Response Content Units: Evaluating Completeness and Proactiveness in Medical Open-Response Question Answering

Yongsin Park
University of Washington
yongsinp@uw.edu

Wen-wai Yim
Microsoft Health AI
yimwenwai@microsoft.com

Emma McKibbin
University of Washington
echmckibbin@gmail.com

Asma Ben Abacha
Microsoft Health AI
abenabacha@microsoft.com

Fei Xia
University of Washington
fxia@uw.edu

Abstract

Remote clinical care has significantly increased the workload for healthcare professionals managing digital inquiries. While automated systems aim to alleviate this burden, consumer health questions present unique challenges due to their linguistic complexity and the need for proactive clinical guidance, which traditional question-answering models often overlook. We introduce the medical Response Content Units (RCU) schema, a framework that facilitates automatic analysis to identify question-answer completeness and critical answer subparts, which can then be used as tools for supporting clinician response or for automatic metric evaluation. Our analysis using this schema reveals a 16.4% gap in response completeness in professional replies and demonstrates that essential medical directives are provided 2.4 to 12.1 times as frequently as direct answers. We provide baseline results and publicly release our annotations and source code to offer an evaluation framework that is more closely aligned with real-world clinical requirements.

1 Introduction

The expansion of remote clinical care has offered unprecedented convenience for patients and providers alike. However, it has also resulted in a significant increase in communicative workload for healthcare professionals, who must now manage a high volume of digital inquiries (Yim et al., 2024). Although automated systems attempt to reduce this workload by generating responses on the doctor's behalf, consumer health questions (CHQs) present unique challenges that distinguish them from standard question-answering tasks. These inquiries are linguistically complex, emotionally charged, and frequently contain multiple distinct questions within a single messy message (Ben Abacha et al., 2017), or lack explicit questions entirely. Furthermore, professional responses often include proactive clinical guidance that is not directly requested by the patient.

Current systems trained on clean, 1:1 question answering datasets may struggle with the context provided by patients, fail to address all issues when multiple concerns are raised (Ben Abacha and Demner-Fushman, 2019), or fail to provide an answer when no explicit question is found. Another critical hurdle is the discrepancy between what a patient asks and what they actually need to know. The best clinical response often includes essential information outside the immediate focus of the patient's question, a requirement that current models often fail to meet, as they are typically trained to generate answers strictly conditioned on explicit questions and evaluated with metrics that penalize additional information. Given the high-stakes nature of the medical domain, it is imperative that professionals manually verify every automated output, which increases cognitive load and effectively undermines the system's intended usefulness.

To bridge the gap between current AI capabilities and clinical requirements, we propose the Response Content Units (RCU) schema, which captures the complexity of real-world medical consultations by modeling open-ended interactions between patients and professionals, as well as provider-to-provider inquiries, and structuring them into units suitable for automated analysis. The schema decomposes the interactions into fine-grained, semantically linked, multi-part question-answer pairs with domain-motivated labels (as shown in Figure 1), enabling assistive tools that flag unaddressed concerns or missing clinical components while preserving professional judgment. It may also support more clinically aligned evaluation beyond traditional reference-based text generation scores, and can help patients navigate complex, jargon-heavy responses by highlighting diagnoses, action items, and prognoses.

In this paper, we present baseline results for ex-

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 231–252

ORIGINAL SOURCE MATERIAL		
Patient Query: Regarding wound inquiry; Regarding a wound inquiry: my heel got scraped by a cart, resulting in a cut about 3 cm long, as shown in the picture. At First, I thought it was just a scrape, but now the wound has sunken in. What's going on? Does this need stitches? Will it heal flat afterwards?		
Expert/System Response: No stitches are necessary. The wound will take 4-6 weeks to heal and will flatten over time.		
RCU SCHEMA ANNOTATIONS		
Inquiry Mapping	**Decomposed Question-Answer Pairs**	**Attributes**
	Q1: What's going on?	**Polarity**: open; **Type**: identification
	→ *Unanswered*	
	Q2: Does this need stitches?	**Polarity**: binary; **Type**: advice
	→ No stitches are necessary.	**Value**: no
	Q3: Will it heal flat?	**Polarity**: binary; **Type**: outcome
	→ ...will flatten over time.	**Value**: yes
Clinical Content	**Medical Directives and Prognosis**	
	Directive: The wound will take 4-6 weeks to heal...	**Category**: Problem
		is_severe: False; **is_conditional**: False
	Prognosis: ...will flatten over time.	

Figure 1: The RCU Annotation Framework applied to a complex clinical interaction. The top section contains the original source material. The lower section highlights our added annotations, identifying and linking multiple questions and answers, alongside a clinical directive and a future expectation.

traction and classification tasks using this multi-layered annotation scheme. To facilitate future research, we publicly release our annotations and guidelines[1] and our source code[2]. Our contributions include:

- **A novel schema for complex clinical QA**: We introduce the medical Response Content Units (RCU) schema, a framework designed to capture the complexities of patient inquiries, including multi-part queries, implicit concerns, and clinical guidance that standard benchmarks often overlook.
- **Analysis of real-world clinical data**: We provide a comprehensive analysis of our annotated dataset, revealing that medical professionals left 16.4% of unique explicit questions unanswered, highlighting a critical gap in clinical completeness . The analysis also demonstrates that essential medical directives are provided between 2.4 and 12.1 times as frequently as direct answers.
- **Baseline systems**: We release the source code and establish baseline results for associated extraction and classification tasks using modern Large Language Models (LLMs).
- **Granular, multi-layered annotations**: We contribute a sentence-level mapping of professional responses to specific question IDs to verify answer completeness. The annotations also capture medical directives and prognoses, essential components of medical care that are often penalized by reference-based text generation metrics, helping assess whether responses align with clinical requirements.

2 Related Work

The multifaceted nature of consumer health questions (CHQs) poses challenges for automated systems, which are often developed using simplified benchmarks for the sake of evaluative convenience. Existing question-answering (QA) benchmarks frequently rely on clean, 1:1 question-answer pairs with constrained output formats that fail to capture the intricacies found in real-world patient inquiries, which are often messy, context-heavy, and may include multiple or no explicit questions, while responses frequently provide clinically relevant guidance beyond what was explicitly asked.

2.1 Limitations of Existing Benchmarks

While open-domain "reading comprehension" datasets like SQuAD (Rajpurkar et al., 2016) provide a strong foundation for question answering,

[1] https://osf.io/kcv2n/files/osfstorage
[2] https://github.com/yongsinp/RCU-MORQA

Dataset	Domain	Instances	Avg Q / Inst.	Sem. Equiv.	Aligned	Answer Format
SQuAD 2.0	General	151,054	1.0	No	No	Factoid Span
PubMedQA (PQA-L)	Biomedical	1,000	1.0	No	No	Binary/Long
MedQuAD	Consumer Health	47,457	1.0	No	No	Retrieval-based
LiveQA 2017	Consumer Health	738	1.42	No	No	Retrieval-based
***DermaVQA-RCU**	**Consumer Derm.**	**306**	**2.18**	**Yes**	**Yes**	**Open-response**
***WoundcareVQA-RCU**	**Consumer Wound**	**476**	**1.82**	**Yes**	**Yes**	**Open-response**

(*This Work.) Note: "Instances" refers to the number of posts/inquiries. "Avg Q / Inst." denotes the number of identified questions per post/inquiry. "Sem. Equiv." denotes linking of semantically equivalent questions within a single query. "Aligned" indicates sentence-level mapping between the query and response.

Table 1: Comparison of our augmented datasets with existing Medical and General QA benchmarks.

their questions are artificially generated from a given document, making them inherently contingent on the text (Kwiatkowski et al., 2019) and lacking the spontaneous, context-heavy, and emotional characteristics of clinical consultations. Query-based datasets, derived from search engine data, are often keyword-driven and lack the full-sentence structure and emotional context of CHQs (Nguyen et al., 2016). Benchmarks like PubMedQA (Jin et al., 2019) and MedQuAD (Ben Abacha and Demner-Fushman, 2019) utilize medical content, but their schemas are optimized for high-level classification or retrieval rather than the granular extraction of clinical intent.

2.2 Comparison of Categorization Schemas

Another challenge in current QA research is the limitations of existing categorization schemas. Traditional schemas, such as the yes/no/maybe labels in PubMedQA (Jin et al., 2019), the question types in MedQuAD (Ben Abacha and Demner-Fushman, 2019), or the topic types in TREC 2017 LiveQA (Ben Abacha et al., 2017), are well suited for classification or retrieval but fail to capture the multi-part clinical functions of professional responses. Crucially, these frameworks cannot distinguish between a clinician's current assessment, medical guidance, and prognosis within a response. Our RCU schema addresses this by aligning specific response sentences to corresponding clinical intents.

Furthermore, LiveQA (Ben Abacha et al., 2017) includes responses from information specialists that favor general information over direct medical opinions, and utilizes human evaluation to assess correctness and completeness for the entire text. Because it depends on manual scoring, it lacks automated, fine-grained evaluation of how well a response addresses each clinical need within a query. Our work bridges this gap by utilizing verified responses from licensed professionals that include medical opinions, and by aligning response sentences to corresponding question IDs, enabling verification of answer completeness absent in current medical question answering datasets.

2.3 Our Approach

By annotating naturally occurring "messy" clinical inquiries at the sentence and sub-sentence level, we move beyond simplified evaluative formats (Table 1). Our RCU schema identifies semantically linked multi-part question-answer pairs and domain-motivated labels like Medical Directives and Prognosis, which enable the development of assistive tools that alert clinicians to unaddressed concerns while preserving professional judgment.

