# MNLP at PerAnsSumm: A Classifier-Refiner Architecture for Improving the Classification of Consumer Health User Responses

**Jooyeon Lee** and **Luan Huy Pham** and **Özlem Uzuner**
George Mason University, USA
{jlee252,lpham6,ouzuner}@gmu.edu

## Abstract

Community question-answering (CQA) platforms provide a crucial space for users to share experiences, seek medical advice, and exchange health-related information. However, these platforms, by nature of their user-generated content as well as the complexity and subjectivity of natural language, remain a significant challenge for tasks related to the automatic classification of diverse perspectives. The PerAnsSumm shared task involves extracting perspective spans from community users' answers, classifying them into specific perspective categories (Task A), and then using these perspectives and spans to generate structured summaries (Task B). Our focus is on Task A. To address this challenge, we propose a Classifier-Refiner Architecture (CRA), a two-stage framework designed to enhance classification accuracy. The first stage employs a Classifier to segment user responses into self-contained snippets and assign initial perspective labels along with a binary confidence value. If the classifier is not confident, a secondary Refiner stage is triggered, incorporating retrieval-augmented generation to enhance classification through contextual examples. Our methodology integrates instruction-driven classification, tone definitions, and Chain-of-Thought (CoT) prompting, leading to improved F1 scores compared to single-pass approaches. Experimental evaluations on the Perspective Summarization Dataset (PUMA) demonstrate that our framework improves classification performance by leveraging multi-stage decision-making. Our submission ranked among the top-performing teams, achieving an overall score of 0.6090, with high precision and recall in perspective classification.

## 1 Introduction

Community question-answering (CQA) forums have emerged as a pivotal medium for individuals seeking diverse perspectives on health-related issues, encompassing personal anecdotes, medical suggestions, factual information, and experiential insights. While these platforms offer a wealth of user-generated knowledge, extracting structured, perspective-specific content from such discussions remains a complex challenge due to linguistic variability and overlapping semantic cues. Traditional single-pass classification systems often misclassify or overlook nuanced snippets, leading to incomplete or misleading results. These limitations are especially consequential in the healthcare domain, where accurate categorization of user responses can influence subsequent experiences, diagnosis, and/or recommendations (Agarwal et al., 2025).

Our approach, tested on the PUMA (Naik et al., 2024) dataset, demonstrates robust performance across macro-F1, weighted-F1, strict, and proportional evaluation metrics. In particular, we highlight the effectiveness of tone definition and CoT prompting, which bolster classification reliability and interpretability. Moreover, we compare leading large language models (LLMs), specifically GPT-4o, Claude 3, and o1-preview, under various experimental configurations, showing that multi-stage decision-making strategies can streamline complex classification tasks in CQA settings across a variety of LLMs.

## 2 Related Work

Research in multi-stage classification has demonstrated that iterative refinement can improve the accuracy and reliability of NLP models (Zhang et al., 2020). In the context of few-shot or low-resource scenarios, Zhao et al. introduced calibration strategies to bolster classification robustness, while Lewis et al. showed that multi-step prompting methods significantly enhance model performance. Moreover, the concept of CoT prompting has been explored by Wei et al. to elicit more transparent reasoning processes in LLMs. CoT is also re-



*Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 349–358

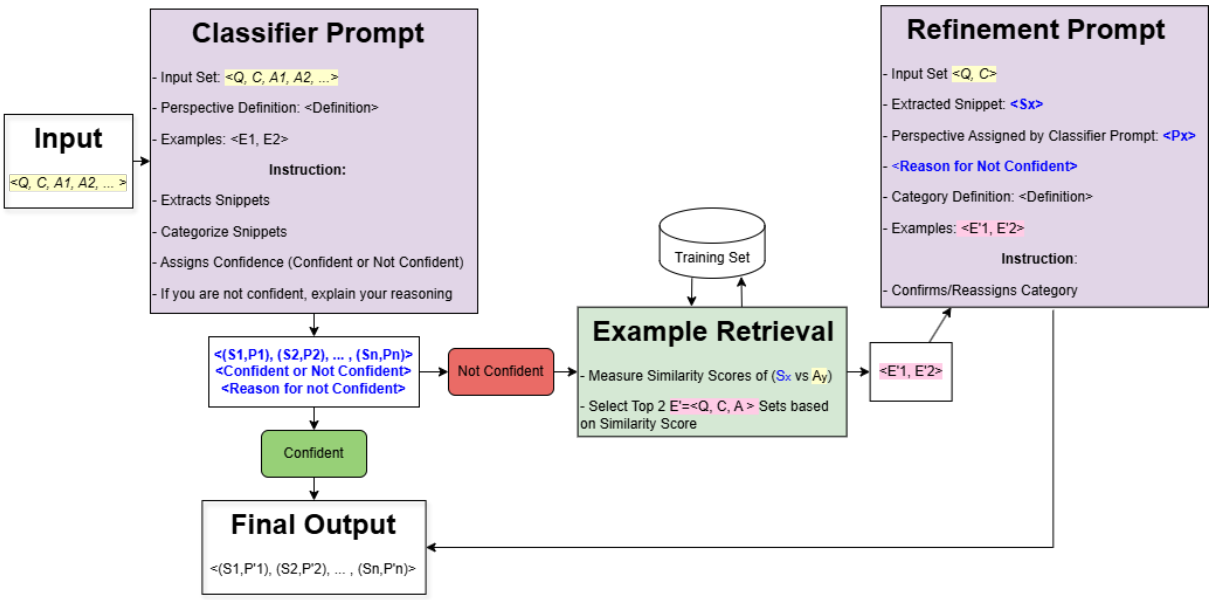


Figure 1: Classifier-Refiner Architecture. Yellow highlight denotes input values extracted from the dataset, including the Question (Q), Context (C), and User Responses (A). Blue text represents the Classifier's output, which subsequently serves as input for the Refiner. Pink highlight indicates the output of Example Retrieval with RAG, which is later incorporated into the Refiner's input.

ceiving attention from Consumer Health Question Answering (CHQA) domain research (Lee et al., 2024).

Recent advances in retrieval-based classification have leveraged the idea of combining external knowledge with model predictions for better handling of uncertain cases (Lewis et al., 2020). Gao et al. demonstrated that retrieval-based prompting can provide relevant context from a structured dataset, thereby improving model understanding. Our method follows these trends by integrating a retrieval-augmented classification and refinement mechanism, in which the system references training data to refine ambiguous labels. This combination of iterative refinement and retrieval augmentation offers a robust alternative to single-pass classification pipelines (Izacard and Grave, 2020).

## 3 Methodology

### 3.1 Task Definition

Given a user's question and a corresponding user-generated health response, we segment the response into self-contained snippets. Each snippet must be assigned one of the following categories (corresponding to the PUMA annotated categories), which are defined by Agarwal et al.:

1. EXPERIENCE (<tone: Personal, Narrative>): Individual experiences or firsthand insights.
2. INFORMATION (<tone: Informative, Educational>): Factual statements or knowledge about health conditions.
3. CAUSE (<tone: Explanatory, Causal>): Explanations of why a condition or symptom might occur.
4. SUGGESTION (<tone: Advisory, Recommending>): Advice or recommendations for resolving or improving a health-related issue.
5. QUESTION (<tone: Seeking Understanding>): Direct inquiries seeking information or clarity.

### 3.2 Dataset

The dataset used in this study is the PUMA dataset, created by independent researchers for the PerAnsSumm shared task (Naik et al., 2024). PUMA was derived from the L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi-part) corpus [1] which contains data up to October 2007, consisting of 3,167 CQA threads. Specifically, Naik et al. filtered Yahoo! Answers for healthcare-related content, randomly selecting 10,000 questions each with up to 10 answers. These records covered a variety of medical topics, including Diabetes, Dental, and Cancer, ensuring broad coverage of health-related discussions.

From this curated set, the authors further refined and annotated specifically for the PerAnsSumm task. The final version of PUMA was then split

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11&guccounter=1

into three subsets for this shared task: a training set of 2,236 question-answer pairs, a validation set of 959 pairs, and a test set of 50 pairs. The annotations were performed by three fluent English speakers (one master's student, one research assistant, and one native English-speaking volunteer) who identified perspective-specific spans in each answer. These spans were categorized into five distinct labels: Cause, Suggestion, Experience, Question, and Information. Multiple annotators cross-validated the labels to ensure reliability and consistency.

