

Safety-Aware Dialogue System for Postoperative Oral Cancer Care with Structured Clarification and a Clinically Curated Dataset*

Tzu-Chi Liu¹, Hui-Ying Yang², Shio-Ching Shun³,
Yu-Chi Chen³, Lu-Yen Anny Chen³, Yong-Sheng Chen¹

¹Department of Computer Science, National Yang Ming Chiao Tung University

²School of Nursing, College of Medicine, Chang Gung University

³Institute of Clinical Nursing, National Yang Ming Chiao Tung University
{jonathan.tliu.en11, scshun, yschen}@nycu.edu.tw

Abstract

Clinical dialogue systems are increasingly vital for patient education and follow-up care; however, effective responses often depend on the clinical context that patients often fail to provide in detail. Responding directly to vague messages can therefore lead to generic or clinically misaligned advice, a challenge that is particularly pronounced in post-op oral cancer (OC) care due to speech impairment and functional limitations. Moreover, post-op OC patients often experience psychological distress, making safety-aware language more likely to arise in dialogue. Dialogue systems in this setting must therefore address both clarifying missing clinical context and ensuring psychological safety. We propose a safety-aware dialogue system that applies information-gain guided clarification before RAG-based response generation and screens user utterances for emotional distress and suicidal ideation. Expert evaluations show that the proposed system improves the quality and clinical appropriateness of generated responses relative to strong baselines, while the safety module closely aligns with expert judgments on clinically concerning utterances. Furthermore, we release a clinically curated Chinese post-op OC QA dataset with expert-validated annotations, which we use throughout our experiments.

1 Introduction

Clinical dialogue systems have significant potential to enhance patient education, symptom management, and follow-up care (Kung et al., 2023; Singhal et al., 2025). However, patients often express their concerns through brief and subjective messages that lack critical clinical context (Alian-nejadi et al., 2019; Zamani et al., 2020). When dialogue systems respond directly to these vague

*Code, data, and prompts are available at <https://github.com/jonathan-tliu/oral-cancer-dialogue-dataset>.

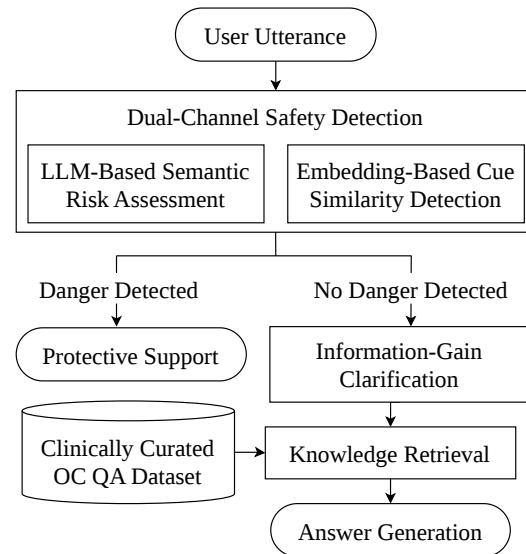


Figure 1: Overview of the proposed safety-aware dialogue system for post-op OC care. Incoming utterances are first screened by the dual-channel safety module. Non-risk cases proceed through information-gain (IG) guided clarification and retrieval-augmented generation, while risk cases are routed to a protected supportive response mode.

inputs, the resulting advice can be generic, mismatched to the condition of the patient, or may overlook concerning psychological signals, particularly in oncology and post-op care settings (Gibson et al., 2024; Pestian et al., 2010).

This issue is amplified in post-op oral cancer (OC) care. Post-surgical speech impairment, oral dysfunction, pain, and fatigue make it difficult for patients to articulate detailed questions, yielding fragmented and underspecified utterances. In this work, we focus on interpreting such underspecified inputs in text-based dialogue settings. While speech-to-text systems can transcribe patient speech, they do not address missing clinical context or ambiguity caused by impaired articulation or fatigue. Our approach, therefore, targets the downstream problem of clarification and in-

terpretation, rather than speech recognition itself. Meanwhile, post-op recovery is frequently accompanied by anxiety, emotional distress, and uncertainty (De Choudhury et al., 2016), increasing the risk of clinically misaligned responses when replying to unclear inputs. The utterances in our dataset reflect these real-world communication characteristics observed in post-op clinical interactions.

To address these challenges, we propose a safety-aware dialogue system that combines psychological screening with structured clarification. Figure 1 illustrates the overall dialogue flow. Incoming utterances are first screened for psychological risk; non-risk cases are handled via information-gain (IG) guided clarification, followed by RAG, while risk cases are routed to a protected supportive mode. Representative dialogue examples for both pathways are provided in Appendix A. We evaluate our approach with expert evaluations in realistic post-op OC scenarios. Our contributions are: (1) an IG-guided clarification policy for clinical dialogue, (2) a dual-channel semantic detector for emotional distress and suicidal ideation, and (3) a Chinese post-op OC QA dataset with expert-validated annotations. To support reproducibility, we release our code, prompts, and dataset resources at <https://github.com/jonathan-tcliu/oral-cancer-dialogue-dataset>.

2 Related Work

2.1 Clinical Dialogue Systems

Clinical dialogue systems have been explored for education, triage, and long-term care. Although recent large language models (LLMs) perform well on medical QA and exam-style tasks, studies consistently report risks such as hallucinations, domain mismatch, and socio-emotional inadequacies in real-world interactions (Kung et al., 2023; Gibson et al., 2024; Singhal et al., 2025). Prior work has introduced safety mechanisms including guardrails, refusal strategies, escalation protocols, and empathy modeling (Shuster et al., 2022; Sorin et al., 2024; Liu et al., 2024). However, many systems still assume user inputs provide enough detail for immediate responses, leaving the problem of structured clarification largely unaddressed in clinical dialogue.