3 Source Datasets and the RCU Schema

We build upon and expand the DermaVQA (IIYI subset) (Yim et al., 2024) and WoundcareVQA (Yim et al., 2025) datasets to address the complexities of open-response clinical interactions.

3.1 Source Datasets

The DermaVQA and WoundcareVQA datasets consist of open-response clinical interactions between patients and providers, as well as provider-to-provider inquiries. While both source datasets are originally designed for visual question answering, our work focuses on the text queries and responses. Both datasets feature human expert responses (Gold) and Large Language Model (LLM) generated responses (System).

DermaVQA (IIYI) This dataset consists of threads from IIYI.com, translated from Chinese into English, featuring patient and professional queries with responses only from validated medical professionals.

WoundcareVQA This corpus focuses on clinical inquiries scraped from Baidu Tieba [3] and Baidu Zhidao [4], which were translated from Chinese to English. The responses were authored by U.S. medical doctors to ensure clinical accuracy.

For further details on these benchmarks, please refer to (Yim et al., 2024) and (Yim et al., 2025).

3.2 RCU Annotation Schema

We extend the two source datasets by adding manual annotations using the medical Response Content Units (RCU) schema, resulting in two newly augmented datasets which we designate as **DermaVQA-RCU** and **WoundcareVQA-RCU**. The schema captures the complexities of real-world medical consultations by decomposing professional responses into discrete "units" (RCUs) at the sentence level. These units are labeled as either direct correspondences, which answer specific questions, or proactive correspondences, which provide clinical directives or prognoses not explicitly requested. Each RCU represents a functional component of the response and is categorized into three primary layers: semantically linked QA pairs, clinical directives, and prognoses. A comprehensive breakdown of all schema labels, attributes, and their permissible values is detailed in Appendix B. This schema enables the development of assistive tools that help professionals identify missing clinical information and help patients navigate complex, jargon-heavy responses.

Multi-part QA Pairs To assess answer completeness, patient queries are decomposed into distinct questions at the sub-sentence level, each assigned a unique ID. Professional response RCUs that address these questions are then mapped to the corresponding IDs, enabling verification of whether every concern has been addressed. For example, in Figure 1, the RCU "No stitches are necessary." is mapped directly to the question "Does this need stitches?", while "What's going on?" is identified as unanswered.

Clinical Directives These identify sentences describing current clinical identification, assessment, or advice, such as problems, required tests, treatments, or follow-up directives. This category describes what is currently happening or current possible problems, often providing essential information outside the immediate focus of the patient's explicit questions. This includes conditional follow-up directives such as: "If antibiotics do not improve healing, it may require debridement".

Prognosis We separately identify sentences that pertain to future outcomes, such as recovery timelines or permanent scarring. These are distinguished from directives because they describe events that have yet to occur. For example, a prognosis RCU identifies future outcomes such as: "Scar remodeling takes approximately one year and will soften and fade over time".

While a sentence is typically categorized as one or the other, these categories are not mutually exclusive; a single sentence can be labeled as both a Medical Directive and a Prognosis if it describes a current condition while predicting a future outcome. Furthermore, these clinical functions often overlap with question answering, as a sentence identified as a Medical Directive or Prognosis may also serve as a direct answer to one or more questions in the query. By labeling these fine-grained attributes, the schema enables an evaluation framework that rewards helpful, proactive clinical guidance rather than penalizing "extra" information.

3.3 Annotation

We used brat (Stenetorp et al., 2012) for annotation, and an example of the interface is shown in Appendix C. The initial guideline creation was performed by two biomedical NLP scientists, and two linguistics students participated in the creation and iteration of the annotation guidelines. Each time the guideline was revised or clarified, a subset of the data was re-annotated and new files were added. The final inter-annotator agreement was calculated using the F1 metric with relaxed match [5] over 51 files. At the entity label level, agreement reached F1 scores of 0.951 for patient questions, 0.918 for answers, 0.964 for medical directives, and 0.957 for prognosis extraction. More detailed agreement scores can be found in the Appendix D. Additionally, four medical annotators, with medical scribe experience, were onboarded and trained to reach these agreement levels. The rest of the corpus was single annotated.

4 Data Statistics and Composition

Table 2 shows that real-world clinical consultations often contain multiple distinct questions. On av-

[3] https://tieba.baidu.com

[4] https://zhidao.baidu.com

[5] Relaxed match considers partial span overlap as correct.

erage, medical queries in this corpus contain 1.96 questions, with some having as many as eight. We analyze these interactions across five key dimensions: question density, implicitness, completeness, proactive guidance, and linguistic structure.

Question Density DermaVQA-RCU inquiries contain an average of 2.18 questions, while WoundcareVQA-RCU inquiries average 1.82, highlighting the multi-part nature of clinical queries. This multiplicity poses a significant hurdle for systems trained on conventional 1:1 question-answering datasets, which are often ill-equipped to identify and address all concerns when multiple distinct issues are raised in a single inquiry.

Implicit Concerns A significant portion of the inquiries lack explicit questions, even though their objective is to solicit some form of diagnosis or medical advice. For instance, 28.4% of posts in the DermaVQA-RCU dataset contain no explicit questions, posing a challenge for traditional question-answering models that require an explicit query to either fail, or rely on a separate process to synthesize a natural language question from the context.

Clinical Completeness Gap While other datasets include multiple responses per query, reducing the likelihood of a question being left unanswered, the training split of WoundcareVQA-RCU corpus contains only a single response per query, providing a unique baseline for measuring answer completeness. In this subset, 16.4% of unique explicit questions were left unanswered by the medical professional. This discrepancy in clinical completeness has significant health implications, as unaddressed concerns may lead to delayed treatment, patient dissatisfaction, and an increased communicative load for healthcare providers as patients follow up for clarification.

Proactive Guidance Professional responses are rich in information not explicitly requested by patients. These essential medical directives, summarized in Table 4, are provided between 2.4 and 12.1 times as frequently as direct answers.

Linguistic Structure The dataset is overwhelmingly open-ended as shown in Table 3. Specifically, 86.8% of DermaVQA-RCU questions and 54.1% of WoundcareVQA-RCU questions require unconstrained, free-text responses, reflecting the diagnostic nature of medical queries where patients provide narrative histories and expect detailed identifications. Although the WoundcareVQA-RCU dataset has a higher percentage of yes/no questions, only 8.7% of binary questions received an explicit "yes" or "no" response. Combined with the rarity of categorical choices across both domains, these findings suggest that consumer health queries are inherently open-ended, and that simplified benchmarks with constrained outputs do not adequately reflect the domain.

5 Building the Baseline System

The following section outlines the overall components of our baseline system, which is designed to decompose complex clinical interactions into fine-grained clinical attributes. The pipeline comprises extraction and classification tasks: identifying patient questions at the sub-sentence level, classifying them by polarity and type, extracting corresponding answers from professional responses, and identifying and classifying clinical directives and prognoses. These tasks support the development of assistive systems that help healthcare providers in composing accurate and complete responses and enable a granular evaluation that aligns more closely with clinical requirements than traditional reference-based text-generation metrics.

5.1 Question Extraction and Classification

To manage the "messy" nature of real-world clinical inquiries, our system first identifies and extracts questions from both the title and body of a query at the sub-sentence level. Each extracted question is assigned a unique ID based on its sequential occurrence, but semantically equivalent questions are assigned identical IDs to track which questions have been answered. When no explicit questions are identified, the inquiry is assumed to be implicit. If multiple questions exist in a single sentence through separable independent clauses, they are labeled separately.

Once extracted, each explicit question is further decomposed into Polarity and Question Type. Polarity categorizes questions as binary (yes/no), categorical (choice-based), or open (open-ended). Simultaneously, the system assigns a question type based on the expected clinical response, such as identification, assessment, advice, or outcome prediction. This granular classification helps the system track whether a patient's specific concerns have been fully addressed and prepares the query for the subsequent answer identification phase.

5.2 Answer Extraction and Classification

To facilitate a granular analysis of clinical communication, our system identifies and extracts professional response components at the sentence level. Answers are linked to specific patient questions using unique question IDs, a mapping that allows the system to verify whether every part of a query has received a direct response. For questions identified with binary polarity, the system further normalizes the response by assigning a logical value of yes or no if the sentence is functionally equivalent to those terms. This classification supports models that help patients understand direct answers or prompt clinicians to provide more explicit replies when the original text is ambiguous.

5.3 Medical Directive and Prognosis Extraction

Beyond direct question-answering, the framework extracts and classifies broader clinical content. Professional responses are annotated for Medical Directives at the sentence level to describe current clinical problems or required actions. These directives are categorized into specific labels (e.g., problem, test, treatment, or followup) and flagged with safety and contextual attributes, `is_severe` for urgent issues or `is_conditional` for followup advice dependent on specific patient conditions. This process provides a basis for evaluating the helpfulness of a response and building systems that help recipients locate diagnoses and specific action items.

Finally, the system identifies Prognoses, which are sentences specifically related to potential future outcomes like healing timelines or permanent effects. While prognoses are distinguished from current clinical directives, the schema allows a single sentence to be labeled as both if it serves both functions, ensuring that proactive clinical guidance is captured accurately.

