Next Utterance Prediction for Mental Health Counseling

Sankalp
IIIT Delhi
sankalp22445@iiitd.ac.in

Sarthak
IIIT Delhi
sarthak22453@iiitd.ac.in

Vivan
IIIT Delhi
vivan22581@iiitd.ac.in

Abstract

This project focuses on next utterance prediction in mental health counseling dialogues, where generating context-aware and emotionally aligned responses is critical. Building upon earlier work with T5-small, we extend our methodology by experimenting with multilingual and instruction-tuned transformer models (`google/mt5-small`, `google/flan-t5-base`). Additionally, we introduce fusion architectures that combine transformer encoders with LSTM/GRU layers to capture conversational flow more effectively. A sentiment classifier is integrated to steer generation toward emotionally consistent replies via post-processing. Evaluation using BLEU, ROUGE, BERTScore, and BLEURT shows varied performance across models, with instruction-tuned T5 demonstrating strong lexical overlap and a fusion model achieving the highest semantic similarity via BERTScore, highlighting the value of combining linguistic structure with emotional context in therapeutic settings. We also report specific results for output based on semantic/lexical criteria and emotion consistency.

1 Introduction

Mental health therapy involves interactive conversations where therapists provide guidance and support to clients seeking help. The quality of these interactions heavily relies on the therapist's ability to understand the client's state and respond appropriately, often by reflecting, questioning, or offering insights. In the context of developing AI-driven support systems for mental health, the ability to accurately predict or generate the next utterance in these therapeutic conversations is a crucial capability. Such systems could potentially assist therapists, provide interim support, or serve as training tools.

The core task involves training a deep learning model to generate relevant and coherent responses based on the preceding conversation history. Given an input sequence representing the dialogue turns so far (e.g., "Therapist: How have you been feeling this week? [SEP] Client: A bit overwhelmed, work has been stressful."), the model must generate an appropriate next utterance that aligns with therapeutic goals and conversational context (e.g., "Therapist: Tell me more about what's been overwhelming at work."). This work aims to advance the state-of-the-art in this domain, moving toward more responsible, context-aware, and emotionally intelligent AI agents suitable for therapy-driven settings.

2 Related Work

The task of next utterance prediction (NUP) in mental health counseling lies at the intersection of behavioral modeling, dialogue generation, and emotionally aware AI. Prior research has explored various angles that inform the design of empathetic and contextually grounded dialogue systems.

Cao et al. (2019) tackled behavioral code prediction in Motivational Interviewing (MI), introducing a dual-task framework that both categorizes and forecasts therapist/client behaviors (e.g., reflections, open questions). Their approach highlights the importance of dialogue context in guiding therapeutic strategy. While their model outputs behavioral labels, our work advances this by generating full-text utterances conditioned on similar contextual cues.

Prior work on multimodal dialogue modeling has shown the value of combining lexical and prosodic features using attention-enhanced BiLSTMs. These models effectively differentiate between therapist and client behaviors, especially by leveraging prosodic cues such as pitch and intonation. Although our system is unimodal (text-based), the idea of enhancing representations using complementary signals inspired our integration of recurrent structures (LSTM/GRU) on top of transformer

encoders.

PsyChat (2024) introduced a shift toward client-centric modeling by predicting client behavioral states to dynamically adapt counselor responses. This aligns with our incorporation of a sentiment classifier to ensure emotionally congruent generation. Unlike strategy selection in PsyChat, our system performs direct text generation, but both share the underlying goal of producing tailored therapeutic responses.

Grespan et al. (2023) addressed annotation scarcity by using session-level risk labels to indirectly supervise utterance-level suicide risk prediction. Their approach underscores the importance of indirect signals and label propagation for counseling domains. In our case, we use sentiment as an auxiliary signal, indirectly shaping the emotional tone of generated utterances.

ConSum (2022) proposed PHQ-9-informed dialogue processing to extract clinically relevant content for summarization. While focused on summarization, their utterance selection mechanism emphasizes the importance of identifying meaningful versus filler content — a principle we implicitly apply by training our models to focus on target utterances that drive the therapeutic exchange forward.

These works collectively inform our approach to building empathetic, context-aware next utterance prediction models. Our contribution extends this landscape by integrating generative modeling with recurrent memory, sentiment alignment, and evaluation across multiple architectures, including instruction-tuned and fusion models.

3 Methodology

This research investigates methods to enhance the quality of generated text in therapeutic dialogue contexts, focusing on semantic accuracy, contextual coherence, and emotional consistency. Our methodology integrates advanced transformer architectures, efficient training strategies, and targeted post-processing techniques.

