Eirene: Addressing End-of-Life Counseling as Constrained Optimization via Multi-Agent Coopetition

Anonymous ACL submission

Abstract

End-of-life counseling for terminal cancer patients constitutes a psychologically intensive clinical interaction that must simultaneously satisfy complex requirements, including the accuracy, safety, and empathetic delivery of medical and advance care planning information. While large language models (LLM) have recently been explored as counseling support tools, single LLM–based response generation struggles to coordinate these competing demands within a single response. To address this challenge, this study proposes Eirene, a multi-agent control architecture that separates goal-oriented response generation from constraint verification. Ablation study results show that Eirene achieves consistent performance improvements across all evaluation metrics. Moreover, analysis of the feedback loop reveals a trade-off between response quality and system cost, indicating that a limited number of iterations functions as an effective balance point. These findings suggest that a multi-agent control framework, which reframes safety and ethical constraints as post-generation judgment and intervention problems, can ensure both stability and conversational quality in high-risk clinical counseling settings.

1 Introduction

End-of-life(EoL) counseling is a core clinical process that supports terminal patients' treatment decisions and the establishment of life values, and Advance Care Planning(ACP) is a primary means of institutionalizing this process. However, in real clinical settings, EoL counseling is difficult to conduct consistently due to time constraints and high emotional burden, and ACP is also often not sufficiently completed (Sailian et al., 2024; Scholz et al., 2020). To compensate for these limitations, the need for automated systems that can structure and support the counseling process has been raised (Cheung et al., 2020; Alanazi et al., 2024). However, each utterance in EoL counseling must simultaneously satisfy medical factuality, ethical and safety requirements, and empathetic responses, and these requirements tend to conflict with one another at the same utterance level (Lechner et al., 2023). Existing single Large Language Model (LLM)-based systems exhibit structural failures in such situations (Lechner et al., 2023). This stems from a structural limitation whereby attempts to simultaneously optimize counseling goal attainment and safety and ethical constraints within a single generation process make it impossible to explicitly control the interactions among these requirements.

Accordingly, this study defines EoL counseling not as a simple dialogue generation problem, but as a problem of controlling multiple simultaneously conflicting objectives, and proposes a multi-agent system, Eirene, that functionally separates goal attainment and constraint verification. Named after the Greek goddess of peace, Eirene reflects the objective of this study to help terminal patients and their families peacefully accept the end of life. The proposed system consists of a Counselor agent that conducts counseling, a Critic agent that verifies semantic constraints, and an Escalation agent that detects high-risk signals and transitions to Human-in-the-loop intervention. This configuration goes beyond a simple cooperative relationship among agents, exhibiting the characteristics of a coopetition relationship in which agents may also engage in critique as needed (Tran et al., 2025). Through this design, the system continuously conducts counseling while maintaining an explicitly controlled structure that ensures each utterance does not violate ethical or semantic constraints.

The main contributions of this paper are as follows. (1) We define the problem of EoL patient counseling as a constrained optimization problem and propose a coopetition structure among a Counselor Agent for achieving counseling goals, a Critic Agent for verifying semantic constraints,