2.2 RAG in High-Risk Domains

RAG grounds model outputs in external documents and reduces hallucinations in knowledge-intensive tasks (Lewis et al., 2020; Izacard and Grave, 2020). Dense retrieval pipelines are generally more reliable than purely parametric models in safety-aware settings (Shuster et al., 2021; Wang et al., 2024), and clinical RAG systems further emphasize provenance and guideline alignment (Lu et al., 2024). However, most RAG frameworks assume that user queries contain sufficient information to guide retrieval. This limitation becomes more pronounced in clinical RAG settings. In post-op OC dialogue, vague patient messages often yield diffuse or generic evidence, highlighting the need for clarification before retrieval rather than attempting to correct ambiguity only after retrieval or generation.

2.3 Clarification in Conversational Retrieval

Clarification has been studied in information retrieval and QA. Prior work shows that well-chosen clarifying questions can significantly improve retrieval quality (Aliannejadi et al., 2019; Zamani et al., 2020). Recent interactive and active RAG approaches further integrate clarification into retrieval and generation processes (Jiang et al., 2023; Kim et al.). Unlike open-domain settings, clarification in clinical dialogue focuses on a small set of clinically meaningful facets, where follow-up questions directly affect safety and decision quality. Our work extends this line by introducing an IG-guided clarification policy for safety-aware clinical dialogue.

2.4 Slot Filling and Clinical Triage Pipelines

Task-oriented dialogue systems commonly use slot schemas to capture user goals and contextual information (Larson and Leach, 2022; Qin et al., 2023). In the medical domain, prior work applies slot filling and symptom extraction for triage, dialogue modeling, and summarization (Zeng et al., 2020; Zhou et al., 2023; Tang et al., 2023). However, many approaches assume complete user input or generate recommendations without explicit clarification. Our framework addresses this gap by introducing clinically grounded OC facets and IG to guide clarification before retrieval and response generation.

2.5 Detection of Concerning Language

Detection of language indicating psychological distress or suicidal ideation has been studied using lexicon-based methods (Pestian et al., 2010; Copersmith et al., 2014; De Choudhury et al., 2016) and transformer-based classifiers (Sawhney et al., 2021; Mitsios et al., 2024). While keyword matching approaches are transparent, they struggle with paraphrased or implicit expressions; transformer-based models improve semantic sensitivity but remain unreliable with short, ambiguous clinical utterances. Consequently, safety-aware dialogue systems often adopt conservative guardrail-based detection strategies (Shuster et al., 2022; Sorin et al., 2024). Our work aligns with this perspective by framing concerning-language detection as a semantic judgment task that supports safety-aware dialogue control.

3 Dataset

We release a clinically curated dataset for post-op OC dialogue: a facet-annotated QA corpus and a clinically concerning utterance cue lexicon, both reviewed by clinical experts.

3.1 Data Collection and Annotation

FOCUSP QA Corpus The corpus includes 867 Chinese QA pairs that address the informational and psychosocial needs of OC patients, which are usually unmet in general datasets. We use the FOCUSP framework (Shun, 2020) to organize OC support across six key domains: Family Involvement, Optimistic Attitude, Coping Effectiveness, Uncertainty Reduction, Symptom Management, and Physical Enhancement. Content was sourced from three qualitative channels:

- **Interviews and Focus Groups:** This study comprised 13 sessions, including 11 individual interviews with 9 OC patients and 2 caregivers, along with 2 focus groups of 8 and 6 patients, respectively. Sessions lasted 45 to 120 minutes, and communication accommodations were provided for participants with speech impairments. All sessions were recorded and transcribed. Clinical experts thematically coded the transcripts to identify informational needs, symptom-related concerns, and psychosocial issues, transforming patient responses into 720 QA pairs that represent realistic post-op dialogue.
- **Clinical Observations:** Over three months,

we observed 32 patient-clinician interactions in both inpatient and outpatient OC units, covering various treatment phases: pre-op (n=3), post-op recovery (n=14), and adjuvant chemoradiotherapy (n=15). These observations resulted in 83 distinct patient questions.

- **Expert-Evaluated Questions:** This work gathered 64 frequently asked questions from an OC case manager and a patient-education nurse, based on authentic clinical encounters.

Across the three sources, interview- and focus-group-derived QA items constitute the largest portion of the corpus, as they most closely reflect naturally occurring post-op dialogue. Clinical observations and expert-curated questions contribute lower-volume but complementary subsets that capture clinically salient scenarios and frequently encountered informational needs.

All QA items were thematically organized and expert-reviewed to ensure clinical consistency. Responses were composed or reviewed by clinical experts using educational materials. Annotation followed a double-review protocol, in which each item was independently reviewed by multiple clinical experts, and discrepancies were resolved by consensus. Table 1 summarizes the distribution across domains. A single QA item may be assigned to multiple FOCUSP domains when it reflects overlapping informational and psychosocial needs. As a result, domain counts may exceed the total number of QA pairs.

Example QA item. To illustrate the structure of the dataset, we provide a representative example of a post-op OC QA pair with facet annotations. Full dialogue examples demonstrating how such QA items are used within the system are provided in Appendix A:

Patient question: “Recently, I feel very dry when eating, and I have some difficulty swallowing. Is this normal?”

