Modeling Low-Resource Health Coaching Dialogues via Neuro-Symbolic Goal Summarization and Text-Units-Text Generation

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu and Nikolaos Agadakos
University of Illinois Chicago, Illinois, United States
{yzhou232, bdieugen, bziebart, sharpl, liub, nagada2}@uic.edu

Abstract

Health coaching helps patients achieve personalized and lifestyle-related goals, effectively managing chronic conditions and alleviating mental health issues. It is particularly beneficial, however cost-prohibitive, for low-socioeconomic status populations due to its highly personalized and labor-intensive nature. In this paper, we propose a neuro-symbolic goal summarizer to support health coaches in keeping track of the goals and a text-units-text dialogue generation model that converses with patients and helps them create and accomplish specific goals for physical activities. Our models outperform previous state-of-the-art while eliminating the need for predefined schema and corresponding annotation. We also propose a new health coaching dataset extending previous work and a metric to measure the unconventionality of the patient's response based on data difficulty, facilitating potential coach alerts during deployment.

Keywords: Dialogue Systems, Neuro-Symbolic AI, NLP in Healthcare

1. Introduction

Health coaching is a patient-centered clinical practice that aims to help patients achieve personalized and lifestyle-related goals to enhance their health behaviors. It has demonstrated efficacy in managing chronic conditions like diabetes and cardiovascular disease, as well as alleviating mental health issues such as anxiety and depression (Butterworth et al., 2006; Ghorob, 2013; Kivelä et al., 2014; Thom et al., 2016). Health coaching is particularly advantageous for low-socioeconomic status (SES) populations, who endure a disproportionate burden of physical and mental health issues (Thackeray et al., 2004; Kangovi et al., 2014). Nonetheless, it is invariably cost-prohibitive for these populations due to its highly personalized and labor-intensive nature; additionally, it also requires considerable time commitment since it normally spans several sessions across weeks or even months.

Recently, efforts have been undertaken to enhance the effectiveness of health coaching through natural language processing techniques. Our group (Gupta et al., 2020a,b, 2021) collected real-world datasets of health coaching conversations and proposed a goal summarization approach based on slot-filling and dialogue state tracking. We further improved goal summarization and introduced an efficient health coach dialogue system that supports empathetic response generation (Zhou et al., 2022). An excerpt of a health coaching conversation is illustrated in Table 1.

However, these works are based on simplified dialogue states (a set of goal attribute-value pairs) and require a designed schema with corresponding human annotations, which are not only labor-intensive but also lose the global contextual information that is established over several sessions (in our case, from 4 to 8 weeks). In addition, the considerable length of each sample, together with the limited data size, poses significant challenges in healthcare dialogue modeling. To address these issues, we propose a neuro-symbolic goal summarization approach that eliminates the need for a pre-designed schema or corresponding annotations yet preserves interpretability. The model is optimized to (i) summarize the goal as effectively as possible, given the current week's dialogue, and (ii) generate an executable instruction on how to modify the summarized goal by referencing the goal from the previous week(s). We also introduce a text-units-text dialogue generation approach that considers as input a sequence of discrete units symbolizing the dialogue history. Moreover, we propose a more generalized approach to detect patient's unconventional responses by extending Point-wise $\mathcal{V}$-usable Information (PVI) (Ethayarajh et al., 2022) for dialogue generation, without using any external datasets, contrasting with our previous work (Zhou et al., 2022). Finally, we introduce a novel health coaching dataset consisting of 1880 dialogue turns from 22 patient-coach conversations, each spanning up to 8 weeks. The dataset is enriched with Fitbit data, which tracks the progress of patients' goals. We aim to augment

LREC-COLING 2024, pages 11498–11509
20-25 May, 2024.

Excerpt of Patient-Coach Conversation # 17 Week # 7
Coach: Hi [Name]! I hope you had a great weekend :) You are now in the second to last week of the study!
Coach: Your goal last week was to reach 7000 steps everyday - you got close!
Coach: Did you want to keep your walking goal the **same this week or lower it a bit?**
Patient: You can lower it to 5000
Coach: Okay, sounds good - and what days are you thinking to do 5,000 steps?
Patient: 7 days
Coach: awesome - and what's your confidence on a scale from 1-10 that you can reach this goal?
... ...
Coach: Hi [Name] - you're doing a really great job with your goal this week.
Coach: You got over 11,000 steps on Monday and yesterday - that's over double your goal! Amazing job :)
Patient: Thanks and you too
Coach: Hi [Name] - Great job reaching your goal last week. You got over 5,000 steps all 7 days! You should be really proud of yourself :)
Excerpt of Patient-Coach Conversation # 17 Week # 8
Coach: This is your last week in the study - what would you like your final walking goal to be?
Patient: Stay the **same as last week**.
... ...

Table 1: An excerpt from a health coaching conversation from our dataset. Each week the coach and patient collaboratively establish a feasible goal. Then, the coach monitors the patient's progress, maintains the patient's engagement, and addresses the patient's concerns. The coach may discuss the goal by referring to the goal settled before.

the existing health coaching datasets and provide a more robust testing benchmark for health coaching modeling, particularly in potential domain-shift scenarios.

