On the Role of Context in LLM Alignment to Mental Health Counseling Competencies

Sadiya Sayara Chowdhury Puspo, Marcos Zampieri, Özlem Uzuner
George Mason University, VA, USA
spuspo@gmu.edu

Abstract

As Large Language Models (LLMs) demonstrate strong performance on clinical benchmarks, it remains unclear whether this reflects true patient-specific reasoning or reliance on generalized symptom patterns. To address this question, we evaluate LLMs on a counseling competency benchmark to assess their use of patient-specific contextual information. Through controlled experiments with contextual ablations, role framing, Thread-of-Thought (ThoT) prompting, and input perturbations, we find that removing contextual details results in only modest performance drops, and predictions remain relatively stable under input variations. Error analysis further reveals systematic patterns where models favor general clinical associations over context-specific cues, even when such cues are correctly identified during intermediate reasoning. Our findings suggest that achieving passing-level performance may not reflect robust context-sensitive decision-making, revealing a potential gap between benchmark-level clinical competence and reliable utilization of patient-specific contextual information. These results highlight the need for evaluation frameworks that more directly assess context integration in mental health applications.

1 Introduction

More than 1 billion people worldwide live with mental health disorders, according to recent data from the World Health Organization (WHO).[1] In the United States alone, one in five adults experience mental illness each year, as reported by SAMHSA[2]. At the same time, access to professional care remains severely constrained. According to HRSA[3], approximately 40% (137 million) Americans lived in designated mental health professional shortage areas by the end of 2025. Additionally, 6 in 10 psychologists do not accept new patients, and the national average wait time for behavioral health services is 48 days.

These gaps highlight the urgent need for scalable solutions that can support mental health assessment and care delivery. In response, recent studies have explored how LLMs might help with various mental health applications such as identifying disorders (Guo et al., 2024; Ge et al., 2025), generating empathetic responses (Loh and Raamkumar, 2023; Lee et al., 2023), and suggesting treatment approaches (Ghorbian and Ghobaei-Arani, 2025; Levkovich, 2025). While these models show promise, it remains unclear whether they can effectively utilize patient-specific contextual information when making clinical decisions. In this work, we use the term *patient-specific context* to refer to information contained in clinical case vignettes beyond the question itself. We distinguish this from the broader NLP usage of "context", which may refer to conversational history, retrieved documents, or prompt conditioning.

Mental health counseling is not merely a pattern recognition task. It involves interpreting complex, context-rich narratives including patients' lived experiences, stressors, and symptoms that are often ambiguous or evolving. Effective decision-making requires integrating information about a person's age, background, culture, and situational factors (Robles-Piña and McPherson, 2002). The key question is whether LLMs meaningfully incorporate such patient-specific context into their decisions, or whether their predictions are primarily driven by general knowledge and pattern-based associations learned during training. At the same time, it remains unclear whether existing benchmarks are explicitly designed to require robust patient-specific contextual reasoning for successful performance. If many questions can be answered correctly using generalized clinical knowledge alone, benchmark accuracy may overestimate context-sensitive clinical reasoning capabilities.

We investigate these questions[4] through a sys-

[1] WHO Mental Health Conditions Report
[2] SAMSHA Mental Health Facts
[3] HRSA Health workforce Report
[4] GitHub Repository

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 76–87

tematic evaluation of LLMs on CounselingBench, a dataset of case-based questions spanning core counseling competencies tested in the National Clinical Mental Health Counseling Examination (NCMHCE), a U.S. licensing exam for mental health counselors. The dataset reflects realistic clinical case vignette (as shown in *Example Case Vignette* Box 1) encountered by students preparing for licensure, with a typical *passing benchmark of approximately 63% accuracy*. Prior work shows that larger, instruction-tuned LLMs can approach or exceed this threshold (Nguyen et al., 2024), though their reasoning capabilities remain underexplored.

Example Case Vignette

Patient Demographics: Male, 26, Caucasian, single

Mental Status Examination: Irritable affect, disorganized and pressured speech, audiovisual hallucinations, tangential thinking, paranoid ideation, poor insight and judgment.

Presenting Problem: First session: Medication nonadherence post-discharge; client believes he is being poisoned; verbal altercations with residents; refuses medication due to side effects.

Other Contexts: Client was stable on medication prior to stopping; resides in assisted living; caseworker accompanied; becomes angry when hospitalization is mentioned.

Question: You administer the Scale for the Assessment of Positive Symptoms (SAPS) to determine the severity of which of the following?

Answer Choices:

(A) Avolition
(B) Diminished speech
(C) Agitation
(D) Social withdrawal

To examine these limitations, we evaluate models across different scales and training paradigms, including general-purpose models (LLaMA-3-8B-Instruct, LLaMA-3-70B-Instruct) and a medically fine-tuned model (OpenBioLLM-70B). These models were selected based on prior work, where they achieve performance at or above the passing threshold on the benchmark. Building on this, we investigate whether such performance reflects genuine use of patient-specific context or whether predictions remain driven by general knowledge and pattern-based associations. This selection further allows us to assess whether increased model capacity and domain-specific training improve the integration of contextual information.

Building on this motivation, we design experiments to test whether LLM predictions are sensitive to patient-specific context. In this work, we define *patient-specific contextual information* as four key components of the dataset: (1) patient demographics, (2) presenting problems, (3) mental status examination, and (4) other contexts that provide clinically relevant details essential for informed decision-making. To probe this, we conduct ablation experiments to assess whether removing such details changes model predictions, answer option shuffling to test sensitivity to option positioning and detect positional bias, and question–answer ordering perturbations to evaluate robustness to broader input structure. We further evaluate prompting strategies, including role framing (e.g., assigning the model as a student or counselor) and context-oriented Thread-of-Thought (ThoT) prompting, to determine whether explicitly encouraging attention to patient-specific details improves contextual reasoning.

