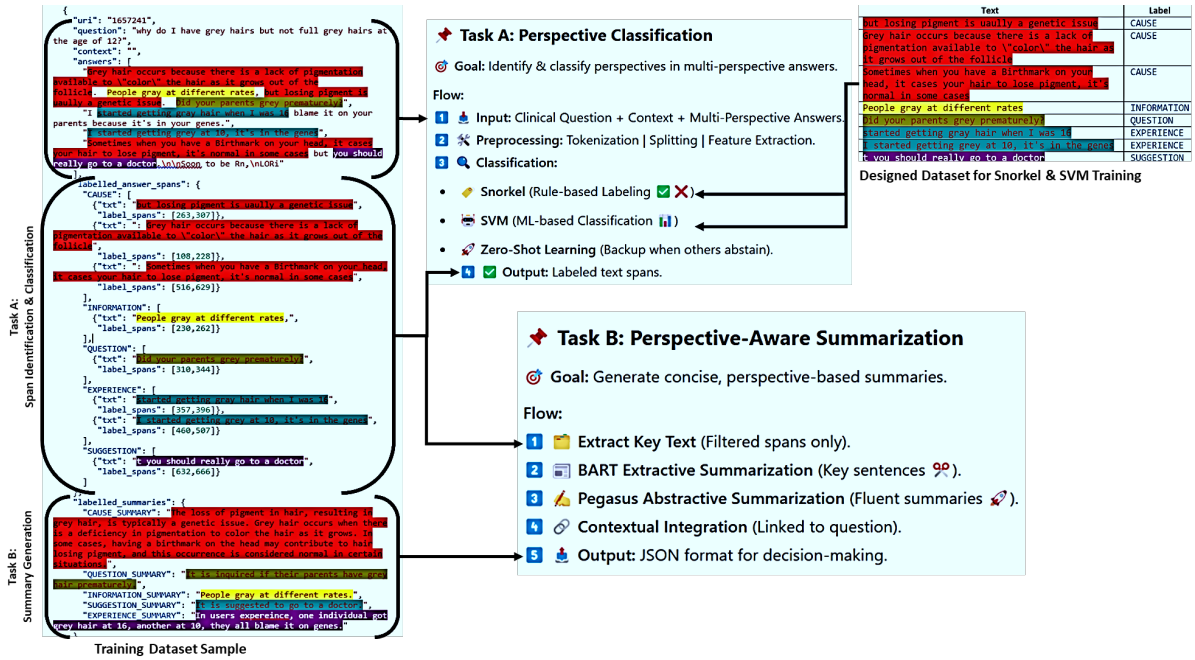

Figure 2: Training sample utilization for weak supervision. Known text spans from labeled data are used to train an SVM classifier, construct Snorkel labeling functions, and refine heuristic rules. The zero-shot model is excluded from direct training and is used as a fallback during classification.

2.2.3 Abstractive Refinement Using Pegasus

The extractive summary is refined with google/pegasus-xsum (Zhang et al., 2020):

$$S' = \text{Pegasus}(S) \tag{10}$$

where S' is the final abstractive summary. Refinement involves following steps:

1. Tokenizing extractive summaries.
2. Using the summarize: prompt.
3. Truncating input to 512 tokens.
4. Applying beam search with: *max_length = 100, min_length = 30, length_penalty = 1.8, num_beams = 6*

For our experiments, we utilize a dataset labeled with five perspective categories P in which *EXPE* and all others relate to the perspective of Experience, Information, Cause, Suggestion, and Question respectively (in Equation 1). Task A involves hierarchical classification, where unlabeled responses are processed using a combination of weak supervision, Support Vector Machines (SVM), and zero-shot learning (ZSL) (as Equation 8). We employ Snorkel for weak supervision, training its label model for 500 epochs to aggregate multiple labeling sources. Sentence embeddings are generated using SentenceTransformer (*all-MiniLM-L6-v2*) (Lewis, 2019), which serves as input to an SVM classifier trained with a linear kernel and default hyperparameters. For ZSL, we use Facebook's BART-Large-MNLI to directly infer category labels from textual descriptions.