We evaluate our model by both automatic metrics and expert-based human evaluation. Experimental results show that our neuro-symbolic goal summarizer outperforms the current state-of-the-art by up to $\sim$30% in semantic frame accuracy. In addition, our text-unit-text dialogue generation achieved the best performance compared to previous work in all metrics for all datasets. Health coaches prefer our generated responses over the previous state-of-the-art 33.9% of the time, while the vice versa happens 19.6% (the rest are ties).

Our contributions are (1) a data-efficient neuro-symbolic goal summarization model and text-units-text dialogue generation model for health coaching, which outperform previous state-of-the-art while eliminating the need for predefined schema and corresponding annotation; (2) a metric measuring the unconventionality of the patient's response in terms of data difficulty, facilitating potential coach alert during deployment and data characterization during training; and (3) a novel health coaching dataset.

2. Related Work

Conversational Agents in Healthcare. Conversational agents have been employed in healthcare to enhance the efficiency and scalability of interactions between healthcare professionals and patients. For instance, chatbots have been utilized in various healthcare contexts, such as chronic disease monitoring (Chaix et al., 2020), cognitive behavior therapy (Fitzpatrick et al., 2017), and physical activity promotion (Mohan et al., 2020; Kocielnik et al., 2018). Nevertheless, these systems often exhibit limitations in their natural language understanding and generation capabilities. More advanced approaches have been proposed for mental health counseling (Althoff et al., 2016; Shen et al., 2020) and health coaching. Our group (Gupta et al., 2020a,b, 2021) collected two real-world health coaching conversation datasets, focusing on the NLU components that summarize weekly goals to support health coaches. Building upon these datasets, we developed a data-efficient health coaching dialogue system with a simplified NLU and NLG framework and mechanism-conditioned empathetic response generation (Zhou et al., 2022). In recent years, there has been a shift in domain-specific dialogue systems from modularized and NLU-NLG component-based designs (Jokinen and McTear, 2009; Williams et al., 2016; Budzianowski et al., 2018; Mrkšić et al., 2017; Wen et al., 2015) towards end-to-end architectures to reduce human effort and error propagation between modules (Hosseini-Asl et al., 2020; Peng et al., 2021).

Neuro-symbolic Approaches have recently gained significant attention due to their ability to facilitate end-to-end while leveraging symbols for interpretability and data-efficient training. Mao et al. (2019) proposed a neuro-symbolic concept learner that combines the strengths of both neural networks and symbolic logic, demonstrating improved performance in visual question-answering tasks. Lamb et al. (2019) introduced the Neuro-Symbolic Transformer, which leverages symbolic reasoning within a transformer-based architecture in text classification tasks. De Raedt et al. (2019) proposed a neuro-symbolic approach that demonstrates the potential of neuro-symbolic systems in handling complex reasoning tasks. Another work in this area is the exploration of end-to-end differentiable natural logic modeling (Feng et al., 2020). Dong and Lapata (2018) proposed a coarse-to-fine decoding method for neural semantic parsing, utilizing a structure-aware neural architecture. Arabshahi et al. (2021) explores a neuro-symbolic approach for enhancing conversational dialogue systems with commonsense reasoning capabilities. Garcez et al. (2018) provides a comprehensive survey and taxonomy of approaches that combine neural and symbolic methods for learning and reasoning in AI systems.

3. Health Coaching Datasets

Background A health coaching process often starts with a goal-setting stage where the coach discusses creating a S.M.A.R.T. goal with the patient, namely a goal that is specific, measurable, achievable, relevant, and time-bound (Doran, 1981). Once the goal is settled, the coach will monitor the patient's progress and maintain patient engagement. Our research group (Gupta et al., 2020a) collected two datasets (dataset 1 and 2) of health coaching dialogues between patients and coaches via text messages. Previously, we defined ten slots for the goal's attributes (types of activity, amount, time, days, location, duration, and the confidence score for the activity). We also used a stage-phase schema for additional turn-level annotation and added dialogue act annotations in our later work (Gupta et al., 2021).

New Human Data Collection Emulating their settings, we recruited four health coaches trained in SMART goal setting and a cohort of 22 patients. The study was conducted over eight weeks. The communication between the coach and the patient was facilitated through text messaging applications, and Fitbits were provided to monitor patients' activity progress. However, six patients withdrew in the early weeks of the study. Consequently, the dataset comprised 1880 dialogue turns over 102 weeks. The datetime of messages, interlocutors, dialogue content, Fitbit action data, and patient and health coach IDs are available for each dialogue. The data were thoroughly anonymized. In contrast to the first two rounds of data, we retained the emoji tokens in the dataset for future analysis. Data statistics are shown in 2. A sample of the health coaching conversation between the patient and health coach is presented in Table 1. The full dataset is available at https://github.com/uic-nlp-lab/virtualcoachdata/.

	#P/C	#W	#T	#T/(W,P)
Dataset 1	27/1	4	2,853	26.6
Dataset 2	28/3	5-8	4,134	18.8
Ours	22/4	2-8	1,880	18.4

Table 2: Dataset Statistics. #P/C, #W, #T, #T/(W,P) refer to the number of patients/health coaches involved, study weeks, total dialogue turns, average turns per patient per week.

