



Universiteit  
Leiden

# Master Computer Science

Prompting a Smaller-Scale LLM for Goal-Oriented  
Emotional Support Dialogue Generation

Name: Joël Buter  
Student ID: 3373347  
Date: 12/04/2025  
Specialisation: Bioinformatics  
1st supervisor: Joost Broekens  
2nd supervisor: Suzan Verberne

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)  
Leiden University  
Niels Bohrweg 1  
2333 CA Leiden  
The Netherlands

## Abstract

Goal-oriented dialogues form an important aspect of human communication, in which speakers guide conversations towards specific outcomes. An important application is emotional support dialogues, where conversations aim to build barriers against psychological distress and improve well-being. These dialogues precede the provision of professional mental health care, which effectively provides distress relief but is constrained by workforce shortages and limited accessibility. Emotional support chatbots address these challenges by using large language models to provide support anywhere and at any time. Recent research has focused on training models for goal-oriented settings on synthetically generated dialogues to address the limited availability of high-quality data. These dialogues are produced by enabling large language models to simulate conversations between two speakers, in ways that align with real-world practices. However, these methods suffer from limited controllability and high computational demands. In this thesis, we show that a smaller-scale, open-source language model can produce synthetic emotional support dialogues comparable to human conversations entirely through prompting. We find that sufficiently constraining the model’s output allows it to provide support that aligns with established psychological frameworks. Furthermore, the results show that the model consistently outperforms humans in accurately implementing support skills. Overall, these results demonstrate that given sufficient guidance, a smaller-scale language model is capable of generating high-quality dialogues, meaning our approach can overcome limitations in existing methods.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Background</b>	<b>7</b>
2.1	Dialogue systems . . . . .	7
2.2	Self-chats . . . . .	7
2.3	In-context learning . . . . .	8
2.4	Prompting for structured reasoning . . . . .	8
2.5	Self-consistency . . . . .	8
2.6	Prompting . . . . .	9
<b>3</b>	<b>Related work</b>	<b>10</b>
3.1	Emotional chatbots . . . . .	10
3.1.1	Natural language understanding . . . . .	10
3.1.2	Dialogue planning . . . . .	11
3.1.3	Natural language generation . . . . .	12
3.2	Synthetic dialogue generation . . . . .	12
<b>4</b>	<b>Methodology</b>	<b>14</b>
4.1	Problem definitions . . . . .	14
4.2	Topic seeding . . . . .	15
4.3	Dialogue policy planning . . . . .	16
4.3.1	Strategy selection . . . . .	16
4.3.2	Strategy feature analysis . . . . .	18
4.3.3	Decision making . . . . .	21
4.3.4	Resolving strategy selection ties . . . . .	22
4.3.5	Support skill selection . . . . .	23
4.4	Response generation . . . . .	24
4.4.1	Coarse-grained context analysis . . . . .	25
4.4.2	Fine-grained context analysis . . . . .	25
4.4.3	Response formulation . . . . .	26
4.5	Conversation termination . . . . .	26
4.6	Self-chat . . . . .	27
<b>5</b>	<b>Experiments and results</b>	<b>29</b>
5.1	General characteristics . . . . .	29
5.2	Diversity and coherence analysis . . . . .	29
5.2.1	Lexical analysis . . . . .	29
5.2.2	Dialogue diversity . . . . .	30
5.2.3	Topic similarity . . . . .	31
5.2.4	Topic analysis . . . . .	31
5.3	Dialogue policy planning . . . . .	32
5.3.1	Conversational features . . . . .	32
5.3.2	Strategy selection in human dialogues . . . . .	34

5.3.3	General support skill statistics . . . . .	34
5.3.4	Support skill trends . . . . .	35
5.3.5	Human evaluation of dialogue planning . . . . .	36
5.4	Response generation . . . . .	38
5.4.1	Relevance and accuracy . . . . .	38
5.4.2	Conversational quality . . . . .	39
<b>6</b>	<b>Discussion</b>	<b>40</b>
6.1	General findings . . . . .	40
6.2	Implications . . . . .	42
6.3	Limitations . . . . .	42
<b>7</b>	<b>Conclusion</b>	<b>43</b>
<b>A</b>	<b>Response repetition</b>	<b>52</b>
<b>B</b>	<b>Qualitative LLM evaluation prompt</b>	<b>59</b>
<b>C</b>	<b>Differences in response generation between prompts</b>	<b>60</b>
<b>D</b>	<b>Human skill and response evaluation interface</b>	<b>61</b>

# 1 Introduction

Goal-oriented dialogues are a core aspect of human communication, in which speakers proactively take initiative to guide conversations towards desired outcomes (Deng et al., 2025). These interactions can range from simple situations, such as ordering coffee, to more complex settings like customer support or counseling. One relevant application of goal-oriented dialogue is in professional mental healthcare, where conversations are structured to help resolve psychological suffering in individuals. However, despite effectively relieving distress, access to care is often limited by workforce shortages or time constraints (Campion et al., 2022). Therefore, everyday emotional support from someone’s social circle forms an important alternative, since it can improve coping abilities, self-esteem, and subjective well-being (Liu et al., 2021a; Azpiazu et al., 2023), thus acting as a buffer against psychological distress (Burlison, 2003).

Emotional support chatbots aim to artificially provide such support, addressing gaps in accessibility by being remotely accessible and available at all times (Vaidyam et al., 2019; Schyff et al., 2023). The abilities of these chatbots range from empathetically responding to help-seekers’ messages (Rashkin et al., 2019; Xie et al., 2021; Sabour et al., 2022), to offering support, e.g., by providing advice or sharing information (Liu et al., 2021b; Zheng et al., 2023a; Zheng et al., 2024). Furthermore, recent studies have begun to investigate approaches to improve support chatbots’ capabilities to proactively guide conversations towards goals, such as distress reduction, by selecting context-dependent support skills like asking questions or reframing perspectives (Cheng et al., 2024; He et al., 2024b; Deng et al., 2024b).

Unlike task-specific domains (e.g., hotel reservations), open-domain dialogues like emotional support conversations can cover any topic. To behave proactively, the chatbot has to understand dialogue context rather than follow predefined templates or fixed patterns (Huang et al., 2020), making it a challenging problem.

To understand language and generate responses, modern support chatbots primarily depend on (large) language models, which generally perform significantly better on conversational tasks when trained on dialogue datasets (Liu et al., 2021b). However, goal-oriented dialogue datasets are scarce; instead, training data often consist of large volumes of unannotated conversations from social platforms like Reddit (Xie et al., 2021; Welivita et al., 2023) or small sets of human-annotated conversations (Rashkin et al., 2019; Liu et al., 2021b). While the latter often better align with real emotional support settings, their curation requires training of human annotators. This makes the curation process time consuming and expensive, which limits dataset size and diversity (Zheng et al., 2023a). For instance, Liu et al. (2021b) curated `ESCONV`, a dataset that consists of only 1,300 emotional support dialogues, but remains the largest of its kind due to the high costs associated with its curation. This relatively small size could in turn limit downstream model performance, which was indeed demonstrated by Zheng et al. (2023a) and Zheng et al. (2024).

Synthetic dialogue generation has become a promising approach to addressing these challenges. For instance, it can be used for upscaling human-annotated datasets with artificial samples (Zheng et al., 2023a; Zheng et al., 2024; Askari et al., 2024), or simulating dialogues to train models in conversational guidance (Zheng et al., 2023a; He et al., 2024b; Deng et al., 2024b). For instance, Zheng et al. (2024) used ChatGPT to implement a prompting method that augments a human-annotated conversational dataset to obtain a tenfold increase in samples. They subsequently showed that downstream large language models (LLMs) trained on their synthetic dataset were better at providing emotional support than those trained on the original dataset.

Several limitations remain despite these advances. Existing open-domain dialogue generation approaches lack mechanisms that align the conversational flow with evidence-based psychological frameworks defining effective support. Training on misaligned dialogues could lead to downstream systems providing ineffective emotional support and potentially harming the well-being of the help-seeker (Burleson, 2003). Furthermore, existing methods require larger-scale ( $\geq 100\text{B}$  parameters), closed-source LLMs (Xu et al., 2023; Zheng et al., 2024; Deng et al., 2024b; He et al., 2024b), which are computationally expensive and lack transparency on their biases (Deng et al., 2024a).

To address these limitations, we propose prompting a LLAMA 3-8B model to generate synthetic emotional support dialogues (Dubey et al., 2024). While such a smaller-scale ( $\leq 20\text{B}$  parameters), open-source LLM likely does not completely match the contextual understanding capabilities of bigger, closed-source models, it could be a more computationally efficient and transparent alternative.<sup>1</sup>

Our method focuses on prompt engineering for both generating responses and guiding conversations. By grounding response generation in emotional support skill descriptions proposed by Hill (2020), we aim to align the outputs with clinical best practices (Chen et al., 2023b; Deng et al., 2024b). Similarly, the conversational structure follows a psychological framework proposed by Liu et al. (2021b) to provide effective goal-oriented emotional support. This prompting approach can be adapted to other languages or domains by translating prompts or modifying them to align with different frameworks. Furthermore, alignment of the model’s outputs with guidelines can substantially reduce its biases (Li et al., 2024).

We integrated guidance into a self-chat process, where the model generates dialogues by simulating a conversation between a help-seeker and a supporter by role-playing as both parties (Deng et al., 2024b). The utility of the dialogues was improved by sampling conversation topics from a dataset that includes distress-related scenarios (Zheng et al., 2023a; Zheng et al., 2024). Based on this methodology, we aim to answer the following research question:

*How well can smaller-scale language models, when prompted to follow psychological frameworks, produce emotional support dialogues of human-level quality, comparable to those in ESCONV?*

We evaluated the quality of the dialogues using a mixed-method approach that combines human evaluations (e.g., response relevance, support skill appropriateness) and automatic evaluations (e.g., distinct- $n$  scores, PCA). The main research question is investigated in more detail by answering the following subquestions:

- RQ 1:* How can smaller-scale language models be prompted for context understanding to enable support skill selection that aligns with psychological frameworks for goal-oriented conversational guidance?
- RQ 2:* How well do smaller-scale language models generate contextually relevant responses that adhere to emotional support skill descriptions specified in their prompts?
- RQ 3:* To what extent does a smaller-scale language model produce dialogues comparable to human conversations when integrated into a self-chat framework that combines goal-oriented guidance with response generation?

---

<sup>1</sup><https://huggingface.co/meta-llama/Llama-3-8B>

## 2 Background

This section provides an overview of concepts important to the synthetic dialogue generation approach proposed in this thesis. First, Section 2.1 introduces artificial dialogue systems, after which Section 2.2 elaborates on the self-chat framework that allows LLMs to generate conversations without human supervision. Subsequently, Section 2.3 describes in-context learning as an approach for using LLMs to solve unseen problems without fine-tuning and Section 2.4 details chain-of-thought prompting and least-to-most prompting, two variants of in-context learning that help improve LLM reasoning. Finally, Section 2.5 discusses self-consistency, which is used to improve the performance of in-context learning.

### 2.1 Dialogue systems

Dialogue systems are artificial intelligence systems that communicate with humans, with early versions performing language generation and dialogue management through symbolic reasoning — processing language through manually defined, human-readable rules (Winoograd, 1971; Goddeau et al., 1996; McTear, 2002). These systems can be divided into three distinct, but related categories: open-domain, goal-oriented, and information-seeking systems (Deng et al., 2023a).

Open-domain dialogue systems can engage in conversations about any topic, whereas goal-oriented dialogues systems focus on accomplishing domain-specific tasks like travel planning or restaurant booking. Finally, information-seeking systems aim to satisfy a user’s requests for information.

The current study focuses on a hybrid dialogue system: goal-oriented *chatbots* capable of engaging in open-domain conversations. Specifically, for a predefined objective, such as relieving the user’s distress, a language model is used to output responses that guide the conversation to obtain this objective. The chatbot achieves this by relying on components required for language understanding, dialogue state tracking, policy learning (the latter two may be grouped under dialogue management, see Section 3.1.2), and response generation (Chen et al., 2017; Bilquise et al., 2022). In modern dialogue systems, these components depend on (L)LMs, due to their abilities to understand context and generate relevant responses. An important challenge in training these models for conversational tasks is the scarcity of task-specific, high-quality training data; human-annotated datasets are generally limited in size due to their high collection costs (Zheng et al., 2023a), a problem that can be partially addressed by letting language models simulate conversations between two speakers (*self-chats*) to generate artificial training data.

### 2.2 Self-chats

The *self-chat* methodology allows language models to generate synthetic dialogues by role-playing as dialogue agents that interact in a conversational manner (Chen et al., 2024). This is especially useful in domains with limited conversational training data or availability of human supervision; the model can create high-quality training samples on many subjects by using a small set of conversation seeds. Downstream models, or even the generating model itself, can subsequently be trained on this data to improve their capabilities on specific tasks (Xu et al., 2023). Methods like least-to-most prompting and self-consistency can improve self-chat frameworks by enabling dialogue agents to guide conversations and enhance their response quality without requiring extensive task-specific training data (see Sections 2.3, 2.4 and 2.5 for details). These methods are particularly relevant when simulating emotional support conversations, where multiple valid conversational directions exist for complex or

ambiguous dialogue scenarios. Moreover, they allow the agents to generate more nuanced and contextually relevant responses, leading to more diverse and high-quality dialogues.

### 2.3 In-context learning

*In-context learning* allows LLMs to perform new language processing tasks without additional training (Brown et al., 2020). Prompting the model with written task instructions, either with (few-shot prompting) or without task-specific examples (zero-shot prompting), lets models learn through analogy; prior knowledge obtained during training is used to generalise information provided in the model’s prompt to new situations (Dong et al., 2024). Given a sufficiently large model and effective prompts, in-context learning allows LLMs to achieve performance comparable to fine-tuned versions on tasks involving logic and reasoning (Touvron et al., 2023). This is especially useful in low-resource contexts, where enough training data for fine-tuning is often not available. Sections 2.4 and 2.5 elaborate on methods that have been proposed to improve the in-context learning abilities of LLMs.

### 2.4 Prompting for structured reasoning

*Chain-of-thought* (CoT) prompting is an important variant of in-context learning. It improves the solutions generated by an LLM by instructing it to mimic human thinking (Wei et al., 2022). Before giving its final answer, the model responds with a step-by-step process to address the problem before outputting a solution. While this method often improves performance on logical and commonsense reasoning tasks, the effectiveness of CoT prompting is primarily limited by the size of the model, with smaller models often generating invalid reasoning steps.

This limitation is addressed by *least-to-most* (LtM) prompting. Similar to CoT prompting, it makes LLMs use structured reasoning to solve complex problems. The approach consists of two stages: (1) breaking down the original problem into easier subproblems, and (2) solving these subproblems in any order using earlier results. The solutions obtained during the second stage are then used to solve the original problem (Zhou et al., 2023). These subproblems can be addressed with any prompting method, e.g., CoT prompting could be used within LtM prompting. Furthermore, the decomposition of the original problem into subproblems during the first stage can be performed either by the model itself or by human experts. Since the performance of CoT prompting can suffer from the complexity of generating reasoning steps, leaving this to humans allows the model to focus on only solving simpler subtasks, which often results in improved performance on complex problems.

### 2.5 Self-consistency

*Self-consistency* improves the accuracy of LLM responses to questions by generating multiple outputs from the same prompt and aggregating their results (Wang et al., 2023b). It performs well when combined with LtM prompting; generating diverse reasoning paths using the model’s inherent variability allows for selecting the most consistent result, which is intuitively more likely to be correct. At the same time, it reduces the impact of random errors that might occur in individual outputs. Furthermore, self-consistency improves the reliability of outputs in scenarios where multiple answers are correct, but only one is optimal (Deng et al., 2024b). These features make it a strong tool for our use-case: handling ambiguous tasks such as emotion understanding with smaller-scale LLMs.



## 2.6 Prompting

Lastly, we distinguish between the two main prompting approaches used in this thesis. The *response prompting* approach lets the model generate unconstrained responses (as the “assistant”) to requests from the “user”. While straightforward to use, this approach naturally leads to a degree of variability in the model’s output that can be problematic when this needs to be parsed or follow a certain format. Therefore, *completion prompting* constrains the output by letting the model finish an already partially structured assistant response. Our preliminary experiments showed this approach is especially useful for tasks requiring specifically formatted responses, or when using smaller-scale models with limited abilities to understand instructions.

The difference between the two prompting methods is demonstrated using a task where the model role-plays as a supporter to generate a continuation to an emotional support dialogue:

**User:**  
The following is a conversation between a supporter and a help-seeker:  
Help-seeker: “I’m just feeling so terrible right now...” (1)  
The supporter should respond by asking the help-seeker to elaborate on their situation.

Although we might prefer only an empathetic response from the supporter, LLMs will instead often generate a contextually correct, but incorrectly formatted output like this:

**Assistant:**  
The supporter could ask the help-seeker what is making them feel this way. (2)  
Supporter: What is making you feel that way?

This makes it difficult to reliably process the model’s output, potentially causing problems when using techniques such as LtM prompting or self-consistency. Luckily, this problem can be addressed by extending Prompt 1 to include a completion prompt, where the model is tasked to complete a partial output. Here, we guide the model by including the beginning of an enquoted response:

**Assistant (completing):**  
Supporter: “**What is making you feel that way?**” (3)

The unbolded text represents the our partially defined response, whereas the bolded text shows the model’s completion of it. By including the structured output in the prompt and letting the model complete it, its outputs become more predictable and reliable compared to letting it generate unconstrained responses. Completion prompting is therefore useful for structured generation tasks, especially when using smaller-scale models, which generally have limited instruction following abilities and thus generate incorrect outputs more often.

### 3 Related work

The current section provides an overview of literature relevant to creating emotional support chatbots. The core components of chatbots are reviewed in Section 3.1: language understanding, dialogue planning, and text generation are discussed in Sections 3.1.1, 3.1.2, and 3.1.3, respectively. Section 3.2 explains how language models can be used for synthetic dialogue generation to address data scarcity challenges.

#### 3.1 Emotional chatbots

ELIZA was the first chatbot able to engage in realistic conversations with its users by using pattern matching and substitution rules to create its responses (Weizenbaum, 1966), which popularized the use of rule-based approaches to natural language processing (Woods, 1970; Shortliffe, 1974; Carbonell Jr, 1978). The introduction of algorithms using statistical methods, such as *hidden Markov models* and *n-gram models* (Brown et al., 1992; Merialdo, 1994), followed by neural network-based approaches like *LSTMs* and *GRUs* removed the need for hand-coded rules to understand context (Hochreiter et al., 1997; Cho et al., 2014). Modern chatbots rely on language models called *transformers* for language understanding and generation, which are neural networks that excel at interpreting patterns in sequential data through an *attention mechanism* (Vaswani, 2017).

An important advancement in conversational AI was the development of BlenderBot by Roller et al. (2021). It was able to show conversational abilities like empathy and knowledge by depending on multiple fine-tuned transformer models. Ouyang et al. (2022) then popularised a method for aligning outputs of language models with human preferences, allowing for more natural interactions between chatbots and humans (Christiano et al., 2017).

