# Generating Unexpected yet Relevant User Dialog Acts

**Lucie Galland**
ISIR
Sorbonne University
Paris, France
galland@isir.upmc.fr

**Catherine Pelachaud**
CNRS - ISIR
Sorbonne University
Paris, France
pelachaud@isir.upmc.fr

**Florian Pecune**
CNRS - SANPSY
Bordeaux University
Bordeaux, France
pecune@u-bordeaux.fr

## Abstract

The demand for mental health services has risen substantially in recent years, leading to challenges in meeting patient needs promptly. Virtual agents capable of emulating motivational interviews (MI) have emerged as a potential solution to address this issue, offering immediate support that is especially beneficial for therapy modalities requiring multiple sessions. However, developing effective patient simulation methods for training MI dialog systems poses challenges, particularly in generating syntactically and contextually correct, and diversified dialog acts while respecting existing patterns and trends in therapy data. This paper investigates data-driven approaches to simulate patients for training MI dialog systems. We propose a novel method that leverages time series models to generate diverse and contextually appropriate patient dialog acts, which are then transformed into utterances by a conditioned large language model. Additionally, we introduce evaluation measures tailored to assess the quality and coherence of simulated patient dialog. Our findings highlight the effectiveness of dialog act-conditioned approaches in improving patient simulation for MI, offering insights for developing virtual agents to support mental health therapy.

## 1 Introduction

The demand for mental health services has surged in recent years, resulting in a significant gap between demand and available resources (Cameron et al., 2017). Consequently, patients often face prolonged wait times before accessing therapy (Cameron et al., 2017; Denecke et al., 2020). To mitigate this challenge, virtual agents capable of emulating Motivational Interviews (MI) have emerged as a potential solution, offering immediate support, especially in therapy modalities requiring multiple sessions (Fiske et al., 2019). These agents are not meant to replace therapists but rather supplement therapy. Designing such agents can follow either a rule-based or data-driven approach. Rule-based systems entail complex development and the creation of intricate rule sets. Conversely, data-driven methods leverage large datasets to train models, potentially yielding optimal performance but requiring substantial data. Given the difficulty in obtaining therapy data, patient simulation emerges as a viable alternative for generating large quantities of synthetic data, traditionally generated at the dialog act level. However, patient simulation relies on a high-quality simulation capable of generating dialog acts that differ enough from the existing dataset to create novel data and be contextually and syntactically correct. Such a simulation should also explore all the possible dialog acts and produce diversified ones. However, the new data should also respect the structure of a real dialog. The objective is not merely to copy the observed behaviors in the dataset but to generate new ones with the following properties: be diversified, syntactically correct, and coherent in the context of the dialog. Evaluating such a simulation poses challenges because traditional accuracy metrics for supervised models may not suffice, as they measure only how accurately the original data is reproduced. Indeed, a generated dialog act may be different from the ones observed in the data but still be syntactically and contextually correct. This is particularly true in open dialog settings, such as MI, where the user's goal is unclear, unlike in task-based scenarios like booking systems. This paper investigates modeling methods to generate such patient dialog acts and explores evaluation methods for open-ended dialog user simulations.

Our contributions include:

- Development of a dialog manager for simulating motivational interviewing patients.
- Proposal of evaluation measures for open-ended dialog user simulation.



*Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 192–203
September 18–20, 2024. ©2024 Association for Computational Linguistics

## 2 Background and Related Works

*Motivationnal Interviewing (MI)* is a collaborative communication style employed by therapists and educators to foster change. The goal of MI is to drive the patient towards wanting to change one of their unhealthy behaviors without giving them any solutions (Miller and Rollnick, 2012). The patient realizes what and how to change through a series of dialog strategies characteristic of MI, such as reflection, where the therapist reformulates what the patient just said to help them take a new perspective. In MI, therapists also create relationships with patients through social behaviors such as empathic reactions (Jani et al., 2012).

*Virtual agents in healthcare* is a developing area of research due to their proven effectiveness and acceptance as support tools (Mercado et al., 2023; Bickmore et al., 2009, 2018). Recently, MI conversational agents have been created in the form of chatbots (Fitzpatrick et al., 2017) and embodied conversational agents (Bickmore et al., 2018). These agents have shown promise in providing social support alongside therapy (Ring et al., 2016). Some studies have also investigated adding empathetic behavior (Lisetti et al., 2013) and humor (Olafsson et al., 2020a) to these agents.

*Adaptability* in such agents is important, as each patient requires a tailored approach (Galland et al., 2024a). One way of managing dialogs is by using a rule-based dialog manager, which necessitates expert knowledge and a complicated set of rules (Pecune et al., 2020). On the contrary, a data-driven dialog manager learns from data to anticipate the best therapist dialog acts based on context (Olafsson et al., 2020b). However, this approach requires a significant amount of data that is difficult to obtain due to the private nature of therapy.