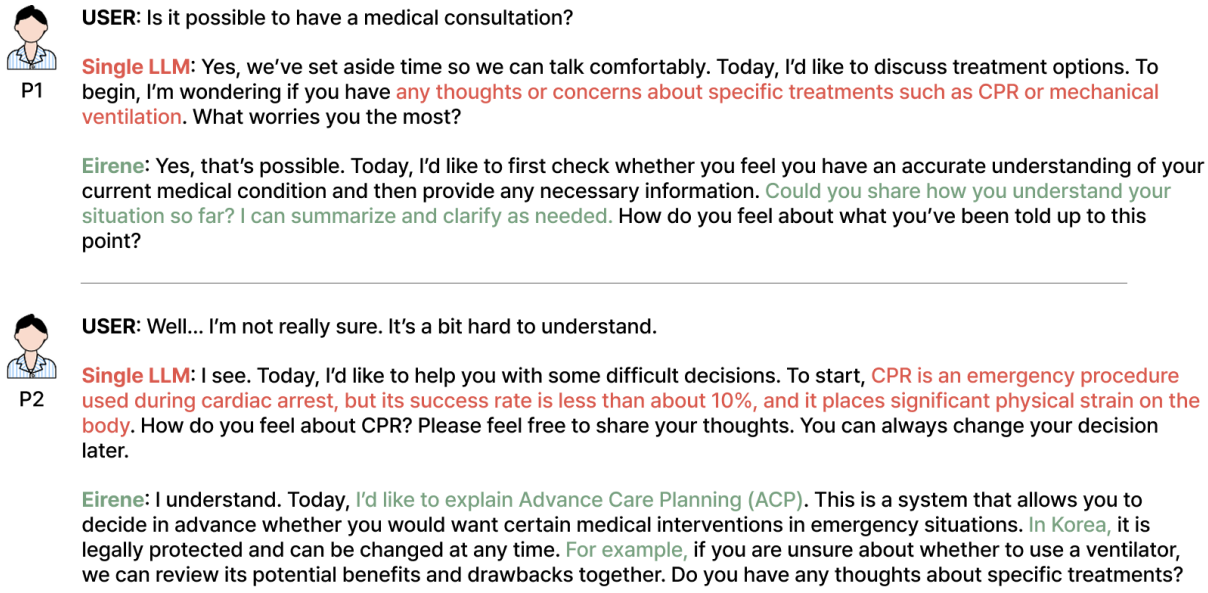

Figure 1: Qualitative analysis of model responses. Unlike the single LLM that uses difficult terminology, the proposed Eirene demonstrates empathy and explains concepts using easy-to-understand examples.

and an Escalation Agent for checking ethical constraints, thereby systematically structuring the conflicting relationships among the various requirements. (2) We evaluate the performance of the proposed multi-agent approach in comparison with a single LLM and demonstrate statistically significant performance improvements across key metrics. In particular, through an ablation study, we demonstrate that each agent module plays a critical role in ensuring response safety. (3) Considering the sociocultural clinical context of Korea, we established a six-stage counseling workflow through co-design with medical professionals, and examined the effectiveness of the proposed methodology through virtual patient simulations reflecting the characteristics of the Korean elderly population. These results demonstrate that the proposed system can function as an effective counseling support tool even in high-risk clinical environments, and suggest the potential to bridge technology and palliative care practice. Throughout the design process, we maintained continuous clinical consultation with palliative care clinicians in Korea to ensure that the system's assumptions, dialogue tone, and safety boundaries remain aligned with real-world end-of-life counseling practice.

2 Related Work

2.1 LLM in End-of-Life care

End-of-life counseling aims to align patients' values with their treatment preferences. While prior studies have proposed theoretical foundations (Bernacki et al., 2014, 2019; Uemura et al., 2024) and web-based tools (Sudore et al., 2017b), their reliance on predefined rules limits flexibility in responding to emotions and context. Although LLMs have been introduced to address this gap, Single LLM systems—where generation and safety verification share a single pathway—face a structural dilemma: prioritizing safety often leads to defensive, formal responses that reduce empathy or impede dynamic control (Lechner et al., 2023). This dilemma arises from jointly optimizing the conflicting objectives of Helpfulness and Safety (Sabour et al., 2023). Moreover, safety-oriented fine-tuning incurs an 'alignment tax' that degrades dialogue quality (Long Ouyang and Lowe, 2022), resulting in evasive behavior (Bai et al., 2022) or over-refusal of benign requests (Röttger et al., 2024). These limitations underscore the need for a structural approach beyond a single generation pathway.

2.2 Self-Correction Multi-Agent Frameworks

Self-critic mechanisms have been proposed to improve LLM performance by enabling models to review and revise their own outputs (Shinn et al., 2023; Kim et al., 2023). Among these, Direct Prompting (DP) induces critique through prompts (Li and Zhao, 2025), but its reliance on intrinsic intuition limits effective self-correction without an external gold signal (Zheng et al., 2025; Huang et al., 2023). In EoL counseling, where answers are inherently ambiguous, single models therefore risk reit-

erating biased responses or performing only shallow verification. To mitigate this, multi-agent systems (MAS) have been introduced (Li et al., 2023). By separating generation and verification roles and enabling collaborative interaction, MAS extends intuitive single-model generation into deeper logical reasoning (Du et al., 2023). Although recent work in mental health shows enhanced empathetic responses (Xu et al., 2025), EoL counseling additionally demands strict medical accuracy and clinical safety (Akdeniz et al., 2021; Sudore et al., 2017a). Existing models (Qiu et al., 2024), optimized for general counseling, thus remain structurally limited in preventing subtle risks or clinical errors in life-critical settings.

2.3 LLM as a judge

With recent advances in reasoning capabilities, the use of LLMs as evaluators, referred to as 'LLM-as-a-judge,' has been expanding in the field of natural language processing. The contextual embeddings of LLMs effectively capture subtle semantic differences and contextual nuances in human language (Kjell et al., 2022, 2024). In pairwise comparison-based evaluations using GPT-4 as a judge on MT-Bench, a high agreement rate of over 80% with human experts was reported (Zheng et al., 2023). Such performance has also been demonstrated in clinical domains that require a high level of expertise (Seo et al., 2024). In particular, although medical expert evaluations themselves exhibit high reliability, a tendency to systematically underestimate certain types of errors has been reported, raising the need for complementary use of automated or model-based evaluators. In the 'AI Hospital' study conducted in a multi-agent environment (Fan et al., 2025), LLM-based evaluations of treatment plans showed a high correlation with physicians' judgments ($r = 0.86$). Moreover, in the mental health domain, LLMs demonstrated meaningful performance in evaluating levels of empathy toward clients (Gabriel et al., 2024) and in identifying key utterances that serve as the basis for judgments (Blanco-Cuaresma, 2024), and achieved substantial agreement with experts (Krippendorff's $\alpha = 0.78$) in tasks involving the identification of high-risk crisis situations (Thomas et al., 2025). These prior studies suggest that LLM-based evaluation has potential applicability even in complex and sensitive clinical environments.

