# COOPER: Coordinating Specialized Agents towards a Complex Dialogue Goal

**Yi Cheng**[1], **Wenge Liu**[2], **Jian Wang**[1], **Chak Tou Leong**[1],
**Yi Ouyang**[3], **Wenjie Li**[1], **Xian Wu**[3], **Yefeng Zheng**[3]
[1]The Hong Kong Polytechnic University
[2]Baidu Inc., Beijing, China
[3]Jarvis Research Center, Tencent YouTu Lab
{alyssa.cheng, jian-dylan.wang, chak-tou.leong}@connect.polyu.hk, kzllwg@gmail.com,
{yiouyang, kevinxwu, yefengzheng}@tencent.com, cswjli@comp.polyu.edu.hk

**Abstract**

In recent years, there has been a growing interest in exploring dialogues with more complex goals, such as negotiation, persuasion, and emotional support, which go beyond traditional service-focused dialogue systems. Apart from the requirement for much more sophisticated strategic reasoning and communication skills, a significant challenge of these tasks lies in the difficulty of objectively measuring the achievement of their goals in a quantifiable way, making it difficult for existing research to directly optimize the dialogue procedure towards them. In our work, we emphasize the multifaceted nature of complex dialogue goals and argue that it is more feasible to accomplish them by comprehensively considering and jointly promoting their different aspects. To this end, we propose a novel dialogue framework, COOPER, which coordinates multiple specialized agents, each dedicated to a specific dialogue goal aspect separately, to approach the complex objective. Through this divide-and-conquer manner, we make complex dialogue goals more approachable and elicit greater intelligence via the collaboration of individual agents. Experiments on persuasion and emotional support dialogues demonstrate the superiority of our method over a set of competitive baselines. Our codes are available at https://github.com/YiCheng98/Cooper.

## Introduction

The use of human language is intentional and purposeful (Austin 1975; Grice 1975). In daily communication, we use language deliberately to achieve various goals, ranging from simple inquiries about a product's pricing to complex objectives like resolving conflicts. Developing goal-oriented dialogue systems has also been a prominent research topic.

In the past few years, there has been growing research interest in dialogue tasks with more complex objectives, such as persuasion (Wang et al. 2019), negotiation (He et al. 2018), and emotional support (Liu et al. 2021b). Compared to traditional service-focused goal-oriented dialogue systems (Rieser and Moore 2005; Boyer et al. 2011; Wen et al. 2016; Liu et al. 2022), these tasks require much more sophisticated strategic reasoning and communication skills. Recent studies show that even state-of-the-art Large Language Models (LLMs) struggle with these tasks, where they exhibit weak awareness of the overall dialogue progression and fail to accomplish a complex dialogue goal through multi-turn interactions strategically (Zhao et al. 2023a). Moreover, another major challenge lies in the difficulty of objectively measuring the achievement of such complex dialogue goals in a quantifiable and reliable way. Consequently, most existing research stays overly focused on how to fit the ground-truth data, without explicit consideration of how each utterance could contribute to the final objective (Zhou et al. 2019a; Joshi et al. 2021; Chen et al. 2023). In the few works that attempt to model these dialogue goals explicitly, it remains highly challenging to optimize the dialogue procedure towards them directly due to their inherent intangibility (Cheng et al. 2022; Sanders et al. 2022; Zhou et al. 2023).

In this work, we highlight the multifaceted nature of complex dialogue goals, which typically encompass multiple interdependent aspects that must be collectively promoted to approach the final objective. For instance, psychological guidelines suggest that Emotional Support Conversations (ESC) should include three key aspects:[1] *exploration* (identify the support-seeker's problem), *comforting* (comfort the seeker's emotion through expressing empathy), and *action* (help the seeker solve the problem) (Hill 2009; Liu et al. 2021b). These aspects are interdependent. For example, exploring the seeker's situation lays the foundation for conveying appropriate empathy, while comforting the user to be in a better emotional state makes them more willing to share details about their experiences and feelings.

Compared with directly optimizing towards the complex dialogue goal, it is more feasible to accomplish it by comprehensively considering and jointly promoting its different aspects. Nonetheless, due to the interdependence among different aspects, the interlocutor still needs to address the challenge of how to strategically coordinate their priority during the conversation. To achieve this, they must dynamically track the states of all the aspects and analyze their progression, that is, how much progress has been achieved so far and where the state of each aspect is heading. As in ESC, a seasoned supporter would continuously record information about the seeker's situation and keep estimating the under-

[1]Some works may refer to the "aspects" here as "stages", but they also emphasize that these "stages" are closely interwoven in practice rather than sequential (Liu et al. 2021b). Given that, we choose to call them as "aspects" uniformly in our work to avoid misunderstanding about their sequential nature.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

lying root problem for further exploration. They would also monitor the progression of the *comforting* and *action* aspects simultaneously. Through comprehensive analysis, the supporter could determine which aspect to prioritize at each point of the conversation.

Based on the above insight, we propose a novel dialogue framework, COOPER, which coordinates multiple specialized agents, each dedicated to a specific aspect separately, to approach a complex dialogue goal. Specifically, by tracking the current state of its assigned aspect, each agent analyzes the progression of this aspect and suggests several topic candidates for the next utterance that can further promote the aspect (e.g., the agent responsible for the *exploration* aspect in ESC will suggest questions to ask the seeker). Then, we coordinate the specialized agents by ranking all the topic candidates with consideration of the overall dialogue progression. Finally, the top-ranked topic candidates are used to guide the generation of the next utterance.

Through this divide-and-conquer manner, we make the complex dialogue goal more approachable and elicit greater intelligence via the collaboration of individual agents. Experiments on ESC and persuasion dialogues demonstrate the superiority of COOPER over a set of competitive LLM-based methods and previous state-of-the-art.

In summary, our contributions are as follows:

- To this best of knowledge, this is the first work that explores how to achieve a complex dialogue goal by coordinating the joint promotion of its different aspects.
- We propose COOPER, an innovative framework that coordinates multiple specialized agents to collaboratively work towards a complex dialogue goal.
- Extensive experiments demonstrate the effectiveness of our approach and also reveal the limitations of current LLMs in handling complex dialogue goals.