Bilquise et al. (2022) argued that chatbots generally need three components to understand and respond to dialogue contexts: (1) natural language understanding, which underlies the next two components and is explained in Section 3.1.1, (2) dialogue planning, which is covered in Section 3.1.2, and (3) natural language generation, which is detailed in Section 3.1.3. These components can be implemented in an end-to-end manner or separately to give a chatbot convincing conversational abilities.

##### 3.1.1 Natural language understanding

Natural language understanding (NLU) plays an important role in creating chatbots for domain-specific contexts such as distress support conversations (Bilquise et al., 2022). Allowing systems to better interpret help-seeker intents, emotions, and needs is a necessity for providing relevant and empathetic responses, and effective conversational guidance (Liu et al., 2021b; Cheng et al., 2024). Therefore, recent literature has explored a number of approaches to improving domain-aware NLU. The majority of these focused on the usage of pre-trained (large) language models (Zhang et al., 2024).

For instance, Demszky et al. (2020) introduced a two-turn conversational dataset with emotion annotations to fine-tune a *BERT-model* for emotion understanding (Devlin et al., 2019).<sup>2</sup> Likewise, Sharma et al. (2020) improved the empathy understanding of a language model that was pre-trained on a mental health support dataset by fine-tuning it on empathy-annotated data. Ji et al. (2022) proposed to pre-train a BERT-model on Reddit posts revolving around mental health, with this model used by Ding et al. (2022) for cognitive disorder classification. Burkhardt et al. (2022) showed that emotion features extracted from the

<sup>2</sup>BERT-models were ground-breaking due to their ability to understand the semantic context of words in a sentence.

embeddings of a fine-tuned BERT-model can be used to identify symptoms of depression and anxiety.

However, these methods all require model retraining or new datasets when problem requirements change, which poses challenges in low-resource domains like mental health care (Zhang et al., 2021). Promising zero-shot and few-shot learning approaches have been developed to address these challenges. Rastogi et al. (2020) introduced a method for intent classification and contextual information extraction by using multiple BERT-models that semantically compare user utterances with natural language descriptions. Balaraman et al. (2021) improved this process by developing a single BERT-model capable of both identifying user intents and extracting context. More recently, research has focused on using instruction-tuned LLMs, hereafter referred to simply as LLMs. Sun et al. (2023b) used two LLMs that collaborate to analyse sentiment in a dialogue, while Broekens et al. (2023) demonstrated that sufficiently large LLMs can understand affective content through zero-shot prompting. Similarly, Fei et al. (2023) used zero-shot least-to-most prompting to guide an LLM through a process similar to human reasoning, with Wang et al. (2023a) following a similar approach to guide a model through intermediate reasoning steps to better interpret linguistic cues. Deng et al. (2023b) showed that in proactive dialogue tasks adding an additional goal planning step to this process further improved language understanding.

Overall, while fine-tuned models often outperform zero-shot and few-shot approaches on language understanding problems, their need for large amounts of training data can make them unsuitable for use in low-resource domains (Wang et al., 2024). Therefore, zero-shot and few-shot prompting can be more flexible solutions that allow adaptation to new tasks without (re)training.

### 3.1.2 Dialogue planning

Due to their flexible generative abilities and strong contextual understanding, chatbots often rely on LLMs to engage convincingly in conversations. However, an open problem for LLMs is generating proactive responses that guide conversations towards a predefined goal like resolving distress during open-domain dialogues, as a result of their training process (Deng et al., 2023a).

Therefore, various approaches to improving LLM capabilities in proactive dialogues have been investigated. An important framework is *dialogue management* (or *dialogue planning*), which selects actions based on dialogue context that let the chatbot achieve its goals (Chen et al., 2017; Deng et al., 2024b). For instance, goal-oriented dialogue planning can intuitively be formulated as a Markov decision process that can be solved using reinforcement learning. In an open-domain emotional support context, support skills can function as actions, the dialogue context as the environment state (updated by role-playing language model agents), and changes in the help-seeker’s distress are captured by rewards (Liu et al., 2020; Tiwari et al., 2022; Deng et al., 2024b). After training, the LLM’s responses can be grounded in actions selected using the learned policy. However, making reward functions that accurately capture conversational dynamics is challenging (Deng et al., 2024b), and role-playing agents typically lack emotional complexity, leading to unrealistic dialogues (He et al., 2024b).

An alternative approach is combining LLM-based dialogue simulations with tree search algorithms (Yu et al., 2023; V  th et al., 2023; He et al., 2024b). By simulating dialogue trajectories, actions can be selected by estimating their long term impact on achieving a goal. However, this method is impractical for real-time systems due to its need for many expensive dialogue simulations to select an action (He et al., 2024b).

Similar to the approaches discussed in Section 3.1.1, in-context learning has been applied to tasks that require dialogue planning (Wang et al., 2023a; Zhang et al., 2023; Deng et al., 2023b; Cheng et al., 2024). Through prompting LLMs to reason about conversational progress towards dialogue goals, the models’ outputs can be used to guide conversations either directly or indirectly. However, while researchers have demonstrated that in-context learning for dialogue planning is promising, it depends on the strong reasoning capabilities of larger-scale LLMs, which can pose challenges for smaller models (Wei et al., 2022; Zhou et al., 2023).

### 3.1.3 Natural language generation

To end-users, the most tangible aspect of a chatbot are the responses it generates, the quality of which mainly depends on expressing relevant emotions and addressing the user’s feelings (Liu et al., 2021b).

Recent literature has shown that fine-tuning language models on annotated empathetic or emotional conversations improved their ability to produce empathic responses. Moreover, extracting commonsense knowledge like the help-seeker’s intents and needs helped to generate contextually relevant responses (Rashkin et al., 2019; Sabour et al., 2022; Tu et al., 2022; Cheng et al., 2023). Further improvements were achieved by pre-training models on large, unstructured, domain-related datasets before fine-tuning on smaller, annotated, task-specific datasets (Xie et al., 2021; Ji et al., 2022; Welivita et al., 2023). These findings show the importance of both the quantity and quality of training data.

However, the proposed approaches and datasets lacked the necessary structure to more systematically address help-seeker needs (Burlison, 2003; Hill, 2020; Liu et al., 2021b). Liu et al. (2021b) addressed this by creating `ESCONV` through a crowdsourcing task. This dataset contains 1.3K conversations that follow the Helping Skills framework proposed by Hill (2020) and are annotated with support skill labels. Language models fine-tuned on `ESCONV` showed big improvements in their abilities to provide structured emotional support and resolving distress, which highlighted the benefits of including support skills in the training process. Zheng et al. (2023a) improved `ESCONV` by using GPT-J 6B to generate `AUGESC`, which addressed the limited size and topic coverage of `ESCONV`, but lacked support skill annotations.<sup>3</sup> Zheng et al. (2024) subsequently used ChatGPT to generate `EXTES`, which contains support skill-annotated dialogues that cover diverse topics.<sup>4</sup> Models trained on `EXTES` were better at emotionally supporting help-seekers than those trained on `ESCONV` and `AUGESC`, demonstrating the promise of synthetic datasets. Section 3.2 further elaborates on the methodologies used to create both `AUGESC` and `EXTES`.

However, challenges similar to those discussed in Section 3.1.1 still remain: (1) annotated data is limited, (2) when task requirements change, new or more data and model re-training are necessary, and (3) pre-training depends on the availability of massive datasets from related domains. Chen et al. (2023b) demonstrated that these issues can be addressed by grounding LLMs’ responses on support skills through prompting. Related work from Askari et al. (2024), Deng et al. (2024b) and He et al. (2024b) further showed the potential of prompting to address limitations resulting from a lack of high-quality training data.

## 3.2 Synthetic dialogue generation

From Sections 3.1.1, 3.1.2, and 3.1.3 it becomes clear that the effectiveness of language understanding, dialogue planning, and response generation depends on the knowledge that is

<sup>3</sup>[https://huggingface.co/docs/transformers/model\\_doc/gptj](https://huggingface.co/docs/transformers/model_doc/gptj)

<sup>4</sup><https://openai.com/index/chatgpt>

present within the model powering them. Currently, fine-tuning language models on large, annotated datasets remains the best way to embed such information (Rashkin et al., 2019; Zhang et al., 2021; Liu et al., 2021b; Zheng et al., 2023a; Zheng et al., 2024).

However, the emotional support domain faces significant data scarcity challenges. Popular support chatbots like Woebot<sup>5</sup>, Replika<sup>6</sup>, and Wysa<sup>7</sup> are proprietary systems with private datasets. Moreover, researchers investigating real-world therapy conversations do not share these because of ethics- or privacy-related reasons (Althoff et al., 2016; Zhang et al., 2020).

Synthetic dialogue generation has therefore become a promising alternative for creating high-quality data (Wang et al., 2023c; Zheng et al., 2023b; Tan et al., 2024; Yuan et al., 2024; Chen et al., 2024; Liang et al., 2025). Recent advancements have mainly focused on *knowledge distillation*, which is a method used to transfer abilities from complex systems to simpler systems to address downstream tasks more efficiently and effectively (Xu et al., 2023; Chen et al., 2023a; Askari et al., 2024; Deng et al., 2024b).

For instance, Askari et al. (2024) improved intent prediction by using ZEPHYR-7B- $\beta$  to generate dialogues grounded on human-annotated intent sequences, demonstrating the ability of smaller LLMs in generating high-quality synthetic data.<sup>8</sup> Deng et al. (2024b) followed a similar methodology, but instead combined a reinforcement learning-based intent selection algorithm with ChatGPT to generate dialogues. Zheng et al. (2023a) fine-tuned GPT-J 6B on ESCONV samples and prompted it to complete dialogues from an open-domain dataset comprised of empathetic conversations. Zheng et al. (2024) prompted ChatGPT with in-context example dialogues to generate conversations about any topic and annotate utterances with predefined support skills.

However, while these approaches produce high-quality data, they either face limitations in controlling model output or lack computational efficiency (Zheng et al., 2024; He et al., 2024b). To address these limitations, we propose a framework for combining the components described in Sections 3.1.1, 3.1.2, and 3.1.3. Following previous work from Chen et al. (2023b), Wang et al. (2023a), Askari et al. (2024) and Deng et al. (2024b), our approach relies only on prompting a smaller-scale LLM to balance flexibility, interpretability, and computational efficiency. To our knowledge, no studies have investigated such an approach.

---

<sup>5</sup><https://woebothealth.com>

<sup>6</sup><https://replika.com>

<sup>7</sup><https://wysa.com>

<sup>8</sup><https://huggingface.co/HuggingFaceH4/zephyr-7b-beta>

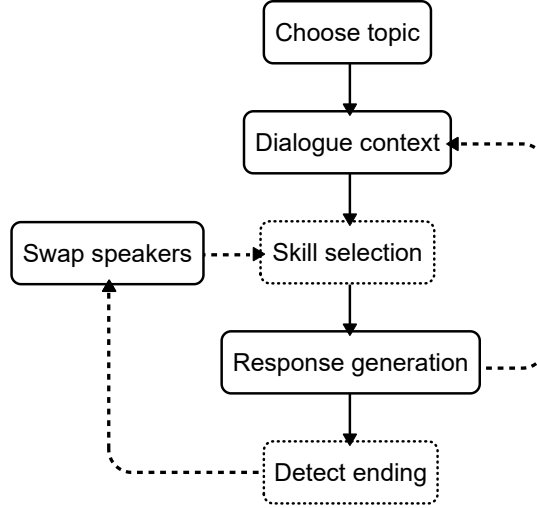


FIGURE 1: Illustration of the proposed methodology for generating a dialogue. A conversational topic initializes the dialogue context with the help-seeker as first speaker. When the speaker is a supporter, a support skill is selected based on the dialogue context. After a response is generated, speaker roles are swapped, and the process continues until an ending is detected. Solid arrows indicate the main steps of the generation process. Dashed arrows represent updates of the dialogue context and speaker roles. Dotted arrows stand for optional steps like skill selection (only for the supporter) or conversation termination (only for the help-seeker).

## 4 Methodology

Generating realistic dialogues relies on dialogue planning and language generation, with their effectiveness depending on strong contextual understanding. Seeding dialogues with conversational topics and allowing the conversation to end naturally further improves realism. This section elaborates on the methodology used to combine these components into a self-chat framework.

First, we formalize the definition of dialogue generation in Section 4.1. The proposed methodology is subsequently broken down into five sections: Section 4.2 focuses on seeding and initiating conversations, Section 4.3 elaborates on dialogue planning, Section 4.4 describes response generation, and Section 4.5 outlines a method for naturally concluding conversations. Finally, Section 4.6 provides an approach for combining them all into a self-chat framework. An illustration of the proposed methodology is shown in Figure 1.

### 4.1 Problem definitions

The methodologies described in Sections 4.2, 4.3, 4.4, and 4.5 were integrated into the self-chat framework detailed in Section 4.6 to generate a dataset consisting of  $n$  conversations:

$$d_{\text{ours}} = \{c_1, c_2, \dots, c_n\}, \quad (1)$$

where each conversation  $c_i \in d_{\text{ours}}$  consists of an ordered sequence of  $m$  utterances, and each utterance  $u_j$  corresponds to a speaker  $s_j$  and a support skill  $a_j \in \mathcal{A}$ :

$$c_i = (\langle u_j, s_j, a_j \rangle)_{j=1}^m. \quad (2)$$

Furthermore, each conversation revolves around a distress-related topic  $t$ . The speaker roles were derived from Liu et al. (2021b) and are defined as follows:

$$s_j \in \{\text{supporter, seeker}\}. \quad (3)$$

Likewise, the set of support skills available to the supporter was adapted from those proposed in Liu et al. (2021b):

$$\mathcal{A} = \{\text{QS, RP, RF, AR, PS, IN}\}. \quad (4)$$

The abbreviations correspond to the following support skills: Asking questions (*QU*), Restating or paraphrasing (*RP*), Reflecting on feelings (*RF*), Affirming and reassuring (*AR*), Providing suggestions (*PS*), and Providing information (*IN*). Two additional support skills proposed by Liu et al. (2021b) were excluded as their implementations were ambiguous. The action  $\emptyset$  is exclusive to the help-seeker, denoting the absence of support skills in their utterances:

$$a_j = \begin{cases} a \in \mathcal{A}, & \text{if } s_j = \text{supporter}, \\ \emptyset, & \text{if } s_j = \text{seeker}. \end{cases} \quad (5)$$

Finally, we define the isolation of a conversational context  $c_i$ ; the notation  $c_i^{(j)}$  represents all dialogue turns up to and including turn  $j$  within the conversational context  $c_i$ :

$$c_i^{(j)} = (\langle u_k, s_k, a_k \rangle)_{k=1}^j. \quad (6)$$

We use the notation  $c_i^{(j)}$  for referencing incomplete conversational contexts to provide clearer explanations of updating and analyzing dialogue segments in subsequent sections.

## 4.2 Topic seeding

There are various ways to seed conversations, usually using human-written examples for in-context learning (Chen et al., 2023a; Zheng et al., 2024), or by sampling topics from an external source (Zheng et al., 2023a; Askari et al., 2024). An example or topic  $t$  is substituted into a prompt instructing the model how it should be used to generate a conversation, i.e., given  $t$  sampled from a dataset, a function  $F(t)$  transforms  $t$  into a prompt for starting conversations:

$$p = F(t). \quad (7)$$

Given the generated prompt  $p$ , the model is instructed to generate the help-seeker’s first utterance of the conversation, and the conversational context  $c_i$  is initialized (Askari et al., 2024):

$$u_1 = \text{LLM}(p), \quad c_i^{(1)} = \{\langle u_1, s_1, a_1 \rangle\}, \quad (8)$$

with  $s_1 = \text{seeker}$  and  $a_1 = \emptyset$ . For our purposes, seeding conversations with topics follows a relatively straightforward process. Following Chen et al. (2023a), the model is prompted with the opening-phrase:

**System:**  
The following is a conversation between a supporter and a help-seeker, about  
“{topic\_seed}”.

(4)



, where “{topic\_seed}” is a description  $t$  of the help-seeker’s issue. Descriptions were sampled from the EXTES dataset (Zheng et al., 2024), which contains support dialogues annotated with short summaries of their contents, for instance:

“I received a text from my significant other saying that they need some space. I’m feeling confused and insecure about our relationship.” (5)

Using a one-shot prompt, the opening-phrase (Prompt 4) and the topic description are combined into the final conversation description:

The following is a conversation between a supporter and a help-seeker, about the help-seeker feeling confused and insecure about their relationship following a text from their significant other saying that they need some space. (6)

We note that a number of topic seeds in EXTES are phrased in the third person (e.g., “*The user is anxious about exams.*”), resulting in the conversation descriptions occasionally involving a third speaker: the “user”. An additional prompt was used to automatically identify and ignore these third-person descriptions.

### 4.3 Dialogue policy planning

Emotional support conversations require chatbots to balance responding to the help-seeker’s emotions with guiding the conversation towards practical outcomes (Burleson, 2003). However, LLMs are primarily trained for reactive responses and generally not optimized for long-term conversational planning (He et al., 2024b). These limitations often make them unsuitable for proactively guiding goal-oriented conversations in open-domain settings (Cheng et al., 2024). In contrast, they have demonstrated strong performance in contextually understanding conversations, which task-specific algorithms can use to guide subsequent chatbot actions, a process called *dialogue policy planning* (Deng et al., 2024b; Cheng et al., 2024). Our method therefore uses an LLM for (1) analyzing conversational context, and (2) mapping these analyses to actions.

To allow the model to take proactive actions while role-playing as the supporter, we organize the conversational flow into phases that align with an emotional support framework. Each phase is represented by broad strategies that encompass more nuanced support skills. The remainder of this section elaborates on why this enables effective support skill selection, how strategies are predicted, and finally how support skills are chosen.

#### 4.3.1 Strategy selection

The complexity of selecting support skills results from the significant overlap and nuance in the conditions for applying them in an emotional support setting. We address this problem by making three simplifying assumptions, based on psychological principles:

1. Similar support skills can be grouped under overarching *strategies* ( $\mathcal{S}$ ) that each capture well-defined and distinct conversational phases (Hill, 2020).
2. There exists a many-to-one mapping between support skills and strategies, which follows from the crowd-sourcing task detailed in Liu et al. (2021b).
3. While a diverse set of support skills is necessary for addressing emotional needs, their overarching strategies define the progression of conversations (Burleson, 2003; Hill, 2020).



These assumptions allow us to formulate support skill prediction as a hierarchical process, which reduces problem complexity by limiting the reasoning scope at the support skill selection stage; any support skill that does not belong to a selected strategy does not need to be considered. Specifically, we propose two stages: (1) the selection of an overarching strategy, detailed in Sections 4.3.1 – 4.3.4, and (2) the selection of an appropriate support skill within the chosen strategy, detailed in Section 4.3.5.

We denote the set of strategies  $\mathcal{S}$  as:

$$\mathcal{S} = \{\text{exploration, comforting, action}\}, \quad (9)$$

consistent with the emotional support stages proposed by Liu et al. (2021b):

1. The **exploration** stage: The supporter first *explores* the help-seeker’s emotional state and topic of distress to be able to provide constructive advice.
2. The **comforting** stage: The supporter provides *comfort* during the exploration process to relieve the help-seeker’s feelings of distress.
3. The **action** stage: When the help-seeker’s problems are understood, *actionable* advice can be given to help address the help-seeker’s distress.