3 Preliminaries

Before describing Eirene, the multi-agent system proposed in this study, we clarify the scope of the problem that the system aims to address. EoL cancer patient counseling is a complex domain that must simultaneously satisfy distinct conversational stages and strict clinical constraints. Accordingly, this section defines (1) the six-stage counseling workflow and (2) the objectives and constraints that constitute the targets of optimization, which together form the foundation of the system design.

3.1 Counseling Stage

Eirene's counseling flow is composed of six stages: Opening, Disease Awareness, ACP Awareness, Empathy, Treatment Preferences, and Closing.

In the Opening stage, the counselor establishes rapport and agrees on the discussion topics and their priorities, and then proceeds to the next stage after confirming that the patient understands the goals. In the Disease Awareness stage, the counselor confirms the patient's understanding of prognosis, symptoms, and test results while correcting misunderstandings in real time. In the ACP Awareness stage, the counselor explains only the key points of advance directives and advance care planning and provides guidance on relevant laws and procedures. The Empathy stage supports emotional acceptance by exploring personal values. In the Treatment Preferences stage, major interventions such as CPR, mechanical ventilation, dialysis, ICU admission, vasopressors, chemotherapy, ECMO, and blood transfusion are addressed sequentially to reduce redundant explanations and information overload. Finally, in the Closing stage, the counselor concludes by summarizing the patient's decisions and questions.

To maintain counseling professionalism, the system adopts a staged linear structure. Transition indicators for each stage are based on linguistic signals defined by the hospice team based on clinical experience, including the patient's explicit consent, confirmation of corrected misunderstandings regarding prognosis, and voluntary articulation of personal values. This structure ensures that the counseling can proceed flexibly in response to the patient's reactions without deviating from clinical guidelines.

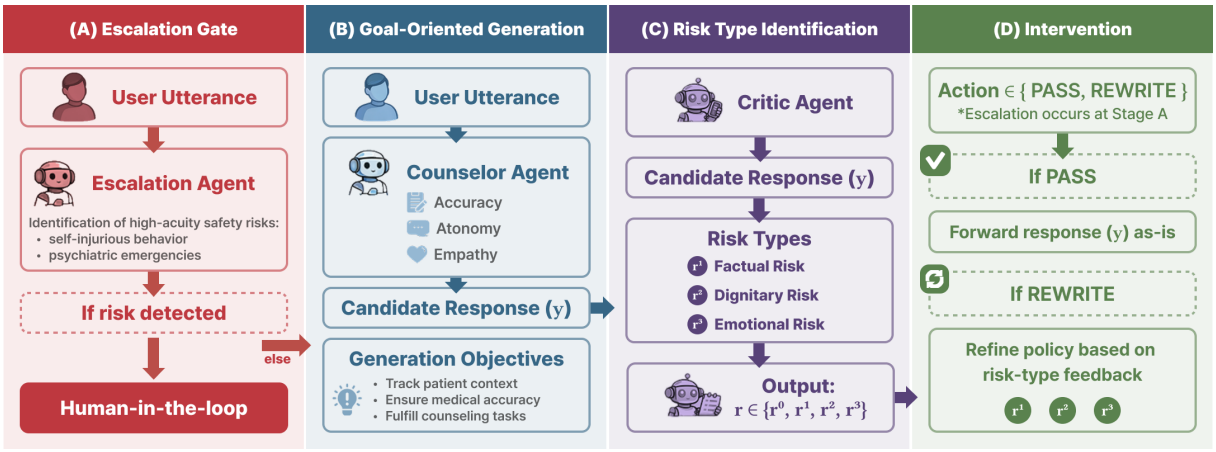

Figure 2: Overall architecture. When the user initiates an utterance, the system (a) checks whether escalation is required, (b) generates candidate responses, (c) verifies the responses based on risk types, and (d) outputs or regenerates the response.

3.2 Objectives & Constraints

EoL counseling for terminal cancer patients is a complex problem that goes beyond conducting dialogue and requires the simultaneous satisfaction of multiple objectives and constraints. This study defines the objectives and constraints of counseling as follows.

Objectives. The counseling system performs three core tasks: (1) contextual understanding through real-time tracking of the patient's emotional state and level of information comprehension, (2) provision of medical information grounded in clinical knowledge, and (3) task execution and utterance generation in accordance with predefined counseling stages.