Task B focuses on response structuring and refinement using transformer-based summarization models. We employ BART-Large-CNN for extractive summarization, generating concise representations of textual responses. To enhance coherence and fluency, we further refine these summaries using Pegasus-XSum (Zhang et al., 2020), an abstractive summarization model designed for extreme summarization tasks. The dataset for Task B consists of both labeled and unlabeled responses, allowing the models to learn from structured examples while refining free-text inputs. Our approach integrates both extractive and abstractive summarization techniques to ensure a well-structured and contextually rich final output.

3 Results and Discussion

In this study, we evaluated multiple hybrid models integrating Few-shot learning, weak supervision (Snorkel), and transformer-based architectures (BART, PEGASUS, and SVMs) for Span Identification & Classification (Task A) and Summarization (Task B). The primary objective was to assess the effectiveness of different learning paradigms in handling biomedical text processing challenges.

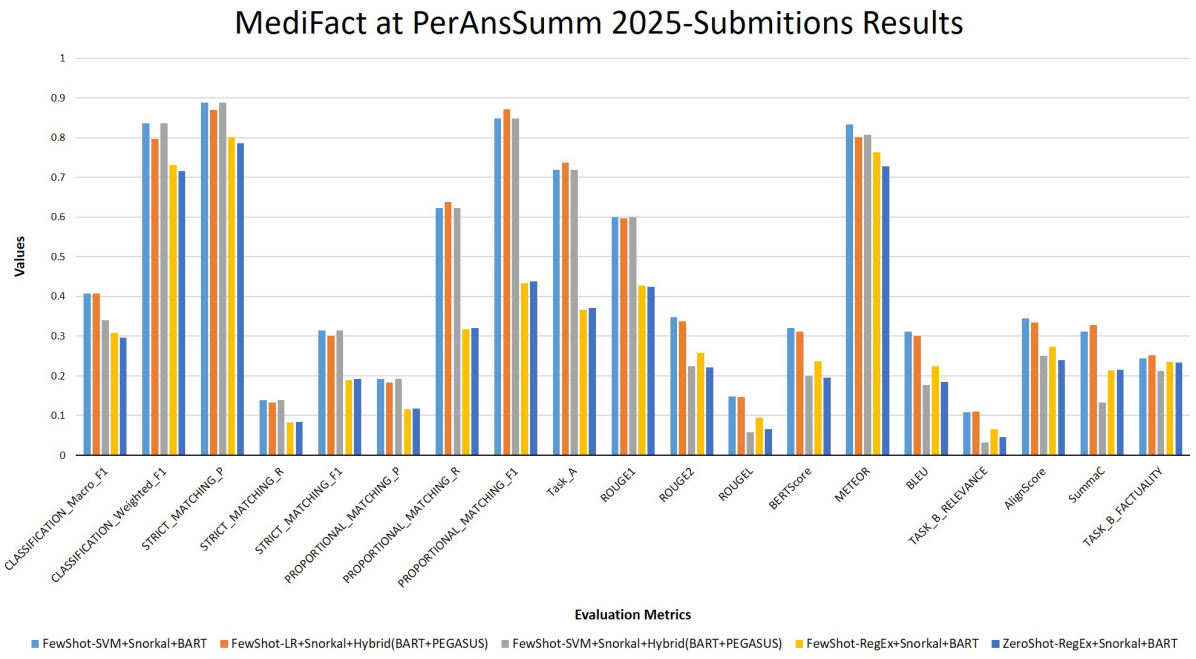

Figure 3: The comparative analysis of MediFact's submitted models on the PerAnsSumm Shared Task - CL4Health@NAACL 2025.

Task A (Perspective Classification) is evaluated using Macro-F1, Weighted-F1, Strict Matching (Precision, Recall, Weighted-F1), and Proportional Matching (Precision, Recall, Weighted-F1). Task B (Perspective Summarization) is assessed using ROUGE (R1, R2, RL), BLEU, Meteor, and BERTScore. The bar graph illustrates a comparative analysis of model performance across both tasks, highlighting strengths and areas for improvement in Figure 3.