5.4 Evaluation

We evaluate extraction tasks using relaxed match and report performance using Precision, Recall, and F1-score. Classification tasks are evaluated using Weighted Average [6] metrics to account for class distribution. We do not provide end-to-end pipeline results for the aforementioned systems, as these experiments were primarily designed to establish baseline performance for a simple system and to identify the expected performance ceiling for each task in isolation.

[6]Weighted average precision, recall, and F1 are the sum of each class's score weighted by the proportion of that class's instances.

6 Baseline Results

We establish baseline performance using a combination of a rule-based system, a Machine Reading Comprehension (MRC) system, and modern Large Language Models (LLMs), including Gemini 2.5 Pro (Comanici et al., 2025), GPT-4o (OpenAI et al., 2024), and Qwen3-VL-Plus (Bai et al., 2025).

6.1 Question Extraction and Classification

Question Extraction We compared a rule-based system (using question marks and question starters) against LLM-based extraction using the prompt in Appendix E.1. As shown in Table 5, the rule-based approach scored an F1 of 0.736 on the DermaVQA-RCU dataset and 0.909 on WoundcareVQA-RCU, matching the performance of the LLM extractors. GPT-4o achieved the highest F1-score of 0.798 on DermaVQA-RCU, while Gemini 2.5 Pro reached 0.930 on WoundcareVQA-RCU.

Question Classification Using the prompt in Appendix E.2, LLMs were used to classify extracted components into the Polarity (binary, categorical, or open) and Question Type (identification, assessment, advice, or outcome_prediction) attributes defined in our RCU schema. Across both datasets, all models demonstrated high performance in the tasks, with Weighted Average F1-scores generally exceeding 0.90, as detailed in Tables 9 and 10.

6.2 Answer Extraction and Classification

Answer Extraction We compared LLMs (Appendix E.3) against a BioBERT-based MRC system (Lee et al., 2019). The BioBERT-MRC model was trained on the combined DermaVQA-RCU and WoundcareVQA-RCU training sets to identify the answer span given a question. Since MRC models require explicit queries, we used template questions based on question type (e.g., "Can you identify the problem?"). In Table 6, Gemini 2.5 Pro led DermaVQA-RCU (0.864 F1) and Qwen3-VL-Plus led WoundcareVQA-RCU (0.883). Despite its smaller size, BioBERT-MRC proved highly competitive, surpassing or performing on par with some foundation models.

Binary Answer Classification For binary questions, LLMs were prompted (Appendix E.4) to infer "yes" or "no" values. As shown in Table 11,

Split	Posts	Impl. Posts (%)	Total Qs (Unique)	Avg / Max Qs	Expl. Qs	Rs (G / S)	Unans. Expl. (G / S)
DermaVQA-RCU Dataset							
Train	150	36 (24.0%)	322 (288)	2.15 / 5	214	1,082 / –	23 / –
Valid	56	17 (30.4%)	124 (111)	2.21 / 4	73	529 / 168	2 / 2
Test	100	34 (34.0%)	220 (202)	2.20 / 4	118	1,126 / 300	3 / 4
Total	**306**	**87 (28.4%)**	**666 (601)**	**2.18 / 5**	**405**	**2,737 / 468**	**28 / 6**
WoundcareVQA-RCU Dataset							
Train	278	29 (10.4%)	518 (447)	1.86 / 7	431	278 / –	59 / –
Valid	105	6 (5.7%)	175 (155)	1.67 / 4	157	210 / 315	11 / 4
Test	93	2 (2.2%)	174 (152)	1.87 / 8	168	279 / 279	7 / 10
Total	**476**	**37 (7.8%)**	**867 (754)**	**1.82 / 8**	**756**	**767 / 594**	**77 / 14**

Note: G = gold standard (human-written) responses; S = system (LLM-generated) responses. Posts without any explicit (Expl.) questions are considered implicit (Impl.) . Average (Avg) and maximum (Max) questions (Q) are based on total number of questions per post. 3 LLMs were used for system response (R) generation. Train splits do not have system responses.

Table 2: Data statistics for the DermaVQA-RCU and WoundcareVQA-RCU datasets. The table presents the number of implicit queries, question density, and a comparison of unanswered (Unans.) explicit questions between human and LLM responses.

Split	Binary (Yes / No)	Categorical	Open	Total Questions
DermaVQA-RCU Dataset				
Train	38 (11.8%)	7 (2.2%)	277 (86.0%)	322
Valid	11 (8.9%)	3 (2.4%)	110 (88.7%)	124
Test	20 (9.1%)	9 (4.1%)	191 (86.8%)	220
Total	**69 (10.4%)**	**19 (2.9%)**	**578 (86.8%)**	**666**
WoundcareVQA-RCU Dataset				
Train	218 (42.1%)	12 (2.3%)	288 (55.6%)	518
Valid	74 (42.3%)	6 (3.4%)	95 (54.3%)	175
Test	82 (47.1%)	6 (3.4%)	86 (49.4%)	174
Total	**374 (43.1%)**	**24 (2.8%)**	**469 (54.1%))**	**867**

Note: All implicit questions are considered to have an Open Polarity.

Table 3: Distribution of Question Polarity across the DermaVQA-RCU and WoundcareVQA-RCU datasets. Polarity labels indicate whether a question expects a binary (yes/no), categorical (choice-based), or open-ended response.

GPT-4o achieved the highest F1-score for both DermaVQA-RCU (0.823) and WoundcareVQA-RCU (0.680).

6.3 Medical Directive and Prognosis

Medical Directive Extraction and Classification Sentences were extracted using the prompt in Appendix E.5 and categorized (e.g. problem, test, treatment, and followup) using the prompt in Appendix E.6. Table 7 shows Gemini 2.5 Pro performed best on both DermaVQA-RCU (0.908) and WoundcareVQA-RCU (0.839) datasets. Models also excelled at classifying attributes like `is_severe` and `is_conditional`, especially for the DermaVQA-RCU dataset with F1-scores exceeding 0.99 as seen in Table 13.

Prognosis Extraction Using the Appendix E.7 prompt, this task distinguished future outcomes (e.g., healing timelines) from current clinical status. Table 12 shows F1-scores ranging from 0.113 to 0.651, illustrating the difficulty of separating future expectations from present clinical directives.

7 Discussion

The statistical and experimental results across the DermaVQA-RCU and WoundcareVQA-RCU datasets highlight the challenges inherent in moving from standard question-answering to clinical consultations. A primary hurdle for automated systems is the prevalence of complex, multi-part queries and implicit questions. As shown in Table 2, many patient inquiries lack explicit questions entirely or contain multiple concerns. Traditional models, trained on clean 1:1 question-answer datasets, are fundamentally ill-equipped to handle this setting.

Split	DermaVQA-RCU		WoundcareVQA-RCU	
	Directive (G / S)	Prognosis (G / S)	Directive (G / S)	Prognosis (G / S)
Train	1,259 / –	18 / –	469 / –	81 / –
Valid	870 / 345	12 / 0	342 / 857	59 / 59
Test	1,640 / 581	11 / 0	573 / 778	79 / 50
Total	**3,769 / 926**	**41 / 0**	**1,384 / 1,635**	**219 / 109**

Note: G = gold standard (human-written) responses; S = system (LLM-generated) responses.

Table 4: Distribution of Medical Directive and Prognosis annotations. Medical Directive identifies current clinical actions, while Prognosis identifies future outcomes. System responses exist only for validation and test splits.

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.708	0.782	0.743
	GPT-4o	0.779	0.818	0.798
	Qwen3-VL-Plus	0.664	0.700	0.681
	Rule-based	0.765	0.709	0.736
WoundcareVQA-RCU	Gemini 2.5 Pro	0.878	0.989	0.930
	GPT-4o	0.838	0.954	0.893
	Qwen3-VL-Plus	0.846	0.977	0.907
	Rule-based	0.903	0.914	0.909

Table 5: Baseline performance for the **Question Extraction** task using relaxed match.

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.863	0.864	0.864
	GPT-4o	0.889	0.712	0.791
	Qwen3-VL-Plus	0.828	0.887	0.857
	BioBERT-MRC	0.876	0.830	0.853
WoundcareVQA-RCU	Gemini 2.5 Pro	0.874	0.769	0.818
	GPT-4o	0.820	0.694	0.752
	Qwen3-VL-Plus	0.887	0.880	0.883
	BioBERT-MRC	0.755	0.749	0.752

Combined results for gold and systems data for each dataset.

Table 6: Baseline performance for **Answer Extraction** task using relaxed match.

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.941	0.878	0.908
	GPT-4o	0.908	0.713	0.798
	Qwen3-VL-Plus	0.911	0.888	0.899
WoundcareVQA-RCU	Gemini 2.5 Pro	0.885	0.797	0.839
	GPT-4o	0.775	0.741	0.758
	Qwen3-VL-Plus	0.814	0.812	0.813

Combined results for gold and systems data for each dataset.

Table 7: Baseline performance for **Medical Directive Extraction** using relaxed match.

7.1 Analysis of the Clinical Completeness Gap

Our findings in Section 4 reveal a critical gap in response completeness, with medical professionals leaving 16.4% of unique explicit questions unanswered. In high-stakes medical settings, unaddressed concerns can delay treatment and increase communicative load through patient follow-ups. While it is difficult to retrospectively speculate on the specific reasoning of clinicians, a qualitative analysis of this gap suggests that some of the unanswered questions may stem from deliberate clinical triage rather than oversight.

