

AI Chatbots as Professional Service Agents: Developing a Professional Identity

Wenwen Li¹, Kangwei Shi^{2,3,4*}, Yidong Chai^{2,3,4}

¹School of Management, Fudan University, Shanghai, China

²Philosophy and Social Sciences Laboratory of Data Science and Smart Society Governance, Ministry of Education, Hefei, China

³Key Laboratory of Philosophy and Social Sciences for Cyberspace Behaviour and Management, Hefei, China

⁴School of Management, Hefei University of Technology, Hefei, China
liwwen@fudan.edu.cn, shikw@mail.hfut.edu.cn, chaidy@hfut.edu.cn

Abstract

With the rapid expansion of large language model (LLM) applications, there is an emerging shift in the role of LLM-based AI chatbots from serving as general inquiry tools to acting as professional service agents. However, current studies often overlook a critical aspect of professional service agents: the act of communicating in a manner consistent with their professional identities. This is of particular importance in the healthcare sector, where effective communication with patients is essential for achieving professional goals, such as promoting patient well-being by encouraging healthy behaviors. To bridge this gap, we propose LAPI (LLM-based Agent with a Professional Identity), a novel framework for designing professional service agent tailored for medical question-and-answer (Q&A) services, ensuring alignment with a specific professional identity. Our method includes a theory-guided task planning process that decomposes complex professional tasks into manageable sub-tasks aligned with professional objectives and a pragmatic entropy method designed to generate professional and ethical responses with low uncertainty. Experiments on various LLMs show that the proposed approach outperforms baseline methods, including few-shot prompting, chain-of-thought prompting, across key metrics such as fluency, naturalness, empathy, patient-centricity, and ROUGE-L scores. Additionally, the ablation study underscores the contribution of each component to the overall effectiveness of the approach.

1 Introduction

AI chatbots have gained great popularity worldwide to serve as professional assistants and provide services across various sectors (Dam et al., 2024; Xi et al., 2023). Integrating AI chatbots is a cost-effective strategy for organizations, delivering value to customers while reducing operational

expenses. For instance, automation in customer interactions is projected to save retailers approximately \$439 billion annually by 2023, as they replace many human-operated customer service roles with chatbots (Williams, 2019). The emergent of generative AI, particularly large language models (LLMs), is revolutionizing the functionality of chatbots by significantly enhancing their interactive communication capabilities. LLMs, such as OpenAI's GPT-series and Llama, enabling chatbots to understand and generate natural language with remarkable fluency and context-awareness (OpenAI et al., 2024; Touvron et al., 2023a).

Recently, there is a trend of a significant transition in chatbot applications, shifting from general inquiry tools to professional service agents capable of delivering specialized support. The increasingly prevalent human-AI interactions in high-stake industries (e.g., healthcare and finance) require professional service agents to offer a more professional and functional experience than simple knowledge-enhanced inquiry tools (Safi et al., 2020). This leads to a critical challenge in human-AI communication that has not been well addressed in existing studies, namely *how AI chatbots can communicate with humans and answer questions in a manner consistent with the chatbots' professional identity*. Beyond simply responding to user queries, professional service agents always have a defined professional identity, such as that of a virtual doctor, which requires their responses to be aligned with professional goals, such as supporting patients' health behavior change (e.g., treatment or medication adherence and reducing substance abuse) (Clark and Bailey, 2024). However, existing AI chatbots primarily emphasize information retrieval and question answering, overlooking the professional intention and goals that are integral to professional fields (Woo Lee et al., 2024; Singhal et al., 2025).

To develop a professional identity, an AI chat-

*Corresponding author

bot should have the ability to think and act like a professional. Firstly, the ability to analyze and break down complex professional tasks into more manageable components is essential. In contrast to mathematical and logical problems, which typically have clear chains of thought and definitive answers(Wei et al., 2022), many question-and-answer (Q&A) tasks in a service scenario are open-ended and require free-form responses(Megahed et al., 2024; Singhal et al., 2025). The challenge lies in effectively guiding these chatbots to adopt a professional thought process. Secondly, given their role as frontline professionals, AI chatbots must have the competence to effectively address user inquiries and deliver reliable services. However, it should be noted that LLMs are non-deterministic, meaning that the same input can generate different outputs, some of which may be incorrect or hallucinated(Farquhar et al., 2024). These drawbacks can result in occasional inaccurate answers or hallucination problems, which is fatal in the high-stakes field(Ji et al., 2023; Liu et al., 2025). Interactions with such chatbots have the potential to trigger undesirable behaviors in users, such as noncompliance among patients, which can negatively impact both service providers and users.

Motivated by to urgent need to provide professional service to users through LLM-based chatbots, our goal is to propose a novel framework, i.e., LAPI (LLM-based Agent with a Professional Identity), which aims to enhance the effectiveness of LLM-based chatbots in serving as professional service agents, particularly in the context of medical Q&A services. To address the aforementioned challenges, we consider a subfield of linguistics, namely pragmatics, which how language is used for social interactions and the attainment of goals(Ciccina and Turkstra, 2002) and the use of information from context to determine the effective means of achieving a given goal(Grice, 1975). From a pragmatic perspective, AI chatbots should provide useful responses and remove meaningless or incorrect information. The proposed approach involves two key steps: first, the comprehension and decomposition of complex professional tasks into manageable subtasks through a theory-guided task planning process; and second, the iterative updating and eventual generation of professional and ethical responses using a pragmatic entropy method. Specifically, we propose a theory-guided task planning process that employs a conceptual framework in the healthcare domain—namely, the

Health Belief Model (HBM)—to guide AI chatbots in how they think and act to support interventions aimed at encouraging healthy behaviors. Furthermore, we propose a pragmatic entropy method that generates professional and ethical responses, with the objective of maximizing the clarity, empathy, and relevance of the chatbot’s outputs while ensuring alignment with ethical standards and professional goals.

In general, our contributions are:

- We introduce a novel framework for developing a professional identity for an LLM-based chatbot, with the objective of employing the chatbot as a professional service agent for medical Q&A services. Specifically, we propose a pragmatic entropy method for the generation of professional and ethical responses. Meanwhile, we provide an effective way to utilize a domain theory to guide task planning process.
- Experiments on various real medical Q&A setting show that our method significantly improves response quality, providing more accurate, empathetic, and professional answers compared to baseline approaches.