3.1 Task A: Span Identification & Classification

The highest Weighted F1 score of 0.8361 was achieved by the *FewShot-SVM+Snorkel+BART* model, demonstrating its robustness in span identification and classification. Additionally, *FewShot-LR+Snorkel+Hybrid (BART+PEGASUS)* exhibited a competitive performance with an F1 score of 0.7961, while also achieving the best proportional match score (0.7373), indicating its capability to identify partially matched spans effectively.

Conversely, models relying on regular expressions (*FewShot-RegEx+Snorkel+BART* and *ZeroShot-RegEx+Snorkel+BART*) underperformed in classification, with *F1* scores of 0.7316 and 0.7161, respectively. This suggests that rule-based approaches lack the generalization needed for complex biomedical text extraction tasks.

3.2 Task B: Summarization Performance

The summarization capabilities of the models were evaluated using ROUGE-1 scores and factuality assessments. The *FewShot-SVM+Snorkel+BART* model achieved the highest ROUGE-1 score of 0.3485, indicating its effectiveness in generating relevant and concise summaries. Interestingly, *FewShot-LR+Snorkel+Hybrid (BART+PEGASUS)* demonstrated superior factuality (0.2897), suggesting that PEGASUS contributes to improved content faithfulness in biomedical text summarization.

Models utilizing regular expression-based classification (FewShot-RegEx and ZeroShot-RegEx variants) performed significantly lower across all summarization metrics. This highlights that statistical and deep learning-based models outperform rule-based approaches in abstractive summarization tasks.

3.3 Comparative Analysis of Model Performance

For a comprehensive evaluation, the combined average score (Task A + Task B performance) was computed for each model (Figure 3). *FewShot-SVM+Snorkel+BART* emerged as the best-performing approach with a combined score of 0.4077, followed by *FewShot-LR+Snorkel+Hybrid (BART+PEGASUS)* with 0.4070. The hybrid mod-

els demonstrated a balanced trade-off between classification accuracy and summarization quality, reinforcing the effectiveness of weak supervision with Snorkel and transformer-based architectures. In contrast, rule-based models (FewShot-RegEx & ZeroShot-RegEx variants) consistently showed inferior performance, suggesting that deep generative models are more suitable for biomedical NLP tasks requiring contextual understanding and content generation.

The experimental results demonstrate that a hybrid learning strategy combining weak supervision (Snorkel), Few-shot learning, and transformer models (BART, PEGASUS) yields optimal performance in biomedical span identification and summarization tasks. The proposed *FewShot-SVM+Snorkel+BART* model outperformed all other configurations, achieving the highest classification accuracy and summarization quality. These findings emphasize the importance of leveraging both structured supervision and deep generative architectures for enhancing biomedical text processing.

3.4 MediFact Performance in PerAnsSumm Shared Task

MediFact secured a position among the **top 12 teams** in the **PerAnsSumm Shared Task - CL4Health@ NAACL 2025**. The final results were officially reported by the task organizers on the shared task website.[2]

In Figure A.1, MediFact's performance across various evaluation metrics demonstrates strong classification capabilities, achieving a competitive Weighted F1-score of 0.8887. However, the Macro F1-score (0.8361) suggests room for improvement in handling class imbalances.

In the matching task, MediFact attains a high Proportional Matching Recall (0.8493), indicating effective identification of relevant matches. However, the Strict Matching Precision (0.1383) and Strict Matching F1 (0.1921) highlight challenges in reducing false positives.

For summarization, the model achieves a BERTScore of 0.8336, reflecting strong semantic alignment. However, lower ROUGE scores (R1: 0.3485, R2: 0.1475, RL: 0.3212) and BLEU (0.1078) suggest the need for more accurate and concise text generation.

Factual consistency metrics, such as AlignScore (0.3121) and Factuality Score (0.2784), indicate areas for improvement in ensuring reliable summarization. Future work should focus on enhancing precision in matching, optimizing summarization coherence, and strengthening factual alignment to ensure more trustworthy outputs.

