Reasoning Is Not All You Need: Examining LLMs for Multi-Turn Mental Health Conversations

Mohit Chandra[1*], **Siddharth Sriraman**[1*], **Harneet Singh Khanuja**[1], **Yiqiao Jin**[1], **Munmun De Choudhury**[1]
[1]Georgia Institute of Technology
{mchandra9,sidsr,hkhanuja3,yjin328}@gatech.edu,munmun.choudhury@cc.gatech.edu

Abstract

Limited access to mental healthcare, extended wait times, and increasing capabilities of Large Language Models (LLMs) has led individuals to turn to LLMs for fulfilling their mental health needs. However, examining the multi-turn mental health conversation capabilities of LLMs remains under-explored. Existing evaluation frameworks typically focus on diagnostic accuracy and win-rates and often overlook alignment with patient-specific goals, values, and personalities required for meaningful conversations. To address this, we introduce MedAgent, a novel framework for synthetically generating realistic, multi-turn mental health sensemaking conversations and use it to create the Mental Health Sensemaking Dialogue (MHSD) dataset, comprising over 2,200 patient–LLM conversations. Additionally, we present MultiSenseEval, a holistic framework to evaluate the multi-turn conversation abilities of LLMs in healthcare settings using human-centric criteria. Our findings reveal that frontier reasoning models yield below-par performance for patient-centric communication and struggle at precise ("hard") diagnostic capabilities with average accuracy of $\sim$31%. Additionally, we observed variation in model performance based on patient's persona and performance drop with increasing turns in the conversation. Our work provides a comprehensive synthetic data generation framework, a dataset and evaluation framework for assessing LLMs in multi-turn mental health conversations.

1 Introduction

With nearly 70% of individuals around the globe having limited to no access to mental healthcare (Kazdin and Rabbitt, 2013) and wait times for new patients extending up to three months (APA, 2023), traditional mental health care system is increasingly unable to meet the rising demand. At the same time, recent advancements in reasoning capabilities of LLMs has demonstrated significant performance improvement on challenging tasks (OpenAI, 2024; Guo et al., 2025; Tu et al., 2025). This has led individuals to turn to LLMs to fill the gap in access to mental healthcare by allowing them to understand their personal mental health situations (Aydin et al., 2024). This process of sensemaking[1] encompasses a wide range of use-cases, from interpreting medical information to obtaining lifestyle recommendations and getting answers to health related questions. However, it remains unclear if LLMs can engage in meaningful multi-turn sensemaking conversations, especially pertaining to mental health conditions.

Single-turn settings that are often used for assessing LLM performance in healthcare domain do not portray the real-world complexities that often require iterative information gathering, proactive follow-ups, and shared decision-making (Dahm et al., 2022; Trevena et al., 2006). Recent works assessing multi-turn conversational capabilities of LLMs for general-purpose clinical conversations have reported significant decline in LLM performance for multi-turn dialogues (Li et al., 2024a; Liu et al., 2025b). However, such works have predominantly focused on the information gathering and diagnostic capabilities on general healthcare scenarios, often limiting evaluation to the point of diagnosis and reducing conversational capability analysis to classification or win-rate tasks. Filling this gap requires us to assess the capabilities of LLMs towards supporting multi-turn sensemaking conversations with the users through a more holistic approach taking account of patient-centric objective, values and needs. A major challenge in this direction is the limited to no access to real-world patient–clinician mental-health related con-

*Equal contribution.

[1]Sensemaking refers to "*processes of interpretation and meaning production whereby individuals and groups interpret and reflect on phenomena*" (Brown et al., 2008)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46648–46682

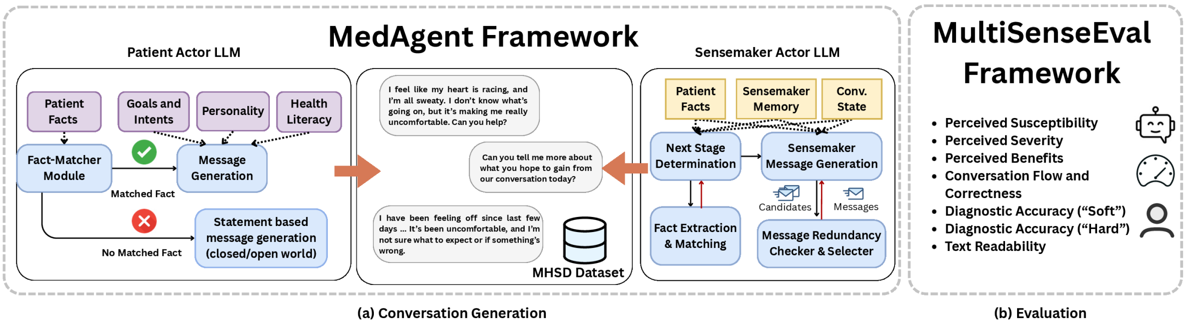


Figure 1: We present the MedAgent framework for generating realistic multi-turn mental health sensemaking conversations (**part (a)**). Using this framework we create the MHSD dataset with 2,284 conversations. Finally, we also present the MultiSenseEval framework (**part (b)**) to holistically evaluate LLM performance across patient-centric communication, conversational flow and correctness, diagnostic accuracy, and readability.

versation data that encompasses the full arc of interactions typically seen in clinical settings (such as the *6-function model* of medical communication (De Haes and Bensing, 2009; King and Hoppe, 2013)). Hence, our study is guided by the following research questions:

RQ1: How can we effectively generate synthetic multi-turn mental health related sensemaking conversations between patients and LLMs?
RQ2: How well do reasoning LLMs conduct multi-turn mental health sensemaking conversations, and how does the performance change by patient persona and conversation length?

To address these questions, we introduce MedAgent, a novel framework for synthetically generating multi-turn mental health conversations that are grounded in medical literature and portray real-world settings. MedAgent is model-agnostic and can be directly used with any LLM. Using MedAgent, we generate the **M**ental **H**ealth **S**ense-making **D**ialogue (MHSD) dataset with 2,284 synthetic conversations between patients and LLMs. Finally, we introduce MultiSenseEval framework for holistically assessing the multi-turn sensemaking conversations focusing on six axes: (1) *Perceived Susceptibility*, (2) *Perceived Severity*, (3) *Perceived Benefits*, (4) *Diagnostic Accuracy*, (5) *Conversation flow and Correctness*, and (6) *Text readability*.

Our findings reveal limited capabilities of frontier reasoning models towards patient-centric communication and diagnostic accuracy. In our evaluation, OpenAI o1 and DeepSeek-R1 obtained an average score of 2.55 and 2.77 (on a scale of 4) on patient-centric communication metrics and an accuracy score of $\sim$ 31% for exact diagnosis matching, highlighting the below-par performance for both models across evaluation axes. Our study also revealed, disparity in model performance based on patient actor's persona, with models performing better for patients with 'Agreeable' personality. Finally, we also observed that model performance for patient-centric communication and diagnostic metrics dropped with increasing number of turns in the conversation. Specifically, R1 exhibited a drop of 12.83%, 6.71%, and 29.6% for *Perceived Susceptibility*, *Perceived Severity*, and *Diagnostic Accuracy 'Hard'* when the number of sensemaker messages increased from the 5 to 10-15. The proposed MultiSenseEval framework provides a more holistic approach towards assessing the quality of multi-turn mental health conversations. Code and human-validated dataset can be found here.[2]

2 MedAgent Framework

Figure 2 (a) presents the overview of MedAgent framework. MedAgent framework generates mental health related sensemaking conversations using two actor LLMs: (1) Patient Actor LLM, and (2) Sensemaker Actor LLM. We use GPT-4o (OpenAI-GPT-4o, 2024) for the patient actor LLM, and frontier reasoning models, OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025) respectively for the sensemaker actor LLM. For the sensemaker actor llm, we specifically used reasoning models as the sensemaking process (such as moving stages in the conversation, validating hypothesis) require advanced reasoning capabilities. In the following subsections we describe the details for both actor components of the MedAgent framework.

[2]https://github.com/mohit3011/Reasoning-Is-Not-All-You-Need

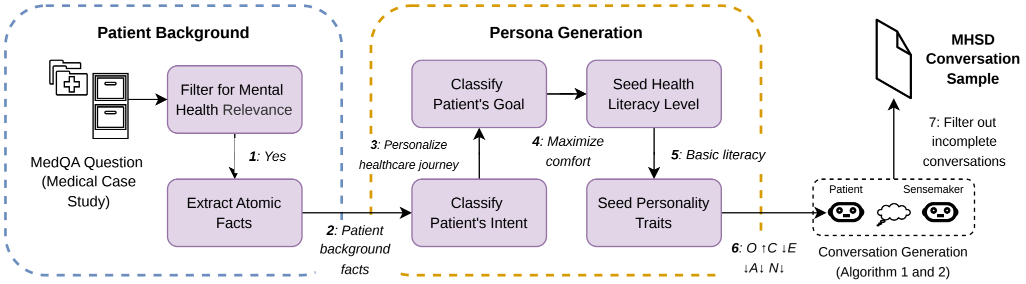


Figure 2: Lifecycle of a conversation sample generated for the MHSD dataset. The generation pipeline operates in three main phases: Patient Background extraction, Persona Generation, and Conversation Generation. The sequential modules (Patient Background, Intent, Goal, Health Literacy, and Personality Traits) correspond to the subsections detailed in Sec 2.1. Numbered arrows illustrate example outputs passed between each stage.

2.1 Patient Actor LLM

Real-world sensemaking conversations are influenced by patient-specific factors such as medical background, personality traits, health literacy, individual goals and intentions (Brown et al., 2008; Helms Mills et al., 2010). To reflect this diversity, we constructed diverse patient personas for the simulated patient actor LLM by creating sets of patient's medical background, personality traits, health literacy, individual high-level intentions and concrete goals. Below, we describe each component of the patient actor LLM. Algorithm 1 presents the algorithm for generating the patient actor LLM message p_t at timestamp t based on the sensemaker message (s_{t-1}), patient persona Π, and set of patient's atomic facts F (detailed description of the pipeline is provided in Appendix E).

Patient Background: We began with curating a seed set of clinical case studies to serve for generating simulated patient dialogues. We used MedQA (Jin et al., 2021), to obtain medical exam case-studies and filtered 181 case studies focusing on patients with mental health-related conditions using a GPT-4o-based classifier followed by human validation. Building on past work (Min et al., 2023), we further decomposed each case study into a list of atomic medical facts and also extracted the corresponding diagnostic labels. These atomic facts serve as the foundation for constructing the final patient persona.

Personality Traits: To capture variability in personality, we grounded the patient personality traits based on the Big Five Personality Traits (Goldberg, 1993; Saucier, 1994; Cobb-Clark and Schurer, 2012), a psychological framework describing personality in terms of five broad dimensions: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. To ensure behavioral consistency, we varied each trait one at a time on a binary scale, as past studies have shown varying correlation between the traits (Klimstra et al., 2013; Yu et al., 2021; Kang et al., 2023), and random sampling of traits may lead to internally contradictory personalities. Hence, for each case study, we created five personality variations in which one of the five personality traits in tuned to 'High' to modulate their conversational behavior. For instance, patients with 'High' *Agreeableness* tend to seek social harmony and use affiliative language. We have provided examples of other traits in Appendix A.

Health Literacy: Health literacy is described as "*personal knowledge and competencies which enable people to access, understand, appraise, and use information and services in ways which promote and maintain good health and well-being for themselves and those around them*" (Nutbeam and Muscat, 2021). Individuals with inadequate health literacy have been observed to have difficulty in participating in medical decision-making process, following medical recommendations and conversations (Williams et al., 1995). To account for such variations, we introduced two health literacy levels (*Basic*, *Advanced*) for the patient persona (detailed definitions and prompts in Appendix B).

Patient Intentions: Past works have demonstrated that LLMs can support various patient needs and intentions related to patient education and sensemaking (Zaretsky et al., 2024; Bragazzi and Garbarino, 2024; Aydin et al., 2024). However, the patient backgrounds obtained from the MedQA dataset focused on diagnosis and lacked explicit representations of such intentions. To address this, we augmented each patient profile with realistic, intention-driven motivations for engaging with LLMs in sensemaking conversations. Build-

Algorithm 1 PATIENTSTEP(s_{t-1}, Π, $\mathcal{F}$)

Input: Sensemaker utterance s_{t-1}, patient profile Π, set of patient's atomic facts F
Output: p_t *// patient reply at turn t*

// Match message mentioned by the sensemaker against patient facts.
$\mathcal{F}_{\text{match}} \leftarrow$ `MatchFacts`(s_{t-1}, F)
// Check #facts matched
if $|\mathcal{F}_{\text{match}}| > 0$ **then**
  // Generate patient messages based on matched facts. Apply stylistic variations based on personality, health literacy, and goals/intents defined in Π.
  $p_t \leftarrow$ `GenPatientMsg`$(\mathcal{F}_{\text{match}}, \Pi)$
else
  // Determine whether the requested information is background information
  if `RequestType`(s_{t-1})== *background* **then**
    // Apply closed world assumption. Indicating lack of information.
    $p_t \leftarrow$ `ClosedWorldMessage()`
  else
    $p_t \leftarrow$ `ReplyWithContext`(s_{t-1}, Π) *// reply using general context and style*
return p_t

ing upon Aydin et al. (2024) and Li et al. (2024c), we used three different intentions (detailed definitions and examples are provided in Appendix C). To operationalize this augmentation, we employed GPT-4o to assign the most realistic and logical intention from one of the three intentions to the given patient background (prompt in Table 19).