## Related Works

In the past few years, there has been growing interest in dialogue generation tasks with complex objectives, such as negotiation (Lewis et al. 2017; He et al. 2018; Zhou et al. 2019b), persuasion (Wang et al. 2019; Li et al. 2020; Samad et al. 2022), and emotional support (Liu et al. 2021a; Peng et al. 2022; Xu, Meng, and Wang 2022; Zhao et al. 2023b). Previous methods in these tasks can be mainly grouped into three categories: dialogue strategy learning (Zhou et al. 2019a; Joshi et al. 2021), user modeling (Yang, Chen, and Narasimhan 2021; Shi et al. 2021; Tran, Alikhani, and Litman 2022), and fusing external knowledge (Tu et al. 2022; Chen et al. 2022; Deng et al. 2023b). Among these works, only very few of them have an explicit consideration of the dialogue goal and how each generated utterance contributes to achieving the final objective. For example, Cheng et al. (2022) predicted the support strategy in ESC by estimating how much the user emotion would be improved with an A*-like algorithm. Zhou et al. (2023) optimized the ESC process through reinforcement learning, using the extent of how much the user's positive emotion is elicited as reward. Sanders et al. (2022) conducted persuasive dialogue generation by measuring the distance of the current dialogue state relative to the desired outcome. However, it is challenging to measure the achievement of these complex dialogue goals objectively in a quantifiable way. For example, assessing how much the user's positive emotion is elicited simply based on the dialogue is extremely difficult in ESC. Directly optimizing towards a complex dialogue goal can be exceptionally hard, even for humans. In real scenarios, the guidelines for these challenging dialogue tasks usually recommend breaking down the complex goals into multiple aspects and jointly promoting them to work towards the broad objective (Petty et al. 1986; Fershtman 1990; Hill 2009).

More recently, several works have applied LLMs to complex goal-oriented dialogues, by directly prompting the LLM to generate utterances (Zhao et al. 2023a; Deng et al. 2023a) or further improving the performance via iterative revision (Fu et al. 2023). Current LLMs exhibit remarkable improvement compared to the previous methods on these tasks, but it is also found that they tend to lack a larger picture of the overall dialogue progression and fail to achieve the dialogue objective strategically through multi-turn interactions (Deng et al. 2023a). For example, on the task of ESC, they often continuously offer coping suggestions and overlook the critical process of exploring the user's situation and expressing empathy (Zhao et al. 2023a).

## Preliminaries

**Problem Formulation** We consider the problem of how to achieve a complex dialogue goal that encompasses multiple aspects, denoted as $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_{n_T}\}$, where $n_T$ is the number of aspects. Given the dialogue history $\mathcal{H}^t$ at the $t$-th dialogue round, the system generates the next utterance $\mathcal{U}^t$, which promotes one or several dialogue goal aspects.

**ESC Framework** Following the ESC framework defined by Liu et al. (2021b), our implementation considers the following aspects for effective emotional support: 1) *Exploration*: identify the support-seeker's problems that cause their distress; 2) *Comforting*: comfort the seeker's emotion by expressing empathy and understanding; 3) *Action*: help the seeker conceive actionable plans to resolve the problems.

**Persuasion Dialogues** Referring to the elaboration likelihood model of persuasion proposed by Petty et al. (1986), we consider the following aspects within the broad goal of persuasion in our implementation: 1) *Attention*: capture the persuadee's attention and elicit their motivation to discuss the related topic; 2) *Appeal*: present persuasive arguments via different strategies and encourage the persuadee to think deeply about the arguments; 3) *Proposition*: explicitly state the persuader's position or call to action, and seek confirmation of the persuadee's attitude towards the proposition.

## Method

Figure 1 presents an overview of our proposed framework. In this section, we illustrate the three major steps within it, as well as its training procedure.

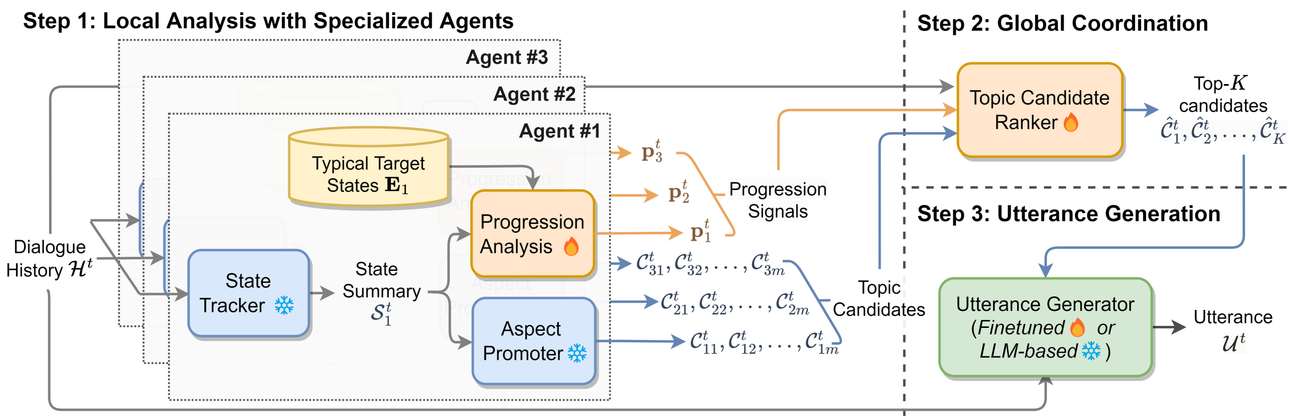


Figure 1: Illustration of our proposed framework COOPER (suppose the number of aspects within the dialogue goal $n_T$=3). The icons of snowflake and flame denote that the module is frozen (LLM prompt-based) or finetuned, respectively.

