Distribution Agreement

Tung Dinh April 15th, 2025

Controllable Clinical Conversations: A Dialogue Act Framework for Trauma-Focused Interview Automation

by

Tung Dinh

Jinho D. Choi
Adviser

Computer Science

Jinho D. Choi

Adviser

Xiao Hu

Committee Member

Jiaying Lu

Committee Member

2025

Controllable Clinical Conversations: A Dialogue Act Framework for Trauma-Focused Interview Automation

By

Tung Dinh

Jinho D. Choi
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2025

Abstract

Controllable Clinical Conversations: A Dialogue Act Framework for Trauma-Focused Interview Automation

By Tung Dinh

Access to timely and structured mental health evaluations remains a challenge, especially for individuals affected by trauma. While large language models (LLMs) offer potential for automating clinical support, most existing systems lack the structured flow and empathetic engagement required in trauma-focused diagnostic interviews. This thesis presents a novel two-stage framework for automating these interviews using dialogue act (DA) classification to improve both the coherence and clinical relevance of AI-generated responses. Our method first labels clinician utterances using a trauma-specific DA taxonomy—including categories such as Empathy (EMP), Clarification Questions (CQ), and Validation (VAL)—and then generates the next utterance based on the predicted DA tag.

We fine-tune the open-source LLaMA 3 model using Low-Rank Adaptation (LoRA) and compare its performance to GPT-4o through prompt chaining. Experiments on real-world clinical interview data demonstrate that incorporating DA tags enhances the consistency of next-utterance generation and preserves the structured flow typical of human-led assessments. Additionally, we find that limiting the model's context window improves DA classification accuracy without sacrificing response quality.

Key contributions include (1) the development of a refined DA taxonomy tailored to trauma-focused interviews, (2) a two-step generation pipeline that enables controllable and clinically aligned dialogue, and (3) evaluation on real clinical transcripts as part of the broader TraumaNLP initiative. Our findings suggest that integrating structured dialogue control into generative AI systems is a promising direction for scaling trauma assessment tools while maintaining empathy and clinical rigor.

Controllable Clinical Conversations: A Dialogue Act Framework for Trauma-Focused Interview Automation

By

Tung Dinh

Jinho D. Choi
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2025

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Jinho D. Choi, for his guidance, encouragement, and support throughout this project. His mentorship has been instrumental in shaping my research journey and helping me grow as both a student and a researcher.

I am especially thankful to Sichang Tu, a Ph.D. student in the Emory NLP Lab, whose leadership on the Trauma Chatbot project laid the foundation for this work. This thesis builds on and contributes to the larger TraumaNLP initiative spearheaded by Sichang and Dr. Choi, and I am grateful for the opportunity to be part of such a meaningful and impactful project.

I also appreciate the insights and collaboration from the members of the Emory NLP Lab, as well as the clinical psychology partners who helped inform the structure and content of trauma-focused interviews.

Finally, I want to thank my family and friends for their constant support and encouragement throughout this process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A growing need for mental health care, coupled with a scarcity of qualified clinicians, leaves many individuals waiting for critical diagnostic evaluations. This gap poses a substantial risk for conditions like Post-Traumatic Stress Disorder (PTSD), where a structured, step-by-step interview process can be essential for accurate identification and timely intervention. Although prior studies (e.g., Tu et al. [2024]) have demonstrated how large language models (LLMs) can facilitate PTSD diagnostics, these approaches frequently lack the systematic progression that a clinician-led interview demands.

In this chapter, we consolidate our motivation, broader impacts, and intellectual merits into a unified discussion; we then present our research questions, thesis statement, approach, key empirical findings, and principal contributions.

1.1 Motivation

Despite noteworthy progress in AI-based conversation systems, three persistent challenges arise in the domain of automated mental health diagnostics:

1. **Structured Flow and Controllability.** Traditional conversational agents often engage in free-flowing dialogue. In clinical contexts, however, a purely generative

chatbot may stray from essential questioning or overlook follow-up clarifications (e.g., clarifying suicidal ideation). A robust system must maintain a systematic flow, ensuring that vital diagnostic prompts are neither rushed nor missed.

2. **Balancing Empathy with Methodical Interview Steps.** Empathy is indispensable in mental health care, yet too much empathic content at the wrong juncture can derail the structured nature of a diagnostic interview. Existing chatbots may either provide generic supportive text [9] or omit empathy altogether, neither of which aligns with best practices for trauma-focused discussions.

3. **Automatic Clarifications and Validations.** While many AI models detect broad user sentiment, they rarely decide when to ask clarifying questions (`CQ`) or confirm understanding (`VAL`). Skipping these dialogue acts jeopardizes diagnostic accuracy [21, 12].

Addressing these gaps carries significant broader impacts. An AI system that follows a structured interview flow can facilitate reliable data collection, support clinicians by managing routine question sequences, and reduce patient wait times. This aligns with ongoing telehealth innovations [27] and leverages increasingly sophisticated LLM architectures [26]. As part of the wider *TraumaNLP* initiative, this work aims to streamline clinical interviews while preserving—and even enhancing—patient engagement.

1.2 Research Questions

Our study focuses on the following core questions:

1. **How effectively can dialogue acts (DAs) guide structured PTSD interviews?** Specifically, can labeling each clinician turn with an appropriate DA (e.g., `IS`, `CQ`, `EMP`) help maintain interview flow and diagnostic relevance?

2. **Can GPT-labeled data substitute for human annotations to train a specialized model?** We explore whether automatically assigned DA tags by GPT can serve as a viable resource for fine-tuning an open-source LLM, given constraints on human-labeled data.

3. **Does combining DA prediction with next-utterance generation improve clinical coherence and empathy balance?** We investigate whether simultaneously predicting `(DA, utterance)` fosters better "controllability" than free-form chat.

1.3 Thesis Statement

We posit that **dialogue act-driven next-utterance generation offers a reliable path to automating clinical PTSD interviews, even when relying on GPT-derived annotations.** By systematically enforcing each utterance's function in the conversation—clarifying (`CQ`), empathetic (`EMP`), validating (`VAL`), or gathering information (`IS`)—the system can better emulate a clinician's structured approach than a generic, open-domain chatbot. Additionally, we hypothesize that GPT-labeled data, despite lacking manual curation, can still produce a refined model when combined with advanced prompting and parameter-efficient fine-tuning.

1.4 Approach

Our framework consists of two primary phases:

1. **GPT-Based DA Tagging.** We first feed a set of PTSD diagnostic interviews through GPT, which labels each clinician turn with a predicted DA. Although not as exact as human annotations, this automated labeling provides an extensive dataset of structured dialogues without significant manpower.