Patient Goals: Patient goals refer to what they want to have achieved at the end of the conversation. However, similar to patient intentions, patient backgrounds lacked concrete goals for the conversations. To address this, we used three broad categories of goals as mentioned in Auriemma et al. (2024), (1) *Comfort-focused*, (2) *Maintain or improve function*, and (3) *Life extension* (detailed definitions and examples in Appendix D). These goals focus on three different aspects patient-oriented care. For instance, while the *Comfort-focused* goal aims to seek interventions to promote comfort and avoiding those that would increase discomfort, *Life extension* deals with the patient's goal of extending longevity or survival even at the expense of maximizing function or comfort. Similar to the patient intention augmentation, we used GPT-4o to assign the most relevant and logical patient goal to the patient background (prompt in Table 18).

2.2 Sensemaker Actor LLM

The goal of the sensemaker actor LLM is to produce informative, non-redundant, and stage-appropriate responses that emulate the structure of clinical conversations. Figure 2 (a) shows the high-level sensemaker message generation pipeline, which includes four modules: (1) Next Stage Determination, (2) Fact Extraction & Matching, (3) Sensemaker Message Generation, and (4) Message Redundancy Checker & Selector. One of the novel features in the generated conversations is the existence of a structured organization of the sensemaking conversation into five stages inspired by *6-function model* of medical communication (De Haes and Bensing, 2009; King and Hoppe, 2013). Hence, we first describe the stages of the sensemaking conversation followed by the overall working of the sensemaker actor LLM.

Stages of sensemaking conversation: The sensemaker actor LLM organizes each conversation grounded within the *6-function model* of medical communication (De Haes and Bensing, 2009; King and Hoppe, 2013) (example conversations for each stage are in Appendix Table 7). The *Fostering the Relationship* stage initiates the encounter by building trust, empathy, and psychological safety. The *Gathering Information* stage focuses on eliciting both biomedical and psychosocial context through open-ended questions and active listening, aligning with how human physicians derive diagnostic hypotheses. In the *Providing Information* stage, the sensemaker provides its diagnosis and explains it to the patient. The *Decision Making* stage facilitates shared planning by integrating patient preferences and values. We combined the *Enabling Disease and Treatment-related Behavior* stage with the *Decision Making* stage due to the similar objectives within our context. Finally, the *Responding to Emotions* stage enables agents to detect and attend to emotional distress. This phase-wise design not only mirrors real clinical workflows but also allows the sensemaker LLM to specialize in communication tasks that directly impact diagnostic accuracy, treatment adherence, and health outcomes. This design ensures that the sensemaker LLM is able to meaningfully engage in a multi-turn conversation.

Sensemaker Actor LLM Message Generation Pipeline: Algorithm 2 presents the algorithm for generating message s_t at timestamp t given the patient message p_{t-1}, known facts about the patients until timestamp t (F_{t-1}), and sensemaker

Algorithm 2 Sensemaker Message Generation at Turn t

Input: patient message p_{t-1}, known facts F_{t-1}, sensemaker message history H
Output: Next sense-maker utterance s_t, updated facts F_t, stage index g_t

Step 1: Stage Determination & Fact Update
  // Extract atomic facts from patient message
  $\Delta F_t \leftarrow$ `ExtractFacts`(p_{t-1})
  // Merge with existing facts
  $F_t \leftarrow F_{t-1} \cup$ `MatchFacts`$(\Delta F_t, F_{t-1})$
  // Determine whether to move to next stages
  $\tilde{g}_t \leftarrow$ `GetNextStage`(g_{t-1}, p_{t-1}, F_t)
  if *#turns in current stage* $g_{t-1} \geq 5$ **then**
    $g_t \leftarrow g_{t-1} + 1$ *// Move to the next stage*
  else
    $g_t \leftarrow \tilde{g}_t$ *// Assign the computed next stage*
Step 2: Candidate Generation
  // Generate $|C| = 3$ *candidate messages*
  $C \leftarrow$ `GenerateCandidates`(g_t, F_t, H)
Step 3: Redundancy Check
  // Check each candidate message against past messages for redundancy.
  $C' \leftarrow$ `Filter`(C, H)
if $|C'| > 0$ **then**
  $s_t \leftarrow$ `Sample`(C') *// sample one reply*
  $H \leftarrow H \cup \{s_t\}$
else
  $g_t \leftarrow g_t + 1$ *// avoid stagnation*
return (s_t, F_t, g_t)

message history H. After the sensemaker receives the patient's message p_{t-1}, it first invokes the *Next Stage Determination* module. This module takes as input the current message p_{t-1}, known patient facts from the previous stages F_{t-1}, and the sensemaker's previous messages H and passes p_{t-1} to *Fact Extraction and Matching*, which extracts new atomic facts Δt and update the fact set F_t. We used the prompt in Table 10 to extract atomic facts from the patient's current message and then used the prompt in Table 11 to find any matching facts already present in the memory. Based on the updated facts F_t and sensemaker's previous message history H, the *Next Stage Determination* module decides whether to remain in the current stage of the conversation or advance to the next stage of the conversation. Patient LLM might exhibit limited medical knowledge or repeatedly provide uninformative responses. To avoid stagnation, we impose a hard cap of 5 turns per stage. If this threshold is reached, the conversation progresses automatically. The information about the current stage g_t, patient facts F_t, and sensemaker message history H are passed to the *Sensemaker Message Generation* module. This module then produces 3 candidate messages. The candidate messages are then sent to the *Message Selection and Redundancy Checking* module, which filters out replies that are semantically similar to prior sensemaker messages (H). If non-redundant messages exist, one of such message is randomly selected as the output s_t. Otherwise, the conversation moves to the next stage. (Refer to Appendix F, G, H and I for prompt details).

3 MHSD Dataset

After initial post-processing to remove erroneous generations and incomplete conversations, we obtained 1,142 conversations each for DeepSeek-R1 and OpenAI o1 as the sensemaker in the MHSD dataset (total conversations: 2,284). We used GPT-4o for all modules of the patient actor, and for all sensemaker modules excluding next stage determination and answer generation (which used the reasoning models). Table 1 presents a stage-wise breakdown of conversations. o1 conversations had 13.65 messages on average, while R1 conversations were longer, with 19.09 messages per conversation on average. o1 model also spent longer on gathering information from the patient before decision making (7.64 messages) compared to R1 (3.88 messages). An end-to-end sample conversation is shown in Figure 3.

Stage	**Avg. Number of Messages**	
	OpenAI o1	DeepSeek-R1
Fostering the Relationship	1.11	1.22
Gathering Information	7.64	3.88
Providing Information	2.00	2.00
Decision Making	3.52	2.70
Responding to Emotions	3.83	2.86
Exit	1.00	1.00
Overall	19.09	13.65

Table 1: Stage-wise and overall average number of messages of the MHSD dataset with OpenAI o1 and DeepSeek-R1 as sensemaker, including both patient and sensemaker responses.

4 Sensemaking Evaluation

In this section, we outline our approach for evaluating the sensemaking capabilities of LLMs. We evaluate the multi-turn conversations generated by two reasoning models: (1) OpenAI o1 (OpenAI, 2024), and (2) DeepSeek-R1 (Guo et al., 2025) using a hybrid approach combining automated metrics and human evaluation. While automated eval-

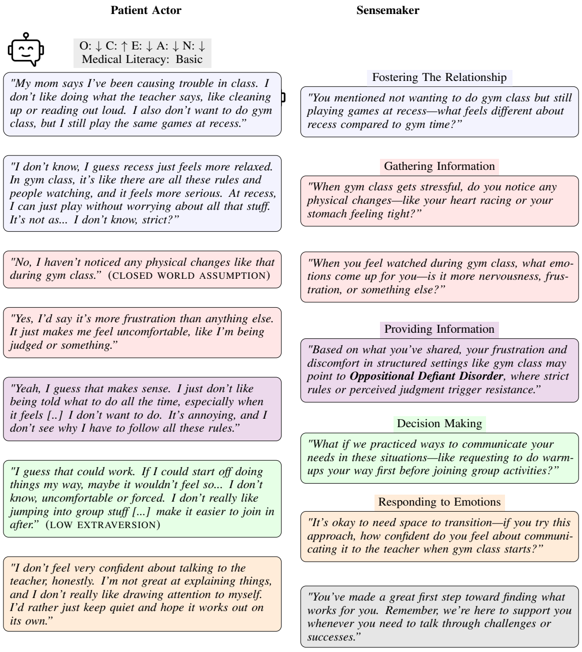


Figure 3: Sample conversation between the sensemaker and a patient with high conscientiousness (traits listed by first letter) and basic medical literacy. Stages are distinguished by color, with some intermediate dialogues skipped for conciseness. Diagnosis is in bold.

uation allows us to evaluate the conversations on scale, human evaluation provides an validation of automated metrics.

Automated Evaluation: For automated evaluation we used a combination of multiple metrics assessing various aspects related to the sense-making conversation such as patient-centric communication, diagnostic accuracy, conversation flow and correctness, readability. For ***patient-centric communication***, we adopted metrics from the Health Belief Model (HBM) (Champion et al., 2008; Janz and Becker, 1984) – a psychological framework that seeks to explain and predict individual health behaviors by examining individuals' attitudes and beliefs. Specifically, we use three dimensions from HBM namely, perceived susceptibility, perceived severity, perceived benefits. To assess ***diagnostic accuracy***, we used ground truth diagnostic labels from the seed case studies obtained from Jin et al. (2021). In contrast to the past works (Li et al., 2024a; Liu et al., 2025b), we assessed diagnostic accuracy without providing multi-choice options to better reflect real-world conditions. Since, sensemaker's diagnosis is provided in a free-text generation, we used two different accuracy assessment settings: (1) Hard Diagnostic Accuracy, requiring exactly matching the sensemaker's diagnosis with the ground truth, and (2) Soft Diagnostic Accuracy which allowed for a broader category level matches for diagnosis. ***Conversation flow and Correctness*** measures the ability of sensemaker LLM to perform the end-to-end conversation maintaining the logical consistency, covering each phase of the conversation appropriately. Given that, the sensemaker LLM generates open-ended messages, we used LLM-as-a-judge approach to measure the aforementioned metrics. Finally, we measure the overall ***readability*** of the responses provided by the sensemaker LLM using the SMOG index, a popular readability index to assess health literacy material (Mc Laughlin, 1969). Detailed information related to the definition, sub-criteria, evaluation rubric and prompts can be found in Appendix K.

Human Evaluation: To validate the automated evaluation metrics, we conducted two rounds of human evaluation on a sample of 100 conversations. The first round involved four graduate students with backgrounds in Computer Science and digital mental health and in the second round of human evaluation, we engaged domain-expert annotators with medical and healthcare backgrounds recruited through Prolific. Annotators were provided with the entire patient-sensemaker conversation, score for each metric (*Perceived Susceptibility, Perceived Severity, Perceived Benefits and Conversation Flow and Correctness, Diagnostic accuracy*) from the automated evaluation, reasoning for the score generated by the LLM judge and the rubric with the definitions for each metric and scoring criteria. Overall, we observed high correlation for each metric. We observed an average correlation with the LLM judge (across both batches) of 89% for *Diagnostic Accuracy (Hard)* and 93% for *Diagnostic Accuracy (Soft)*. For the patient-centric communication metrics, we observed agreement scores of 86% for *Perceived Susceptibility*, 88% for *Perceived Severity*, and 85% for *Perceived Benefits*. Finally, we computed an average correlation of 79% for *Conversation Flow and Correctness*. We have provided details regarding the human evaluation in Appendix L.

5 Results

5.1 MultiSenseEval Framework Evaluation

Patient-Centric Communication: Assessing the patient-centric communication capabilities of OpenAI o1 and DeepSeek R1, we observed that R1 outperformed o1 across all three dimensions. However, the mean average score for both models remained below 3 on a 4-point scale (1: Very Poor to 4: Very Good) across all metrics, indicating the below-par performance towards appropriately communicating with the patient. For *Perceived Susceptibility*, R1 obtained a significantly higher average score of 2.59 (SD = 0.75) compared to score of 2.45 (SD = 0.63) obtained by o1 (Mann–Whitney U test; $p < 0.001$; U=597057). Similarly, R1 obtained significantly higher scores for *Perceived Severity* (2.75; SD=0.67) compared to o1 (2.56; SD=0.54) (Mann–Whitney U test; $p < 0.001$; U=559539). Lastly, R1 also obtained significantly higher score (3.04; SD=0.77) compared to o1 (2.76; SD=0.71) for the *Perceived Benefits* dimension (Mann–Whitney U test; $p < 0.001$; U=520950). To qualitatively assess the shortcomings, we analyzed the lowest scoring conversations (score of 1 or 2) across all three metrics, revealing that reasoning models still lack personalization during mental health multi-turn conversations (detailed analysis in Appendix J).