Finally, we perform a structured error analysis to identify cases where models recognize but fail to prioritize relevant contextual factors. For instance, when symptoms follow a recent traumatic event, a context-sensitive model should favor trauma-related diagnoses, whereas a context-insensitive model may default to more common disorders based on surface-level symptom overlap.

Based on this, we address the following Research Questions (RQ):

RQ1 Do LLMs utilize patient-specific contextual information when answering clinical questions?

RQ2 Can prompting strategies, including role framing and context-oriented prompting (ThoT), enhance the use of patient-specific contextual information in LLM decision-making?

2 Related Work

The use of computational methods to analyze mental health signals has evolved from rule-based approaches to transformer-based models capable of processing social media posts (Greco et al., 2023; Raihan et al., 2024; Bucur et al., 2025a,b; Bucur, 2026). With the emergence of LLMs, recent work has examined their capacity for identifying mental health risk signals and extended evaluation to multilingual (Elboardy et al., 2025; Raihan et al., 2024, 2026; Bucur et al., 2026) and low-resource settings. More recently, the focus has shifted from detection to interaction, where LLMs are studied as agents that simulate counseling through role-based

(Qiu and Lan, 2024; Cho et al., 2026) and multi-agent reasoning (Ozgun et al., 2025; Shafi, 2025). This shift raises a key question: can these models support clinically meaningful reasoning rather than simply produce plausible responses.

From a clinical perspective, effective psychotherapy requires integrating general knowledge with patient-specific context. The American Psychological Association (APA) guidelines for evidence-based practice emphasize that treatment decisions must account for individual characteristics such as developmental history, sociocultural background, and environmental stressors (on Evidence-Based Practice et al., 2006). Clinical reasoning is therefore inherently contextual, requiring not only symptom recognition but also interpretation of patient history, situational factors, and individualized needs (Cook et al., 2017; Moggia et al., 2024).

Current LLM evaluations do not fully capture this requirement. Benchmarks such as CounselingBench (Nguyen et al., 2024) show that larger LLMs can surpass the NCMHCE passing threshold and generate coherent explanations. However, such performance does not guarantee effective use of case-specific information, as predictions may rely on generalized symptom patterns rather than patient details. Recent work further highlights limitations of LLMs in psychotherapy settings (Chandra et al., 2025; Arnaout et al., 2026), including failures in contextual understanding (Wang et al., 2025), misinterpretation of user experiences, and reliance on generic responses (Iftikhar et al., 2025). Our work directly evaluates whether LLMs utilize patient-specific context in counseling scenarios.

3 Methods

3.1 Data

The dataset, named CounselingBench[1], is curated from National Clinical Mental Health Counseling Examination (NCMHCE) mock exams, sourced from mometrix.com, tetsst.vcm, CounselingExam.com, which is taken part by people who seek to become licensed clinical mental health counselors. CounselingBench contains 1,612 unique questions across 138 case studies consisting patient demographics, mental status examination, presenting problems, questions with answers, and expert-generated rationales. The benchmark assesses five key mental health counseling competencies identified through a national job analysis of over 16,000 credentialed counselors, representing empirically-validated work behaviors for effective counseling practice. To ensure fair use compliance, Nguyen et al. (2024) adapt procedures from Jin et al. (2021) by shuffling answer options while tracking correct answers. Two medical doctors specializing in psychiatry independently annotated all 1,612 questions to map each question to its corresponding competency assessment area, their distributions are shown in Table 1.

Competency	#
Counseling Skills & Interventions (CS&I)	599
Intake, Assessment, & Diagnosis (IA&D)	460
Professional Practice & Ethics (PP&E)	274
Treatment Planning (TP)	253
Core Counseling Attributes (CCA)	23

Table 1: Distribution of questions across competencies

The dataset includes various demographic backgrounds corresponding to categories present in the US Census data. Several cases involve patients not born in the United States, requiring cultural competency for accurate assessment.

3.2 Modeling

To evaluate whether LLMs utilize patient-specific contextual information or rely on pattern-based associations, we design controlled experiments that systematically manipulate both contextual and structural aspects of the input.

We begin by defining what constitutes *patient-specific context* in the CounselingBench dataset. As illustrated in Figure 1, clinical reasoning requires integrating four dimensions of information beyond surface-level symptoms when making counseling decisions, including demographic attributes, presenting conditions, mental status examination, and other contexts. These contextual dimensions form the basis of our ablation experiments, where specific components are removed to evaluate their influence on model predictions.

Although Figure 1 presents these components as structured elements for clarity, these elements are embedded within unstructured narratives, requiring models to extract and integrate relevant signals rather than relying on explicitly structured inputs.