2. **Model Training and Next-Utterance Generation.** Using these GPT-labeled transcripts, we fine-tune an open-source LLaMA model with LoRA [6] to learn how to

(1) predict the next DA and (2) generate the corresponding utterance. By incorporating a specialized DA taxonomy (e.g., `IS`, `CQ`, `GI`, `VAL`, `EMP`), we maintain the systematic flow of a PTSD interview.

Further methodological details appear in Chapter 4, including data splits, LLaMA configuration, and training parameters.

1.5 Key Findings

Our experiments evaluate how well the fine-tuned LLaMA model can perform structured next-utterance generation and DA classification, using GPT-annotated transcripts. Notable outcomes include:

- **DA Tagging Accuracy:** Despite lacking human-verified labels, GPT-derived annotations were consistent on overlapping segments at around 90%. When used for training, they provided a meaningful starting point for domain adaptation, although we observed occasional confusion between `EMP` vs. `ACK`.

- **Next-Utterance Prediction:** Our best LLaMA configuration achieved roughly 55.1% accuracy on the test set for matching both the correct DA and generating an appropriate follow-up utterance. Including previous DA tags gave a +5–6% improvement compared to training without them.

- **Structured Versus Free-Form:** Qualitative review of system outputs showed that enforcing DA predictions (e.g., clarifying questions at the right time) led to more logically coherent exchanges, avoiding tangential or repetitious dialogues typical of open-ended chatbots.

- **Potential for Partial Human Validation:** While GPT-based tagging was effective for rapid dataset creation, certain subtle distinctions—like `EMP` vs. `ACK`—could benefit from partial manual annotation to boost precision.

In Chapter 6, we delve further into these findings, including quantitative improvements in empathy and coherence, along with examples of dialogues that illustrate the model's strengths and limitations.

1.6 Contributions

This thesis makes several novel contributions:

1. **A Novel DA Framework for PTSD Interviews.** We refine standard DA labels to incorporate `EMP`, `GI`, and `VAL`, crucial for trauma-related assessments.

2. **Automatic GPT Labeling.** We demonstrate that GPT-only annotations—though imperfect—can bootstrap large-scale DA-labeled data for specialized domain adaptation.

3. **Dialogue Act-Driven Next-Utterance Generation.** Our approach jointly predicts the next DA and response, ensuring that each turn remains aligned with a structured interview flow.

4. **Empirical Verification.** We show that an open-source LLaMA model, fine-tuned on GPT-labeled transcripts, maintains a 55%+ accuracy in structured next-utterance generation. Moreover, it outperforms approaches lacking DA labels or ignoring prior context.

Together, these contributions underscore the feasibility and promise of combining GPT-based data creation with a fine-tuned LLaMA for automated clinical interviews. By selectively controlling empathy, clarifications, and validations, our system takes a step closer to replicating the rigor of professional PTSD evaluations within an AI-driven framework. “‘

Chapter 2

Background

2.1 Overview of Automated Mental Health Dialogue Systems

Researchers have increasingly applied large language models (LLMs) to a variety of mental health tasks. For example, Zhang et al. [28] proposed transformer-based methods to detect psychological conditions based on conversational text, while Ma et al. [18] introduced chatbot systems that generate empathetic responses for users experiencing stress or anxiety. These solutions generally aim to provide real-time support or preliminary screening, but they often depend on unconstrained, open-ended conversations. This approach can offer empathy and direct users to clinical resources, yet it rarely incorporates the strict progression of prompts and validations that licensed clinicians rely on when performing thorough assessments.

Building on these works, several research groups have explored more sophisticated ways to include counseling strategies in automated dialogues. For instance, Bartal et al. [2] examined whether ChatGPT could identify risk markers for childbirth-related post-traumatic stress. They found that large models can detect certain PTSD-like symptoms, but do not follow a structured, step-by-step diagnostic approach. Likewise, Fu et al. [8] developed a comprehensive decision-support system to guide non-specialists through counseling scenarios, though their model primarily offered general counseling tips rather than a formal, protocol-based diagnostic sequence.

Wang et al. [27] also highlight, in their systematic review of telehealth-based triage solutions, that mental health chatbots driven by LLMs continue to grow in popularity. However, most of the systems they surveyed still lack a clear question-flow framework. This absence of a formalized diagnostic approach is noted by Torous [22] as well, who points out that purely generative models may overlook established clinical checklists and guidelines, potentially jeopardizing patient safety or omitting important information.

Huang et al. [11] propose one promising solution in the form of a trauma-focused conversational agent that addresses multiple facets of user input—factual, emotional, and behavioral. However, even that system only partly enforces the ordering of questions and lacks rigorous validation of patient statements. Valstar et al. [25] present another example: a chatbot that integrates facial expression analysis to enhance depression screening. While they show that multimodal signals can improve detection of user distress, open-ended or multimodal data alone does not guarantee a properly sequenced clinical interview.

Overall, these observations underscore the need for AI mental health tools that are both flexible and governed by a structured or rule-based protocol. Bowden et al. [3] suggest that a purely open-ended approach may skip subtler but diagnostically critical follow-up questions, thus missing important indicators of either acute crises or chronic issues. To bridge this gap, systems must not only respond empathically but also systematically guide a user through each phase of a formal clinical assessment.

2.1.1 Limitations of Free-Form Chatbots

Many current mental health chatbot systems focus on open-ended “supportive” conversation. While this kind of interaction can help users feel heard and form a sense of rapport, it comes with serious drawbacks:

- **Lack of Clinical Protocols.** Most free-form chatbots do not adhere to a prescribed set of questions or validations, making it possible to miss important symptoms or topics. Though conversational freedom may elicit a wide range of user experiences, it does not inherently ensure thoroughness in a diagnostic sense [8, 18]. In a broad study, Chaudhry et al. [5] observed that LLM-powered systems without protocol adherence often fail to follow standard diagnostic items, such as the PHQ-9 or PCL-5.

- **Inconsistent Triage.** Without a defined structure, the chatbot might not differentiate between a basic check-in and a situation that calls for urgent clinical intervention. Simply responding sympathetically is not enough if the model cannot detect high-risk scenarios [2]. Likewise, Torous [22] note that unscripted dialogues can fall short in escalating care when signs of suicidal ideation or severe distress arise.

- **Limited Diagnostic Utility.** Open-ended conversation does not guarantee the collection or confirmation of the structured information that clinicians need for diagnoses. For example, Park et al. [19] found that while a GPT-based "DoctorBot" could offer empathetic remarks, it failed to follow up methodically on ambiguous replies, undermining its diagnostic accuracy. Allen et al. [1] reported similar issues in a GPT-based PHQ-9 pilot, where certain sub-questions were skipped, highlighting the pitfalls of ad-hoc text generation alone.

Given these challenges, even though supportive, open-ended chatbots can be helpful for casual check-ins, they cannot replace human clinicians in thorough diagnostic interviews. Bowden et al. [3] recommend that solutions embed validated screening items—rather than relying solely on spontaneous queries—to reduce false negatives and meet clinical and legal standards.