4 Conclusion

This research introduces a modular and resource-efficient approach for perspective-aware classification and summarization. We combine weak supervision, machine learning, and pre-trained transformers to balance accuracy and computational cost (Ratner et al., 2017; Rueping, 2010; Lewis, 2019). Instead of training a model from scratch, we fine-tune pre-trained models on our dataset. This approach reduces resource demands and speeds up adaptation to new tasks.

One major motivation for our method is overcoming computational limitations. Training large models from the ground up requires extensive hardware and time (Touvron et al., 2023; Floridi and Chiriatti, 2020; Lewis, 2019). To handle this, we use pre-trained models that can be fine-tuned efficiently. We also apply weak supervision with heuristic labeling, reducing the need for manual annotation (as shown in Figure 2). This makes our approach scalable and practical.

Our study shows that strong results can be achieved even with limited resources. We propose a modular and adaptable solution that does not depend entirely on commercial large language models (LLMs). While proprietary models offer high performance, they lack flexibility and accessibility (Team et al., 2023; Lee and Hsiang, 2020). Instead, we demonstrate how open-source models and targeted fine-tuning provide robust results without heavy computational costs.

In conclusion, this work highlights the importance of resource-aware AI research. It proves that effective NLP solutions can be built without expensive models. Open-source tools played a key role in making this study possible (Wolf, 2019; Lewis, 2019; Zhang et al., 2020). By selecting the right model and designing a modular workflow, we achieve high-quality classification and summarization even with limited resources. This research encourages future work to focus on scalable, adaptable, and cost-effective AI solutions instead of relying solely on commercial LLMs.

[2]PerAnsSumm Shared Task - CL4Health@ NAACL 2025: https://peranssumm.github.io/docs/#leaderboard

5 Limitations

Weak supervision relies on heuristic rules, which may introduce bias or inconsistencies. While pre-trained models reduce the computational burden, further improvements can be made. Future research can explore lightweight architectures, efficient fine-tuning methods (such as LoRA (Hu et al., 2021) and quantization (Yang et al., 2019)), and retrieval-augmented generation (RAG) (Notarangelo et al., 2016) to handle unseen perspectives.

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.

Al-Zoubi Ala'M, Antonio M Mora, and Hossam Faris. 2023. A multilingual spam reviews detection based on pre-trained word embedding and weighted swarm support vector machines. *IEEE Access*.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2020. Trove: Ontology-driven weak supervision for medical entity classification. *arXiv preprint arXiv:2008.01972*.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. *arXiv preprint arXiv:2210.17541*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Emily Jiang. 2024. *Clinical Question-Answering over Distributed EHR Data*. Ph.D. thesis, Massachusetts Institute of Technology.

Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. 2024. Question answering systems for health professionals at the point of care—a systematic review. *Journal of the American Medical Informatics Association*, 31(4):1009–1024.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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.

Luigi D Notarangelo, Min-Sung Kim, Jolan E Walter, and Yu Nee Lee. 2016. Human rag mutations: biochemistry and clinical implications. *Nature Reviews Immunology*, 16(4):234–246.

Alexander J Ratner, Stephen H Bach, Henry R Ehrenberg, and Chris Ré. 2017. Snorkel: Fast training set generation for information extraction. In *Proceedings of the 2017 ACM international conference on management of data*, pages 1683–1686.

Stefan Rueping. 2010. Svm classifier estimation from group probabilities. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 911–918.

Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. 2021. End-to-end weak supervision. *Advances in Neural Information Processing Systems*, 34:1845–1857.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pages 6–15.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wee Hyong Tok, Amit Bahree, and Senja Filipi. 2021. *Practical Weak Supervision*. " O'Reilly Media, Inc.".

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. 2019. Quantization networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7308–7316.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

A MediFact Performance Detail

This section provides additional insights into MediFact's performance, complementing the discussion in Section 3. Figure A.1 presents a detailed breakdown of evaluation metrics across different tasks, including classification, matching, and summarization. The results highlight MediFact's strong classification capabilities, particularly in achieving a competitive Weighted F1-score. However, performance in strict matching and summarization coherence suggests potential areas for improvement. These findings provide direction for future optimizations, focusing on enhanced precision and factual consistency.