3.1 Dataset and Preprocessing

We utilize structured therapy dialogues containing `input_text` and `target_text` fields. The data is partitioned into standard training, validation, and test sets. Loading and manipulation are handled using Hugging Face's `datasets` library for efficient transformation and batching.

3.2 Data Augmentation via Back-Translation

To increase dataset diversity and mitigate overfitting, we employ back-translation. Each English `input_text` in the training set is translated to French and subsequently translated back to English using the `MarianMTModel`. This process generates paraphrased inputs while preserving the original `target_text`. The augmented examples are then combined with the original training data, creating a richer corpus that exposes the model to greater semantic variance.

3.3 Core Models and Architectural Enhancements

Our approach leverages state-of-the-art transformer models and explores architectural modifications:

- **Baseline Transformer Models:** We employ pre-trained transformer models as foundational components. Specific models include `google/mt5-small` (a multilingual T5 variant) and `google/flan-t5-base` (an instruction-tuned T5 variant), chosen for their text-to-text capabilities. We also evaluate `t5-small` as an initial baseline.

- **LoRA Fine-Tuning:** For efficient adaptation, we utilize Low-Rank Adaptation (LoRA). LoRA injects trainable low-rank matrices into the transformer's attention modules, enabling effective fine-tuning with significantly fewer trainable parameters compared to full fine-tuning. This is particularly beneficial in resource-constrained settings.

- **Fusion Models with RNN Layers:** To enhance sequential modeling capabilities, we introduce a custom `FusionModel` architecture. This model integrates the encoder outputs of a transformer backbone (T5 or mT5) with recurrent layers (LSTM, GRU, or a hybrid LSTM→GRU sequence). Bidirectional RNN layers capture context from both past and future tokens. The output from these fusion layers is compressed via a linear layer and then fed as input embeddings (`inputs_embeds`) to the transformer's decoder, aiming to improve coherence and context tracking in multi-turn dialogues.

3.4 Training Strategies

We implement several strategies to optimize the training process:

- **Tokenization and Prompt Engineering:** Inputs are tokenized to a fixed maximum length. To align with instruction-following models, inputs for models like Flan-T5 are prepended with an instructional prompt, e.g., "Respond appropriately:".

- **Difficulty-Based Curriculum Learning:** We rank training examples by the length of the `target_text` (as a proxy for difficulty) and divide the dataset into three stages: easy, medium, and hard. Models are trained sequentially on these stages. This allows the model to first learn basic conversational patterns before tackling more complex responses, potentially improving stability and convergence.

- **Multi-Stage Curriculum Training:** Training proceeds sequentially through the curriculum stages using Hugging Face's `Seq2SeqTrainer` or a custom PyTorch loop. Checkpoints are saved after each stage to preserve progress.

- **Custom Training Loop:** For models not directly compatible with the `Trainer` (like the custom `FusionModel`), we implement a standard PyTorch training loop using the Adam optimizer, allowing granular control over the training process and validation.

3.5 Generation and Post-processing Refinement

Techniques are applied during and after generation to improve output quality:

- **Beam Search Decoding:** During inference, beam search is used to explore multiple potential output sequences. Repetition penalties and n-gram constraints are applied to discourage repetitive or generic outputs and promote diversity.

- **Sentiment Detection & Emotional Alignment (Rule-Based Rewriting):** To ensure emotional coherence, a critical aspect in therapeutic settings, we implement a post-processing step.

    1. **Sentiment Classification:** A DistilBERT-based sentiment classifier labels the sentiment (e.g., positive, neutral, negative) of both the generated response and the ground truth target.

    2. **Alignment and Rewriting:** If the predicted sentiment mismatches the target sentiment, the generated response is rewritten using a T5-based emotion rewriting model (`mrm8488/t5-base-finetuned-emotion`). The rewriter is prompted with natural language instructions (e.g., "rephrase to be positive: <response>").

    This step aims to align the emotional tone of the output with the expected therapeutic context. Response generation functions are designed to incorporate this step for all model types.

4 Dataset, Experimental Setup, and Results

4.1 Dataset

The dataset consists of structured therapy conversations extracted from publicly available counseling dialogue corpora. Each data point contains:

- **Input Text:** A sequence of dialogue turns between a therapist (T) and a patient (P), separated by a special token (e.g., '[SEP]'). Example: `"T: Hi, how are you today? [SEP] P: Great. How are you? [SEP] T: I'm doing well. Thanks for asking."`

- **Target Text:** The subsequent utterance in the conversation that the model needs to predict. Example: `"So you're doing great."`

The dataset was split into standard training, validation, and test sets.