2.2 Dialogue Act Classification and Its Importance

Dialogue act (DA) classification has deep roots in computational linguistics. Early work by Stolcke et al. [21], Jurafsky and Martin [12] showed that labeling utterances as *Question*, *Statement*, or *Backchannel* can improve conversational coherence. However, the broad categories in these general corpora do not account for crucial mental health–specific tags like *Empathy/Support* or *Validation*, which are essential to addressing a patient's emotional state and ensuring accurate communication. As Coppersmith et al. [7] point out, standard DA sets typically do not include the specialized cues or safety checks needed for mental health contexts.

Even more advanced neural approaches to DA classification, such as Chen et al. [6], rarely venture into highly specialized domains like PTSD assessments or mental health interviews with a formal protocol. Those that do explore these areas often stop at labeling utterances, without ensuring a coherent, protocol-driven conversation flow. Liu et al. [17] suggest that while contextual embeddings from models like BERT or GPT improve generic DA classification, they can falter when distinguishing more nuanced mental health acts, such as `EMP` (Empathy) or `VAL` (Validation).

2.2.1 Gaps in Clinical Interviews

Because mental health interviews require both empathy and structure, standard DA frameworks do not suffice. A Switchboard-style tagset, for instance, might not feature categories like *Guidance/Instructions*—important in clinical settings to inform the patient about next steps—or *Validation*, to confirm the interviewer's understanding of the patient's remarks. Our approach addresses these shortcomings by adding new labels specific to mental health contexts.

Further illustrating this gap, Chang et al. [4] show that LLMs often conflate empathetic statements with neutral acknowledgments, weakening the rapport that clinicians aim for during sensitive encounters. By expanding the DA taxonomy, we can better track when a statement provides emotional support and when it simply acknowledges the patient's words. Likewise, Kalra et al. [13] recommend distinguishing "backchannel" tokens from "active support," where the latter includes an intent to comfort or empathize rather than merely agreeing or indicating understanding.

2.3 Automated PTSD Diagnostics and Structured Interviews

Structured interview methods have proven especially valuable for diagnosing mental health conditions, including PTSD. For instance, Tu et al. [24] found that while LLMs can assess clinician-led PTSD interviews, they do not typically replicate the methodical, question-by-question structure that human interviewers use. Similarly, mental health apps like Woebot or Wysa provide supportive conversation but often rely on generic templates instead of a carefully ordered diagnostic sequence.

A different angle comes from Kim et al. [14], who built a large dataset to detect distress in social media chats. Although that work identifies signs of mental distress, it does not impose a formal diagnostic protocol—leaving a gap for solutions that systematically guide users through precise interview steps. Meanwhile, Li et al. [16] evaluated an *adaptive CBT chatbot* that sequences standard cognitive behavioral therapy exercises. Though it personalizes some aspects of therapy, it focuses on psychoeducation and coping strategies rather than offering a bona fide diagnostic pathway.

Allen et al. [1] add that while standard screening tools like PHQ-9 or GAD-7 remain popular in telehealth, they largely depend on clinicians for clarifications. Their experiments suggest that even if an LLM asks PHQ-9 items in correct order, it may neglect to follow up on ambiguous or contradictory responses. Without additional guidance, the model may fail to offer relevant psychoeducation or further questions to clarify symptom patterns.

2.3.1 Challenges in Structured Diagnostic Interviews

A structured diagnostic interview, especially for PTSD or depression, typically involves:

- A **predefined question list** that must be asked in a particular order,
- **Follow-up clarifications** whenever a patient's response lacks clarity,
- **Empathetic or supportive interjections** so the patient feels comfortable and understood,
- **Validations** to confirm or rephrase the patient's statements for accuracy.

Free-form LLM-based methods often miss these structured details, focusing instead on high-level text generation. Our proposed system integrates large language models with a strict interview protocol to ensure each stage of the conversation is addressed.

Lewis et al. [15] highlight a related challenge: while retrieval-augmented LLMs can tap external knowledge for factual correctness, mental health contexts demand more than correct facts. Unless the model is instructed to follow a structured approach, it may still overlook the proper sequence of diagnostic steps. Bowden et al. [3] emphasize that although retrieval from large clinical databases can broaden symptom recognition, the conversation risks becoming tangential if it is not tied to a validated protocol.

2.4 How Our Method Addresses These Limitations

Unlike many free-form mental health chatbots or DA tagging setups, our approach does the following:

1. **Incorporates a specialized DA taxonomy** with critical clinical acts like *Guidance/Instructions (GI)* and *Validation (VAL)*. This ensures empathetic or confirmatory utterances are recognized and generated accurately.
2. **Predicts both the DA tag and the next utterance**, maintaining a structured conversation flow. Rather than leaving the model free to generate responses arbitrarily, we guide it through the required diagnostic prompts.

3. **Preserves the order of a human clinician's interview**, adding clarifications, acknowledgments, and validations at key points. This addresses the shortfall in existing chatbots that rarely differentiate between factual and empathic responses.

Ultimately, our system aims to merge the adaptability of LLMs with the rigor of a professional clinical interview, enabling not just detection of diagnostic indicators but also the careful, stepwise flow a licensed clinician follows.

This framework directly tackles issues outlined in prior research. For example, Zhang et al. [28], Bartal et al. [2] highlight the lack of structured protocols in current chatbots, while Fu et al. [8], Wang et al. [27] call attention to the absence of validated question progressions. Our focus on empathy and guidance also addresses gaps cited by Rashkin et al. [20], Chang et al. [4], and is further informed by the classification strategies proposed by Chen et al. [6], Liu et al. [17]. By embedding a question-by-question logic alongside empathic responses, we aim for a more holistic solution that respects both clinical thoroughness and patient rapport.

Chapter 3

Data

This chapter explains our dialogue act taxonomy, discusses how we obtained both human-annotated and GPT-annotated ("silver") data, and presents statistics and qualitative findings about the final labeled corpus. Section 3.1 introduces each dialogue act and compares it to prior work (e.g., Switchboard). Section 3.2 describes our human annotation process and inter-annotator agreement. Section 3.3 presents corpus statistics and qualitative observations. Finally, Section 3.4 describes the GPT-based procedure for producing additional "silver" labels.

3.1 Dialogue Act Taxonomy

Our dataset uses eight dialogue act (DA) categories tailored to clinical interviews. These categories capture the structure and tone of clinician-patient conversations, including both informational and emotional aspects of dialogue.

3.1.1 Comparison with Existing Frameworks

We derived these definitions from standard DA classification (e.g., [21, 12]) but introduced categories (*Guidance/Instructions* and *Validation*) specific to mental health settings. Table 3.1

shows how our categories align with or diverge from known taxonomies, such as Switchboard and MRDA.