Semantic Constraints. In the process of achieving these objectives, the system adheres to the following constraints to protect patient trust. First, it maintains medical accuracy by excluding incorrect prognostic information or unfounded claims. Second, it preserves patient autonomy by preventing coercive expressions and focusing on the role of an information provider. Third, it practices respect for emotions by not overlooking the patient's psychological distress.

Ethical Constraints. Finally, when the patient exhibits risk signals, the system must immediately suspend automated responses and apply a Human-in-the-loop mechanism. This serves as a safety mechanism that enables the AI system to recognize the limits of clinical judgment and appropriately escalate situations requiring expert intervention.

In conclusion, the EoL counseling problem is defined as a constrained optimization problem that simultaneously pursues functional objectives—contextual understanding of dialogue, provision of medical information, and stage-wise task accomplishment—while satisfying semantic constraints such as preservation of autonomy, medical accuracy, and respect for emotions, as well as ethical constraints involving safe escalation in crisis situations.

4 Methods

4.1 Overall architecture

This study assigns responsibility for each problem to independent agents in order to simultaneously satisfy the system objectives and two types of constraints. When a user utterance is received, the Escalation agent first checks whether the ethical constraints are satisfied. If Human-in-the-loop intervention is deemed necessary, a predefined message is output and the conversation is immediately terminated. If the ethical constraints are satisfied, the Counselor agent analyzes the user utterance and generates an appropriate response to achieve the objectives of the current counseling stage. This response is then passed to the Critic agent, which checks whether the response violates semantic constraints and provides feedback. If the response is judged to require revision due to constraint violations, the Counselor receives the feedback and regenerates the response. Through this feedback loop, a final response that satisfies the constraints while simultaneously addressing the objectives is derived. The pseudo code is presented in Algorithm 1.

Algorithm 1: Main workflow of the multi-agent counseling system

Input : $userInput$
Output : $response$

// Ethical Constraints
$isRisky \leftarrow EscalationAgent.check(userInput, memory)$
if $isRisky$ **then**
    return "Human-in-the-loop"
$feedback \leftarrow$ ""
for $t = 1$ **to** N **do**
    // Objectives
    $userSession \leftarrow CounselorAgent.plan(userInput, userSession)$
    $medicalInfo \leftarrow CounselorAgent.useTool(userInput, memory)$
    $response \leftarrow$
     $CounselorAgent.generate(userInput, userSession, memory, medicalInfo, feedback)$
    // Semantic Constraints
    $riskType, feedback \leftarrow CriticAgent.evaluate(userInput, response, memory)$
    if $riskType = 0$ **then**
        break
return $response$

4.2 Multi-agent system

The Escalation Agent receives the user utterance x and a memory m containing the conversation history as input, and performs binary classification of acute risk $e \in \{0, 1\}$. When $e = 1$, the system prioritizes ethical constraints over response generation and transitions to a safe output pathway involving human expert intervention. Given the characteristics of EoL counseling, expressions related to acceptance of or preparation for a natural death are designed not to be overclassified as risk signals.

$$e = Agent(x, m \mid \theta_{escalation})$$

The Counselor agent is the generative entity that conducts the actual counseling. It receives the user utterance x, the current counseling stage s, the memory m, the tools τ available to the agent, and the feedback fb from the previous step, and generates a response y. Beyond simple response generation, the Counselor agent performs planning by analyzing the user's previous counseling stage s and the current input x. Through the planning step, it analyzes which counseling stage is currently required and modifies the instructions within the prompt according to the decision, a design intended to ensure that the tasks required at each counseling stage are achieved.

The core components supporting the inference of the Counselor agent are as follows. The memory m consists of Short-term Memory (STM), which stores and summarizes conversation history to accurately capture conversational context (Zhang and Luo, 2024), and Long-term Memory (LTM), which stores characteristic information specific to the patient(Chhikara et al., 2025). The tool τ enables Retrieval Augmented Generation (RAG) to provide medically accurate and evidence-based information. For this purpose, a vector database composed of 135 cancer-related documents from the National Cancer Institute was constructed (National Cancer Institute, 2025). The feedback fb is provided by the Critic agent introduced below when the Counselor agent violates semantic constraints. The initial feedback fb_0 is provided as an empty string.

$$y_t = Agent(x, s, m, \tau, fb_{t-1} \mid \theta_{counselor})$$

The Critic agent receives the user utterance x, the Counselor agent's response y, and the memory m as input, and outputs a risk type r and feedback fb. Specifically, $r \in \{r^0, r^1, r^2, r^3\}$ indicates which semantic constraint has been violated, and the risk types are defined as follows. r^1 denotes medical

errors or unfounded prognostic assertions, r^2 denotes violations of patient autonomy or the use of coercive expressions, and r^3 denotes neglect of emotional considerations, while r^0 is output when no constraints are violated. When $r_t \neq r^0$, the Counselor regenerates the response by reflecting the feedback fb. The use of a finite set of risk types is intended to clearly distinguish semantic constraint violations and enable repeatable evaluation. The feedback fb contains instructions on how the Counselor agent should revise the response based on the violated risk type.

$$(r_t, fb_t) = Agent(x, y_t, m \mid \theta_{critic})$$

Response generation by the Counselor agent and feedback generation by the Critic agent are repeated until the risk type r^0 is produced. To prevent infinite loops and ensure system stability, the number of iterations is limited to N. In summary, the proposed system introduces a Counselor agent to achieve objectives and a Critic agent and an Escalation agent to verify constraints, thereby addressing the EoL counseling problem. Implementation details are provided in the Appendix.