Based on these definitions, we can formalize strategy selection. First, the function  $F_s$  creates a series of prompts based on a conversation  $c_i^{(j)}$ . These prompts guide the model in understanding context while it processes the conversation to produce a textual analysis  $t_s$ :

$$t_s = \text{LLM}(p_s), \quad p_s = F_s(c_i^{(j)}). \quad (10)$$

Because preliminary experiments showed that the model often failed to select an effective strategy from  $\mathcal{S}$  (even when manually provided with a strong textual analysis  $t_s$ ), we define  $M_s$  to prompt the model to convert  $t_s$  into a structured format:

$$v_s = \text{LLM}(p_v), \quad p_v = M_s(t_s). \quad (11)$$

For example, the model can select a short description that best aligns with an analysis of the help-seeker’s emotional state (Deng et al., 2024b):

**User:**

Following is an analysis of the help-seeker’s emotional state: “{analysis}”

What is the emotional state of the help-seeker? You can choose the following options:

1. The help-seeker is relatively calm. 2. The help-seeker is experiencing some anxiety. 3. The help-seeker is experiencing high emotional distress. (7)

**Assistant (completing):**

The help-seeker **is relatively calm.**

An example of a continuation generated by the model is formatted in bold text. This structured output can be mapped to a scalar that is used by a decision tree to more reliably select the most appropriate strategy  $r_{j+1}$ .<sup>9</sup>

$$r_{j+1} = S_s(v_s, \mathcal{S}). \quad (12)$$

<sup>9</sup>Combining the strong contextual understanding of language models with the transparency of symbolic reasoning is a form of *neuro-symbolic AI*.

Recent studies found that general-purpose LLMs struggle with domain-specific tasks that require logic, such as reasoning or classification (Sun et al., 2023a; Wu et al., 2024), making it difficult to prompt them to choose optimal strategies directly. However, our preliminary experiments demonstrated that they can be prompted to effectively characterize qualitative features in conversational contexts, consistent with results from Deng et al. (2024b). Given these limitations in reasoning but strengths in characterizing features, we propose to prompt the model to analyze conversational features that guide strategy selection rather than select strategies directly.

We thus derive two conversational features based on the earlier definitions of the exploration, comforting, and action strategies: (1) the extent to which the supporter has explored the help-seeker’s emotional state and topic of distress (feature  $f_1$ ), and (2) the help-seeker’s expressed level of distress (feature  $f_2$ ). We note that usage of the action strategy primarily depends on the exploration strategy: it should only be used after there has been enough exploration. It was therefore decided not to define an additional feature to assess readiness for receiving, e.g., advice.

#### 4.3.2 Strategy feature analysis

To enable LLM-based feature analysis, we used the *least-to-most prompting* methodology proposed by Zhou et al. (2023): the model solves a series of easier subtasks and combines their intermediate results to address more complex problems. This aligns with the definition of our prompt generation function  $F_s$ ; context analysis can be decomposed into subprompts that guide the model. Specifically, we define a function  $F_{f_i}$  for each feature  $f_i$  that creates prompts from context  $c_i^{(j)}$  to guide the model in analyzing the respective feature, resulting in a textual analysis  $t_{f_i}$  output by the model:

$$t_{f_i} = \text{LLM}(p_{f_i}), \quad p_{f_i} = F_{f_i}(c_i^{(j)}). \quad (13)$$

Clearly, the quality of the feature analyses depends on the model’s contextual understanding of the help-seeker’s and supporter’s statements. Therefore, subprompts include the conversational history in their respective system prompts when necessary. Similar to Section 4.2, the conversational history is included as follows:

**System:**

The following is a conversation between a supporter and a help-seeker, about “{topic\_seed}”:

Help-seeker: “Hey, I’ve been feeling really overwhelmed lately.”

Supporter: “Can you tell me more about what’s going on?”

...

(8)

Additionally, to focus the model’s attention on the active topic of discussion, instead of the entire conversation, we explicitly reference the help-seeker’s most recent message and the supporter’s corresponding message. We note that due to the asynchronous nature of chat conversations, the supporter’s most recent message is not necessarily the message the help-seeker responded to. The model is therefore guided to identify the supporter’s relevant message by making it continue a *completion prompt*:

**Assistant (completing):**

Based on the conversational history and on the help-seeker's last message  
 "{seeker\_message}" in response to the supporter's message "**Maybe we could explore  
 writing a heart-felt letter or having a calm conversation.**" (9)

After processing the model's output from Prompt 9, the help-seeker's and supporter's messages can be included in a completion prompt to focus the model's attention on the relevant context when necessary:

**Assistant (completing):**

Based on the conversation history and on the help-seeker's last message  
 "{seeker\_message}" in response to the supporter's message "{supporter\_message}",  
*[model completes]* (10)

This improves the quality of the analyses compared to assuming that the last message from the supporter is the one being responded to, which we found introduces conversational inconsistencies. The italicized text "*[model completes]*" represents any possible continuation generated by the model.

Building on these prompts that improve context-awareness, we now elaborate on the two conversational features proposed in Section 4.3.1.

**Feature  $f_1$ : Exploration extent**

The assessment of the *exploration extent* depends on a sequence of prompts to determine the depth of exploration into the topic of distress. The first task entails identifying the current topic of distress:

**User:**

What is the current topic of distress? Answer in a paragraph. (11)

The model's attention is focused on the active topic of discussion using Prompt 10, which it completes to identify the distress topic. It is important to note that leading the model to answer with a single paragraph, instead of its otherwise very lengthy answers, results in fewer reasoning errors. This is consistent with findings from Askari et al. (2024).

The predicted topic is used to sequentially evaluate two distress-related aspects: (1) whether the help-seeker has provided enough information on the given topic, and (2) if the supporter has sufficient understanding of the help-seeker's distress to provide actionable advice:

1. **Help-seeker's elaboration:** The distress topic predicted through Prompt 11 is used to guide the model in assessing the help-seeker's elaboration on said topic:

**User:**

The topic of discussion is "{distress\_topic}". Has the help-seeker elaborated  
 on causes of distress regarding the topic of discussion? Answer in a paragraph. (12)

2. **Supporter's understanding:** Prompt 12 and its output are chained to Prompt 13, which allows the model to assess the supporter's understanding of the help-seeker's distress. Specifically, the information shared by the help-seeker should be sufficient for the supporter to start addressing the topic of distress:

**User:**

Does the supporter have enough understanding of the causes of distress to start resolving the help-seeker’s distress? Answer in a paragraph. (13)

We refer the reader to Prompt 16 to see the complete version.

However, when using Prompts 12 and 13, we found that the model often responds with “Yes, because...” or “No, because...”, and constructs its subsequent output justifying this initial, often incorrect decision. Instead, it should provide an unbiased analysis of the conversation, which was achieved by prefixing the model’s output for both the help-seeker and supporter “perspectives” as follows:

**Assistant (completing):**

The help-seeker [model completes] (14)

**Assistant (completing):**

Based on the conversation, the help-seeker [model completes] (15)

An additional problem with these prompts is the model’s tendency to always prioritize exploration. This results in too much probing into the help-seeker’s emotional state, which hinders the main goal of emotional support conversations: reducing distress (Burlison, 2003). Following Deng et al. (2023b), we addressed this problem by explicitly defining the conversation’s objective using a system prompt that precedes, e.g., Prompts 12 and 14. The complete prompt combining all components is then structured as follows:

**System:**

The goal of the supporter is to reduce the help-seeker’s distress through providing guidance.

**User:**

The topic of discussion is “{distress\_topic}”. Has the help-seeker elaborated on causes of distress regarding the topic of discussion? Answer in a paragraph.

**Assistant (completing):**

The help-seeker [model completes] (16)

**User:**

Does the supporter have enough understanding of the causes of distress to start resolving the help-seeker’s distress? Answer in a paragraph.

**Assistant (completing):**

Based on the conversation, the help-seeker [model completes]

The combined analyses generated by the model reflect whether sufficient exploration of the help-seeker’s distress has occurred.

**Feature  $f_2$ : Distress level**

The help-seeker’s *distress level* was determined by analyzing the tone in their most recent message, as well as throughout the conversation. We found that this resulted in a better analysis of the help-seeker’s distress compared to directly prompting the model to analyze their distress level.

We first prompted the model to assess the help-seeker’s emotional state, resulting in descriptions of, e.g., their stress or anxiety:

**User:**  
What is the emotional state of the help-seeker? Answer in a paragraph. (17)

To guide the analysis of the help-seeker’s tone and focus the model’s attention on the active topic of discussion, an additional completion prompt is combined with Prompt 10:

**Assistant (completing):**  
, the help-seeker’s tone [*model completes*] (18)

Directly prompting the model for tone analysis with a user prompt often lead to it ignoring its instructions. However, the quality of the analyses improved substantially after updating the instructions to include Prompt 18, which was very important for effective strategy selection.

### 4.3.3 Decision making

The decision making process involved translating the qualitative feature analyses into quantitative scores that allow for strategy selection. Its structure deviates slightly from the selection definition provided in Section 4.3.1, as it instead involves two scores to make a choice.

The function  $M_{f_i}$  creates a prompt similar to Prompt 7 that instructs the model to convert a textual analysis  $t_{f_i}$  to a structured output by letting the model select a short description that best aligns with the analysis (Deng et al., 2024b). This output is then mapped to a scalar  $v_{f_i}$ :

1. **Exploration extent mapping:** The exploration extent is evaluated by analysing the supporter’s understanding of the help-seeker’s distress. The feature score is defined as:

$$v_{f_1} = \begin{cases} -0.5, & \text{for limited understanding,} \\ 0.0, & \text{for partial understanding,} \\ 0.5, & \text{for clear understanding.} \end{cases} \quad (14)$$

2. **Distress level mapping:** The help-seeker’s emotional state is determined through their distress level, assigning a score based on the following conditions:

$$v_{f_2} = \begin{cases} -0.5, & \text{when calm,} \\ 0.0, & \text{when slightly anxious,} \\ 0.5, & \text{when overwhelmed.} \end{cases} \quad (15)$$

To increase the reliability of the analyses, the model considers one conversational feature at a time, instead of evaluating multiple features simultaneously. The inherent variability in model outputs is dealt with using *self-consistency* (Wang et al., 2023b; He et al., 2024b). For each feature  $f_i$ , multiple analyses and scalars are generated and their results are averaged to calculate the final feature score  $\bar{v}_{f_i}$ :

$$\bar{v}_{f_i} = \frac{1}{n} \sum_{m=1}^n \frac{1}{k} \sum_{j=1}^k \text{LLM}(M_{f_i}(t_{f_i}^m)). \quad (16)$$

Here,  $n$  is the number of generated analyses for feature  $f_i$ ,  $k$  is the number of scores sampled for an analysis, and  $t_{f_i}^m$  is the  $m$ -th analysis of a feature. Both scores are subsequently categorized using predefined thresholds  $\epsilon_{f_1}$  and  $\epsilon_{f_2}$ :

Exploration extent	Distress level	Optimal strategies
Insufficient	Low	Exploration
Sufficient	Low	Action
Insufficient	High	Exploration or comforting
Sufficient	High	Action or comforting

TABLE 1: Framework similar to a decision-tree for finding the most appropriate strategies given the supporter’s extent of exploration and the help-seeker’s distress.

$$\text{Exploration extent} = \begin{cases} \text{sufficient,} & \text{if } \bar{v}_{f_1} \geq \epsilon_{f_1}, \\ \text{insufficient,} & \text{otherwise,} \end{cases} \quad (17)$$

$$\text{Distress level} = \begin{cases} \text{high,} & \text{if } \bar{v}_{f_2} \geq \epsilon_{f_2}, \\ \text{low,} & \text{otherwise,} \end{cases} \quad (18)$$

After both features are categorized, we use a decision tree  $S_s(\bar{v}_{f_1}, \bar{v}_{f_2})$  to select an optimal strategy  $r_{j+1}$  (Medeiros et al., 2019; Xiong et al., 2024). Table 1 represents this tree and shows how categories map to optimal strategies. When multiple strategies are optimal, we use the tie-breaking approach described in Section 4.3.4.

#### 4.3.4 Resolving strategy selection ties

Table 1 shows that some situations allow for multiple valid strategies. For example, both exploring and comforting strategies are suitable for help-seekers who have insufficiently elaborated on their situation, but also appear distressed. Therefore, a second step is necessary to break these ties.

Unlike the feature-based selection process described in Section 4.3.2, this tie-breaking step involves choosing between two strategies. Choosing between two well-defined and contextually appropriate options allows the model to reason about their effectiveness, unlike situations with overlapping or unsuitable alternatives.

The function  $F_b$  is used to create the series of prompts necessary for tie-breaking. It is passed the conversation context  $c_i^{(j)}$  and the available strategies  $\mathcal{S}^* \subset \mathcal{S}$  applicable to  $c_i^{(j)}$ , resulting in an analysis  $t_b$  output by the model:

$$t_b = \text{LLM}(p_b), \quad p_b = F_b(c_i^{(j)}, \mathcal{S}^*), \quad (19)$$

The approach is similar to the process followed in Section 4.3.2. First, the model receives the supporter’s goal, i.e., to reduce the help-seeker’s distress, along with the conversational history. Then, it is instructed to reason about which strategy is most appropriate to use based on brief descriptions of when to use either one:

$$\begin{array}{|l} \text{User:} \\ \text{Should the supporter use \{strategy\_one\} or \{strategy\_two\}?} \\ \text{Choose \{strategy\_one\} if [...], or choose \{strategy\_two\} when [...].} \end{array} \quad (19)$$

Before the model analyses the conversational context, it is informed of its available choices. This is combined with Prompt 19 to produce the following:

**User:**

Should the supporter use {strategy\_one} or {strategy\_two}?  
Choose {strategy\_one} if [...], or choose {strategy\_two} when [...].

(20)

You can choose from the following options: (1) The supporter should use {strategy\_one}. (2) The supporter should use {strategy\_two}. Answer in a paragraph.

We found that this improved the model’s ability to identify relevant features in the conversational context, potentially because repeating the possible options near the end of the prompt reduces ambiguity. To focus the model’s attention on the active topic of discussion, Prompt 10 was used to include the relevant messages from the help-seeker and the supporter.

The function  $M_b$  creates a prompt that is used by the model to convert  $t_b$  to a strategy label in  $S^*$ . In line with our earlier approach, self-consistency was used to improve the consistency of the decision-making process. As mentioned earlier, the reliability of the context analyses improved because the model can more easily distinguish between two appropriate strategies. Symbolic reasoning is therefore not necessary and the strategy labels output by the model after processing  $M_b$  can be used directly instead. The results were aggregated through the mode of the sequence of strategy labels:

$$r_{j+1} = \text{mode} \bigcup_{m=1}^n \bigcup_{l=1}^k \text{LLM}(M_b(t_b^m)). \quad (20)$$

Here,  $n$  is the number of analyses,  $k$  is the number strategy labels sampled per analysis, and  $t_b^m$  is the  $m$ -th analysis.

#### 4.3.5 Support skill selection

After selecting the best strategy  $r_{j+1} \in \mathcal{S}$  to guide the conversation, a support skill that best implements the strategy needs to be chosen. Thus, we define a many-to-one mapping from support skills to strategies, following Liu et al. (2021b):

$$a_{j+1} = \begin{cases} a \in \{\text{QS}, \text{RP}\}, & \text{if } r_{j+1} = \text{exploration}, \\ a \in \{\text{RF}, \text{AR}\}, & \text{if } r_{j+1} = \text{comforting}, \\ a \in \{\text{PS}, \text{IN}\}, & \text{if } r_{j+1} = \text{action}. \end{cases} \quad (21)$$

The many-to-one mapping limits the decision space for support skill selection to only the skills belonging to the chosen strategy. To guide the model in selecting an appropriate skill  $a_{j+1}$  from subset  $\mathcal{A}_r \subset \mathcal{A}$ , where  $\mathcal{A}_r$  denotes the skills belonging to  $r_{j+1}$  (see Equation 21), we follow the strategy selection approach described in Section 4.3.1: the prompting function  $F_a$  creates a series of prompts that guide the model in producing a context analysis  $t_a$ , based on the current conversation context  $c_i^{(j)}$ , the chosen strategy  $r_{j+1}$ , and the available skill set  $\mathcal{A}_r$ :

$$t_a = \text{LLM}(p_a), \quad p_a = F_a(c_i^{(j)}, r_{j+1}, \mathcal{A}_r). \quad (22)$$

Time-constraints led us to hypothesize that the reduced reasoning scope for skill selection ( $|\mathcal{A}_r| \ll |\mathcal{A}|$ ) might allow the model to directly assess the appropriateness of each skill. Therefore, we tasked the model with directly explaining its skill preferences rather than limiting it to analyzing conversational features, as in Section 4.3.2.

To enable this, the model is first informed of the goal of the chosen strategy, similar to the feature analyses in Section 4.3.2. For example, the following prompt was used to select support skills belonging to the action strategy:

**System:**  
The supporter’s goal is to provide guidance to reduce the help-seeker’s distress. (21)

The model is then presented with the available support skill labels and their descriptions and instructed to explain its reasoning for choosing one. For instance, we used the following prompt to select actionable skills:

**User:**  
Should the supporter share information or provide suggestions?  
Choose *information* when the help-seeker needs facts [...]. Choose *suggestions* when the help-seeker [...] needs help exploring potential actions or solutions. (22)  
  
Choose from the following options: (1) The supporter should share information. (2) The supporter should provide suggestions. Answer with a paragraph.

After processing these prompts, the model’s output  $t_a$  contains an explanation of which support skill it prefers. The function  $M_a$  then creates a prompt that guides the model in converting its assessment into a skill label. Like strategy selection, self-consistency was used to improve consistency. The mode of the label sequence was used to aggregate mappings from analyses to selections, similar to the tie-breaking method in Section 4.3.4.

$$a_{j+1} = \text{mode} \bigcup_{m=1}^n \bigcup_{l=1}^k \text{LLM}(M_a(t_a^m)). \quad (23)$$

Here,  $n$  is the number of analyses,  $k$  the number of skill labels sampled per analysis, and  $t_a^m$  is the  $m$ -th analysis.

#### 4.4 Response generation

The chosen support skill  $a_{j+1}$  and its corresponding conversational context  $c_i^{(j)}$  are used to create a prompt that guides the supporter’s response generation and the model generates utterance  $u_{j+1}$  given prompt  $p_r$ :

$$u_{j+1} = \text{LLM}(p_r), \quad p_r = G_r(s_{j+1}, a_{j+1}, c_i^{(j)}), \quad (24)$$

where  $s_{j+1} = \text{supporter}$  and  $G_r$  is a function that includes the speaker, support skill, and context into the generation prompt. After this, the conversational context is updated by appending the generated utterance, the speaker, and the used support skill:

$$c_i^{(j+1)} = c_i^{(j)} \cup \{ \langle u_{j+1}, s_{j+1}, a_{j+1} \rangle \}. \quad (25)$$

In contrast, utterances from the help-seeker depend only on the current conversational context  $c_i^{(j)}$ , a distress topic  $t$ , and a fixed description of the help-seeker’s character present in  $G_h$ :

$$u_{j+1} = \text{LLM}(p_h), \quad p_h = G_h(s_{j+1}, \emptyset, c_i^{(j)}), \quad (26)$$



where  $s_{j+1} = \text{seeker}$  and  $a_{j+1} = \emptyset$  indicates that the help-seeker does not use support skills. The utterance is used to update the conversational context:

$$c_i^{(j+1)} = c_i^{(j)} \cup \{\langle u_{j+1}, s_{j+1}, \emptyset \rangle\}. \quad (27)$$

Similar to Section 4.3, dynamic natural language instructions are used to help the model implement support skills. Because each support skill addresses a unique support aspect, a “one-size-fits-all” approach for utterance generation does not exist. For instance, whereas some skills require emphasis on the contents of the topic of discussion, others prioritize the emotions expressed by the help-seeker. We found that ensuring the model focuses on the right aspects is important for generating appropriate responses.