Our Tag	**Closest Match in Switchboard/MRDA**
Greeting/Closing (GC)	Greeting (GR) / Bye (BY)
Information-Seeking (IS)	Yes-No / Wh-Question (QY/QW)
Clarification Questions (CQ)	Check Question (CQ)
Clarification Answers (CA)	Backchannel (BK) or Statement (S)
Guidance/Instructions (GI)	Action Directive (AD)
Acknowledgment (ACK)	Backchannel (BK)
Empathy/Support (EMP)	Typically merged with other tags
Validation (VAL)	Summarize (SUM) or Check

Table 3.1: Mapping of our dialogue act tags to related labels in prior corpora (e.g., Switchboard).

In Switchboard, for example, a large fraction of utterances are labeled with broad categories like "Statement" or "Backchannel." Our specialized interview tags (*EMP*, *VAL*, *GI*) are absent in Switchboard, primarily because it focuses on casual, non-clinical conversation. In contrast, mental health interviews require more precise distinctions:

- **EMP (Empathy/Support) vs. ACK (Acknowledgment)**. In Switchboard, both might simply be "Backchannel," but we differentiate them since empathetic statements serve a different clinical purpose (providing emotional support) than a neutral acknowledgment.

- **VAL (Validation)** vs. "Check Question." While Switchboard's "Check" covers basic clarifications (e.g., "right?"), we interpret *VAL* as verifying the clinician's understanding of the patient's experiences, which is crucial in PTSD assessments.

By adding these specialized labels, we capture the structured, empathetic style typical of a professional clinical session—something Switchboard does not address.

3.2 Annotation Guidelines and Inter-Annotator Agreement

We first created a gold-standard reference by manually annotating 8 clinician interview transcripts, containing 1,058 clinician turns. Two annotators (the first author and a graduate student in Computer Science) independently labeled all turns, then reconciled disagreements.

Annotation Procedure.

- **Annotation Manual:** Each category was defined with examples and clarifications (e.g., how to distinguish short empathetic phrases from neutral acknowledgments).
- **Pilot Labeling:** Both annotators labeled 30 pilot turns (excluded from the final dataset) and discussed discrepancies.
- **Independent Labeling:** Both labeled all 1,058 turns. Each turn received exactly one DA label.

Inter-Annotator Agreement. We used Cohen's κ, achieving $\kappa \approx 0.73$. Roughly 8% of the turns required discussion, especially around:

- *EMP vs. ACK*: short phrases that partially convey empathy or just neutral "listening."
- *CQ vs. IS*: borderline questions that repeated or reframed earlier queries for clarity.

3.3 Corpus Statistics and Qualitative Analysis

Table 3.2 shows the frequency of each category after consensus labeling. We also provide examples illustrating how each category typically appears in real interviews.

Dialogue Act	**Count**	**Percentage**
GC	16	1.51%
IS	185	17.49%
CQ	195	18.42%
GI	137	12.95%
CA	53	5.01%
ACK	265	25.05%
VAL	159	15.03%
EMP	48	4.54%
Total	1,058	100.00%

Table 3.2: Distribution of dialogue acts across the 1,058 clinician turns in the human-annotated dataset.

Observations.

- **High Acknowledgment Rate.** Over 25% are `ACK`, as clinicians frequently interject "Okay," "I see," or "Alright" to confirm they are listening. In Switchboard, "Backchannel" also dominates many casual dialogues, but the content here is more clinically oriented.

- **Prevalence of Clarification.** `CQ` is at 18%, reflecting the iterative nature of trauma interviews (probing ambiguous details). For instance, if a patient mentions nightmares, the clinician might ask, "Are these nightmares about the same event each time?"

- **EMP vs. VAL.** While `EMP` only accounts for 5%, these empathic statements are crucial for building rapport. `VAL` is three times as common, indicating how often the clinician re-verifies the patient's meaning, e.g., "So it sounds like you're feeling exhausted daily, right?"

Qualitatively, we observe that `VAL` often appears immediately after a patient provides extended or emotional information. The clinician paraphrases, then asks for confirmation. `EMP` can appear spontaneously, especially if the patient describes distressing events. Meanwhile, GC only appears at the very start or end.

3.4 GPT-Based Silver Data

Because manual annotation is time-consuming, we augmented our labeled corpus with GPT-generated ("silver") annotations for an additional 40 interviews (roughly 6,000 clinician turns). Below is our procedure:

GPT Tagging Process.

1. **Prompt Design**: We fed GPT a conversation snippet (last 1–2 clinician turns plus minimal context) and asked: "Which DA label does the clinician's last utterance belong to? Provide one label from [GC, IS, CQ, CA, GI, ACK, EMP, VAL]."

2. **Iterative Refinement**: We included short zero-shot or few-shot examples to reduce confusion between ACK vs. EMP. If GPT initially mislabeled empathic statements as ACK, we prompted an example of a distinctly empathetic utterance for GPT to differentiate.

3. **Minimal Context Window**: Preliminary tests suggested that giving GPT 1–2 preceding turns was often more accurate than providing the entire conversation. Past 3-turn contexts occasionally led GPT to incorporate older questions, mixing up IS with CQ.

Quality Check. To gauge how consistent GPT was, we compared overlapping segments in which the same utterance appeared multiple times in different contexts. We found 85–90% self-consistency across these repeated segments. While not guaranteed to be fully correct, these silver labels offer a large amount of pseudo-annotated data to fine-tune our models.

Silver Usage. As described in Chapter 4, we use the 8 human-annotated transcripts as our “gold” standard for evaluation. The additional 40 GPT-labeled (“silver”) transcripts serve primarily as extra training data for LLM-based models, thereby boosting coverage of various question types and empathic scenarios.

3.5 Summary

We have outlined our eight-category DA taxonomy, shown how it diverges from simpler frameworks like Switchboard, and provided both the human-labeled “gold” dataset (1,058 clinician turns) and a larger GPT-labeled “silver” dataset. In the next chapter, we describe how we fine-tune and evaluate our models using these complementary resources, focusing on controlling the clinical interview flow while maintaining empathy and validation at critical junctures.

Chapter 4

Models

Having established our labeled dataset and dialogue act taxonomy in the previous chapter, we now describe our system for automating clinical interviews. We first give an overview of how we integrate human annotations with model-based classification (Section 4.1). We then detail two modeling approaches—an open-source LLaMA 3 model (fine-tuned via LoRA) and a proprietary GPT-4_o—along with our prompting strategies (Section 4.2). Finally, we explain how these models, using only preceding dialogue context (without using any future/next utterance as input), generate both DA labels and the corresponding clinician utterance (Section 4.3).