*Simulating users* has emerged as a viable approach to generate simulated data for training conversational systems (Schatzmann et al., 2006). Traditionally, users are simulated through a dialog manager utilizing statistical inference (Schatzmann et al., 2007), inverse reinforcement learning (Chandramohan et al., 2011), or transformers (Lin et al., 2021, 2022) to select the next dialog act, enabling controllability and integration of expert or task-specific knowledge. Recently, social aspects have been incorporated into such user simulations, featuring different user types (Pecune et al., 2020) and engagement simulations (Galland et al., 2022). However, these techniques mainly focus on limited task domains and rely on template-based utterance generation. This approach is impractical for open application domains such as MI, where patients' responses can vary. The emergence of Large Language Models (LLMs) has led to a new approach to simulated patients that addresses this challenge. This method uses LLMs as black boxes for user simulation, with the model generating the next patient utterance based on the dialog context (Chiu et al., 2024). However, this technique lacks controllability and may significantly diverge from actual data without being coherent. We propose a hybrid approach that utilizes conditioned LLMs to overcome these issues.

*Evaluating simulated users* poses challenges as simulated users are intended to create novel data with our desired properties (i.e., syntactically correct, coherent in the dialog context, and diversified). Existing works mainly evaluate their simulated users using accuracy metrics such as the F1 score (Lin et al., 2022; Schatzmann et al., 2007) that measures only the similarity with ground truth leaving aside novelty. Another commonly used evaluation method involves computing the task success rate of systems trained with simulated users (Lin et al., 2022, 2021). While this method works well for task-based dialog, it is more complicated to apply to open-domain dialogs such as MI where social acts matter also. Another evaluation method is to compare the distribution of the characteristics of generated dialogs with those of the ground truth, such as dialog length (Chandramohan et al., 2011) or dialog act distribution (Galland et al., 2022). However, these metrics do not capture the quality of the generated data. Therefore, we propose metrics measuring how well user simulators fit the data and their capabilities to generate novel, syntactically and contextually correct data. To this aim, we adapt the serendipity measure to the dialog system domain.

In the subsequent sections, we provide the context of our study (Section 3), introduce our proposed method (Section 4), present our proposed measures (Section 5), and evaluate objectively and subjectively the method (Section 6.2).

## 3 Context

Motivational Interviewing (MI) is a therapeutic approach that prioritizes collaboration and fosters behavioral change. Within MI sessions, therapists employ various strategies to facilitate patients' expression of motivation for change (Miller and Rollnick, 2012). Consequently, the study of MI focuses on the language of change. The language of change is defined in the motivational interviewing skill code (MISC) (Miller et al., 2003) that classifies patient behaviors into three categories: **Change Talk (CT)**: reflecting actions toward behavior change, **Sustain Talk (ST)**: reflecting actions away from behavior change, **Follow/Neutral (F/N)**: unrelated to the target behavior. This classification of the client's multimodal behavior is interesting as it predicts the therapy outcome. Indeed, **ST** is associated with poorer treatment results (Magill et al., 2014). Furthermore, **CT** is linked to risk behavior reduction during follow-up assessments (Magill et al., 2018). These results make MISC a promising tool for studying the efficacy of MI.

### 3.1 Dataset

This paper relies on the HOPE dataset (Malhotra et al., 2022), a corpus of transcribed therapy sessions. HOPE is composed of $\sim$12.9K utterances departed into 212 sessions. The sessions are publicly available videos collected from the web. The transcripts were produced automatically and then corrected by the authors (Malhotra et al., 2022). The data is separated into a train (85%), validation (5%), and test set (10%).

#### 3.1.1 Dialog acts

Each utterance is classified into a dialog act to label the corpus in terms of dialog acts using a schema and classifier presented in (Galland et al., 2024a) and derived from (Malhotra et al., 2022). Patient's utterances are classified into nine different dialog acts presented in Table 1, and therapist's utterances are separated into 13 different dialog acts presented in Table 2. There are 22 dialog acts in total; some of these dialog acts are oriented towards change ("Changing unhealthy behavior", "Sharing positive feeling or emotions") while others are oriented towards sustain ("Sustaining unhealthy behavior", "Sharing negative feeling or emotions"). The classifier is based on a few-shot prompting of Mistral 7B instruct, an open-source LLM, and yields an F1 score of 0.69 for the client and 0.7 for the therapist, which is equivalent to state-of-the-art results for such task (Malhotra et al., 2022).

### 3.2 Patient types

Patients in MI may manifest diverse reactions concerning their readiness to alter behaviors. Patients engaged in MI sessions may be classified into distinct types, as outlined in (Galland et al., 2024a), categorized as Open-to-Change, Receptive, or Resistant-to-Change:

- **Open-to-Change**: These patients are more willing to alter unhealthy behaviors.
- **Resistant-to-Change**: Patients in this category are inclined to maintain unhealthy behaviors.
- **Receptive**: Characterized by initially displaying low motivation to change, receptive patients transition towards a high motivation to change their unhealthy behaviors towards the end of the conversation.