Both literature and our preliminary experiments indicated that models, even at the scale of GPT-4o, often struggle to combine multiple instructions within a single prompt into a relevant output (He et al., 2024a), a limitation that is even worse in smaller-scale language models. Therefore, the response quality is improved by decomposing the generation task into three subtasks, which are outlined in Sections 4.4.1, 4.4.2 and 4.4.3.

#### 4.4.1 Coarse-grained context analysis

Some support skills require an understanding of the entire conversation to respond appropriately to the help-seeker’s last message. The first subtask therefore consist of a coarse-grained analysis of the conversation. This prioritizes context at the dialogue level and reduces the model’s confusion on which speaker made what statement.

For example, the model extracts the help-seeker’s emotional state throughout the conversation when formulating a comforting response. After including the conversational history in its system prompt, the model is instructed to articulate and describe the help-seeker’s feelings:

**User:**  
Briefly and concisely articulate and describe the help-seeker’s experience, based on the conversational history. Answer in a single paragraph. (23)

The relevance of the model’s summary is improved by referencing the help-seeker with a completion prompt:

**Assistant (completing):**  
The help-seeker [model completes] (24)

#### 4.4.2 Fine-grained context analysis

In the second subtask, the specific emotions or tone expressed in the help-seeker’s latest message are extracted to improve the relevance of the supporter’s response. For instance, while the dialogue level context might focus on job-related anxiety, the help-seeker’s latest message might express gratitude. The model is primed to describe the emotions expressed in the help-seeker’s message by including it in a completion prompt:

**Assistant (completing):**  
Help-seeker: “{seeker\_message}” (expressing **feelings of hope and gratitude, but also uncertainty**) (25)

This approach helps the model to focus on the current emotional context to ensure that the generated response aligns with the help-seeker’s expressed feelings. Interestingly, we found that this “parenthetical method” lead to concise but accurate descriptions of the help-seeker’s feelings, which resulted in more empathetically appropriate responses.

#### 4.4.3 Response formulation

In the third subtask, the coarse- and fine-grained analyses from the previous two subtasks are combined with detailed instructions on implementing a support skill:

**User:**  
 “{support\_skill}” focuses on the supporter [...]  
 Give an example of “{support\_skill}” applied in a response from the supporter to the help-seeker’s last message. (26)

Importantly, prompting the model to generate an example of a response rather than querying it to reply as the supporter improved output quality: responses were concise, more strictly followed instructions, and avoided excessive empathy. Moreover, the coarse-grained analysis (Prompt 24) and fine-grained analysis (Prompt 25) were included into a completion prompt to improve the relevance of the response. Finally, the model’s attention is primed to formulate a response from the perspective of the supporter:

**Assistant:**  
 1. Help-seeker’s experience: “{experience}”  
 2. Help-seeker’s last message: “{seeker\_message}” (expressing {feelings}) (27)  
 Here’s an example of a response using {support\_skill}:  
 Supporter: “**You’re welcome, did you speak with your boss about your anxiety?**”

When formulating utterances, this combined approach of including conversation-level and utterance-level insights lets the model consider both the conversation’s overarching theme and more local details.

#### 4.5 Conversation termination

Conversations should not have fixed lengths, as this can lead to situations where conversations end prematurely or become repetitive. Instead, they should end when the help-seeker’s issue has been resolved. Therefore, the self-chat process should continue to generate utterances until a dynamic terminal state is reached: a maximum number of turns is exceeded, or the supporter has successfully helped the help-seeker.

Following Deng et al. (2024b), the model was prompted to assess the help-seeker’s status and determine whether the supporter successfully resolved the help-seeker’s problems. Following this analysis, the model was instructed to select a short description that best aligned with its assessment:

$$v_r = \begin{cases} -1.0, & \text{if the help-seeker feels worse,} \\ -0.5, & \text{if the help-seeker feels the same,} \\ 0.5, & \text{if the help-seeker feels better,} \\ 1.0, & \text{if the supporter’s help is sufficient,} \end{cases} \quad (28)$$

where  $v_r$  is a mapping of a description to a scalar. For  $n$  status scores generated by model, the help-seeker's issue was considered addressed if their mean passed threshold  $\epsilon_r$ :

$$\frac{1}{n} \sum_{i=1}^n v_r^i \geq \epsilon_r. \quad (29)$$

However, we found that the model often falsely assumed the problem was resolved based on positive words like expressions of gratitude. Therefore, the model was additionally prompted to assess if the help-seeker appeared done with speaking:

$$v_f = \begin{cases} -1.0, & \text{if the help-seeker wants to continue talking,} \\ 0.0, & \text{if the help-seeker might continue talking,} \\ 1.0, & \text{if the help-seeker is done talking.} \end{cases} \quad (30)$$

The conversation was considered complete if these  $n$  generated status scores also exceeded a second threshold  $\epsilon_f$ :

$$\frac{1}{n} \sum_{i=1}^n v_f^i \geq \epsilon_f. \quad (31)$$

This ensured that conversation termination occurred only when it was clear that the help-seeker's issue had been addressed and they were ready to end the conversation, leading to a natural dialogue ending.

## 4.6 Self-chat

To create synthetic dialogues, the components covered in Sections 4.2, 4.3, 4.4, and 4.5 — namely topic generation, support skill selection, response formulation, and conversation termination — were combined into a self-chat framework.

Given a conversation topic, the model assumes both the role of the help-seeker and the supporter. Following an opening message from the help-seeker, the model alternates between these roles. The utterances generated during the supporter's turns depend on the conversational context and a support skill based on the optimal support strategy, while the help-seeker's utterances depend only on the conversational context and its topic. After each turn, the model assess whether the help-seeker seems satisfied with the supporter's advice,

---

### Algorithm 1 Self-chat framework

---

```

1: function GENERATE()
2:    $j \leftarrow 1$  ▷ Initialize the step counter
3:    $t \leftarrow \text{TOPIC}()$  ▷ Generate a conversation topic
4:    $s_j \leftarrow \text{seeker}, u_j \leftarrow \text{INITIATE}(t), a_j \leftarrow \emptyset$  ▷ Start with an opener from the seeker
5:    $c_i^{(j)} \leftarrow \{\langle u_j, s_j, a_j \rangle\}$ 
6:   repeat
7:      $s_{j+1} \leftarrow \text{SWAP}(s_j)$  ▷ Swap speakers (e.g., seeker  $\leftrightarrow$  supporter)
8:      $r_{j+1} \leftarrow \text{STRATEGY}(c_i^{(j)}, s_{j+1})$  ▷ Select a conversational strategy
9:      $a_{j+1} \leftarrow \text{ACTION}(c_i^{(j)}, r_{j+1}, s_{j+1})$  ▷ Select an appropriate support skill
10:     $u_{j+1} \leftarrow \text{RESPONSE}(c_i^{(j)}, a_{j+1}, t)$  ▷ Generate the next utterance ( $t$  is optional)
11:     $c_i^{(j+1)} \leftarrow c_i^{(j)} \cup \{\langle u_{j+1}, s_{j+1}, a_{j+1} \rangle\}$ 
12:     $j \leftarrow j + 1$ 
13:  until  $s_{j-1}$  is seeker and  $\text{ENDING}(c_i^{(j)})$  ▷ Check ending if the seeker was speaking
14:  return  $c_j$ 

```

---

to dynamically end the conversation. The complete pseudo-algorithm is outlined in Algorithm [1](#).

## 5 Experiments and results

The current section focuses on the evaluation individual components of the proposed methodology and the dialogues resulting from it. We combine automatic and human evaluations to capture both the objective and subjective aspects of our approach.

Experiments were performed using a Python 3.11.3 implementation of the proposed methodology. Each experiment was run on a Red Hat Enterprise Linux 9.4 system, with an Intel Xeon Gold 6126 @ 2.6GHz CPU and an NVIDIA A100 GPU. We use the Meta LLAMA 3-8B model implementation from the Hugging Face transformers library. The output of the model is controlled using the Outlines framework.<sup>10</sup>

### 5.1 General characteristics

To compare the characteristics of the generated synthetic dialogues with those of the human crowdsourced `ESCONV` dataset, we provide some high-level dataset statistics, which are shown in Table 2.

Our final dataset consists of 100 dialogues, each consisting of approximately 459.2 words. On average, the utterances consist of 25.0 words, indicating that utterances from our dialogues are of similar length as those in `ESCONV` dialogues ( $\approx 50\%$  longer). Furthermore, dialogues consist of 18.3 utterances, suggesting that many conversations do not dynamically terminate naturally, as the maximum number of turns was set to 22. Manual assessment of such dialogues showed that later conversational stages contain loops where supporters provide suggestions that help-seekers acknowledge without progressing the conversation, thus preventing a natural ending to it. This likely reduces the diversity of dialogues, as well as their topics, which we discuss in Section 5.2. Furthermore, the evaluations of the dialogue policy planning approach in Section 5.3 may be affected by this pattern of using a single support skill in later conversational stages.

### 5.2 Diversity and coherence analysis

The dataset’s dialogue diversity and coherence are evaluated with the goal of knowledge distillation in mind. The training dataset’s diversity affects a downstream model’s abilities to understand elaborate conversational contexts; limited diversity may lead to poor generalization and potential overfitting. Similarly, the dialogues’ coherence, i.e., the maintenance of realistic conversational flows, consistency of speaker roles, and adherence to topic seeds, improves a downstream model’s ability to interpret context across multiple turns and allows for better adaptation to shifts in conversational dynamics (Liu et al., 2021b). Furthermore, coherent training data helps generative models learn how to provide more fluent and contextually relevant responses, which leads to an improved end-user experience.

#### 5.2.1 Lexical analysis

Following Chen et al. (2023a), the lexical diversity of our synthetic dataset is measured using the Distinct-2/3 metric (Li et al., 2016). It is defined as the ratio of unique  $n$ -grams to the total number of  $n$ -grams in a text, with scores closer to 1.0 indicating more diverse or creative writing (Xing et al., 2017). In line with Bao et al. (2020), the text was preprocessed by lowercasing, tokenizing, and removing punctuation to calculate the score. Distinct-2/3 scores are calculated at (1) the utterance level to reduce dialogue length bias when analyzing response diversity, and (2) the dataset level to measure how repetition potentially affects

<sup>10</sup><https://github.com/dottxt-ai/outlines>

	ESCONV		Ours	
	Mean	$\sigma$	Mean	$\sigma$
# Dialogue words	483.8	185.8	459.2	152.6
# Utterances	29.5	10.2	18.3	4.6
# Utterance words	17.0	5.9	25.0	4.9
<b># Dialogues</b>	<b>1,300.0</b>		<b>100.0</b>	

TABLE 2: Comparison of dataset-, dialogue-, and utterance level statistics between ESCONV and our synthetic dataset.

Dataset	Distinct-2		Distinct-3		S-BERT
	(S)	(D)	(S)	(D)	
ESCONV	0.88	0.48	0.97	0.81	0.11
Ours	0.72	0.27	0.83	0.50	0.21

TABLE 3: Comparison of syntactic and semantic diversity between ESCONV and our synthetic dataset through Distinct-2/3 and S-BERT embedding cosine similarity scores, respectively. Distinct-2/3 scores were calculated at the utterance level (S) and dataset level (D).

dialogue diversity. Furthermore, we control for length when calculating dataset level diversity, since longer texts generally have lower Distinct-2/3 scores than shorter texts for similarly diverse vocabularies (Liu et al., 2022). We therefore sampled 100 random dialogues from ESCONV to ensure approximately equal input lengths.

To assess the effect of the previously mentioned conversational loops on dialogue diversity, we also evaluated the semantic similarity of utterances using Sentence-BERT (S-BERT) embeddings. For both datasets, we encode each unique pair of responses into semantic vectors using S-BERT and compute their cosine similarity scores, with values closer to 0.0 indicating lower semantic similarity (Reimers et al., 2019). We then take the average pairwise similarity of all response pairs as a dataset’s semantic utterance level diversity.

The results are shown in Table 3. The lower Distinct-2/3 scores suggest that, compared to human dialogues from ESCONV, our synthetic dialogues have lower lexical diversity at both the utterance level (0.88 vs 0.72 and 0.97 vs 0.83) and the dataset level (0.48 vs 0.27 and 0.81 vs 0.50). Additionally, the average S-BERT similarity scores for our responses (0.21) and human responses (0.11) hint at greater similarity between our utterances and indicate a higher level of repetition in our synthetic dataset. An assessment of the 20 “least original” utterances (each at least 60 characters in length) in our dataset and their 4 closest matches shows that the model indeed repeats responses with little to no variation within and between dialogues (see Appendix A).

### 5.2.2 Dialogue diversity

Content diversity between dialogues was measured by constructing TF-IDF vectors for each dialogue in our synthetic dataset. These vectors encode the relative importance of words in a dialogue given their frequency across the entire dataset and thus represent a dialogue’s vocabulary. For all dialogue pairs, the cosine similarity score of their TF-IDF vectors was calculated. Scores closer to 0.0 indicate a higher dissimilarity between the dialogues’ contents, meaning different themes were covered (Zheng et al., 2023a). To prepare the dialogues for analysis, we followed the text preprocessing pipeline proposed in Egger et al. (2022); using the Scikit-learn library, stopwords and irrelevant text (e.g., punctuation and numbers) were removed, followed by text tokenizing and lemmatizing using WordNet Lemmatizer.<sup>11</sup>

Figure 2a displays a histogram illustrating the distribution of cosine similarity scores for dialogue pairs in ESCONV and our synthetic dataset. Synthetic dialogues appear to have a moderate similarity, which suggests that our dataset has a reasonable degree of content

<sup>11</sup><https://scikit-learn.org/>

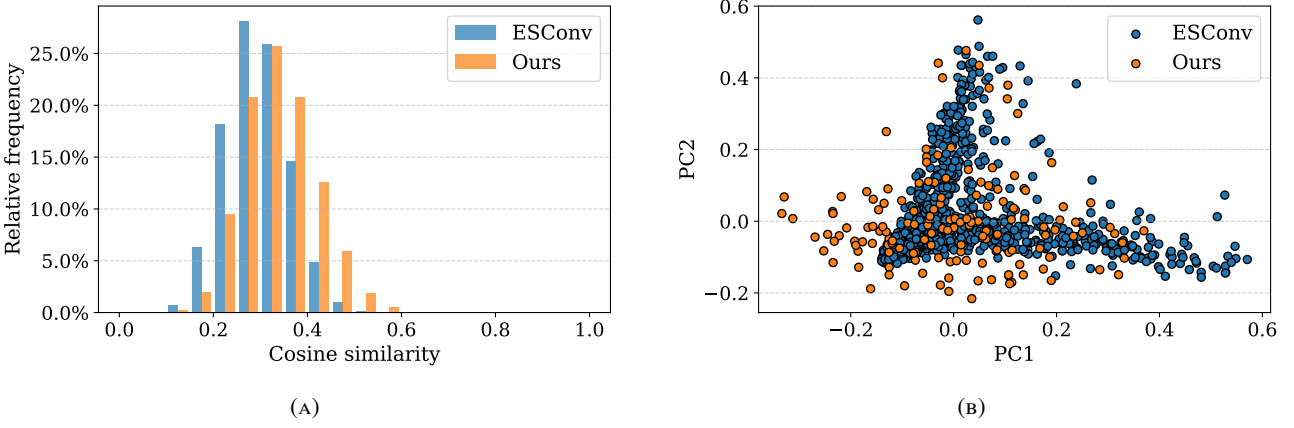


FIGURE 2: (A) Histogram showing the distribution of cosine similarity scores between dialogue pairs for `ESConv` and our synthetic dataset. (B) Scatter plot resulting from 2D PCA dimensionality reduction of TF-IDF vectors for dialogues from `ESConv` and our synthetic dataset.

diversity, with overlaps between dialogues likely being a result of the shared emotional support setting. The distribution for `ESConv` dialogues is more narrow and peaks at lower similarity scores, indicating that these dialogues are more diverse.

### 5.2.3 Topic similarity

Zheng et al. (2023a) and Zheng et al. (2024) proved that topic diversity is important in enabling a downstream model to contextually understand a large range of topics. We therefore apply a number of topic modelling methods in this section and Section 5.2.4 to investigate how topics are discussed.

The topic similarity between dialogues can be compared by visualizing 2D PCA embeddings of TF-IDF vector representations of dialogues (Zheng et al., 2023a). Neighboring points will share similar TF-IDF features, indicating similar dialogue contents. Preparing the dialogues for analysis followed the same steps as those described in Section 5.2.2.

Figure 2b shows the results of this process. Each point represents a single dialogue. It appears that our dialogues cover a broader range of topics than `ESConv` dialogues, as suggested by their wider spread. However, the substantial overlap between points from both datasets indicates that both cover similar themes, likely as a result of the shared psychological framework underlying the conversations. This implies that while our synthetic dataset improves topic coverage, it still shares the conversational structural with `ESConv`.

### 5.2.4 Topic analysis

Dialogue themes were analysed by using BERTopic to identify topic patterns.<sup>12</sup> These are modeled by creating clusters of topics based on BERT-embeddings of conversations (Grootendorst, 2022). Representative words are then extracted for each cluster, resulting in easily interpretable descriptions of the topics covered in the conversations. BERTopic was chosen over Latent Dirichlet Allocation, because the former generated more interpretable topic representations compared to the topic-word distributions produced by the latter. Because BERTopic uses embeddings from transformer models, maintaining the original structure of conversations is necessary (Egger et al., 2022). Therefore, no preprocessing was applied to the text.

<sup>12</sup><https://github.com/MaartenGr/BERTopic>

ESCONV		Ours	
Topic	Count	Topic	Count
depression_depressed_suffered_debilitating	192	undervalued_valued_recognized_aren	18
anxiety_attacks_guard_panic	137	creative_platforms_behance_creatives	17
stressed_stress_stressful_stressing	123	rut_stuck_get_don	14
holidays_holiday_christmas_thanksgiving	115	us_our_differences_find	13
vaccine_vaccines_vaccination_vaccinated	113	concerns_friend_okay_inquiry	12
breakup_breakups_ups_break	105	cousin_support_uncertain_rather	11
cat_dog_dogs_pet	99	parents_be_critical_fears	11
discuss_else_anything_talk	98	yeah_right_true_hard	11
books_book_reading_read	92	silence_rejection_isolating_friend	10
fear_scared_fears_scary	91	fear_boss_as_seen	10
yes_yessir_tons_cause	88	interests_share_who_groups	10
walks_walk_walking_air	85	anxiety_waiting_stress_edge	10
therapist_therapy_therapists_uninsured	82	concerns_conversation_different_reframing	10
zoom_facetime_meetings_conferencing	75	running_hopeless_cycle_trapped	9

TABLE 4: Most common topics identified by BERTopic in the ESCONV dataset and our dataset, with their frequencies.