We select LLaMA-3-8B-Instruct, LLaMA-3-70B-Instruct, and OpenBioLLM-70B to capture variation across both model scale and domain specialization. Prior work on CounselingBench (Nguyen et al., 2024) shows that larger instruction-

[1] https://github.com/cuongnguyenx/CounselingBench

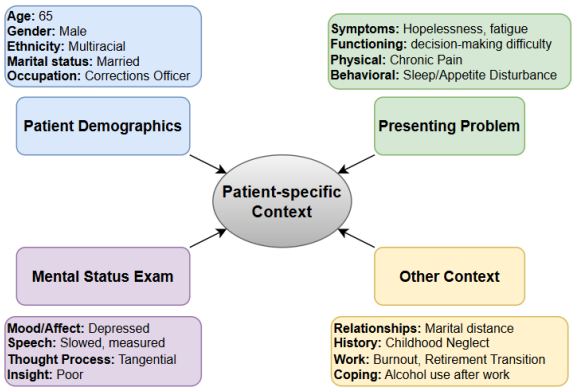

Figure 1: Breakdown of patient-specific contextual information in a representative CounselingBench case. Patient-specific context spans four dimensions: demographic attributes, presenting problems, mental status examination, and other context.

tuned models are more likely to exceed the NCMHCE passing threshold of 63% accuracy, with LLaMA-3-70B-Instruct achieving strong performance while LLaMA-3-8B-Instruct remains closer to the threshold. This enables comparison across models with differing levels of task proficiency and assess whether higher accuracy corresponds to improved utilization of patient-specific context.

Additionally, prior work suggests diminishing performance gains from scaling within newer LLM families, motivating a comparison between 8B and 70B variants to examine whether increased model capacity corresponds to differences in contextual reasoning behavior rather than overall benchmark accuracy alone. We further include OpenBioLLM-70B, a LLaMA-based model additionally instruction-tuned on biomedical and medical-domain data, to investigate whether domain-specific adaptation improves sensitivity to patient-specific contextual information beyond general-purpose instruction tuning. However, because OpenBioLLM-70B is primarily optimized for biomedical reasoning rather than psychotherapy or counseling-specific decision-making, any improvements in contextual counseling performance are not necessarily expected. This comparison therefore allows us to examine whether higher capacity or medical-domain specialization meaningfully changes how models utilize patient-specific context during clinical decision-making.

3.3 Experimental Design

To evaluate whether model predictions are sensitive to patient-specific context or driven by pattern-based associations, we introduce controlled perturbations to the input. These perturbations are designed to test whether model predictions remain stable under changes that preserve semantic content but alter presentation, providing evidence of context-sensitive reasoning versus reliance on superficial input patterns.

Ablation Study: We conduct a structured ablation study to evaluate the extent to which model predictions depend on patient-specific contextual information. Specifically, we progressively remove contextual components from the case descriptions while preserving the question and answer options, allowing us to isolate the contribution of different types of information to model decision-making.

Contextual attributes are removed in a staged manner, following a hierarchy from less to more diagnostically informative components. We begin by removing *patient demographic* attribute, which provide general background information such as, name, age, gender etc, about the patient. Next, we remove *other contexts* attribute which contains relationship history, and occupational stress, these may influence interpretation but are not directly diagnostic. We then remove the *mental status exam* attribute, which captures clinical observations such as appearance, speech, thought process, and orientation. Finally, we remove the *presenting problem*, which contains the core symptom descriptions and key elements of the counseling interaction most directly tied to diagnosis.

This progressive removal enables a graded analysis of model sensitivity to different layers of patient-specific context. A context-sensitive model is expected to exhibit increasing changes in predictions as more diagnostically relevant information is removed. In contrast, stable predictions under such perturbations suggest that model behavior is driven by pattern-based associations rather than the integration of patient-specific contextual information.

Surface Perturbations: In this setting, we randomly reorder the answer choices to assess sensitivity to option positioning (*ShuffledOP*) to detect id models have any positional bias. On the other hand, in the question-first condition (*Q-first*), the question and answer options are presented before the case description, reversing the standard context-first format. This design allows us to analyze the primacy effect and assess whether model predictions depend on the relative position of contextual information. A context-sensitive model should integrate patient-specific details regardless of their position, while

strong sensitivity to ordering suggests reliance on input structure rather than contextual reasoning.

Role Framing: We adopt the prompting template from the previous study as our baseline, where the model is assigned the role of a student attempting to answer NCMHCE-style questions. To evaluate the effect of role specification on model behavior, we systematically vary the identity assigned to the LLM. Specifically, we consider four role configurations:

(1) *Role:STU*– the baseline student role,
(2) *Role:LC*– a licensed mental health counselor role,
(3) *Role:LC-DSM*– a licensed counselor role with explicit DSM-5 oriented diagnostic reasoning,
(4) *Role:EXP*– an expert-identity prompt

For the expert-identity setting, following the framework of Xu et al. (2023), we prompt GPT-5.1 to generate a specialized expert profile tailored to solving NCMHCE-style multiple-choice questions, which is then used to define the role and response expectations for all evaluated models.

Finally, to assess whether aggregating across role specifications improves sensitivity to patient-specific context, we apply role-based ensembling (*Role:Ensemble*). Each role configuration is evaluated independently, and final predictions are obtained via majority voting across role-conditioned outputs.

Thread-of-Thought (ThoT) Prompting: To investigate whether structured reasoning improves patient-specific contextual integration, we adapt Thread-of-Thought (ThoT) prompting (Zhou et al., 2023) to the clinical counseling domain. The original ThoT formulation is designed for retrieval-based settings, where models must filter relevant information from noisy inputs. However, clinical vignettes differ in that most contextual information is already relevant, and the primary challenge lies in integrating and prioritizing these signals for diagnosis. To better reflect this setting, we modify the ThoT trigger to explicitly structure reasoning around key clinical dimensions.

We follow a two-step ThoT pipeline. In Step 1, the model produces a structured intermediate summary from the patient-specific context. In Step 2, this summary replaces the original vignette and is used for answer selection. Unlike prior work that relies on free-text answer extraction, we use a consistent scoring-based approach for answer selection to ensure comparability across prompting conditions.