4.2 Experimental Setup

Our experimental design involves two primary codebases, allowing for a comparative analysis of different architectural and training paradigms.

4.2.1 Dataset

The dataset used for all experiments was provided to us.

4.2.2 Evaluation Metrics

We employ a comprehensive suite of metrics to evaluate model performance from multiple perspectives:

- **BLEU:** Assesses n-gram precision overlap for surface-level fluency.

- **ROUGE (ROUGE-1, ROUGE-2, ROUGE-L):** Measures recall-oriented n-gram and subsequence overlap.
- **BERTScore (F1, Precision, Recall):** Evaluates semantic similarity using contextual embeddings.
- **BLEURT:** A learned metric trained on human ratings to predict generation quality, aiming for better correlation with human judgment. Negative scores, as observed, indicate outputs rated poorly relative to references.

Metrics like BLEU and BERTScore were computed both before and after the sentiment alignment step (which involves sentiment classification via DistilBERT and rule-based rewriting using a T5 emotion model) to quantify its impact.

4.2.3 Experimental Configurations

We evaluated configurations derived from two main experimental setups:

Setup A: Curriculum Learning with LoRA-Tuned Transformer

- **Focus:** Efficient fine-tuning using structured learning progression and post-hoc emotional alignment.
- **Model Backbone:** `google/flan-t5-base`.
- **Fine-Tuning:** Low-Rank Adaptation (LoRA) was used for efficient parameter tuning by injecting trainable low-rank matrices into attention modules.
- **Training:** Employed difficulty-based curriculum learning (ranking examples by target length into easy, medium, hard stages) with back-translation data augmentation (paraphrasing inputs via English-French-English translation).
- **Post-processing:** Applied sentiment alignment and rewriting, where outputs with sentiment mismatching the target were rewritten using an emotion-conditioned T5 model.
- **Evaluation:** Primarily BLEU and BERTScore were computed, assessed both before and after the sentiment correction step.

Setup B: Fusion Models with LSTM/GRU-Based Sequential Memory

- **Focus:** Enhancing transformer encoders with RNN-based sequential memory (using custom `FusionModel` architecture integrating LSTM/GRU layers) for improved temporal modeling.
- **Base Models Tested:**
  - `google/mt5-small` (referred to as *model1* in results)
  - `google/flan-t5-base` (referred to as *model2* in results)
- **Fusion Variants Tested:**
  - *model1* + LSTM→GRU Fusion (*fusion_m3* in results)
  - *model2* + LSTM→GRU Fusion (*fusion_m4* in results)
- **Training:** Utilized a custom PyTorch training loop with the Adam optimizer for granular control.
- **Post-processing:** Applied the same sentiment alignment and rewriting procedure as in Setup A (DistilBERT classification + T5 rewriting).
- **Evaluation:** Employed the comprehensive metrics suite listed above, including BLEU, ROUGE, BERTScore, and BLEURT.

This dual-setup approach allows for comparing the effectiveness of efficient fine-tuning strategies versus architectural enhancements involving recurrent memory.

4.2.4 Implementation Details

For the LoRA configuration used in Setup A (`google/flan-t5-base`), we applied the following settings via the PEFT library: The rank (`r`) was set to 8, and `lora_alpha` was 32. LoRA was applied to the query (`"q"`) and value (`"v"`) matrices within the attention modules. The `lora_dropout` rate was 0.1, and bias terms were set to `"none"`. The task type was specified as sequence-to-sequence language modeling (`TaskType.SEQ_2_SEQ_LM`). Other hyperparameters included [Specify training epochs, batch size, learning rate, optimizer if available, e.g., "training for 3 epochs with a batch size of 16 and a learning rate of 2e-5 using the AdamW optimizer"]. Experiments were conducted on [Specify hardware, e.g., "NVIDIA A100 GPUs"]. For Setup B, standard PyTorch training procedures were followed using

the Adam optimizer [Specify learning rate, batch size etc. if different from Setup A].

4.3 Results: Baseline Model Performance

We first evaluated two standard baseline models, T5-small and Flan-T5-base, to establish initial performance benchmarks. The results are shown in Table 1.

Model	BLEU	BERT F1	BERT P	BERT R
T5-small	0.0020	0.8521	0.8659	0.8393
Flan/T5-basic	0.0033	0.8580	0.8706	0.8463

Table 1: Evaluation results for baseline models. BERT F1/P/R are BERTScore F1/Precision/Recall means.