5 Experiments

5.1 Metrics

To evaluate the counseling quality, we defined a set of metrics through a co-design process with medical professionals. These metrics, designed to verify the alignment with our objectives and constraints, were rated on a 1–10 Likert scale: Empathy, Appropriateness, Professionalism, Accuracy, Specificity, User-centeredness, Ethicality. Detailed explanations are provided in the Appendix.

5.2 Dataset

Due to the nature of EoL counseling, acquiring real patient dialogue data raises ethical concerns and makes it difficult to construct a sufficiently large dataset. Therefore, to ensure safety and reproducibility, an LLM-based patient simulator was used instead of real patient interactions, modeling ten virtual patients with diverse sociocultural and emotional characteristics. The simulator was designed to reflect the sociocultural context of Korea. Personas were geographically distributed to include linguistic and dialectal differences among major cities (e.g., Seoul, Busan, Jeonju). Occupational and socioeconomic diversity was configured to reflect the demographic composition of the elderly population in Korea. Each persona embodied culture-specific attributes such as Confucian family values, emotion suppression influenced by social norms, religious differences in perceptions of death, and concerns about burdening family members. Language-specific elements were also incorporated, including the use of honorific and informal speech, hierarchical family expressions, and dialectal nuances. This design aimed to realistically capture the cultural identity and communication styles of Korean terminal cancer patients (Rim et al., 2021; Tarbi et al., 2021). Each dialogue session consisted of 10–20 turns, with each turn composed of a patient utterance and an Eirene response pair. This design reflected clinical safety standards observed in real clinical environments, ensuring emotional acceptance and information comprehension within the simulation setting.

5.3 Experiment Design

To verify the performance and efficiency of Eirene, this study designed experiments by setting the following two key research questions (RQs).

RQ1. *Does Eirene improve clinical response quality in EoL counseling?* To verify whether Eirene achieves the objectives and constraints, the seven evaluation metrics defined above are assessed using a 1–10 Likert scale. Scores are measured using an LLM-as-a-judge evaluation method, with GPT-4 employed as the evaluation model (Gu et al., 2024). For additional comparison, an ablation study is conducted on the agents used in the system to examine the contribution of each component. Specifically, comparative evaluations under identical conditions are performed for (1) removal of the Critic agent, (2) removal of the Escalation agent, (3) removal of both the Critic and Escalation agents, and (4) a Single LLM configuration. Because the Counselor agent is the entity that generates the final response, direct ablation was not feasible; instead, the contribution of the Counselor agent is indirectly evaluated using a single LLM with only a prompt.

RQ2. *What level of intervention by the Critic Agent is optimal?* We analyze the impact of the maximum number of iterations N, set to prevent infinite stagnation of the feedback loop, on system performance and resource efficiency. If N is too small, response correction may be insufficient, whereas if it is excessively large, latency and token costs may increase sharply. Accordingly, response quality, average and P95 latency, and token

Table 1: Scores of the seven evaluation metrics for the ablation experiments of the main agents

Models	Empathy	Approp.	Profess.	Accuracy	Specif.	User-cent.	Ethicality
Eirene	**7.9**	**8.1**	**9.0**	**8.9**	**7.3**	**8.2**	**9.4**
Single LLM	7.6	6.8	8.3	8.3	6.3	7.8	8.5
w/o Critic & Escalation	7.8	7.7	8.6	8.5	7.2	7.8	8.9
w/o Critic	**7.9**	7.6	8.4	8.6	6.8	7.8	9.1
w/o Escalation	7.7	7.5	8.8	8.6	7.1	7.8	9.0

usage are comprehensively analyzed for settings of $N = 1, 3, 5, 7, 10$. Response quality is evaluated using the same metrics and evaluation methods defined above. The P95 metric indicates that 95% of all responses fall within the corresponding value, representing an upper bound of performance that the system can stably guarantee in real-time counseling environments while excluding the top 5% of extreme values.

6 Result

6.1 Ablation Study

This section quantitatively compares and analyzes the performance of the Eirene multi-agent system across seven evaluation metrics. All experiments were conducted using identical evaluation settings and inputs, and result interpretation was based on the average scores reported in the table.