2 Research Background

2.1 Medical AI Chatbots

The medical field represents a domain in which the implementation of AI chatbots is gaining traction as a means of facilitating access to information from the patient perspective and alleviating the workload on doctors(Safi et al., 2020). A variety of methods and techniques have been employed in the development of medical chatbots for diverse objectives, including statistical methods, deep learning, and LLMs(Mihailidis et al., 2008; Jothi et al., 2022; woo Lee et al., 2024). These bots have been developed for diverse applications, including various aspects of health and medicine. These range from chatbots that support general health and activities of daily living to chatbots that assist specific disease management, such as heart disease and cancer(Wong et al., 2011; woo Lee et al., 2024). Prior to the advent of LLMs, conventional chatbots exhibited constrained contextual and language understanding, resulting in inaccurate responses and suboptimal interaction in a human-like manner. Recently, significant advancements have been made

in the development of LLM-based medical chatbots, which are designed to facilitate human-like interactions with patients(Singhal et al., 2025).

2.2 Professional LLM-based Chatbots

Chatbots have stepped into the age of LLM empowerment. The rapid development of LLMs enhances the capabilities of chatbots by improving their conversational proficiency and facilitating human-like interactions. ChatGPT exemplifies this category of chatbot, optimized to generate natural, human-like dialogue. Other widely used LLMs include GPT4, Llama, Mistral, ChatGLM2, and Gemma(OpenAI et al., 2024; Touvron et al., 2023a; Jiang et al., 2023; GLM et al., 2024; Team et al., 2024). Chatbots have been employed in customer service to address customer inquiries regarding products or services, either as an initial point of contact or as an alternative to speaking with a human representative(Hong et al., 2025). In addition to healthcare, LLM-based chatbots have gradually become powerful tools in areas such as retail, education, research, and many others(Megahed et al., 2024; Benzinho et al., 2024; Giudici et al., 2024). To improve chatbot’s ability on professional tasks, such as programming education and home designing, various knowledge-enhanced strategies, such as fine-tuning, prompt engineering, and Retrieval-Augmented Generation (RAG), have been employed(Lewis et al., 2021).

3 Theoretical Background of Professional Service Agents

3.1 Professional Service Agents

From a philosophical standpoint, the term "agent" traditionally describes entities possessing desires, beliefs, intentions, and the capacity to action(Xi et al., 2023; Davidson, 1963). Human agents, for instance, are frequently employed by companies to handle customer service operations, focusing on addressing customer queries and ensuring satisfaction through effective communication and problem-solving skills. With the progression of AI, the concept of an "agent" has expanded to describe AI entities that are characterized by intelligent behavior, such as autonomy, proactiveness, and social capabilities. AI agents, particularly chatbots powered by large language models, have demonstrated remarkable versatility across various scenarios, attributable to their extensive inherent knowledge and human-like communication abilities(Huang et al.,

2024; Chong et al., 2021).

LLM-based agents are typically structured around three key components. Agent Core serves as the main brain of the system, typically utilizing an LLM with general-purpose capabilities. Task Planning is responsible for decomposing complex tasks into manageable subtasks, thereby enhancing the agent’s reasoning and reliability of responses. Actions of a chatbot are communication with humans. A professional AI agent’s primary actions are aimed at high-level information transmission to fulfill a robot’s intent, a process referred to as “intention action.” A typical example of such agents is applications in the healthcare domain, where empathy-aware agents assist in diagnostics and knowledge retrieval to provide a user-friendly service(Mao et al., 2022).

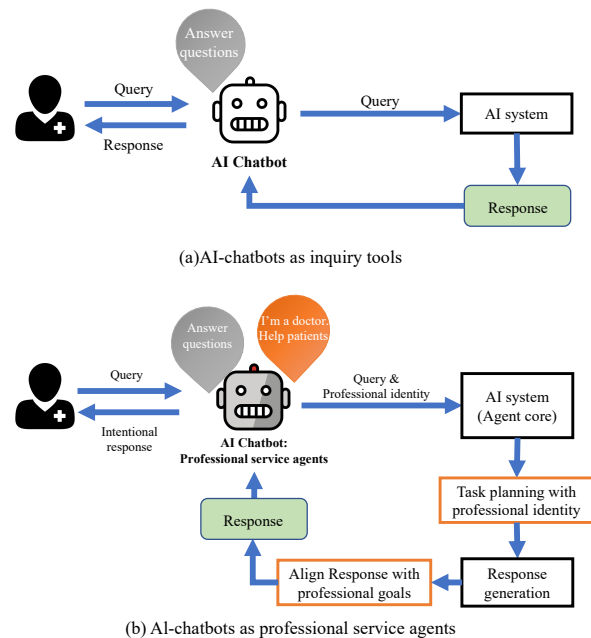


Figure 1: Distinction Between AI Chatbots as Inquiry Tools and Professional Service Agents

Responding to the trend of chatbots shifting from general inquire tools to professional service agents, we identify the significant distinction between AI-chatbots as inquiry tools and as professional service agents, as shown in Figure 1. Professional identity is a crucial attribute of a professional service agent, significantly shaping the agent’s goals, task planning, and action execution. Chatbots with professional identity have distinct professional intentions and goals, which in turn influence their task planning and action execution. Using medical Q&A as an example, inquiry tools primarily focus on answering medical questions(Safi et al., 2020).