4. Methods

This section presents the two integral components of our health coaching dialogue system: the Neuro-Symbolic Goal Summarizer and the Neuro-Symbolic Dialogue Generator.

4.1. Neuro-Symbolic Goal Summarizer

The task of goal summarization involves processing the dialogue history between the health coach and the patient to generate the negotiated goal in natural language. The dialogues in health coaching often span multiple weeks. When the coach discusses the current week's goal, they may refer back to the attributes of the goal implemented in the previous week. However, incorporating all text from previous weeks can result in excessively long inputs, which, given the limited dialogue samples, can be challenging to train on. Our previous research (Zhou et al., 2022; Gupta et al., 2021) approached this task by formulating it as dialogue state tracking, where the state is defined as a set of goal attribute-value pairs. However, these methods require a pre-defined schema with corresponding annotations and a separate slot-filling module, which do not optimally utilize the entire context, resulting in compromised information.

To address these challenges, we aim to provide a data-efficient and end-to-end yet interpretable solution and propose a neuro-symbolic approach for goal summarization. The approach considers the dialogue of the current week while referencing the previous week's goal. In addition, it does not require a schema or corresponding annotations. We aim to optimize the model to summarize the goal, focusing on two key aspects: (i) summarizing the goal as effectively and comprehensively as possible, given the *current* dialogue, and (ii) generating a feasible instruction on how to modify the summarized goal from (i) based on the previous goal. As an example depicted in Figure 1, the coach elaborates on the goal for week w_t by referencing the goal established in the preceding week w_{t-1} (*"same days?"*). Our model learns to generate an executable instruction (`Copy {Days}`) and extract the partial goal (*"Walk 2,500 steps"*) from the current week's dialogue, and execute the instruction on the referenced previous week's goal to generate the complete goal summarization: *"Walk 2,500 steps from Monday to Friday."*

To achieve that, we formulate the problem within the context of Reinforcement Learning (RL), specifically utilizing the Proximal Policy Optimization (PPO) (Schulman et al., 2017) framework. We maximize the following training objective:

$$\text{objective}(\phi) = \mathbb{E}(x, y) \sim D_{\pi_\phi^{RL}}[r(y, y^*|x) - \lambda \, \text{KL}(\pi_\phi(y|x)||\pi_{base}(y|x))] \quad (1)$$

where π_ϕ is the learned RL policy (e.g., the language model to be fine-tuned) and π_{base} is the untouched pre-trained language model. r is the reward function (we use ROUGE score (Lin, 2004)), and the KL term is the Kullback–Leibler Divergence as regularization. We aim to optimize the likelihood of the generated sequences, comprising partial goals and instructions, that `Instruction{Partial Goal, Reference}` subsequently results in ground truth summarization. The full set of instructions (e.g., `Add {Num}`, `Copy {Times}`) will be available in the Appendix.

4.2. Text-Units-Text Dialogue Generator

4.2.1. Generation with Unit Symbols

A prevalent issue in healthcare dialogue datasets is the extensive length of each sample, associated with the limited number of samples. Training a sequence-to-sequence generation model using a long context window presents a significant challenge. However, an interesting observation is that healthcare dialogues typically follow a similar pattern. For instance, health coaching invariably begins with the coach discussing a realistic goal with the patient, confirming each aspect of the goal, such as activity, amount, day, and times. Once the goal is established, the coach monitors the patient's progress and sustains patient engagement.

Inspired by this observation and the recent success of speech-to-unit modules that predict the discrete representations of the speech (Lee et al., 2022; Lakhotia et al., 2021), we explore the possibility of utilizing a short context window while preserving the dialogue history information through symbolic abstraction in text-to-text dialogue generation. Specifically, our response generation incorporates the following two models:

Dialogue-History-to-Unit Encoder encodes the long dialogue history into discrete unit symbols using out-of-box pre-trained language models and unsupervised approaches. We first encode each turn of the conversation in the training dialogues with SBERT, then run k-means clustering to obtain K clusters; we subsequently designate the cluster indices as units. Despite the unique details of each dialogue, we anticipate a similar sequence of units to emerge as each dialogue progresses. Furthermore, the total number of units in a dialogue can indicate the dialogue's progression length.

Units-to-Text Generation We propose a sequence-to-sequence model that generates responses for a virtual coach. The model takes as inputs a sequence of discrete units $N_1, ..., N_{t-1}$ symbolizing the dialogue history, the most recent dialogue turns from the coach C_{t-1} and the patient U_{t-1}, and a partial goal G_{t-1} summarized as discussed in Section (4.1) based on the current dialogue. All inputs are concatenated as a single sequence. Then the generated virtual coach response R_t is defined as:

$$R_t = \text{Seq2Seq}([N_1, ..., N_{t-1}, G_{t-1}, R_{t-1}, U_{t-1}]) \quad (2)$$

4.2.2. Measuring Data Difficulty of User Responses

Identifying unconventional user responses is vital in domain-specific dialogue systems to circumvent system failures due to unsupported user inputs. It can be particularly critical in healthcare settings, where patients often express concerns and emotions that exceed the system's capabilities or do not respond directly to the system's prompts. The unconventional inputs from patients, coupled with the limited availability of healthcare training data, can pose significant challenges to the model's training process. In this work, we show the potential of leveraging data difficulty to identify unusual patient inputs.