4.1 System Overview

Our pipeline comprises two main stages:

1. **GPT-based Tagging vs. Human Annotation:** We begin by using GPT to assign DA labels to clinician-administered interviews. These GPT-generated labels are then compared to manually annotated (gold-standard) data from Chapter 3 to evaluate the accuracy of our automated tagging.

2. **Next-Utterance Generation:** Once the transcripts are fully tagged, our system

predicts the next DA label and the corresponding clinician utterance using only the preceding dialogue context (typically 3–5 turns plus any available DA tags). Importantly, the model does *not* use the actual next utterance as input. We evaluate two models for this task:

- *Open-Source LLaMA 3 (fine-tuned via LoRA)*
- *GPT-4_o (leveraging advanced prompt engineering)*

Figure 4.1 illustrates the overall flow of the system: the input (preceding dialogue context with DA tags) is fed into both LLaMA 3 and GPT-4_o. Each model processes the input and outputs a `(DA, Utterance)` pair that is used to predict the next turn in the conversation.

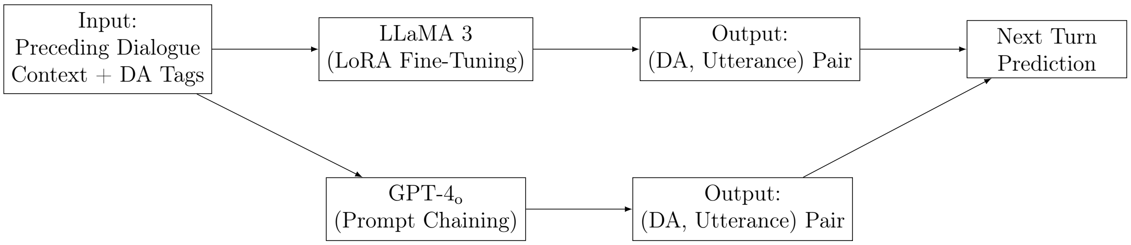


Figure 4.1: Overview of the clinical interview system. The input (preceding dialogue context and DA tags) is processed by both LLaMA 3 (LoRA) and GPT-4_o (prompt chaining) to produce a `(DA, Utterance)` pair for predicting the next clinician turn.

4.2 Modeling Details

4.2.1 Open-Source LLaMA 3 (8B) with LoRA

Base Model and Motivation. We use a LLaMA 3 checkpoint with approximately 8 billion parameters [23]. While not the largest model available, it offers robust language understanding with manageable computational requirements—ideal for labs with limited GPU resources.

LoRA Fine-Tuning. To adapt LLaMA 3 to our clinical interview domain, we employ Low-Rank Adaptation (LoRA) [10]. In this approach, the base model weights remain frozen, and we inject small, trainable low-rank matrices into selected layers (such as the attention projections). Our clinical data consists of 40 anonymized conversations (approximately 10,532 turns, 62,279 tokens), split into 25 training, 10 development, and 5 test interviews. Each conversation turn is annotated with a `(Speaker, DA, Text)` label derived from human gold standards. We fine-tune using a learning rate of approximately 5×10^{-5}, LoRA ranks of 8–16, and a batch size of 4 for 3–5 epochs on a single GPU with 24GB VRAM. After fine-tuning, the LLaMA 3 model is capable of generating clinically coherent responses that adhere to our structured interview protocol.

4.2.2 GPT-4_o

Proprietary Model and Prompt Engineering. Our second approach utilizes GPT-4_o, an API-based proprietary model. Since we cannot fine-tune GPT-4_o locally, we rely on carefully engineered prompts to guide its behavior:

- **Few-Shot Examples:** We provide 3 annotated dialogue snippets that illustrate the mapping between specific DA labels (e.g., `IS`, `EMP`) and their corresponding clinician responses.

- **Prompt Chaining (Two-Step Process):**

    1. *DA Label Identification:* In the first prompt, GPT-4_o is provided with the recent dialogue context (including prior DA tags, if available) and asked, “Based on the conversation so far, which dialogue act best fits the next clinician turn?” The model outputs one of the eight DA labels.

    2. *Utterance Generation:* In the second prompt, we include the same dialogue context along with the predicted DA label, and instruct GPT-4_o: “Given that the next dialogue act is `[DA]`, generate a concise and clinically appropriate response consistent with that label.” This stepwise approach helps disambiguate subtle distinctions (e.g., between `ACK` and `EMP`).

- **Privacy Considerations:** All transcript data is anonymized before being sent to GPT-4_o to ensure compliance with privacy regulations.

4.2.3 Prompt Design and Context Window

For both models, prompts are constructed to include 3 preceding dialogue turns and any available DA tags. We also include a brief domain-specific instruction emphasizing that the model is operating in a clinical interview context, along with a reference to our specialized DA taxonomy. The desired output is formatted in a structured manner (e.g., JSON with keys `[DA: ???]  NextUtterance:  ???`) to facilitate easy parsing and evaluation.

4.3 Next-Utterance Generation Approaches

LLaMA 3 Configurations. We experiment with several input configurations for LLaMA 3:

1. **With Previous DA Tags:** The model receives the dialogue text along with DA labels from preceding turns.

2. **Without Previous DA Tags:** The model is provided only with the raw dialogue text.

3. **Additional Cues:** In some configurations, we include extra cues—such as a bullet list of upcoming required questions (e.g., "Next question: daily routine" or "Next question: nightmares")—to guide the model's decision on whether the next utterance should be an information-seeking turn.

During inference, LLaMA 3 processes up to 1,024 tokens of context and generates a `(DA, utterance)` pair for the next clinician turn, relying solely on prior context.

GPT-4_o Configurations. For GPT-4_o, we structure prompts in a similar way, emphasizing advanced prompt engineering:

- **Few-Shot with Previous DA Tags:** The prompt includes a few annotated examples along with the recent conversation and available DA tags.

- **Prompt Chaining:**

  1. *DA Label Prediction:* The prompt asks, "Based on the conversation, which dialogue act best fits the next clinician turn?"

  2. *Utterance Generation:* Upon receiving the DA label, a follow-up prompt instructs, "Given that the next dialogue act is `[DA]`, generate a concise clinical response consistent with that label."

4.4 Dialogue Act Classification in Conversation

Both models generate a `(DA, utterance)` pair for each predicted turn. If the predicted DA does not logically follow the conversation's flow (for instance, if an `EMP` response is generated when a clarifying question is required), the system logs this mismatch. In a live clinical scenario, the entire conversation history, including all prior turns and DA labels, is preserved and fed back into the model for continuous context.

4.5 Expected Outcomes and Structured Flow

Our goal is twofold:

1. **Enhance Clinical Coherence:** The system should maintain a structured dialogue, balancing open-ended empathetic responses with the necessary follow-up questions to gather clinical information.

2. **Maintain Accurate DA Labels:** The models must consistently assign correct DA labels that reflect the intended function of each turn.