Expert answer: “Experiencing dryness and difficulty swallowing after surgery is relatively common, and may be related to reduced saliva production or ongoing tissue recovery. You may try eating smaller, more frequent meals and choosing softer, more moist foods, while ensuring adequate fluid intake. If the symptoms persist or worsen, it is recommended to consult your medical team.”

Facet annotations: *Treatment Phase:* post-op recovery; *Functional & Symptom Impact:* dys-

FOCUSP Domain	#	
	QA Pairs	Example Focus
Family Involvement	354	Family support
Optimistic Attitude	84	Emotional encouragement
Coping Effectiveness	84	Managing stress/anxiety
Uncertainty Reduction	351	Prognosis, side effects
Symptom Management	354	Eating difficulties, dryness
Physical Enhancement	36	Rehabilitation exercises

Table 1: QA distribution across FOCUSP domains.

phagia, xerostomia; *FOCUSP Domain*: symptom management.

Clinically Concerning Utterance Set We additionally release a de-identified lexicon of clinically verified cue phrases associated with emotional distress and suicidal ideation in post-op OC dialogues. The lexicon is derived from anonymized patient-clinician messages, validated by OC specialists, and used as a lightweight trigger resource in our safety module.

3.2 Schema and Slot Design

We organize the annotation schema into three complementary facets: clinical state, care-related context, and information and support intent. While facets may co-occur and are queried independently, these groups are conceptual rather than hierarchical, reflecting how clinicians consider multiple dimensions of patient context during post-op consultations.

The schema serves two purposes: it provides a consistent annotation framework for the QA dataset, and it defines the action space for clarification during dialogue by specifying which clinically meaningful aspects can be queried to reduce ambiguity. For instance, a symptom report may trigger clarification on *Treatment Phase* (e.g., post-op recovery vs. adjuvant chemoradiotherapy) and *Functional & Symptom Impact* (e.g., dysphagia or xerostomia), while psychosocial needs can be captured through the *FOCUSP* support domain (e.g., uncertainty reduction vs. coping effectiveness).

Detailed definitions of these facets and their intended clinical roles are included in Appendix B. The clarification prompts designed by clinicians are included in the released dataset, along with the schema that specifies the exact question formulations used by the controller, not just the facet names. This approach enables accurate replication and fair ablations across different clarification de-

signs.

During dataset construction, each QA item was annotated with one or more relevant facets through a two-stage process. Initial labels were assigned via rule-based thematic matching and subsequently reviewed and refined by clinical experts to ensure alignment with underlying clinical intent. These annotations are used to guide clarification decisions and also to support retrieval control in downstream response generation. All data were collected under institutional ethical approval, fully de-identified prior to release, and are made available for academic research use.

4 Methodology

The system is designed for post-op OC dialogue, where responding immediately to underspecified patient messages may be unsafe when psychological vulnerability is present. To address this, we adopt a safety-aware RAG controller that governs the entire dialogue flow. Each incoming utterance is first screened for clinically concerning meanings. If psychological risk is detected, the system switches to a protected supportive mode. Otherwise, it proceeds through an information-seeking pathway consisting of structured clarification, evidence retrieval, and the generation of a controlled response. This design reflects the clinical principle that patient safety takes precedence over completeness of information.

4.1 Dual-Channel Safety Detection

Building on this safety-first routing principle, we design a dual-channel safety detection module consisting of: (i) an *LLM-Based Semantic Risk Assessment* channel, and (ii) an *Embedding-Based Cue Similarity Detection* channel. The LLM-based channel uses a safety-oriented prompt with GPT-4o to identify overwhelming emotional distress and suicidal ideation at the semantic level. The embedding-based channel compares utterance embeddings generated by the text-embedding-3-large model against a curated set of clinically relevant cue phrases, enabling the detection of paraphrased or non-literal expressions. A positive signal from either channel flags the utterance as requiring safety-aware handling.

For the embedding-based channel, we embed the incoming utterance and compute its maximal cosine similarity to the curated cue set, separately for extreme distress and suicidal ideation; the ut-

terance is flagged when the score exceeds a fixed threshold τ (with $\tau = 0.5$). For auditing, the system logs trigger signals and a brief LLM rationale without exposing risk labels to users.

When risk is detected, the system enters a protective mode. Clarification and retrieval are suppressed, and response generation is restricted to a pre-validated set of supportive messages that emphasize emotional validation, avoid prescriptive advice, and encourage seeking professional or emergency care when appropriate.

If no risk cues are detected, the system proceeds with structured clarification and RAG. This gatekeeping mechanism ensures that information-seeking behavior is engaged only when it is clinically appropriate, aligning system behavior with real-world post-op OC care practices.

4.2 Information-Gain Guided Clarification

Overview. After passing safety screening, the system enters an information-seeking mode in which it must retrieve clinically appropriate information before generating an answer. We therefore perform *structured clarification* over a clinician-defined set of *facets* (Section 3.2). Given a user query, dense retrieval produces a *candidate set* $C = \{c_1, \dots, c_n\}$ of potentially relevant QA pairs. A clarification question selects one facet f and asks the user to choose value(s) for f , thereby partitioning C into subsets and reducing ambiguity. The central problem is deciding which facet to clarify next, using an IG criterion that quantifies the expected reduction in retrieval ambiguity.