Both models achieve high BERTScore F1 values, indicating strong semantic understanding relative to the target text, despite low BLEU scores which suggest lexical divergence. Flan-T5-base shows slightly higher scores on both metrics compared to T5-small.

4.4 Results: Model Comparison

The evaluation results for the different models compared in Setup B on the test set are presented in Table 2.

Key findings from this comparison include:

- **Lexical Overlap (BLEU/ROUGE):** The scores are generally very low across all models, suggesting generated responses differ significantly from the ground truth in terms of exact word sequences. 'Model2' (Flan-T5 from Setup B) achieved the highest scores on these metrics, indicating better surface-level similarity compared to MT5 and the fusion models. Its performance here differs notably from the baseline Flan-T5-base in Table 1, likely due to differences in training (e.g., Setup B vs Setup A, or other variations).
- **Semantic Similarity (BERTScore):** BERTScore F1 scores for these models ($\sim$0.81-0.825) are slightly lower than the baseline results in Table 1. 'Model1 + LSTM+GRU' achieved the highest BERTScore in this comparison, slightly surpassing 'Model2'. This suggests the fusion approach might enhance semantic capture for the MT5 base.
- **Learned Quality (BLEURT):** All models received negative BLEURT scores, with 'Model2' (Flan-T5 from Setup B) performing the least poorly. Negative scores typically indicate the generated text is considered dissimilar or lower quality compared to reference texts according to the BLEURT model.
- **Fusion Model Performance:** The fusion architecture yielded mixed results. Adding LSTM+GRU layers improved the BERTScore for 'Model1' (MT5) but decreased performance across almost all metrics for 'Model2' (T5).

The sentiment alignment step was applied before these final evaluations, aiming to ensure emotional consistency.

4.5 Results: Strategy Evaluation

- **Semantic and Lexical :**
  - BERTScore F1: 0.8482
  - BERTScore Precision: 0.8499
  - BERTScore Recall: 0.8469
  - BLEU Score: 0.0080
- **Semantic Quality and Emotion Consistency :**
  - BERTScore F1: 0.8332
  - BERTScore Precision: 0.8409
  - BERTScore Recall: 0.8261
  - BLEU Score: 0.0087

5 Discussion/Analysis/Observations

Our findings highlight several key aspects and challenges in next utterance prediction for mental health counseling:

- **Baseline Performance:** The initial baseline results (Table 1) show that standard T5 models (T5-small, Flan-T5-base) can achieve high semantic similarity (BERTScore > 0.85) but struggle with lexical match (BLEU < 0.004). Flan-T5-base slightly outperforms T5-small as a baseline.
- **Instruction Tuning vs. Base Models (Setup B):** In the Setup B comparison (Table 2), the instruction-tuned 'Model2' (Flan-T5) generally outperformed the base 'Model1' (MT5) on metrics sensitive to lexical overlap (BLEU, ROUGE) and the learned quality metric (BLEURT), consistent with baseline trends.

Model	BLEU	R-1	R-2	R-L	BERT F1	BLEURT
Model1 (MT5)	0.0012	0.0257	0.0035	0.0237	0.8193	−1.3568
Model2 (T5)	0.0054	0.0723	0.0143	0.0642	0.8229	−1.3229
Model1 + LSTM+GRU	0.0028	0.0538	0.0092	0.0475	0.8250	−1.3762
Model2 + LSTM+GRU	0.0021	0.0503	0.0055	0.0453	0.8146	−1.3860

Table 2: Evaluation results comparing base models and fusion models (Setup B) on the test set. Model1=mt5-small, Model2=flan-t5-base. Fusion models use LSTM→GRU layers. R-1/R-2/R-L are ROUGE scores. BERT F1 is BERTScore F1.

However, the absolute scores, particularly BERTScore, were lower than the initial baselines, suggesting potential variations due to the specific training setup (Setup B vs Setup A or baseline setup).

- **Impact of Fusion Architecture:** The addition of LSTM/GRU layers did not yield consistent improvements. While it slightly boosted the semantic similarity (BERTScore) for the MT5 base model (‘Model1‘), it negatively impacted the stronger Flan-T5 base model (‘Model2‘), particularly reducing its lexical overlap scores. This suggests that for a well-tuned transformer like Flan-T5, the additional sequential modeling from the RNN layers might not add significant value or could even interfere with the transformer's learned representations for this specific task and architecture.