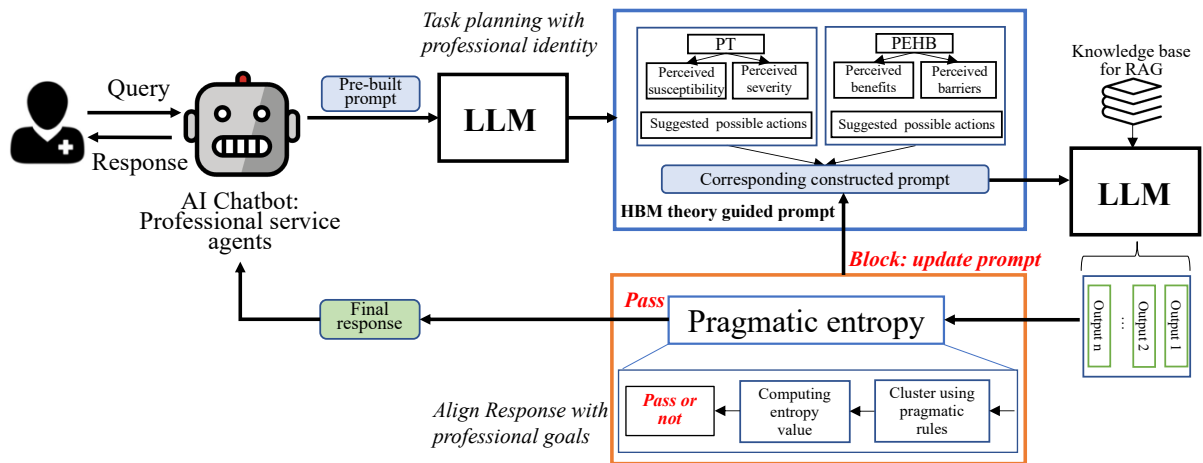


Figure 2: Overview of LLM-based Agent with a Professional Identity (LAPI). The patient’s query is first assigned weights for the two HBM domains—Perceived Threat and Perceived Effectiveness of Health Behavior. An initial prompt is then generated based on these weights, followed by the addition of pragmatic rules. Multiple outputs are generated by the LLM, and pragmatic entropy optimization is applied to obtain the most appropriate response.

In contrast, professional service agents serve two purposes: responding to medical inquiries and promoting patient well-being by encouraging healthy behaviors, aligning with their professional identity (e.g., doctor). For task planning, the decomposition of complex professional tasks should adhere to established professional standards, informed by either theories or expert guidelines. Regarding actions, the chatbot’s responses should align with professional goals. This alignment ensures that the chatbot not only addresses patient questions but also offers suggestions and persuasive guidance to foster healthy behaviors.

3.2 Health Belief Model

The Health Belief Model (HBM) is a widely used conceptual framework in health behavior research, which supports interventions aimed at encouraging healthy behaviors (Champion and Skinner, 2008). Initially proposed in the 1950s, the model has evolved in response to practical public health concern and healthcare management. Rooted in a well-established body of psychological and behavioral theories, the HBM posits that health-related behavior depends primarily influenced by two factors: (1) the individual’s desire to avoid illness or recover from it if already affected; and (2) the belief that a particular health action care prevent or mitigate the illness (Janz and Becker, 1984).

The HBM comprises four key constructions: Perceived Susceptibility, Perceived Severity, Perceived Benefit, and Perceived Barrier. These constructs can be broadly categorized into two main aspects:

Perceived Threat (including Perceived Susceptibility and Perceived Severity) and Perceived Effectiveness of Health Behavior (including Perceived Benefit and Perceived Barrier). Perceived Susceptibility refers to an individual’s belief about the likelihood of contracting a particular condition. Perceived Severity refers to the belief regarding the seriousness of contracting an illness or the consequences of leaving it untreated. Perceived Benefit relates to the belief in the effectiveness of the recommended action to mitigate the disease threat. Perceived Barrier refers to the belief regarding the potential negative aspects or costs associated with the recommended health action. Individuals tend to weigh the anticipated benefits of an action against the perceived barriers, resulting in a belief in the effectiveness of the health behavior. These four constructions interact to guide individuals toward a preferred course of action.

The HBM has been employed to inform health intervention designs aimed at promoting healthy behaviors (Orji et al., 2012), particularly as patient-centric (proactive) healthcare gains popularity. Previous research has shown that interventions grounded in established theories and models tend to be more effective than those developed based on intuition (Glanz, 1997).

4 Methods

4.1 Problem Formulation

Given a patient’s question q and a set of predefined pragmatic rules aligning with a professional iden-

tity $R = \{r_1, r_2, r_3, \dots, r_m\}$, the objective is to generate a response s that not only provides an professional answer to the query but also adheres to these established rules, thereby reflecting the professional identity (i.e., the doctor in our experiment setting). The desired response s should satisfy the following condition: it adheres to the pragmatic rules with a minimal level of uncertainty in the response distribution. The problem can be formally expressed as:

$$s = \text{LLM}(q, R, \tau, \epsilon) \quad (1)$$

where τ is the satisfaction rate threshold, and ϵ is the pragmatic entropy threshold.

Figure 2 illustrates the proposed framework, i.e., LLM-based Agent with a Professional Identity (LAPI), which includes a theory-guided task planning process and a pragmatic entropy method. In the following sections, we introduce details on our specific implementation of these two components.

4.2 Theory-guided Task Planning

We employ the Health Belief Model (HBM) as a guiding framework to structure our design, decomposing the professional goal of promoting healthy behavior into manageable sub-tasks. These sub-tasks include: fostering individuals' accurate perception of disease risk, aiding individuals in developing a realistic understanding of the consequences associated with a condition or recommended action, enhancing individuals' awareness of the potential benefits of adopting healthy behaviors, and offering strategies to overcome perceived barriers to implementing these behaviors.

The task planning process begins by assessing the relevance of the user's query to the two core categories of HBM: *Perceived Threat* and *Perceived Effectiveness of Health Behavior*. To achieve this, we first prompt the LLM to estimate the proportions of the query related to these two categories.

Based on the prompt shown in Figure 4 of Appendix A, the LLM assesses the query's relevance to each category and assigns a corresponding weight to each. Next, we generate the initial prompt p_{initial} for the LLM by constructing two specific prompts: one fully focused on the *Perceived Threat* category, and another fully focused on *Perceived Effectiveness of Health Behavior*. These prompts are pre-configured to ensure that the LLM generates an initial prompt that accurately reflects the correct placement of the question within the HBM

theory framework. The content of the prompt is detailed in Figure 5 of Appendix A.

4.3 Pragmatic Entropy

Pragmatic entropy aims to assist the AI chatbot in generating professional and ethical responses by reducing the entropy of model outputs across different pragmatic rule categories and maximizing the proportion of responses that satisfy all predefined pragmatic rules $R = \{r_1, r_2, r_3, \dots, r_m\}$.

Given a prompt p , the LLM generates a set of n responses by sampling from its output distribution with temperature 1, which enables the model's outputs to more authentically reflect its internal representations and reasoning processes:

$$S = \text{LLM}(p, \text{temperature} = 1) \quad (2)$$

where $S = \{s_1, s_2, \dots, s_n\}$ represents the generated responses.

Each response $s \in S$ is classified into categories C_j based on its satisfaction of specific subsets of the pragmatic rules R . This classification process is judged by the LLM itself. The following metrics are computed to evaluate the quality of the prompt.