## Local Analysis with Specialized Agents

We devise multiple specialized agents to separately tackle different dialogue goal aspects. We denote them as $\{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_{n_T}\}$, with agent $\mathcal{A}_i$ dedicated the aspect $\mathcal{T}_i$ ($i$=1, 2, .., $n_T$). Each agent consists of three modules: a *state tracker*, an *aspect promoter*, and a *progression analysis* module.

Given the context $\mathcal{H}^t$ at the $t$-th dialogue round, the state tracker of $\mathcal{A}_i$ utilizes an LLM to summarize the current state of its assigned aspect, producing a summary $\mathcal{S}_i^t$. For example, in order to get the state summary for the *exploration* aspect in ESC, we prompt the LLM to "*summarize the seeker's experience that caused their emotional distress*".[2]

The aspect promoter in $\mathcal{A}_i$ then suggests $m$ topic candidates $\{\mathcal{C}_{i1}^t, \mathcal{C}_{i2}^t, ..., \mathcal{C}_{im}^t\}$ that can be used to further promote the assigned aspect, based on $\mathcal{H}^t$ and $\mathcal{S}_i^t$. This module is also realized by prompting an LLM. The topic candidates here can be seen as a brief content outline for the following utterance. For instance, the aspect promoter of the *exploration* agent in ESC is implemented by instructing an LLM to "*list* <*m*> *questions that the supporter can ask the seeker to further understand their situation (each less than 20 words)*".

The progression analysis module in $\mathcal{A}_i$ produces a signal $\mathbf{p}_i^t$ for its assigned aspect. This signal is expected to indicate *how much progress has been achieved so far* regarding this aspect and its *estimated target state* at the end of conversation. To achieve this, we construct a state embedding space to consider the evolving path of the past states in this space and estimate the position of the potential target state regarding each aspect. Specifically, given the state summary $\mathcal{S}_i^t$, we map it into the state embedding space by encoding it with a pretrained sentence encoder, MPNet (Song et al. 2020). We denote the encoded embedding of $\mathcal{S}_i^t$ as $\mathbf{s}_i^t \in \mathbb{R}^{n_d}$, where $n_d$ is the dimension of the state embedding. Intuitively, the information in $\mathbf{s}_i^t$ summarizes the progress has been made so far regarding the aspect $\mathcal{T}_i$.

To estimate the target state of $\mathcal{T}_i$, we first resort to the dialogues in the training set and record the states of each aspect at the end of these conversations to obtain the typical target states of this aspect. For instance, to obtain the typical target states for the *exploration* aspect in ESC, for each dialogue in the training set, we adopt the same practice as in the state tracker to summarize the seeker's problem based on the complete dialogue. Then, we map these summaries to the state embedding space. Denote the matrix that encompasses all the obtained target state embeddings of this aspect as $\mathbf{E}_i \in \mathbb{R}^{N_D \times n_d}$, where $N_D$ is the number of dialogues in the training set. After that, we cluster the embeddings in $\mathbf{E}_i$ through the $k$-means algorithm (Hartigan and Wong 1979), where the number of clusters $k_i$ is determined based on the silhouette score (Rousseeuw 1987) of the clustering results. We denote the centroids of these clusters as $\{\mathbf{e}_i^1, \mathbf{e}_i^2, ..., \mathbf{e}_i^{k_i}\}$. Intuitively, these centroids represent the typical final states of the aspect $\mathcal{T}_i$. The above clustering process is finished offline before inference. At the inference stage, we estimate the potential target state of $\mathcal{T}_i$ for the current dialogue by attending the state embedding $\mathbf{s}_i^t$ to the above centroids. Formally, we calculate the estimated target state $\mathbf{v}_i^t$ as follows:

$$h_{ij} = (\mathbf{W}_i \mathbf{s}_i^t) \cdot (\mathbf{W}_i \mathbf{e}_i^j),$$
$$\alpha_{ij} = \frac{\exp(h_{ij})}{\sum_{l=1}^{k_i} \exp(h_{il})},$$
$$\mathbf{v}_i^t = \mathrm{ReLU}(\sum_{j=1}^{k_i} \alpha_{ij} \mathbf{e}_i^j),$$

where $\mathbf{W}_i \in \mathbb{R}^{n_d \times n_d}$ is a trainable matrix. Finally, we get the progression signal $\mathbf{p}_i^t = [\mathbf{v}_i^t; \mathbf{s}_i^t]$, where $\mathbf{p}_i^t \in \mathbb{R}^{2 \times n_d}$ and [;] represents the vertical concatenation operation of vectors.

## Global Coordination

With the local analysis results from the specialized agents, we conduct global coordination among them by ranking all the topic candidates with consideration of the progression signals. Specifically, we learn a scoring function $f(\cdot)$ and conduct ranking based on the scoring results of the topic candidates. Here, we mainly explain the inference process in the global coordination module, and will leave the illustration of its training procedure at the end of this section.

[2] For all the prompt-based methods mentioned in this paper, we provide the detailed prompt templates in the appendix.

During inference at the $t$-th round, we calculate the score $f(\mathcal{H}^t, \mathcal{C}^t_{ij})$ for each topic candidate $\mathcal{C}^t_{ij}$ ($i$=1, 2, ..., $n_T$; $j$=1, 2,..., $m$). To achieve this, we first concatenate $\mathcal{C}^t_{ij}$ with $\mathcal{H}^t$ and encode them with a Transformer (Vaswani et al. 2017):

$$\mathbf{B}^t_{ij} = \text{TRS}[\text{Emb}(\texttt{[CLS]} \oplus \mathcal{H}^t \oplus \mathcal{C}^t_{ij})],$$

where TRS denotes the Transformer encoder, Emb($\cdot$) represents the operation of the embedding layer, and $\oplus$ refers to the operation of text concatenation. We take the encoded hidden vector corresponding to the [CLS] token, denoted as $\widetilde{\mathbf{b}}^t_{ij}$. Then, to take the progression signals into account, we pass all the progression signals through a multilayer perceptron (MLP), denoted as $\text{MLP}_{\text{PRG}}$:

$$\widetilde{\mathbf{p}}_t = \text{MLP}_{\text{PRG}}(\mathbf{p}_1; \mathbf{p}_2; ...; \mathbf{p}_{n_T}),$$

where $\widetilde{\mathbf{p}}_t \in \mathbb{R}^{n_d}$. Finally, we obtain the score $f(\mathcal{H}^t, \mathcal{C}^t_{ij})$ by passing $\widetilde{\mathbf{p}}_t$ and $\widetilde{\mathbf{b}}^t_{ij}$ through a single feedforward layer:

$$f(\mathcal{H}^t, \mathcal{C}^t_{ij}) = \text{FF}(\widetilde{\mathbf{p}}_t \mid \widetilde{\mathbf{b}}^t_{ij}),$$

where FF($\cdot$) represents the feedforward layer and | refers to the horizontal concatenation operation of two vectors into one long vector. By sorting the scores of all the topic candidates, we obtain the top-$K$ candidates $\{\hat{\mathcal{C}}^t_1, \hat{\mathcal{C}}^t_2, ..., \hat{\mathcal{C}}^t_K\}$, where the subscripts represent their ranking (i.e. $\hat{\mathcal{C}}^t_1$ is the candidate with the highest score).

## Utterance Generation

The top-$K$ ranked topic candidates are then used to guide the utterance generation. We experiment with two ways of implementing the utterance generator: a finetuned approach and an LLM prompt-based approach. Intuitively, the former way can learn the nuanced patterns specific to the complex dialogue task directly from the dataset, while the latter can leverage the remarkable performance of the LLM, which is supposed to have better generalization in various scenarios. The finetuned approach is developed upon BART (Lewis et al. 2020). Specifically, we concatenate the top-$K$ topic candidates, the state summaries of all the aspects $\{\mathcal{S}^t_1, \mathcal{S}^t_2, ..., \mathcal{S}^t_{n_T}\}$, and the dialogue context $\mathcal{H}^t$ as its input, separated with the special token [SEP]. For the prompt-based approach, we directly utilize an LLM to generate the next utterance $\mathcal{U}^t$, where the prompt includes the dialogue history $\mathcal{H}^t$ and the top-$K$ topic candidates.

In the following, we will refer to our framework that uses the finetuned generator as **COOPER$_{\text{(FT-G)}}$** and the one that adopts the LLM prompt-based generator as **COOPER$_{\text{(PT-G)}}$**.

## Training

For COOPER$_{\text{(PT-G)}}$, we train the progression analysis modules and the ranker in an end-to-end manner, optimizing with the weighted sum of the triplet ranking loss (Schroff, Kalenichenko, and Philbin 2015) and the pointwise loss. Specifically, the triplet loss is defined as:

$$\mathcal{L}_t = \sum_{\hat{g}(\mathcal{C}^t_{ij}) < \hat{g}(\mathcal{C}^t_{i'j'})} \max(0, f(\mathcal{H}^t, \mathcal{C}^t_{ij}) - f(\mathcal{H}^t, \mathcal{C}^t_{i'j'}) + \tau),$$

where $\tau$ represents the margin enforced between the positive and negative pairs, and $\hat{g}(\cdot)$ returns the ranking label of the given topic candidate. The pointwise loss is defined as:

$$\mathcal{L}_p = \frac{1}{n_T \cdot m} \sum_{i,j} (\hat{g}(\mathcal{C}^t_{ij}) - g(\mathcal{C}^t_{ij}))^2,$$

where $g(\cdot)$ returns the predicted ranking position of the given topic candidate from our method. The overall ranking loss function is the combination of them:

$$\mathcal{L}_R = \alpha \cdot \mathcal{L}_t + (1 - \alpha) \cdot \mathcal{L}_p,$$

where $\alpha$ is a hyperparameter that balances the two losses. Since the experimental datasets do not contain the ground-truth labels for topic candidate ranking, we conduct pseudo-labeling and determine whether $g(\mathcal{C}^t_{ij}) < g(\mathcal{C}^t_{i'j'})$ using the following criteria. First, we compare if one of the two candidates aims to promote the ground-truth dialogue goal aspect[3] while the other does not. In such cases, the former is ranked higher than the latter. If this criterion cannot enable a comparison, we then consider the text similarity between the candidate and the ground-truth utterance, ranking the more similar one as superior. The text similarity is measured by computing the inner product of their sentence embeddings encoded with MPNet.

For COOPER$_{\text{(FT-G)}}$, we also need to finetune the utterance generator. We train it separately from the progression analysis modules and the ranker in a pipeline way. It is optimized with the generation loss $\mathcal{L}_G$, defined as the negative log-likelihood of the ground-truth token.

# Experiments

## Experimental Setup

**Datasets** Our experiments are conducted on the **ESConv** dataset (Liu et al. 2021b) and the **P4G** dataset (Wang et al. 2019). ESConv is an ESC dataset, including 1,300 conversations. We follow the setting in (Cheng et al. 2022) for its data preprocessing and data split. After preprocessing, there are 1,040/130/130 conversations in the training/validation/test sets, with an average of 11.7 rounds of interactions in each dialogue. P4G is a persuasion dialogue dataset, including 1,017 dialogues with an average of 10.4 dialogue rounds. We distribute 867/50/100 conversations into the training/validation/test sets. Both datasets include the annotation of which dialogue strategies are adopted by the supporter/persuader, based on which we can infer which dialogue goal aspects are promoted in a ground-truth utterance (more details are included in the appendix).