### 3.3 Classifier-Refiner Framework

**Classifier** In the first stage, we use a prompting technique with a language model (e.g., a GPT-based or other LLM) to process each user response and produce potential snippet boundaries, as well as initial category labels. This Classifier is instructed to highlight text segments that can meaningfully stand alone. The output of the prompt follows the JSON format:

```
[
  {
    "text": "<Extracted Snippet>",
    "confidence": "CONFIDENT",
    "reason not confident": "",
    "category": "INFORMATION"
  },
  ...
]
```

In cases where the LLM is uncertain about the correct category, "confidence" is set to "NOT_CONFIDENT", and an additional "reason_not_confident" field is provided.

**Refiner** The Refiner operates by leveraging a retrieval-augmented generation (RAG) mechanism, which enhances classification accuracy by incorporating contextual examples from the training set. Specifically, when triggered, the Refiner first retrieves the two most similar sentences from the training set using a sentence similarity model (all-MiniLM-L6-v2) (Wang et al., 2020). This allows us to use different examples from the Classifier, thus we expect different results from the Classifier output. The all-MiniLM-L6-v2 model was used in an unsupervised approach in this task. It is a lightweight transformer-based model for semantic similarity comparison, optimized for model size and faster inference. The model has 66 million parameters, compressed in a Student-Mimicking Teacher network relationship. By utilizing self-attention distribution, the training of the student model is guided using the teacher's last layer, ensuring effective and flexible results across 12 different languages.

These retrieved examples are then inserted into the Refiner prompt as few-shot examples, allowing the model to refine the classification by comparing the uncertain snippet with previously labeled cases. This iterative approach ensures that the classification process incorporates relevant training instances, thereby improving overall classification reliability and mitigating ambiguity in nuanced cases. The Refiner finally returns JSON format result:

```
{
  "previous_category":
        "<category from previous step>",
  "confidence":
        "<CONFIDENT or NOT_CONFIDENT>",
  "refinement_reasoning": "<brief explanation>",
  "refined_category": "<final label>"
}
```

By incorporating the different context and referencing prior examples, this step significantly reduces misclassification in borderline scenarios.

### 3.4 Language Models

We experimented with multiple language models to evaluate the effectiveness of different architectures in classification refinement:

**GPT-4o** An omni-modal autoregressive model capable of processing text, audio, image, and video inputs while generating text, audio, and image outputs. GPT-4o demonstrates exceptional multilingual proficiency and enhanced computational efficiency, making it significantly faster and more cost-effective compared to GPT-4 Turbo. Its advanced speech-to-text capabilities and safety alignment mechanisms enhance reliability in consumer health discussions by reducing misinformation and bias. This model was evaluated in multiple prompting setups, including single-prompt classification, instruction-based CRA, and CoT refinement. (OpenAI et al., 2024).

**Claude 3** Developed by Anthropic, Claude 3 (Opus, Sonnet, and Haiku) represents a family of LLMs optimized for cognitive reasoning, nuanced contextual understanding, and expansive token processing (up to 1 million tokens in specialized tasks). Claude 3 Opus demonstrated self-awareness in controlled testing environments, particularly in needle-

| Module | Feature | Example |
|---|---|---|
| Classifier | Extracted Text | If you took the prescribed antibiotics as recommended you are no longer contagious. |
| | Perspective | CAUSE |
| | confidence | NOT CONFIDENT |
| | reason not confident | Could be either information or suggestion, as it implies both diagnosis and recommendation to investigate further |
| Refiner | refinement reasoning | After reviewing the full context and similar examples, this statement is clearly providing factual information about contagiousness in relation to antibiotic treatment, similar to example 4 'For the first 24 to 48 hours after you start taking an antibiotic, you are still contagious.' The statement is not explaining why something happens (CAUSE), but rather stating a medical fact about contagiousness after antibiotic treatment. |
| | Refined category | INFORMATION |

Table 1: Example Output.

in-a-haystack tasks, making it an ideal candidate for refining ambiguous classifications in CQA settings. This model was primarily used in CRA with tone definitions, providing insights into subjective aspects of user responses (Anthropic, 2024).