Professionals frequently prioritize the highest-stakes issue when consumer health inquiries bury multiple concerns within a single message. For instance, they may prioritize a critical safety inquiry regarding a tetanus shot or a visit to the ER while dropping secondary questions about topical disinfection. Furthermore, proactive clinical guidance sometimes renders a patient's question irrelevant. Clinicians may bypass vague diagnostic questions ("What is happening?") or requests for home remedies to issue referrals for in-person evaluation.

However, we also notice some legitimate questions regarding prognosis ("How long until I can move freely?"), identification ("Why is it turning black?"), and patient experience ("Will it cause pain?") are also left unanswered. This indicates that while triage is a primary factor, the high question density and disorganized nature of these inquiries may still result in clinicians overlooking concerns that remain central to the patient's stated information needs.

7.2 Proactive Guidance and Metric Penalties

There is often a stark discrepancy between what a patient explicitly asks and what they actually need to know. Comparing the density of question-linked answers (Table 2) to broader medical directives (Table 4) demonstrates that professional responses contain a wealth of essential information that falls outside the immediate focus of the patient's inquiry. In fact, these medical directives appear 2.4 to 12.1 times as frequently as direct answers. As shown in Table 8, the patient explicitly asks only for a diagnosis ("What kind of rash is this?"). Despite this

Patient Query: What kind of rash is this?; It happens recently for a few days. It is itchy, and torn up after scratching. It mainly concentrate in the limbs. I have cat in the house and I am not sure if the problem was spread from the cat. It has been raining heavily these days and I am not able to get into the city because of transport problems. I would like to have experts here to take a look on it. Thank you very much.

Gold Response #1
Based on the description provided, I suggest several possible conditions, such as allergic dermatitis, parasitic infestations (e.g., flea bites or scabies), or fungal infections (photos are not clear enough). Avoid Scratching and use a calamine lotion to soothe the skin. Wash the affected areas with lukewarm water and a mild soap or soap-free cleanser. Apply a fragrance-free, hypoallergenic moisturizer and use a mild hydrocortisone cream on itchy areas, and an oral antihistamine tablet once daily to reduce inflammation. Keep your cat away from sleeping areas or direct skin contact until the issue is resolved and try to keep it clean and treated for fleas or mites regularly. Wash bed linens and clothing in hot water and vacuum carpets and furniture thoroughly. Contact your dermatologist for prescription-strength treatments if over-the-counter options are insufficient.

Gold Response #2
You have pruritic erythematous rash and papules over extensor surface of forearm. It can be insect bite hypersensitivity, papular urticaria or eczema. Apply a cream having combination of fusidic acid and betamethasone valerate twice daily for 10 days, then make it once for 10 days. Apply moisturiser in meantime. Take tab allegra 180 mg on sos basis for itching.

Table 8: An example of a multi-part clinical interaction from the dataset, showcasing a messy patient query addressed by two distinct professional responses that provide both direct answers and proactive clinical directives.

limited request, the clinicians proactively provide multi-step treatment regimens, topical medication instructions, and behavioral interventions.

Traditional automated systems struggle to provide this information when it is not directly requested. Worse, reference-based text-generation metrics often penalize "extra" information, actively discouraging the safety-first, proactive behavior essential to medicine. These metrics struggle because proactive components are either missing from the reference answers or they vary significantly by clinician style. For instance, Gold Response 1 focuses heavily on hygiene and environmental intervention, whereas Gold Response 2 targets a concise prescription strategy. Because reference answers vary based on individual focus, standard metrics are ill-suited for evaluation. Ultimately, the decision of whether an inquiry requires a direct answer, or if unrequested guidance is necessary, must remain at the medical professional's discretion.

8 Conclusion

Medical question-answering is a high-stakes domain where inquiries are inherently complex, and defining a "good" or "bad" answer is difficult. Evaluation remains challenging because experts frequently hold differing opinions and standard metrics struggle to assess free-form clinical responses.

Our findings reveal that real-world clinical queries typically contain multiple concerns, and responses may fail to address them all, which may pose unique risks. Effective clinical communication requires proactive guidance, a critical element often missing from traditional datasets. To navigate these challenges, systems must assist medical professionals rather than replace them, ensuring clinicians retain final discretion in patient care.

One promising avenue for improvement is utilizing multiple professional responses. Our dataset's multi-layered annotation scheme allows us to analyze how experts prioritize different medical directives. By extracting RCUs, we can identify distinct pieces of content that can be compared separately, rather than together with entire chunks of text. Future iterations of assistive tools could synthesize these differing professional perspectives to provide a more robust, response for the patient.

In summary, our RCU schema provides a novel framework for analyzing real-world clinical consultations. Our analysis reveals critical gaps in clinical completeness while highlighting the high frequency of proactive clinical guidance. The high performance achieved across baseline tasks suggests that these classification models are sufficiently robust to be deployed in assistive systems, such as provider-patient portal applications. Such tools could identify unaddressed concerns in real-time, reducing communicative load for healthcare providers. However, while the technical viability is demonstrated, seamless integration and usability for medical professionals will require further study and experimentation to ensure these tools effectively assist rather than hinder clinical judgment.

Limitations

This work has several limitations. First, different foundation models exhibit different behaviors on the same prompt, and each could have performed better under another prompt. Our prompts are limited in that they were designed for a simple system to establish baseline performance and were not optimized for each specific model. Second, while our Response Content Units schema is generalizable to other medical subdomains, our released annotations are limited to dermatology and wound care. Lastly, the dataset includes not only patient-to-professional inquiries but also provider-to-provider posts, mirroring real-world clinical settings. Although these can often be distinguished based on content, we do not explicitly annotate the professional status of the inquirer, nor do we analyze potential differences between the two groups.

Ethical Considerations

Our augmented dataset expands upon DermaVQA (Yim et al., 2024) and WoundcareVQA (Yim et al., 2025) benchmarks, all of which are publicly available. The annotation was done by in-house paid annotators with medical scribe experience (according to fair wages of the state) or linguistics students participating as part of class projects.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful feedback and suggestions.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *Text Retrieval Conference*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

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, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan

Yu, Fei Xia, M. Kennedy Hall, and Meliha Yetisgen. 2025. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024. DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland.

A More Baseline Results

In this section, we show baseline performances on question polarity classification (Table 9), question type classification (Table 10), binary answer classification (Table 11), medical directive classification (Table 13), and prognosis extraction (Table 12). The results demonstrate that while models achieve high scores on most classification tasks, they struggle with more nuanced tasks like prognosis extraction, highlighting the difficulty of distinguishing future medical outcomes from present clinical guidance.

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.942	0.923	0.928
	GPT-4o	0.956	0.940	0.944
	Qwen3-VL-Plus	0.959	0.957	0.958
WoundcareVQA-RCU	Gemini 2.5 Pro	0.983	0.982	0.982
	GPT-4o	0.938	0.939	0.938
	Qwen3-VL-Plus	0.973	0.969	0.970

Table 9: Baseline performance for the **Question Polarity Classification** task (weighted average metrics).

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.965	0.932	0.948
	GPT-4o	0.973	0.915	0.943
	Qwen3-VL-Plus	0.965	0.940	0.951
WoundcareVQA-RCU	Gemini 2.5 Pro	0.942	0.939	0.939
	GPT-4o	0.935	0.871	0.885
	Qwen3-VL-Plus	0.925	0.908	0.912

Table 10: Baseline performance for the **Question Type Classification** task (weighted average metrics).

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.817	0.819	0.807
	GPT-4o	0.826	0.831	0.823
	Qwen3-VL-Plus	0.788	0.789	0.764
WoundcareVQA-RCU	Gemini 2.5 Pro	0.736	0.686	0.671
	GPT-4o	0.732	0.692	0.680
	Qwen3-VL-Plus	0.719	0.646	0.627

Combined results for gold and systems data for each dataset.

Table 11: Baseline performance for the **Binary Answer Classification** task (weighted average metrics).

Dataset	Model	Prec.	Rec.	F1
DermaVQA-RCU	Gemini 2.5 Pro	0.076	0.909	0.141
	GPT-4o	0.078	0.727	0.142
	Qwen3-VL-Plus	0.061	0.727	0.113
WoundcareVQA-RCU	Gemini 2.5 Pro	0.329	0.805	0.467
	GPT-4o	0.514	0.891	0.651
	Qwen3-VL-Plus	0.434	0.844	0.573

Combined results for gold and systems data for each dataset.

Table 12: Baseline performance for the **Prognosis Extraction** task using relaxed match. Results illustrate the difficulty of distinguishing future outcomes from current clinical directives.

Dataset	Category	Gemini 2.5 Pro			GPT-4o			Qwen3-VL-Plus		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
DermaVQA-RCU	Problem	0.952	0.951	0.951	0.398	0.631	0.488	0.398	0.631	0.488
	Test	0.984	0.975	0.977	0.923	0.961	0.942	0.923	0.961	0.942
	Treatment	0.986	0.985	0.986	0.737	0.859	0.793	0.737	0.859	0.793
	Follow-up	0.947	0.949	0.945	0.782	0.885	0.830	0.782	0.885	0.830
	Severe	0.996	0.994	0.995	0.991	0.996	0.993	0.991	0.996	0.993
	Conditional	0.997	0.992	0.994	0.991	0.996	0.993	0.991	0.996	0.993
WoundcareVQA-RCU	Problem	0.851	0.853	0.852	0.476	0.690	0.563	0.476	0.690	0.563
	Test	0.994	0.965	0.977	0.988	0.994	0.991	0.988	0.994	0.991
	Treatment	0.965	0.964	0.965	0.479	0.692	0.566	0.479	0.692	0.566
	Follow-up	0.934	0.935	0.933	0.576	0.759	0.655	0.576	0.759	0.655
	Severe	0.969	0.944	0.954	0.928	0.963	0.945	0.928	0.963	0.945
	Conditional	0.954	0.956	0.955	0.776	0.881	0.825	0.776	0.881	0.825

Combined results for gold and systems data for each dataset.