These typologies capture variances in both patient and therapist behavior (Galland et al., 2024a). Consequently, the ability to simulate these three distinct patient types would be advantageous for training

| | Definition |
|---|---|
| Changing unhealthy behavior | The patient explicitly expresses their willingness to change |
| Sustaining unhealthy behavior | The patient explicitly expresses their unwillingness to change |
| Sharing negative feeling or emotion | The patient shares a negative feeling or vision of the world |
| Sharing positive feeling or emotion | The patient shares a positive feeling or vision of the world |
| Realization or Understanding | The patient realizes or understand something about their problem |
| Share personal information | The patient shares factual personal information about their situation or background |
| Greeting or Closing | The patient opens or closes the conversation |
| Backchannel | The patient acknowledges that they heard the last therapist's statement |
| Asking for medical information | The patient asks for medical information |

Table 1: List and definitions of patient's dialog acts (Malhotra et al., 2022; Galland et al., 2024a)

| | Definition |
|---|---|
| **Task oriented Dialog acts** | |
| Ask for consent or validation | The therapist checks that their last statement was correct or that the patient consented to move forward |
| Medical Education and Guidance | The therapist provides the patient with medical or therapeutic facts |
| Planning with the patient | The therapist builds a plan with the patient to modify their unhealthy behavior/thoughts |
| Give Solution | The therapist provides the patient with solutions to solve their problem |
| Ask about current emotions | The therapist asks the patient what they are feeling during the therapy session |
| Invite to shift outlook | The therapist asks the patient to imagine their reaction to a future event or to change their perspectives on a past even |
| Ask for Information | The therapist asks the patient factual information about their background or situation |
| Reflection | The therapist summarizes or reformulates the patient statement without judgment |
| **Socially oriented Dialog acts** | |
| Empathic reaction | The therapist expresses empathy to the patient |
| Acknowledge progress and encourage | The therapist praises the patient for their achievements or encourages them |
| Backchannel | The therapist acknowledges that they heard the last patient's statement |
| Greeting or Closing | The therapist open or closes the conversation |
| Experience Normalization and Reassurance | The therapist normalizes the patient experience and reassure them |

Table 2: List and definitions of therapist's dialog acts (Malhotra et al., 2022; Galland et al., 2024a)

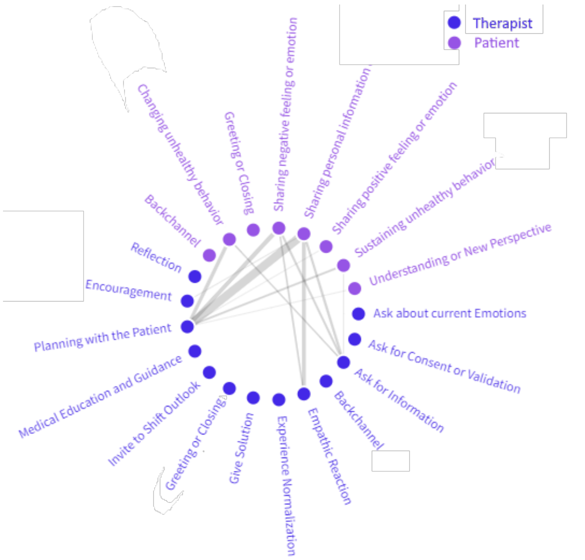


Figure 1: Local (therapist DA to patient DA) patterns in the HOPE dataset.
A pattern is a sequence of dialog acts that appear at least twice in the dataset. The width of the lines is proportional to the number of occurrences of the patterns in the dataset.

our virtual therapist's dialog model in subsequent stages.

Patients tend to gravitate towards or away from change during a dialog, influenced by their types. This sets a broad (or seasonal) trend in the evolution of dialog acts, with an average increase or decrease in change-oriented and sustain-oriented dialog acts, representing the user's inherent goal. Concurrently, we observe specific local patterns in dialog acts, where certain therapist utterances are often followed by particular patient responses (see Figure 1). For instance, an empathic reaction from the therapist often leads to sharing personal information or the patient's negative feelings. Consequently, our dialog manager must be capable of capturing both the global trends in dialog acts and these localized patterns.

### 3.3 Patient simulation

We aim to develop a patient simulation capable of generating natural and coherent actions akin to an actual patient's without exactly copying the corpus, thus creating novel data. To this intent, we propose to simulate the patient by combining a dialog manager that selects the next dialog act and a conditioned LLM that generates the associated utterance. This paper focuses on the development and evaluation of the dialog manager (see Fig.2).

The conditioned utterances are generated

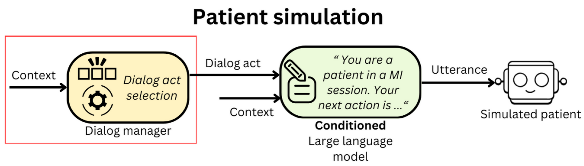


Figure 2: Patient simulation approach

through Mistral 7B instruct, an open-source large language model that can be run offline. Using a few shot-learning techniques, the model is prompted to act as a patient in an MI simulation and to perform a given dialog act. The definition of the dialog act, as well as some examples, are provided in the prompt. The associated prompt is available in Appendix A.1, and the related code is available on Github[1]. Examples of generated utterances are visible in Appendix A.2. This generation method was validated in (Galland et al., 2024b). The generated utterances are correctly classified into the instructed dialog acts. Moreover, utterances generated with the ground truth dialog act condition are perceived as more coherent and natural than those generated by an unconditioned LLM (presented in Section 6.1), motivating the development of a dialog manager that produces appropriate dialog acts to condition the LLM. In the following of this paper, we focus on presenting our dialog manager and its evaluation.