**o1-preview** A state-of-the-art language model developed by OpenAI, extensively tested on complex reasoning tasks spanning multiple domains, including computer science, mathematics, medicine, linguistics, and social sciences. The model exhibits superior performance in competitive programming, high school-level mathematical reasoning, and radiology report generation. Additionally, o1-preview excels in natural language inference tasks, sentiment analysis, and financial modeling. This model was particularly effective in CQA classification due to its ability to integrate contextual cues across diverse perspectives (Zhong et al., 2024).

## 4 Results

We adopted macro-F1 (C-MF1), weighted-F1 (C-WF1), Strict Precision/Recall/F1 (S-P, S-R, S-F1), and Proportional Precision/Recall/F1 (P-P, P-R, P-F1) - which are the official metrics used for the PerAnsSumm shared task - to capture a range of performance aspects. C-MF1 and C-WF1 are Macro-averaged and weighted F1 scores for the classification task, focusing on how well the system balances performance across categories. S-P, S-R, S-F1 are Strict metrics to gauge performance under the assumption that each snippet clearly belongs to one category. P-P, P-R, P-F1 are proportional metrics to evaluate partially correct classifications, recognizing that user-generated health content often spans multiple categories or perspectives.

### 4.1 Evaluation

Single-Prompt vs. CRA: The single-pass methods (rows 1-2) show lower C-MF1 and C-WF1 scores. Once the CRA approach is introduced (rows 3-6), the metrics consistently improve, indicating the effectiveness of a multi-stage classification pipeline.

Tone Definition Impact: Including explicit tone definitions tends to increase both strict and proportional F1 scores by helping the model distinguish subtle differences (e.g., between EXPERIENCE vs. INFORMATION or SUGGESTION vs. INFORMATION).

CoT Influence: CoT reasoning further refines the model's decision-making, especially in complex or overlapping perspectives. This is reflected in higher macro-F1 scores for the CRA + CoT configurations.

o1-preview (MNLP Final Submission) achieves the best overall score of 0.6090, setting a strong benchmark. Notably, its P-R (0.8406) and P-F1 (0.7382) values are significantly higher than the other configurations.

In the broader context of the PerAnsSumm shared task, our team (MNLP) ranks among the top five, as shown in Table 3. Although not topping every sub-metric, MNLP's approach demonstrates a balanced performance across multiple dimensions, showcasing the strength of the CRA pipeline.

## 5 Discussion

The results underscore several key insights.

| Idx | Model | Method Description | C-MF1 | C-WF1 | S-P | S-R | S-F1 | P-P | P-R | P-F1 | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT 4o | Single Prompt | 0.7985 | 0.8651 | 0.1459 | 0.1448 | 0.1453 | 0.4773 | 0.6013 | 0.5322 | 0.5142 |
| 2 | GPT 4o | Single prompt+ Removed Question | 0.6991 | 0.8101 | 0.1438 | 0.0800 | 0.1028 | 0.4508 | 0.5775 | 0.5064 | 0.4731 |
| 3 | GPT 4o | CRA+Instr+tone def | 0.8126 | 0.8771 | 0.1852 | 0.1429 | 0.1613 | 0.5874 | 0.6342 | 0.6099 | 0.5494 |
| 4 | GPT 4o | CRA+ CoT | 0.8292 | 0.8879 | **0.1896** | 0.1524 | 0.1690 | 0.5963 | 0.5942 | 0.5953 | 0.5507 |
| 5 | GPT 4o | CRA+ CoT+ tone def | 0.8387 | 0.8948 | 0.1809 | 0.1371 | 0.1560 | 0.5925 | 0.6005 | 0.5965 | 0.5491 |
| 6 | Claude 3 | CRA+Instr+tone def | 0.7963 | 0.8718 | 0.1168 | 0.0914 | 0.1026 | 0.6113 | 0.3847 | 0.4722 | 0.4822 |
| 7 | o1-preview | CRA+Instr+tone def | **0.8524** | **0.9061** | 0.1376 | **0.2724** | **0.1829** | **0.6580** | **0.8406** | **0.7382** | **0.6090** |