While this formulation encourages more explicit references to patient-specific details, it allows us to directly assess whether improved structure translates into better diagnostic decisions, or merely more coherent explanations without corresponding gains in accuracy.

4 Results & Observations

4.1 Ablation Study

Table 2 presents the results of our staged contextual ablation study across three models. A key observation is that removing patient-specific context components produces only modest accuracy degradation across all models. For LLaMA-3.1-8B-Instruct, the full context condition achieves 0.664, while the question-only condition – with all patient context removed – yields 0.625, a drop of only 3.9 percentage points. A similar pattern holds for LLaMA-3.3-70B-Instruct (0.732 $\rightarrow$ 0.691, Δ = 4.1%) and OpenBioLLM-70B (0.722 $\rightarrow$ 0.689, Δ = 3.3%). These relatively modest performance drops suggest that many questions in the benchmark may be answerable using generalized clinical knowledge and pattern-based associations acquired during pre-training, without requiring substantial reliance on patient-specific contextual information. However, the results do not conclusively determine whether the observed behavior primarily reflects limitations in the models' contextual reasoning abilities, limitations of the benchmark's ability to isolate context-sensitive reasoning, or the inherent difficulty and heterogeneity of context-dependent clinical questions. Instead, the findings indicate that benchmark-level performance alone may be insufficient for determining the extent to which LLMs reliably utilize patient-specific context during clinical decision-making.

Examining the per-competency breakdown reveals more nuanced patterns. The Core Counseling Attributes (CCA) competency shows notably stable performance across ablation conditions for LLaMA-3.1-8B-Instruct and OpenBioLLM-70B, suggesting this competency may be particularly amenable to pattern-based responses that do not require contextual integration.In contrast, the Intake Assessment and Diagnosis (IA&D) competency shows a consistent decline from full-context to question-only settings across all models, from 0.676 to 0.614 for LLaMA-3.1-8B-Instruct, 0.721 to 0.667 for LLaMA-3.3-70B-in, and 0.704 to

Condition	Overall Accuracy	Δ	IA&D (N=466)	TP (N=254)	CS&I (N=600)	PP&E (N=275)	CCA (N=23)
LLaMA-3.1-8B-Instruct							
Demographics + Other Contexts + Mental Status Exam + Presenting Problem + Question with Options	0.664	-	0.676	0.665	0.695	0.575	0.696
Other Contexts + Mental Status Exam + Presenting Problem + Question with Options	0.654	-0.010	0.655	0.650	0.690	0.575	0.696
Mental Status Exam + Presenting Problem + Question with Options	0.655	-0.009	0.682	0.650	0.663	0.589	0.696
Presenting Problem + Question with Options	0.642	-0.022	0.644	0.654	0.662	0.578	0.739
Question with Options	0.625	-0.039	0.614	0.598	0.645	0.618	0.696
LLaMA-3.3-70B-Instruct							
Demographics + Other Contexts + Mental Status Exam + Presenting Problem + Question with Options	0.734	-	0.721	0.720	0.757	0.720	0.783
Other Contexts + Mental Status Exam + Presenting Problem + Question with Options	0.735	0.001	0.730	0.736	0.753	0.709	0.739
Mental Status Exam + Presenting Problem + Question with Options	0.732	-0.003	0.717	0.756	0.748	0.709	0.696
Presenting Problem + Question with Options	0.727	-0.008	0.721	0.740	0.735	0.709	0.739
Question with Options	0.691	-0.044	0.667	0.677	0.722	0.673	0.739
OpenBioLLM-70B							
Demographics + Other Contexts + Mental Status Exam + Presenting Problem + Question with Options	0.722	-	0.704	0.713	0.742	0.727	0.696
Other Contexts + Mental Status Exam + Presenting Problem + Question with Options	0.724	0.002	0.697	0.736	0.748	0.713	0.696
Mental Status Exam + Presenting Problem + Question with Options	0.711	-0.011	0.689	0.740	0.723	0.705	0.696
Presenting Problem + Question with Options	0.713	-0.008	0.682	0.728	0.737	0.709	0.652
Question with Options	0.689	-0.033	0.646	0.685	0.723	0.695	0.696

Table 2: Ablation study evaluating the impact of removing patient-specific contextual components on model performance. Context is progressively reduced from full clinical vignette (demographics, other context, mental status exam, and presenting problem) to question-only input. Performance is reported as accuracy and across competency domains. Δ denotes change relative to the full-context condition for each model.

0.645 for OpenBioLLM-70B, indicating greater sensitivity to the removal of contextual information, though the overall sensitivity remains low.

Notably, the medically fine-tuned OpenBioLLM-70B does not exhibit substantially greater sensitivity to context removal compared to the general-purpose LLaMA models, suggesting that domain-specific training does not meaningfully improve reliance on patient-specific information. Overall, these results indicate that models approach this task largely as a knowledge retrieval rather than one requiring patient-specific reasoning, consistent with our central hypothesis.