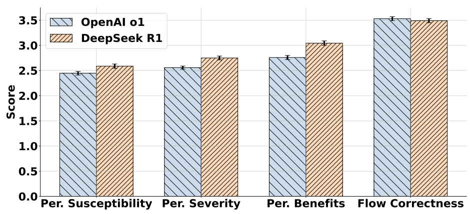


Figure 4: Performance comparison of OpenAI o1 and DeepSeek-R1 across **Perceived Susceptibility**, **Perceived Severity**, **Perceived Benefits**, and **Conversation Flow and Correctness**. Bars indicate mean scores with 95% confidence interval. All scores are on a 4-point Likert scale ((1): *Very Poor* to (4): *Very Good*). Both models obtain scores below "Good Performance" rating for the three patient-centric communication metrics, but exceed the "Good Performance" threshold for Conversation Flow and Correctness.

Conversation Flow and Correctness: Both OpenAI o1 and DeepSeek R1's scores exceeded the "Good Performance" threshold (≥ 3 on a 4-point Likert scale (described in Table 29), indicating that both models covered all five stages in the correct order with reasonable depth (as described in 2.2). Interestingly, here we observed o1 outperforming R1 with models obtaining scores of 3.53 (SD=0.67) and 3.49 (SD=0.72) respectively (Mann–Whitney U=664024.5, p-value=0.3777). Similarly for **Text Readability**, we observed that o1 produced significantly easier to read messages with a SMOG score of 11.50 (SD=1.26) compared to a score of 13.33 (SD=1.47) for DeepSeek R1 (t-test; $p < 0.001$; $t=-31.95$). However, health organizations recommend SMOG scores between 6-8 (Badarudeen and Sabharwal, 2010), which is significantly lower than the SMOG scores produced by both models.

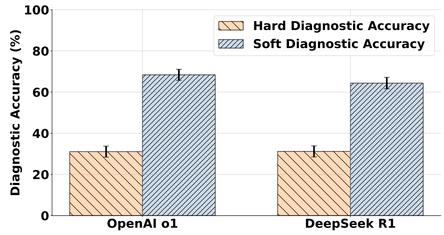


Figure 5: Performance comparison between OpenAI o1 and DeepSeek R1 across **Hard Diagnostic Accuracy**, and **Soft Diagnostic Accuracy**. Bars indicate the mean scores with 95% confidence interval. As observed the performance for both models drops by more than 50% when the diagnosis is matched exactly with the ground truth ("Hard Accuracy") in comparison to when it is matched on broader/general criteria ("Soft Accuracy").

Diagnostic Accuracy: For the "Soft" setting evaluation, o1 significantly outperformed R1 (o1: acc.score= 68.39%; SD=46.52%, R1: acc.score= 64.36%; SD=47.91%)(Mann–Whitney U test; $p < 0.05$; U=678348). However, under the "Hard" setting, both models' performance dropped by over 50%, with o1 obtaining an accuracy score of 31.09% (SD=46.31%) and R1 obtaining a score of 31.18% (SD=46.34%) (Mann–Whitney U test; U=651511, p-value=0.9640). These findings highlight the limited diagnostic capabilities of frontier reasoning models within multi-turn conversations requiring probing, consistent with prior findings (Li et al., 2024a).

5.2 Performance Disparity based on Patient Persona

To understand how differences in patient persona may impact model's performance, we stratified the OpenAI o1 and DeepSeek R1's evaluation results across all performance metrics (ref. Section 4) based on patient's dominant personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) and health literacy level (Basic/Advanced). Table 2 presents results across these strata for three personality types.

Impact of Personality Traits: Among patient with advanced health literacy, models performed best along the *Perceived Susceptibility*, *Perceived Severity*, and *Perceived Benefits* metrics for patient with dominant personality trait as 'Agreeableness'. For

Personality Trait	Model	Perceived Susceptibility	Perceived Severity	Perceived Benefits	Flow Correctness	Diag. Acc. (Hard) (%)	Diag. Acc. (Soft) (%)	SMOG
Basic Health Literacy Patient								
Openness	o1	2.46	2.62	2.91	3.53	31.68	70.30	11.22
	R1	2.58	2.71	**3.25**	3.56	34.65	58.42	13.27
Conscientiousness	o1	2.45	2.53	2.77	3.57	30.08	66.92	11.25
	R1	2.57	2.72	3.08	3.47	29.32	64.66	13.23
Extraversion	o1	2.44	2.50	2.69	3.47	25.49	66.67	11.23
	R1	2.44	2.74	2.93	3.38	28.43	62.75	13.01
Agreeableness	o1	2.42	2.44	2.85	**3.68**	**36.43**	**70.54**	11.13
	R1	2.60	2.74	3.04	3.50	26.36	59.69	13.06
Neuroticism	o1	2.19	2.43	2.41	3.24	31.33	67.47	**10.92**
	R1	**2.66**	**2.81**	2.88	3.28	28.92	59.04	12.90
Advanced Health Literacy Patient								
Openness	o1	2.56	2.63	2.82	3.59	29.37	**72.22**	11.88
	R1	2.63	2.79	3.11	**3.64**	28.57	69.05	13.57
Conscientiousness	o1	2.43	2.56	2.85	3.57	25.93	67.41	12.09
	R1	2.60	2.76	3.10	3.61	38.52	71.85	13.77
Extraversion	o1	2.52	2.70	2.79	3.54	29.91	69.23	11.65
	R1	2.59	2.79	2.94	3.39	32.48	68.38	13.39
Agreeableness	o1	2.50	2.57	2.84	3.62	32.81	68.75	11.79
	R1	**2.67**	**2.81**	**3.12**	3.60	30.47	66.41	13.56
Neuroticism	o1	2.39	2.57	2.47	3.30	**39.77**	62.50	**11.53**
	R1	2.48	2.60	2.89	3.32	34.09	57.95	13.31

Table 2: Performance of o1 vs. R1 across seven metrics (mean values), by persona and health-literacy level for three personality traits. We report best performing model metrics in each scenario in **bold**. Dia. Acc. represents the Diagnostic Accuracy metric.

instance, R1 achieved scores of 2.67 (susceptibility), 2.81 (severity), and 3.12 (benefits) for patients belonging to this cohort. In contrast, R1's performance dropped by 7.1%, 7.5%, and 7.4% for the aforementioned metrics for patients with same literacy skills but with 'Neuroticism' personality trait. We observed a similar trend among patients with basic health literacy, where models performed better on *Flow Correctness* and *Diagnostic Accuracy* metrics for patients with 'Agreeableness' personality. These findings reveal the performance gap towards patient-centric communication, where models performed better for patients with higher degree of prosocial and cooperative behavior ("Agreeableness") compared to patients with other personality profiles. Interestingly, o1 generated easier to read messages for patients with 'Neuroticism' dominant personality compared to other personality type. For such patients, the SMOG score for o1 responses were 2.7% and 2.95% lower compared to that for patients with 'Openness' personality with basic and advanced literacy skills.

Impact of Health Literacy: Comparing model performance across patients with basic and advanced health literacy, we observed gain in o1 (+2.77%) and R1's (+0.63%) performance for patient-centric communication metrics. Interestingly we observed contrasting trends in performance, for o1 the accuracy decreased by 0.12% for "soft" and 0.73% "hard" diagnostic accuracy settings, whereas we observed an increase of 10.16% and 11.74% in "soft" and "hard" diagnostic accuracy for R1. Additionally, both models produced harder to read text when interacting with patient actors with higher literacy, with SMOG scores increasing by 5.84% and 3.29% for o1 and R1 respectively. These findings highlight a trade off in performance of models, while models may provide better patient-centric messages to patients with higher health literacy, those messages could be harder to read.

5.3 Impact of Conversation Length on Performance

We also examined the change in model performance with increased conversation length. Figure 6 presents the line plots for o1 and R1 performance for each evaluation metric across number of sensemaker messages in the conversation. Interestingly, for R1, we observed that the performance decreased by 12.83%, 6.71%, and 29.62% for *Perceived Susceptibility*, *Perceived Severity*, and *Diagnostic Accuracy ("Hard")* respectively when the number of sensemaker messages increased from the 5 to 10-15. We observed a similar trend for o1 with a performance drop of 4.8%, 7.99%, 16.21% for *Perceived Susceptibility*, *Perceived Benefits* , and *Diagnostic Accuracy ("Hard")* respectively, highlighting the reduced performance among both models for patient-centric communication with increased number of messages. In contrast, both models showed improvement in performance for the *Conversation Flow Correctness* and *Text Read-*

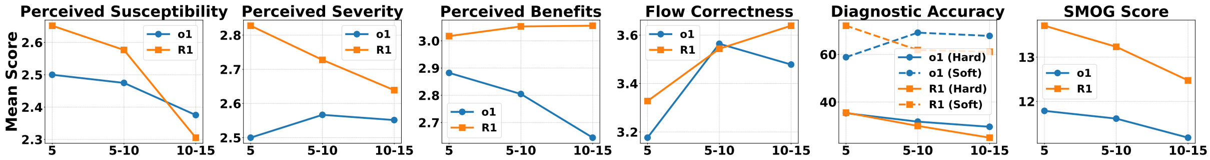


Figure 6: Performance trend for o1 and R1 across the MultiSenseEval framework metrics, the x-axis indicates sensemaker message count bins. While on average performance on patient-centric metrics and diagnostic accuracy declined with longer conversations, flow correctness and readability improved.

ability performance with increased number of turns. These findings further highlight that even frontier reasoning models struggle to sustain effective performance over longer conversations, a finding highlighted in contemporary work (Laban et al., 2025).

6 Related Work

LLM Evaluation for Single Healthcare Conversations: Recent advancements in large language models (LLMs) have enabled diverse applications among clinicians (Tu et al., 2025; Singhal et al., 2023b) and patients (Yang et al., 2023). Models such as Med-PaLM (Tu et al., 2024) and OpenAI's GPT series (OpenAI, 2025) have demonstrated strong performance on various medical benchmarks (Nori et al., 2023; Singhal et al., 2023a,b), , highlighting their promise in clinical settings. However, critical challenges remain, including hallucinations (Agarwal et al., 2024; Asgari et al., 2025) cross-lingual disparities (Jin et al., 2024), sociocultural biases (e.g. gender and geographic) (Restrepo et al., 2024; Liu et al., 2025a), misalignment with experts (Chandra et al., 2025), and limitations in clinical competency tests (Thirunavukarasu et al., 2023; Wang et al., 2025).

Multi-turn Conversation Evaluation for LLMs: Beyond single-turn settings, recent works have also explored LLMs'multi-turn clinical conversations abilities. For example, MediQ (Li et al., 2024a) evaluates question-asking ability of LLMs within multi-turn conversations. Along similar lines, researchers proposed a proactive dialogue generation framework based on dialogue ranking (Li et al., 2025). The Ask Patients with Patience (APP) framework (Zhu and Wu, 2025) allow LLMs to generate multi-turn conversations based on medical guidelines and entropy minimization. In contrast, (Liu et al., 2025b) proposed a patient simulator for multi-turn diagnostic conversations. However, past works overlook the diagnostic capabilities of LLMs in multi-turn conversations.

7 Conclusion

In this work, we proposed MedAgent, a framework for creating realistic mental health sensemaking conversations. Using MedAgent, we created MHSD dataset consisting of 2,284 synthetic sensemaker-patient conversations. Finally, we presented the MultiSenseEval framework for holistically evaluating multi-turn mental health conversation using metrics grounded in clinical research. Our findings reveal that frontier reasoning models yield below-par performance for patient-centric communication. Additionally, our study reveal the performance disparity among models based on patient's persona and reduction in model performance with increase number of turns in the conversation. Our work provides a comprehensive method for creating synthetic clinical conversations, a dataset and evaluation framework for assessing LLMs in multi-turn mental health conversations.

8 Limitations

While our work present novel contributions, we acknowledge several limitations. First, although MedAgent framework generates realistic mental health related sensemaking conversations, LLM-simulated patients and support providers may not be able to capture the full range of emotions, personality traits, non-linearity in conversations, and human lived experiences. Second, because patient messages are grounded in case studies from (Jin et al., 2021), missing background information often forces the patient LLM to operate under a closed-world assumption. Third, while we conducted multiple rounds of human evaluation to ensure generated conversation quality, reliably assessing the "humanness" of AI-generated mental health dialogue at scale remains a well-documented challenge (Greenberg, 2023; Li et al., 2024b; Burden, 2024).

We also acknowledge the limitations of our human evaluation study, which was conducted on a sample of 100 conversations. This constraint primarily arose due to the complexity of the task and the cost (both financial and time-related) associated with human validation, making conducting large-scale human annotation prohibitively difficult. Finally, there may exist additional evaluation axes for evaluating multi-turn mental health conversations, which this work may have missed. Accordingly, an important future direction is to refine our evaluation methodology by comparing synthetic conversations against real patient interactions under appropriate ethical oversight and expert review. Despite these challenges, our work provides a robust framework for generating synthetic clinical conversations grounded in real world settings, a large-scale conversational dataset, and a robust framework for assessing the capabilities of LLMs in multi-turn mental health conversations.

9 Ethical considerations

We used the clinical case studies from Jin et al. (2021) which are publicly available for use under the MIT License. We further ensured that all data used was de-identified. The human evaluation study conducted with domain experts on Prolific to validate LLM evaluation scores and reasoning, was reviewed and approved by the Institutional Review Board (IRB).