- **Metric Discrepancies:** The stark contrast between very low BLEU/ROUGE scores and high BERTScore values persists across all experiments. It indicates that the models generate responses that are semantically similar to the target but use different wording (paraphrases). This is common in dialogue generation but poses a challenge for evaluation. The negative BLEURT scores across all models in Table 2 are concerning, suggesting potential issues with fluency, coherence, or relevance that are not captured by simple overlap metrics.

- **Importance of Emotional Alignment:** While its quantitative impact wasn't isolated in the main model comparison table, the inclusion of the sentiment detection and rewriting step addresses a critical requirement for therapeutic dialogue systems. Ensuring emotional congruence, even as a post-processing step, is vital for responsible deployment in sensitive applications. Observations during development indicated this step corrected potentially jarring emotional mismatches in generated text. The results presented in Section 4.5 also touch upon the balance between semantic quality and emotional consistency.

- **Curriculum Learning:** The training logs for ‘Model2‘ (Setup A) showed decreasing validation loss across curriculum stages, supporting the hypothesis that starting with simpler examples aids model training and stability for this complex generation task.

Overall, the Flan-T5 model (‘Model2‘ in Table 2) fine-tuned with LoRA and curriculum learning appears to offer a good balance based on the provided metrics, particularly ROUGE and BLEURT within that comparison set. However, the high BERTScore of the MT5-Fusion model (‘Model1 + LSTM+GRU‘) and the baseline results (Table 1) warrant attention regarding semantic capture. The generally low n-gram overlap scores and negative BLEURT scores highlight the ongoing difficulty of generating human-like, contextually perfect therapeutic responses.

6 Conclusion and Future Work

This project successfully explored advanced techniques for next utterance prediction in mental health counseling dialogues. We established baseline performance using T5-small and Flan-T5-base, demonstrating high initial semantic similarity but low lexical overlap. We further investigated the effectiveness of combining instruction-tuned transformers (Flan-T5) with efficient fine-tuning (LoRA), curriculum learning, and data augmentation. Additionally, we introduced novel fusion architectures incorporating RNN layers and implemented a crucial sentiment alignment mechanism via post-processing. Our evaluations revealed that the instruction-tuned Flan-T5 model provided

strong performance relative to MT5 within specific experimental setups, particularly on lexical overlap metrics, while a fusion model based on MT5 achieved competitive semantic similarity (BERTScore). Dedicated strategy experiments showed potential for achieving high semantic quality. The results underscore the complexity of the task, highlighted by the discrepancy between different automatic metrics and variations across experimental setups.

Future work could explore several directions:

- **Incorporating Behavioral Codes:** Explicitly conditioning generation on predicted therapeutic behavioral codes (e.g., reflection, open question) could lead to more strategically aligned responses.
- **Advanced Fusion Models:** Experimenting with different fusion mechanisms, such as attention between transformer and RNN layers.
- **Personalization:** Adapting models to individual client histories or therapist styles.
- **Multi-Modal Inputs:** Integrating non-verbal cues (e.g., prosody from audio, facial expressions from video) if available.
- **Human Evaluation:** Conducting thorough human evaluations with domain experts to assess therapeutic appropriateness, empathy, fluency, and overall quality beyond automatic metrics, especially given the low BLEU/ROUGE and negative BLEURT scores.
- **Refining Emotion Control:** Developing more sophisticated methods for fine-grained emotion control integrated directly into the generation process, rather than post-processing.
- **Analyzing Training Variations:** Investigating why BERTScore results differed between the baseline evaluation and the Setup B comparison for Flan-T5-base.

Continued research in this area holds promise for developing AI tools that can responsibly and effectively support mental health professionals and clients.

Limitations

The current work relies solely on text-based dialogues and does not incorporate multi-modal signals (audio, video) which are important in real therapy. The evaluation primarily uses automatic metrics, which may not fully capture the nuances of therapeutic quality like empathy or clinical appropriateness; extensive human evaluation by experts is needed for validation, especially given the metric scores obtained. The sentiment alignment is rule-based and relies on separate classifiers and rewriters, which might introduce errors or unnatural phrasing. The datasets used, while structured, may not fully represent the diversity and complexity of all real-world counseling scenarios. Generalization to unseen therapeutic styles or client issues needs further investigation. The models generate responses based on patterns learned from data and do not possess true understanding or clinical judgment. The fusion architecture's effectiveness seemed dependent on the base model and did not provide universal gains. Performance variations observed for the same base model (Flan-T5-base) across different result tables suggest sensitivity to the specific training setup or implementation details not fully captured.