Table 4 shows the 15 most common topics in dialogues from our synthetic dataset and the ESCONV dataset. Conversations from the ESCONV dataset seem to center around mental health problems like anxiety and depression, problems in relationships, coping mechanisms like walking or reading books, and challenges related to COVID-19. Likewise, synthetic conversations revolve around emotional problems like anxiety and fear of judgement, and conflicts in settings related to family, friends, and the workplace. These findings indicate that our synthetic dialogues are well grounded on their topic seeds, resulting in realistic conversations that have strong overlaps in content with human dialogues.

### 5.3 Dialogue policy planning

Our methodology uses context analysis and support skill selection (see Section 4.3) to enable an LLM to guide conversations towards predefined goals. In this section, the changes in conversational features and support skills throughout conversations are evaluated using the ESCONV dialogue dataset and our synthetic dialogues. Based on the evaluation results, we aim to: (1) understand how effectively our model captures conversational dynamics to provide support, (2) gain insights into alignment of emotional support in human and synthetic conversations, and (3) assess the model’s implementation of the support framework proposed by Liu et al. (2021b).

#### 5.3.1 Conversational features

The selection of strategies relies on the assessment of conversational features described in Section 5.3.1. The ESCONV dialogue dataset ( $d_{\text{esc}}$ ) was used to evaluate how well these features reflect changes in the dynamics of human conversations over time.

To obtain representations of conversational stages in dialogues, we split the dialogues into increasingly long segments. Specifically, each randomly sampled dialogue,  $c_i \sim d_{\text{esc}}$ , was divided into  $h$  equally sized intervals, with every interval containing roughly the same number of turns. For each interval  $v$ , we represented the help-seeker’s first utterance by index  $m_v$  denoting its position in the original dialogue  $c_i$ . Dialogues containing intervals without utterances from the help-seeker were excluded from consideration. Recall that



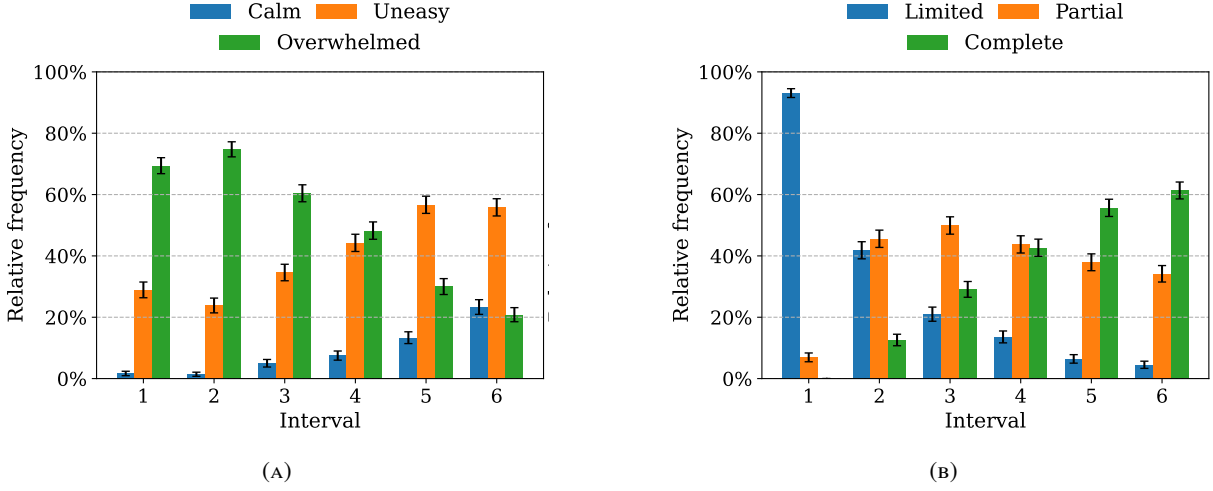


FIGURE 3: Relative frequencies of the model’s predictions of the (A) help-seekers’ emotional distress and (B) supporters’ understanding across intervals in dialogues from the ESCONV dataset. Error bars represent 95% confidence intervals.

Equation 6 defined a dialogue truncated at turn  $j$  as follows:

$$c_i^{(j)} = (\langle u_k, s_k, a_k \rangle)_{k=1}^j. \quad (32)$$

Thus, the evaluation set  $Y_f$  is constructed by sampling  $q$  dialogues from  $d_{\text{esc}}$  and extracting  $h$  segments from each of these  $q$  dialogues, resulting in a total of  $q \times h$  segments:

$$Y_f = \bigcup_q \bigcup_h \left\{ c_i^{(m_v)} \mid c_i \sim d_{\text{esc}} \right\}. \quad (33)$$

To analyse the progression of the supporter’s understanding of the help-seeker’s distress ( $f_1$ ), as well as the perceived changes in the emotional state of the help-seeker ( $f_2$ ), the numbers of dialogues sampled from ESCONV and intervals were set to  $q = 80$  and  $h = 6$ , respectively. This resulted in a test set  $Y_f$  consisting of 480 segments. Scores resulting from running feature analyses on these segments were used to obtain Figures 3a and 3b.

Figure 3a shows the evaluation of distress levels across intervals. Most help-seekers feel *overwhelmed* in the first two intervals, but in the following intervals the less severe *uneasy* category increases strongly while the *overwhelmed* category decreases. Distress appears to be present across all conversational stages, as the *calm* category remains low in every interval. These observations seem to align with real human support conversations, where the decrease of overwhelmedness and increase of more manageable emotional states reflect the natural progression of emotional support (Burleson, 2003; Hill, 2020).

The extent to which the supporter understood the help-seeker’s experience as the conversation progressed is presented in Figure 3b. The supporter initially had a *limited* understanding, which turned into *partial* or even *complete* understanding over time. Given the high frequency of exploratory skills used by humans at the start of conversations (see Figure 6), this suggests that the supporter’s understanding improved as the help-seeker’s experience and feelings were explored. Supporters start conversations with limited knowledge of the help-seeker’s situation, but over time gain more understanding through exploration, which intuitively aligns with how human conversations progress.

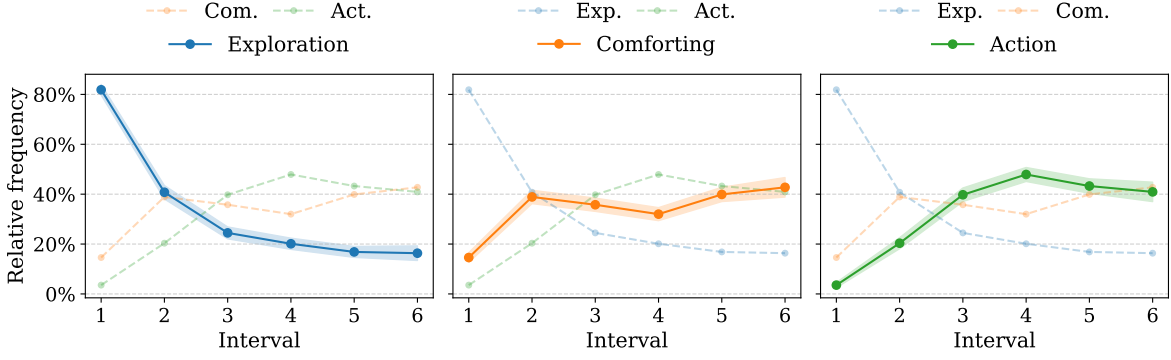


FIGURE 4: Progression of strategy usage over time by crowdworkers in dialogues sampled from `ESCONV`. The graphs show the relative frequency per interval of three strategies: exploration (Exp.), comforting (Com.), and action (Act.). The shaded areas around lines represent 95% confidence intervals.

### 5.3.2 Strategy selection in human dialogues

To evaluate the conversion of conversational feature scores into support strategies, the same testing parameters from the experiments in Section 5.3.1 were used: before sampling dialogues from `ESCONV`, we set  $q = 80$  and  $h = 6$  to obtain  $Y_S$ . The strategies selected by crowdworkers and the model are illustrated in Figures 4 and 5, respectively. Each subplot shows a strategy, with the shaded areas representing 95% confidence intervals, assuming a normal distribution of the relative frequencies.

Both figures show that exploration is prioritised at the start of the conversation, suggesting that the model initially focuses on gathering knowledge about the help-seeker’s experience. This aligns with the high number of *limited* knowledge scores observed in Figure 3b. The strong increase in using actionable support skills seems to coincide with a decrease in exploration, indicating that the model prioritises resolving distress once it has enough information on the help-seeker’s situation.

However, crowdworkers to show an increasing use of comforting over time, whereas the model’s usage of comforting strategies remains at a low but constant level throughout the conversations. This aligns with the predicted distress levels seen in Figure 3a, with the model prioritizing comfort only when it perceives high overwhelm from the help-seeker. Furthermore, humans show a more balanced use of strategies; the increase of the comforting and action strategies is coinciding with a steep decline of exploration. In contrast, the model’s selections show a more gradual decline in exploration, with a strong increase in actionable support, and small changes in comforting.

### 5.3.3 General support skill statistics

For a birds-eye view on the alignment between our model’s and human usage of support skills, we collected general statistics on support skill frequencies and patterns. These statistics additionally provide insights into the model’s adherence to the support framework proposed by Liu et al. (2021b).

Table 5 compares support skill usage in `ESCONV` and synthetic conversations. The model relies primarily on asking questions and providing suggestions to guide conversations, whereas humans use support skills in a more balanced way. Furthermore, human supporters seem more inclined to use affirming or reassuring words to provide comfort, while the model prioritises reflecting on the help-seeker’s feelings. When engaging in actionable support,

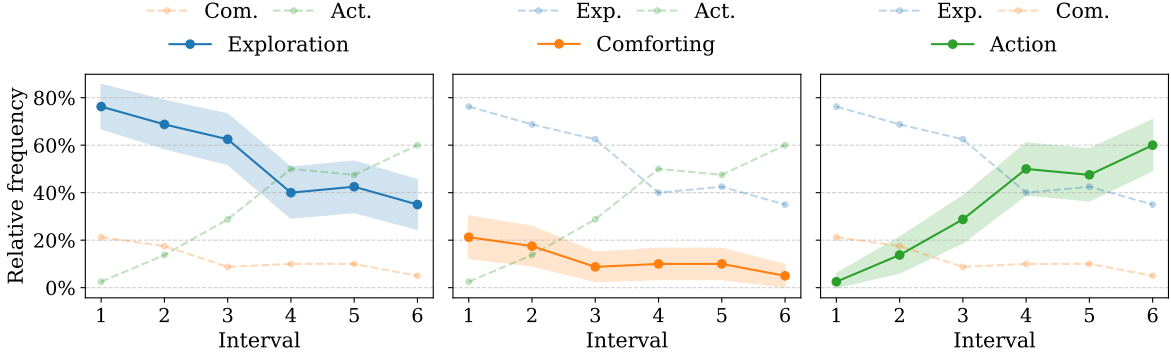


FIGURE 5: Strategies to use in dialogues sampled from *ESCONV*, predicted using our selection methodology. The graphs show the relative frequency per interval of three predicted strategies: exploration (Exp.), comforting (Com.), and action (Act.). The shaded areas around lines represent 95% confidence intervals.

providing suggestions is preferred over information sharing by both our model as well as humans; however, the model almost never shares information.

Table 6 shows the three most common support skills sequences at 3 and 4 hops (Liu et al., 2021b). These hops show both how support skills depend on each other, as well as frequent patterns in guiding conversations. The model appears to mainly alternate between asking questions and providing suggestions, which is similar to the results derived from Table 5. These skills are combined with restating or paraphrasing the help-seeker’s statements to reflect on their experiences, indicating that the model primarily focuses on exploration and problem resolution. However, the comforting support skills are not present, which could mean that the model lacks the ability to determine when to use an emotionally validating response.

### 5.3.4 Support skill trends

The distribution of support skills at different conversational stages was evaluated to complement the insights gained in Section 5.3.3. In line with the experimental set-ups proposed in Sections 5.3.1 and 5.3.2, we sampled  $q = 80$  dialogues from our synthetic dataset, which were each divided into  $h = 6$  intervals. The relative frequency of support skills within each segment was calculated to understand how support skills follow each other in our dialogues (Liu et al., 2021b).

Support skill	<i>ESCONV</i>	Ours
Questions ( <i>QU</i> )	20.69%	33.98%
Restatement or Paraphrasing ( <i>RP</i> )	5.93%	6.55%
Reflection of feelings ( <i>RF</i> )	7.82%	17.88%
Affirmation and Reassurance ( <i>AR</i> )	15.38%	2.02%
Providing Suggestions ( <i>PS</i> )	16.07%	39.55%
Information ( <i>IN</i> )	6.61%	0.02%

TABLE 5: Distribution of the relative frequencies of used support skills between dialogues from the *ESCONV* dataset and our synthetic dataset.

Hops	Sequence	Proportion
4-Hops	QU → PS → QU → PS	12.09%
	QU → RP → QU → PS	5.59%
	PS → QU → PS → QU	5.49%
3-Hops	QU → PS → QU	12.03%
	QU → RP → QU	8.86%
	PS → QU → PS	7.59%

TABLE 6: Proportions of the top-3 support skill sequences at 3-hops and 4-hops in our synthetic dataset. The abbreviations align with those in Table 5.

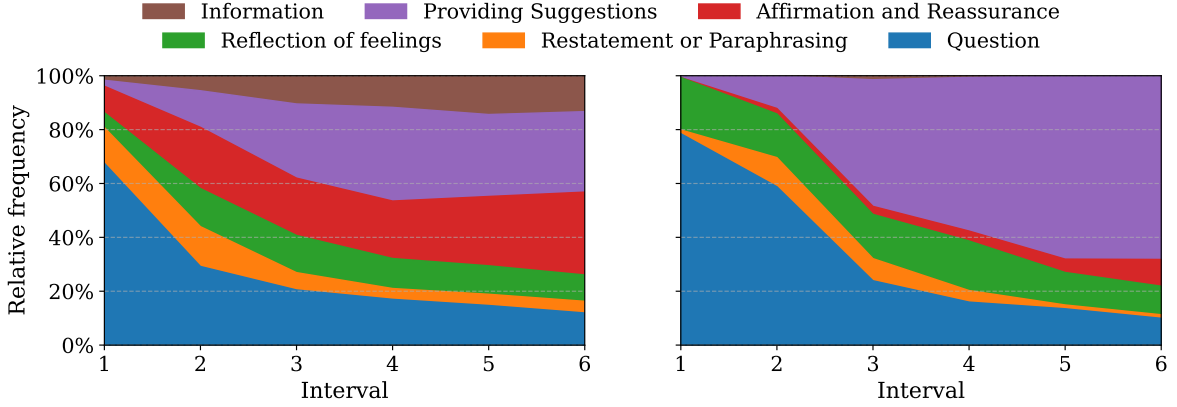


FIGURE 6: Relative frequency distribution of six support skills across intervals. The left graph shows the distribution in `ESCONV` dialogues, while the right graph shows distribution in our dialogues.

Figure 6 shows that the support skill usage in our synthetic dialogues generally aligns with the results obtained from strategy selection evaluation on human dialogues in Section 5.3.2. During the beginning stages of the conversations, asking questions is the primary support skill, in line with the model’s predicted use of exploration at the start of human conversations (see Figure 3b). Over time, asking questions becomes less dominant and the model starts to provide suggestions, which aligns with the transition to providing guidance proposed in the support framework from Liu et al. (2021b).

However, our dialogues show a more constrained use of support skills compared to `ESCONV` dialogues. As conversations progress, the model overuses Providing Suggestions and Reflection of Feelings compared to humans, while underusing both Affirmation and Reassurance, and Information. Additionally, the model’s support appears to collapse into using a single skill, which contrasts with the more balanced progression of support skill usage in human dialogues. These results suggest that although the model is able to follow the support framework in terms of following exploration with guidance, it lacks the nuance of human support.

### 5.3.5 Human evaluation of dialogue planning

The reliability of support skill annotations in `ESCONV` is limited, since they were assigned by individual, non-expert raters (Liu et al., 2021b). To address this issue, we invited a psychology master’s student to re-annotate dialogue contexts from `ESCONV`. Using our model as a third rater allowed us to assess the skill prediction agreement between three raters on human data, improving the reliability of our experiments.

To construct an evaluation dataset  $Y_{\mathcal{A}}$  using dialogues from `ESCONV`, we let  $d_{\text{esc}}^a \subseteq d_{\text{esc}}$  denote the subset of dialogues in which (1) a supporter used a support skill  $a$  during at least one turn, and (2) these turns were directly preceded by a message from the help-seeker. This ensures alignment of the support skill selection contexts for human evaluators and our model, since the latter depends on the help-seeker’s latest message to make decisions. For every support skill  $a \in \mathcal{A}$ , we sampled  $q$  dialogues. In each randomly sampled dialogue,  $c_a \sim d_{\text{esc}}^a$ , a turn satisfying the above conditions was randomly selected and the dialogue truncated at this turn. Using Equation 6, the evaluation set  $Y_{\mathcal{A}}$  can be denoted as:

$$Y_{\mathcal{A}} = \bigcup_{a \in \mathcal{A}} \bigcup_q \left\{ c_a^{(r_{c_a})} \mid c_a \sim d_{\text{esc}}^a \right\}, \quad (34)$$

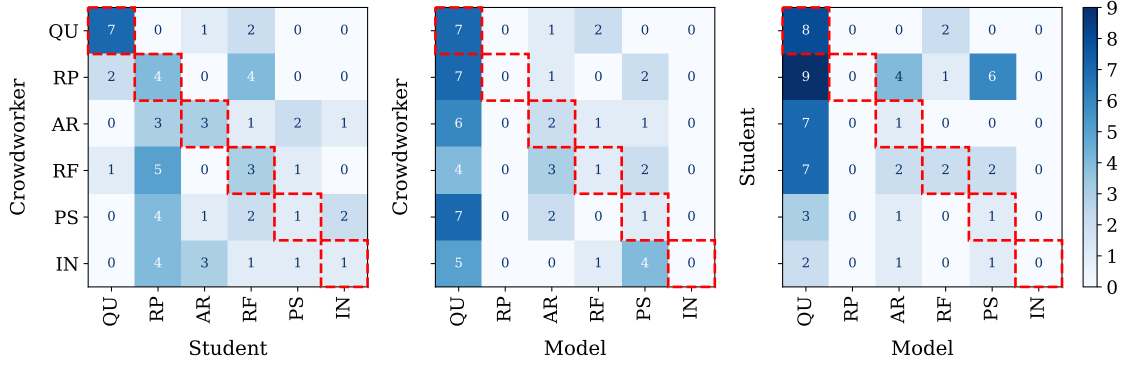


FIGURE 7: Confusion matrices comparing support skill selections between raters. The matrix on the left shows agreement between the psychology student and crowdworkers, the matrix in the middle shows agreement between crowdworkers and the model, and the matrix on the right shows agreement between the psychology student and the model. The highlighted diagonal elements denote selections where both raters agreed. The abbreviations on the axes align with those in Table 5.

where  $r_{c_a}$  is the index of the supporter’s selected turn. To compare support skill selections between all three raters, the last turn was removed from every truncated conversational context in  $Y_A$ . Our model and the psychology student were subsequently instructed to select a support skill for responding to the help-seeker’s latest message.