**Baselines** Our baselines include several LLM prompt-based methods and the previous state-of-the-art methods on two experimental datasets. Specifically, we consider the following prompt-based methods: **GPT-3.5** prompts an LLM to generate the next utterance based on a brief task description and the dialogue history, following the similar format as in (Zheng et al. 2023); **GPT-3.5+CoT** prompts an LLM to conduct chain-of-thought reasoning (Wei et al. 2022) about the

[3]We infer which aspects are promoted by a ground-truth utterance based on the dialogue strategy annotation in the dataset.

| Dataset | Generation Paradigm | Model | BL-1 | BL-2 | BL-4 | RG-L | MET | Dist-1 | Dist-2 | Dist-3 |
|---|---|---|---|---|---|---|---|---|---|---|
| ESConv | Prompt-based | GPT-3.5 | 17.16 | 5.04 | 1.02 | 15.44 | 9.12 | 4.50 | 25.53 | 47.72 |
| | | GPT-3.5+CoT | 15.86 | 4.66 | 0.94 | 14.42 | 9.36 | 4.29 | 24.61 | 47.62 |
| | | MixInit | 16.26 | 4.65 | 0.93 | 14.52 | 9.32 | 3.64 | 20.88 | 40.33 |
| | | **COOPER$_{(PT-G)}$** | **17.62** | **5.42** | **1.11** | **15.86** | **9.36** | **5.22** | **29.45** | **54.40** |
| | Finetuned | KEMI | 20.94 | 8.71 | 2.67 | 17.48 | 8.31 | 2.77 | 15.26 | 30.22 |
| | | MultiESC | 21.30 | 9.19 | 3.06 | **20.24** | 8.69 | 3.54 | 16.70 | 31.07 |
| | | **COOPER$_{(FT-G)}$** | **22.76** | **9.54** | **3.11** | 20.18 | **9.22** | **5.02** | **24.22** | **43.55** |
| P4G | Prompt-based | GPT-3.5 | **21.05** | 8.31 | 2.01 | 16.19 | 10.55 | 4.50 | 19.66 | 34.33 |
| | | GPT-3.5+CoT | 18.74 | 7.37 | 1.99 | 15.86 | **10.71** | 3.86 | 19.34 | 36.68 |
| | | MixInit | 16.83 | 6.22 | 1.36 | 14.56 | 10.69 | 3.42 | 17.39 | 32.94 |
| | | **COOPER$_{(PT-G)}$** | 20.76 | **8.68** | **2.48** | **16.84** | 10.55 | **5.28** | **23.38** | **41.16** |
| | Finetuned | ProAware | 18.40 | 7.60 | 2.61 | 16.92 | 7.92 | 4.78 | 23.25 | 42.90 |
| | | ARDM | 21.17 | 9.73 | 3.73 | 17.19 | 8.98 | 4.99 | 24.20 | 45.19 |
| | | **COOPER$_{(FT-G)}$** | **23.88** | **11.44** | **4.67** | **18.83** | **9.96** | **5.35** | **25.58** | **46.90** |

Table 1: Static evaluation results on the ESConv and P4G datasets.

progression state of each dialogue goal aspect and determine which aspect needs to be prioritized in the current round before utterance generation; **MixInit** (Chen et al. 2023) explicitly indicates what dialogue strategies are used by the interlocutors in the dialogue history and requires the LLM to predict which strategy to adopt in the next utterance before generation. We also compare with several state-of-the-art methods that adopt finetuned generators, which are **MultiESC** (Cheng et al. 2022) and **KEMI** (Deng et al. 2023b) for ESC; **ARDM** (Wu et al. 2021) and **ProAware** (Sanders et al. 2022) for persuasion dialogues. More details about the baselines are provided in the appendix.

**Implementation Details** All the prompt-based modules in COOPER and the prompt-based baselines are implemented with `gpt-3.5-turbo`. On both datasets, there are three specialized agents focusing on different dialogue goal aspects.[4] We set $m$=4 on the ESConv dataset (i.e., each agent needs to produce four topic candidates) and $m$=3 on the P4G dataset. We set $K$=3 on both datasets (i.e., the top-3 topic candidates are used to guide utterance generation). In the global coordination module, we set $\alpha$=0.9 and $\tau$=0.2. For KEMI, MultiESC, ProAware, and ARDM, we use their released codes to conduct the experiments. More implementation details are provided in the appendix.

## Static Evaluation

We conduct static evaluation on the generated utterances, by comparing them with the ground-truth ones in the datasets. We use the following automatic metrics: BLEU-1/2/4 (**BL-1/2/4**) (Papineni et al. 2002), which measure the $n$-gram precision; ROUGE-L (**RG-L**) (Lin 2004), which measures the recall of longest common subsequences; METEOR (**MET**) (Lavie and Agarwal 2007), which further considers stem match or synonymy match; Distinct-1/2/3 (**Dist-1/2/3**), which calculates the ratios of unique $n$-grams.

**Comparison with Baselines** The evaluation results are presented in Table 1. For clarity, we classify the compared models into two categories with respect to their utterance generation paradigm: the LLM prompt-based and the finetuned ones. On both datasets, the two variants of our framework (COOPER$_{(PT-G/FT-G)}$) outperform the baselines within the same category in terms of the overall performance, demonstrating the effectiveness of our proposed method.

Among the prompt-based methods, COOPER$_{(PT-G)}$ performs significantly better in Dist-1/2/3, which indicates superior diversity of the generated content. A very likely reason is that the other prompt-based methods tend to be biased towards one specific aspect of the dialogue goal, which we will further discuss in later experiments. In comparison, our method can comprehensively consider all the aspects by brainstorming topic candidates from each of them and fusing the most appropriate ones in the generated utterance. Surprisingly, the two baselines that are deliberately prompted to reason about the dialogue progression and dialogue strategy (GPT-3.5+CoT and MixiInit) perform even worse than GPT-3.5. It demonstrates that the LLM is poor at reasoning about how to approach a complex dialogue goal strategically. The explicit reasoning process even magnifies their differences from human behavior. In our framework, we bridge this gap with the global coordination module, which learns to select the most appropriate topic candidates produced by LLMs with supervision from the ground-truth data.

In the finetuned category, COOPER$_{(FT-G)}$ also performs the best, although it does not implement any complex mechanisms in the utterance generator as some baseline models do. This mainly benefits from the state summaries and the appropriate topic candidates produced by the other LLM-based modules, which are concatenated in the input. The finetuned methods generally achieve better scores than the prompt-based ones in the static evaluation, but as they receive much more supervision from the training data, we cannot arrive at the conclusion that they are more competitive. We conduct the interactive evaluation for further analysis.