Table 2: Task A Results.

| Team Name | C-MF1 | C-WF1 | S-P | S-R | S-F1 | P-P | P-R | P-F1 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| yxyx | 0.8697 | **0.9173** | **0.2205** | 0.2781 | **0.2460** | 0.6215 | 0.8029 | 0.7006 | 0.6213 |
| **MNLP** | 0.8524 | 0.9061 | 0.1376 | 0.2724 | 0.1829 | 0.6580 | **0.8406** | **0.7382** | **0.6090** |
| AICOE | 0.8656 | 0.9140 | 0.1765 | 0.2743 | 0.2148 | 0.6597 | 0.7159 | 0.6866 | 0.6052 |
| YALENLP | 0.8439 | 0.8902 | 0.1571 | **0.2857** | 0.2027 | 0.6372 | 0.8218 | 0.7178 | 0.6036 |
| LTRC | **0.9033** | 0.9239 | 0.1915 | 0.2229 | 0.2060 | **0.6774** | 0.6833 | 0.6803 | 0.6034 |

Table 3: Top 5 Team Results for Task A

### 5.1 Two-Stage Decision-Making Improves Reliability

Incorporating a secondary Refiner model significantly reduces classification uncertainty. In single-pass systems, difficult or ambiguous snippets often receive incorrect labels. The Refiner leverages additional context (e.g., new examples, reason not confident) to resolve ambiguities.

### 5.2 Role of Tone Definitions

Empirical evidence suggests that explicitly including tone information—such as labeling a snippet as 'personal/narrative' or 'informative/educational'—guides the model to distinguish subtle semantic differences between EXPERIENCE and INFORMATION categories. This additional guidance appears to yield more consistent performance.

### 5.3 Impact of CoT

CoT prompts give the language model intermediate reasoning steps, leading to more thorough snippet analysis. While adding CoT marginally increases computational cost, it provides a measurable boost in precision, particularly for borderline cases where multiple categories overlap. These findings align with prior research on the benefits of explicitly prompting large models to articulate their reasoning steps (Wei et al., 2022; OpenAI et al., 2024).

### 5.4 Model Comparison

As outlined in the Methodology section, three models (GPT-4o, Claude 3, and o1-preview) were evaluated under configurations tailored to multi-stage classification in healthcare QA. Below, we highlight the core empirical findings and discuss how each model responded to different prompt designs.

#### 5.4.1 Prompting Strategies and Performance

**GPT-4o.** GPT-4o's best performance emerged from a "CRA + CoT" setup, yielding an overall score of 0.5507. Removing the explicit CoT steps and instead relying on "Instruction + Tone Definition" resulted in only a marginal decrease (0.5494). This near-parity suggests that GPT-4o effectively processes step-by-step reasoning, even without direct user guidance, provided instructions remain sufficiently structured and detailed.

**Claude 3.** For consistency with o1-preview, Claude 3 was primarily tested under "CRA + Tone

Definition." The model's performance varied more substantially than GPT-4o, likely reflecting Claude 3's sensitivity to domain-specific nuances and question complexity. Despite such fluctuations, Claude 3 did exhibit strong alignment with user instructions, consistent with its "Constitutional AI" training paradigm—and demonstrated robust comprehension in tasks demanding nuanced responses. Future refinements or domain-specific tuning may further enhance its stability.

**o1-preview.** Unlike GPT-4o, o1-preview internally implements COT reasoning and prohibits external user-directed CoT prompts. Consequently, we restricted prompts to "CRA + Tone Definition" for a fair comparison. Under these conditions, o1-preview achieved the highest performance across our evaluation metrics. Its internally generated reasoning appears mature enough to parse complex instructions, enforce safety considerations, and incorporate tonal guidelines, without requiring explicit step-by-step instructions from the user.