Narrowing of candidate set. Let C denote the current candidate set at a clarification round. For a facet f , let V_f denote the set of all possible values of facet f observed in C . Let $V_f(c_i)$ denote the annotated value set of facet f for candidate c_i . Intuitively, if we ask about facet f and the user selects value v , the candidate set C is narrowed to C_v , where $C_v = \{c_i \in C \mid v \in V_f(c_i)\}$.

Retrieval ambiguity of candidate set. The selection of the best-matching candidate in C is based on the retrieval ambiguity of C . We adopt a uniform prior over candidates and define the ambiguity of C by $H(C) = \log |C|$, which increases with the number of remaining plausible candidates. The ambiguity of the narrowed candidate set C_v after receiving the user response v becomes $H(C_v)$, where $H(C_v) = \log |C_v|$.

Information gain. The empirical frequency of each response value v within C is $p_v = \frac{|C_v|}{|C|}$. The expected retrieval ambiguity for each facet f is then

$$H(C \mid f) = \sum_{v \in V_f} p_v H(C_v).$$

We define the IG of querying facet f as the expected reduction in ambiguity:

$$IG(f) = H(C) - H(C \mid f).$$

At each clarification round, we compute $IG(f)$ for all facets that exhibit sufficient variation in the current candidate set. We select the facet \hat{f} with the largest IG, and check whether $IG(\hat{f}) \geq \epsilon$ ($\epsilon = 0.05$ in this study); otherwise, clarification stops.

Question presentation. The selected facet \hat{f} is presented using a concise, clinician-designed prompt with representative options. For example, when the *treatment phase* facet is selected, the system may ask “Which stage of treatment are you currently in?” with options such as *post-surgery*, *undergoing radiotherapy*, or *follow-up care*. Similarly, for facets such as functional impact or psychosocial concerns, the prompt provides a set of clinically meaningful choices. Users can choose one or multiple options based on their preferences, or skip the question if unsure.

Soft re-ranking with user responses. Rather than hard-filtering candidates, which can be brittle in the presence of annotation noise or occasional user misselection, we apply a soft re-ranking strategy that rewards candidates consistent with the users clarification responses. Once a facet \hat{f} is selected for clarification according to the IG criterion described above, the system presents a small set of representative value options derived from the current candidate set. Specifically, we present at most five values in $V_{\hat{f}}$ with the most empirical frequency values within C as selectable options, as well as an additional “*Other*” option that captures all remaining or unspecified values. This design balances coverage and practicality while allowing users to indicate uncertainty or cases not covered by the dominant values.

When the user selects the “*Other*” option, the response is interpreted as indicating that none of the explicitly presented top-five values applies. In this case, the consistency score, described below,

rewards candidates whose facet values fall outside the presented options, i.e., whose $V_{\hat{f}}(c_i)$ does not intersect with the top five values. This strategy allows clarification to remain informative even when the user’s situation lies in the long tail of the facet-value distribution, without forcing an incorrect commitment to a dominant category.

Let $A_{\hat{f}} \subseteq \mathcal{V}_{\hat{f}}$ denote the set of values selected by the user, and let $V_{\hat{f}}(c_i)$ denote the annotated value set of the facet \hat{f} for candidate c_i . We define a facet-level matching score for candidate c_i as

$$m_{\hat{f}}(c_i) = \frac{|A_{\hat{f}} \cap V_{\hat{f}}(c_i)|}{|A_{\hat{f}}|}.$$

This formulation measures the degree to which facet annotations align with the user selection, without penalization of multi-valued annotations.

Let \mathcal{F} denote the set of best facets queried during clarification rounds. The overall bonus score for candidate c_i is computed as the average consistency across queried facets:

$$b_i = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} m_f(c_i).$$

The final ranking score $S(c_i)$ combines the FAISS similarity score s_i (Douze et al., 2025) and the clarification-based bonus b_i :

$$S(c_i) = \alpha s_i + \beta b_i,$$

where s_i is normalized to $[0, 1]$ within the current candidate set, and $\alpha = \beta$ in our experiments. This soft re-ranking maintains candidate coverage while steering retrieval considering the consistency with the user-clarified clinical context.

At the final clarification round, the user has provided sufficient facet-level information to meet the clarification need. Instance-level commitment is then performed automatically by the system through selecting the candidate \hat{c}_i with the highest final ranking score $S(\hat{c}_i)$. This clarification design allows users to provide guidance through clinically meaningful context, while the system retains responsibility for selecting the specific QA instances used to generate responses.

Clarification behavior in practice. In our experimental setting, clarification is triggered in approximately 97% of user interactions, reflecting the prevalence of underspecified patient utterances in post-op OC dialogue. The average clarification

depth is around three turns, indicating that clinically relevant context can typically be resolved within a small number of interactions. Clarification terminates when the IG of the best candidate facet falls below a predefined threshold ($\epsilon = 0.05$), as described above. This design balances effective ambiguity reduction with efficient interaction.

4.3 RAG-Based Answer Generation

After the clarification phase, the system proceeds to response generation grounded in the selected QA pair \hat{c}_i . At this stage, the relevant clinical question has already been determined through user clarification, and the role of RAG is to construct a safe and patient-specific response based on the associated expert-curated answer.

Grounded context construction. The generation of the final response depends only on the selected QA pair \hat{c}_i . Its question text, expert-defined answer, structured facet annotations, and provenance information are assembled together with the original user utterance to form a grounded input for the generation model. This design ensures that all generated responses are explicitly anchored to a single, clinically validated source, supporting interpretability and auditability.