## 4 Dialog Manager

In this section, we discuss the architecture of our dialog manager that selects the next dialog act given the context (see Fig. 2).

### 4.1 DA2Vec

We introduce DA2Vec, leveraging a Word2Vec approach (Mikolov et al., 2013), to represent dialog acts. Each dialog act is encoded as a vector within a latent space, facilitating proximity for dialog acts frequently occurring together in conversations. Our model employs a window size of 3 and an embedding dimension of 8.

### 4.2 Global Model Architecture

The global model architecture, depicted in Figure 3, operates on three types of input: the last three speaking turns in transcript (Text) form, the 20 most recent speaking turns in dialog act (DA) form and the type of user to simulate (akin to its goal).

[1] https://github.com/l-Galland/Patient_simulation

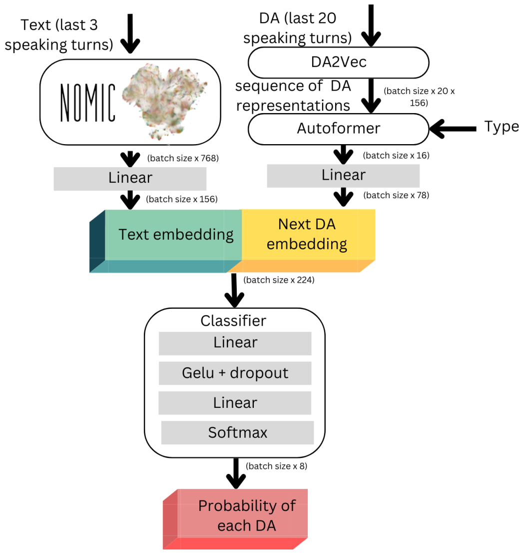


Figure 3: Dialog manager architecture

Textual data undergoes embedding using the Nomic embedding's text model version 1.5 (Nussbaum et al., 2024). The DA context is embedded through DA2Vec and further processed using Autoformer (Wu et al., 2021), a transformer-based architecture adapted for time series forecasting tasks. Autoformer aims to disentangle seasonal trends from local patterns, aligning with our context where global trends and local dialog patterns influence patient outcomes (see Section 3.2). Autoformer also takes the type of patient to simulate as a static categorical variable as input.[2] The produced embedding is processed by linear layers, contained, and classified by two linear layers interposed with a Gelu activation function and dropout layer. We train the model for 150 epochs with a learning rate of 1e-4, utilizing an Adam optimizer and a OneCycleLR scheduler. We use the sum of the cross entropy loss as a loss function for the final classification and reconstruction loss in the Autoformer's output.

## 5 Definition of Evaluation Metrics

Assessing the performance of simulated users presents a challenge, as the objective is to generate behavior that aligns with real patient behavior while also introducing novel interactions. The aim is not to precisely replicate patient behavior but to produce novel data. Consequently, a comprehensive analysis should involve multiple measures to evaluate the effectiveness of simulated users. These problems are similar to those encountered in recommender systems evaluation, where the goal is to recommend diverse, novel, and relevant items to a particular user. Metrics such as diversity, unexpectedness, relevance, or serendipity are commonly used to address these challenges (Kaminskas and Bridge, 2016). Here, we propose translating these measures to the realm of user simulation.

[2] https://github.com/l-Galland/UnexpectedRelevantUserSimulation

### 5.1 Accuracy

Accuracy serves as a conventional metric for appraising simulated users. High accuracy suggests that the generated behaviors closely resemble real users, demonstrating consistency across a substantial portion of the dataset. We use the macro F1 score to account for unbalanced classes in our dataset.

### 5.2 Diversity

In addition to accuracy, the behaviors generated by simulated users must exhibit diversity, ensuring that trained models encounter a broad spectrum of dialog acts. We propose employing the Simpson index (Simpson, 1949) to quantify diversity. This index assesses the likelihood that the model generates the same dialog act given two randomly selected contexts from the dataset, defined as $\lambda = \sum_{i=1}^{N_{da}} p_i^2$.

Here, $N_{DA}$ represents the number of distinct dialog acts, and $p_i$ denotes the proportion of dialog acts $i$. The Simpson index ranges from $\frac{1}{N_{DA}}$ to 1, with lower values indicating greater diversity in generation.