Figure 7 presents the confusion matrices that compare the support skill selections by crowdworkers, the psychology student, and our model. The diagonal elements, highlighted in red, show where raters agree. The strongest agreement between all raters is on *questions* (QU), whereas the agreement on *comforting skills* (AR and RF) and *actionable skills* (PS and IN) is much lower. This is both the case between human raters and between humans and the model. For instance, the psychology student often chose *restatement or paraphrasing* (RP) where crowdworkers used comforting or actionable skills, while the model never selected RP or IN at all. The latter observation aligns with results from Figure 6, where the model favors QU or *providing suggestions* (PS) rather than balancing these with RP and IN, respectively. This suggests that the model has problems with differentiating between skills that belong to the same strategy. Furthermore, the model predicted QU 36 times, which exceeds the psychology student’s (30) and crowdworkers’ (20) total usage of exploratory skills and is in line with the finding that the model often relies on a single skill per strategy.

The frequency distribution of the model’s predictions of exploratory and actionable skills aligns more closely with that from the student. Both the model and the student selected exploratory skills more than 30 times, compared to just 20 times by the crowdworkers. Similarly, the model’s use of actionable skills (10) is more in line with that of the student (9) than of the crowdworkers (20). The model selected comforting skills only 14 times, compared to more than 20 times by the human raters, suggesting the model underuses empathetic skills.

However, aggregating support skills into their overarching strategies and computing F1-scores, reflecting how well the model’s choices match those of humans, showed the model’s decisions align more closely with those of crowdworkers than of the student. While the model performs exploration similar to both humans (0.50 with crowdworkers and 0.52 with the student), it shows better alignment with crowdworkers on comforting (0.41 vs 0.29) and actionable strategies (0.33 vs 0.21).

Finally, the confusion matrices show that the humans demonstrate moderate agreement on strategy selection. When computing F1-scores between human raters, we find 0.52 on

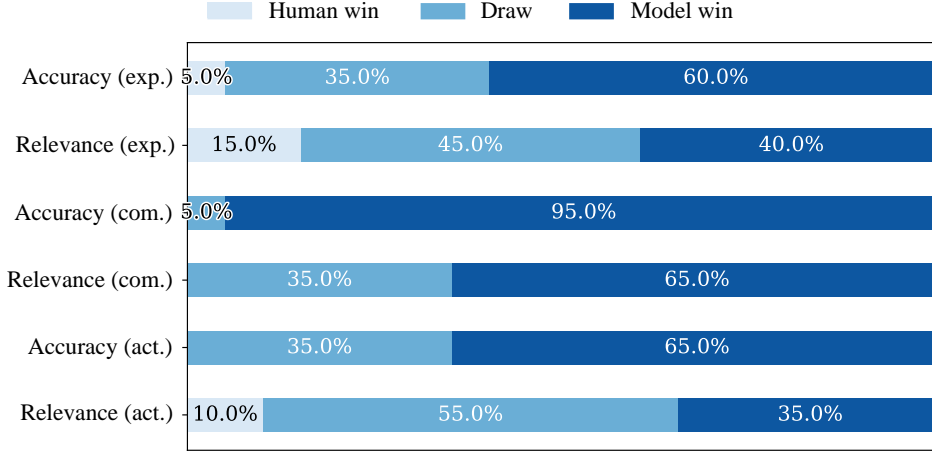


FIGURE 8: Win-rates for responses generated by the model and written by humans across support strategies. The bar-plot displays accuracy and relevance metrics for exploration (exp.), comforting (com.), and action (act.) strategies.

exploration, 0.34 on comforting, and 0.35 on action. This highlights the subjectivity in support skill selection, which likely contributes to the model’s limitations.

## 5.4 Response generation

We evaluated the performance of the response generation component at both the response level and dialogue level. At the response level, human evaluations were used to assess how well utterances adhered to their support skills and how relevant they were to the conversational context. At the dialogue level, the LLM-as-a-judge paradigm was followed to measure multiple qualitative features, which were compared between human and synthetic dialogues.

### 5.4.1 Relevance and accuracy

To assess the model’s ability to apply support skills during a conversation, we evaluate two metrics: the response’s *relevance* and *accuracy*. Accuracy reflects how responses adhere to skill descriptions, while relevance measures how well responses fit within to the conversation and logically follow from the context (Chen et al., 2023b). The evaluations were run on the same test dataset  $Y_A$  as used in Section 5.3.5; however, the supporter’s last response was not truncated from the context.

While automatic metrics are important for evaluating the proposed methodology, they often poorly correlate with human judgements when features such as context, tone, or empathy are involved (Lee et al., 2021). Additionally, we found that for a given support skill, the *ESCONV* dataset lacks reliable references for evaluating generated responses. We therefore decided to include human evaluations in our response generation analysis. Both accuracy and relevance were analysed through preference ranking, where the psychology student was instructed to select their preference between 60 response pairs (or select a tie) based on which response better suited our specified criteria (see Appendix D).

For an isolated conversational context  $c_a^{(r_{ca})} \in Y_A$ , the last utterance  $u_\ell$  and support skill  $a_\ell$  in context  $c_a^{(r_{ca})}$  were selected. The model was then prompted to generate an utterance  $\hat{u}_\ell$  grounded on skill  $a_\ell$ . The generated utterance was combined with the original utterance to let the evaluator assess the pair  $\langle u_\ell, \hat{u}_\ell \rangle$  based on the proposed response evaluation metrics.

Dataset	Informativeness	Understanding	Helpfulness	Consistency	Coherence
ESCONV	3.90	3.60	3.52	4.20	4.23
Ours	<u>4.22</u>	<u>4.93</u>	<u>4.45</u>	<u>4.95</u>	<u>4.97</u>

TABLE 7: LLM-as-a-judge evaluation results for quality of dialogues from ESCONV and our synthetic dataset. Ratings are based on a 1 to 5 Likert-scale. The highest score on a metric is underlined.

Figure 8 shows human and model win rates for the generated responses. The model scores high in accuracy with better win rates on all strategies, suggesting that it more accurately adheres to support skill descriptions. The relevance results contain many draws, indicating similar performance between humans and our model in generating contextually relevant responses. See Appendix C for a comparison of our prompting approach and other methods. Overall, the model often matches or exceeds human response quality, although it does lose a small percentage (5% to 15%) of the time.

#### 5.4.2 Conversational quality

The evaluation framework proposed by Zheng et al. (2023a) was followed to evaluate the quality and dynamics of dialogues generated through the self-chat methodology by assessing the following metrics: (1) informativeness evaluates to what extent a help-seeker elaborated on the details of their distress, (2) understanding assesses how well the supporter subsequently understood these details, (3) helpfulness measures the degree to which the supporter was able to reduce the help-seeker’s emotional distress, (4) consistency ensures that both speakers followed their assigned roles, and (5) coherence assesses if conversations remained on topic and topic transitions are natural.

We rated the quality of 80 randomly sampled dialogues from both our synthetic dataset and the ESCONV dataset using a five-level Likert scale that ranges from 1 to 5. Since LLMs with strong reasoning capabilities show high agreement with human judgement on evaluation tasks, the *LLM-as-a-judge* paradigm was followed due to the high costs associated with obtaining ratings from humans (Liu et al., 2023). Therefore, we used DEEPSEEK-R1 671B, a cheap reasoning model, to replace human evaluators for rating dialogue quality (Guo et al., 2025). The used prompt can be found in Appendix B.

Table 7 compares the evaluations of dialogues from the ESCONV dataset and our synthetic dataset for the proposed metrics. Dialogues from our dataset obtain higher scores than human dialogues on all qualities, most notably on the supporter’s understanding (4.93 vs 3.60) and helpfulness (4.45 vs. 3.52). Furthermore, the informativeness scores (4.22 vs. 3.90) are similar for both datasets, indicating that human and artificial help-seekers elaborate on their distress to similar extents. The higher scores on coherence (4.97 vs. 4.23) and consistency (4.95 vs. 4.20) indicate that our dialogues are more structured than human dialogues and that the role-playing agents better adhere to their assigned roles than human speakers.



## 6 Discussion

This thesis investigated how well a LLAMA 3-8B model, when prompted to follow psychological frameworks, can produce emotional support dialogues of human-level quality. To investigate this, we formulated three subquestions: (RQ1) how can the model be prompted for context understanding to enable support skill selection that aligns with psychological frameworks for goal-oriented conversational guidance, (RQ2) how well does the model generate contextually relevant responses that adhere to emotional support skill descriptions in its prompts, and (RQ3) to what extent does the model produce dialogues comparable to human conversations when integrated into a self-chat framework.

We addressed these questions by performing experiments on a dataset of synthetic dialogues, generated using the framework described in Section 4.6. We answer RQ1 by analyzing how our approach captures conversational progression and enables strategy and skill predictions, and compare these against selections by human raters, RQ2 by evaluating response coherence using automatic metrics, while a psychology student assessed relevance and skill adherence by comparing synthetic and human responses, and RQ3 by comparing synthetic and real dialogues in terms of structure, topic and content diversity, support skill usage patterns, and overall quality, using LLM-based and human evaluations.

### 6.1 General findings

The synthetic dialogues were grounded in distress-related topics sampled from the EXTES dataset to ensure the dataset enables downstream models to generalize to diverse conversations. A PCA on TF-IDF vector representations of the dialogues showed that they cover a wider range of topics than ESCONV, with a BERTopic analysis confirming that both datasets cover similar themes (e.g., relational conflict, anxiety, coping strategies). These results suggest that our dataset has better topic diversity, while maintaining the themes of the emotional support setting.

Although dialogues display broad topic coverage, evaluations showed that the transition from providing guidance to ending the conversation frequently stagnated. This resulted in repetitive dialogues, which was reflected in pairwise cosine similarity scores of synthetic dialogues peaking at higher values than ESCONV, indicating less diverse content. Despite these shortcomings in conversational progress and diversity, the generated dialogues demonstrate a high level of quality; an analysis using DEEPSEEK-R1 671B as a rater showed that synthetic dialogues scored higher than ESCONV on all qualitative measures, especially on supporter understanding and helpfulness. These results emphasize the promise of the synthetic dataset’s utility in training downstream models to understand emotional context or provide goal-oriented guidance.

Evaluations of the dialogue planning approach confirm that capturing conversational features enables the model to postpone providing guidance until sufficient understanding is achieved. This aligns with the emotional support framework proposed by Liu et al. (2021b) and contrasts strongly with unguided LLMs, which provide guidance prematurely (Kang et al., 2024; Cheng et al., 2024). However, comforting skills remained underused, likely because (1) the comforting strategy can only be chosen after winning tie-breakers, reducing its likelihood of being selected; and (2) the model defaults to actionable skills in later conversational stages because the help-seeker’s distress is understood and resolved. This likely contributed to the earlier mentioned conversational stagnation, since use of skills like affirmation and reassurance, which help transition conversations to a conclusion, was restricted.



Furthermore, comparison of support skill usage in human and synthetic dialogues, and analyses of our confusion matrices revealed collapse of support strategies into a single support skill. For instance, when the model explored, it defaulted to using *questions* and underused *restatement and paraphrasing*. Similarly, it practically never used *providing information* when selecting actionable skills, and relied almost exclusively on *reflection of feelings* to provide comfort. These findings contradict our hypothesis in Section 4.3.5, where we proposed that a reduced reasoning scope might enable the model to directly reason about which skill is optimal. Instead, the results indicated that direct skill selection by an LLM remains challenging, suggesting that conversational features at the skill level, similar to those for strategy selection, may be necessary for decision-making. The dialogue planning analyses should be interpreted with care, since the model’s abilities may be under- or overestimated depending on how the evaluated conversational contexts align with the biases in the hierarchical skill selection process.

Human evaluations of strategy and support skill selection in truncated `ESCONV` dialogues demonstrated limited one-to-one agreement between human raters and the model, but also among human raters themselves. This suggests that support skill usage is subjective and multiple skills are appropriate for identical support contexts. F1-scores showed that the model aligned more closely with crowdworkers than with the psychology student, especially for comforting and actionable skills. Interestingly, the model’s frequency of exploration and action strategy use matched the student’s, but it underused comforting skills compared to both human raters, in line with our previous observations.

The evaluation setup could partially explain these results. The model tends to default to actionable skills in later conversational stages, but rarely selected them in human evaluations, which suggests the truncated dialogues were skewed towards earlier stages where actionable skills are not appropriate. The alignment between the student and model may therefore reflect this bias in selection of these dialogues.

In addition to the insights into dialogue planning, human evaluations demonstrated that our task decomposition helped the model generate contextually relevant responses that were accurately grounded in support skill descriptions. We also found that the approach allows the model to avoid excessively empathetic replies, a frequently observed problem in unprompted LLMs (Cheng et al., 2024). Its responses were generally as relevant as those from crowdworkers, and it achieved high win-rates on support skill accuracy, especially for comforting skills. On relevance the model had lower win-rates and a higher number of draws. This suggests that while the model likely generates responses that are contextually relevant, these may touch upon more superficial topics than those from humans. This could be caused by the model’s small size, limiting its ability to identify meaningful conversational directions.

While individual responses were generally contextually appropriate and adhered to their skill descriptions, utterance level Distinct-2/3 scores for synthetic utterances showed slightly lower lexical diversity than those in `ESCONV`, while lexical diversity at the dataset level was substantially lower for our dataset. Additionally, S-BERT cosine similarity scores indicated that synthetic utterances were more semantically similar than human ones. This combination of human-like lexical utterance diversity, limited lexical dataset diversity, and high semantic similarity suggests repetition within and between synthetic dialogues. Further evaluations confirm that when the same support skill is used successively, the model repeats earlier implementations of that skill with little to no variation, rather than introducing new perspectives or conversational directions.

## 6.2 Implications

We consider several implications of our findings on the field of dialogue system development. First, while certain aspects of the synthetic dialogues require improvements, their overall quality is considerable. Our method therefore offers a promising alternative to collecting large quantities of sensitive real-world data for training goal-oriented dialogue systems. Since proactive conversational guidance is achieved solely through prompting, the same model could be used to create data for related goal-oriented domains like persuasion, tutoring, or customer support. By modifying the prompts to align with domain-specific frameworks, additional retraining or data collection is not necessary, which reduces overall dataset curation costs and makes development of context-aware and proactive systems more broadly accessible.

Furthermore, our approach is especially valuable in low-resource domains like mental healthcare and, more generally, academic research, where political, privacy, ethical, and cost constraints can limit access to high-quality data. Because researchers and organizations with limited computational resources could host smaller-scale language models themselves, dependencies on external sources, such as proprietary models from large companies, are reduced. This democratizes dataset curation, reinforces institutional autonomy, and enables innovation in under-studied areas.

Finally, the prompting approach enables research into localized dialogue systems for other languages. By translating prompts and (if necessary) fine-tuning LLMs for specific languages, researchers could remove language barriers, making these systems globally accessible.

## 6.3 Limitations

We conclude the discussion by outlining four limitations of our study. First, the dialogue generation method is computationally inefficient. Because the model’s small size affected its contextual understanding, extensive prompting and many self-consistency iterations were required. This led to an inefficient process that limited the number of samples used during evaluations, which potentially affected the statistical power of our study and limited the effectiveness of our approach in generating large-scale datasets.

Second, although our thesis investigated how a smaller-scale LLM could be used to generate a synthetic dataset for training downstream models, time constraints and the limited size of our dataset prevented us from actually training downstream models and evaluating changes in their performance.

Third, our human evaluations involve only one rater. Support skill evaluations indirectly included a second rater through crowdworkers’ annotations, but the response (and thereby dialogue) evaluations could be biased. We aimed to address this by using a larger-scale reasoning model as a rater, but acknowledge that this does not replace the quality that multiple human experts could provide.

Fourth, the automatic metrics used to assess dialogues provided insights into the dialogues’ characteristics generally failed to capture whether they “made sense”, which is a common challenge when attempting to capture subjective qualities using automated metrics.

Future work can address these experimental limitations by improving the scalability of our method, either by reducing the complexity of the prompting approach or investigating whether larger-scale models provide significantly better contextual understanding despite a small increase in computational cost. Additionally, involving multiple human evaluators could improve the reliability and coverage of the qualitative analyses.

## 7 Conclusion

In this thesis, we proposed a prompt-based, goal-oriented synthetic dialogue generation approach using a smaller-scale open-source language model. We investigated whether this methodology could be used to generate emotional support dialogues of human-level quality aligned with a psychological framework by answering three subquestions that focused on prompt-based support skill selection and response generation, as well as the integration of these components into a self-chat framework.

We performed our experiments on a synthetic dataset that was generated using our method. The evaluations encompassed three main aspects: the alignment between the model’s and human conversational guidance, the response quality in terms of contextual relevance and adherence to skill descriptions, and the dialogues’ diversity, structure, and content. We assessed these qualities through automatic, LLM-based, and human evaluations to analyze the effectiveness of our approach.

Our findings suggest that our approach indeed enables the model to produce goal-oriented dialogues that align with the used emotional support framework, and show promising quality when compared against human dialogues. Specifically, we answer our research questions as follows. For RQ1, we found that extracting conversational features like distress level and exploration extent enables strategy selection that aligns with the support framework. This demonstrates that smaller-scale LLMs can understand context effectively through prompting, provided that the prompts offer sufficient guidance. However, using the model to directly reason about support skill selection led to “skill collapse”, where the model struggled to differentiate between skills and often defaulted to a single one. Additionally, biases in strategy selection led to stagnation in dialogue progress, resulting in repetitive conversations. For RQ2, the model adhered to emotional support skill descriptions more accurately than humans and produced responses that were contextually relevant but potentially superficial. Interestingly, when the same skill was used successively, the model reused prior implementations of that skill, thus failing to introduce new conversational directions. For RQ3, synthetic dialogues outperformed human ones in qualitative evaluations performed by DEEPSEEK-R1 671B and covered a wider range of distress topics, while maintaining alignment with common emotional support themes. Furthermore, the dialogues followed the expected transition from exploration to guidance, but lacked in comforting and had substantially lower lexical diversity, likely caused by conversational stagnation.

Based on these findings, we propose several directions for future research. One involves introducing additional conversational features at the strategy selection level, which could resolve conversational stagnation. Additionally, implementing feature analysis at the skill selection level could address skill collapse, with more realistic skill usage likely also improving response diversity, though this could alternatively be addressed by optimizing the model’s generation parameters. Another interesting direction is exploring the trade-off between higher computational costs and improved reasoning capabilities in larger-scale language models, which might improve the method’s efficiency by requiring a less complicated prompting approach and thereby lowering its total computational cost. Additionally, it remains important to evaluate how training downstream models on our synthetically generated datasets actually influences their performance. Finally, while our experiments comprised only static evaluations, future work could investigate how our dialogue planning and response generation approaches affect the distress of real users.