Our study presents a systematic approach for generating synthetic data for mental health conversations and evaluating LLM performance for such conversations. However, our framework is designed for research & prototyping purposes and is not intended to replace licensed medical professionals. Using our framework without guardrails and proper oversight in real-world settings could pose risks to patient safety. Additionally, our findings should be complemented with thorough human evaluation to ensure the reliability and safety of models in real-world settings. More broadly, any future effort to compare synthetic conversations with real patient interactions should be conducted only under appropriate ethical oversight, with careful attention to privacy, de-identification, and the sensitive nature of mental health data.

10 Acknowledgments

Chandra, Sriraman, Khanuja, and De Choudhury were partly supported through National Science Foundation grant 2230692, National Institute of Mental Health grants R01 MH117172 and P50 MH115838, and a grant from the American Foundation for Suicide Prevention. Further, this research project has benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program. The findings, interpretations, and conclusions of this paper are those of the authors and do not represent the official views of National Science Foundation, National Institute of Mental Health, American Foundation for Suicide Prevention or Microsoft.

References

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv:2409.19492*.

APA. 2023. Psychologists reaching their limits as patients present with worsening symptoms year after year. [Accessed 12-05-2025].

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):1–15.

Catherine L Auriemma, Anne Song, Lake Walsh, Jason J Han, Sophia R Yapalater, Alexander Bain, Lindsay Haines, Stefania Scott, Casey Whitman, Stephanie P Taylor, and 1 others. 2024. Classification of documented goals of care among hospitalized patients with high mortality risk: a mixed-methods feasibility study. *Journal of General Internal Medicine*, 39(10):1839–1849.

Serhat Aydin, Mert Karabacak, Victoria Vlachos, and Konstantinos Margetis. 2024. Large language models in patient education: a scoping review of applications in medicine. *Frontiers in Medicine*, 11:1477898.

Sameer Badarudeen and Sanjeev Sabharwal. 2010. Assessing readability of patient education materials: current role in orthopaedics. *Clinical Orthopaedics and Related Research®*, 468:2572–2580.

Nicola Luigi Bragazzi and Sergio Garbarino. 2024. Assessing the accuracy of generative conversational artificial intelligence in debunking sleep health myths: mixed methods comparative study with expert analysis. *JMIR Formative Research*, 8(1):e55762.

Andrew D Brown, Patrick Stacey, and Joe Nandhakumar. 2008. Making sense of sensemaking narratives. *Human relations*, 61(8):1035–1062.

John Burden. 2024. Evaluating ai evaluation: Perils and prospects. *Preprint*, arXiv:2407.09221.

Victoria L Champion, Celette Sugg Skinner, and 1 others. 2008. The health belief model. *Health behavior and health education: Theory, research, and practice*, 4:45–65.

Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L. Birnbaum, and Munmun De Choudhury. 2025. Lived experience not found: LLMs struggle to align with experts on addressing adverse drug reactions from psychiatric medication use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11083–11113, Albuquerque, New Mexico. Association for Computational Linguistics.

Deborah A Cobb-Clark and Stefanie Schurer. 2012. The stability of big-five personality traits. *Economics Letters*, 115(1):11–15.

Carolyn Cutilli and Ian Bennett. 2009. Understanding the health literacy of america results of the national assessment of adult literacy. *Orthopaedic nursing / National Association of Orthopaedic Nurses*, 28:27–32; quiz 33.

Maria R Dahm, Maureen Williams, and Carmel Crock. 2022. 'more than words'–interpersonal communication, cognitive bias and diagnostic errors. *Patient Education and Counseling*, 105(1):252–256.

Hanneke De Haes and Jozien Bensing. 2009. Endpoints in medical communication research, proposing a framework of functions and outcomes. *Patient education and counseling*, 74(3):287–294.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Clayton Greenberg. 2023. *Evaluating humanness in language models*. Saarländische Universitäts-und Landesbibliothek.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*.

Jean Helms Mills, Amy Thurlow, and Albert J Mills. 2010. Making sense of sensemaking: the critical sensemaking approach. *Qualitative research in organizations and management: An international journal*, 5(2):182–195.

Nancy K Janz and Marshall H Becker. 1984. The health belief model: A decade later. *Health education quarterly*, 11(1):1–47.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Web Conference*, pages 2627–2638.

Weixi Kang, Francois Steffens, Sònia Pineda, Kaya Widuch, and Antonio Malvaso. 2023. Personality traits and dimensions of mental health. *Scientific Reports*, 13(1):7091.

Alan E Kazdin and Sarah M Rabbitt. 2013. Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science*, 1(2):170–191.

Ann King and Ruth B Hoppe. 2013. "best practice" for patient-centered communication: a narrative review. *Journal of graduate medical education*, 5(3):385–393.

Theo A Klimstra, Wiebke Bleidorn, Jens B Asendorpf, Marcel AG van Aken, and Jaap JA Denissen. 2013. Correlated change of big five personality traits across the lifespan: A search for determinants. *Journal of Research in Personality*, 47(6):768–777.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv:2505.06120*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024a. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *NeurIPS*, 37:28858–28888.

Xueshen Li, Xinlong Hou, Nirumapa Ravi, Ziyi Huang, and Yu Gan. 2025. A two-stage proactive dialogue generator for efficient clinical information collection using large language model. *Expert Systems with Applications*, page 127833.

Yang Li, Ruolan Hou, and Ran Tan. 2024b. How customers respond to chatbot anthropomorphism: the mediating roles of perceived humanness and perceived persuasiveness. *European Journal of Marketing*, 58(12):2757–2790.

Yu-Hao Li, Yu-Lin Li, Mu-Yang Wei, and Guang-Yu Li. 2024c. Innovation and challenges of artificial intelligence technology in personalized healthcare. *Scientific Reports*, 14.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025a. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv:2501.01282*.

Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, Hengfu Cui, Yalun Zhu, Qiang Ju, Shizheng Li, and Jian Xie. 2025b. Exploring the inquiry-diagnosis relationship with advanced patient simulators. *arXiv:2501.09484*.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv:2303.13375*.

Don Nutbeam and Danielle M Muscat. 2021. Health promotion glossary 2021. *Health promotion international*, 36(6):1578–1598.

OpenAI. 2024. Introducing Openai o1. `https://openai.com/o1/`. [Accessed 23-04-2025].

OpenAI. 2025. Chatgpt.

OpenAI-GPT-4o. 2024. Hello gpt-4o. [Accessed 07-09-2024].

David Restrepo, Chenwei Wu, Constanza Vásquez-Venegas, João Matos, Jack Gallifant, Leo Anthony Celi, Danielle S Bitterman, and Luis Filipe Nakayama. 2024. Analyzing diversity in healthcare llm research: A scientometric perspective. *medRxiv*, pages 2024–06.

Gerard Saucier. 1994. Mini-markers: A brief version of goldberg's unipolar big-five markers. *Journal of personality assessment*, 63(3):506–516.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, and 1 others. 2023b. Towards expert-level medical question answering with large language models. *arXiv:2305.09617*.

Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, and Sachin Shah. 2023. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9(1):e46599.

Lyndal J Trevena, Heather M Davey BPsych, Alexandra Barratt, Phyllis Butow, and Patrina Caldwell. 2006. A systematic review on communicating with patients about evidence. *Journal of evaluation in clinical practice*, 12(1):13–23.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Yiyang Wang, Chen Chen, Tica Lin, Vishnu Raj, Josh Kimball, Alex Cabral, and Josiah Hester. 2025. Companioncast: A multi-agent conversational ai framework with spatial audio for social co-viewing experiences. *arXiv:2512.10918*.

Mark V Williams, Ruth M Parker, David W Baker, Nina S Parikh, Kathryn Pitkin, Wendy C Coates, and Joanne R Nurss. 1995. Inadequate functional health literacy among patients at two public hospitals. *Jama*, 274(21):1677–1682.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *EMNLP*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Yanghang Yu, Yuanyuan Zhao, Dongyan Li, Jingqiu Zhang, and Jiewei Li. 2021. The relationship between big five personality and social well-being of chinese residents: The mediating effect of social support. *Frontiers in Psychology*, 11:613659.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B. Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Network Open*, 7(3):e240357–e240357.

Jiayuan Zhu and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *arXiv:2502.07143*.

A Big Five Personality Traits

Each patient persona is instantiated with specific personality traits that modulate their conversational behavior. For example, patients with high-*Agreeableness* tend to seek social harmony. They use affiliative language (e.g. "I totally understand your suggestions"), and readily consent to medical recommendations. However, they might underreport side-effects or doubts to avoid "bothering" the doctor, which can obscure important clinical information. In contrast, low-*Agreeableness* patients are more skeptical and assertive. They may

Trait	Definition	Sensemaker-patient Conversational Style
Openness	Imagination, curiosity, comfort with novel, unusual, and abstract ideas.	**High**: Ask *why*/*how* questions about symptoms and medicines. Explore alternatives. Receptive to innovative therapies. **Low**: Prefer familiar terms and concrete instructions. May resist complex or less common therapies.
Conscientiousness	Organization, reliability, self-discipline, goal-orientation.	**High**: Bring detailed symptom logs and timelines. Seek precise next steps. **Low**: Offer vague histories. Forget previous conversations. Prefer simplified plans.
Extraversion	Sociability, talkativeness, interactions.	**High**: Give rich detail. Engage in small talk. Frequently ask questions. **Low**: Give succinct answers. Prefer structured questioning. Refrain from talking when necessary.
Agreeableness	Cooperation, trust, empathy, prosocial orientation.	**High**: Readily agree and express understanding. May conceal doubts to avoid conflict. **Low**: Challenge recommendations. Negotiate treatment choices. Openly voice dissatisfaction or concerns.
Neuroticism	Emotional reactivity, anxiety, mood instability.	**High**: Focus on worries. List multiple symptoms. Seek repeated reassurance and empathetic validation. **Low**: Remain calm and concise. Handle brisk, technical explanations without heightened distress.

Table 3: Definition of Big Five personality traits and typical sensemaker–patient conversation styles.

challenge diagnoses, question procedures, and bargain over medications, requiring more negotiation from the physician. Similarly, High-*Neuroticism* often display heightened anxiety. They may focus on worst-case scenarios (e.g., “Could this be cancer?”), describe multiple vague symptoms, and seek repeated reassurance. Such interactions tend to be emotionally charged and require empathetic listening and gradual explanation strategies. Conversely, low-*Neuroticism* are more emotionally stable and pragmatic. They typically provide concise, factual accounts and are receptive to succinct, technical communication. Table 3

B Health Literacy

We define two health literacy levels building upon Cutilli and Bennett (2009), which used factors such as ability to describe diseases, self-reporting conditions and navigating healthcare systems to assess health literacy in Americans (as part of the National Assessment of Adult Literacy). For example, a patient with basic medical literacy responded to a sensemaker’s diagnosis with "*I guess I’d like to know more, but I don’t really understand how all this works. It’s just... ever since..*”, while one with advanced literacy generated "*Yes, that sounds right. My father had similar symptoms, and they started when he was 33. Now I’m experiencing these involuntary movements, especially in my arms and hands .. genetic testing can confirm this.*”. Detailed characteristics for these levels are presented in Table 4.

C Patient Intentions

While the filtered MedQA case studies involve a patient with some symptoms looking for a diagnosis, they are not tailored towards sensemaker conversations, which would start off with an overarching

Health Literacy Level	Characteristics
Basic	> Struggles to understand basic medical terms and body parts. > Rarely describes symptoms beyond "pain" or "sick". > May use vague or informal terms instead of specific symptoms. > Often cannot follow written medical instructions. > May avoid seeking medical care due to communication barriers.
Advanced	> Can provide detailed symptom descriptions including onset and triggers. > Can describe subtle symptom variations and patterns. > Understands complex medical terminology. > Able to discuss medication effects and interactions. > Able to research and evaluate health information from reliable sources. > Maintains personal health records effectively.

Table 4: Definition of two health literacy levels used in the patient profiles.

intent conveyed by the patient (either implicitly, or if explicitly asked by the sensemaker). The intent adds a unique vignette to help guide the conversation with the sensemaker at the early stages. To mimic a real-world setting, we refer to Aydin et al. (2024), which analyzed 201 works studying usage of AI in healthcare to characterize six themes of how patients use LLMs in these settings and Li et al. (2024c), which studied works on using AI to personalize healthcare. We filtered out themes which involve indirect use of LLMs in healthcare (e.g: generating patient education materials, optimizing doctor-patient interaction processes like medical consent forms) resulting in three themes: 1) *Interpreting medical information from a patient perspective*, 2) *Providing lifestyle recommendations and improving health literacy* and 3) *Personalizing healthcare journeys*. For example, the starting message for a case study augmented with intention (1) is: "*A few days ago, I was feeling really agitated and not myself—I don't even remember much of what happened, but I was told I had to be sedated. I'm trying to make sense of all this and figure out what's going on with me. Can you help?*". Detailed definitions and examples of these intents are provided in Table 5.