4.2 Surface Perturbations

Table 3 presents results across prompting strategies including role framing input controls and ThoT prompting for LLaMA-3.1-8B-Instruct, LLaMA-3.3-70B-Instruct, and OpenBioLLM-70B. Shuffling answer options (*ShuffledOP*) results in negligible changes in accuracy across all models (LLaMA-3.1-8B: 0.660, LLaMA-3.3-70B: 0.735, OpenBioLLM-70B: 0.707), indicating limited reliance on answer position. In contrast, presenting the question before the vignette (*Q-first*) leads to the largest performance drop across models (LLaMA-3.1-8B: 0.629, LLaMA-3.3-70B: 0.709, OpenBioLLM-70B: 0.682), suggesting sensitivity to input ordering rather than clinical content. One possible explanation is that presenting the question and answer options first encourages models to form an initial prediction based on generalized clinical associations and learned response patterns before processing the patient vignette, thereby reducing the influence of subsequent patient-specific contextual information during answer selection. Together, these findings reveal an asymmetry: while models are robust to changes in answer option position, they are sensitive to prompt structure, a pattern inconsistent with robust clinical reasoning, where decisions should be driven by clinical content rather than sensitivity to input ordering.

4.3 Role Framing & ThoT Prompting

Across all three models, different role-framing strategies and their ensemble combinations produce only marginal accuracy differences. For LLaMA-3.1-8B-Instruct, the four role-framed prompts span a narrow range of 0.655 to 0.663, and a similar pattern holds for LLaMA-3.3-70B-Instruct (0.734-0.737) and OpenBioLLM-70B (0.720-0.722). This near-uniform performance suggests that assigning different clinical personas, from student to licensed counselor to domain expert, does not meaningfully alter the model's underlying reasoning process, consistent with our hypothesis that predictions are driven by knowledge-based associations rather than persona-specific clinical reasoning.

ThoT prompting consistently underperforms the role-framed prompts across all three models (LLaMA-3.1-8B-Instruct: 0.628, LLaMA-3.3-70B-Instruct: 0.730, OpenBioLLM-70B: 0.709). Despite explicitly instructing the model to process patient-specific context in a structured, step-by-step manner, ThoT does not yield accuracy gains over simpler role-framed prompts. This suggests that structured contextual summarization alone may be insufficient to substantially improve context-

	Accuracy	IA&D (N=466)	TP (N=254)	CS&I (N=600)	PP&E (N=275)	CCA (N=23)
LLaMA-3.1-8B-Instruct						
Role: STU	0.663	0.670	0.657	0.697	0.578	0.696
ShuffledOP	0.660	0.667	0.634	0.698	0.585	0.652
Q-first	0.629	0.627	0.634	0.635	0.615	0.609
Role: LC	0.658	0.674	0.673	0.680	0.567	0.696
Role: LC-DSM	0.655	0.657	0.665	0.683	0.575	0.739
Role: EXP	0.658	0.652	0.673	0.693	0.571	0.739
Role: Ensemble	0.663	0.672	0.665	0.690	0.582	0.696
ThoT	0.628	0.633	0.626	0.652	0.571	0.609
LLaMA-3.3-70B-Instruct						
Role: STU	0.735	0.721	0.720	0.758	0.720	0.783
ShuffledOP	0.735	0.730	0.728	0.755	0.709	0.739
Q-first	0.709	0.693	0.705	0.738	0.680	0.652
Role: LC	0.737	0.727	0.752	0.755	0.709	0.696
Role: LC-DSM	0.734	0.730	0.732	0.750	0.716	0.696
Role: EXP	0.737	0.719	0.752	0.753	0.727	0.696
Role: Ensemble	0.732	0.719	0.748	0.750	0.705	0.696
ThoT	0.730	0.734	0.713	0.745	0.716	0.696
OpenBioLLM-70B						
Role: STU	0.722	0.704	0.713	0.742	0.727	0.696
ShuffledOP	0.707	0.689	0.705	0.720	0.720	0.652
Q-first	0.682	0.659	0.693	0.693	0.687	0.696
Role: LC	0.721	0.697	0.705	0.750	0.724	0.696
Role: LC-DSM	0.720	0.697	0.717	0.743	0.720	0.696
Role: EXP	0.721	0.695	0.705	0.755	0.716	0.696
Role: Ensemble	0.721	0.697	0.713	0.748	0.720	0.696
ThoT	0.709	0.687	0.713	0.732	0.698	0.696

Table 3: Performance across prompting strategies and input controls for three LLMs on CounselingBench. Results are reported as accuracy and across competency domains. Role: STU = student persona, ShuffledOP = shuffled answer options, Q-first = question before context, Role: LC = licensed counselor, Role: LC-DSM = licensed counselor with DSM-5-TR reasoning, Role: EXP = clinical expert, Role: Ensemble = majority vote across all four role-framed prompts, ThoT = Thread-of-Thought two-step pipeline.

sensitive reasoning in clinical counseling tasks. However, it remains unclear whether this reflects limitations in the models' ability to effectively utilize patient-specific information from the generated summaries, or whether many benchmark questions can already be answered without substantial reliance on such context.

LLaMA-3.3-70B-Instruct consistently outperforms LLaMA-3.1-8B-Instruct by roughly 7 percentage points across all conditions, aligning with prior findings that larger models achieve higher accuracy on CounselingBench (Nguyen et al., 2024). However, OpenBioLLM-70B, despite being fine-tuned on medical literature, does not surpass the general-purpose LLaMA-3.3-70B-Instruct and often performs comparably. This suggests that medically fine-tuned models do not provide substantial advantages over instruction-tuned models for utilizing patient-specific context. Overall, these results indicate that counseling competency relies more on integrating patient-specific contextual information than on biomedical knowledge, and models may not effectively utilize patient-specific context despite access to domain knowledge.

ThoT Summary - Decision Mismatch

In a pediatric counseling vignette, a 10-year-old boy presents with anxiety symptoms following a recent family relocation and separation from caregivers. The vignette emphasizes fears about parental safety, sleep disturbance, and distress during separation, and explicitly links the symptoms to this relocation.