Table 13: Baseline performance for **Medical Directive Classification** (weighted average metrics).

B The RCU Annotation Schema

This appendix defines the attributes and values for the Response Content Units (RCU) schema. The Question Level captures the informational needs (IDs, polarities, types) of the inquirer, while the Response Level maps answers to question IDs to verify completeness and extracts proactive Medical Directives and Prognoses.

Level	Label	Attribute	Value	Description
Question	`question`	`id`	1, 2, 3...	Sequential ID of a question found in a query. Semantically equivalent questions share the same ID.
		`is_implicit`	1	When no explicit question is found. IDs for implicit questions are assigned according to the alphabetical order of question types.
		`polarity`	binary, categorical, open	Expected answer format (yes/no, choices, or free-text).
		`type`	*multiple*	The expected clinical response (e.g., advice, assessment, identification, outcome_prediction).
Response	`shortest_answer`	`id`	0, 1, 2...	Maps the answer to the corresponding question ID.
		`val`	yes, no	Logical binary value (only for binary questions).
	`medical_iaa`	`problem`	1	Sentence relates to a diagnosis or current condition.
		`test`	1	Sentence relates to a required diagnostic test.
		`treatment`	1	Sentence relates to a required treatment.
		`followup`	1	Refers the patient to a specialist or asks them to follow up.
		`severity`	1	Problem requires immediate medical attention.
		`conditional`	1	Indicates that a recommended followup is conditional.
	`prognosis`	-	-	Relates to possible future outcomes that have yet to happen.

Table 14: Consumer Health QA Open Response Schema Definitions and Attributes.

C Annotation Interface

This appendix shows the brat annotation interface used for dataset labeling.

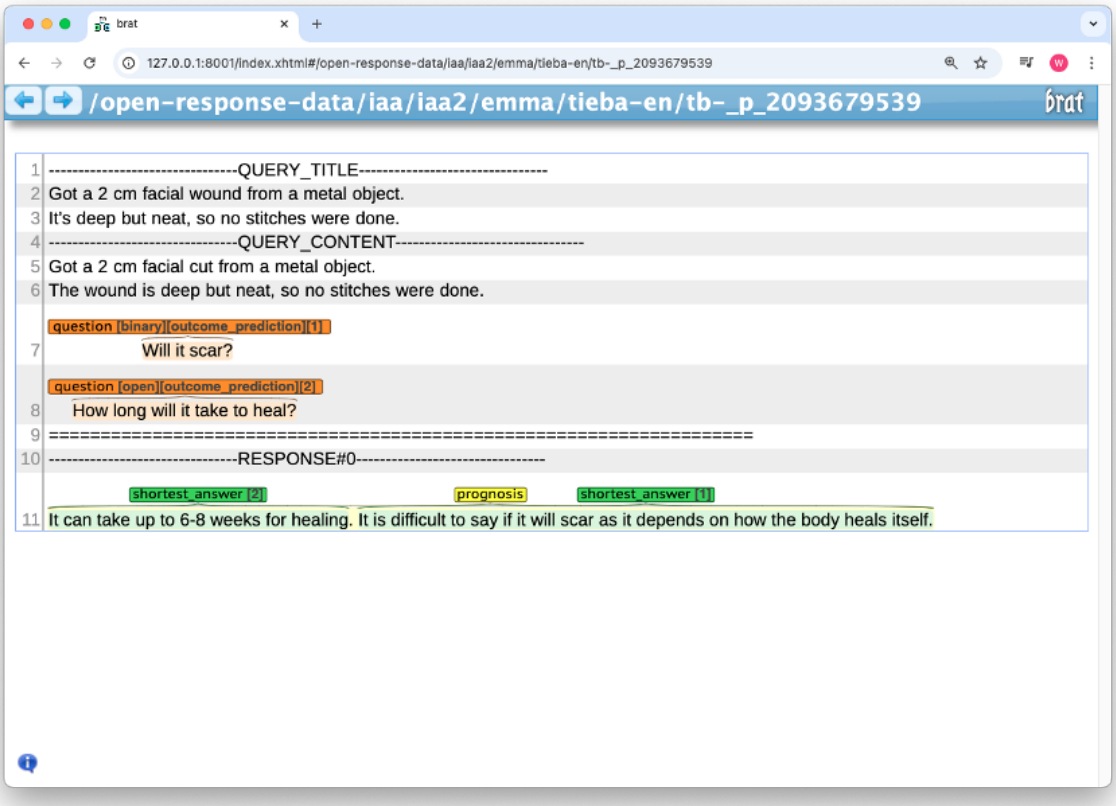

Figure 2: The brat annotation interface.

D Inter-annotator Agreement

This appendix provides a comprehensive breakdown of the inter-annotator agreement scores achieved during the validation of the Response Content Units (RCU) schema. The final agreement was calculated using the F1 metric with relaxed match across 51 files.

Annotation Category	**Average F_1**
Question ID Mapping (q-id)	0.8858
Question Type Mapping (q-type)	0.8248
Question Polarity Mapping (q-polarity)	0.9021
Answer ID Mapping (a-id)	0.6889
Answer Binary Value Mapping (a-yn)	0.9206
Medical Directive Attributes (iaa)	0.8579

Table 15: Aggregate inter-annotator agreement results.

Attributes	**F_1 Score**
Entity Labels	
question	0.951
shortest_answer	0.918
medical_iaa	0.964
prognosis	0.957
Question Attributes	
1 (Top ID)	0.922
2 (Top ID)	0.726
3 (Top ID)	0.667
advice	0.869
assessment	0.737
identification	0.982
outcome_prediction	0.870
binary	0.971
categorical	0.800
open	0.935
is_implicit	0.761
Answer Attributes	
1 (Top ID)	0.851
2 (Top ID)	0.733
3 (Top ID)	0.194
yes	0.900
no	0.941
Medical Directive Attributes	
is_followup	0.602
is_prob	0.960
is_test	0.884
is_treat	0.986
is_severe	0.842
is_conditional	0.750

Table 16: Inter-annotator agreement for the individual Response Content Units (RCU).

E System Prompts for Baseline Tasks

This section details the system prompts used for the Large Language Model (LLM) baselines described in Section 6.

E.1 Question Extraction Prompt

```
You are an advanced Medical Scribe. Your task is to identify and extract all text segments where the user is soliciting medical advice, diagnosis, opinion, or help.

GUIDELINES
1. Focus on Intent: Extract any text where the user is seeking an answer or a solution. This includes direct questions (e.g., Is this normal?) and implicit requests (e.g., Please help me identify this).
2. Verbatim Extraction: Extract the text exactly as it appears in the source, including the trailing punctuations, if any.
3. Context: Split compound sentences. If a user asks "What is this and how do I treat it?", extract them as multiple entries.

OUTPUT FORMAT
Return a JSON object just containing a list of strings. If no inquiries are found, return an empty list.

EXAMPLES
Example 1
Input: Urgent!!! Is this Dermatitis due to Blattella???
Output: ["Is this Dermatitis due to Blattella???"]

Example 2
Input: The patient is a 49-year-old female with papules on her face. She has a history of rosacea.
Output: []

Example 3
Input: Lower limb eczema (with picture), please provide diagnosis and prescription.
Output: ["please provide diagnosis", "prescription"]
```

E.2 Question Classification Prompt

```
You are an expert Medical Linguistic Analyzer. Your task is to classify a given medical question based on two specific dimensions: Polarity and Type.

DEFINITIONS
1. Polarity:
 - binary: Questions that can be logically answered with a simple "Yes" or "No" (e.g., "hyperkeratosis, can it appear on the lower limbs?").
 - categorical: Questions presenting a choice between specific options (e.g., "Oral mucosal disease, eczema? Herpes?", "Is this urticaria or a skin allergy?").
 - open: Questions requiring a descriptive response, explanation, or list (Who, What, Where, When, Why, How) (e.g., "what tests need to be done.").

2. Type:
 - identification: Asking to identify the wound's cause, current state, pathology, or developments (e.g., "please provide diagnosis", "May I ask what kind of skin disease is this?").
 - assessment: Asking to evaluate severity or urgency (e.g., "Is this a serious issue?", "Is there a problem with this wound?").
 - advice: Asking for actionable medical steps, including tests, treatments, or prescriptions (e.g., "Treatment for Chronic Urticaria", "Can someone give me a suggestion?").
 - outcome_prediction: Asking for predictions regarding recovery time or permanent effects (e.g., "How many days will it take to cure this disease approximately?", "Will this lead to tetanus?").

OUTPUT FORMAT
Return a single JSON object containing "polarity" and "type".

EXAMPLES
Example 1
Input: Please help to identify what this is on my hand.
Output: {"polarity": "open", "type": "identification"}

Example 2
Input: What kind of topical medication works best?
Output: {"polarity": "open", "type": "advice"}

Example 3
Input: Is it eczema or acute impetigo-like pityriasis versicolor?
Output: {"polarity": "categorical", "type": "identification"}

Example 4
Input: how long will it take to heal?
Output: {"polarity": "open", "type": "outcome_prediction"}

Example 5
Input: Will it heal without deforming?
Output: {"polarity": "binary", "type": "outcome_prediction"}

Example 6
Input: is the condition severe?
Output: {"polarity": "binary", "type": "assessment"}
```