Data difficulty metrics categorize data into easy-to-learn and difficult-to-learn examples based on training dynamics (Pleiss et al., 2020; Swayamdipta et al., 2020) or mutual information (Xu et al., 2020; Ethayarajh et al., 2022). Previous research primarily focuses on classification tasks, where a difficult-to-learn example typically signifies that the example is mislabeled. Xu et al. (2020) propose a framework called **predictive $\mathcal{V}$-information** to measure how much information can be extracted or learned from input X about its label Y when constrained to functions or a model family $\mathcal{V}$. The predictive $\mathcal{V}$-information, $I_\mathcal{V}(X \to Y)$ is defined as the difference between *predictive $\mathcal{V}$-entropy* $H_\mathcal{V}(Y)$ and *conditional $\mathcal{V}$-entropy* $H_\mathcal{V}(Y|X)$:

$$I_\mathcal{V}(X \to Y) = H_\mathcal{V}(Y) - H_\mathcal{V}(Y|X) \quad (3)$$

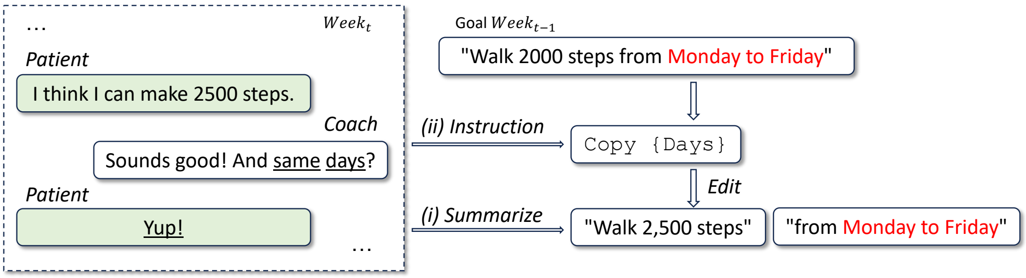

Figure 1: A simplified demonstration of Neuro-Symbolic Goal Summarization. The health coach discusses the goal for week w_t by referring to the goal set in the previous week w_{t-1} (*"same days?"*). Our model is trained to generate an executable instruction (`Copy {Days}`) and to extract the partial goal (*"Walk 2,500 steps"*) from the dialogue of the *current* week. The model then edits the partial goal by applying the instruction to the reference previous goal, resulting in the comprehensive goal summary: *"Walk 2,500 steps from Monday to Friday."*

The greater the $I_{\mathcal{V}}(X \to Y)$, the easier the dataset is for $\mathcal{V}$. Extending from Xu et al. (2020), Ethayarajh et al. (2022) propose **Point-wise $\mathcal{V}$-usable Information (PVI)** to measure the information in the *individual* instance usable by a model family $\mathcal{V}$ with respect to the data distribution. The PVI of an instance (x, y) is defined as:

$$\mathrm{PVI}(x \to y) = -\log_2 g'_y(\varnothing) + \log_2 g_y(x) \quad (4)$$

where $g' \in \mathcal{V}$ s.t. $\mathbb{E}[-\log g'_Y(\varnothing)] = H_{\mathcal{V}}(Y)$ and $g \in \mathcal{V}$ s.t. $\mathbb{E}[-\log g_Y(X)] = H_{\mathcal{V}}(Y|X)$. If $\mathcal{V}$ were the BERT function family, g and g' would be the models after finetuning BERT with and without the input, respectively. An instance with a large negative PVI value is considered as "difficult," showing that the model can better predict the majority class when ignoring X, which often indicates the instance is mislabeled.

Inspired by their work, we aim to explore the difficulty in predicting user responses given the most recent dialogue context, which will provide insights into the degree of unconventionality or surprise in the user's response and its divergence from the dataset distribution. This approach could additionally serve as an easy-to-implement and general out-of-domain detection, eliminating the need to construct a separate domain-specific detection classifier (e.g., empathy detector (Zhou et al., 2022)). To achieve this, we propose an *extension to the PVI* in sequence-to-sequence *generation* as:

$$\mathrm{PVI}(x \to y) = -\underbrace{\sum_{t=1}^{n} log(p(y_t|y_{<t}, \varnothing))}_{\text{by } g'} + \underbrace{\sum_{t=1}^{n} log(p(y_t|y_{<t}, x))}_{\text{by } g} \quad (5)$$

where x and y are the sequence of the dialogue context and the user response, and g and g' are the generative models after fine-tuning with and without the dialogue context, respectively. An instance (user response) with a large negative PVI value is considered "difficult," showing that the model can better predict the user response when ignoring the context, indicating the unconventionally in the user response.