ThoT Summary (excerpt): *"The family has recently relocated, which could be a significant stressor for Michael."*

When asked which statement is most indicative of the child's condition, the correct answer is *family relocation*, as it represents the precipitating cause. However, all models instead select *refusal to go to school*.

Despite explicitly identifying the correct contextual factor, the final answer prioritizes a general symptom over the patient-specific cause.

5 Understanding Context-Insensitive Errors

While earlier analyses reveal that models underutilize patient-specific context, they do not explain how this failure manifests at the level of individual clinical decisions. To investigate, we analyze the 490 cases where all four role-framed prompts converge on the same incorrect answer for LLaMA-3.1-8B-Instruct model, the clearest signature of systematic pattern matching rather than knowledge-based patient-specific reasoning. Table 4 summarizes consensus behavior across all 1,621 questions; the 30.2% unanimous-wrong category is our focus here.

Through qualitative inspection of a representative sample of the 490 unanimously incorrect cases, we identify three recurring error patterns that illustrate how context-insensitive reasoning manifests in practice.

Consensus Type	Count	Percentage
All correct	1028	63.4%
All wrong (same answer)	490	30.2%
All wrong (different answer)	29	1.8%
1 correct + 3 wrong	24	1.5%
2 correct + 2 wrong	23	1.4%
3 correct + 1 wrong	27	1.7%

Table 4: Role-framing consensus analysis for LLaMA-3.1-8B-Instruct: all correct and all wrong (same answer) categories indicate unanimous agreement across four role-framed prompts, while partial disagreement rows reflect cases where prompts diverge in their predictions.

Broad-to-specific Failure: Models consistently select general-purpose instruments over the specific tool the patient's context demands. *A 28-year-old female attorney (Robin) presenting with alcohol use disorder states: "I can't live with the pain of*

our separation much longer, and I don't know how to cope with it". All four prompts select the Hamilton Anxiety Rating Scale (HAM-A), a general anxiety measure. The correct answer is the Columbia Suicide Severity Rating Scale (C-SSRS), directly indicated by the patient's language. The contextual signal distinguishing a general anxiety presentation from one with suicidal ideation is present and unambiguous, the model bypasses it in favor of the categorically familiar instrument.

Generic-priority Substitution: When contextual features point in multiple directions, models default to the higher-frequency clinical association rather than integrating the full patient picture. *A 65-year-old male (Alex) presenting with persistent depression, hopelessness, feeling "down in the dumps" and worsening affect is asked what to discuss first at intake*. All four roles select substance use history, a reasonable intake topic, but one whose priority is overridden by the patient's age, sex, and explicit hopelessness. The correct answer is suicidal ideation. The patient's demographics and affect are available; but they do not influence the prediction.

Relational Context Ignored: PP&E shows the highest unanimous error rate at 38.9% (shown in Table 5), reflecting a tendency to apply canonical ethical heuristics without integrating case-specific relational history. *In Robin's fifth session, the client, who has always rescheduled in advance, fails to appear without contact for the first time. A signed release for her mother and an established safety contact are documented in the vignette.* All four roles select "drive to the client's house to perform a wellness check", a generic crisis response. The correct answer is to contact the client's mother, as the established protocol dictates. The model defaults to a dramatic but contextually inappropriate action, ignoring the relational history that determines the right one.

Across all three error types, a consistent pattern emerges: model predictions appear to rely more on general clinical associations than on integrating patient-specific contextual details. This behavior is consistent with our earlier findings, where removing contextual information led to only modest performance degradation and input perturbations had minimal effect on predictions. The clinical consequences are substantial, missed suicidality screening, misprioritized intake assessment, and ethically inappropriate interventions, and in each case, the information required to arrive at the correct decision is explicitly present in the vignette but not effectively utilized. While we do not directly analyze training data distributions, this behavior is consistent with reliance on high-frequency associations rather than patient-specific reasoning.

Competency	Incorrect Predictions / Total Questions	Percentage
CS&I	162/600	27.0%
IA&D	136/466	29.2%
PP&E	107/275	38.9%
TP	79/254	30.7%
CCA	6/23	26.1%

Table 5: Distribution of the 490 unanimously incorrect predictions (all four role-framed prompts selecting the same wrong answer) across counseling competency areas, for LLaMA-3.1-8B-Instruct.

6 Conclusion & Future Directions

We revisit the research questions (RQ) posed in the introduction and present the main findings of our review below:

RQ1: Do models utilize patient-specific information when answering clinical questions?

Findings: *Ablation and input-structure control experiments suggest limited observable dependence on patient-specific context for benchmark performance. Under progressive removal of contextual components, from demographics to presenting problem, performance drops only modestly. Shuffling answer options yields largely stable predictions, suggesting minimal sensitivity to answer position, while varying question order introduces more noticeable changes in performance. Together, these findings indicate that model predictions may be influenced not only by patient-specific information, but also by prompt structure and generalized clinical associations learned during training. However, the extent to which this reflects limitations in contextual reasoning, properties of the benchmark itself, or the inherent difficulty of context-dependent questions remains unclear. A qualitative error analysis of the 490 cases where all four role-framed prompts converge on the same incorrect answer reveals three recurring failure patterns: broad-to-specific failures in assessment selection, generic-priority substitution at intake, and relational context ignored in ethical decision-making. In these cases, contextual information relevant to the correct answer was present in the vignette, yet the models*

consistently selected alternatives more strongly associated with high frequency or generalized clinical patterns.