**Ablation Study** To examine the effects of different modules in our framework, we conduct ablation studies by comparing the complete COOPER$_{(FT-G)}$ framework with its following variants on the ESConv dataset: (1) ***w/o* GCord** does not incorporate topic candidate ranking and directly passes all the topic candidates to the utterance generator; (2) ***w/o***

[4]Please refer to the "Preliminaries" section about the dialogue goal aspects that we consider in ESC and persuasion dialogues.

| Compared Models | Coherence | | Natural | | Identification | | Empathy | | Suggestion | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Win | Lose | Win | Lose | Win | Lose | Win | Lose | Win | Lose |
| COOPER$_{(FT-G)}$ vs. MultiESC | 24.2 | **27.5** | **36.9**$^‡$ | 19.6 | **17.3**$^†$ | 12.7 | **45.0**$^‡$ | 21.9 | **38.1**$^‡$ | 28.8 |
| COOPER$_{(PT-G)}$ vs. GPT-3.5 | **20.8** | 17.7 | **78.5**$^‡$ | 10.0 | **41.5**$^†$ | 36.9 | **67.7**$^‡$ | 19.2 | **25.4**$^†$ | 18.5 |
| COOPER$_{(PT-G)}$ vs. COOPER$_{(FT-G)}$ | **83.8**$^‡$ | 13.1 | **75.4**$^‡$ | 14.6 | **81.5**$^‡$ | 13.1 | **74.6**$^‡$ | 10.0 | **82.3**$^‡$ | 10.8 |

Table 2: Interactive evaluation results (%). The columns of "Win/Lose" indicate the proportion of cases where the former model in that set of comparisons wins/loses. †/‡ denote $p$-value $< 0.1/0.05$ (statistical significance test).

| Model | BL-1 | BL-2 | RG-L | MET | Dist-2 |
|---|---|---|---|---|---|
| **COOPER$_{(FT-G)}$** | **22.76** | **9.54** | **20.18** | **9.22** | **29.42** |
| *w/o* GCord | 19.73 | 8.28 | 19.94 | 8.51 | 24.27 |
| *w/o* ProAna | 21.11 | 8.55 | 19.36 | 8.77 | 26.17 |
| *w/o* TProm | 20.51 | 8.80 | 20.03 | 8.28 | 22.03 |
| *w/o* STrack | 20.07 | 8.76 | 19.86 | 7.99 | 25.85 |

Table 3: Ablation study on the ESConv dataset.

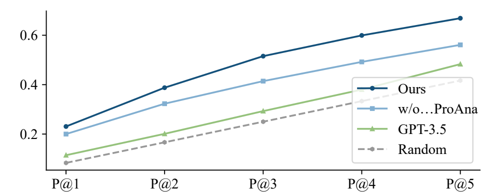


Figure 2: Precision@$n$ of our topic candidate ranking approach and the baseline methods on the ESConv dataset.

ProAna performs topic candidate ranking without progression signals; (3) ***w/o* TProm** does not produce topic candidates and the input of the utterance generator only includes dialogue history and state summaries; (4) ***w/o* STrack** does not concatenate the state summaries to the input of the utterance generator. As shown in Table 3, the ablation of any component leads to a decrease in performance, indicating the indispensability of each component in contributing to the overall performance. Comparatively, the performance decline in "*w/o* GCord" is the most significant. It means that some low-quality topic candidates produced by the LLM can only introduce noise for utterance generation, which underscores the importance of conducting global coordination and filtering these low-quality candidates. The performance drop in "*w/o* STrack" is also notable, suggesting their importance in capturing the key information in the long context.

## Interactive Evaluation

We simulate realistic conversations with the systems to further assess their performance in an interactive setting. We adopt a similar practice as done in (Li et al. 2023), using ChatGPT to play the role of an emotional support seeker and converse with the evaluated system. Specifically, for each dialogue in the test set of ESConv, we summarize the seeker's problem in it and then prompt ChatGPT to simulate their process of seeking emotional support based on the summary. Given a pair of dialogues produced by conversing with two compared systems about the same problem, we manually assess which one is better (or select *tie*) in the following dimensions: (1) **Coherence**: which model generates more coherent content with the context; (2) **Natural**: which model is more natural and human-like; (3) **Identification**: which model can more effectively explore the seeker's problem; (4) **Empathy**: which model shows better empathy to the seeker; (5) **Suggestion**: which model provides more practical suggestions tailored to the seeker's situation. Five graduate students with linguistic backgrounds are recruited as the annotators. We compare COOPER$_{(FT-G)}$ and COOPER$_{(PT-G)}$ with MultiESC and GPT-3.5, two representative baselines in different categories, respectively. We also conduct a comparison between the two variants of COOPER to evaluate which kind of implementation is better for utterance generation.

As shown in Table 2, COOPER$_{(PT-G)}$ outperforms GPT-3.5 in all metrics, especially in the dimensions of "natural" and "empathy". It is because GPT-3.5 often generates too much advice in a didactic tone and largely overlooks the comforting process. Their generations also often follow a similar pattern, which seems unnatural, as we will show in the case study. In contrast, our method can balance all aspects more appropriately. Besides, despite much more advice generated by GPT-3.5, they are still slightly worse in terms of "suggestions", as their suggestions are usually too general and unable to tailor to the seeker's situation. COOPER$_{(FT-G)}$ also outperforms the competitive finetuned baseline, MultiESC, in terms of the overall performance. Nonetheless, compared with the LLM-based methods, neither of the two methods that use small language models as backbones for generation can facilitate multi-turn interactions very effectively. Their generated content is usually very repetitive and general, making it difficult for the annotators to determine the better one, so the proportion of ties is relatively high in this set of comparisons. For the two variants of our method, we can see that COOPER$_{(PT-G)}$ performs significantly better than COOPER$_{(FT-G)}$, demonstrating that LLM-based methods are a better choice for demanding dialogue tasks like ESC.