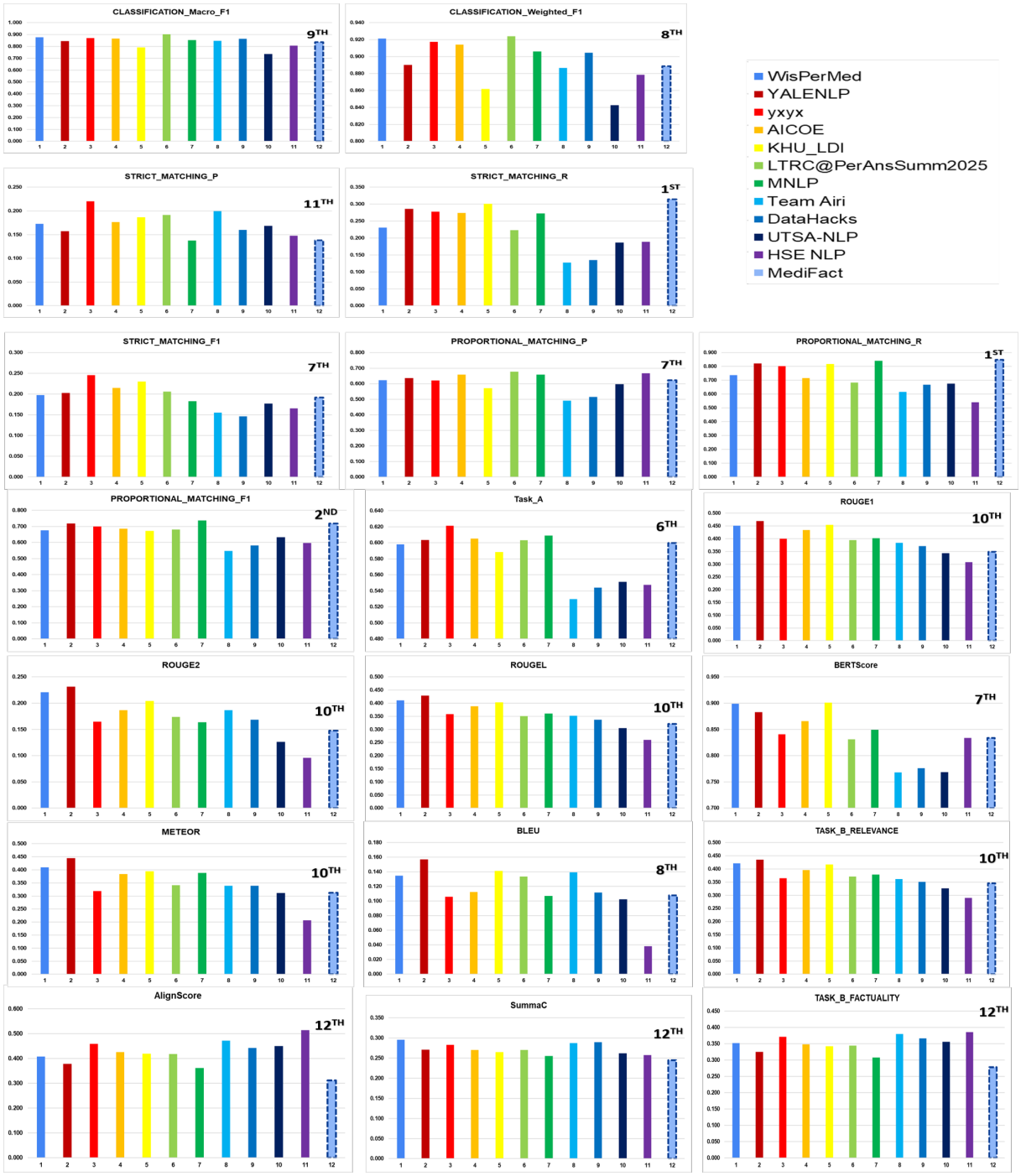

Figure A.1: Comparative Performance Analysis of MediFact Among the Top 12 Models in the PerAnsSumm Shared Task CL4Health@NAACL 2025.