RQ2: Can prompting strategies, including role framing and context-oriented prompting (ThoT), enhance the use of patient-specific contextual information in LLM decision-making?

Findings: *Role framing has no measurable effect on accuracy, with models converging on the same wrong responses 30.2% of the time regardless of assigned persona. ThoT encourages more structured representations and increases explicit reference to patient-specific details, yet does not improve decision outcomes. Even when relevant context is correctly identified in intermediate reasoning, models fail to prioritize it during answer selection, revealing a gap between structured explanation and actual clinical reasoning that prompting alone cannot close.*

These findings have important implications for the deployment of LLMs in mental health support contexts. Achieving a passing score on aggregate benchmarks may not necessarily indicate robust utilization of patient-specific contextual information during clinical decision-making. Models may perform well overall while still exhibiting difficulty appropriately prioritizing contextual cues that distinguish between clinically appropriate and inappropriate responses in certain cases.

Future work should pursue several directions. First, *contrastive case vignette design*, where identical symptom profiles require different decisions based solely on demographic or situational context, would provide a more direct evaluation of context-sensitive reasoning than ablation-based approaches. Second, the *error taxonomy* identified in this work should be systematically validated across the full set of 490 cases and extended to additional models to determine whether these patterns reflect model-specific behaviors or broader limitations of LLM-based clinical reasoning. Third, given the concentration of errors in Professional Practice and Ethics (PP&E), future work should investigate whether targeted *fine-tuning on ethics-focused cases* with rich relational context can reduce reliance on generalized heuristics and improve context-sensitive decision-making. Finally, future evaluations should examine clinically appropriate uncertainty behaviors, such as asking clarifying questions, expressing insufficient confidence, or declining definitive recommendations when critical contextual information is missing.

Limitations

This study has several limitations. First, our evaluation is restricted to LLaMA-3.1-8B-Instruct, LLaMA-3.3-70B-Instruct, and OpenBioLLM-70B; findings may not generalize to proprietary models such as GPT-4 or Claude, which may exhibit different contextual reasoning behaviors. Second, CounselingBench is based on U.S. licensing examination questions (NCMHCE), which reflect a specific cultural and regulatory context. Performance on this benchmark may not represent LLM capabilities in counseling settings outside the United States or in languages other than English. Third, our evaluation uses multiple-choice questions, which differ substantially from real counseling interactions that involve open-ended dialogue, therapeutic alliance, and moment-to-moment clinical judgment. High benchmark accuracy should not be interpreted as evidence of clinical competence. Finally, the ThoT experiments use a maximum of 512$\sim$768 generated tokens for Step 1 summaries, which may truncate complex multi-session vignettes and limit the quality of the intermediate clinical reasoning captured.

Another limitation is the possibility of benchmark contamination or indirect exposure during pre-training or instruction tuning. Because NCMHCE-style preparation materials are publicly available online, models may partially rely on memorized exam-style patterns rather than robust contextual reasoning, potentially contributing to stable performance under contextual ablation.

Ethical Considerations

This study is purely analytical and does not involve new data collection or human participants. All analyses are conducted on the publicly available CounselingBench dataset (Nguyen et al., 2024), derived from U.S. mental health counseling licensing examinations (NCMHCE). No personally identifiable information is used, and the study follows the ACL Code of Ethics.[1]

Acknowledgments

We thank the CounselingBench authors for releasing this interesting dataset to the research community. We further thank the anonymous CLPSych reviewers for their valuable feedback.

[1] ACL Code of Ethics

References

Hiba Arnaout, Anmol Goel, H Andrew Schwartz, Steffen T Eberhardt, Dana Atzil-Slonim, Gavin Doherty, Brian Schwartz, Wolfgang Lutz, Tim Althoff, Munmun De Choudhury, and 1 others. 2026. Responsible evaluation of ai for mental health. *arXiv preprint arXiv:2602.00065*.

Ana-Maria Bucur. 2026. Computational approaches to mental health disorders detection from social media texts, images and videos. *Procesamiento del Lenguaje Natural*, 76:311–314.

Ana-Maria Bucur, Andreea Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur Khudabukhsh, and Liviu P Dinu. 2025a. Datasets for depression modeling in social media: An overview. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 116–126.

Ana-Maria Bucur, Andreea-Codrina Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur R KhudaBukhsh, and Liviu P Dinu. 2025b. On the state of nlp approaches to modeling depression in social media: A post-covid-19 outlook. *IEEE Journal of Biomedical and Health Informatics*.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–918.

Mohit Chandra, Siddharth Sriraman, Harneet Singh Khanuja, Yiqiao Jin, and Munmun De Choudhury. 2025. Reasoning is not all you need: examining llms for multi-turn mental health conversations. *arXiv preprint arXiv:2505.20201*.

Ha Na Cho, Jiayuan Wang, Di Hu, and Kai Zheng. 2026. Large language model–based chatbots and agentic ai for mental health counseling: Systematic review of methodologies, evaluation frameworks, and ethical safeguards. *JMIR AI*, 5(1):e80348.

Sarah C Cook, Ann C Schwartz, and Nadine J Kaslow. 2017. Evidence-based psychotherapy: Advantages and challenges. *Neurotherapeutics*, 14(3):537–545.

Ahmed Tamer Elboardy, Ziad Mohamed, Ghada Khoriba, Tamer Arafa, and Essam A Rashed. 2025. Bridging gender disparities in mental health: A bilingual large language model for multilingual therapeutic chatbots. In *2025 22nd International Learning and Technology Conference (L&T)*, volume 22, pages 251–256. IEEE.