#### 5.4.2 Observations and Implications

**Internal vs. User-Supplied CoT** GPT-4o benefits from explicit CoT prompts, whereas o1-preview inherently manages its own CoT. The near-equivalence of GPT-4o's "CRA + CoT" (0.5507) and "CRA + Instruction + Tone Definition" (0.5494) underscores that well-crafted instructions can closely approximate explicit CoT. By contrast, o1-preview excels through its internalized reasoning approach, obviating the need for user-provided CoT altogether. This design choice can be seen as advantageous for developers seeking a lower cognitive overhead when engineering prompts, although it also reduces direct user control over the model's reasoning process.

**Tone Definition and Stylistic Constraints** "Tonal" or "stylistic" labels did not show significant improvement with GPT-4o. However, these could be mitigated through additional fine-tuning or domain adaptation.

**Practical Considerations for Multi-stage Healthcare QA** Real-world healthcare QA systems demand predictable model behavior and ease of prompt design. While GPT-4o may need user-defined CoT to reach peak performance, o1-preview's autonomous internal reasoning streamlines the developer experience. Choices between these models must weigh the trade-off between direct CoT control (GPT-4o) and fully internalized reasoning (o1-preview) against the complexity of the tasks at hand.

In summary, GPT-4o demonstrated strong capability with user-supplied CoT prompts, whereas o1-preview's internally managed reasoning and refined alignment led to consistently higher performance without explicit CoT instructions. Claude 3, meanwhile, remained competitive but was more sensitive to prompt variations. These findings underscore the importance of prompt engineering, built-in CoT, and alignment strategies in deploying LLMs for complex tasks such as multi-stage classification in healthcare QA.

#### 5.4.3 Potential Explanations for o1-preview's Superior Results

o1-preview's top performance may stem from both architectural refinements and advanced alignment protocols. First, o1-preview likely benefits from curated training data tailored to tasks requiring fine-grained reasoning and tone management. Second, improved alignment techniques (building on GPT-4o's foundation) may enhance the balance between correctness, recall, and user-centric instructions. Notably, o1-preview's resilience to prompt alterations, including variations such as "CRA + CoT + tone def," suggests that it integrates complex instructions and stylistic requirements without sacrificing coherence.

Taken together, the differing performances of GPT-4o, Claude 3, and o1-preview highlight the interplay between model architectures, alignment strategies, and prompt design. While both GPT-4o and Claude 3 demonstrate robust capabilities under certain configurations, o1-preview's refined integration of reasoning and tone guidance appears to yield superior classification outcomes.

### 5.5 Error Analysis

Although the two-stage classification approach proved effective overall, a closer inspection of the 21 instances where the Refiner was triggered (out of 1039 total snippets) offers valuable insights into recurring error patterns and the advantages of iterative refinement. Table 4 presents representative examples where the Classifier's initial label differed from the Refiner's final judgment, along with corresponding reasoning ("thought") from both stages. Three principal themes emerged:

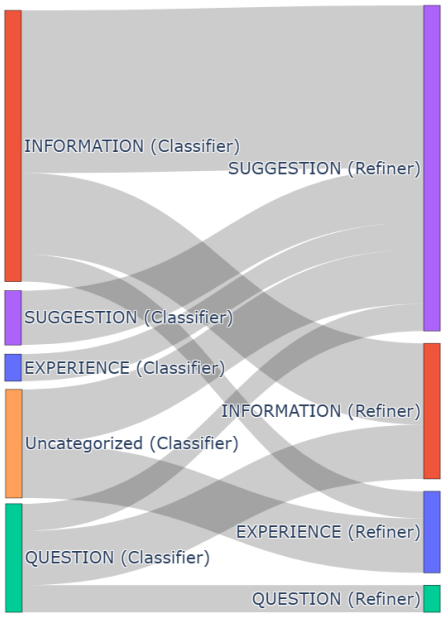


Figure 2: Sankey diagram illustrating the flow of snippet labels from the Classifier to the Refiner. Each node represents a classification label, with left-side nodes corresponding to the Classifier's initial labels and right-side nodes representing the Refiner's final labels. The thickness of each link is proportional to the number of snippets that transitioned between categories. Notably, the Refiner frequently corrected INFORMATION to SUGGESTION and reclassified certain QUESTION and EXPERIENCE snippets, indicating that these categories were more prone to initial misclassification. This visualization highlights the value of iterative refinement in improving classification accuracy.