Eirene achieved the highest scores across all evaluation metrics. Compared with the Single LLM, Eirene demonstrated consistent performance improvements on all metrics. Appropriateness increased by approximately 19%, and specificity increased by approximately 16%. Accuracy also improved by approximately 7%. These results indicate that the multi-agent architecture generated higher-quality responses overall than the single LLM-based approach.

In the configuration without the Critic agent, appropriateness decreased by approximately 6% compared to Eirene, and specificity decreased by approximately 7%. This indicates that the removal of the Critic agent leads to score degradation in certain metrics. In the configuration without the Escalation agent, ethicality decreased by approximately 4%, and user-centeredness decreased by approximately 5%. In the configuration where both the Critic agent and the Escalation agent were removed, slight performance degradation compared to Eirene was observed across all metrics. However, this configuration still maintained overall higher scores than the Single LLM, suggesting that the Counselor agent alone is capable of generating more appropriate responses than a Single LLM.

In summary, each agent makes a meaningful contribution to overall response quality, and the best performance was observed in the Eirene configuration that incorporates all agents. Notably, the contributions of individual agents do not simply accumulate independently; rather, additional performance gains emerge when all three agents are combined. In fact, for six out of the seven evaluation metrics, Eirene achieved higher scores than all ablation settings in which any single agent was removed, and this level of performance was not reproduced by any partial combination. Considering that EoL counseling is a task requiring the simultaneous satisfaction of complex objectives and constraints, these findings suggest that a multi-agent collaborative architecture incorporating role specialization and mutual oversight among agents is an effective approach.

6.2 Optimal Number of Iterations

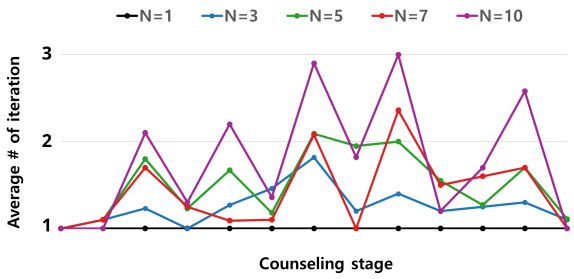

Figure 3: Average iterations per session across 13 counseling stages. The average feedback frequency is less than 3 for all stages, including sub-stages.

A comprehensive analysis of the experimental results with respect to the upper bound N on repeated feedback reveals a clear trade-off between response quality and system cost. Under the $N = 1$ setting, the average overall score was the lowest, while the average token usage and average latency remained at minimal levels, making it the most lightweight configuration in terms of efficiency.

Table 2: Analysis of response quality and cost with respect to N

N Limit	Overall Score(↑)	Avg. Tokens(↓)	P95 Tokens(↓)	Avg. Latency(s, ↓)	P95 Latency(s, ↓)	Avg. Iteration
1	7.99	**3568.7**	**5578.9**	**8.33**	**11.59**	1.0
3	8.24	4346.3	7715.1	10.40	18.01	1.28
5	7.91	5201.5	13510.4	11.45	26.40	1.57
7	**8.33**	4733.3	13114.6	11.30	31.11	1.46
10	7.87	5923.8	26427.0	13.56	46.36	1.87

When increased to $N = 3$, the overall score rose by 0.25 points to 8.24, while the average token usage and latency increased to 4346.3 and 10.4 seconds, respectively. This indicates that a limited number of Critic agent feedback iterations contributes to response quality improvement while keeping cost increases relatively moderate. In contrast, at $N = 5$, although the average token usage increased to 5201.5 and the P95 token count reached 13510.4, the overall score decreased to 7.91. This suggests that increasing the number of feedback iterations does not directly translate into quality improvement and may instead negatively affect response quality due to instability in the revision process or excessive repetition. At $N = 7$, the Overall score reached its highest value of 8.33, but the P95 latency increased to 31.11 seconds. When $N = 10$, not only did the average token usage increase markedly to 5923.8, but the P95 latency also increased to 46.36, representing a fourfold increase compared to $N = 1$.

As shown in Figure 3, which illustrates the actual number of feedback iterations per session, the average number of iterations remained below three even as N increased. This indicates that the Critic feedback process converges in most cases before reaching the upper bound N, and that actual repetitions remain limited even when the upper bound is set high. Taken together, $N = 3$ serves as a balanced point that provides a clear quality improvement over $N = 1$ while effectively suppressing the excessive increases in token usage and latency observed at $N \geq 5$. These results demonstrate that rather than increasing the intensity of Critic feedback without bound, setting an appropriate upper limit that accounts for actual convergence behavior is critical for maintaining the trade-off between response quality and system cost.