## Analysis of Global Coordination

**Analysis of Topic Candidate Ranking** We analyze the topic ranking performance of the global coordination module in COOPER by comparing it with the following methods: (1) ***w/o* ProAna** is a variant of our method, which conducts topic ranking without progression signals; (2) **GPT-3.5** prompts `gpt-3.5-turbo` to select the top-$k$ topic candidates given the dialogue history; (3) **Random** ranks the topic candidates randomly. We use Precision@$n$ as our evaluation metric, which measures the proportion of relevant items among the top $n$ results. Figure 2 displays the evaluation results on the ESConv dataset. We can see that our method for topic ranking performs the best in terms of Precision@$n$. Comparing our method with "*w/o* ProAna", we can observe that the performance improvement brought

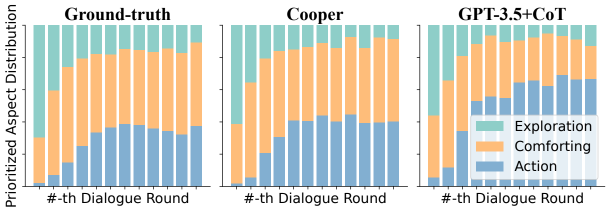


Figure 3: The distribution of the prioritized dialogue goal aspects with respect to the dialogue progress, in the ground-truth data, COOPER, and GPT-3.5+CoT on ESConv.

by progression signals is significant, which underscores the importance of analyzing the current progression of each dialogue goal aspect when determining the topic of the next utterance. GPT-3.5 exhibits limited performance in topic candidate ranking, with only a marginal advantage over the random method. It demonstrates that GPT-3.5's inclination towards dialogue content planning diverges greatly from human behavior in complex dialogue tasks like ESC, thus being unable to address them very effectively.

**Analysis of Prioritized Aspects** For further examination, we analyze which dialogue goal aspect is more frequently prioritized with respect to the dialogue progress. To this end, we visualize the distribution of the prioritized dialogue goal aspects from the first to the twelfth dialogue rounds on the ESConv dataset. Specifically, since each topic candidate is produced by one agent responsible for a particular dialogue goal aspect in COOPER, we regard the aspect of the top-1 ranked candidate as the primarily prioritized aspect in the current round. For comparison, we also visualize the distributions in the ground-truth data and GPT-3.5+CoT, which is prompted to explicitly reason about the prioritized aspect.

As shown in Figure 3, we can see that, in the ground-truth data, the *exploration* aspect is more frequently promoted at the beginning of the conversation and its frequency gradually decreases over time, while the proportion of *action* gradually increases. In contrast, the proportion of *comforting* remains relatively stable throughout the dialogue, consistently accounting for approximately 50%. In COOPER and GPT-3.5+CoT, we can observe a similar variation trend. However, the proportional distributions in COOPER align more closely with those in the ground-truth data than GPT-3.5+CoT does. In contrast to the ground-truth data, GPT-3.5+CoT demonstrates a significantly higher presence of the *action* aspect and a considerably lower presence of *comforting*. This result is consistent with the findings in (Zhao et al. 2023a), wherein it was observed that ChatGPT often rushes to provide advice and overlooks the comforting process. The same problem also manifests in persuasion dialogues, where the three prompt-based baselines often hasten to propose the call to action before the time is ripe. It demonstrates that current LLMs still lack the capability of approaching a complex dialogue goal strategically through multiple interactions.

## Case Study

The case study in Table 4 compares the utterances generated by COOPER$_{\text{PT-G}}$ and GPT-3.5 in an ESC. We can see that

| |
|---|
| *(1st Round)* **User**: I am stressed that my 25-year-old son is moving back. My boyfriend and I have a home but my son hates him. |
| **COOPER**$_{\text{(PT-G)}}$: It must be difficult if there's tension between your son and your boyfriend. Is there anything specific that makes your son doesn't like him?<br>**GPT-3.5**: I'm sorry to hear that you're getting through a stressful situation. It's important to approach this with open communication. Have you tried ... It's also essential to ... Remember, ... |
| *(3rd Round)* **User**: My boyfriend thinks my son should speak to me more respectfully. He gets upset when my son shouts at me. |
| **COOPER**$_{\text{(PT-G)}}$: It's important to address any tensions and find a balance that works for everyone involved. How has the presence of your son affected your relationship with your boyfriend?<br>**GPT-3.5**: It sounds like you're dealing with a challenging situation. . . It might be helpful to have an open and honest conversation ... Encourage them to ..., and try to ... Remember,. . . |
| *(6th Round)* **User**: My son is going back to school and I am supporting him in that. I am not going to put him out of the house. But he sometimes is verbally aggressive. |
| **COOPER**$_{\text{(PT-G)}}$: It's understandable that you're torn between your son and your boyfriend. You want to support your son, but also maintain a healthy relationship with your boyfriend. It might be helpful to have a calm and open conversation with both of them to express your concerns and find a way to improve the situation.<br>**GPT-3.5**: It sounds like you're in a difficult situation. It's important to communicate openly ... Maybe you can ... It might be helpful to find a compromise ... Remember, ... |

Table 4: Case study (the first, third, and sixth rounds of an example dialogue on the ESConv dataset).

COOPER$_{\text{PT-G}}$ outperforms GPT-3.5 in many aspects. One significant problem with GPT-3.5 is its repetitive response pattern, briefly acknowledging the user's problem followed by much generic advice. In contrast, COOPER$_{\text{PT-G}}$ demonstrates a deeper understanding of the user's situation and provides more varied responses tailored to the user's situation. For example, at the third round of interaction, it identifies the son's behavior might have an impact on the relationship between the user and her boyfriend; at the sixth round, it points out the dilemma between supporting her son and maintaining a healthy relationship with her boyfriend. Moreover, COOPER$_{\text{PT-G}}$ can more effectively guide the emotional support procedure by asking open-ended questions and providing personalized insights, which helps facilitate a more productive and meaningful exchange.