#### 5.5.1 Reclassification of Short or Polite Snippets

In multiple cases, polite expressions or brief well-wishes (e.g., "Be well," "good luck") were initially labeled as INFORMATION or left as Uncategorized by the Classifier. The Refiner, however, recognized these statements as advisory or encouraging in nature, as aligning with training set (e.g., I hope that you keep on going, and that you realize how important you are to our world.: SUGGESTION) thereby reassigning them to SUGGESTION. This suggests the Classifier's tendency to default to INFORMATION when textual clues are minimal, whereas the Refiner incorporates context (e.g., prior labeled examples) to identify the statement's tone and intent.

#### 5.5.2 Distinguishing Rhetorical Questions from Genuine Questions

Several snippets contained rhetorical or illustrative "questions" (e.g., "Is it because of the antibiotics?") that the Classifier labeled as QUESTION. Upon refinement, these snippets were deemed INFORMATION once the system determined they functioned more as explanatory remarks rather than genuine queries. This underscores the importance of discourse context in discerning the pragmatic function of a statement.

#### 5.5.3 Personal Commentary and Narrative Content

Certain snippets expressing personal opinions or narrative remarks were originally labeled as INFORMATION or EXPERIENCE. The Refiner identified that these statements often warrant EXPERIENCE, particularly when they reflect an individual's personal viewpoint or emotive stance rather than a factual claim. For instance, "What a great question." was recognized as more personal/relational than purely informational, leading to reclassification from INFORMATION to EXPERIENCE.

#### 5.5.4 Implications for Multi-stage Classification

These illustrative examples highlight how the Refiner adds a crucial layer of context-awareness, correcting labels when the Classifier defaults to INFORMATION or encounters snippets with ambiguous linguistic cues. Notably, the number of triggers (21) is small relative to the overall dataset (N=1039), yet it plays a disproportionate role in improving the accuracy of borderline or confusing snippets.

#### 5.5.5 Practical Outcomes

Practical outcomes of this CRA include:

- **Reduced Misclassification**: The second stage captures subtle differences (e.g., well-wishes vs. factual statements) that single-pass models often overlook.
- **Context Utilization**: By referencing the full user response or previously labeled snippets, the Refiner more accurately infers intent behind brevity, politeness, or indirect language.
- **Efficiency Consideration**: Triggering the Refiner only for ambiguous or contradictory Classifier outputs mitigates computational overhead compared to always running two stages.

In summary, this error analysis underscores that ambiguous linguistic cues, limited context in short

| Classifier Result | Refiner Result | Error Case Examples |
|---|---|---|
| EXPERIENCE | SUGGESTION | A snippet initially labeled EXPERIENCE was reclassified after the Refiner noted its advisory content ("... a personal request aiming to persuade selection..."), fitting better in SUGGESTION. |
| INFORMATION | SUGGESTION | The statement "Best of good luck from Italy" was interpreted as INFORMATION until the Refiner interpreted it as a supportive or advisory comment, upgrading it to SUGGESTION. |
| QUESTION | INFORMATION | Rhetorical questions (e.g., "can you picture a fish out of the water?") were reframed as INFORMATION once the Refiner deduced they conveyed illustrative content rather than genuinely seeking an answer. |
| UNCATEGORIZED | SUGGESTION | Extremely short snippets like "Geez! How terrible for her!!! Good luck to her & you." lacked a Classifier label. 'good luck' serves as a supportive and advisory statement, the Refiner assigned it to SUGGESTION. |

Table 4: Example cases of Refiner modifying the classification label.

snippets, and the pragmatic function of rhetorical questions remain primary sources of error. However, iterative refinement significantly alleviates these issues, resulting in higher fidelity categorizations. Future enhancements might include more explicit discourse modeling or leveraging external knowledge bases for context augmentation, particularly for healthcare-related queries, where subtle nuances can have significant implications for the quality of advice or information provided.