7 Conclusion

This study defines EoL counseling as a constraint-aware optimization problem that simultaneously seeks to achieve the objective of generating appropriate responses and the constraints of not violating safety and ethical requirements. To address this problem, we designed agents capable of handling each component and proposed a multi-agent architecture in which agents engage in coopetition. Experimental results show that applying the multi-agent architecture outperformed all comparison baselines across seven evaluation metrics, demonstrating its contribution to improving clinical response quality.

Ethical Considerations

The Eirene proposed in this study is intended to support clinical constraints and assist medical professionals' decision-making, rather than to replace expert-led counseling. It was developed through a clinical co-design process with medical staff at a tertiary hospital, ensuring alignment with established medical ethics and clinical protocols. To address the risks associated with large language models, the system adopts a multi-agent architecture in which a Critic Agent performs real-time validation of generated responses, while an Escalation Agent suspends automated interactions and prompts professional intervention upon detecting crisis signals. To protect vulnerable terminal patients from ethical risks, the framework was evaluated exclusively using virtual patient simulations grounded in Korean socio-cultural clinical scenarios, thereby safeguarding sensitive health information while maintaining experimental reproducibility.

Limitations

This study limits semantic constraints in end-of-life cancer patient counseling to three types: medical or factual errors, violations of patient autonomy, and emotional harm. This scope was set to analyze the operating principles and initial effectiveness of the proposed mechanism, and it does not encompass all risks present in real clinical environments. Discussions with medical professionals also indicated that clinical settings involve more complex situations, suggesting that future work should strengthen the

clinical applicability of intervention strategies by expanding risk types based on clinical expertise.

In addition, this study employs a structure in which responses are iteratively improved through repeated interactions between the Counselor Agent and the Critic Agent until risk types are eliminated. However, it does not address how to handle cases in which constraints are not satisfied even after N iterations. Future research should examine strategies such as escalation to human experts and adaptive termination policies based on risk types.

Finally, this study relies on LLM-based automatic evaluation and does not include assessments by human experts. This choice was made because the focus was on verifying the effectiveness of the control structure for coordinating conflicting objectives and constraints, rather than on absolute evaluation of counseling quality. Future work plans to further validate the impact on actual clinical counseling quality through evaluations conducted by human experts.

References

Melahat Akdeniz, Bülent Yardımcı, and Ethem Kavukcu. 2021. Ethical considerations at the end-of-life care. *SAGE open medicine*, 9:20503121211000918.

Majed Awad Alanazi, Marwa Mamdouh Shaban, Osama Mohamed Elsayed Ramadan, Mohammed Elsayed Zaky, Huda Hamdy Mohammed, Fatma Gomaa Mohamed Amer, and Mostafa Shaban. 2024. Navigating end-of-life decision-making in nursing: a systematic review of ethical challenges and palliative care practices. *BMC nursing*, 23(1):467.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Rachelle Bernacki, Joanna Paladino, Bridget A Neville, Mathilde Hutchings, Jane Kavanagh, Olaf P Geerse, Joshua Lakin, Justin J Sanders, Kate Miller, Stuart Lipsitz, and 1 others. 2019. Effect of the serious illness care program in outpatient oncology: a cluster randomized clinical trial. *JAMA internal medicine*, 179(6):751–759.

Rachelle E Bernacki, Susan D Block, and 1 others. 2014. Communication about serious illness care goals: a review and synthesis of best practices. *JAMA internal medicine*, 174(12):1994–2003.

Sergi Blanco-Cuaresma. 2024. Psychological assessments with large language models: A privacy-focused and cost-effective approach. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 203–210, St. Julians, Malta. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Johnny TK Cheung, Doreen Au, Anthony HF Ip, Jenny Chan, Kenway Ng, Lok Cheung, Jacqueline Yuen, Elsie Hui, Jenny Lee, Raymond Lo, and 1 others. 2020. Barriers to advance care planning: a qualitative study of seriously ill chinese patients and their families. *BMC Palliative Care*, 19(1):80.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *Preprint*, arXiv:2504.19413.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–21. ACM.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677.

Oscar NE Kjell, Katarina Kjell, and H Andrew Schwartz. 2024. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333:115667.

Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, 12(1):3918.

Fabian Lechner, Allison Lahnala, Charles Welch, and Lucie Flek. 2023. Challenges of gpt-3-based conversational agents for healthcare. *arXiv preprint arXiv:2308.14641*.

Baoxue Li and Chunhui Zhao. 2025. Self-reflection enhances large language models towards substantial academic response. *npj Artificial Intelligence*, 1(1):42.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Xu Jiang Diogo Almeida Carroll Wainwright Pamela Mishkin Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell Peter Welinder Paul F Christiano Jan Leike Long Ouyang, Jeffrey Wu and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pages 27730–27744. Curran Associates, Inc.

National Cancer Institute. 2025. Cancer topics. https://www.cancer.gov. Accessed: 2025-07-23.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Kim Yeh Rim, Eun-ju Lee, and Hwang MinJi. 2021. The comprehension of indirect speech acts with aging: Focusing on young, middle-aged and elderly group. *Journal of Special Education*, 37(3):143–161.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2023. A chatbot for mental health support: exploring the impact of emohaa on reducing mental distress in china. *Frontiers in digital health*, 5:1133987.