### 5.3 Unexpectedness

The unexpectedness captures how far the generated dialog act is from the target dialog act, hence how expected the generated data is. If the generated selected data is really different from the target, then the unexpectedness is high. Unexpectedness is traditionally gauged by the Cosine Similarity of a recommended item $i$ with historical interactions $H$. Adapting this concept, we compute the Cosine Similarity of the Da2Vec representation (see Section 4.1) of the generated dialog act $DA_g$ and the expected target dialog act $DA_t$: $\text{Unexpectedness}(DA_g|DA_t) = \text{CosineSimilarity}(\text{DA2Vec}(DA_g), \text{DA2Vec}(DA_t))$

### 5.4 Relevance

For recommender systems, the relevance of a proposed item is binary and based on user interactions. The relevance is 1 if the user interacts with the proposed item and 0 otherwise. However, determining the relevance of a dialog act is more nuanced. It isn't easy to assess a dialog act's relevance without the associated utterance. Moreover, the patient simulation output is an utterance generated from the dialog act (see Figure 2). Therefore, the generated utterance should be relevant, i.e., fit well with the context and be syntactically correct. Each utterance is rated in coherence and syntactic correctness with a score between 0 and 1. Automatic measures of syntactical correctness and coherence of utterances have been shown not always to be correlated with subjective measures(van der Lee et al., 2021). Therefore, we present two ways to measure syntactical correctness and coherence: automatically and subjectively.

#### 5.4.1 Automatic measures

We measure the coherence and syntactic correctness of the generated dialog acts automatically using the framework Unieval (Zhong et al., 2022). This framework was developed to uniformize the evaluation of natural language generation. It evaluates generated utterances given a dialog context into five dimensions: naturalness, coherence, engaging, understandability, and groundedness. We measure syntactic correctness through naturalness. Naturalness and coherence are the two most important dimensions for patient-simulated utterances. Indeed, patients ought to be natural and coherent in their discourse. However, they are not necessarily engaging or understandable and do not have any particular information to be grounded. We generate an utterance using the method presented in Section 3.3 for each predicted dialog act in the test set. Using the Unieval framework, we attribute a naturalness and coherence score to each utterance.

#### 5.4.2 Subjective measures

Another way to measure naturalness and coherence is through subjective measures. We select 27 ground truth utterances, and their context (2 preceding turns of speech of the HOPE database), which represent the different possible dialog acts evenly. The utterances are transformed into audio using the Bark TTS (Charles, 2024). We transform the utterances into audio for subjective evaluation as the contexts are transcribed from face-to-face interactions, and such utterances are not consistently evaluated by crowdsourced when prompted as transcripts (Galland et al., 2024b). We recruited 30 participants per condition to evaluate 27 stimuli. 2 attention checks were performed at one-third and two-thirds of the task. The participants with English as a primary language and an approval rate higher than 99% are recruited and rewarded through the Prolific platform (Prolific, 2023). Sample audio is available on OSF[3]. To evaluate naturalness, participants rate their perception of the quality of the synthesized voice and of the wording of the utterance on two items of the Godspeed scale (Bartneck et al., 2009), from 1 (Fake) to 7 (Natural) and from 1 (Machine-like) to 7 (Human-like). Coherence is evaluated by asking participants to rate their agreement on a 7-step Likert scale with the following statements derived from a questionnaire proposed in (Fitrianie et al., 2020) to standardize virtual agents' evaluation: "The sentence fits harmoniously into the surrounding context." and "The sentence does not make sense." The participants answered the questions through a website derived from WebMushra (Schoeffler et al., 2018).

### 5.5 Serendipity

While accuracy and diversity are essential, the ability to generate novel behaviors that are both unexpected and relevant is equally crucial. In recommendation systems, this concept is encapsulated by the serendipity (Ge et al., 2010), defined as discovering unforeseen yet relevant items. In our context, serendipity pertains to generating dialog acts that deviate from the corpus while remaining coherent and natural patient behavior, particularly facilitating novel behavior for dialog model training.

We define the serendipity of a generated dialog act $DA_g$ given context $c$ and the associated target dialog act $DA_t$ from the dataset as:

$$\text{Serendipity}(DA_g|\text{c},DA_t) = \text{Unexpectedness}(DA_g|DA_t) * \frac{\left(\text{Naturalness}(DA_g|\text{c}) + \text{Coherence}(DA_g|\text{c})\right)}{2} \quad (1)$$

Here, unexpectedness quantifies the distance of the generated dialog act $DA_g$ relative to the expected target dialog act $DA_t$, while naturalness and coherence assess the appropriateness of the utterance generated with $DA_g$ given context $c$.

## 6 Evaluation

### 6.1 Baseline

As a baseline for our evaluation, we employ a non-conditioned Large Language Model (LLM) tasked with responding as the patient. The LLM, Mistral

[3]https://osf.io/4mt7s/

7B instruct, is prompted to act as a patient in an MI session and to produce the next utterance given the context. The associated prompt is visible in the appendix and the related code on Github[4]. The resulting utterances are then classified into dialog acts using the classifier presented in (Galland et al., 2024a) and Section 3.1.1.

### 6.2 Results

Measures values on the test set are visible in Table 3. We compute the average value of each metric as well as the 95% confidence interval. The unexpectedness, naturalness, coherence, and serendipity measures are averaged only on the utterances where the predicted dialog act differs from the target dialog act to evaluate how natural, coherent, and unexpected novel data is. We performed an ablation study to study the impact of the dialog act and text inputs with two models: one using only text input and one using only DA and types as inputs.