The proportion of responses satisfying all rules is given by:

$$P(C_{\text{all}}|S) = \frac{|\{s_i \in S \mid s_i \in C_{\text{all}}\}|}{n} \quad (3)$$

where $|\{s_i \in S \mid s_i \in C_{\text{all}}\}|$ denotes the number of responses in S that fall into category C_{all} , and n is the total number of generated responses.

The entropy of the output distribution across all categories C_j :

$$PE(S) = - \sum_{C_j} P(C_j|S) \log P(C_j|S) \quad (4)$$

where $P(C_j|S) = \frac{|\{s_i \in S \mid s_i \in C_j\}|}{n}$ represents the proportion of responses in S that fall into C_j .

To refine the prompt iteratively, the process consists of generating an improved prompt, evaluating it quantitatively, and iterating until either the desired criteria are met or the maximum number of iterations is reached.

To improve the prompt, the current prompt p along with information on unmet conditions (Ucs) is fed into the large language model to generate a refined version p_{new} :

$$p_{\text{new}} = \text{LLM}(p, Ucs, \text{temperature} = 0.3) \quad (5)$$

where p_{new} is the updated prompt. Ucs refer to the missing pragmatic rules identified within the category that has the largest proportion in categories that have not satisfied all the rules. The parameter temperature = 0.3 (Moslem et al., 2023; Wassie, 2024) is chosen to balance stability and flexibility in the prompt generation process.

If the satisfaction rate $P(C_{\text{all}}|S)$ meets or exceeds a predefined threshold τ , and the pragmatic entropy $PE(S)$ is below a threshold ϵ , the prompt p is considered optimal, and the process terminates with $p^* = p$. Otherwise, the model proceeds to refine the prompt. The overall process can be seen in Algorithm 1.

Algorithm 1 Pragmatic entropy

Input: Current prompt p , pragmatic rules R , satisfaction threshold τ , entropy threshold ϵ , maximum iterations T .

Output: Final response s^* .

```

1: Set iteration counter  $t \leftarrow 0$ ;
2: while  $t < T$  do
3:   Generate response set  $S$  using Eq.(2);
4:   Compute  $P(C_{\text{all}}|S)$  using Eq.(3);
5:   Compute  $PE(S)$  based on Eq.(4);
6:   Select  $s^*$  randomly from  $C_{\text{all}}$ ;
7:   if  $P(C_{\text{all}}|S) \geq \tau$  and  $PE(S) \leq \epsilon$  then
8:     Return  $s^*$ ;
9:   else
10:    Identify  $Ucs$  from categories that have
      not satisfied all the rules;
11:    Get prompt  $p_{\text{new}}$  using Eq.(5);
12:    Set  $p \leftarrow p_{\text{new}}$ ;
13:   end if
14:   Increment iteration counter  $t \leftarrow t + 1$ 
15: end while
16: Return  $s^*$ 

```

5 Experiments

5.1 Experimental Settings

Dataset and Metrics The dataset used in our experiments was collected from a reputable ophthalmology hospital and comprises 400 real-world patient inquiries paired with concise and professional answers provided by expert ophthalmologists. To support Retrieval-Augmented Generation (RAG), we utilized 300 patient-question-answer pairs as the knowledge base for retrieval. To assess the efficacy of the proposed method, a test set of 10 patient inquiries was randomly selected. This test set was

used to evaluate the performance of the method in handling real-world medical question-and-answer scenarios.

To comprehensively evaluate the performance of our proposed method, we utilized two categories of metrics: LLM-based and non-LLM-based metrics. For LLM-based evaluation, we employed G-EVAL (Liu et al., 2023), a state-of-the-art metric designed to assess various qualitative aspects of large language model outputs, including Fluency, Coherence, Naturalness, Empathy, and Patient-Centricity, where higher scores indicate better performance. Detailed descriptions of these five evaluation metrics can be found in Appendix B. G-EVAL has been shown to exhibit strong consistency with human evaluations (Zhu et al., 2025; Gao et al., 2024), making it a reliable tool for qualitative assessment. For non-LLM-based evaluation, we adopted ROUGE-L recall and precision (Lin, 2004) as objective n-gram matching-based metrics. ROUGE-L Recall measures the extent to which the generated responses cover the content of the doctor-provided answers, with higher values indicating better coverage, while ROUGE-L Precision evaluates the informativeness of the generated responses, with lower values indicating that the generated responses contain more information.

Baselines We categorize the baseline methods into two groups: prompt design-based methods and quality inspection-based methods. For the prompt design-based methods, we include the following: **Zero-Shot**, where the model generates responses directly without additional guidance; **Zero-Shot + Pragmatic Rules**, which enhances the zero-shot approach by incorporating explicit pragmatic rules tailored to professional contexts; **RAG-based Few-Shot** (Ram et al., 2023; Yu et al., 2023), a retrieval-augmented generation strategy that retrieves relevant examples from a knowledge base to assist response generation; and **Chain-of-Thought (CoT)** (Kojima et al., 2023), which encourages the model to generate step-by-step reasoning paths to improve response accuracy and relevance. For the quality inspection-based methods, we employed **Semantic Entropy** (Farquhar et al., 2024), a technique designed to evaluate the confidence of the model regarding specific knowledge points in its generated responses. The confidence scores are then fed back into the model, prompting it to revise and improve its outputs.