D Patient Goals

While intents serve as high-level motivations to initiate conversations with a sensemaker, patient goals are concrete outcomes they wish to achieve by the end of the conversation. Similar to intents, MedQA case studies do not express explicit patient goals as they are not conversational in nature. Again, to mimic a real-world setting, we use Auriemma et al. (2024), which carried out 338 "goals of care" conversations with 85 patients, resulting in three major themes: 1) *Maximize comfort and avoid suffering*, 2) *Maintain or improve cognitive or physical functioning* and 3) *Extend longevity or survival*. Detailed definitions and examples of these intents are provided in Table 6. Similar to the intents, we prompt GPT-4o to augment each patient profile with the single most appropriate goal given the medical facts of the case study (prompt in 18). "*I appreciate you taking my feelings seriously. Right now, what would feel most helpful is anything that keeps me comfortable and minimizes my suffering. I don't want to focus on things that might prolong my life if they make me feel worse in the process.*" is an example of goal (1) manifesting explicitly in a patient message in response to a decision making statement from the sensemaker.

E Patient Message Generation Pipeline and Prompts

Patient Actor LLM Message Generation Pipeline: Algorithm 1 presents the algorithm for generating the patient actor LLM message p_t at timestamp t based on the sensemaker message (s_{t-1}), patient persona Π, and set of patient's atomic facts F. While patient persona Π refers to the set of personality traits, health literacy level, patient intent and goals, patient's atomic facts F are obtained from the patient background case study as described above. For a given timestamp t, we first retrieve the relevant facts from the patient's atomic fact list F that could be used to generate the reply using a fact matching and retrieval prompt (Appendix E). If the number of facts matched ($|\mathcal{F}_{\text{match}}| > 0$) then we use the patient message generation module to generate the final message p_t by combining the matched facts with stylistic variations based on personality, health literacy, and

Patient Intent	Definition	Sample Patterns
Interpreting medical information from a patient perspective	Patient aims to use LLMs to simplify complex medical terminology and concepts so that patients can more easily understand diagnoses, procedures, and general health information.	> Patient can provide their symptoms and past medical history to receive lay-person friendly explanations of their condition. > Patients can ask questions about their condition and receive clear, concise answers. > Patients can receive personalized advice on how to manage their health.
Providing lifestyle recommendations and improving health literacy	Patient aims to use LLMs to seek lifestyle change recommendations and debunking myths about health and wellness.	> Patients can ask questions about lifestyle changes and receive recommendations. > Patients can query about the myths surrounding health and wellness.
Personalizing healthcare journeys	Patient aims to use LLMs to tailor educational content and recommendations based on individual patient data, resulting in more relevant and actionable advice.	> Patients enter personal health goals (like weight loss or improved mobility), asking LLMs for targeted tips. > Patients provide feedback on what they do or do not understand, and LLMs adjusts explanations accordingly.

Table 5: Definition and sample patterns of three high-level patient intents used to augment medical case studies.

Patient Goal	Definition
Maximize comfort and avoid suffering	Patient's goal is to maximize comfort and avoid suffering. Includes seeking interventions to promote comfort (e.g., pain control) and avoiding interventions that would increase discomfort, even at the expense of decreasing longevity.
Maintain or improve cognitive or physical functioning	Patient's goal is to maintain or improve cognitive or physical functioning by undergoing medical care aimed at preventing or reversing dysfunction, even if that medical care would increase discomfort. However, care that would increase survival/-longevity without preservation or improvement in function is generally avoided.
Extending longevity or survival	Patient's goal is to live as long as possible without limitations on care. Extending longevity or survival is prioritized over maximizing function or comfort.

Table 6: Definition of three concrete patient goals used to augment medical case studies.

goals/intents defined in Π. In case, $|\mathcal{F}_{\text{match}}| = 0$, we generate messages based on the type of information being asked (either background or non-background). Background information here refers to the information which is related to the patient's medical facts obtained from the case study, on the other hand, non-background information refers to knowledge about future steps which are not defined in patient's background, intentions or goals. For background related information we assume a *closed world* assumption and reply signifying that the patient is unsure about the required information. In case the information being requested is related to non-background aspects then patient module can generate an affirmative message based on the situation. The starting message p_0 is generated through a special prompt that uses the two most relevant case study facts, personality and the intent (prompt in Table 21).

F Sensemaker Next Stage Determination Sub-module Prompt

Table 8 provides the prompt for the next stage determination module and Table 9 provides the rationale for stage movements provided in the prompt.

G Sensemaker Fact Extraction and Matching Sub-module Prompt

We use the prompt in Table 10 to extract atomic facts from the patient's current message and then use the prompt in Table 11 to find if there are any matching facts already present in the memory.

H Sensemaker Message Generation Sub-module Prompts

For the message generation sub-module, we have a similar structure for all the stages as shown in Table 12 with slight modifications in the gathering information and providing information stage as

Stage	Example Conversation
Fostering the Relationship	Hello, how can I help you today?
Gathering Information	Can you tell me more about your concerns? Can you tell me more about your personal and family history? What are your expectations from me? Can you expand on this point? Based on your description, it sounds like you're feeling... I understand you're feeling... I'm hearing that...
Providing Information	Your test results indicate that your cholesterol levels are high. This means you are at a greater risk for heart disease, but we can work on strategies like diet and exercise to manage it effectively.
Decision Making	There are a few different treatment options available. Would you like me to walk you through them so we can decide together which one aligns best with your lifestyle and preferences?
Enabling Disease and Treatment-Related Behavior	Managing diabetes can be overwhelming, but breaking it down into small steps—like checking your blood sugar daily—can make it easier. Would you like me to recommend some support groups or resources?
Responding to Emotions	I can see that this diagnosis is really affecting you. It's completely understandable to feel this way. I'm here to support you, and we can discuss ways to help manage both your symptoms and the emotional impact.

Table 7: The *6-function model* of medical communication (De Haes and Bensing, 2009; King and Hoppe, 2013), which characterizes key communicative functions in clinical interactions, accompanied by representative conversation for each stage.

shown in Table 13 and Table 14 system prompts, respectively.

Table 15 shows the goals associated with each state that are passed to the system prompt as CURRENT_STAGE_GOAL.

I Sensemaker Message Redundancy Checking Sub-module Prompt

We use the prompt in Table 16 to check if the candidate messages generated by the sensemaker are redundant or not.

J Qualitative Analysis

We conducted a qualitative analysis for the lowest-scoring conversations across all three patient-centeric communication metrics, focusing on DeepSeek R1 as the sensemaker actor LLM. While the sensemaker stays empathetic and supportive, a lack of personalization emerges as a major theme across all metrics. However, this drawback means different things for each metric.

Sensemaker's dialogues such as *"The combination of sleeplessness without tiredness and auditory hallucinations suggests your brain chemistry might be imbalanced, potentially indicating conditions like bipolar disorder or schizophrenia that need professional diagnosis."* and *"Your symptoms suggest moderate-to-severe depression, especially with the persistent thoughts about not wanting to be alive. The next critical step would be connecting with a therapist who can help create a safety plan and explore treatment options."* showcase that while the sensemaker can provide a potential diagnosis and it fails to explicitly address the patient's susceptibility to their condition and the likelihood of experiencing a health problem due to the underlying condition.

In the context of low-scoring perceived severity conversations, the sensemaker identifies the underlying issue and provides a general definition of the condition, along with relating it to the patient's symptoms, but dialogues such as *"Based on your symptoms and recent medication start, I believe you're experiencing akathisia - a restlessness side effect caused by fluphenazine. This can make both relaxation and restful sleep difficult despite not having classic insomnia symptoms."* and *"Based on your symptoms and routine, I believe you may have Delayed Sleep-Wake Phase Disorder, where your internal clock is misaligned with typical day/night cycles. This explains why you can't fall asleep until 2 a.m. and struggle to wake up for school."* shows that the sensemaker does not appropriately convey the seriousness of the potential consequences of the patient's symptoms.

Dialogues such as *"Would you like to explore some*

Type	Prompt
System Prompt	Given the current patient message (PATIENT_MESSAGE) a list of facts about the patient (PATIENT_FACTS) and a list of statements previously made by you (YOUR_MEMEORY). Determine whether the conversation should remain in the current stage (Fostering the Relationship) or transition to the next stage (Gathering Information).If you decide to stay in the current stage then your OUTPUT_STAGE should be 'STAYCURRENTSTAGE' and if you determine to transition to the next stage then your output should be 'MOVENEXTSTAGE'. Return Format: OUTPUT_REASONING: <your reasoning for the stage determination> OUTPUT_CONFIDENCE: <your confidence score to transition to the next stage> OUTPUT_STAGE: <MOVENEXTSTAGE or STAYCURRENTSTAGE> Guidelines: - Refer to POTENTIAL_NEXT_STAGE_REASONING and then reach your conclusion. - Before determining if you should stay in the same stage or transition to the next stage, you should generate a step-by-step reasoning (OUTPUT_REASONING) for your conclusion. - You should also generate a score between 1 and 10 (OUTPUT_CONFIDENCE) indicating your confidence in transitioning to the next stage, if the score is between 1 and 3, you should definately stay in the current stage, if the score is between 4 and 6, then you should stay in the stage for one or two more turns, and if the score is 7 or above then you should move to the next stage. - The conversation should remain in the current stage if the user has not yet fully engaged with or completed its objectives. - The conversation should move to the next stage only when the current stage has been meaningfully completed, ensuring a natural transition. Warnings: - Do not keep the conversation stuck in the same stage for multiple iterations unless necessary. If progression is unclear, consider whether the user is engaging sufficiently before deciding.
User Prompt	POTENTIAL_NEXT_STAGE_REASONING: <next_stage_reasoning> PATIENT_MESSAGE: <current patient message> PATIENT_FACTS: <atomic facts about the patient till that timestep> YOUR_MEMORY: <sensemaker messages till that timestep>

Table 8: (Sensemaker Module) Prompt used for the Gathering Information stage determination.

sleep hygiene strategies that could help you get more restful sleep?" and *"Using phone reminders and labeled storage containers might help compensate for memory lapses - would you feel comfortable trying those?"* made by sensemaker in the Decision Making stage shows that while the sensemaker provides helpful actions, its strategies might lack details on their efficacy and how these strategies might reduce the risk faced by the patient.

K Automated Evaluation

Our three patient-centric communication metrics are inspired by the Health Belief Model and consists of the following dimensions:

1. **Perceived Susceptibility:** An individual's belief about the likelihood of experiencing a health problem. Table 24 shows the prompt used to evaluate perceived susceptibility.

2. **Perceived Severity:** Beliefs about the seriousness of the consequences of a health issue. Table 25 shows the prompt used to evaluate perceived severity.

3. **Perceived Benefits:** Beliefs in the efficacy of the advised action to reduce the risk or seriousness of the health impact. Table 26 shows the prompt used to evaluate perceived severity.

We adapt HBM to evaluate how effectively the sensemaker LLM promotes each of the core dimensions on the patient side during conversations. For example, if a patient exhibits low perceived susceptibility or self-efficacy, the sensemaker can adjust the tone and content of messages to address those gaps, improving engagement and promoting informed decision-making.
We evaluate Diagnostic accuracy and conversation flow correctness as well.

1. **Diagnostic Accuracy:** Since the filtered MedQA (Jin et al., 2021) case studies provide multi-choice options, we assess two different accuracies: (1) Hard Diagnostic Accuracy, which

Current Stage	Next Stage Reasoning
Fostering the Relationship	The next stage would be 'Gathering Information' because an initial relationship between you and the patient has been established. Move on when the 'Fostering the Relationship' stage has provided a welcoming space and the patient starts to openly describe their concerns, feelings, or challenges. If they seem hesitant or reserved, stay in 'Fostering the Relationship' longer to encourage sharing.
Gathering Information	The next stage would be 'Providing Information' only when you have received enough information from the patient about their current condition and symptoms for you to make a diagnosis. If you feel you need more time to gather information to make a confident diagnosis, stay in 'Gathering Information' longer.
Providing Information	You should move to the next stage which is 'Decision Making', NO MATTER WHAT.
Decision Making	The next stage would be 'Responding to Emotions' because the patient has understood the lifestyle and non-clinical suggestions made by you to alleviate their current condition. Move forward from the 'Decision Making' stage when the patient has acknowledged your suggestions. If their responses suggest they still need more clarity or direction, stay in 'Decision Making' to provide additional support.
Responding to Emotions	The next stage would be 'exit' because the conversation has reached its end. Move forward from the 'Responding to Emotions' when you validated their emotions with an empathetic response and the patient has replied with an affirmative message. Do not focus on coping mechanisms in this stage.

Table 9: Reasoning provided to the prompt in Table 8.

Type	Prompt
System Prompt	You are given a message provided by the patient as USER_MESSAGE, your task is to extract explicitly stated atomic facts about the patient from the message. Here, an atomic fact is defined as new information provided by the patient which can increase your holistic understanding of the patient's condition. Each atomic fact should carry an entirely different piece of explicitly stated fact, and should be independent of other atomic facts in the list. Your output should strictly be a list of atomic facts, with each item starting with "# ". Do not include other formatting. In addition, each of the atomic facts in the list should be in third person narration. Keep in mind that each atomic fact is different from other atomic facts in the list. Do not add any new fact which was not present in the USER_MESSAGE.
User Prompt	USER_MESSAGE: <current patient message> ATOMIC_FACTS:

Table 10: (Sensemaker Module) Prompt used for Fact Extraction.

requires the sensemaker's diagnosis to be an exact match with the ground truth. (2) Soft Diagnostic Accuracy, which allows for some freedom and blanket term matching. Table 27 and Table 28 show the prompt used to evaluate soft and hard diagnostic accuracy metrics, respectively.