The difficulty metric we propose serves two functions: Firstly, it assists in identifying unconventional patient responses during deployment. If such a response were detected, the system could alert the human coach, indicating potential patient concerns that require attention or suggesting that the patient's utterance or question may not be within the system's capabilities. Secondly, the difficulty metric acts as a data filter during training. It categorizes our limited data into subsets of easy-to-learn and hard-to-learn instances, facilitating possible curriculum learning or denoising approaches for our system.

5. Experiments

5.1. Experimental Settings

To ensure comparability, we use dataset 1 and dataset 2 from Gupta et al. (2021) for training/development and testing, following our previous work (Zhou et al., 2022). We also employ identical model architecture backbones as before, specifically utilizing T5 (Raffel et al., 2020) for the goal summarization task and GPT-2 (Radford et al., 2019) for the response generation task. We additionally evaluate model response generation performance on our newly collected dataset 3 for generalizability in possible domain-shifting. To mitigate inefficient sampling, we manually annotate 40 positive examples. We also use two T5-base models for our PVI-generation metrics. We choose k = 15

Model	D1 (F)	D1 (B)	D2
Gupta et al. (2021)	15	13.1	-
Zhou et al. (2022)	21.7	31.6	13.6
Ours	**52.3**	**44.8**	**46.7**

Table 3: Goal summarization performance by semantic frame correctness. D1 (F), D1 (B), and D2 refer to dataset 1 (Forward), dataset 1 (backward), and dataset 2. Our model outperforms previous work by a significant margin in all three datasets.

for deriving the discrete units. Further details and model parameters will be available in the Appendix.

Evaluation Metrics We use semantic frame correctness for evaluating the goal summarization, which is identical to goal correctness@k with k = 10, used by our previous work (Gupta et al., 2021; Zhou et al., 2022). For dialogue response generation, we use BLEU (Papineni et al., 2002), BertScore (Zhang et al., 2020), and Perplexity (PPL). We measure fluency as perplexity (PPL) of the generated response using a pre-trained GPT2 model that has not been fine-tuned for this task, following previous work (Ma et al., 2020; Sharma et al., 2021).

.

5.2. Results

We compare our Neuro-symbolic Summarization model with the best performance reported by Gupta et al. (2021) and Zhou et al. (2022) in Table 3. D1 (F), D1 (B), and D2 refer to dataset 1 (Forward), dataset 1 (backward), and dataset 2. Forward and Backward refer to the two points where they labeled the goals for each week: one at the end of the goal-setting stage (forward) and the other at the end of the goal-implementation stage (backward). Our approach improves semantic frame correctness by a significant margin ($\sim$ 20$\%$ - $\sim$30$\%$), compared with previous work that utilizes the traditional slot-filling and state-tracking framework.

Table 4 shows the model performance on dialogue response generation. We employ two settings to train our response generation model: (1) Original. Using the original data, and (2) Low-PVI-Replace. substituting the patient's response, which has been assigned a large negative PVI value, with an alternative response from a similar context with a positive PVI in the training set. This context is identified by locating the semantically closest previous coach utterance using sentence-BERT embeddings. We substitute 5% of the patient's utterances, prioritizing those with the most negative PVI values in the training set. We acknowledge that patient utterances with low negative PVI values are not necessarily "noise," and such substitutions could potentially disrupt the coherence of subsequent dialogues. However, we have found that this replacement strategy generally aids our model's convergence, particularly with limited training data. Our model (both with and without PVI fix) outperforms previous work on both dataset 2 and our newly collected dataset in all metrics. We have identified a domain shift in our collected dataset, characterized by patients exhibiting less verbal communication and increased usage of emojis. Coaches have adapted to maintain the continuity of the dialogue by becoming more talkative. The domain shift negatively impacted our model's performance.

Human Evaluation We conducted an expert-led assessment of the generated coach responses via A/B testing. Four health coaches were consulted to evaluate our model's outputs compared to those from previous state-of-the-art (Zhou et al., 2022), given the same dialogue histories. The coaches were asked to select one of the following options: (a) Response A is most suitable in the given context; (b) Response B is most suitable in the given context; (c) Tie; both responses are equally good, or (d) Neither is appropriate. Among 56 collected samples, our model's outputs have a 14.29$\%$ preference over previous state-of-the-art (Zhou et al., 2022) (as the difference between the preferences for one or the other). However, approximately 30$\%$ of the generated responses were not deemed appropriate by our experts. This indicates the ongoing challenge of generating high-quality health coaching responses, particularly given the constraints of limited data and the sensitive nature of health-related interactions. See Figure 2.

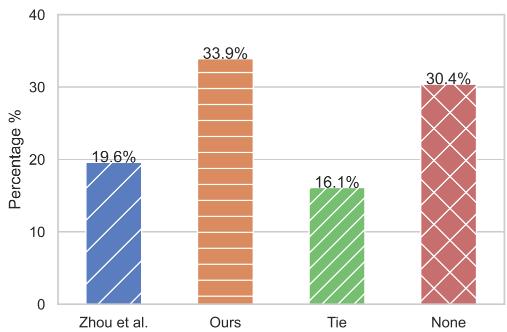

Figure 2: Human evaluation of generated response by health coaches. Our model's outputs have a 14.29% preference over previous state-of-the-art (Zhou et al., 2022).