Models	Approach	Flu.	Coh.	Nat.	Emp.	Pat-C.	Rec.	Pre.
Llama-2-7B-Chat	Zero-Shot	3.90	3.60	2.95	2.10	2.45	0.39	0.07
	Zero-Shot + Pragmatic rules	4.00	3.50	3.40	2.70	3.00	0.33	0.10
	RAG-based Few-Shot	3.60	2.50	2.65	2.20	2.45	0.38	0.07
	Chain-of-Thought	4.00	3.60	2.80	2.20	2.40	0.40	0.07
	LAPI	4.00	3.60	3.05	2.90	3.10	0.35	0.07
Llama-2-13B-Chat	Zero-Shot	3.75	3.60	2.85	2.25	2.65	0.38	0.08
	Zero-Shot + Pragmatic rules	4.00	3.75	3.10	3.05	3.25	0.33	0.12
	RAG-Based Few-Shot	3.60	2.65	2.75	2.35	2.25	0.31	0.07
	Chain-of-Thought	3.65	3.65	2.95	2.10	2.50	0.39	0.07
	LAPI	4.00	3.75	3.15	3.40	3.45	0.36	0.07
Llama-2-70B-Chat	Zero-Shot	3.90	3.45	3.15	2.15	2.80	0.35	0.09
	Zero-Shot + Pragmatic rules	4.00	3.75	3.40	2.80	3.35	0.35	0.08
	RAG-Based Few-Shot	3.80	3.20	2.95	2.35	2.95	0.31	0.11
	Chain-of-Thought	3.95	3.45	3.05	2.40	2.70	0.38	0.07
	LAPI	4.00	3.75	3.50	3.35	3.35	0.36	0.07
GPT-3.5-Turbo	Zero-Shot	4.00	3.60	3.30	2.35	3.15	0.25	0.15
	Zero-Shot + Pragmatic rules	4.00	3.60	3.45	3.15	3.55	0.26	0.16
	RAG-Based Few-Shot	4.00	3.45	3.25	2.10	3.30	0.26	0.24
	Chain-of-Thought	4.00	3.70	3.10	2.30	3.10	0.29	0.16
	Semantic Entropy	4.00	3.40	3.10	2.30	3.15	0.21	0.21
GPT-4	LAPI	4.00	3.85	3.60	3.50	3.70	0.29	0.13
	Zero-Shot	4.00	3.70	3.05	2.10	2.90	0.35	0.08
	Zero-Shot + Pragmatic rules	4.00	3.90	3.50	3.55	3.80	0.35	0.08
	RAG-Based Few-Shot	4.00	3.65	3.25	2.30	3.25	0.36	0.11
	Chain-of-Thought	4.00	3.70	3.10	2.35	2.85	0.37	0.08
	Semantic Entropy	4.00	3.65	3.05	2.40	2.80	0.35	0.07
	LAPI	4.00	3.90	3.95	3.85	3.80	0.38	0.07

Table 1: Evaluation results. The table presents the performance of different prompting approaches across various evaluation metrics. Flu., Coh., Nat., Emp., and Pat-C. represent Fluency, Coherence, Naturalness, Empathy, and Patient-Centricity, respectively. Rec. and Pre. refer to ROUGE-L Recall and Precision. Higher scores for the first six metrics indicate better performance, while a lower score for Pre. reflects a more desirable outcome. **Bold:** Best-performing paradigm per metric.

Experiment Details We evaluate our proposed method on five LLMs: Llama-2-7B-Chat, Llama-2-13B-Chat, Llama-2-70B-Chat (Touvron et al., 2023b), GPT-3.5-Turbo (gpt-3.5-turbo-0125), and GPT-4 (gpt-4-0125-preview), selected for their diverse scales and capabilities, representing both open-source and closed-source models. To enhance the alignment of responses with professional and human-like communication, we incorporated pragmatic rules developed through a streamlined process: we reviewed medical literature, identified key communication principles (focusing on reply format, information content, and user emotional impact), and distilled these into actionable rules, which are detailed in Appendix C.

For the RAG-based Few-Shot method, we employed a retrieval-augmented generation strategy using a BERT encoder to retrieve the three most relevant question-answer pairs based on cosine similarity, which were then used as prompts to guide response generation. In the LAPI method, we set n to 5, τ to 0.5, and ϵ to 0.8. For the Semantic Entropy method, due to the high text comprehension capability required for this method, which was only applied to GPT models.

5.2 Experimental Results

Our proposed approach LAPI demonstrates superior performance across various models, as evidenced by the results in Table 1. From the results perspective, LAPI consistently outperforms other methods in terms of Fluency, Coherence, Naturalness, Empathy, Patient-Centricity, ROUGE-L Recall, and Precision. For instance, in the same model family (e.g., Llama-2-7B-Chat), LAPI provides the highest scores across multiple metrics highlighting the effectiveness of LAPI in generating more precise, empathetic, and coherent responses, tailored to professional service agent tasks. From the model perspective, we observe that LAPI performs even better on larger models such as GPT-3.5-Turbo and GPT-4. The enhanced textual comprehension capabilities of these models enable them to generate more accurate prompts and subsequently produce better results. LAPI benefits from this ability, yielding more contextually rich and reliable outputs.

The combination of pragmatic rules and the HBM framework allows LAPI to produce responses that are not only accurate and fluent but also deeply empathetic and patient-centric. By in-

After laser surgery for high myopia, there is a possibility of the degree of myopia increasing again over time. This is known as regression, and it can happen in some cases. It is important to have regular follow-up appointments with your ophthalmologist to monitor any changes in your vision and to discuss any concerns you may have. Additionally, continuing to follow good eye habits and maintaining a healthy lifestyle can help minimize the risk of regression. If you notice any changes in your vision, make sure to schedule an appointment with your eye doctor for further evaluation.	After undergoing laser surgery for high myopia, there is a possibility that the degree of myopia may increase again over time. This is known as regression, and it can occur for various reasons such as natural changes in the eye, age-related changes, or the initial surgery not being completely effective. It is important to follow up with your eye doctor regularly after laser surgery to monitor any changes in your vision and to discuss any concerns you have about the progression of your myopia. They will be able to provide guidance on the best course of action to manage any potential increase in myopia.	It's possible for the degree of myopia to fluctuate, even after undergoing laser surgery. Some studies have shown that a small number of people may have some regression of the initial treatment result over time. However, it's important to consult with your eye surgeon for a personalized assessment of your eyes and to discuss any potential risks or concerns related to your specific situation. Regular follow-up appointments with your eye care provider are recommended to monitor any changes in your vision after laser surgery.	After undergoing laser surgery for high myopia, there is a possibility that the degree of myopia may increase again over time. This is known as regression, and it can happen in a small percentage of cases. It is important to follow up with your eye doctor regularly after surgery to monitor any changes in your vision and address any potential issues promptly.
LAPI	Chain-of-Thought	Zero-Shot	RAG-Based Few-Shot

Figure 3: Overview of GPT-3.5-Turbo’s responses using different methods to answer the patient query: "After undergoing laser surgery for high myopia, will the degree of myopia increase again?"

tegrating step-by-step reasoning and continuous feedback mechanisms, our method ensures the delivery of high-quality, professional, and reliable responses. These results demonstrate the potential of LAPI to improve interactions in professional service applications, especially in the healthcare sector where empathy and clarity are paramount.

5.3 Ablation Study

We conducted an ablation study using GPT-3.5-Turbo as the base model to evaluate the effectiveness of each component in our proposed method. Table 2 presents the results of the ablation study, where we compare three different setups: the full LAPI method, the model without the HBM framework w/o HBM, and the model without the Pragmatic Entropy component w/o PE. As shown in the table, the full LAPI method outperforms both variants in all evaluation metrics, including Fluency, Coherence, Naturalness, Empathy, Patient-Centricity, ROUGE-L Recall, and Precision.