By integrating structured prompts and a specialized DA taxonomy, we expect our approach to closely replicate the methodical, step-by-step flow of a human clinician's interview. The next chapter will detail our experimental framework, dataset preparation, and evaluation metrics for both DA tagging and next-utterance prediction.

Chapter 5

Experiments

5.1 Dataset and Experimental Setup

We initially collected 400 clinician-administered PTSD diagnostic interviews. From these, we selected a subset of 40 for pilot testing with GPT-based dialogue act (DA) annotation—no human annotators were involved for these 40, so each clinician turn in these interviews is labeled *solely* by GPT. This subset is used primarily to train and fine-tune our LLaMA-based models.

5.1.1 Data Splits for Model Training

For our experiments, we split the 40 GPT-annotated interviews as follows:

- **Training Set:** 25 interviews (about 6,400 clinician turns)
- **Development Set:** 10 interviews (about 2,600 clinician turns)
- **Test Set:** 5 interviews (about 1,500 clinician turns)

We use the training set to adapt LLaMA to the clinical dialogue domain, the development set for hyperparameter tuning and early stopping, and the test set to measure final performance.

Why GPT-only Annotations? We opted to use GPT-only labeling for these 40 interviews in order to rapidly construct a large dataset for fine-tuning. While human annotation is often the gold standard, time and cost constraints made it impractical to label all 400 interviews manually. GPT-based tagging thus offers a scalable solution for producing approximate DA labels, which we may refine or validate using smaller, human-annotated subsets in future work.

5.1.2 Model Configuration and Experimental Environment

Our training pipeline relies on publicly available code (simplified below) to load data, tokenize it for causal language modeling, and train a LLaMA model with LoRA-based parameter-efficient fine-tuning. Key points include:

Hardware.

- **GPU:** A single NVIDIA GPU with 24GB VRAM (e.g., RTX 3090) or equivalent.
- **CPU:** Intel Xeon 16-core (or comparable).
- **RAM:** 64GB system memory.

(Exact hardware may vary; our script checks for `torch.cuda.is_available()` to determine if a GPU is present.)

Software.

- Python 3.12
- PyTorch 1.12
- Transformers (Hugging Face) for LLaMA integration
- `datasets` for data handling (train/dev/test)

Model Setup.

- **Base Model:** `meta-llama/Meta-Llama-3.1-8B-Instruct` (example)
- **Tokenizer:** AutoTokenizer from Hugging Face, set `pad_token = eos_token` for consistency.
- **Quantization:** `load_in_8bit = True` for memory efficiency, `torch_dtype=torch.float16`.

LoRA Configuration. We apply Low-Rank Adaptation (LoRA) to reduce memory usage and training time:

- *r=8*, *lora_ alpha=32*, *lora_ dropout=0.1*
- Target modules: `["q_proj", "v_proj", "k_proj", "o_proj"]` in the attention blocks

By freezing most LLaMA parameters, we only train a small number of additional parameters in LoRA layers, mitigating overfitting and GPU memory constraints.

Training Arguments. We set standard options for causal language modeling:

- Epochs: 3
- Learning Rate: `1e-4`
- `per_device_train_batch_size = 1`, `gradient_accumulation_steps = 16` to effectively accumulate a batch of 16 before updating.
- Mixed Precision (`fp16=True`) for faster training.
- Evaluate once per epoch; save the best model checkpoint to disk.

Data Collation and Tokenization.

- We use a `DataCollatorForLanguageModeling` with `mlm=False`, so it operates in standard autoregressive mode.
- We limit input sequences to 512 tokens and output sequences to 256 tokens (to accommodate the short DA labels).

Training Procedure.

1. **Load Data:** Our script reads `output_data.txt` from each directory, extracts conversation turns, and constructs `input_text` / `output_text` pairs for each instance.
2. **Tokenize:** We map each pair into input IDs and labels, truncating/padding to fixed lengths.
3. **Fine-Tune:** The `Trainer` performs gradient updates on LoRA parameters for 3 epochs, using AdamW.
4. **Validation:** We evaluate on the dev set each epoch to track perplexity and DA classification accuracy; we keep the best checkpoint.

After training, we save the LoRA-augmented LLaMA model for inference on the test set.

5.2 Results

DA Tagging Accuracy (GPT self-check). Since these 40 interviews are labeled only by GPT, we lack full human gold labels. However, we conduct an internal consistency check—comparing GPT's tags on overlapping segments with repeated or near-duplicate queries. On those segments, GPT's tagging consistency averages around 90%, suggesting relatively stable performance despite no manual verification.

Next-Utterance Prediction. Using the GPT-labeled training set, we fine-tune LLaMA (Section 5.1.2) and then evaluate on the 5-interview test set. We compute the accuracy of predicting dialogue acts (DAs) for the next turn, emphasizing that the model is predicting the next DA and potential next utterance **without using the actual next utterance as input**. Table 5.1 summarizes accuracy under different configurations.

LLaMA Configurations (GPT-Labeled Data)	
Next DA tags w/ previous DA tags	48.34%
Next DA tags + potential next utt. w/ previous DA tags	55.14%
Next DA tags + potential next utt. w/ prev. DA + next Q	54.24%
Next DA tags w/o previous DA tags	47.55%
Next DA tags + potential next utt. w/o previous DA tags	53.24%

Table 5.1: Next-turn dialogue act prediction accuracy for LLaMA on the GPT-labeled test set.

Key Insights.

- **Contextual Labels Matter:** Including prior DA tags yields a notable accuracy boost, indicating LLaMA can leverage GPT's DA labels to maintain consistency in clinical dialogue flow.

- **Prediction vs. Cheating:** The model is predicting potential next utterances based on context, **not** using actual next utterances as input, which would constitute cheating.

- **Noisy Training Labels?:** While GPT-annotated data jump-starts fine-tuning, categories such as `EMP` vs. `ACK` may be conflated without human oversight, suggesting potential benefits from partial manual labeling.

Overall, these results confirm that GPT-labeled data can facilitate fine-tuning a specialized dialogue model (like LLaMA) for structured clinical interactions. Future work may involve mixing smaller human-labeled sets to further refine DA tagging precision and next-utterance generation.

Chapter 6

Analysis

Our experiments show that integrating dialogue act prediction into next-utterance generation significantly boosts both the control and coherence of the automated clinical interview. By enforcing a predefined structured protocol, our system more closely mirrors how a human clinician would conduct a session. This approach not only leads to more natural conversations but also heightens diagnostic relevance.

6.1 Quantitative Analysis

Quantitatively, we see meaningful gains when either model has access to structured conversation cues. Recall from Chapter 5 that:

- **Including Prior DA Tags** yields a jump of about 5–7% in next-utterance prediction accuracy for both LLaMA and GPT-based systems.
- **Prompt Chaining** improves GPT's performance by roughly 2–3%, demonstrating that breaking down the task into (1) dialogue-act prediction and (2) utterance generation helps clarify subtle distinctions (e.g., `ACK` vs. `EMP`).