Constrained and safety-aware generation. Response generation is guided by a clinician-designed system prompt, released along with the dataset, which constrains the model to produce patient-appropriate responses grounded strictly in the selected QA content \hat{c}_i , and to refrain from individualized diagnoses or unsupported extrapolation. To ensure psychological safety, response generation remains subject to the upstream risk detection module. If a safety event is detected during generation, the system immediately suppresses evidence-grounded answering and falls back to the protected supportive response mode described in Section 4.1. This design ensures that grounded generation never takes precedence over safety-aware dialogue control.

5 Experiments and Results

We evaluate our system along two complementary dimensions. First, we evaluate the clinical quality and appropriateness of generated responses through expert preference judgments. Second, we evaluate the systems ability to identify clinically concerning meanings in patient utterances, focusing on emotional distress and suicidal ideation as

two critical safety concerns in post-op OC dialogue. Together, these evaluations examine both response effectiveness and safety-related reliability in realistic clinical scenarios. Given the high cost and limited availability of clinical expert annotation, we design these studies as targeted evaluations that prioritize depth and clinical validity over scale.

5.1 Experimental Setup

For clinical answer quality evaluation, we randomly sampled 30 patient questions from a curated pool of 50 candidates. For semantic evaluation of clinically concerning utterances, we sampled 30 utterances, each independently annotated by experts along two binary dimensions: emotional distress and suicidal ideation. Both studies were conducted with the same panel of 18 clinical experts. The expert panel consisted of multidisciplinary clinical professionals, including three oncologists, eight oncology nurses, three clinical psychologists, two nutritionists, one physical therapist, and one nursing informatics specialist. Clarification behavior in our system follows the statistics described in Section 4.2, where clarification is triggered in most interactions and typically resolves within a few turns.

5.2 Evaluation of Clinical Answer Quality

We evaluate the clinical quality of responses generated by models through an expert ranking study. Our goal is to determine whether structured clarification and safety-aware retrieval enhance answer quality beyond what can be achieved by strong general-purpose or medical-domain language models alone.

Evaluation protocol. For each question, experts ranked responses from all four systems based on five criteria: (1) clarity and coherence, (2) readability, (3) suitability for post-op OC patients, (4) professional medical tone, and (5) overall quality. The top-ranked response received 4 points, and the lowest-ranked response received 1 point. Scores were averaged across experts and questions to obtain mean scores for each system and criterion. All systems were evaluated using the same user queries without access to expert annotations.

Models compared. We compare four systems representing complementary design choices in clinical dialogue: DeepSeek-R1 (strong Chinese-language baseline), MedGamma-27B (medical-

domain LLM without explicit retrieval control), GPT-4o (general-purpose conversational model), and OC Bot (ours), a safety-aware RAG-based dialogue system with IG-guided clarification and risk-aware control; GPT-4o is used only for final response generation, with retrieval, clarification, and safety decisions governed by our controller.

Results. Table 2 presents the mean scores aggregated across experts and questions, computed from the ranking-based judgments. OC Bot achieves the highest overall mean score and the top average performance across four of five criteria, including clarity, readability, professional medical tone, and overall quality.

For the criterion of *suitability for OC patients*, MedGamma-27B slightly outperforms the other models, with OC Bot showing comparable performance. This result is expected, as this dimension primarily reflects general medical plausibility and tone, where large medical-domain or conversational models already perform strongly.

OC Bot consistently outperforms the baseline models in dimensions that require structured reasoning and contextual understanding. This indicates that improvements are not solely due to model capacity, even when the same backbone is utilized. Instead, these enhancements result from the system-level integration of IG-guided clarification, retrieval grounding, safety-aware control, and the clinically curated facet-annotated dataset.

The results indicate consistent improvements across various evaluation criteria and among experts. However, the number of evaluation questions is limited due to the high cost of expert annotation. Therefore, the findings of this study should be viewed as evidence from a focused expert assessment rather than strong claims of statistical significance or broad generalization.

5.3 Semantic Evaluation of Clinically Concerning Utterances

We evaluate the systems ability to identify clinically concerning meanings in post-op OC utterances, focusing on semantic signals of psychological vulnerability. Rather than treating this task as risk prediction, we frame it as a semantic interpretation problem aligned with clinical judgment, distinguishing between (i) emotional distress and (ii) suicidal ideation.

Data and annotation protocol. We evaluated 30 utterances annotated by 18 clinical experts. For

Criterion	GPT-4o	Deep-Seek	Med-Gamma	OC Bot
Clarity	2.471	2.427	<u>2.534</u>	2.575
Readability	2.460	2.479	<u>2.498</u>	2.580
OC Suitability	2.480	2.403	2.579	<u>2.563</u>
Professionalism	2.388	2.519	<u>2.546</u>	2.566
Overall	2.443	<u>2.502</u>	2.496	2.583

Table 2: Expert evaluation results for clinical answer quality. Values indicate mean ranking-based scores averaged across experts and questions. For each criterion, the best-performing system is shown in **bold**, and the second-best is underlined. Higher scores indicate better performance, where the top-ranked response receives 4 points, and the lowest-ranked response receives 1 point. Model names are abbreviated here for readability.

each utterance, experts independently assigned two binary labels corresponding to emotional distress and suicidal ideation. Annotations were aggregated by expert agreement, and the final labels were determined by majority vote.