Approach	Flu.	Coh.	Nat.	Emp.	Pat-C.	Rec.	Pre.
LAPI	4.00	3.85	3.60	3.50	3.70	0.29	0.13
w/o HBM	4.00	3.70	3.75	3.30	3.60	0.27	0.15
w/o PE	3.90	3.70	3.50	2.95	3.60	0.26	0.16

Table 2: Ablation study on GPT-3.5-turbo.

The results indicate that each component contributes positively to the overall performance. Specifically, removing the HBM framework (w/o HBM) leads to a slight decrease in the Coherence and Empathy metrics, while the absence of the Pragmatic Entropy (w/o PE) reduces the overall Naturalness and Empathy. These findings suggest that combining both the HBM framework and Pragmatic Entropy is more effective than using them separately, as the combined approach leverages the strengths of both components, resulting in more fluent, coherent, and empathetic responses tailored

to professional service applications.

5.4 Case Study

To further illustrate the capabilities of our proposed method, we present a case study using GPT-3.5-Turbo to answer the patient query. Figure 3 provides a comparison of the responses generated by four different prompting strategies: LAPI, CoT, Zero-Shot, and RAG-Based Few-Shot. In the responses, green text directly answers the question, yellow text promotes regular doctor contact, and blue text provides actionable patient advice. As observed in the case study, LAPI (left) outperforms others, providing a more comprehensive and detailed response that directly answers the question, stresses the importance of regular ophthalmologist follow-ups, and offers advice on healthy eye habits. This showcases the benefit of integrating the HBM framework and Pragmatic Entropy. In contrast, other methods provide direct answers and some patient advice but lack LAPI’s level of detail and actionable guidance. For instance, CoT misses lifestyle advice, and RAG-Based Few-Shot doesn’t emphasize continuous patient-doctor communication as strongly as LAPI.

The LAPI framework also demonstrates superior performance compared to various CoT variants, with detailed results presented in Appendix D.

6 Conclusion

In this work, we propose a novel approach for designing LLM-based professional service agents specifically tailored for medical Q&A services. Our method promotes communication aligned with professional identities by integrating theory-guided task planning and a pragmatic entropy method to generate ethical, coherent, and patient-centric responses. Experimental results demonstrate that our approach outperforms baseline methods in flu-

ency, empathy, patient-centricity, and ROUGE-L scores. This study underscores the potential of enhancing AI chatbot communication in healthcare and encourages the continued development of more effective methods in future research.

7 Limitations

While our LAPI framework shows significant promise for enhancing professional communication in medical Q&A services, several aspects require further refinement. The current evaluation primarily relies on large language models to assess output quality, which aligns closely with human judgment. However, incorporating real-world testing, such as deploying the model in clinical settings, would further validate its practical effectiveness and applicability. While this didn't affect the validity of the proposed method—since results across different LLM sizes showed notable improvements over baseline methods in attributes like Empathy—the smaller sample size might limit how broadly the findings can be applied. Furthermore, the experiments focused solely on medical Q&A scenarios. While the framework is designed to apply to other domains like law, education, or finance, its effectiveness in these areas requires further validation.

8 Ethical statements

The dataset was provided by medical professionals from a reputable ophthalmology hospital and consists solely of anonymized question-answer pairs without any identifiable patient information. During the data collection process, patients provided verbal consent for their questions to be used for academic research purposes and were advised to avoid including personal or identifiable information in their queries. This process ensured that no personal health information was collected or used. Furthermore, the dataset was reviewed and approved by the Institutional Review Board (IRB) of the collaborating ophthalmology hospital prior to its use in our study, ensuring compliance with ethical standards.

9 Acknowledgements

We would like to thank all the reviewers for their helpful feedback, and EMNLP 2025 and ACL Rolling Review organizers for their efforts. This work was supported by the grants from National Natural Science Foundation of China (No.72201068, 72131004, 72342011).

References

- Katherine A Allen, Victoria Charpentier, Marissa A Hendrickson, Molly Kessler, Rachael Gotlieb, Jordan Marmet, Emily Hause, Corinne Praska, Scott Lunos, and Michael B Pitt. 2023. [Jargon be gone – patient preference in doctor communication](#). *Journal of Patient Experience*, 10:23743735231158942.
- José Benzinho, João Ferreira, Joel Batista, Leandro Pereira, Marisa Maximiano, Vítor Távora, Ricardo Gomes, and Orlando Remédios. 2024. Llm based chatbot for farm-to-fork blockchain traceability platform. *Applied Sciences*, 14(19):8856.
- Victor L. Champion and Christine S. Skinner. 2008. *The health belief model*, 4 edition.
- Shuqing Chen, Xitong Guo, Tianshi Wu, and Xiaofeng Ju. 2020. [Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis](#). *Information Processing & Management*, 57(5):102253.
- Terrence Chong, Ting Yu, Debbie Isobel Keeling, and Ko de Ruyter. 2021. [Ai-chatbots on the services frontline addressing the challenges and opportunities of agency](#). *Journal of Retailing and Consumer Services*, 63:102735.
- Angela H Ciccio and Lyn S Turkstra. 2002. Cohesion, communication burden, and response adequacy in adolescent conversations. *Advances in Speech Language Pathology*, 4(1):1–8.
- Michael Clark and Sarah Bailey. 2024. Chatbots in health care: Connecting patients to information. *Canadian Journal of Health Technologies*, 4(1).
- Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. [A complete survey on llm-based ai chatbots](#). *Preprint*, arXiv:2406.16937.
- Donald Davidson. 1963. Actions, reasons, and causes. *Journal of Philosophy*, 60(23):685–700.
- Stephen Farquhar, Johan Kossen, Lukas Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Dairui Liu, Tianwei She, Yang Jiang, and Irene Li. 2024. [Large language models on wikipedia-style survey generation: an evaluation in nlp concepts](#). *Preprint*, arXiv:2308.10410.
- Mathyas Giudici, Luca Padalino, Giovanni Paolino, Ilaria Paratici, Alexandru Ionut Pascu, and Franca Garzotto. 2024. Designing home automation routines using an llm-based chatbot. *Designs*, 8(3):43.
- Karen Glanz. 1997. *Theory at a glance: A guide for health promotion practice*. 97. US Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute.

- Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.
- Mengze Hong, Chen Jason Zhang, Di Jiang, Yuanfeng Song, Lu Wang, Yuanqin He, Zhiyang Su, and Qing Li. 2025. [Expanding chatbot knowledge in customer service: Context-aware similar question generation using large language models](#). *Preprint*, arXiv:2410.12444.
- Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, Ashley Llorens, John Langford, Hoi Vo, Li Fei-Fei, Katsu Ikeuchi, and Jianfeng Gao. 2024. [Position paper: Agent ai towards a holistic intelligence](#). *Preprint*, arXiv:2403.00833.
- Nancy K. Janz and Marshall H. Becker. 1984. The health belief model: A decade later. *Health Education Quarterly*, 11(1):1–47.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Shaohai Jiang, Zhengyu Wu, Xiaoyu Zhang, Ying Ji, Jingyong Xu, Peng Liu, Yan Liu, Jie Zheng, Liang Zhao, and Jingxi Chen. 2024. [How does patient-centered communication influence patient trust?: The roles of patient participation and patient preference](#). *Patient Education and Counseling*, 122:108161.
- J. Nirmal Jothi, S.Poongodi, V.Chinnammal, L.Kannagi, M.Panneerselvam, and R. Thandaiah Prabu. 2022. Ai based humanoid chatbot for medical application. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1135–1140. IEEE.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. [Large language models are zero-shot reasoners](#). *Preprint*, arXiv:2205.11916.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen tau Yih, Tim Rockt  schel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Feng Liu, Jiaqi Jiang, Yating Lu, Zhanyi Huang, and Jiuming Jiang. 2025. [The ethical security of large language models: A systematic review](#). *Frontiers of Engineering Management*, 12:128–140.
- Wenlong Liu, Min Jiang, Wangjie Li, and Jian Mou. 2024. [How does the anthropomorphism of ai chatbots facilitate users’ reuse intention in online health consultation services? the moderating role of disease severity](#). *Technological Forecasting and Social Change*, 203:123407.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2022. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.
- Fouad M. Megahed, Y. J. Chen, James A. Ferris, and 1 others. 2024. How generative ai models such as chatgpt can be (mis) used in spc practice, education, and research? an exploratory study. *Quality Engineering*, 36(2):287–315.
- Alex Mihailidis, Jason N. Boger, Tim Craig, and Jesse Hoey. 2008. The coach prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics*, 8:1–18.
- Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. [Fine-tuning large language models for adaptive machine translation](#). *Preprint*, arXiv:2312.12740.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Rochelle Orji, Judith Vassileva, and Regan Mandryk. 2012. Towards an effective health interventions design: an extension of the health belief model. *Online Journal of Public Health Informatics*, 4(3):e61050.

- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-context retrieval-augmented language models](#). *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Ziad Safi, Ahmad Abd-Alrazaq, Mohamed Khalifa, and Mowafa Househ. 2020. Technical aspects of developing chatbots for medical applications: scoping review. *Journal of Medical Internet Research*, 22(12):e19127.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohammed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Aman Kassahun Wassie. 2024. [Machine translation for ge’ez language](#). *Preprint*, arXiv:2311.14530.
- Jianwei Wei, Xuezhi Wang, Dale Schuurmans, and Yarin Gal. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Ruth Williams. 2019. Study: Chatbots to drive \$112 b in retail sales by 2023. [MarketingDive](#).
- Wai Wong, Jothi Thangarajah, and Leendert Padgham. 2011. Health conversational system based on contextual matching of community-driven question-answer pairs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2577–2580.
- Jae woo Lee, In-Sang Yoo, Ji-Hye Kim, Won Tae Kim, Hyun Jeong Jeon, Hyo-Sun Yoo, Jae Gwang Shin, Geun-Hyeong Kim, ShinJi Hwang, Seung Park, and Yong-June Kim. 2024. [Development of ai-generated medical responses using the chatgpt for cancer patients](#). *Computer Methods and Programs in Biomedicine*, 254:108302.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 10 others. 2023. [The rise and potential of large language model based agents: A survey](#). *Preprint*, arXiv:2309.07864.
- Guoxin Yu, Lemao Liu, Haiyun Jiang, Shuming Shi, and Xiang Ao. 2023. [Retrieval-augmented few-shot text classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6721–6735, Singapore. Association for Computational Linguistics.
- Xin Zhang, Linzi Li, Quan Zhang, Long Hoang Le, and Yijin Wu. 2024. [Physician empathy in doctor-patient communication: A systematic review](#). *Health Communication*, 39(5):1027–1037.
- Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025. [Factual dialogue summarization via learning from large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4474–4492, Abu Dhabi, UAE. Association for Computational Linguistics.

A Prompts Used in Theory-guided Task Planning

[QUESTION]: {question}
 [HBM theory]: The Health Belief Model focuses on understanding the cognition of vulnerability to disease along four dimensions... Does the QUESTION pertain to 'Perceived Threat' or 'Perceived Effectiveness of Health Behavior.'?
 There are five types of proportions. Please select your answer:
 (A) 100% perceived threat; 0% perceived effectiveness of health behavior
 (B) 70% perceived threat; 30% perceived effectiveness of health behavior
 (C) 50% perceived threat; 50% perceived effectiveness of health behavior
 (D) 30% perceived threat; 70% perceived effectiveness of health behavior
 (E) 0% perceived threat; 100% perceived effectiveness of health behavior
 Answer:

Figure 4: Prompt used to assign weights to a question based on its relevance to the HBM categories.

If a question is about "100% diagnosis and treatment of diseases", the prompt will be : "You are a doctor, please answer the question of the visitor. These include: the likelihood and severity of the patient's illness; Provide actionable advice to patients."
 If a question about "100% perceived effectiveness of health behavior ", the prompt will be : "You are a doctor, Please respond to visitors' questions by addressing the health behaviors they can adopt regarding their concerns, the potential barriers to these behaviors, and provide feasible advice."
 Now, the proportion is {question_kind}, please give the prompt according to the proportion of the two category.
 New prompt:

Figure 5: Prompts used to generate an initial prompt that accurately reflects the correct placement of the question within the HBM theory framework.

B Evaluation Metrics

In the following content, we provide a detailed description of the five key evaluation metrics used to assess the performance of the proposed method, all of which are evaluated using the G-EVAL(Liu et al., 2023) framework with GPT-4. This approach ensures a comprehensive and consistent evaluation of the generated responses.