## 6 Conclusion

In this paper, we presented a CRA that addresses the intrinsic complexity of health-related user-generated content by employing a two-stage decision-making pipeline. Our experiments on the PUMA dataset, curated for the PerAnsSumm shared task (Task A: span extraction and perspective classification), underscored how iterative refinement, retrieval-augmented generation, and CoT prompting collectively enhance classification confidence and accuracy. Comparative analyses across leading LLMs (GPT-4o, Claude 3, and o1-preview) revealed that multi-stage approaches deliver more robust handling of ambiguous or overlapping categories. While our findings highlight significant gains in classification metrics such as macro-F1 and weighted-F1, improvements are likely possible with key future directions include model interpretability enhancements, domain-specific fine-tuning for nuanced medical conditions, and cross-lingual adaptations that can scale to diverse user populations. Furthermore, integrating external medical knowledge bases or discourse-level context could refine the Refiner's decision boundaries, especially for borderline snippets that require deeper inference. By unifying advanced prompting techniques with context-driven refinement, the proposed CRA framework can be extended to broader, multi-turn QA and summarization tasks in healthcare, ultimately improving the reliability of automated systems designed to navigate the ever-evolving landscape of health information exchange.

## Limitations

Although our CRA significantly improves classification accuracy for user-generated health content, there are notable limitations that warrant attention. First, the approach relies heavily on the availability of high-quality labeled data in the training set. If the training set lacks examples that closely resemble an ambiguous snippet, the Refiner may fail to retrieve contextually relevant instances, leading to suboptimal classification. Second, while the inclusion of CoT prompting and tone definitions enhances interpretability, it does not fully guarantee factual correctness, particularly critical in healthcare scenarios. Our system is not designed to validate medical claims or detect misinformation, so erroneous or potentially harmful suggestions could persist if they align with patterns seen in the training data. Additionally, the current pipeline has been tested on a single domain-specific dataset and language, limiting its generalizability to other languages or more specialized medical domains. Future research could explore cross-lingual implementations or adapt the method to incorporate external medical knowledge bases for deeper validation. Finally, despite demonstrating improvements in computational efficiency by triggering the Refiner only when the Classifier is uncertain, the iterative nature of our approach incurs additional inference time for borderline cases, which might not be desirable for large-scale, real-time applications.

## 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.

Anthropic. 2024. Introducing the next generation of claude.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *CoRR*, abs/2012.15723.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *CoRR*, abs/2007.01282.

Jooyeon Lee, Luan Huy Pham, and Özlem Uzuner. 2024. Enhancing consumer health question reformulation: Chain-of-thought prompting integrating focus, type, and user knowledge level. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 220–228, Torino, Italia. ELRA and ICCL.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Mi-

nal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *CoRR*, abs/2002.10957.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Rong Zhang, Revanth Gangi Reddy, Md. Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. *CoRR*, abs/2010.05904.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *CoRR*, abs/2102.09690.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, Chao Cao, Hanqi Jiang, Hanxu Chen, Yiwei Li, Junhao Chen, Huawen Hu, Yihen Liu, Huaqin Zhao, Shaochen Xu, Haixing Dai, Lin Zhao, Ruidong Zhang, Wei Zhao, Zhenyuan Yang, Jingyuan Chen, Peilong Wang, Wei Ruan, Hui Wang, Huan Zhao, Jing Zhang, Yiming Ren, Shihuan Qin, Tong Chen, Jiaxi Li, Arif Hassan Zidan, Afrar Jahin, Minheng Chen, Sichen Xia, Jason Holmes, Yan Zhuang, Jiaqi Wang, Bochen Xu, Weiran Xia, Jichao Yu, Kaibo Tang, Yaxuan Yang, Bolun Sun, Tao Yang, Guoyu Lu, Xianqiao Wang, Lilong Chai, He Li, Jin Lu, Lichao Sun, Xin Zhang, Bao Ge, Xintao Hu, Lian Zhang, Hua Zhou, Lu Zhang, Shu Zhang, Ninghao Liu, Bei Jiang, Linglong Kong, Zhen Xiang, Yudan Ren, Jun Liu, Xi Jiang, Yu Bao, Wei Zhang, Xiang Li, Gang Li, Wei Liu, Dinggang Shen, Andrea Sikora, Xiaoming Zhai, Dajiang Zhu, and Tianming Liu. 2024. Evaluation of openai o1: Opportunities and challenges of agi. *Preprint*, arXiv:2409.18486.