We found that the baseline model tends to be more accurate than our model and its ablations. However, the accuracy achieved by the Full Model and the ablation OT, taking text as input, is comparable to the baseline's. The ablation ODA, taking only dialog acts as input, is significantly less accurate than the Full Model, the Baseline, and the ablation OT, highlighting the importance of text input for predicting the next dialog act. While adding dialog acts in addition to text (Full model) seems to improve accuracy, the results are not significant. All of our models are significantly more diverse and unexpected than the baseline. However, the automatic measure of naturalness and coherence does not indicate any differences between conditions. The measure of naturalness and coherence performed with Unieval is mainly impacted by the context and not the targeted utterances. Therefore, we compute subjective naturalness and coherence as described in Section 5.4.2. We recompute the measures on a subset of the test set used for subjective measures, composed of 27 utterances (see Table 4). The measures are computed for the LLM Baseline, our model, and the text ablation. For every condition, we have a set of identical utterances, as the predicted dialog acts were the same. The subjective naturalness and coherence ratings are corrected to have the same average on the common utterances to account for differences between groups of participants. The models have no significant differences in naturalness and coherence (Baseline, ablation OT, and Full model).

### 6.3 Discussion

Using our proposed metrics, we were able to highlight differences between models that are not captured by traditional metrics. Indeed, although all text-based models achieve similar accuracy in dialog act prediction, significant differences are observed in other metrics. The baseline, an LLM generating the next utterance based on the context, is significantly less diverse than our proposed model. This highlights that LLMs produce data that, although of high quality (good accuracy), represents an average of the data used to train them. Consequently, they make data similar to what an average user would generate, diminishing the diversity of produced dialog acts. They always tend to answer the same way, whereas our proposed method can generate dialog acts across the entire spectrum of possible dialog acts with more diversity. Similarly, when the baseline differs from the target, it produces dialog acts that are significantly more expected than our proposed method. This underscores the quality of the data generated by LLMs as they remain close to the target dialog act, even if it is not the targeted one. However, unexpectedness can be beneficial if it is also natural and coherent, which is why we compute serendipity. The utterances generated with the dialog acts predicted by our Full Model tend to be subjectively rated on average as less natural and coherent than those from the baseline. The difference in the subjective naturalness and coherence values is not significant, so no conclusion can be drawn. However, the serendipity of our Full Model is significantly higher than the baseline, meaning that when the dialog acts produced by our model are unexpected, they are also natural and coherent. In contrast, unexpected dialog acts produced by the baseline are not as natural and coherent. This underlines our model's ability to create novel data that is also natural and coherent. In contrast, the baseline performs well in replicating data but struggles to generate novel, unexpected, natural, and coherent data. All these results highlight the averaging quality of LLMs, whereas our model, trained on target dialog data, better understands the structure of the dialog and can generalize. Our model allows us to explore user's reactions that are absent from the data but still natural and coherent,. The ablation study high-

[4]https://anonymous.4open.science/r/Patient_simulation-3DE3/README.md

| Model | F1 score | Diversity | Unexpectedness | **Automatic** Coherence | **Automatic** Naturalness | Serendipity |
|---|---|---|---|---|---|---|
| Baseline (LLM) | 0.40[0.35, 0.44] | 0.23[0.21, 0.24] | 0.57[0.54, 0.60] | 0.83[0.82, 0.85] | 0.92[0.91, 0.93] | 0.31[0.29, 0.33] |
| Ablation Only DA (ODA) | 0.20[0.18, 0.22] | **0.18**[0.18, 0.19] | **0.65**[0.62, 0.67] | 0.86[0.85,0.87] | 0.93[0.92,0.93] | **0.58**[0.55,0.60] |
| Ablation Only Text (OT) | 0.35[0.32, 0.37] | **0.16**[0.15, 0.16] | **0.70**[0.67, 0.73] | 0.86[0.84,0.87] | 0.93[0.92,0.93] | **0.62**[0.59,0.65] |
| Full model (input Text + DA + Type) | 0.37[0.34, 0.39] | **0.16**[0.15, 0.16] | **0.66**[0.63, 0.69] | 0.86[0.84, 0.87] | 0.93[0.92, 0.94] | **0.59**[0.56,0.62] |

Table 3: Measures value on the test set of HOPE. The 95% confidence intervals are computed using the bootstrap method and 1000 runs (Efron and Tibshirani, 1994). Results in **bold** are significantly better than the Baseline.

| Model | F1 score | Diversity | Unexpectedness | **Subjective** Coherence | **Subjective** Naturalness | Serendipity |
|---|---|---|---|---|---|---|
| Baseline (LLM) | 0.43[0.27, 0.60] | 0.26[0.20, 0.33] | 0.48[0.29, 0.65] | 0.75[0.70,0.80] | 0.72[0.68, 0.76] | 0.36[0.25, 0.46] |
| Ablation Only Text (OT) | 0.44[0.30, 0.60] | 0.19[0.15, 0.24] | 0.54[0.36, 0.71] | 0.67[0.62, 0.70] | 0.66[0.60,0.71] | **0.55**[0.46, 0.61] |
| Full model (input Text + DA + Type) | 0.44[0.29, 0.59] | 0.19[0.15, 0.25] | 0.54[0.38, 0.73] | 0.69[0.63,0.75] | 0.75[0.63,0.81] | **0.60**[0.51,0.69] |