E.3 Answer Extraction Prompt

```
You are a precise linguistic analysis engine specialized in medical context extraction. Your task is to extract only the minimum
```

```
    necessary sentences that directly answer
    the provided questions.

INPUT DATA
You will receive:
1. Questions: A list of strings (phrased
    differently but asking the same thing).
2. Polarity: (binary, categorical, or open).
3. Question Type: (identification, assessment,
     advice, outcome_prediction).
4. Responses: A list of strings to analyze.

DEFINITIONS
Use these definitions strictly to determine
    relevance:
1. Polarity:
 - binary: Questions logically answerable with
     "Yes" or "No" (e.g., "Is it X?"). A
    relevant answer might not say "Yes/No"
    explicitly but provides the confirmation
    or refutation (e.g., "It is Y" implies "No
    " to "Is it X?").
 - categorical: Questions presenting specific
    choices.
 - open: Questions requiring description,
    explanation, or lists.

2. Question Type:
 - identification: Identifying the wound/
    disease cause, state, pathology, or name.
 - assessment: Evaluating severity, urgency,
    or current status.
 - advice: Actionable steps, tests, treatments
    , or prescriptions.
 - outcome_prediction: Predictions on recovery
     or permanent effects.

PROCESSING LOGIC
For each item in the "Responses" list, perform
     the following steps:
Step 1: Determine Relevance Strategy
 - IF the "Questions" list contains valid
    strings: You are looking for sentences
    that specifically address the semantic
    intent of those questions.
 - IF the "Questions" list is empty ("") or
    contains only empty strings: You are
    looking for any sentences within the
    response that match the provided "Question
     Type".

Step 2: Sentence-Level Extraction
Split the response into individual sentences.
    Analyze each sentence to see if it
    qualifies as an answer.
 - Condition A (Type Match): The sentence
    content must semantically align with the
    provided "Question Type". (e.g., If
    Question Type is "advice", but the
    sentence is a diagnosis/identification
    like "It is Eczema", it is NOT a match).
 - Condition B (Answer Match): If questions
    are provided, the sentence must directly
    answer the inquiry.
 - Note: For Binary Identification questions (
    e.g., "Is it Tinea?"), a sentence
    identifying a different disease (e.g., "It
     is Psoriasis") IS a valid answer because
    it implicitly answers "No".

Step 3: Formatting
 - Extract the qualifying sentences verbatim.
    Do not paraphrase.
 - Combine multiple relevant sentences from a
    single response with their original
    punctuation.
 - If no sentences in a response meet the
    criteria, the result for that index is an
    empty string "".
 - Maintain a strict one-to-one mapping with
    the input "Responses" list.

EXAMPLES
Example 1
Input: Questions: ["Pygmy Moss?"], Polarity:
    binary, Type: identification, Response: ["
    Frictional Lichenoid Eruption", "
    Frictional lichenoid rash doesn't seem to
    be the case, but it appears to be some
    kind of lichenoid rash.", "Glossy Moss"]
Output: ["Frictional Lichenoid Eruption", "
    Frictional lichenoid rash doesn't seem to
    be the case, but it appears to be some
    kind of lichenoid rash.", "Glossy Moss"]

Example 2
Input: Questions: ["What skin disease?", "What
     is this emergency?"], Polarity: open,
    Type: identification, Response: ["Tension
    blister", "Papular urticaria", "For the
    itching diagnosis, a blister puncture will
     suffice.", "Papular urticaria, topical
    use of Lugen Shi lotion."]
Output: ["Tension blister", "Papular urticaria
    ", "For the itching diagnosis, a blister
    puncture will suffice.", "Papular
    urticaria, topical use of Lugen Shi lotion
    ."]

Example 3
Input: Questions: [""], Polarity: open, Type:
    identification, Response: ["Elbow Black
    Acanthosis Nigricans", "Frictional
    Hyperkeratosis"]
Output: ["Elbow Black Acanthosis Nigricans", "
    Frictional Hyperkeratosis"]

Example 4
Input: Questions: ["I suspect it's tinea,
    could an expert please confirm this?"],
    Polarity: binary, Type: identification,
    Response: ["The original poster is
    requested to provide more medical history.
     A sudden increase in blood count may
    suggest erysipelas, while a chronic
    process could indicate tuberculosis or
    even leprosy."]
Output: ["A sudden increase in blood count may
     suggest erysipelas, while a chronic
    process could indicate tuberculosis or
    even leprosy."]

Example 5
Input: Questions: ["What should we do?"],
    Polarity: binary, Type: advice, Response:
    ["Early Stage of Eczema in Children", "
    Eczema", "Eczema.............", "Eczema
    is easy to treat for some, but not for
    others."]
Output: ["", "", "", ""]
```

```
OUTPUT FORMAT
Return strictly a JSON list of strings.
    Maintain a strict 1:1 mapping with the
    Response list from the Input.
```

E.4 Binary Answer Classification Prompt

```
You are an expert Medical Linguistic Analyzer.
     Your task is to determine the logical
    binary value (val) of specific answers in
    response to a medical Yes/No question.

INPUT DATA
You will receive:
1. Questions: A list of semantically
    equivalent strings asking the same medical
     "Yes/No" question.
2. Answers: A list of strings, each being a
    distinct answer segment to be labeled.

DEFINITIONS AND LABELS
Assign one of the following labels to each
    answer:
"yes": The answer affirms the question. Use
    this if the answer can be logically
    replaced with the word Yes.
"no": The answer negates the question. Use
    this if the answer is equivalent to No.
"" (empty string): The answer is ambiguous,
    explicitly states it depends, or provides
    information without a clear affirmation or
     negation.

ANNOTATION LOGIC
The Replacement Test Ask: Can this sentence be
     replaced with Yes or No without changing
    the meaning?
IF the answer explicitly says "Yes" or affirms
     the premise ASSIGN "yes".
IF the answer explicitly says "No" or negates
    the premise ASSIGN "no".

The Doctor Intent Test Ask: Based on your
    judgment, does the doctor believe they are
     answering the question?
If the doctor provides an alternative
    diagnosis they are implicitly answering No
    . ASSIGN "no".
If the doctor provides reassurance that
    implies a negative they are answering No.
    ASSIGN "no".
If the doctor provides a recommendation that
    replaces the expected outcome they are
    answering No. ASSIGN "no".

The Conditional Rule If an answer depends on a
     condition (e.g., If X happens, then Y),
    analyze the recommendation:
Action-Oriented Intent: If the doctor
    recommends an action based on a likely
    condition (e.g., "If it is rusty, I
    recommend a shot"), interpret this as an
    affirmative recommendation. ASSIGN "yes".
True Ambiguity: If the doctor says "It is
    difficult to say" or "It depends on how
    the body heals," they are refusing to
    answer yes/no. ASSIGN "".

EXAMPLES
Example 1
Input: Questions: ["Is there such a thing as
    urticaria?"], Answers: ["Erythema annulare
     centrifugum?", "I think it still looks
    like urticaria, continue with the anti-
    allergy treatment", "Urticaria", "I think
    the likelihood of urticaria is the highest
    , but the skin lesions at the root of the
    thigh are hard to explain, so erythema
    annulare cannot be ruled out either."]
Output: ["no", "yes", "yes", "yes"]
(Reasoning: For the last answer, while the
    doctor believes other diagnosis is
    possible, they still believe urticaria is
    the most likely diagnosis so it's a yes.)

Example 2
Input: Questions: ["Will it leave a scar?"],
    Answers: ["The scar will continue to
    remodel for a year and soften/fade over
    time."]
Output: ["yes"]

Example 3
Input: Questions: ["Are there any ideal
    topical medication for hands like this"],
    Answers: ["Apply moisturizing ointment or
    cream as much as possible on the hands.
    You can try over the counter
    hydrocortisone 1% cream 2 times a day for
    1 week to help speed up the healing."]
Output: ["yes"]

Example 4
Input: Questions: ["Is it necessary to remove
    this mole?", "is it necessary to remove it
    ?"], Answers: ["Pigmented nevus -
    intradermal nevus, often occurs in
    childhood, benign, surgery is recommended
    for removal.", "It's better to have it cut
    .", "Simple excision for pathology is
    sufficient.", "It's not a good thing, cut
    it off."]
Output: ["yes", "yes", "yes", "yes"]

OUTPUT FORMAT Return strictly a JSON list of
    strings corresponding 1:1 with the input
    Answers. Valid values are only "yes", "no
    ", or "".
```