## Conclusion

This paper investigated how to construct dialogue systems that can achieve complex dialogue goals. We highlighted the importance of comprehensively considering the multiple aspects within a complex dialogue goal, as it is more feasible to accomplish it by jointly promoting its different aspects. Accordingly, we proposed a novel dialogue framework, COOPER, which coordinates multiple specialized agents, each dedicated to a specific dialogue goal aspect, to approach the complex objective. The empirical results on emotional support and persuasion dialogues demonstrated the effectiveness of our proposed approach.

## Acknowledgments

This work was supported by the Research Grants Council of Hong Kong (PolyU/5204018, PolyU/15207920, PolyU/15207122, PolyU/15213323) and National Natural Science Foundation of China (62076212). We also thank Wenjun Hou and Kaishuai Xu for their helpful comments.

## References

Austin, J. L. 1975. *How to do things with words*, volume 88. Oxford University Press.

Boyer, K.; Grafsgaard, J. F.; Ha, E. Y.; Phillips, R.; and Lester, J. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1190–1199.

Chen, M.; Shi, W.; Yan, F.; Hou, R.; Zhang, J.; Sahay, S.; and Yu, Z. 2022. Seamlessly integrating factual information and social content with persuasive dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 399–413.

Chen, M.; Yu, X.; Shi, W.; Awasthi, U.; and Yu, Z. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 951–966. Association for Computational Linguistics.

Cheng, Y.; Liu, W.; Li, W.; Wang, J.; Zhao, R.; Liu, B.; Liang, X.; and Zheng, Y. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3014–3026.

Deng, Y.; Lei, W.; Liao, L.; and Chua, T.-S. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Deng, Y.; Zhang, W.; Yuan, Y.; and Lam, W. 2023b. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4079–4095. Association for Computational Linguistics.

Fershtman, C. 1990. The importance of the agenda in bargaining. *Games and Economic Behavior*, 2(3): 224–238.

Fu, Y.; Peng, H.; Khot, T.; and Lapata, M. 2023. Improving language model negotiation with self-play and in-context learning from AI feedback. *arXiv preprint arXiv:2305.10142*.

Grice, H. 1975. Logic and conversation. *Foundations of Cognitive Psychology*, 719.

Hartigan, J. A.; and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108.

He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2333–2343.

Hill, C. E. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Joshi, R.; Balachandran, V.; Vashishth, S.; Black, A.; and Tsvetkov, Y. 2021. DialoGraph: Incorporating interpretable strategy-graph networks into negotiation dialogues. In *International Conference on Learning Representations*.

Lavie, A.; and Agarwal, A. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 228–231. Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; and Batra, D. 2017. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2443–2453. Association for Computational Linguistics.

Li, G.; Hammoud, H. A. A. K.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. Camel: Communicative agents for "mind"' exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Li, Y.; Qian, K.; Shi, W.; and Yu, Z. 2020. End-to-End trainable non-collaborative dialog system. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 8293–8302. AAAI Press.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021b. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483.

Liu, W.; Cheng, Y.; Wang, H.; Tang, J.; Liu, Y.; Zhao, R.; Li, W.; Zheng, Y.; and Liang, X. 2022. "My nose is running." "Are you also coughing?": Building A Medical Diagnosis Agent with Interpretable Inquiry Logics. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 4266–4272.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Peng, W.; Hu, Y.; Xing, L.; Xie, Y.; Sun, Y.; and Li, Y. 2022. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*.

Petty, R. E.; Cacioppo, J. T.; Petty, R. E.; and Cacioppo, J. T. 1986. The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, 19: 123–205.

Rieser, V.; and Moore, J. D. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 239–246.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Samad, A. M.; Mishra, K.; Firdaus, M.; and Ekbal, A. 2022. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL*, 844–856.

Sanders, A.; Strzalkowski, T.; Si, M.; Chang, A.; Dey, D.; Braasch, J.; and Wang, D. 2022. Towards a progression-aware autonomous dialogue agent. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1194–1212.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Shi, W.; Li, Y.; Sahay, S.; and Yu, Z. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP*, 3478–3492. Association for Computational Linguistics.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857–16867.

Tran, N.; Alikhani, M.; and Litman, D. 2022. How to ask for donations? Learning user-specific persuasive dialogue policies through online interactions. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 12–22.

Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.-R.; and Yan, R. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 308–319. Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5635–5649. Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wen, T.-H.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojas-Barahona, L. M.; Su, P.-H.; Ultes, S.; and Young, S. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Wu, Q.; Zhang, Y.; Li, Y.; and Yu, Z. 2021. Alternating Recurrent Dialog Model with Large-scale Pre-trained Language Models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1292–1301.

Xu, X.; Meng, X.; and Wang, Y. 2022. Poke: Prior knowledge enhanced emotional support conversation with latent variable. *arXiv preprint arXiv:2210.12640*.

Yang, R.; Chen, J.; and Narasimhan, K. 2021. Improving dialog systems for negotiation with personality modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 681–693. Association for Computational Linguistics.

Zhao, W.; Zhao, Y.; Lu, X.; Wang, S.; Tong, Y.; and Qin, B. 2023a. Is ChatGPT equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Zhao, W.; Zhao, Y.; Wang, S.; and Qin, B. 2023b. TransESC: Smoothing emotional support conversation via turn-Level state transition. In *Findings of the Association for Computational Linguistics: ACL*, 6725–6739. Association for Computational Linguistics.

Zheng, C.; Sabour, S.; Wen, J.; Zhang, Z.; and Huang, M. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1552–1568. Association for Computational Linguistics.

Zhou, J.; Chen, Z.; Wang, B.; and Huang, M. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1714–1729. Association for Computational Linguistics.

Zhou, Y.; He, H.; Black, A. W.; and Tsvetkov, Y. 2019a. A dynamic strategy coach for effective negotiation. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, 367–378. Association for Computational Linguistics.

Zhou, Y.; He, H.; Black, A. W.; and Tsvetkov, Y. 2019b. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 367–378.