Zhuohan Ge, Nicole Hu, Darian Li, Yubo Wang, Shihao Qi, Yuming Xu, Han Shi, and Jason Zhang. 2025. A survey of large language models in mental health disorder detection on social media. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, pages 164–176. IEEE.

Mohsen Ghorbian and Mostafa Ghobaei-Arani. 2025. Large language models for mental health diagnosis and treatment: a survey. *Artificial Intelligence Review*, 59(1):9.

Candida M Greco, Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. 2023. Transformer-based language models for mental health issues: a survey. *Pattern Recognition Letters*, 167:204–211.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.

Zainab Iftikhar, Amy Xiao, Sean Ransom, Jeff Huang, and Harini Suresh. 2025. How llm counselors violate ethical standards in mental health practice: A practitioner-informed framework. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1311–1323.

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.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Inbar Levkovich. 2025. Evaluating diagnostic accuracy and treatment efficacy in mental health: A comparative analysis of large language model tools and mental health professionals. *European Journal of Investigation in Health, Psychology and Education*, 15(1):9.

Siyuan Brandon Loh and Aravind Sesagiri Raamkumar. 2023. Harnessing large language models' empathetic response generation capabilities for online mental health counselling support. *arXiv preprint arXiv:2310.08017*.

Danilo Moggia, Wolfgang Lutz, Eva-Lotta Brakemeier, and Leonard Bickman. 2024. Treatment personalization and precision mental health care: Where are we and where do we want to go? *Administration and Policy in Mental Health and Mental Health Services Research*, 51(5):611–616.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J Soled, Michael L Birnbaum, Srijan Kumar, and 1 others. 2024. Do large language models align with core mental health counseling competencies? *arXiv preprint arXiv:2410.22446*.

APA Presidential Task Force on Evidence-Based Practice and 1 others. 2006. Evidence-based practice in psychology. *The American Psychologist*, 61(4):271–285.

Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, and Junxiao Wang. 2025. Trustworthy ai psychotherapy: Multi-agent llm workflow for counseling and explainable mental disorder diagnosis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2263–2272.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, and Marcos Zampieri. 2026. Large language models for mental health: A multilingual evaluation. In *Proceedings of LoResLM 2026*.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Mentalhelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203.

Rebecca A Robles-Piña and Robert H McPherson. 2002. The relationship between educational and demographic variables and supervisor's multicultural counseling competencies. *The Clinical Supervisor*, 20(2):67–79.

Fozle Rabbi Shafi. 2025. Wellbeingagent: An llm-driven agentic framework for personalized mental health support. Master's thesis, Queen's University (Canada).

Yinzhou Wang, Yimeng Wang, Ye Xiao, Liabette Escamilla, Bianca Augustine, Kelly Crace, Gang Zhou, and Yixuan Zhang. 2025. Evaluating an llm-powered chatbot for cognitive restructuring: Insights from mental health professionals. *arXiv preprint arXiv:2501.15599*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. 2023. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*.

Appendix

A Prompts

Role:STU (Baseline):

<System>
You are a helpful, respectful, honest, and knowledgeable **student** studying to become a licensed therapist. You must answer a series of multiple-choice questions provided by the user from a **US mental health counselor licensing exam**. Based on the question text and the context provided, you must answer with either "A", "B", "C", or "D".

</System>

<User>
Question:
The patient demographic is as follows: *[PATIENT DEMOGRAPHICS]*
[PRESENTING PROBLEM]
[MENTAL STATUS EXAMINATION]
[OTHER CONTEXTS]
Given the context above and your expert-level knowledge of mental health counseling, please answer the following question: *[QUESTION]*
[CANDIDATE ANSWERS]

<Assistant>
Correct Answer is (

Role:LC: You are a helpful, respectful, honest, and knowledgeable **licensed mental health counselor**. You must answer a series of multiple-choice **questions similar to those on a licensing board exam**. Based on the question text and the context provided, you must answer with either "A", "B", "C", or "D".

Everything else is as the baseline prompt.

Role:LC-DSM: You are a helpful, respectful, honest, and knowledgeable **licensed mental health counselor**. You must answer a series of multiple-choice questions. Based on the question text and the context provided, you must answer with either "A", "B", "C", or "D", **applying DSM-5 consistent reasoning**.

Everything else is as the baseline prompt.

Prompt to Generate Expert Identity: Suppose you need to answer MCQs for the NCMHCE exam that include extensive client demographics and detailed case studies. Provide an expert identity prompt that can guide the model in responding effectively.

ROLE:EXP (Generated Expert Identity): You are a clinically trained Licensed Mental Health Counselor with comprehensive expertise in DSM-5-TR diagnosis, assessment, ethics, crisis intervention, and evidence-based treatment across diverse populations. You use advanced clinical reasoning to answer multiple-choice questions with the best possible response. Based on the question text and the context provided, you must answer with either "A", "B", "C", or "D"

ThoT Prompt: Walk me through this patient case in manageable parts step by step – consider their demographics, presenting problem, mental status, and other contextual factors – summarizing and analyzing each as we go.

B Inference Settings

Table 6 summarizes the inference settings used across all experiments. For ThoT Step 1, we employ the *Role: LC* prompt to encourage more comprehensive summarization of the full context.

Parameter	**Value**
Temperature (answer selection)	0.0∼0.7
Top-p	1.0
Decoding strategy	Greedy
Max tokens (answer selection)	1
Max tokens (ThoT Step 1)	512∼768
Stop tokens	) and \n
Serving framework	vLLM

Table 6: Inference configuration for all experiments.