E.5 Medical Directive Extraction Prompt

```
You are a precise linguistic analysis engine
    specialized in medical context extraction.
     Your task is to identify and extract all
    'sentences' that qualify as Medical
    Identification, Assessment, or Advice (IAA
    ) from a response, distinguishing it from
    Prognosis.

INPUT DATA
You will receive:
1. Response: A list of strings (medical
    answers) to analyze.

DEFINITIONS
Use these definitions strictly to determine
    relevance. A sentence is IAA if it
```

```
    describes what is currently happening,
    current possible problems, or current test
    /treatments required. It must fall into at
     least one of these four categories:
1. Problem: Describes the diagnosis, current
    condition or problem.
2. Test: Describes a required or recommended
    diagnostic test, imaging, or lab work.
3. Treatment: Describes a required treatment.
4. Followup: Describes a referral to a
    particular department or specialist or
    asking the patient to follow up in a
    certain amount of time.

EXCLUSION CRITERIA
Prognosis: Exclude sentences that predict
    possible future outcomes. These are
    classified as Prognosis, not IAA. However,
     they are not mutually exclusive; a
    sentence can be both IAA and Prognosis if
    it describes the current condition while
    also predicting future outcomes.

PROCESSING LOGIC
For each item in the Response list, perform
    the following steps:
Step 1: Analyze the entire text of the
    response. A single response may contain
    multiple distinct sentences that qualify
    as IAA (e.g., a diagnosis at the beginning
     and a treatment recommendation at the end
    ), separated by non-relevant text.

Step 2: Analyze every sentence. Do not stop
    after finding the first relevant sentence.
     Keep the sentence only if it meets the
    IAA definition (Problem, Test, Treatment,
    or Followup).
Multi-Sentence Spans: A single piece of advice
     or assessment may span multiple
    consecutive sentences. Extract them as one
     unless they require different sets of
    labels (e.g. separate as different IAAs if
     first sentence satisfies Problem but the
    next sentence satisfies both Problem and
    Treatment). If they are non-consecutive (e
    .g. they have Prognosis or irrelevant
    sentences in between), extract them as
    separate IAAs even if they have an
    identical IAA attribute. Include all
    sentences that contribute to the
    actionable advice or current assessment.

Step 3: Extract the qualifying sentences
    verbatim. Group ALL extracted IAA
    sentences from a single response into a
    list (e.g. ["IAA Sentence 1", "IAA
    Sentence 2", ...]). If no sentences meet
    the criteria, the result is an empty list
    [].

EXAMPLES
Example 1
Input: ["After the treatment of contact
    dermatitis and scabies, many patients show
     changes in dermatitis. On one hand, it is
     related to scabies itself, on the other
    hand, the treatment drugs mainly based on
    sulfur have a significant impact on the
    skin. Therefore, during and after the
    treatment, attention should be paid to
    avoid further damage and protect the skin.
     However, it is necessary to first confirm
     whether the scabies has been cured. If
    the scabies has been cured, the main focus
     should be on anti-allergy treatment.", "
    Pay attention to hygiene and frequently
    air out your underwear and bedding.
    Ventilate the room and keep the
    environment clean. \n\nMaintain a light
    diet and avoid spicy and greasy foods. Eat
     more vegetables and fruits rich in
    vitamin C, and drink more milk."]
Output: [["After the treatment of contact
    dermatitis and scabies, many patients show
     changes in dermatitis.", "On one hand, it
     is related to scabies itself, on the
    other hand, the treatment drugs mainly
    based on sulfur have a significant impact
    on the skin.", "Therefore, during and
    after the treatment, attention should be
    paid to avoid further damage and protect
    the skin.", "However, it is necessary to
    first confirm whether the scabies has been
     cured.", "If the scabies has been cured,
    the main focus should be on anti-allergy
    treatment."], ["Pay attention to hygiene
    and frequently air out your underwear and
    bedding. Ventilate the room and keep the
    environment clean", "Maintain a light diet
     and avoid spicy and greasy foods. Eat
    more vegetables and fruits rich in vitamin
     C, and drink more milk."]]
(Reasoning:
In the first response, the extraction is
    segmented into multiple distinct strings
    because the classification attributes
    change from sentence to sentence. Per the
    processing logic, consecutive IAA
    sentences must be separated if they
    require different label sets.
Segment 1 & 2: While the first sentence ("
    After the treatment...") describes the
    Problem, the second sentence ("On one hand
    ...") discusses both the Problem and the
    impact of drugs (Treatment). This addition
     of the 'Treatment' attribute necessitates
     a new span.
Segment 3: "Therefore, during..." shifts focus
     strictly to preventative advice (
    Treatment), dropping the 'Problem' label.
Segment 4: "However, it is necessary..."
    introduces a requirement for confirmation,
     introducing Followup attribute in
    addition to 'Problem' (the mention of
    scabies).
Segment 5: "If the scabies..." returns to '
    Problem' and 'Treatment'.
Because the specific combination of labels (
    Problem, Test, Treatment, Followup) shifts
     at each sentence boundary, they are
    returned as individual strings rather than
     a single merged block.
In the second response, the separation is
    caused by the double newline (\n\n). This
    formatting is non-clinical whitespace and
    does not qualify as Medical Identification
    , Assessment, or Advice (IAA). Because you
     should extract only valid IAA content and
     strictly excludes non-relevant text, the
```

```
    \n\n is not captured. This creates a gap
    in the extraction, resulting in two
    distinct, non-contiguous spans of advice.)

Example 2
Input: ["Recommend evaluation in ER for Xray
    to evaluate for fracture. The nail will
    take up to 6 months to grow back."]
Output: [["Recommend evaluation in ER for Xray
     to evaluate for fracture.", "The nail
    will take up to 6 months to grow back."]]
(Reasoning: The first sentence qualifies as an
     IAA because it recommends evaluation (
    Test, Followup). While the second provides
     information on the expected recovery
    timeline for the nail (Prognosis), but
    also describes current condition (Problem)
    .)

Example 3
Input: ["First of all, you can't scratch it
    anymore. Apply some anti-inflammatory
    topical cream. Observe it for a few days
    .", "Eczema?", "It is estimated to be a
    disease related to capillary hemangioma.
    Continue to observe, and if it enlarges,
    surgical removal is recommended.", "Is it
    folliculitis?", "It is recommended to
    first use Band-Aid externally and pay
    attention to cleanliness! Observe for a
    few days and see. If there is no
    improvement, go to a regular hospital for
    a check-up.", "Considering it is a
    capillary hemangioma, laser treatment is
    recommended.", "Capillary hemangioma???",
    "Hemangioma...", "Considered to be a
    capillary hemangioma.", "The possibility
    of purulent granuloma is relatively high
    .", "The description is about capillary
    hemangioma. Try using ionization to burn
    it, liquid nitrogen is also acceptable.
    Don't squeeze it anymore, it's prone to
    infection.", "Pyogenic granuloma, laser
    treatment.", "The possibility of a skin
    hemangioma is still relatively high, laser
     or liquid nitrogen therapy should be
    considered.", "Capillary hemangioma, apply
     Mupirocin externally, observe!", "
    Hemangioma, is it possible?", "Don't rush
    to pick at it yet.", "After scratching the
     papular urticaria, closely follow up and
    revisit after a week.", "Consider multiple
     angiomas.", "Consider doing a color
    Doppler ultrasound.", "Can't pick with
    hands anymore.", "Considering hemangioma,
    it's very common in the chest area. Use
    laser after infection control.", "I am
    considering diseases related to angioma. I
     suggest that there is currently no need
    for medication and we should observe first
    . It could also be pigmentation.", "
    Artificial dermatitis, it will get better
    on its own in a few days."]
Output: [["First of all, you can't scratch it
    anymore. Apply some anti-inflammatory
    topical cream. Observe it for a few days
    ."], ["Eczema?"], ["It is estimated to be
    a disease related to capillary hemangioma
    .", "Continue to observe, and if it
    enlarges, surgical removal is recommended
    ."], ["Is it folliculitis?"], ["It is
    recommended to first use Band-Aid
    externally and pay attention to
    cleanliness! Observe for a few days and
    see.", "If there is no improvement, go to
    a regular hospital for a check-up"], ["
    Considering it is a capillary hemangioma,
    laser treatment is recommended"], ["
    Capillary hemangioma???"], ["Hemangioma
    ..."], ["Considered to be a capillary
    hemangioma."], ["The possibility of
    purulent granuloma is relatively high."],
    ["The description is about capillary
    hemangioma", "Try using ionization to burn
     it, liquid nitrogen is also acceptable.
    Don't squeeze it anymore, it's prone to
    infection."], ["Pyogenic granuloma, laser
    treatment."], ["The possibility of a skin
    hemangioma is still relatively high, laser
     or liquid nitrogen therapy should be
    considered."], ["Capillary hemangioma,
    apply Mupirocin externally, observe!"], ["
    Hemangioma, is it possible?"], ["Don't
    rush to pick at it yet."], ["After
    scratching the papular urticaria, closely
    follow up and revisit after a week."], ["
    Consider multiple angiomas."], ["Consider
    doing a color Doppler ultrasound."], ["Can
    't pick with hands anymore."], ["Use laser
     after infection control.", "Considering
    hemangioma, it's very common in the chest
    area."], ["It could also be pigmentation
    .", "I suggest that there is currently no
    need for medication and we should observe
    first.", "I am considering diseases
    related to angioma."], ["Artificial
    dermatitis, it will get better on its own
    in a few days."]]
(Reasoning: Sentences "Don't squeeze it
    anymore, it's prone to infection." and "
    Artificial dermatitis, it will get better
    on its own in a few days." are both
    prognoses because they describe possible
    outcomes. However, since the former is
    part of a broader treatment recommendation
     and the latter includes a diagnosis, they
     also qualify as IAA and are therefore
    included in the output.)

OUTPUT FORMAT
Return strictly a JSON list of lists of
    strings. Maintain a strict 1:1 mapping
    with the Response list from the Input.
```