## References

- Althoff, Tim, Kevin Clark and Jure Leskovec (2016). “Large-scale analysis of counseling conversations: An application of natural language processing to mental health”. In: *Transactions of the Association for Computational Linguistics* 4, pp. 463–476.
- Askari, Arian, Roxana Petcu, Chuan Meng, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas and Suzan Verberne (2024). “Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking dialogs”. In: *arXiv preprint arXiv:2402.11633*.
- Azpiazu, Lorea, Iratxe Antonio-Agirre, Arantza Fernández-Zabala and Naiara Escalante (2023). “How does social support and emotional intelligence enhance life satisfaction among adolescents? A mediational analysis study”. In: *Psychology Research and Behavior Management*, pp. 2341–2351.
- Balaraman, Vevake and Bernardo Magnini (2021). “Domain-aware dialogue state tracker for multi-domain dialogue systems”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 866–873.
- Bao, Siqi, Huang He, Fan Wang, Hua Wu and Haifeng Wang (2020). “PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 85–96.
- Bilquise, Ghazala, Samar Ibrahim and Khaled Shaalan (2022). “Emotionally intelligent chatbots: a systematic literature review”. In: *Human Behavior and Emerging Technologies* 2022.1, p. 9601630.
- Broekens, Joost, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard and Aske Plaat (2023). “Fine-grained affective processing capabilities emerging from large language models”. In: *2023 11th international conference on affective computing and intelligent interaction (ACII)*. IEEE, pp. 1–8.
- Brown, Peter F, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai and Robert L Mercer (1992). “Class-based n-gram models of natural language”. In: *Computational linguistics* 18.4, pp. 467–480.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. (2020). “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33, pp. 1877–1901.
- Burkhardt, Hannah, Michael Pullmann, Thomas Hull, Patricia AreÄ;n and Trevor Cohen (July 2022). “Comparing emotion feature extraction approaches for predicting depression and anxiety”. In: *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Seattle, USA: Association for Computational Linguistics, pp. 105–115.
- Burleson, Brant R (2003). “Emotional support skills”. In: *Handbook of communication and social interaction skills*. Routledge, pp. 569–612.
- Campion, Jonathan, Afzal Javed, Crick Lund, Norman Sartorius, Shekhar Saxena, Michael Marmot, John Allan and Pichet Udomratn (2022). “Public mental health: required actions to address implementation failure in the context of COVID-19”. In: *The Lancet Psychiatry* 9.2, pp. 169–182.

- Carbonell Jr, Jaime G (1978). “POLITICS: Automated ideological reasoning”. In: *Cognitive Science* 2.1, pp. 27–51.
- Chen, Hongshen, Xiaorui Liu, Dawei Yin and Jiliang Tang (2017). “A survey on dialogue systems: Recent advances and new frontiers”. In: *Acm Sigkdd Explorations Newsletter* 19.2, pp. 25–35.
- Chen, Maximillian, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu and Dilek Hakkani-Tur (2023a). “PLACES: Prompting Language Models for Social Conversation Synthesis”. In: *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 844–868.
- Chen, Maximillian, Xiao Yu, Weiyan Shi, Urvi Awasthi and Zhou Yu (2023b). “Control-able Mixed-Initiative Dialogue Generation through Prompting”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 951–966.
- Chen, Zixiang, Yihe Deng, Huizhuo Yuan, Kaixuan Ji and Quanquan Gu (2024). “Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models”. In: *Forty-first International Conference on Machine Learning*.
- Cheng, Jiale, Sahand Sabour, Hao Sun, Zhuang Chen and Minlie Huang (2023). “PAL: Persona-Augmented Emotional Support Conversation Generation”. In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 535–554.
- Cheng, Yi, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu and Yefeng Zheng (2024). “Cooper: Coordinating specialized agents towards a complex dialogue goal”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16, pp. 17853–17861.
- Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio (2014). “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078*.
- Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg and Dario Amodei (2017). “Deep reinforcement learning from human preferences”. In: *Advances in neural information processing systems* 30.
- Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade and Sujith Ravi (July 2020). “GoEmotions: A Dataset of Fine-Grained Emotions”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4040–4054.
- Deng, Yang, Wenqiang Lei, Wai Lam and Tat-Seng Chua (2023a). “A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects”. In: *IJCAI*, pp. 6583–6591.
- Deng, Yang, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei and Tat-Seng Chua (2023b). “Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration”. In: *arXiv preprint arXiv:2305.13626*.
- Deng, Yang, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam and Tat-Seng Chua (2025). “Proactive conversational ai: A comprehensive survey of advancements and opportunities”. In: *ACM Transactions on Information Systems* 43.3, pp. 1–45.
- Deng, Yang, Lizi Liao, Zhonghua Zheng, Grace Hui Yang and Tat-Seng Chua (2024a). “Towards human-centered proactive conversational agents”. In: *Proceedings of the 47th*

- International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 807–818.
- Deng, Yang, Wenxuan Zhang, Wai Lam, See-Kiong Ng and Tat-Seng Chua (2024b). “Plug-and-play policy planner for large language model powered dialogue agents”. In: *The Twelfth International Conference on Learning Representations*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186.
- Ding, Xiruo, Kevin Lybarger, Justin Tauscher and Trevor Cohen (July 2022). “Improving Classification of Infrequent Cognitive Distortions: Domain-Specific Model vs. Data Augmentation”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, pp. 68–75.
- Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li and Zhifang Sui (2024). “A Survey on In-context Learning”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1107–1128.
- Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan et al. (2024). “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783*.
- Egger, Roman and Joanne Yu (2022). “A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts”. In: *Frontiers in sociology* 7, p. 886498.
- Fei, Hao, Bobo Li, Qian Liu, Lidong Bing, Fei Li and Tat-Seng Chua (2023). “Reasoning Implicit Sentiment with Chain-of-Thought Prompting”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1171–1182.
- Goddeau, David, Helen Meng, Joseph Polifroni, Stephanie Seneff and Senis Busayapongchai (1996). “A form-based dialogue manager for spoken language applications”. In: *Proceeding of Fourth International Conference on Spoken Language Processing. IC-SLP’96*. Vol. 2. IEEE, pp. 701–704.
- Grootendorst, Maarten (2022). “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. In: *arXiv preprint arXiv:2203.05794*.
- Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi et al. (2025). “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning”. In: *arXiv preprint arXiv:2501.12948*.
- He, Qianyu, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang and Yanghua Xiao (2024a). “Can Large Language Models Understand Real-World Complex Instructions?” In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.16, pp. 18188–18196.

- He, Tao, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen and Bing Qin (2024b). “Planning Like Human: A Dual-process Framework for Dialogue Planning”. In: *arXiv preprint arXiv:2406.05374*.
- Hill, Clara E (2020). *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.
- Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.
- Huang, Minlie, Xiaoyan Zhu and Jianfeng Gao (2020). “Challenges in building intelligent open-domain dialog systems”. In: *ACM Transactions on Information Systems (TOIS)* 38.3, pp. 1–32.
- Ji, Shaoxiong, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari and Erik Cambria (June 2022). “MentalBERT: Publicly Available Pretrained Language Models for Mental Health-care”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 7184–7190.
- Kang, Dongjin, Sunghwan Mac Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee and Jinyoung Yeo (2024). “Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15232–15261.
- Lee, Chris van der, Albert Gatt, Emiel van Miltenburg and Emiel Krahmer (2021). “Human evaluation of automatically generated text: Current trends and best practice guidelines”. In: *Computer Speech & Language* 67, p. 101151.
- Li, Jingling, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi and Yang Liu (2024). “Steering llms towards unbiased responses: A causality-guided debiasing framework”. In: *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
- Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao and Bill Dolan (2016). “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 110–119.
- Liang, Yiming, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, Wenhao Huang and Jiajun Zhang (2025). “I-SHEEP: Self-Alignment of LLM from Scratch through an Iterative Self-Enhancement Paradigm”. In: *Submitted to ACL Rolling Review - December 2024*. under review.
- Liu, Jianfeng, Feiyang Pan and Ling Luo (2020). “Gochat: Goal-oriented chatbots with hierarchical reinforcement learning”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1793–1796.
- Liu, Qiaolan, Min Jiang, Shiyong Li and Yang Yang (2021a). “Social support, resilience, and self-esteem protect against common mental health problems in early adolescence: A nonrecursive analysis from a two-year longitudinal study”. In: *Medicine* 100.4, e24334.
- Liu, Siyang, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu and Minlie Huang (2022). “Rethinking and Refining the Distinct Metric”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 762–770.

- Liu, Siyang, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang and Minlie Huang (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)*. Association for Computational Linguistics, pp. 3469–3483.
- Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu and Chenguang Zhu (2023). “G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 2511–2522.
- McTear, Michael F (2002). “Spoken dialogue technology: enabling the conversational user interface”. In: *ACM Computing Surveys (CSUR)* 34.1, pp. 90–169.
- Medeiros, Lenin, Charlotte Gerritsen and Tibor Bosse (2019). “Towards humanlike chatbots helping users cope with stressful situations”. In: *Computational Collective Intelligence: 11th International Conference, ICCCI 2019, Hendaye, France, September 4–6, 2019, Proceedings, Part I 11*. Springer, pp. 232–243.
- Merialdo, Bernard (1994). “Tagging English text with a probabilistic model”. In: *Computational linguistics* 20.2, pp. 155–171.
- Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray et al. (2022). “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35, pp. 27730–27744.
- Rashkin, Hannah, Eric Michael Smith, Margaret Li and Y-Lan Boureau (2019). “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 5370–5381.
- Rastogi, Abhinav, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta and Pranav Khaitan (2020). “Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 8689–8696.
- Reimers, Nils and Iryna Gurevych (2019). “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084*.
- Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau and Jason Weston (2021). “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, pp. 300–325.
- Sabour, Sahand, Chujie Zheng and Minlie Huang (2022). “Cem: Commonsense-aware empathetic response generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10, pp. 11229–11237.
- Schyff, Emma L van der, Brad Ridout, Krestina L Amon, Rowena Forsyth and Andrew J Campbell (2023). “Providing self-led mental health support through an artificial intelligence-powered chat bot (Leora) to Meet the demand of mental health care”. In: *Journal of Medical Internet Research* 25, e46448.



- Sharma, Ashish, Adam Miner, David Atkins and Tim Althoff (Nov. 2020). “A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5263–5276.
- Shortliffe, Edward H (1974). “A rule-based computer program for advising physicians regarding antimicrobial therapy selection”. In: *Proceedings of the 1974 annual ACM conference-Volume 2*, pp. 739–739.
- Sun, Xiaofei, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang and Guoyin Wang (2023a). “Text Classification via Large Language Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, pp. 8990–9005.
- Sun, Xiaofei, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang and Guoyin Wang (2023b). “Sentiment analysis through llm negotiations”. In: *arXiv preprint arXiv:2311.01876*.
- Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng and Huan Liu (Nov. 2024). “Large Language Models for Data Annotation and Synthesis: A Survey”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 930–957.
- Tiwari, Abhisek, Sriparna Saha and Pushpak Bhattacharyya (2022). “A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning”. In: *Knowledge-Based Systems* 242, p. 108292.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar et al. (2023). “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971*.
- Tu, Quan, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen and Rui Yan (May 2022). “MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 308–319.
- Vaidyam, Aditya Nrusimha, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan and John Blake Torous (2019). “Chatbots and conversational agents in mental health: a review of the psychiatric landscape”. In: *The Canadian Journal of Psychiatry* 64.7, pp. 456–464.
- Vaswani, A (2017). “Attention is all you need”. In: *Advances in Neural Information Processing Systems*.
- Väth, Dirk, Lindsey Vanderlyn and Ngoc Thang Vu (2023). “Conversational Tree Search: A New Hybrid Dialog Task”. In: Version Number: 1. arXiv.
- Wang, Hongru, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu and Kam-Fai Wong (2023a). “Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore, pp. 12047–12064.

- Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery and Denny Zhou (2023b). “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: *The Eleventh International Conference on Learning Representations*.
- Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi and Hannaneh Hajishirzi (2023c). “Self-Instruct: Aligning Language Models with Self-Generated Instructions”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 13484–13508.
- Wang, Zengzhi, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang and Rui Xia (2024). “Is ChatGPT a Good Sentiment Analyzer?” In: *First Conference on Language Modeling*.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le and Denny Zhou (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., pp. 24824–24837.
- Weizenbaum, Joseph (1966). “ELIZA—a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1, pp. 36–45.
- Welivita, Anuradha, Chun-Hung Yeh and Pearl Pu (2023). “Empathetic Response Generation for Distress Support”. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, pp. 632–644.
- Winograd, Terry (1971). “Procedures as a representation for data in a computer program for understanding natural language”. In:
- Woods, William A (1970). “Transition network grammars for natural language analysis”. In: *Communications of the ACM* 13.10, pp. 591–606.
- Wu, Zhaofeng, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Na-joung Kim, Jacob Andreas and Yoon Kim (2024). “Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1819–1862.
- Xie, Yubo and Pearl Pu (2021). “Empathetic Dialog Generation with Fine-Grained Intents”. In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 133–147.
- Xing, Chen, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou and Wei-Ying Ma (2017). “Topic aware neural response generation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1.
- Xiong, Sichao, Yigit Ihlamur, Fuat Alican and Aaron Ontoyin Yin (2024). “GPTree: Towards Explainable Decision-Making via LLM-powered Decision Trees”. In: *arXiv pre-print arXiv:2411.08257*.
- Xu, Canwen, Daya Guo, Nan Duan and Julian McAuley (2023). “Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data”. In: *Proceedings of the*

- 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 6268–6278.
- Yu, Xiao, Maximillian Chen and Zhou Yu (2023). “Prompt-Based Monte-Carlo Tree Search for Goal-oriented Dialogue Policy Planning”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 7101–7125.
- Yuan, Weizhe, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu and Jason E Weston (2024). “Self-Rewarding Language Models”. In: *Forty-first International Conference on Machine Learning*.
- Zhang, Haode, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Albert Y.S. Lam and Xiao-Ming Wu (Nov. 2021). “Effectiveness of Pre-training for Few-shot Intent Classification”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1114–1120.
- Zhang, Justine and Cristian Danescu-Niculescu-Mizil (2020). “Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5276–5289.
- Zhang, Qiang, Jason Naradowsky and Yusuke Miyao (2023). “Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models”. In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6665–6694.
- Zhang, Yiqun, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song and Ge Yu (Jan. 2024). “Affective Computing in the Era of Large Language Models: A Survey from the NLP Perspective”. en. In: *CoRR*.
- Zheng, Chujie, Sahand Sabour, Jiaxin Wen, Zheng Zhang and Minlie Huang (2023a). “AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, pp. 1552–1568.
- Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez and Ion Stoica (Dec. 2023b). “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. en. In: *Advances in Neural Information Processing Systems* 36, pp. 46595–46623.
- Zheng, Zhonghua, Lizi Liao, Yang Deng, Libo Qin and Liqiang Nie (2024). “Self-chats from Large Language Models Make Small Emotional Support Chatbot Better”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11325–11345.
- Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le et al. (2023). “Least-to-Most Prompting Enables Complex Reasoning in Large Language Models”. In: *The Eleventh International Conference on Learning Representations*.

## A Response repetition

- [59] Supporter (*Providing Suggestions*): “Hmm, maybe exploring flexible work arrangements or delegating tasks to colleagues could be a good starting point. What do you think?”
  - [61] Supporter (*Providing Suggestions*): “Hmm, maybe exploring flexible work arrangements or delegating tasks to colleagues could be a good starting point. What do you think?” (similarity=1.0000)
  - [65] Supporter (*Providing Suggestions*): “Hmm, maybe exploring flexible work arrangements or delegating tasks to colleagues could be a good starting point.” (similarity=0.9736)
  - [67] Supporter (*Providing Suggestions*): “Hmm, maybe exploring flexible work arrangements or delegating tasks to colleagues could be a good starting point.” (similarity=0.9736)
  - [63] Supporter (*Providing Suggestions*): “That sounds like a good starting point. Have you considered talking to your supervisor about flexible work arrangements or delegating tasks to colleagues?” (similarity=0.8340)
- [88] Help-Seeker: “That sounds like a good idea. I think it could really help me gain some insight into my decision-making process.”
  - [90] Help-Seeker: “Yeah, that sounds like a good idea. I think it could really help me gain some insight into my decision-making process.” (similarity=0.9951)
  - [86] Help-Seeker: “That sounds like a good idea. I think it could really help me clarify my thoughts and gain confidence in my decision-making abilities.” (similarity=0.9269)
  - [82] Help-Seeker: “That sounds like a great idea. I think it could really help me gain some confidence in my decision-making abilities.” (similarity=0.8840)
  - [81] Supporter (*Providing Suggestions*): “What do you think about trying a small decision journal’ to track your choices and see how it feels?” (similarity=0.7184)
- [146] Supporter (*Providing Suggestions*): “Maybe you could try framing your concerns as a genuine inquiry, like ‘Hey, I’ve noticed we haven’t talked about [specific topic] in a while. Is everything okay?’ or ‘I’ve been feeling a bit uncertain about our friendship lately. Can we talk about what’s going on?’ What do you think about that?”
  - [152] Supporter (*Providing Suggestions*): “Maybe framing your concerns as a genuine inquiry, like ‘Hey, I’ve noticed we haven’t talked about [specific topic] in a while. Is everything okay?’ or ‘I’ve been feeling a bit uncertain about our friendship lately. Can we talk about what’s going on?’” (similarity=0.9944)
  - [148] Supporter (*Providing Suggestions*): “Maybe we could try framing your concerns as a genuine inquiry, like... ‘Hey, I’ve noticed we haven’t talked about [specific topic] in a while. Is everything okay?’ or... ‘I’ve been feeling a bit uncertain about our friendship lately. Can we talk about what’s going on?’” (similarity=0.9898)

- [142] Supporter (*Providing Suggestions*): “Have you thought about framing your concerns as a genuine inquiry, like ‘Hey, I’ve noticed we haven’t talked about [specific topic] in a while. Is everything okay?’ It might help your friend feel more comfortable opening up.” (similarity=0.8635)
- [136] Supporter (*Providing Suggestions*): “Have you considered... reaching out to your friend and asking if everything’s okay, or... sharing your concerns with them in a non-accusatory way?” (similarity=0.6871)
- [149] Help-Seeker: “Yeah, that sounds like a good idea. I’ll try that. Thank you for your help.”
  - [970] Help-Seeker: “Yeah, that sounds like a good idea. I’ll try that. Thanks for your help.” (similarity=0.9944)
  - [434] Help-Seeker: “Yeah, that’s a great idea. I’ll try that. Thanks for your help.” (similarity=0.9791)
  - [968] Help-Seeker: “That sounds like a good idea. I’ll try that. Thanks for your help.” (similarity=0.9754)
  - [1305] Help-Seeker: “That sounds like a good idea. I’ll try it. Thanks for your help.” (similarity=0.9322)
- [158] Supporter (*Restatement or Paraphrasing*): “So, it seems that your friend has stopped responding to your calls and messages, and you’re left feeling ignored and unsure of what’s changed in their life that could have led to this sudden silence.”
  - [160] Supporter (*Restatement or Paraphrasing*): “Your friend has stopped responding to your calls and messages, and you’re left feeling ignored and unsure of what’s changed in their life that could have led to this sudden silence.” (similarity=0.9922)
  - [155] Help-Seeker: “My friend has been ignoring my calls and messages. We used to talk regularly, but suddenly they stopped responding.” (similarity=0.8114)
  - [156] Supporter (*Question*): “What do you think might have changed in your friend’s life or priorities that could have led to this sudden silence?” (similarity=0.7457)
  - [164] Supporter (*Reflection of Feelings*): “It sounds like you’re feeling a deep sense of rejection and abandonment, like your friend’s silence is a rejection of your friendship and your worth as a person. Is that a feeling that’s been lingering for you, or is it a new one that’s emerged in this situation?” (similarity=0.7059)
- [192] Supporter (*Providing Suggestions*): “That’s an interesting idea. Have you considered searching for online communities or social media groups centered around your interests? Maybe that could be a good starting point for connecting with others who share your passions.”
  - [188] Supporter (*Providing Suggestions*): “That’s an interesting idea. Have you considered searching for online communities or social media groups centered around your interests? Maybe that could be a good starting point for connecting with others who share your passions.” (similarity=1.0000)
  - [1261] Supporter (*Providing Suggestions*): “That’s an interesting idea. What do you think about exploring a book club or a similar group, or even attending a networking event to connect with others who share similar passions?” (similarity=0.8946)