Silva Dakessian Sailian, Yakubu Salifu, and Nancy Preston. 2024. Dignity enhanced through faith & family support in palliative care: a qualitative study. *BMC Palliative Care*, 23(1):142.

Brett Scholz, Liza Goncharov, Nathan Emmerich, Vinh N Lu, Michael Chapman, Shannon J Clark, Tracey Wilson, Diana Slade, and Imogen Mitchell. 2020. Clinicians' accounts of communication with patients in end-of-life care contexts: A systematic review. *Patient education and counseling*, 103(10):1913–1921.

Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, and Edward Choi. 2024. Evaluation framework of large language models in medical documentation: Development and usability study. *Journal of Medical Internet Research*, 26:e58329.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Rebecca Sudore, Hillary Lum, John You, Laura Hanson, Diane Meier, Steven Pantilat, Daniel D Matlock, Judith Rietjens, Ida J Korfage, Christine Ritchie, and 1 others. 2017a. Defining advance care planning for adults: a consensus definition from a multidisciplinary delphi panel (s740). *Journal of Pain and Symptom Management*, 53(2):431–432.

Rebecca L Sudore, John Boscardin, Mariko A Feuz, Ryan D McMahan, Mary T Katen, and Deborah E Barnes. 2017b. Effect of the prepare website vs an easy-to-read advance directive on advance care planning documentation and engagement among veterans: a randomized clinical trial. *JAMA internal medicine*, 177(8):1102–1109.

Elise C Tarbi, Robert Gramling, Christine Bradway, and Salimah H Meghani. 2021. "if it's the time, it's the time": Existential communication in naturally-occurring palliative care conversations with individuals with advanced cancer, their families, and clinicians. *Patient education and counseling*, 104(12):2963–2968.

Julia Thomas, Zohar Elyoseph, Lars Kuchinke, and Gunther Meinlschmidt. 2025. Large language model performance versus human expert ratings in automated suicide risk assessment. *Scientific Reports*, 15(1):39231.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Takeshi Uemura, Kaori Ito, Misuzu Yuasa, Youkie Shiozawa, Hirono Ishikawa, Shunichi Nakagawa, Eriko Onishi, and Kei Ouchi. 2024. Enduring positive impact of a virtual communication skills workshop of vitaltalk pedagogy in a non-us setting. *American Journal of Hospice and Palliative Medicine®*, 41(4):424–430.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, and 1 others. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

XiuYu Zhang and Zening Luo. 2024. Advancing conversational psychotherapy: Integrating privacy, dual-memory, and domain expertise with large language models. *arXiv preprint arXiv:2412.02987*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. 2025. Critic-cot: Boosting the reasoning abilities of large language model via chain-of-thought critic. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1768–1806.

A Implementation Details

Eirene was implemented using Python 3.12 (PSF License) and FastAPI 0.116.1 (MIT License) in an Ubuntu 24.04.1 LTS environment. Facebook AI Similarity Search (MIT License) was used as the vector DB, and HyperClovaX HCX-005, which is specialized for understanding the Korean language and culture, was employed for language modeling (Yoo et al., 2024). Sentence embeddings were generated using bge-m3 (Chen et al., 2024). The contents of the STM constituting the memory are periodically compressed into 200-character summaries every 10 turns (Zhang and Luo, 2024), while the LTM continuously accumulates and updates personalized information following the mem0-style memory operation framework (Apache 2.0 License) (Chhikara et al., 2025). These mechanisms promote user self-disclosure and induce higher engagement during the counseling process (Jo et al., 2024). All other auxiliary libraries and dependencies were also confirmed to be used in accordance with their authorized licenses.

B Evaluation Metrics

The detailed descriptions of the seven metrics co-designed with medical professionals are as follows:

- Empathy: The degree to which the system accurately recognizes the patient's emotional state, respects dignity, and responds warmly.
- Appropriateness: The degree to which the dialogue proceeds naturally according to the seven counseling stages.
- Professionalism: The degree to which medical knowledge and terminology are appropriately simplified and presented at a level understandable to the patient.
- Accuracy: The degree to which medical information and treatment guidelines are factually correct and based on clinical evidence.
- Specificity: The degree to which tailored details or examples are provided in accordance with the patient's questions and situation.
- User-Centeredness: The degree to which responses are adjusted by considering the patient's level of understanding and emotional state, while ensuring the patient's opportunity to speak.
- Ethicality: The degree to which patient safety is ensured through protection of personal information, maintenance of a non-judgmental attitude, and appropriate responses in high-risk situations.

C Declaration of Generative AI

During the preparation of this work, the authors used ChatGPT in order to translate the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.