E.6 Medical Directive Classification Prompt

```
You are an expert Medical Annotation
    Classifier. Your task is to analyze
    extracted "Medical Identification,
    Assessment, or Advice" (IAA) text and
    classify them according to specific
    clinical labels and attributes.

INPUT DATA
1. Context: The entire response text from
    which the IAA sentences were extracted.
2. IAA Texts: A list of strings (extracted IAA
     sentences) to classify.
```

```
DEFINITIONS AND LABELS
For each string, determine which of the
    following 4 labels apply. A single string
    should have at least one label.
problem: The text relates to a diagnosis,
    current condition, or identifying the
    problem.
test: The text relates to a required or
    recommended diagnostic test, imaging, or
    lab work.
treatment: The text relates to a required
    treatment, medication, or procedure.
followup: The text relates to referring the
    patient to a specific department or
    specialty, or asking them to follow up
    after a certain time.

ATTRIBUTES
For each string, determine the integer value
    (0 or 1) for the following two attributes:
1. is_severe (0 or 1): Set to 1 if the context
     implies the problem needs immediate
    medical attention. Mentions of "Emergency
    Room," "ER," or "Urgent Care" are strong
    clues for severity. If advice to see a
    specialist or follow up is given without
    urgency cues, set to 0. You must interpret
     the entire context instead of relying on
    keywords. If a mention of "Urgent Care" is
     qualified by a non-mandatory clue (e.g.,
    "I recommend Urgent Care if you are
    interested in getting an X-ray"), set to 0
     because the visit is optional.
2. is_conditional (0 or 1): Set to 1 if and
    only if the labels include 'followup' AND
    the action is explicitly optional or
    conditional. Otherwise, set to 0. If the
    overall response says seeking medical
    attention is mandatory (the visit itself)
    but contains a condition for a specific
    procedure, set to 0. For example, "I would
     recommend that you go to the nearest
    Urgent Care... and get a tetanus vaccine
    if it has been over 5 years." Here, the
    visit is not conditional, only the vaccine
     is. Therefore, is_conditional should be
    0.

OUTPUT FORMAT
Return a single JSON list of objects. Each
    object must contain:
"labels": A list of applicable strings ["
    problem", "test", "treatment", "followup
    "].
"is_severe": Integer 0 or 1.
"is_conditional": Integer 0 or 1.

EXAMPLES
Example 1
Input: Context: "After the treatment of
    contact dermatitis and scabies, many
    patients show changes in dermatitis. On
    one hand, it is related to scabies itself,
     on the other hand, the treatment drugs
    mainly based on sulfur have a significant
    impact on the skin. Therefore, during and
    after the treatment, attention should be
    paid to avoid further damage and protect
    the skin. However, it is necessary to
    first confirm whether the scabies has been
     cured. If the scabies has been cured, the
     main focus should be on anti-allergy
    treatment.", IAA Texts: ["After the
    treatment of contact dermatitis and
    scabies, many patients show changes in
    dermatitis.", "On one hand, it is related
    to scabies itself, on the other hand, the
    treatment drugs mainly based on sulfur
    have a significant impact on the skin.", "
    Therefore, during and after the treatment,
     attention should be paid to avoid further
     damage and protect the skin.", "However,
    it is necessary to first confirm whether
    the scabies has been cured.", "If the
    scabies has been cured, the main focus
    should be on anti-allergy treatment."]
Output: [{"labels": ["problem"], "is_severe":
    0, "is_conditional": 0}, {"labels": ["
    problem", "treatment"], "is_severe": 0, "
    is_conditional": 0}, {"labels": ["
    treatment"], "is_severe": 0, "
    is_conditional": 0}, {"labels": ["followup
    ", "problem"], "is_severe": 0, "
    is_conditional": 0}, {"labels": ["problem
    ", "treatment"], "is_severe": 0, "
    is_conditional": 0}]

Example 2
Input: Context: "Recommend evaluation in ER
    for Xray to evaluate for fracture. The
    nail will take up to 6 months to grow back
    .", IAA Texts: ["Recommend evaluation in
    ER for Xray to evaluate for fracture.", "
    The nail will take up to 6 months to grow
    back."]
Output: [{"labels": ["followup", "test"], "
    is_severe": 1, "is_conditional": 0}, {"
    labels": ["problem"], "is_severe": 0, "
    is_conditional": 0}]

Example 3
Input: Context: "It is estimated to be a
    disease related to capillary hemangioma.
    Continue to observe, and if it enlarges,
    surgical removal is recommended.", IAA
    Texts: ["It is estimated to be a disease
    related to capillary hemangioma.", "
    Continue to observe, and if it enlarges,
    surgical removal is recommended."]
Output: [{"labels": ["problem"], "is_severe":
    0, "is_conditional": 0}, {"labels": ["
    followup", "treatment"], "is_severe": 0, "
    is_conditional": 1}]

Example 4
Input: Context: "The laceration is not
    obviously infected but the stitches were
    likely removed prematurely. Clean with
    soap and water daily and cover with dry
    dressing. If there is pus or spreading
    redness seek evaluation in urgent care.",
    IAA Texts: ["The laceration is not
    obviously infected but the stitches were
    likely removed prematurely.", "Clean with
    soap and water daily and cover with dry
    dressing.", "If there is pus or spreading
    redness seek evaluation in urgent care."]
Output: [{"labels": ["problem"], "is_severe":
    0, "is_conditional": 0}, {"labels": ["
    treatment"], "is_severe": 0, "
```

```
    is_conditional": 0}, {"labels": ["followup
    "], "is_severe": 1, "is_conditional": 1}]

Example 5
Input: Context: "The bleeding is likely caused
     by a blood vessel at the site that has
    not yet clotted. Apply a pressure dressing
     with gauze and tape but you can wait
    until tomorrow to see a doctor unless you
    have symptoms such as continuous bleeding
    not controlled by the dressing.", IAA
    Texts: ["Apply a pressure dressing with
    gauze and tape but you can wait until
    tomorrow to see a doctor unless you have
    symptoms such as continuous bleeding not
    controlled by the dressing.", "The
    bleeding is likely caused by a blood
    vessel at the site that has not yet
    clotted."]
Output: [{"labels": ["problem", "treatment", "
    followup"], "is_severe": 0, "
    is_conditional": 1}, {"labels": ["problem
    "], "is_severe": 0, "is_conditional": 0}]

OUTPUT FORMAT
Return strictly a JSON list of objects as
    specified above. Maintain a strict 1:1
    mapping with the IAA Texts list from the
    Input.
```

E.7 Prognosis Extraction Prompt

```
You are a precise linguistic analysis engine
    specialized in medical context extraction.
     Your task is to identify and extract all
    'sentences' that qualify as Prognosis from
     a response, distinguishing it from IAA (
    Identification, Assessment, or Advice).

INPUT DATA
You will receive:
1. Response: A list of strings (medical
    answers) to analyze.

DEFINITIONS
A sentence is Prognosis if it describes future
     outcomes, predictions, or expectations.

EXCLUSION CRITERIA
IAA: Exclude sentences that describe the
    current diagnosis/condition, current test
    requirements, current treatment steps, or
    current referral/follow-up instructions.
    These are classified as IAA (present), not
     Prognosis (future). However, they are not
     mutually exclusive; a sentence can be
    both IAA and Prognosis if it describes the
     current condition while also predicting
    future outcomes.

PROCESSING LOGIC
For each item in the Response list, perform
    the following steps:
Step 1: Analyze the entire text of the
    response. Do not stop after finding the
    first relevant sentence as multiple
    Prognosis can exist.
Step 2: Analyze every sentence. Keep the
    sentence only if it meets the Prognosis
    definition.
Multi-Sentence Spans: A single prognosis
    prediction may span multiple consecutive
    sentences. Extract consecutive Prognosis
    sentences as one single string.
Split/Gap: If relevant sentences are non-
    consecutive (separated by IAA, irrelevant
    text, or structural breaks like \n\n),
    extract them as separate strings.

Step 3: Extraction Extract the qualifying
    sentences verbatim. Group ALL extracted
    Prognosis strings from a single response
    into a list. If no sentences meet the
    criteria, the result is an empty list [].

EXAMPLES
Example 1
Input: ["The description is about capillary
    hemangioma. Try using ionization to burn
    it, liquid nitrogen is also acceptable.
    Don't squeeze it anymore, it's prone to
    infection.", "After scratching the papular
     urticaria, closely follow up and revisit
    after a week.", "Artificial dermatitis, it
     will get better on its own in a few days
    ."]
Output: [["Don't squeeze it anymore, it's
    prone to infection."], [], ["Artificial
    dermatitis, it will get better on its own
    in a few days."]]
(Reasoning: "Don't squeeze it anymore, it's
    prone to infection." and "Artificial
    dermatitis, it will get better on its own
    in a few days." are both IAAs because the
    former is part of a broader treatment
    recommendation and the latter includes a
    diagnosis. However, since they also
    describe possible outcomes, they qualify
    as Prognosis and are therefore included in
     the output.)

Example 2
Input: ["The nail will likely grow back but
    can take up to 6 months. It is a good sign
     that the nail bed is growing back. It is
    difficult to say if there will be a
    deformity but no further intervention is
    recommended at this time."]
Output: [["The nail will likely grow back but
    can take up to 6 months.", "It is
    difficult to say if there will be a
    deformity but no further intervention is
    recommended at this time."]]
(Reasoning: The first and third sentences
    describe future outcomes regarding nail
    growth and potential deformity, thus
    qualifying as Prognosis. The second
    sentence does not predict future outcomes,
     so it is excluded.)

OUTPUT FORMAT
Return strictly a JSON list of lists of
    strings. Maintain a strict 1:1 mapping
    with the Response list from the Input.
```