- 
- [190] Supporter (*Providing Suggestions*): “Have you considered exploring on-line communities or social media groups centered around your interests? Maybe that could be a good starting point for connecting with others who share your passions.” (similarity=0.8800)
  - [186] Supporter (*Providing Suggestions*): “That sounds like a really tough spot to be in. Have you considered trying online communities or social media groups centered around your interests? Maybe that could be a good starting point for connecting with others who share your passions.” (similarity=0.7948)
  - [279] Supporter (*Restatement or Paraphrasing*): “So, you’re facing a high volume of tasks, which you’re struggling to manage, and this is causing you to worry about not meeting expectations and disappointing others.”
    - [277] Supporter (*Restatement or Paraphrasing*): “So, it seems that you’re facing a high volume of tasks, which you’re struggling to manage, and this is causing you to worry about not meeting expectations and disappointing others.” (similarity=0.9953)
    - [685] Supporter (*Restatement or Paraphrasing*): “So, you’re feeling overwhelmed by the demands of your responsibilities, and you’re struggling to keep up with multiple tasks, which is causing you to feel stuck and unable to make progress.” (similarity=0.7385)
    - [1061] Supporter (*Affirmation and Reassurance*): “You’re not alone in feeling overwhelmed, and it’s normal to worry about not meeting expectations. Your motivation and capabilities are not defined by one setback or failure.” (similarity=0.6915)
    - [1084] Supporter (*Providing Suggestions*): “Have you considered prioritizing tasks, setting smaller goals, and seeking guidance from colleagues or supervisors?” (similarity=0.6869)
  - [313] Supporter (*Providing Suggestions*): “Have you considered exploring a ‘gray area’ approach, where you both try to find common ground?”
    - [311] Supporter (*Providing Suggestions*): “Have you considered exploring the ‘gray area’ approach, where you both try to find common ground?” (similarity=0.9957)
    - [34] Help-Seeker: “I think we’re not communicating effectively. We both have different opinions, and we don’t know how to find common ground.” (similarity=0.7760)
    - [257] Supporter (*Providing Suggestions*): “Have you considered having a conversation about active listening and trying to understand each other’s perspectives?” (similarity=0.7719)
    - [309] Supporter (*Providing Suggestions*): “What if you tried seeking common ground or exploring a ‘gray area’ approach?” (similarity=0.7398)
  - [331] Help-Seeker: “I think that could be really helpful. It might help us stay on track and avoid getting sidetracked by other issues. And maybe we can also try to listen more actively to each other, instead of just focusing on our own points.”
    - [335] Help-Seeker: “I think that could be really helpful. It might help us stay on track and avoid getting sidetracked by other issues. And maybe we can also try

- to listen more actively to each other, instead of just focusing on our own points.” (similarity=1.0000)
- [329] Help-Seeker: “I think that could be really helpful. It might help us stay on track and avoid getting sidetracked by other issues.” (similarity=0.7984)
  - [332] “That’s a great idea. Maybe we could also explore having a ‘pre-conversation’ to clarify the issues and establish a clear agenda for the discussion?” (similarity=0.7804)
  - [312] Help-Seeker: “I think that’s a great idea. I never thought about trying to find common ground instead of just focusing on our differences.” (similarity=0.7741)
- [669] Supporter (*Providing Suggestions*): “Maybe we could start by identifying one small, achievable goal you’d like to work towards. Just one step at a time, and we can explore that together.”
    - [671] Supporter (*Providing Suggestions*): “Maybe we could start by identifying one small, achievable goal you’d like to work towards. Just one step at a time, and we can explore that together.” (similarity=1.0000)
    - [1605] Supporter (*Providing Suggestions*): “Perhaps you could start by exploring what advice or support you do feel capable of providing?” (similarity=0.6153)
    - [497] Supporter (*Providing Suggestions*): “That sounds like a great place to start. Perhaps we could explore some specific phrases or questions you could use to clarify your needs and desires. What do you think?” (similarity=0.6126)
    - [582] Supporter (*Providing Suggestions*): “Perhaps we could explore some small steps you could take to start moving forward, like setting aside time for self-reflection or reaching out to friends for support.” (similarity=0.6009)
  - [687] Supporter (*Reflection of Feelings*): “So, it sounds like you’re feeling overwhelmed by multiple responsibilities, and you’re struggling to make progress, which is causing you to feel stuck and frustrated.”
    - [689] Supporter (*Reflection of Feelings*): “So, it sounds like you’re feeling overwhelmed by multiple responsibilities, and you’re struggling to make progress, which is causing you to feel stuck and frustrated.” (similarity=1.0000)
    - [683] Supporter (*Reflection of Feelings*): “So, it sounds like you’re feeling overwhelmed by the demands of your responsibilities, and that’s causing you to feel stuck and unable to make progress.” (similarity=0.8933)
    - [685] Supporter (*Reflection of Feelings*): “So, you’re feeling overwhelmed by the demands of your responsibilities, and you’re struggling to keep up with multiple tasks, which is causing you to feel stuck and unable to make progress.” (similarity=0.8867)
    - [695] Supporter (*Reflection of Feelings*): “You’re feeling overwhelmed and unsure about how to move forward. Can you tell me more about what’s causing this frustration and helplessness?” (similarity=0.7623)
  - [882] Supporter (*Providing Suggestions*): ““Perhaps taking a break from the arguments and having a calm conversation about what’s bothering you both could be a step towards reconnecting?”

- [884] Supporter (*Providing Suggestions*): ““Perhaps taking a break from the arguments and having a calm conversation about what’s bothering you both could be a step towards reconnecting?” (similarity=1.0000)
- [1288] Supporter (*Reflection of Feelings*): “You’re feeling stuck and unable to move forward, reliving the pain and hurt of the argument. That sounds really painful and stuck. Can you tell me more about what it’s like to be stuck in this place?” (similarity=0.7066)
- [1446] Help-Seeker: “We just have trouble understanding each other and communicating effectively. We always seem to be on different pages, and it’s led to a lot of conflict and frustration.” (similarity=0.6338)
- [567] Supporter (*Reflection of Feelings*): “It sounds like you’re looking for a way to reconnect and clear the air. Have you considered reaching out to them with a calm and open message, expressing your feelings and concerns, and asking if they’re willing to listen and work together to find a resolution?” (similarity=0.6294)
- [933] Help-Seeker: “That’s a good idea. I never thought about it that way. Maybe I can find new ways to connect with people and create new routines.”
  - [937] Help-Seeker: “That’s a great idea. I never thought about it that way. Maybe I can find new ways to connect with people and create new routines.” (similarity=0.9906)
  - [187] Help-Seeker: “That’s a great idea. I never thought of that. Maybe I can find some people who share my interests and connect with them that way.” (similarity=0.7503)
  - [191] Help-Seeker: “I think that’s a great idea. I never thought of that. Maybe I can find some people who share my interests and connect with them that way.” (similarity=0.7448)
  - [775] Help-Seeker: “I think that could be a good way to start reconnecting with myself and finding a sense of purpose here.” (similarity=0.6844)
- [940] Supporter (*Providing Suggestions*): “Maybe exploring local community centers or volunteer opportunities could be a good starting point?”
  - [938] Supporter (*Providing Suggestions*): “Perhaps exploring local community centers or volunteer opportunities could be a good starting point?” (similarity=0.9975)
  - [774] Supporter (*Providing Suggestions*): “That’s an interesting thought. Have you considered exploring local volunteer opportunities or community groups that align with your passions?” (similarity=0.7434)
  - [772] Supporter (*Providing Suggestions*): “That sounds like a powerful goal. Have you considered exploring local volunteer opportunities or joining a community group that aligns with your passions?” (similarity=0.7202)
  - [1202] Supporter (*Providing Suggestions*): “Perhaps you could also look into local animal shelters or rescue organizations to explore volunteer opportunities?” (similarity=0.7158)
- [993] Supporter (*Providing Suggestions*): “That sounds like a beautiful idea. Would you like to explore some creative ways to honor your dog’s memory, like writing, drawing, or making a memory box?”



- [987] Supporter (*Providing Suggestions*): “That’s a beautiful idea. Would you like to explore some creative ways to honor your dog’s memory, like writing, drawing, or making a memory box?” (similarity=0.9936)
- [786] Supporter (*Providing Suggestions*): “That’s an interesting idea. Have you considered making a memory book or a small ritual to honor your dog’s memory?” (similarity=0.8933)
- [788] Supporter (*Providing Suggestions*): “What do you think about creating a memory book or a small ritual to honor your dog’s memory?” (similarity=0.8307)
- [985] Supporter (*Providing Suggestions*): “Have you considered creating a memory book or a scrapbook to honor your dog’s memory and process your emotions?” (similarity=0.8265)
- [1067] Supporter (*Providing Suggestions*): “Perhaps exploring prioritization techniques or delegating tasks could help you regain control. What do you think?”
  - [1065] Supporter (*Providing Suggestions*): “Maybe exploring prioritization techniques or delegating tasks could help you regain control. What do you think?” (similarity=0.9990)
  - [1069] Supporter (*Providing Suggestions*): “Maybe considering a stop doing’ list or delegating tasks to others could help you regain control. What do you think?” (similarity=0.9117)
  - [1234] Supporter (*Providing Suggestions*): “Have you considered breaking down your tasks into smaller, manageable chunks? That might help you regain a sense of control and confidence.” (similarity=0.7998)
  - [1607] Supporter (*Providing Suggestions*): “Perhaps taking small, specific actions could help you feel more in control?” (similarity=0.7079)
- [1226] Supporter (*Question*): “What’s been the most challenging part of your day/week/lately that’s contributing to this feeling of being overwhelmed?”
  - [1116] Supporter (*Question*): “What’s been the most challenging part of your day/week lately that’s contributing to this feeling of being overwhelmed?” (similarity=0.9954)
  - [1182] Supporter (*Question*): “What’s been the most challenging aspect of your life lately that’s contributing to this feeling of being overwhelmed?” (similarity=0.9494)
  - [504] Supporter (*Question*): “What’s been the most challenging part of your day/week lately that’s contributing to this feeling of overwhelm?” (similarity=0.9269)
  - [463] Supporter (*Question*): “What’s been on your mind lately that’s causing you to feel so overwhelmed?” (similarity=0.8747)
- [1284] Supporter (*Providing Suggestions*): “Maybe we could explore writing a heartfelt letter or having a calm conversation... or maybe something else entirely?”
  - [1282] Supporter (*Providing Suggestions*): “Maybe we could explore writing a heartfelt letter or having a calm conversation... or maybe something else entirely?” (similarity=1.0000)

- 
- [1280] Supporter (*Providing Suggestions*): “Maybe we could explore some potential steps you could take to reach out, like... writing a heartfelt letter or having a calm conversation. What do you think?” (similarity=0.7126)
  - [1455] Supporter (*Providing Suggestions*): “Maybe experimenting with setting aside dedicated time for open conversations, focusing on empathic listening, or sharing small gratitudes could bring a sense of connection and understanding. How do you feel about that?” (similarity=0.6601)
  - [866] Supporter (*Providing Suggestions*): “If you feel comfortable, maybe we could explore some guided meditations or writing prompts to help you process your emotions?” (similarity=0.6545)
- [1372] Help-Seeker: “Yeah, I think that’s a good idea. I’ll try journaling and see if it helps me process my emotions better.”
    - [1366] Help-Seeker: “I think that’s a good idea. I’ll try journaling and see if it helps me process my emotions better.” (similarity=0.9931)
    - [587] Help-Seeker: “I think that could help. I’ve been trying to talk to my friends, but it’s hard for them to understand what I’m going through. Maybe writing things down will help me process my emotions better.” (similarity=0.8035)
    - [591] Help-Seeker: “I think that could be a good way to start processing my emotions and moving forward.” (similarity=0.7686)
    - [866] Supporter (*Providing Suggestions*): “If you feel comfortable, maybe we could explore some guided meditations or writing prompts to help you process your emotions?” (similarity=0.7569)

## B Qualitative LLM evaluation prompt

Prompt used for qualitative evaluations using DEEPSEEK-R1:

```
### Dialogues:  
{conversation}
```

```
## Instruction:
```

```
Rate the informativeness, understanding, helpfulness, consistency  
and coherence of the input dialogue on a scale of 1 to 5 and  
just output the corresponding ratings. Informativeness measures  
how well the individual seeking support articulates their  
emotional challenges. Understanding gauges the supporter's grasp  
of the individual's experiences and emotions. Helpfulness  
evaluates the effectiveness of the supporter's efforts in  
mitigating the individual's emotional distress. Consistency  
ensures participants consistently adhere to their roles and  
exhibit non-contradictory behavior. Coherence checks if  
conversations have seamless topic transitions.
```

```
### Output Format:
```

```
informativeness - x  
understanding - x  
helpfulness - x  
consistency - x  
coherence - x
```

```
### Your Response:
```

## C Differences in response generation between prompts

Small example demonstrating how model responds to a conversational context as the supporter, when given (1) no instructions on how to respond, (2) asking a question following the prompting methodology proposed by [Chen et al. \(2023b\)](#), and (3) asking a question using our prompting method.

Help-Seeker: "Hello"

Supporter: "Hello, what would you like to talk about?"

Help-Seeker: "I am having a lot of anxiety about quitting my current job. It is too stressful, but pays well."

Supporter: "What makes your job stressful for you?"

Help-Seeker: "I have to deal with many in hard financial situations, and it is upsetting."

Supporter: "Do you help your clients to make it to a better financial situation?"

Help-Seeker: "I do, but often they are not going to get back to what they want. Many people are going to lose their home when safeguards are lifted."

Supporter: "But you offer them a better future than what they have currently. It may not be what they wanted, but it helps them in the long run."

Help-Seeker: "That is true but sometimes I feel like I should put my feelings and health first."

**[Unprompted]** "It sounds like you're feeling torn between your sense of responsibility towards your clients and your own well-being. It's completely understandable that you're experiencing anxiety about quitting your job, especially since it pays well. Can you tell me more about what's driving your desire to leave your job? Is it the constant stress and emotional toll it's taking on you, or is there something else at play? Also, I want to acknowledge that your work can be incredibly challenging, and it's natural to feel overwhelmed by the financial struggles of your clients. It's admirable that you're committed to helping them, but it's also important to recognize that you can't pour from an empty cup. What do you think would need to change in your life for you to feel more balanced and fulfilled?"

**[Chen et al. prompting]** "I see. It sounds like you're feeling a sense of responsibility towards your clients, which is admirable. But at the same time, you're also feeling overwhelmed and stressed out by the situation. Can you tell me more about what you mean by 'putting your feelings and health first'? What does that look like to you?"

**[Our prompting]** "Can you tell me more about what's driving the sense that your feelings and health are being neglected, especially in moments when you're weighing the pros and cons of staying in your job?"

## D Human skill and response evaluation interface

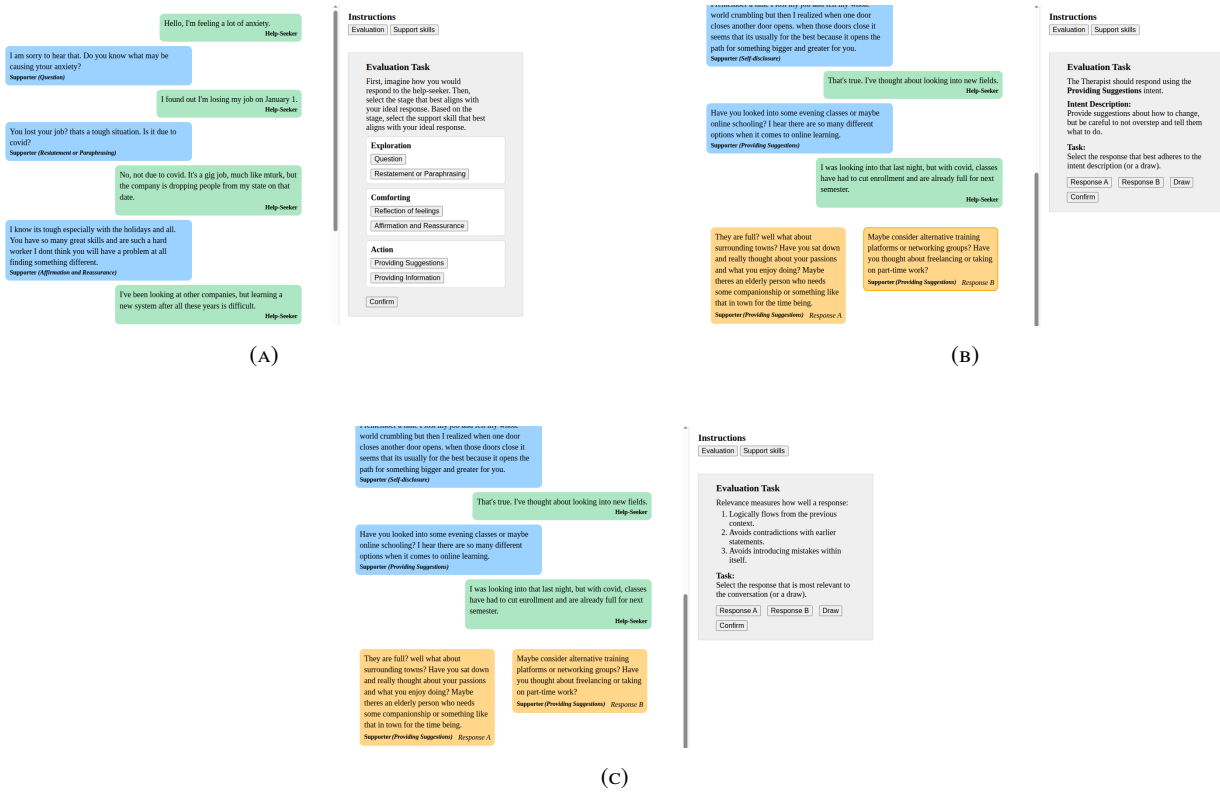


FIGURE 9: (a) Optimal support skill selection. (b) Response adherence to skill description. (c) Response relevance evaluation.