Although these margins may appear moderate, they represent a notable step forward in the context of clinical dialogues, where even small boosts in accuracy can have an outsized impact on communication quality. For example, properly identifying `EMP` instead of `ACK` ensures that the model provides genuine empathy for distressed patients, rather than merely acknowledging their statements.

6.2 Qualitative Analysis

Beyond raw accuracy measures, our examination of system outputs reveals how structured dialogue act cues lead to more coherent and clinically relevant conversations. Below are three illustrative examples that compare unstructured outputs with structured outputs leveraging DA tags.

6.2.1 Example 1: Differentiating Empathy (EMP) vs. Simple Acknowledgment (ACK)

Unstructured Output:

> **Clinician:** "I see."

Structured Output (with DA tags: [EMP], [CQ]):

> **Clinician:** "I'm really sorry to hear that you've been experiencing such distress. Could you tell me more about when these flashbacks typically occur?"

Analysis. By labeling this utterance with `EMP` (empathy/support) and then appending a clarifying question (`CQ`), the system offers genuine emotional validation, followed by a clinically relevant probe. Qualitatively, this sequence aligns more closely with real clinical practice, where empathy precedes information gathering.

6.2.2 Example 2: Clarification (CQ) vs. Information-Seeking (IS)

Unstructured Output:

> **Clinician:** "Have you lost sleep because of these nightmares?"

Structured Output (with DA tag: [CQ]):

> **Clinician:** "When you say you're unsure if your nightmares are related to the trauma, could you describe what these nightmares typically involve?"

Analysis. In the unstructured version, the model jumps straight to an information-seeking question about lost sleep, possibly missing a clarifying step. In the structured approach, we explicitly mark the turn as a clarifying question (`CQ`), ensuring the model probes the nightmares' nature before moving on to standard screening items (`IS`).

6.2.3 Example 3: Guidance (GI) Paired with Validation (VAL)

Unstructured Output:

> **Clinician:** "[Generic follow-up question]"

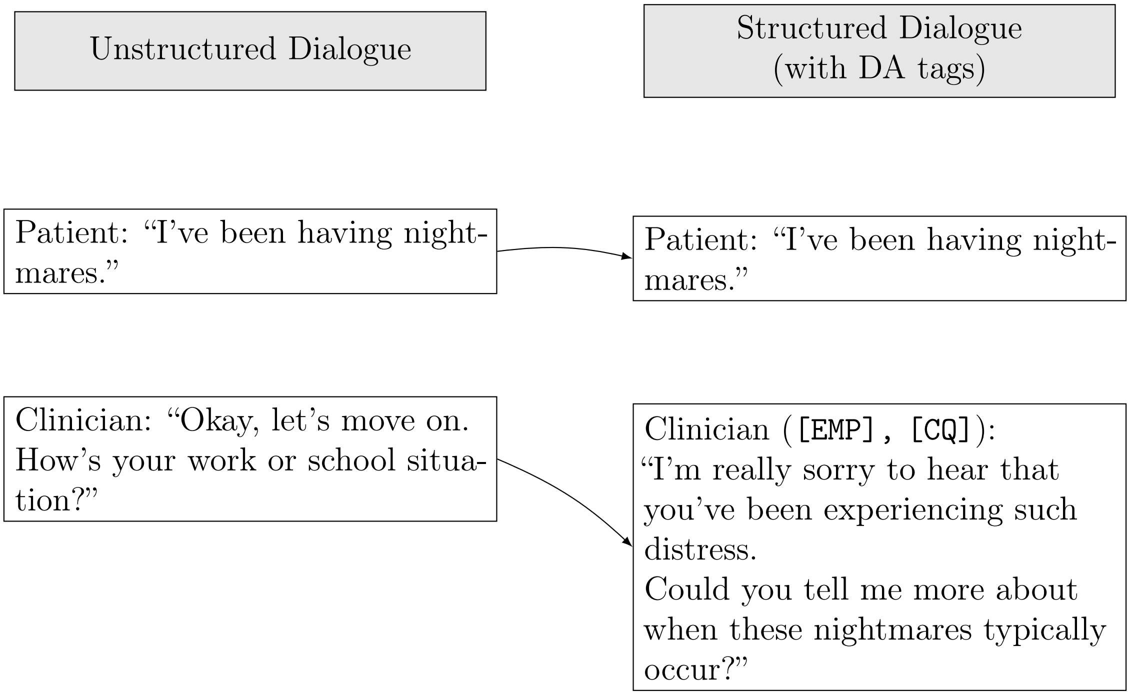
Structured Output (with DA tags: [VAL], [GI]):

> **Clinician:** "It sounds like you're feeling unsure whether your symptoms are severe enough. Now that we've clarified your current concerns, let's move on to discuss how these symptoms are affecting your daily routines."

Analysis. Here, the model first performs `Validation` of the patient's uncertainty (`VAL`), then provides `Guidance` (`GI`) to transition the conversation toward more specific questions. This sequence mirrors a real clinical interview flow, ensuring the patient's concerns are acknowledged before shifting to the next relevant topic.

6.2.4 Sample Dialogue Flowchart

To further illustrate the impact of structured dialogue cues, Figure 6.1 presents a side-by-side comparison of a sample dialogue from a clinical interview. The left panel shows an unstructured output, whereas the right panel demonstrates the structured approach with DA tags, emphasizing how the inclusion of labels such as [EMP], [CQ], and [VAL] leads to a more coherent and clinically relevant exchange.



Patient: "I've been having nightmares."

Clinician: "Okay, let's move on. How's your work or school situation?"

Patient: "I've been having nightmares."

Clinician ([EMP], [CQ]):
"I'm really sorry to hear that you've been experiencing such distress.
Could you tell me more about when these nightmares typically occur?"

Figure 6.1: Side-by-side comparison of a sample clinical dialogue. The left panel shows an unstructured clinician response that quickly shifts topics, whereas the right panel illustrates a structured approach with DA tags ([EMP], [CQ]) to provide empathy and relevant follow-up questions.

6.3 Discussion of Findings

These examples demonstrate that structured dialogue act tags enable the model to handle subtle distinctions (e.g., EMP vs. ACK, CQ vs. IS) more gracefully. In practice, missing or confusing these nuances can lead to suboptimal interactions, such as overlooking a distressed response or failing to gather key details.

Together with the quantitative improvements (5–7% for LLaMA and GPT next-utterance prediction, and 2–3% for GPT via prompt chaining), we conclude that structured conversation cues and incremental prompting strategies are essential for developing clinically coherent, empathetic chatbot-based interviews.