Methods compared. We compare four approaches for detecting clinically concerning utterances: (i) Keyword Matching, a lexicon-based detector that flags utterances containing our curated distress or suicide-related cues; (ii) Emotion Transformer, a transformer-based emotion classifier following Mitsios et al. (Mitsios et al., 2024); (iii) Suicidal Ideation Transformer, a binary semantic classifier adapted from prior suicide ideation detection work (Sawhney et al., 2021); and (iv) Hybrid (Ours), a dual-channel detector that combines embedding-based cue similarity with LLM-based semantic judgment, where a positive signal from either channel triggers detection.

Evaluation metrics. We report agreement with expert majority judgments for each semantic dimension. Accuracy is computed against majority labels, with error patterns analyzed in cases of expert disagreement. Given the safety-aware nature of the task, we emphasize interpretability and conservative behavior over aggressive optimization.

Results. Table 3 summarizes agreement with expert majority judgments across the two semantic dimensions. Overall, the proposed hybrid approach achieves the highest alignment with expert assessments for both emotional distress and suicidal ideation, substantially outperforming all baseline methods.

For emotional distress, keyword matching is not

Method	Emotional Distress	Suicidal Ideation
Keyword Matching	36.7%	73.3%
Emotion Transformer	73.3%	–
Suicidal Transformer	–	46.7%
Hybrid (Ours)	96.7%	86.7%

Table 3: Agreement with expert majority judgments for clinically concerning utterance semantics. Higher values indicate closer alignment with expert analysis.

Method	FPR (EN)	FNR (EN)	FPR (SI)	FNR (SI)
Keyword Matching	0.0%	73.1%	6.7%	53.3%
Emotion Transformer	50.0%	23.1%	–	–
Suicidal Transformer	–	–	100.0%	0.0%
Hybrid (Ours)	0.0%	3.8%	1.3%	6.7%

Table 4: False positive rate (FPR) and false negative rate (FNR) for clinically concerning utterance detection under majority expert labels. EN denotes emotional distress, and SI denotes suicidal ideation.

very effective, achieving only 36.7%, reflecting the fact that severe distress in post-op OC dialogue is often conveyed implicitly rather than through explicit lexical cues. The transformer-based emotion classifier substantially improves alignment, achieving 73.3% accuracy, indicating greater sensitivity to contextual and semantic signals. The hybrid approach further improves performance to 96.7%, demonstrating that combining embedding-based semantic similarity with LLM-based judgment enables robust detection of both paraphrased and context-dependent expressions of distress.

For suicidal ideation, keyword matching shows moderate agreement at 73.3%, benefiting from the curated lexicon of domain-specific suicide-related cues. However, the suicidal ideation transformer alone performs markedly worse, achieving only 46.7% agreement. This suggests it may be less effective when applied to ambiguous clinical statements. Our hybrid approach achieves the highest agreement of 86.7%, indicating that integrating explicit cue sensitivity with semantic-level interpretation provides more reliable detection of suicidal ideation in post-op OC dialogue.

Table 4 further reports FPR and FNR under expert labels, providing a more detailed view of safety behavior. Lexicon-based keyword matching exhibits extremely high FNRs for emotional distress (73.1%) and suicidal ideation (53.3%), highlighting its limited coverage of paraphrased or im-

Round	Avg. Candidate Size		Avg. Entropy	
	IG	Random	IG	Random
0	37.48	37.48	–	–
1	18.68	30.77	1.299	0.467
2	10.85	25.20	0.927	0.446
3	8.38	20.87	0.858	0.424

Table 5: Comparison of information-gain (IG) guided clarification and random clarification regarding average candidate set size and entropy over clarification turns. Lower values indicate less retrieval ambiguity.

plicit expressions. The emotion transformer reduces false negatives for extreme distress (23.1%) but introduces a high FPR (50.0%), potentially leading to frequent over-triggering in practice. The suicidal ideation transformer achieves zero false negatives but triggers on nearly all non-suicidal utterances, resulting in a 100.0% FPR.

In contrast, the proposed hybrid detector maintains both low FNRs (3.8% for extreme distress and 6.7% for suicidal ideation) and low FPRs (0.0% and 1.3%, respectively). This balance reflects a clinically appropriate trade-off: the system prioritizes minimizing the risk of overlooking critical safety signals while avoiding excessive or unnecessary triggering in ambiguous cases.

5.4 Clarification Effectiveness Evaluation

To verify that our IG-guided clarification policy reduces retrieval ambiguity, we compare it against a random clarification baseline using two turn-level diagnostics: (i) the average size of the retrieval candidate set, and (ii) the entropy of the facet-value distribution over the remaining candidates after each clarification turn. A lower candidate set size indicates stronger narrowing of relevant evidence, while lower entropy indicates reduced ambiguity regarding clinically meaningful facets.

Table 5 shows that IG-guided clarification narrows the candidate set substantially faster than random clarification. Starting from the same initial neighborhood, IG reduces the average candidate size from 37.48 to 8.38 by turn 3, whereas random clarification only reduces it to 20.87. We further observe a consistent decrease in facet-value entropy across turns under IG clarification, indicating progressively reduced ambiguity regarding clinically meaningful facets. In contrast, random clarification shows only a marginal reduction in entropy. The lower initial entropy observed for

random clarification reflects the fact that randomly selected facets may occasionally partition candidates unevenly without systematically targeting clinically relevant sources of ambiguity. Overall, these results indicate that IG-guided clarification better focuses retrieval on clinically relevant evidence under a small interaction budget than unguided random questioning.