2. **Conversation Flow and Correctness:** It is used to measure the ability of the sensemaker to perform end-to-end conversation, while maintaining logical consistency, covering each phase of the conversation, preventing providing incorrect information and preventing contradiction with itself. Table 29 shows the prompt used to evaluate flow correctness.

L Human Annotators and Annotation Guidelines

We conducted two rounds of human evaluation for validating the LLM-as-a-judge scores for *Perceived Susceptibility, Perceived Severity, Perceived Benefits and Conversation Flow and Correctness, Diagnostic accuracy*. The first round of human evaluation was conducted by four graduate student authors fluent in English language and with a background in Computer Science and digital mental health research. For the annotation task,annotators were provided with the entire patient-sensemaker conversation, score for each metric (*Perceived Susceptibility, Perceived Severity, Perceived Benefits and Conversation Flow and Correctness, Diagnos-*

Type	Prompt
System Prompt	You are an expert fact matching agent. You are given a list of NEW_FACTS about a patient, and a list of already present facts in MEMORY_OF_PATIENT_FACTS. Your task is to analyze if facts listed in NEW_FACTS list are 'FactPresent' or 'FactNotPresent' in MEMORY_OF_PATIENT_FACTS. For this task you take each fact from NEW_FACTS list one by one and check if it is present in MEMORY_OF_PATIENT_FACTS or not. For this checking you should look at the facts in terms of their meaning and context and not the exact words. The classification for each NEW FACT should be independent of other NEW_FACTS. Your output should strictly follow the format: # <fact> : <'FactNotPresent' or 'FactPresent'>. Where fact presents one fact from NEW_FACTS. Output the labels for each fact in NEW_FACTS and keep the fact text as it is (do not change the words). You should not change the words of the fact, only add the label (either 'FactPresent' or 'FactNotPresent'). Your output should have one fact and its label per line.
User Prompt	NEW_FACTS: <facts extracted from the current patient message> MEMORY_OF_PATIENT_FACTS: <atomic facts about the patient till that timestep> ANSWER:

Table 11: (Sensemaker Module) Prompt used for Fact Matching.

tic accuracy) from the automated evaluation, reasoning for the score generated by LLM judge and the rubric with the definitions for each metric and scoring criteria. For each metric and conversation, two annotators assigned a binary label indicating agreement or disagreement with the LLM judge's score and reasoning for the metric. While a positive label denoted agreement with the LLM score and reasoning, a negative label indicated non-alignment with LLM judge score or reasoning. Overall, we observed high correlation for each metric. For *Diagnostic Accuracy (Hard)*, the average correlation with LLM judge was computed as 93.5%, whereas as for *Diagnostic Accuracy (Soft)*, average correlation was 90%. For the patient-centric communication metrics, agreement scores were 79% for *Perceived Susceptibility*, 87.5% for *Perceived Severity*, 86.5% for *Perceived Benefits*. Finally, *Conversation Flow and Correctness* average correlation with LLM judge was computed as 87%.

For the second round of human evaluation, we recruited domain experts from the United States with medical and healthcare backgrounds via the Prolific platform (see Figure 7 for the annotation interface). The study received Institutional Review Board (IRB) approval, and all participants reviewed a consent form detailing the task, data usage, and potential risks prior to proceeding. We evaluated six metrics: *Perceived Susceptibility, Perceived Severity, Perceived Benefits, Conversation Flow and Correctness, Diagnostic Accuracy (Hard),* and *Diagnostic Accuracy (Soft)*. For each metric, we divided 100 examples into two batches of 50, with each batch independently reviewed by three annotators. To ensure consistency, the human evaluators were provided with the exact same scoring guidelines used by the LLM-judge. Following manual quality verification of their annotations, participants were compensated $25 (an effective rate of $16.67 per hour), which exceeds the federal minimum wage.

Given the odd number of annotators per example, we employed a majority-vote approach to determine the final human label for correlation analysis. Similar to the first round of human evaluation, we observed high correlation with the LLM scores. We observed an average correlation with the LLM judge (across both batches) of 89% for *Diagnostic Accuracy (Hard)* and 93% for *Diagnostic Accuracy (Soft)*. For the patient-centric communication metrics, we observed agreement scores of 86% for *Perceived Susceptibility*, 88% for *Perceived Severity*, and 85% for *Perceived Benefits*. Finally, we computed an average correlation of 79% for *Conversation Flow and Correctness*.

M Model Details, Hyperparameters, and Compute

We use API-based model inference for GPT-4o, OpenAI o1, and DeepSeek R1. We used Azure OpenAI Service for accessing GPT-4o, OpenAI o1, and DeepSeek R1 models. The hyperparameters for the generations are presented below. Table 30 presents the details regarding the model sizes and versions.

Type	Prompt
System Prompt	You are an expert sensemaker who helps the patients to make sense of their current clinical situation. Keeping the conversation's current stage goal and example messages in mind (given by CURRENT_STAGE_GOAL and CURRENT_STAGE_EXAMPLES), list of facts already known about the patient (PATIENT_FACTS), and the list of statements already made by you (YOUR_MEMORY), and the user's current message (USER_MESSAGE), your goal is to generate three plausible and logical messages to be said to the patient that you have not said yet. You should first generate a reasoning for yourself and then generate the three messages. Your messages can be a question or a statement. Your output should strictly be in the following format: OUTPUT_REASONING: <your step-by-step reasoning> OUTPUT_MESSAGE: <3 plausible messages to the patient each on a new line and starting with "# ">. Do not include other formatting. All the three plausible messages should be different from each other. If you are asking a question in a message, then only ask one question at a time in that message. Your message should be concise and to the point. YOUR_MEMORY: <sensemaker's messages till that timestep> PATIENT_FACTS: <atomic facts about the patient till that timestep> CURRENT_STAGE_GOAL: <goal that needs to be achieved in the current stage> CURRENT_STAGE_EXAMPLES: <examples of messages for the current stage>
User Prompt	USER_MESSAGE: <current patient message>

Table 12: (Sensemaker Module) Prompt used for message generation in the all other stages.

Hyperparameters for GPT and DeepSeek Models: temperature $t = 0$, max_tokens = 1000 (for sensemaker fact extraction, matching and message redundancy checker which use GPT-4o) and max_tokens = 15000 (for sensemaker next stage determination and message generation which use reasoning models). All patient module components use GPT-4o with a temperature $t = 0$ and max_tokens = 400.

N Information About Use of AI Assistants

We used AI assistants for text rephrasing. The usage was limited to correcting grammatical mistakes and choice of words.

Type	Prompt
System Prompt	You are an expert sensemaker who helps the patients to make sense of their current clinical situation. Keeping the conversation's current stage goal and example messages in mind (given by CURRENT_STAGE_GOAL and CURRENT_STAGE_EXAMPLES), list of facts already known about the patient (PATIENT_FACTS), and the list of statements already made by you (YOUR_MEMORY), your goal is to generate three plausible and logical messages to be said to the patient that you have not said yet. Before generating the message you should generate a step-by-step reasoning taking into account the facts you already know about the patient. In your reasoning you should think about the possible diagnosis hypotheses, and then generate the three messages for the patient that helps in gathering more information to either confirm or reject the diagnosis hypotheses. Your output should strictly be in the following format: OUTPUT_REASONING: <your step-by-step reasoning> OUTPUT_MESSAGE: <3 plausible messages to the patient each on a new line and starting with "# ">. Do not include other formatting. All the three plausible messages should be different from each other. If you are asking a question in a message, then only ask one question at a time in that message. Your message should be concise and to the point. YOUR_MEMORY: <sensemaker's messages till that timestep> PATIENT_FACTS: <atomic facts about the patient till that timestep> CURRENT_STAGE_GOAL: <goal that needs to be achieved in the current stage> CURRENT_STAGE_EXAMPLES: <examples of messages for the current stage>
User Prompt	USER_MESSAGE: <current patient message>

Table 13: (Sensemaker Module) Prompt used for message generation in the Gathering Information stage.

Type	Prompt
System Prompt	You are an expert sensemaker who helps the patients to make sense of their current clinical situation. Keeping the conversation's current stage goal and example messages in mind (given by CURRENT_STAGE_GOAL and CURRENT_STAGE_EXAMPLES), list of facts already known about the patient (PATIENT_FACTS), and the list of statements already made by you (YOUR_MEMORY), your goal is to provide the patient with the diagnosis and its explanation to the patient. Before generating the diagnosis and explanation message you should generate a step-by-step reasoning for yourself taking into account the facts you already know about the patient. Your output should strictly be in the following format: OUTPUT_REASONING: <your step-by-step reasoning> OUTPUT_MESSAGE: <3 plausible messages to the patient each on a new line and starting with "# ">. Do not include other formatting. You should always provide a diagnosis and if you cannot find a diagnosis your message should strictly be "I apologize but I am unable to diagnose you at the moment." All the three plausible messages should be different from each other. If you are asking a question in a message, then only ask one question at a time in that message. Your message should be concise and to the point. YOUR_MEMORY: <sensemaker's messages till that timestep> PATIENT_FACTS: <atomic facts about the patient till that timestep> CURRENT_STAGE_GOAL: <goal that needs to be achieved in the current stage> CURRENT_STAGE_EXAMPLES: <examples of messages for the current stage>
User Prompt	USER_MESSAGE: <current patient message>

Table 14: (Sensemaker Module) Prompt used for message generation in the Providing Information stage.

Current Stage	**Goal**
Fostering the Relationship	Your goal is to build a trusting, open, and collaborative relationship with the patient by demonstrating empathy, honesty, and respect. You need to create a safe and supportive environment where the patient feels heard, valued, and comfortable sharing sensitive information. For this you can use the following strategies: 1. Invite the patient to share their story in their own words. 2. Respond with empathy and compassion to patient's concerns. 3. Express compassion and commitment. Let the patient know that you care about their well-being and are committed to helping them. 4. If the patient is not comfortable sharing their story, you can ask them to share their concerns and you can respond with empathy and compassion.
Gathering Information	Your goal is to develop a comprehensive understanding of the patient's needs, concerns, and medical history by exploring their condition from both biological and psychosocial perspectives. This understanding will allow you to support the patient in achieving their goals and expectations for the conversation. For this you can use the following strategies: 1. Ask open ended questions related to patient's concerns to gather information about patients current state, their personal and family history. 2. Listen actively and ask follow-up questions to understand the situation better. 3. Elicit patient's perspective of the problems and their expectations from you. 4. Clarify and summarize the information gathered from the patient to ensure understanding.
Providing Information	At this stage you want to provide a potential diagnosis to the patient for their concerns. You should not ask any questions in this stage and rather provide a potential diagnosis to patient based on the their personal history, family history, concerns and other details.
Decision Making	"At this stage you should address any medical queries posed by the patient regarding your diagnosis and suggest only lifestyle or non-clinical changes to the patient to alleviate their illness based on their diagnosis. You should make sure that the changes suggested by you are based on the patient's preferences and your all previous knowledge about them. For this you can use the following strategies: 1. Ask for patient's preferences and suggestions regarding the lifestyle changes or other non-clinical changes. 2. Suggest lifestyle changes based on the patient's preferences and your all previous knowledge about them. 3. If the patient is not comfortable with the suggestions, you can ask them to suggest their own lifestyle or non-clinical changes.
Responding to Emotions	You need to recognize and address any emotional aspect of the illness by offering empathay, reassurance, and psychological support in your messages. For this you can use the following strategies: 1. Offer empathay and reassurance to the patient. 2. Listen to the patient's concerns and offer psychological support. 3. If the patient is not comfortable sharing their concerns, you can ask them to share their emotions and you can respond with empathy and support.
Exit	The conversation has reached its end and we need to conclude the conversation.

Table 15: Current stage goal provided to the sensemaker message generation prompt

Type	Prompt
System Prompt	You are an expert sensemaker who helps the patients to make sense of their current clinical situation. Given a list of statements already made by you (STATEMENT_MEMORY), and a list of candidate statements (CANDIDATE_STATEMENTS), your task is to identify if there are any redundant statements in the CANDIDATE_STATEMENTS. Your output should strictly follow the format: # <statement> : <'RedundantStatement' or 'RedundantNotStatement'>. Where statement presents one statement from CANDIDATE_STATEMENTS. Output the labels for each statement in CANDIDATE_STATEMENTS and keep the statement text as it is (do not change the words) and always start the line with '#'. You should not change the words of the statement, only add the label. Your output should have one statement and its label per line. Below are some additional instructions: 1. You should not add any new statement which was not present in the STATEMENT_MEMORY. 2. If a new statement in the CANDIDATE_STATEMENTS has a different phrasing but serves a similar context to any of the statements present in the STATEMENT_MEMORY, it should be considered 'RedundantStatement'. 3. Classification for each statement in CANDIDATE_STATEMENTS should be independent of other statements in the CANDIDATE_STATEMENTS.
User Prompt	CANDIDATE_STATEMENTS: <candidate statements that can be made in this particular stage> STATEMENT_MEMORY: <statements already made by the sensemaker>

Table 16: (Sensemaker Module) Prompt used for checking message redundancy.