	Data 2			Data 3		
Model	BLEU (↑)	PPL (↓)	BertS F1 (↑)	BLEU (↑)	PPL (↓)	BertS F1 (↑)
Zhou et al. (2022)	25.1	15.6	87.2	20.6	19.7	83.5
OURS+Original	26.7	14.7	88	22.5	16.7	**84.4**
OURS+Low-PVI-Replace	**28.4**	**14.1**	**89.2**	**23.7**	**15.3**	84.1

Table 4: Evaluation on dialogue generation. BLEU: Average of BLEU-1,-2,-3,-4. BertS F1: Averaged BERT-score F1. Our model (both with and without PVI fix) outperforms previous work on both dataset 2 and our newly collected dataset 3 in all metrics. Nevertheless, the performance of all models diminishes on dataset 3 due to possible distribution shifts.

5.3. Qualitative Analysis

We present the output examples from our neuro-symbolic goal summarizer and dialogue generator, comparing them to previous work. We also demonstrate generation from GPT-3.5-turbo[1] (two-shot in-context exemplars due to the input context limit), highlighting the limitation of utilizing LLMs in low-resource healthcare settings.

Table 5 compares goal summarization results from various models given the same dialogue history example. The dialogue example contains misleading information, such as *"Friday"* and *"2 miles a day"* (highlighted in red in the Table), while the part of the ground-truth information is in the previous goal (*"7 days a week"*), as indicated by the bolded text *"same as last week"* that coach and patient agree to. Our previous methodologies, which rely on slot-filling and update dialogue states as the goal by either the last appearance of the slot-value (Gupta et al., 2021) or a carryover classification (Zhou et al., 2022) and neglect the global context, fail to extract the correct goal information. In contrast, not only does our model summarize *"3 miles"* from the dialogue but also generates the instruction of *copy(days)*, applying this instruction to the previous goal to complete the summarization with *"7 days a week."* Moreover, our model delivers outputs in a natural language format, which is more user-friendly and comprehensible, particularly for health coaches, compared to the slot-values format. Lastly, the result from GPT3.5-turbo only captures partial information but asserts that the goal is a well-defined SMART goal.

As illustrated in Table 6, our Extended Point-wise $\mathcal{V}$-usable Information (PVI) for Generation measures unconventionality in the patient's response given the dialogue context. A significantly negative value indicates that the model struggles to anticipate the response, implying a higher degree of surprise in the patient's response. Conversely, a positive value suggests a more predictable and conventional response. Responses that either fail to directly address the coach's question (example #1-3) or fall outside the domain of the conversation (example #4, the patient intended to show her dog named King watching Youtube) are assigned a negative value. On the other hand, if the patient's response closely aligns with the coach's utterance, PVI will be positive. Interestingly, the response "short of breath" is unconventional while aligning with the coach's question, resulting in a moderately low value.

Table 7 shows response generation outputs from different models. In Dialogue #1, the model by (Zhou et al., 2022) failed to generate a coherent response due to the model's conditioning on the state of goal slot values while disregarding the critical contextual information ("set the same goals"). Interestingly, GPT-3.5 recognizes the patient's adherence to their goals but still prompts for a new goal definition. However, our model generates a coherent response that benefits from accurate Neuro-symbolic goal summarization. In Dialogue #2, Our previous model failed to generate an appropriate response due to an error in updating the goal stage. The model incorrectly predicts the dialogue to be in the goal-setting stage, hence attempting to fill the empty attribute of the goal (e.g., days). In contrast, our NS generation does not utilize or predict stage information. Instead, it infers the dialogue progression, which is more likely to be in the goal implementation stage, based on a sequence of discrete units. It then focuses on the patient's concerns rather than completing the goals. Surprisingly, GPT3.5 failed to respond to the patient's concern, even when explicitly indicated in the prompt. The complete prompt for GPT3.5 dialogue generation can be found in the appendix.

6. Conclusion

We propose a neuro-symbolic goal summarization approach that (i) summarizes the goal as effectively as possible, given the current week's dialogue, and (ii) generates an executable instruction on modifying the summarized goal by referencing the previous goal. We also build a text-units-text dialogue generation approach that considers as input a sequence of discrete units symbolizing the dialogue history. Our models outperform previous state-of-the-art while eliminating the need for a pre-designed schema or corresponding annotations.