6 Conclusion

We developed a safety-aware dialogue system for post-op OC care that incorporates structured clarification, RAG, and semantic-level safety screening. By treating clarification as an IG optimization based on clinically relevant factors, the system avoids making premature responses to vague patient queries while remaining interpretable and controllable.

Expert evaluations indicate that our proposed system enhances the quality of clinical answers compared to strong general-purpose and medical-domain baselines, even when using the same generation framework. In addition, our semantic evaluation demonstrates that the proposed hybrid detector closely aligns with expert judgments on emotional distress and suicidal ideation, supporting its role in guiding safety-aware dialogue control rather than performing clinical diagnosis or risk stratification.

Furthermore, we are releasing a clinically curated Chinese QA dataset and domain-specific distress cue lexicons to support reproducible research on safety-aware clinical dialogue. We believe this work lays a practical foundation for implementing RAG-based dialogue systems in high-risk post-op OC environments.

7 Limitations

This work has several limitations. First, the proposed framework has been evaluated specifically in the context of post-op OC care, and its facet schema and clarification policy are tailored to this area. Although the overall architecture is general, adapting the schema for use in other medical specialties will require additional input from experts and re-annotation.

Second, due to the high cost and limited availability of clinical experts, we were unable to conduct a comprehensive ablation study of clarification strategies within the clinical answer quality evaluation. In particular, evaluating multiple clar-

ification variants (e.g., with vs. without clarification, or different clarification policies) would substantially increase the number of dialogue conditions and generated responses requiring expert review. Given the substantial time and cognitive effort required for each clinical evaluation, we prioritized depth and reliability of expert judgment over breadth of experimental coverage, and restricted expert assessments to a small number of representative systems and scenarios.

As a result, while we separately analyze the effectiveness of clarification through candidate reduction and entropy-based diagnostics, its direct contribution to perceived answer quality is not isolated through expert preference comparisons. We view this as an inherent constraint of expert-in-the-loop evaluation in high-risk clinical dialogue settings, and leave large-scale or longitudinal user studies, as well as broader clarification ablations, to future work.

Third, while our dataset of post-op OC dialogue is constructed in Chinese, the proposed framework is not inherently limited to this specific setting. The facet schema is designed to capture clinically meaningful dimensions of patient context, and is largely language-independent. With appropriate translation and adaptation of facet definitions and clarification prompts, the same clarification-driven approach can be applied to other languages.

More broadly, although the current dataset focuses on post-op OC care, the overall framework, combining safety-aware routing, structured clarification, and retrieval grounding, can be extended to other clinical domains where user inputs are often underspecified and safety considerations are critical. Such adaptation would require domain-specific schema design and expert-validated data.

Another limitation of our approach is the potential user burden introduced by multi-turn clarification. While clarification helps resolve underspecified inputs and improve response relevance, it requires additional user interaction, which may not always be desirable in real-world settings.

In this study, clarification behavior is evaluated through expert-based assessment rather than direct patient interaction. Although the observed clarification depth is relatively small (approximately three turns on average), the actual usability and perceived burden in patient-facing scenarios remain to be validated. Evaluating user experience, interaction efficiency, and acceptance of clarification strategies requires carefully designed clinical

studies. A patient-facing pilot study is currently under ethical review as part of ongoing work.

Finally, the system is intended to support, not replace, professional medical judgment. System responses are informational and supportive in nature and should not be interpreted as personalized medical advice.

References

- Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.
- Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.
- Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.
- Damien Gibson, Stuart Jackson, Ramesh Shanmugasundaram, Ishith Seth, Adrian Siu, Nariman Ahmadi, Jonathan Kam, Nicholas Mehan, Ruban Thanigasalam, Nicola Jeffery, and 1 others. 2024. Evaluating the efficacy of chatgpt as a patient education tool in prostate cancer: multimetric assessment. *Journal of Medical Internet Research*, 26:e55939.
- Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.
- Gangwo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models, 2023. URL <https://arxiv.org/abs/2310.14696>.

- Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2023. Performance of chatgpt on usml: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.
- Stefan Larson and Kevin Leach. 2022. A survey of intent classification and slot-filling datasets for task-oriented dialog. *arXiv preprint arXiv:2207.13211*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, and 1 others. 2024. Large language models in the clinic: a comprehensive benchmark. *arXiv preprint arXiv:2405.00716*.
- Yuxing Lu, Xukai Zhao, and Jinzhao Wang. 2024. Clinicalrag: enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68.
- Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, and 1 others. 2024. Improved text emotion prediction using combined valence and arousal ordinal classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 808–813.
- John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.
- Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2311.09008*.
- Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, pages 2176–2190.
- Shiow-Ching Shun. 2020. Proposing a comprehensive prehabilitation model for individuals with operable pancreatic cancer. *Asia-Pacific Journal of Oncology Nursing*, 7(3):255–258.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, and 1 others. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.
- Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.
- Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Terminology-aware medical dialogue generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing bert for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, 56(7):1–33.
- Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pages 418–428.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- Sitong Zhou, Kevin Lybarger, Meliha Yetisgen, and Mari Ostendorf. 2023. Generalizing through forgetting-domain generalization for symptom event extraction in clinical notes. *AMIA Summits on Translational Science Proceedings*, 2023:622.

A Example Dialogue Flows

The following examples are anonymized and paraphrased for illustration purposes.

A.1 No-Risk Clarification and Response Flow

User: “I have been having some difficulty eating lately. Is this normal?”

System (Clarification 1): “Could you let me know which stage of treatment you are currently in? For example, are you recovering from surgery or undergoing radiotherapy or chemotherapy?”