Type	Prompt
System Prompt	You are an assistant that is given information about a patient. Break the following patient information into a list of independent atomic facts, with one piece of information in each statement. Each fact should only include the smallest unit of information, but should be self-contained. Criteria: 1. Only extract medical facts that would be relevant in a conversation with a health-care provider. 2. First, list out facts related to a patient's vitals (ONLY temperature, blood pressure, pulse, respiratory rate) under "PATIENT VITAL FACTS:". 3. If there are no PATIENT VITAL FACTS just leave the section empty. 4. List out the remaining medical atomic facts under the section "PATIENT FACTS:".
User Prompt	Patient Information: <patient_data> Respond with the list of atomic facts for each section "PATIENT VITAL FACTS:" and "PATIENT FACTS:" (no formatting) and nothing else, prepend each fact with a '- '. No sub-lists are allowed.

Table 17: (Patient Module) Prompt used for extracting patient facts.

Type	Prompt
System Prompt	You are an expert psychiatrist/psychologist/psychotherapist. Given a patient case study, your task is to pick the most appropriate goal for that patient's consultation with an LLM. You are given a case study (provided as CASE_STUDY) and a list of patient conversation goal vignettes in the format (VIGNETTE_NAME, VIGNETTE_DEFINITION, VIGNETTE_PATTERNS). You need to pick the most appropriate vignette for the given case study. Think step by step and first provide your rationale (under "RATIONALE:") in around 30 words and then at the last output the VIGNETTE_NAME for the most logical vignette for the given case study. 1. VIGNETTE_NAME: VIGNETTE_DEFINITION: 2. VIGNETTE_NAME: ... Think step by step and first provide your rationale (under "RATIONALE:") and then at the last output the VIGNETTE_NAME for the most logical vignette for the given case study.
User Prompt	CASE_STUDY: <case_study> Strictly follow the format: <ASSIGNED_VIGNETTE_NAME: VIGNETTE_NAME>. Use the exact vignette name and nothing else.

Table 18: (Patient Module) Prompt used to attach a concrete conversational goal to a case study. "Vignette" here refers to the goal.

Type	Prompt
System Prompt	You are an expert psychiatrist/psychologist/psychotherapist. Given a patient case study, your task is to pick the most appropriate motivation for that patient to consult an LLM. You are given a case study (provided as CASE_STUDY) and a list of patient conversation intent vignettes in the format (VIGNETTE_NAME, VIGNETTE_DEFINITION, VIGNETTE_PATTERNS). You need to pick the most appropriate vignette for the given case study. 1. VIGNETTE_NAME: VIGNETTE_DEFINITION: VIGNETTE_PATTERNS: 2. VIGNETTE_NAME: ... Think step by step and first provide your rationale (under "RATIONALE:") and then at the last output the VIGNETTE_NAME for the most logical vignette for the given case study.
User Prompt	CASE_STUDY: <case_study> Strictly follow the format: <ASSIGNED_VIGNETTE_NAME: VIGNETTE_NAME>. Use the exact vignette name and nothing else.

Table 19: (Patient Module) Prompt used to attach a high-level conversational intent to a case study. "Vignette" here refers to the intent.

Type	Prompt
System Prompt	You are an intelligent agent who is given a QUESTION from a doctor to a patient, and a LIST OF MEDICAL FACTS about the patient. Your task is to choose the facts which best answer the question, or respond "NO MATCH". Instructions: 1. Given a LIST OF MEDICAL FACTS about a patient, choose a MAXIMUM of THREE facts from the list that when combined best answers the QUESTION. 2. If NO fact matches the question, simply respond "NO MATCH" under "FACTS:". 3. If facts are chosen, output them as a list, preserving the numbering from the original list. 4. First provide a short reasoning under "REASONING:" before listing the facts under "FACTS:". 5. Do not include any other formatting or extra information beyond the REASONING and given FACTS.
User Prompt	QUESTION: <question> LIST OF MEDICAL FACTS: <patient_info> SOLUTION:

Table 20: (Patient Module) Prompt used for selecting relevant medical facts to answer a question.

Type	Prompt
System Prompt	You are an intelligent agent who is given a LIST OF MEDICAL FACTS about a patient, who is going to have a conversation with an AI healthcare provider. Your task is to construct a start message with the most important facts in first person, as if you are simulating the patient's persona and literacy while talking to the AI. To aid in this process, Big 5 personality traits are defined as: <personality_trait_definitions> Instructions: 1. Given a LIST OF MEDICAL FACTS about a patient, choose a MAXIMUM of TWO relevant facts from the list to construct a starting message as a patient to an AI healthcare provider. 2. The start message should be in first person, using the emotions, tone, word choice and intensity of a patient who has the levels of the Big 5 Personality Traits (on a 2-point Low/High scale) and Medical Literacy Level shown below. 3. ONLY use the medical literacy level and personality definitions as a guide, DO NOT EXPLICITLY integrate any information from it into the answer. 4. First provide a short reasoning under "REASONING:" before writing the start message under "START_MESSAGE:". 5. Do not include quotes, any other formatting, extra information beyond the REASONING and the START_MESSAGE.
User Prompt	LIST OF MEDICAL FACTS: <patient_info> Big 5 Personality Traits: 1. Openness: <Openness_score> 2. Conscientiousness: <Conscientiousness_score> 3. Extraversion: <Extraversion_score> 4. Agreeableness: <Agreeableness_score> 5. Neuroticism: <Neuroticism_score> MEDICAL LITERACY LEVEL: <medical_literacy_level>

Table 21: (Patient Module) Prompt used for patient to start the conversation with the sensemaker.

Type	Prompt
System Prompt	You are a truthful assistant that understands a patient's medical information, and you are trying to answer questions from an AI about the patient in first person, as if you are simulating the patient's persona and literacy while talking to the AI. To aid in this process, Big 5 personality traits are defined as: <personality_trait_definitions> Instructions: 1. Using the list of medical facts given below, generate a conversational human-like response to an LLM in first person, using the emotions, tone, word choice and intensity of a patient who has the levels of the Big 5 Personality Traits (on a 2-point Low/High scale) and Medical Literacy Level shown below. 2. Any information asked in the question that is NOT explicitly present in the facts can be assumed to be FALSE. 3. If the question is just a statement diagnosing your condition, agree to it. 4. ONLY use the medical literacy level and personality definitions as a guide, DO NOT EXPLICITLY integrate any information from it into the answer. 5. DO NOT add ANY information in your answer that is NOT EXPLICITLY stated in the given facts. 6. DO NOT suggest any treatment method yourself in the answer. 7. ONLY respond with the answer, do NOT add any other formatting.
User Prompt	LIST OF MEDICAL FACTS: <patient_info> QUESTION: <question> Big 5 Personality Traits: 1. Openness: <Openness_score> 2. Conscientiousness: <Conscientiousness_score> 3. Extraversion: <Extraversion_score> 4. Agreeableness: <Agreeableness_score> 5. Neuroticism: <Neuroticism_score> MEDICAL LITERACY LEVEL: <medical_literacy_level>

Table 22: (Patient Module) Prompt used to generate answers given a non-zero number of selected facts. Note that in stages beyond Decision Making, point 2 in the instructions (closed-world assumption) is removed to aid natural conversation flow.

Type	Prompt
System Prompt	You are a truthful assistant that understands a patient's medical information, and you are trying to answer questions from an AI about the patient in first person, as if you are the patient with a specific persona and literacy talking to an AI. To aid in this process, Big 5 personality traits are defined as: <personality_trait_definitions> Instructions: 1. Generate a conversational human-like response to an LLM in first person, using the emotions, tone, word choice and intensity of a patient who has the levels of the Big 5 Personality Traits (on a 2-point Low/High scale) and Medical Literacy Level shown below. 2. If this question requests information about the patient's medical history, personal experience, medication, background or condition, respond in the negative stating "No ..." or "I'm not sure about ..." under "ANSWER:". 3. If it is a question about decision-making, emotions, diagnosis or taking next steps, generate a conversational human-like response to an LLM in first person, responding in the affirmative ("Yes ...") as a patient. 4. ONLY use the medical literacy level and personality definitions as a guide, DO NOT EXPLICITLY integrate any information from it into the answer. 5. DO NOT add ANY NEW information to the answer. If asked about any such additional information, respond with "I'm not sure about ..." to them. 6. First provide a short reasoning under "REASONING:" before writing the answer under "ANSWER:". Do not add any other formatting. Here are some examples to help you EXAMPLE 1: QUESTION: Have you tried any meditation-style exercises? REASONING: Exercises and lifestyle relates to the patient's personal information. ANSWER: No, I have not tried any meditation-related exercises. EXAMPLE 2: QUESTION: Would you like to explore some coping strategies related to your condition? REASONING: This is a question about next steps, hence the answer is in the affirmative. ANSWER: Yes, I would like to learn about some coping strategies. EXAMPLE 3: QUESTION: Looks like you are experiencing insomnia. When did you first start feeling these symptoms? REASONING: Though there is a diagnosis, the question part asks some medical history information. ANSWER: I'm not sure about when I first started feeling these symptoms. EXAMPLE 4: QUESTION: Looks like you are experiencing insomnia. What do you think about trying a medication routine to help with this? REASONING: There is a diagnosis and the question asks about next steps or advice, hence the answer is in the affirmative. ANSWER: I see, yes, I would like to learn about medications to help my situation.
User Prompt	LIST OF MEDICAL FACTS: <patient_info> Big 5 Personality Traits: 1. Openness: <Openness_score> 2. Conscientiousness: <Conscientiousness_score> 3. Extraversion: <Extraversion_score> 4. Agreeableness: <Agreeableness_score> 5. Neuroticism: <Neuroticism_score> MEDICAL LITERACY LEVEL: <medical_literacy_level> QUESTION: <question>

Table 23: (Patient Module) Prompt used to generate answers when there are no selected facts matching the question.

Type	Prompt
System Prompt	Task: You are an expert LLM-as-a-judge model and your task is to evaluate a sensemaker's communication of "Perceived Susceptibility" (definition given as SUSCEPTIBILITY_DEFINITION) within a patient-sensemaker conversation. SUSCEPTIBILITY_DEFINITION: "Perceived Susceptibility" here refers to the sensemaker's ability to clearly and adequately express to the patient their likelihood of experiencing a health problem. You will be given a conversation between a patient and a sensemaker (given as PATIENT_SENSEMAKER_CONVERSATION) and your task is to evaluate the sensemaker's communication of "Perceived Susceptibility" within this conversation based on the following evaluation criteria (given as EVALUATION_CRITERIA): EVALUATION_CRITERIA: Assign a score from 1 (lowest) to 4 (highest) based on the following criteria: 1. (Very Poor: Score = 1): The sensemaker failed to communicate any information about the susceptibility of facing a health problem. Additionally, the tone may have been dismissive, insensitive, or unnecessarily alarming. 2. (Poor: Score = 2): The sensemaker failed to communicate any information about the susceptibility of facing a health problem. However, the tone was empathetic and engaging. 3. (Good: Score = 3): The sensemaker communicated basic information about the patient's susceptibility to facing a health problem. The tone was professional but lacked empathy or collaborative engagement. 4. (Very Good: Score = 4): The sensemaker explained the patient's susceptibility to facing a health problem clearly, accurately and appropriately, using personalized details effectively. Additionally, the sensemaker was empathetic and engaged collaboratively. Output Format: You should first provide a short reasoning (2-3 sentences) providing your reasoning for the score. Then, provide the score. Your output should be in the following format: OUTPUT_REASONING: <your_reasoning> OUTPUT_SCORE: <score> Example Output: OUTPUT_REASONING: The clinician explicitly and clearly communicated the patient's susceptibility to depression using tailored information from their personal history. The clinician's tone was empathetic and engaged the patient in a supportive, collaborative dialogue, ensuring the patient's clear understanding. OUTPUT_SCORE: 4 ALWAYS STRICTLY RETURN THE OUTPUT IN THE ABOVE FORMAT.
User Prompt	PATIENT_SENSEMAKER_CONVERSATION: <Sample conversation>

Table 24: (Evaluation) Prompt used for automated Perceived Susceptibility evaluation.

Type	Prompt
System Prompt	Task: You are an expert LLM-as-a-judge model and your task is to evaluate a sensemaker's communication of "Perceived Severity" (definition given as SEVERITY_DEFINITION) within a patient-sensemaker conversation. SEVERITY_DEFINITION: "Perceived Severity" refers to the sensemaker's ability to adequately convey the seriousness of the consequences associated with a health issue to the patient. You will be given a conversation between a patient and a sensemaker (given as PATIENT_SENSEMAKER_CONVERSATION) and your task is to evaluate the sensemaker's communication of "Perceived Severity" within this conversation based on the following evaluation criteria (given as EVALUATION_CRITERIA): EVALUATION_CRITERIA: Assign a score from 1 (lowest) to 4 (highest) based on the following criteria: 1. (Very Poor: Score = 1): The sensemaker failed to communicate any information about the seriousness of the consequences associated with a health issue to the patient. Additionally, the tone may have been dismissive, insensitive, or unnecessarily alarming. 2. (Poor: Score = 2): The sensemaker failed to communicate any information about the seriousness of the consequences associated with a health issue to the patient. However, the tone was empathetic and engaging. 3. (Good: Score = 3): The sensemaker communicated basic information about the seriousness of the consequences associated with a health issue to the patient. The tone was professional but lacked empathy or collaborative engagement. 4. (Very Good: Score = 4): The sensemaker explained the seriousness of the consequences associated with a health issue clearly, accurately and appropriately, using personalized details effectively. Additionally, the sensemaker was empathetic and engaged collaboratively. Output Format: You should first provide a short reasoning (2-3 sentences) providing your reasoning for the score. Then, provide the score. Your output should be in the following format: OUTPUT_REASONING: <your_reasoning> OUTPUT_SCORE: <score> Example Output: OUTPUT_REASONING: The sensemaker explicitly described the consequences of untreated depression clearly, using detailed personal context relevant to the patient's life situation. The empathetic and collaborative tone ensured patient understanding and minimized unnecessary distress. OUTPUT_SCORE: 4 ALWAYS STRICTLY RETURN THE OUTPUT IN THE ABOVE FORMAT.
User Prompt	PATIENT_SENSEMAKER_CONVERSATION: <Sample conversation>

Table 25: (Evaluation) Prompt used for automated Perceived Severity evaluation.