Table 4: Measures value on 27 utterances of the test set of HOPE. The 95% confidence intervals are computed using the bootstrap method and 1000 runs (Efron and Tibshirani, 1994). Results in **bold** are significantly better than the Baseline.

lights the importance of text inputs for predicting the next dialog act. Indeed, the context captures substantial information relevant to dialog act prediction. Using dialog acts alone (ablation ODA) does not adequately capture the dynamics of the dialog, resulting in less accurate predictions. While including dialog acts and text inputs shows a positive tendency to improve prediction accuracy in the Full Model over the ablation OT, the results are not significant. The serendipity of the Full Model also tends to be better than the serendipity of the ablation OT. In some instances, dialog act information could be beneficial for deciding between multiple possible dialog acts, which explains the observed positive tendency in accuracy. The Full Model might also have learned to reproduce the patterns in the data, which improves the naturalness and coherence, the newly generated data, and the accuracy. This suggests that using dialog acts as input to the Full Model in complement of the text improves comprehension of the structure of the dialog. These results validate our model for patient simulation and highlight the advantages of looking beyond the accuracy metric. Indeed, they show that while our baseline is closer to the original utterances, our proposed model can create novel, syntactically and contextually correct data.

## 7 Conclusion

In this paper, we propose a dialog manager architecture and introduce comprehensive evaluation metrics tailored to open-ended dialog user simulation to address the simulation of MI patients and their evaluation. Our contributions include the development of a dialog manager capable of simulating natural, coherent, and diverse patient behaviors, leveraging a combination of text and dialog act inputs. We have also proposed a set of evaluation metrics—accuracy, diversity, unexpectedness, naturalness and coherence, and serendipity—that provide a more complete assessment of simulated user performance than traditional accuracy measures. These measures have demonstrated the effectiveness of our approach in generating diverse, unexpected, natural, and coherent patient behaviors compared to a baseline LLM model. Our model's ability to capture and generalize from therapy data while generating novel interactions highlights its potential for training dialog models in mental health therapy settings. Our findings underscore the significance of looking beyond conventional metrics and adopting a more comprehensive approach to evaluating simulated users. By focusing on diversity, unexpectedness, naturalness, and coherence, we can ensure that simulated users replicate existing behaviors and generate novel and meaningful interactions, enhancing their effectiveness as tools for supporting mental health therapy. One limitation of our study is the absence of evaluation through interactive sessions, which necessitates the development of a therapist MI dialog model. Additionally, the naturalness and coherence metrics rely on the generated utterances, potentially susceptible to the methodology employed for utterance generation. Nevertheless, the consistent use of the same Large Language Model (LLM) as both the baseline and the generation method mitigates this concern, utilizing highly similar prompts. A stronger baseline, such as GPT-4, would also strengthen these results. Finally, conducting subjective naturalness and coherence measurements on a larger number of utterances and participants would further validate our findings, enabling the detection of significant differences.

## References


Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81.

Timothy W Bickmore, Everlyne Kimani, Ha Trinh, Alexandra Pusateri, Michael K Paasche-Orlow, and Jared W Magnani. 2018. Managing chronic conditions with a smartphone-based conversational virtual agent. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 119–124.

Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1265–1274.

Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a chatbot for digital counselling. In *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017) 31*, pages 1–7.

Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefevre, and Olivier Pietquin. 2011. User simulation in dialogue systems using inverse reinforcement learning. In *Interspeech 2011*, pages 1025–1028.

P.W.D. Charles. 2024. Bark. https://github.com/suno-ai/bark.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.

Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182.

Bradley Efron and Robert J Tibshirani. 1994. An introduction to the bootstrap: Crc press. *Ekman, P., & Friesen, WV (1978). Manual for the facial action coding system*.

Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research*, 21(5):e13216.

Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2022. Adapting conversational strategies in information-giving human-agent interaction. *Frontiers in Artificial Intelligence*, 5:1029340.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024a. Emmi–empathic multimodal motivational interviews dataset: Analyses and annotations. *arXiv preprint arXiv:2406.16478*.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024b. Simulating patient oral dialogues: A study on naturalness and coherence of conditioned large language models. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pages 1–4.

Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260.

Bhautesh Dinesh Jani, David N Blane, and Stewart W Mercer. 2012. The role of empathy in therapy and the physician-patient relationship. *Forschende Komplementärmedizin/Research in Complementary Medicine*, 19(5):252–257.

Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–42.

Hsien-Chin Lin, Christian Geishauser, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gašić. 2022. Gentus: Simulating user behaviour and language in task-oriented dialogues with generative transformers. *arXiv preprint arXiv:2208.10817*.

Hsien-chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, and Milica Gašić. 2021. Domain-independent user simulation with transformers for task-oriented dialogue systems. *arXiv preprint arXiv:2106.08838*.

Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):1–28.