[1] https://platform.openai.com/docs/guides/chat

Dialogue + [Previous Goal: walk 2 miles a day 7 days a week.]
Coach: Also, what would you like to set as your SMART goal this week? Patient: Hey bro last Friday I met my goal but I didn't have the Fitbit on. Patient: On Tuesday I had it on but it went dead. I met my goal on that day also. My goal for this week is 2 miles a day. Coach: No problem, making the goal is most important. Your body is your tracker. Patient: Ok I'm excited and thanks for your help. Coach: Okay so you want to **stick with your same goal** as last week? FYI 2 miles is 4000 steps. Patient: Let's make it 3 miles. Coach: Base on your steps the first week I think 3 miles a day **is doable** if we monitor your step and have a plan... Coach: **How many days do you want to go for?** Patient: **Same as last week?** Coach: **Sure**. And last what might get in the way of you accomplishing your goal? Patient: Nothing. Patient: I'm ready. Coach: Sounds like a plan. Good luck!!
Model Predictions
Gupta et al. (2021): `{Activity:Steps, Amount:[3 miles, 4000 steps], daytimes:Tuesday}` Zhou et al. (2022): `{Activity:Walk, Amount:3 miles, daytimes:Friday}` **Ours**: Walk 3 miles a day 7 days a week. `Instruction:Copy{Days}` GPT-3.5-turbo: "Goal: 3 miles a day." - This goal is Specific (3 miles), Measurable (can be tracked by steps or distance), Achievable (based on the patient's previous performance), Relevant (to the patient's fitness objectives), and Time-bound (for the duration of the week). It's a well-defined and attainable goal.

Table 5: Qualitative examples of summarized goal from different models.

Dialogue Context → Patient Response	**PVI**
... Coach: and one last thing, how would you rate the goal? → Patient: yes and yes.	-0.11
... Coach: What would you like to set as your SMART goal this week? → Patient: Bam! Did you see the weekend?	-0.259
... Coach: Great and what days it will be? → Patient: Yes, but we can move it up to 8,000 steps.	-0.391
... Coach: It's also important to be kind to yourself. → Patient: This is King. Looking at Youtube.	-0.57
... Coach: Great would you be available at 9am or 11am? → Patient: 11am	1.875
... Coach: How confident are you that you can reach your goal on a scale from 1 - 10? → Patient: 10. Absolutely.	1.498
... Coach: what days would you like to do the elliptical? → Patient: Wednesday, Friday, and Saturday.	1.096
... Coach: oh no! I'm sorry to hear that. What happened? → Patient: Short of breath.	0.78

Table 6: Our Extended Point-wise $\mathcal{V}$-usable Information (PVI) for Generation measures unpredictability in the patient's response given the dialogue context. A large negative value shows the model's difficulty in predicting the response, indicating a higher degree of surprise in the patient's response. On the other hand, a positive value indicates a more predictable and conventional response.

Dialogue #1
Coach: Hi [Name], hope you're well. I forgot to message you yesterday to set your goal for Week 6, my apologies! Coach: You met your goal 5 out of 7 days last week, great job What would you like to do for this week? Patient: Gd morning [Name], I am going to **set the same goals**. With a **7 confidence**.
Model Response
Zhou et al. (2022): That's great. And think about what activity would you like to set for your goal? Walks, stairs, etc **Ours**: Ok, great! Can you think of any barriers that might prevent your goal? GPT-3.5-turbo: That's fantastic to hear that you're sticking to your goals! Let's make sure it's a SMART goal for Week 6. Can you provide more details about the goal, like the specific number of steps or any changes in frequency, time, or location?
Dialogue #2
Coach: Hi [Name]! Hope you had a good weekend and were able to stay cool/dry! How are you feeling about your 5,000 steps goal this week? Patient: Good morning. I'm feeling good about it. I will start today (Monday) this weekend I had a party on Saturday and rested on Sunday. Coach: Sounds like a fun weekend! I'm excited to see your steps for today and the rest of the week! Patient: Loved "Sounds like a fun weekend! Iḿ excited to see your steps for today and the rest of the week!" Coach: You are totally crushing your goal this week! Great job! Keep it up! Patient: Thank you. Coach: Reminder to sync your Fitbit! Patient: What do you mean? Are you unable to track me?
Model Response
Zhou et al. (2022): What days? **Ours**: Oh, Just wanted to remind you. GPT-3.5-turbo: That's great to hear you're feeling good about your goal! How many days this week do you plan to aim for those 5,000 steps?

Table 7: Qualitative examples and error analysis for response generation.

This significantly minimizes the annotation labor cost and development cycle, benefiting healthcare and other resource-constrained scenarios. The framework also exhibits the potential for generalization to other applications, including developing dialogue systems in patient education upon discharge or consultations regarding behavioral change issues. However, there remains to be a disparity between the responses generated by the model and those of expert human levels. Given the nature of data limits and sensitivity in healthcare, generating human-level, fair, and faithful outputs while ensuring the model's interpretability is a significant and challenging direction for future exploration.

7. Acknowledgments

The work is supported by the National Science Foundation under Grant IIS-1838770.

8. Bibliographical References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2021. Conversational neuro-symbolic commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4902–4911.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Susan Butterworth, Ariel Linden, Wende McClay, and Michael C Leo. 2006. Effect of motivational interviewing-based health coaching on employees' physical and mental health status. *Journal of occupational health psychology*, 11(4):358.

Benjamin Chaix, Arthur Guillemassé, Pierre Nectoux, Guillaume Delamon, Benoît Brouard, et al. 2020. Vik: a chatbot to support patients with chronic diseases. *Health*, 12(07):804.

Luc De Raedt, Robin Manhaeve, Sebastijan Dumancic, Thomas Demeester, and Angelika Kimmig. 2019. Neurosymbolic= neural+ logical+ probabilistic. In *NeSy'19@ IJCAI, the 14th International Workshop on NeuralSymbolic Learning and Reasoning*.

Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

George T Doran. 1981. There's a SMART way to write management's goals and objectives. *Management review*, 70(11):35–36.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Yufei Feng, Zi'ou Zheng, Quan Liu, Michael Greenspan, and Xiaodan Zhu. 2020. Exploring end-to-end differentiable natural logic modeling. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1172–1185, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Artur d'Avila Garcez, Marco Gori, Luise Lamb, and Luciano Serafini. 2018. Neural-symbolic learning and reasoning: A survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4869–4881.

Amireh Ghorob. 2013. Supplement: Health coaching: Teaching patients to fish. *Family practice management*, 20(3):40–42.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020a. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256, 1st virtual meeting. Association for Computational Linguistics.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2020b.

Goal summarization for human-human health coaching dialogues. In *The Thirty-Third International Flairs Conference*.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2021. Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–289.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1):1–151.

Shreya Kangovi, Frances K Barg, Tamala Carter, Kathryn Levy, Jeffrey Sellman, Judith A Long, and David Grande. 2014. Challenges faced by patients with low socioeconomic status during the post-hospital transition. *Journal of general internal medicine*, 29(2):283–289.

Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.

Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: a conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–26.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Alex Lamb, Anirudh Goyal, Yunyang Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. 2019. Neuro-symbolic transformers for text classification: A case for symbolic reasoning in deep learning. In *NeurIPS*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *ICLR*.

Shiwali Mohan, Anusha Venkatakrishnan, and Andrea L Hartzler. 2020. Designing an ai health coach and studying its utility in promoting regular aerobic exercise. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(2):1–30.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056. Curran Associates, Inc.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019.

Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. *Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach*, page 194–205. Association for Computing Machinery, New York, NY, USA.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Rosemary Thackeray, Ray M Merrill, and Brad L Neiger. 2004. Disparities in diabetes management practice between racial and ethnic groups in the united states. *The Diabetes Educator*, 30(4):665–675.

David H Thom, Jessica Wolf, Heather Gardner, Denise DeVore, Michael Lin, Andy Ma, Ana Ibarra-Castro, and George Saba. 2016. A qualitative study of how health coaches support patients in making health-related decisions and behavioral changes. *The Annals of Family Medicine*, 14(6):509–516.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. 2022. Towards enhancing health coaching dialogue in low-resource settings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 694–706, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A. Instruction Set

The full set of instruction on goal editing based on the reference is shown in Table 8.

Instruction	Explanation
`" "`	Pass
`Copy {Times}`	Copy times, e.g., "after 8 pm"
`Copy {Days}`	Copy days, e.g., "Mon-Fri"
`Copy {Num}`	Copy number, e.g., "3000 steps"
`Add {Num}`	Add number from previous goal.
`Add {Days}`	Add days from previous goal.
`Copy {All}`	Copy all info from previous.

Table 8: The executable instructions we used in the neuro-symbolic goal summarization.

B. Prompts

We show the templates we use when prompting the language model for goal summarization and response generation.

Prompt for Goal Summarization: `Summarize the SMART goal discussed in the health coaching dialogue. The goal attributes include details such as the activity (e.g., walking), the quantity of the activity (e.g., 4000 steps), the schedule (e.g., Monday-Friday, after 4 pm), locations, and more:/n`

Prompt for Response Generation: `As a health coach, your task is to refine the patient's goal into a SMART goal. Ask for details like frequency, time, duration, location, and confidence level, but only one aspect at a time. After establishing the goal, monitor the patient's progress and keep them engaged. Always address their concerns concisely. Now, consider the following dialogue context and formulate your response:/n`

C. Quantitative Results of LLMs

Due to resource constraints, we run the goal summarization and dialogue generation on 100 randomly selected examples with GPT-3.5-turbo. We show the results in Table 9.

D. Training Details

All the following models use Huggingface Transformers Library (Wolf et al., 2020). The hyperparameters are not extensively fine-tuned.

Response Generation	
PPL	23.6
BLEU	14.21
BertScore	85.62
Goal Summarization	
Semantic Frame Acc	41.0

Table 9: Performance of GPT-3.5-turbo on sampled response generation and goal summarization data.

For the goal summarizer, we use T5-base as the model backbone. To mitigate inefficient sampling, we manually annotate 40 positive examples, i.e., the sequences of instructions and partial goals that result in ground truth. The model was first fine-tuned with dialogue-to-goal pairs and then with the 40 examples we labeled. Finally, we contrast the sampled negative examples (the ones that fail) with the positive examples to update the gradient.

We also use two T5-base models to train our PVI-generation metrics. Model g was trained with the context mapping to the patient response for two epochs, while g' was trained with an empty context with one epoch.

For dialogue generation, we choose k = 15 for deriving the discrete units. We use GPT-2-large as the model backbone with a max sequence length set to 128 since the dialogue history has been symbolized as discrete units. The model was trained for 7.0 epochs with a learning rate of 1e-4, with a batch size of 16. We use sampling during decoding with top-k set to 40 and top-p set to 1.