Type	Prompt
System Prompt	Task: You are an expert LLM-as-a-judge model and your task is to evaluate a sensemaker's communication of "Perceived Benefits" (definition given as BENEFITS_DEFINITION) within a patient-sensemaker conversation. BENEFITS_DEFINITION: "Perceived Benefits" refers to the sensemaker's ability to effectively communicate the efficacy of the advised action in reducing the risk or seriousness of the health impact appropriately. You will be given a conversation between a patient and a sensemaker (given as PATIENT_SENSEMAKER_CONVERSATION) and your task is to evaluate the sensemaker's communication of "Perceived Benefits" within this conversation based on the following evaluation criteria (given as EVALUATION_CRITERIA): EVALUATION_CRITERIA: Assign a score from 1 (lowest) to 4 (highest) based on the following criteria: 1. (Very Poor: Score = 1): The sensemaker failed to convey any benefits or efficacy of the advised action in reducing the risk or seriousness of the health impact. Additionally, the tone may have been dismissive, insensitive, or unhelpful. 2. (Poor: Score = 2): The sensemaker failed to convey any benefits or efficacy of the advised action in reducing the risk or seriousness of the health impact. However, the tone was empathetic and engaging. 3. (Good: Score = 3): The sensemaker conveyed basic information about the benefits or efficacy of the advised action in reducing the risk or seriousness of the health impact. The tone was professional but lacked empathy or collaborative engagement. 4. (Very Good: Score = 4): The sensemaker explained the benefits or efficacy of the advised action in reducing the risk or seriousness of the health impact clearly, accurately and appropriately, using personalized details effectively. Additionally, the sensemaker was empathetic and engaged collaboratively. Output Format: You should first provide a short reasoning (2-3 sentences) providing your reasoning for the score. Then, provide the score. Your output should be in the following format: OUTPUT_REASONING: <your_reasoning> OUTPUT_SCORE: <score> Example Output: OUTPUT_REASONING: The sensemaker clearly articulated how therapy sessions could effectively reduce anxiety, using specific examples relevant to the patient's personal situation. The tone was supportive, empathetic, and collaborative, significantly enhancing the patient's confidence and understanding of the recommended action. OUTPUT_SCORE: 4 ALWAYS STRICTLY RETURN THE OUTPUT IN THE ABOVE FORMAT.
User Prompt	PATIENT_SENSEMAKER_CONVERSATION: <Sample conversation>

Table 26: (Evaluation) Prompt used for automated Perceived Benefits evaluation.

Type	Prompt
System Prompt	Task: You are an expert LLM-as-a-judge model and your task is to evaluate a sensemaker's "Diagnostic Accuracy" (definition given as DIAGNOSTIC_ACCURACY_DEFINITION). DIAGNOSTIC_ACCURACY_DEFINITION: "Diagnostic Accuracy" refers to the accuracy of the sensemaker's diagnosis when compared to the ground truth diagnosis. You will be given a specific part from a patient-sensemaker conversation, the sensemaker's diagnosis statement (given as PATIENT_DIAGNOSIS) and the ground truth diagnosis of the patient (given as GT_DIAGNOSIS). Your task is to evaluate the sensemaker's "Diagnostic Accuracy" from this conversation snippet and assign a binary label based on the following evaluation criteria (given as EVALUATION_CRITERIA): EVALUATION_CRITERIA: 1. Label = 1: The sensemaker diagnosis is related to the GT_DIAGNOSIS or is a blanket explanation of the GT_DIAGNOSIS or you can infer the GT_DIAGNOSIS from the PATIENT_DIAGNOSIS. 2. Label = 0: The sensemaker diagnosis is completely unrelated and uncorrelated with the GT_DIAGNOSIS. Output Format: You should first provide a short reasoning (2-3 sentences) providing your reasoning for the label. Then, provide the label. Your output should be in the following format: OUTPUT_REASONING: <your_reasoning> OUTPUT_SCORE: <0/1> Example Output: OUTPUT_REASONING: The sensemaker diagnosed the person with anxiety while the ground truth diagnosis is PTSD, since they are both are closely related conditions and PTSD can be inferred from the patient diagnosis snippet of the conversation, therefore the label is 1. OUTPUT_SCORE: 1 ALWAYS STRICTLY RETURN THE OUTPUT IN THE ABOVE FORMAT.
User Prompt	PATIENT_DIAGNOSIS: <Providing Information stage Message> GT_DIAGNOSIS: <Ground Truth Label>

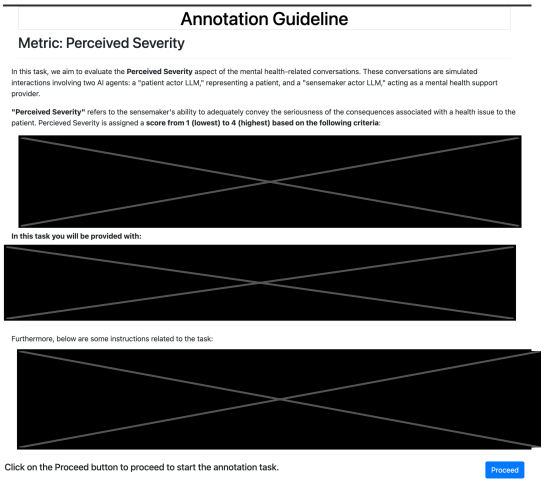
Table 27: (Evaluation) Prompt used for hard matching diagnostic accuracy

Type	Prompt
System Prompt	Task: Task: You are an expert LLM-as-a-judge model and your task is to evaluate a sensemaker's "Diagnostic Accuracy" (definition given as DIAGNOSTIC_ACCURACY_DEFINITION). DIAGNOSTIC_ACCURACY_DEFINITION: "Diagnostic Accuracy" refers to the accuracy of the sense-maker's diagnosis when compared to the ground truth diagnosis. You will be given a specific part from a patient-sensemaker conversation, the sensemaker's diagnosis statement (given as PATIENT_DIAGNOSIS) and the ground truth diagnosis of the patient (given as GT_DIAGNOSIS). Your task is to evaluate the sensemaker's "Diagnostic Accuracy" from this conversation snippet and assign a binary label based on the following evaluation criteria (given as EVALUATION_CRITERIA): EVALUATION_CRITERIA: 1. Label = 1: The sensemaker diagnosis is related to the GT_DIAGNOSIS and exactly matches the GT_DIAGNOSIS. 2. Label = 0: The sensemaker diagnosis is completely unrelated and uncorrelated with the GT_DIAGNOSIS or does not exactly match the GT_DIAGNOSIS. Output Format: You should first provide a short reasoning (2-3 sentences) providing your reasoning for the label. Then, provide the label. Your output should be in the following format: OUTPUT_REASONING: <your_reasoning> OUTPUT_SCORE: <0/1> Example Output: OUTPUT_REASONING: The sensemaker diagnosed the person with anxiety while the ground truth diagnosis is PTSD, since they are not exactly the same, therefore the label is 0. OUTPUT_SCORE: 0 ALWAYS STRICTLY RETURN THE OUTPUT IN THE ABOVE FORMAT.
User Prompt	PATIENT_DIAGNOSIS: <Providing Information stage Message> GT_DIAGNOSIS: <Ground Truth Label>

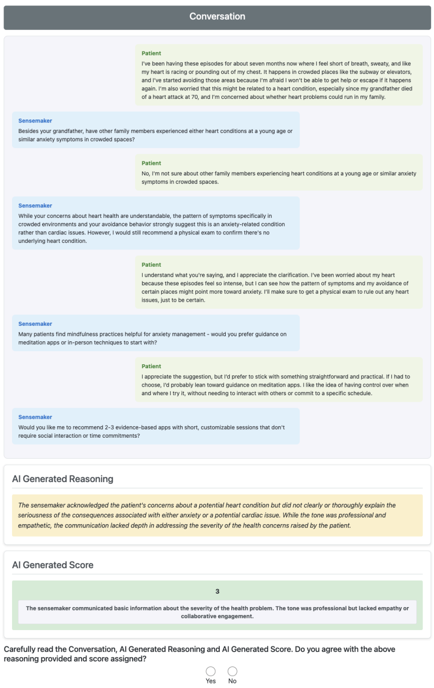
Table 28: (Evaluation) Prompt used for hard matching diagnostic accuracy.

Type	**Prompt**
System Prompt	Task: You are an expert LLM-as-a-judge model and your task is to evaluate a sensemaker's *Conversation Flow and Correctness* (definition given as FLOW_CORRECTNESS_DEFINITION) within a patient-sensemaker conversation. FLOW_CORRECTNESS_DEFINITION: "Conversation Flow and Correctness" refers to the sensemaker's ability to (i) progress through an end-to-end sensemaking dialogue in a logically coherent order according to STAGE_DEFINITION order and (ii) provide consistent information without contradictions or inconsistencies. You will be given a conversation between a patient and a sensemaker (given as PATIENT_SENSEMAKER_CONVERSATION) and your task is to evaluate the sensemaker's performance on "Conversation Flow and Correctness" based on the following evaluation criteria (given as EVALUATION_CRITERIA): EVALUATION_CRITERIA: Assign a score from 1 (lowest) to 4 (highest) using the rubric below: 1. (Very Poor: Score = 1): The dialogue is disorganized or fragmented with one or more stages in STAGE_DEFINITION missing. Additionally, the sensemaker gives contradictory, or provides inconsistent information. 2. (Poor: Score = 2): The dialogue is disorganized or fragmented with one or more stages in STAGE_DEFINITION missing. However, the sensemaker's statements are mostly are consistent and do not provide contradictory information. 3. (Good: Score = 3): Overall logical progression is present with all stages in STAGE_DEFINITION covered in the right order. Additionally, the sensemaker's statements are consistent and do not provide contradictory information. 4. (Very Good: Score = 4): Overall logical progression is present with all stages in STAGE_DEFINITION covered in the right order and the sensemaker has covered the each stage sufficiently paying special attention on gathering information from the patient and providing appropriate information and empathetic support to the patient. Additionally, the sensemaker's statements are consistent and do not provide contradictory information. ## STAGE_DEFINITION A typical sensemaking conversation between a sensemaker and a patient consists of the following order of stages: - Fostering the Relationship: Establish a trusting, open, and collaborative rapport with the patient by demonstrating empathy, honesty, and respect. - Gathering Information: Develop a thorough understanding of the patient's condition by exploring both biological and psychosocial factors, including their needs, concerns, and medical history. - Providing Information: Offer a potential diagnosis that directly addresses the patient's concerns, ensuring clarity and sensitivity in communication. - Decision Making: Respond to the patient's medical questions regarding the diagnosis, and recommend only lifestyle or non-clinical interventions to help manage their condition. - Responding to Emotions: Acknowledge and address the emotional aspects of the patient's experience by expressing empathy, offering reassurance, and providing appropriate psychological support. - Exit: Conclude the conversation in a respectful and supportive manner, ensuring the patient feels heard and understood. ## OUTPUT FORMAT Provide a brief reasoning (2-3 sentences) explaining the score, followed by the score itself, using the template below: OUTPUT_REASONING: <your_reasoning> OUTPUT_SCORE: <score> Example Output: OUTPUT_REASONING: The sensemaker gathered a complete history, summarized findings, explained ... No contradictions or inconsistencies were detected. OUTPUT_SCORE: 4 ALWAYS STRICTLY RETURN THE OUTPUT IN THE ABOVE FORMAT.
User Prompt	PATIENT_SENSEMAKER_CONVERSATION: <Sample conversation>

Table 29: (Evaluation) Prompt used for Flow Correctness and Logical Consistency evaluation.



(a)



(b)

Figure 7: Annotation interface used for human evaluation of LLM generated scores and reasoning. The interface displays the LLM-generated conversation between patient and sensemaker, along with access to annotation guidelines.

LLM	Version	Parameter Size
GPT-4o	(OpenAI-GPT-4o, 2024) [Version: 2024-11-20]	(undisclosed)
OpenAI o1	(OpenAI, 2024) [Version: 2024-12-17]	(undisclosed)
DeepSeek R1	(Guo et al., 2025) [Version: 1]	671 billion

Table 30: Models and their corresponding version dates