Molly Magill, Timothy R Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca EF Gordon, J Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process:

Technical, relational, and conditional process models of change. *Journal of consulting and clinical psychology*, 86(2):140.

Molly Magill, Jacques Gaume, Timothy R Apodaca, Justin Walthers, Nadine R Mastroleo, Brian Borsari, and Richard Longabaugh. 2014. The technical hypothesis of motivational interviewing: A meta-analysis of mi's key causal model. *Journal of consulting and clinical psychology*, 82(6):973.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.

José Mercado, Ismael Edrein Espinosa-Curiel, and Juan Martínez-Miranda. 2023. Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review. *Journal of Medical Internet Research*, 25:e52097.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Stefan Olafsson, Teresa K O'Leary, and Timothy W Bickmore. 2020a. Motivating health behavior change with humorous virtual agents. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*, pages 1–8.

Stefan Olafsson, Byron C Wallace, and Timothy W Bickmore. 2020b. Towards a computational framework for automating substance use counseling with virtual agents. In *AAMAS*, pages 966–974. Auckland.

Florian Pecune, Stacy Marsella, and Alankar Jain. 2020. A framework to co-optimize task and social dialogue policies using reinforcement learning. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*, pages 1–8.

Prolific. 2023. Prolific.

Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. 2016. An affectively aware virtual therapist for depression counseling. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health*, pages 01951–12.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1):8.

Edward H Simpson. 1949. Measurement of diversity. *nature*, 163(4148):688–688.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

# A Appendix

## A.1 Utterance generation prompt

The condition utterances are generated through the prompt visible in Table 5

## A.2 Utterance generation examples

Examples of generated utterances are in Table 6).

| Condition | Prompt |
|---|---|
| **Base** | You are a patient talking with a therapist. Your task is to generate the patient’s next utterance.<br>The dialog is happening orally, use a oral style language with hesitation, repetition, and deviation.<br>You will only respond with one patient’s utterance. Do not provide explanations or notes.<br>Make only one proposition.<br>The response must be short, no more than 2 or 3 utterances. |
| **Unconditioned** | **Base prompt +**<br>{Utterance examples}<br><<<<br>Context : {context}<br>Generate the patient’s next utterance<br>>>><br>Patient’s utterance: |
| **Dialog act conditioned** | **Base prompt +**<br>{Utterance examples with corresponding dialog act}<br><<<<br>Context: {context}<br>Generate the patient’s next utterance with the intent: {intent}<br>>>><br>Patient’s utterance: |

Table 5: Prompts template for each of the two conditions

| Context | Original utterance | Baseline (LLM) | Text (Ours) | DA + Text + Type (Ours) |
|---|---|---|---|---|
| Turn 71/79<br>Therapist: You're quiet down.<br>Patient: yeah<br>Therapist: Are you keeping interest in anythings,<br>Is there anything that that you find enjoyment in? | [*Sustaining unhealthy behavior*]<br>I am not doing the thing<br>there is nothing for me to do | [*Sharing personal information*]<br>I like to watch movies<br>and play video games. | [*Sharing positive feeling or emotion*]<br>Yeah, I love to play basketball.<br>It really makes me happy when I score a goal. | [*Sharing positive feeling or emotion*]<br>Yeah, I love to play basketball.<br>It really makes me happy when I score a goal. |
| Turn 21/80<br>Therapist: Have you criticized other coworkers?<br>Patient: yeah.Therapist: So do you think that her criticism was<br>within normal limits like what you've seen there regularly. | [*Realization/Understanding*]<br>I mean, now that you pointed out<br>I mean, I guess I can view it in that way<br>that it wasn't. | [*Realization/Understanding*]Well, I guess it depends on<br>who it is and what they did.<br>Sometimes I think people<br>deserve criticism.<br>[laughs] But other times,<br>maybe it's not warranted.<br>It's hard to say for sure. | [*Sharing positive feeling or emotion*]<br>Yeah, I guess so.<br>It feels good to be able to express my thoughts<br>openly and honestly without fear of judgment. | [*Realization/Understanding*]<br>I guess so. It depends on the situation.<br>Maybe if she was being too harsh, but<br>sometimes coworkers need a push. |
| Turn 23/67<br>Therapist: For the position. Are you eligible for unemployment?<br>Patient: I'm not sure I'm still looking into that.<br>Therapist: Because if you, I mean, I'm not a labor law expert.<br>But I think if you're, if you're laid off, there's probably decent probability<br>you could get on employment.<br>How much for educational ? | [*Sharing personal information*]<br>I just have my bachelor's. | [*Sharing personal information*]<br>don't know yet. I haven't checked into that.<br>I was just focusing on getting my unemployment<br>situation figured out first.<br>But I might look into educational assistance as well.<br>It depends on the cost and what's available. | [*Sharing personal information*]<br>don't know yet. I haven't checked into that.<br>I was just focusing on getting my unemployment<br>situation figured out first.<br>But I might look into educational assistance as well.<br>It depends on the cost and what's available. | [*Sustaining unhealthy behavior in the future*]<br>I don't really want to depend on that, though.<br>I prefer to be independent. |

Table 6: Examples of generated utterances in each condition given their context