- **Fluency (1–4 points):** Quality of the response in terms of grammar, spelling, punctuation, word choice, and sentence structure.
 - *4 points:* No noticeable issues with grammar, spelling, punctuation, word choice, and sentence structure; the response is concise and clear.
 - *3 points:* The response has few or no grammatical or spelling errors; some word choices or structures may seem slightly unnatural but do not impede understanding.

- *2 points:* There are noticeable grammatical, spelling, or structural issues that affect fluency, but the overall message is understandable.
- *1 point:* Many errors in grammar, spelling, punctuation, or word choice, affecting overall understanding and the fluency of the response.

- **Coherence (1–4 points):** This dimension aligns with structure and coherence in the DUC quality issues, where the answer should be well-structured and organized. The answer should not merely be a compilation of related information but should develop step-by-step into a coherent body of information on a specific topic.

- *4 points:* Clear structure and logical organization, with natural and smooth transitions between sentences. Not only is the information relevant, but it also demonstrates overall coherence of the response, avoiding harsh transitions or topic interruptions.
- *3 points:* Generally clear structure and organization, though occasional logical jumps or transitions may not be natural. The thematic development is mostly complete, but the connection of some content is slightly mechanical or abrupt, slightly diminishing the overall coherence.
- *2 points:* Clear breaks or jumps between sentences, with rather loose logical organization. The thematic development of the information is unclear, with some content appearing irrelevant or repetitive, making the overall presentation disjointed.
- *1 point:* Sentences lack a clear logical relationship, with chaotic content organization, making it difficult to form a coherent theme development. There is a significant feeling of information piling up, lacking structural organization, making it difficult for the reader to understand the core idea.

- **Naturalness (1–4 points):** Assesses whether the answer conforms to the habits of natural language expression, resembling the tone and

logic of human everyday communication. Focus on word choice, tone, sentence structure, and the overall reading experience.

- *4 points*: The expression is smooth, the tone is natural and close to everyday conversation, sentence structures are flexible and varied, with no stiffness. Word choice is precise and engaging, creating a sense of realism in the dialogue.
 - *3 points*: Overall, the expression is natural, but some parts may appear formulaic or mechanized. The tone is slightly flat but does not hinder understanding. Word choice is appropriate but may lack variety.
 - *2 points*: Language expression is partly stiff or awkward, the tone is unnatural, appearing mechanical or overly formal. There may be sentence structures or logical arrangements that do not conform to everyday language habits.
 - *1 point*: The mode of expression is clearly unnatural, the tone is stiff or stilted, completely detached from everyday language habits, difficult for people to understand or accept.
- **Empathy (1–4 points)**: Comprehensive assessment of an individual’s ability to understand, empathize with, and respond to others’ emotions.
 - *4 points*: Able to deeply understand and empathize with others’ emotions, responding appropriately to others’ feelings, demonstrating high emotional sensitivity and care.
 - *3 points*: Can understand others’ emotions and appropriately respond to most situations, but may occasionally fall short in complex or subtle emotional expressions.
 - *2 points*: Sometimes understands others’ emotions, but responses may lack depth or contain misunderstandings, showing some emotional detachment.
 - *1 point*: Struggles to understand others’ emotions or emotional reactions, lacks empathy, responses often seem inappropriate or irrelevant.

- **Patient-Centricity (1–4 points)**: Comprehensive assessment of an individual’s ability to center communication on the patient’s needs, providing clear, practical, and approachable information in a conversational manner.

- *4 points*: Communication is highly patient-centered, using concise and approachable language, delivering practical advice tailored to the patient’s situation, and maintaining a logical and conversational tone that fosters trust and understanding. Without listing ideas.
- *3 points*: Generally patient-centered, but may lack full clarity, practical detail, or conversational ease in some parts while still being largely useful and understandable.
- *2 points*: Moderately patient-centered but may include excessive technical details, lack practical guidance, or have a less approachable tone, making it harder for patients to follow or apply.
- *1 point*: Communication fails to center on the patient, being overly formal, academic, or irrelevant, and lacking clarity, practicality, or approachability.

C Pragmatic Rules

Our pragmatic rule acquisition followed a rigorous and multi-phase process. First, we conducted a systematic review of medical literature. Second, we summarized the principles that doctors should follow in communication with patients into three aspects: what format to use for replies, what information to convey, and how the user will feel. Third, we distilled actionable rules from these principles. The pragmatic rules used in this paper are listed below.

- Without listing ideas, reply like a human being(Liu et al., 2024).
- The reply is easy to understand, and the technical terms have relevant explanations(Allen et al., 2023).
- Showing empathy: a comprehensive assessment of an individual’s ability to understand, empathize with, and respond to others’ emotions(Zhang et al., 2024).
- The reply has authenticity and feasibility(Chen et al., 2020).

- Clear structure and logical organization, with natural and smooth transitions between sentences(Jiang et al., 2024).

D LAPI vs. CoT Variants Performance with GPT-4

To demonstrate the distinct advantages of the LAPI framework over Chain-of-Thought (CoT) prompting, particularly in providing actionable advice, we conducted a comparative analysis between LAPI and several enhanced CoT variants. This evaluation aims to verify whether LAPI’s superior performance persists even when CoT prompts are explicitly designed to align more closely with LAPI’s approach, thus clarifying LAPI’s technical contribution. Specifically, we compared LAPI against four CoT variants using GPT-4: (1) Baseline CoT with *“Let’s think step by step”* prompting; (2) CoT-(P.R.), which incorporates pragmatic rules; (3) CoT-(HBM), which integrates HBM theory; and (4) CoT-(HBM+P.R.), which adds pragmatic rules to CoT-(HBM). As shown in Table 3, LAPI consistently outperforms all CoT variants across multiple metrics, reinforcing its distinct technical advantages.

Approach	Flu.	Coh.	Nat.	Emp.	Pat-C.	Rec.	Pre.
CoT	4.00	3.70	3.10	2.35	2.85	0.37	0.08
CoT-(P.R.)	4.00	3.90	3.75	3.65	3.55	0.36	0.07
CoT-(HBM)	4.00	3.85	3.15	2.80	3.05	0.36	0.08
CoT-(HBM+P.R.)	4.00	3.80	3.90	3.60	3.75	0.35	0.07
LAPI	4.00	3.90	3.95	3.85	3.80	0.38	0.07

Table 3: Performance comparison between LAPI and improved CoT prompts using GPT-4.