User: “I had surgery about a month ago.”

System (Clarification 2): “Are you mainly experiencing difficulty swallowing, dryness in the mouth, or both?”

User: “Both.”

System (Response): “It is quite common to experience both difficulty swallowing and dryness during the post-op recovery phase. These symptoms are often related to tissue healing and reduced saliva production. You may consider choosing softer, more moist foods and eating smaller, more frequent meals. If the symptoms persist or interfere with your daily life, it would be advisable to discuss this with your medical team.”

A.2 Safety-Triggered Supportive Response Flow

User: “I feel like every day is very painful, and sometimes I feel like life is meaningless.”

System (Backend safety detection triggered; no message is shown to the user): The utterance is flagged internally by the dual-channel safety module in the backend due to expressions of severe emotional distress and potential suicidal ideation. This safety trigger is used only for internal system handling and does not generate any message visible to the user.

System (Supportive response): “It sounds like you have been going through a very difficult and painful time. You are not alone in feeling this way, and you do not have to face it by yourself. If possible, consider reaching out to your doctor, a mental health professional, or someone you trust for support. If you feel that your situation is urgent, please consider contacting local emergency services or a crisis support hotline in your area.”

B Clinical Facet Schema for Post-op OC Dialogue

This appendix documents the clinically grounded facet schema used for dataset annotation and clarification control. Rather than enumerating an exhaustive label ontology, the schema defines a small

set of clinically meaningful facet groups that reflect how clinicians interpret underspecified patient messages during post-op OC follow-up. The schema was developed through iterative discussions with clinicians and OC nursing experts and is designed to support both clarification decisions and retrieval control. Importantly, the facet groups described below are conceptual categories rather than hierarchical layers or sequential stages. Each facet can be queried independently, and multiple facets may be relevant to a single patient utterance.

B.1 Clinical State Facets

Clinical state facets describe the patients current medical and functional condition. These facets are essential for interpreting symptoms, assessing the appropriateness of information, and contextualizing responses.

Treatment phase. This facet captures the patients current stage in the treatment trajectory. Treatment phase strongly influences symptom expectations, risk assessment, and response framing. Typical values include pre-op preparation, post-op recovery, concurrent or adjuvant chemoradiotherapy, and survivorship or follow-up care.

Functional and symptom impact. This facet represents functional limitations and symptom burdens commonly observed after OC treatment. Examples include difficulties with chewing or swallowing, speech impairment, limited mouth opening, xerostomia, mucositis, fatigue, reduced stamina, weight loss, and nutritional concerns. These factors are central to post-op assessment and frequently require clarification when patients describe symptoms vaguely.

Comorbid conditions. This facet captures additional health conditions that may influence safety, symptom interpretation, or the appropriateness of responses. Examples include chronic diseases requiring concurrent management or other clinically relevant conditions disclosed by the patient.

B.2 Care-Related Clinical Context Facets

This group of facets captures care-relevant clinical and psychosocial contexts that influence how a response should be framed, prioritized, or interpreted. Importantly, these facets are *not* used for psychological risk screening and do not affect whether the system continues the dialogue.

All facet-based clarification described here is performed *only after* an utterance has passed the initial risk screening stage.

Psychosocial and emotional context. This facet reflects non-crisis psychosocial concerns commonly expressed during post-op care, such as anxiety about recovery, emotional burden, fear of recurrence, or uncertainty about future functioning. These expressions inform response tone and supportive emphasis but are explicitly distinguished from emotional distress or suicidal ideation, which are handled exclusively by the upstream risk detection module.

Care and symptom management context. This facet captures care-related concerns, including wound management, rehabilitation progress, symptom monitoring, medication adherence, and follow-up logistics. Clarifying this context helps tailor responses to the patients practical care needs without implying urgency or safety risk.

Social and occupational context. This facet includes concerns related to return to work, family roles, financial stress, or access to social support. These factors often shape patient questions and expectations, but function solely as contextual information for clarification and response generation.

All facets in this group serve to reduce clinical ambiguity and improve response relevance. They do not trigger safety actions, override prior risk assessments, or suppress clarification once the dialogue has entered the safe information-seeking state.

B.3 Information and Support Intent Facets

Information and support intent facets describe what the patient seeks to understand or address beyond their immediate clinical state. These facets guide clarification by identifying the informational or supportive focus of the dialogue.

Information and education needs. This facet captures requests for medical information or education, such as prognosis, treatment rationale, staging, side effects, vaccination eligibility, or post-treatment lifestyle adjustments. Clarifying this intent helps ensure that responses align with the patients actual informational goals.

Psychosocial support preference. This facet follows the FOCUSP framework and captures the

psychosocial domain that the patient wishes to prioritize. The six domains include family involvement, optimistic attitude, coping effectiveness, uncertainty reduction, symptom management, and physical enhancement. This facet primarily guides response tone and support orientation rather than factual content selection.

B.4 Annotation and Usage

Each QA item in the dataset was annotated with one or more relevant facets based on its semantic content. Annotations were determined by assessing which clinically relevant factors were implicated by the patient's question, following the same reasoning process clinicians use during post-op inquiry. The annotation process employed initial thematic matching followed by expert review to ensure clinical validity.

These facet annotations serve two purposes. First, they provide a consistent and clinically grounded annotation framework for the dataset. Second, they define the action space for clarification during dialogue, specifying which facets may be queried to reduce ambiguity and improve retrieval accuracy. All data were fully de-identified, reviewed by clinical experts, and collected under institutional ethical approval.