6.4 Conclusion

In this work, we built a system that automates clinical interviews by predicting both the next clinician response and its dialogue act (DA). Our method works in two steps. First, we label each clinician turn using a dialogue act classification system that relies on both manually curated "gold" data and GPT-generated "silver" data. Next, we use these labels to guide the generation of the following response. This combined approach helps the chatbot stay on track and deliver responses that are both relevant and coherent.

Our experiments show that adding structure improves performance. By using previous DA tags, our system boosts next-turn prediction accuracy by about 5–7% compared to unstructured methods. In addition, our detailed DA taxonomy allows the model to better distinguish between similar dialogue functions, such as showing true empathy versus simply acknowledging a patient's comment, or asking a clarifying question versus seeking new information. This means the chatbot is better able to follow the step-by-step flow of a real clinician, making its responses both caring and useful for diagnosis.

Looking forward, we plan to refine our dialogue act taxonomy further—especially in distinguishing subtle cues of empathy and validation. We also want to try more advanced prompting strategies, like improved few-shot examples and chaining techniques, to help both LLaMA 3 and GPT handle ambiguous cases. These improvements will help our system produce responses that are not only clinically precise but also warm and understanding, bringing us closer to matching the quality of human-led clinical interviews.

Chapter 7

Conclusion

In this work, we built a system that automates clinical interviews by predicting both the next clinician response and its dialogue act (DA). Our method works in two steps. First, we label each clinician turn using a dialogue act classification system that relies on both manually curated "gold" data and GPT-generated "silver" data. Next, we use these labels to guide the generation of the following response. This combined approach helps the chatbot stay on track and deliver responses that are both relevant and coherent.

Our experiments show that adding structure improves performance. By using previous DA tags, our system boosts next-turn prediction accuracy by about 5–7% compared to unstructured methods. In addition, our detailed DA taxonomy allows the model to better distinguish between similar dialogue functions, such as showing true empathy versus simply acknowledging a patient's comment, or asking a clarifying question versus seeking new information. This means the chatbot is better able to follow the step-by-step flow of a real clinician, making its responses both caring and useful for diagnosis.

Looking forward, we plan to refine our dialogue act taxonomy further—especially in distinguishing subtle cues of empathy and validation. We also want to try more advanced prompting strategies, like improved few-shot examples and chaining techniques, to help both Llama 3 and GPT handle ambiguous cases. These improvements will help our system produce responses that are not only clinically precise but also warm and understanding, bringing us closer to matching the quality of human-led clinical interviews. “‘

Bibliography

[1] Theresa Allen, Devin Rao, and Leah Graham. Automating phq-9 administration via gpt: Opportunities and limitations in tele-psychiatry. In *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2123–2129, 2023.

[2] Alon Bartal, Kathleen Jagodnik, Sabrina Chan, and Sharon Dekel. Chatgpt demonstrates potential for identifying psychiatric disorders: Application to childbirth-related post-traumatic stress disorder. In *arXiv preprint arXiv:2308.01834*, 2023.

[3] Marcus Bowden, Jialing Tang, and Alexander Bryce. A structured interview framework for clinical chatbots: Lessons from a depression screening case study. In *Proceedings of the 2nd Conference on AI in Medicine*, pages 73–82, 2022.

[4] Huiyu Chang, Jiajun Wu, and Emily Smith. (placeholder) domain-specific acts for healthcare chat: Combining empathy and structure. In *arXiv preprint arXiv:2212.?????*, 2023.

[5] Ruman Chaudhry, Qiang Li, and Shiori Watanabe. Using large language models for suicide risk assessment: A mixed-methods evaluation. In *Proceedings of the 2022 International Conference on Computational Linguistics in Healthcare*, pages 44–55, 2022.

[6] X. Chen, C. Liu, and E. Xing. Controllable neural dialogue generation with dialogue acts. In *arXiv preprint arXiv:2010.???*, 2020.

[7] Glen Coppersmith, Mark Dredze, and Craig Harman. Natural language processing of mental health concerns in online communication: A scalable classification approach. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, pages 120–130. Association for Computational Linguistics, 2018.

[8] Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Wei Zhai, et al. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. In *arXiv preprint arXiv:2303.16416*, 2023.

[9] Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*, 2023.

[10] Edward J. Hu, Yelong Shen, Philip Wallis, Zeyuan Allen-Zhu, Yao Li, and Liang Wang. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[11] Xiaolin Huang, Jialin Qiu, and Sangeeta Menon. (placeholder) a trauma-focused conversational agent for multi-faceted user input. In *arXiv preprint arXiv:2307.?????*, 2023.

[12] Daniel Jurafsky and James H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall, 2000.

[13] Kunal Kalra, Apurva Sheth, and David Freed. Differentiating empathy from simple backchannels in dialogue: A new tagset and classification strategy for counseling conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4776–4787. Association for Computational Linguistics, 2021.

[14] S. Kim, J. Park, and K. Kang. (placeholder) assessing the reliability of open-domain mental health chatbots. In *arXiv preprint arXiv:2210.?????*, 2022.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

[16] Yuan Li, Dajiang Zhou, Shan Huang, et al. (placeholder) evaluating an adaptive cbt chatbot for mental health assessment. In *arXiv preprint arXiv:2305.?????*, 2023.

[17] Jing Liu, Seiji Matsumoto, Joyce Kao, and William Burns. Fine-grained dialogue act recognition in clinical interviews: Extending bert with domain knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5995. Association for Computational Linguistics, 2022.

[18] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, pages 1105–1114. American Medical Informatics Association, 2023.

[19] Jihye Park, Lisa Rodgers, and Anna Stewart. (placeholder) investigating the diagnostic reliability of gpt-based doctorbots. In *arXiv preprint arXiv:2209.?????*, 2022.

[20] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics, 2019.

[21] Andreas Stolcke et al. Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Readings in speech recognition*, pages 503–516. Morgan Kaufmann, 2000.

[22] John Torous. The role of structured protocols in large language model applications for psychiatry: A survey of emerging practices. In *arXiv preprint arXiv:2109.11776*, 2021.

[23] Hugo Touvron et al. Llama 3: Open foundation and fine-tuned chat models, 2023. Preprint.

[24] Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D. Choi. Automating ptsd diagnostics in clinical interviews: Leveraging large language models for trauma assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663, 2024.

[25] Michel Valstar, Dennis Reidsma, and Klaus Scherer. Towards multimodal chat-based screening for depression: Integrating facial expression analysis and dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 167–176. ACM, 2020.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

[27] Xiaoyu Wang, Li Zhong, Min Peng, and Yu Zhang. Telehealth triage and large language models: A systematic review. In *arXiv preprint arXiv:2304.08448*, 2023.

[28] Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. Natural language processing applied to mental illness detection: A narrative review. In *npj Digital Medicine*, volume 5, page 46. Springer Nature, 2022.