



Universiteit  
Leiden  
The Netherlands

# Bachelor Data Science & Artificial Intelligence

Questionnaire-driven Dialogue:  
Utilizing Large Language Models for Hallucination-free Conversational AI  
in Elderly Well-being Monitoring

Ruzanna Baghdasaryan  
s3068021

Supervisors:  
Prof.dr. M.R. Spruit & Dr. A.E.J.L. Lefebvre

BACHELOR THESIS  
Leiden Institute of Advanced Computer Science (LIACS)  
[www.liacs.leidenuniv.nl](http://www.liacs.leidenuniv.nl)

27/08/2024

## Abstract

Conversational AI is a promising solution to mitigate the growing shortage of human caregivers for elderly people. However, AI systems, and LLMs in particular often produce hallucinations and inaccurate responses. This thesis proposes the use of a questionnaire-driven conversational AI system to enhance the accuracy and relevance of dialogues in elderly well-being monitoring. The system employs a mixed-initiative dialogue framework that allows for smooth transitions between chit-chat and task-oriented dialogues, minimizing hallucinations. Current methods for hallucination mitigation and personalized dialogue are reviewed through a systematic literature review and the base program is tested after development. Results indicate that a questionnaire-driven approach reduces hallucinations, providing a promising path toward a more reliable and user-friendly experience for elderly users.

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Statement	1
1.2	Thesis overview	2
<b>2</b>	<b>Background Knowledge</b>	<b>3</b>
2.1	Welzijn.AI	3
2.2	Hallucinations	5
<b>3</b>	<b>Method</b>	<b>6</b>
3.1	Literature Review	6
3.2	Base Program	8
3.3	Evaluation	10
<b>4</b>	<b>Results</b>	<b>12</b>
4.1	Systematic Literature Review	12
4.1.1	Hallucinations and knowledge	12
4.1.2	Focus and control	15
4.1.3	Dialogue planning	18
4.1.4	Different approach	20
4.1.5	Unifying task and chitchat skills	21
4.1.6	Persona-based dialogue	22
4.1.7	Emotional support through chit-chat	26
4.2	Base Program Evaluation	28
4.3	Combining Literature and Software	31
<b>5</b>	<b>Conclusions and Discussion</b>	<b>33</b>
5.1	Conclusion	33
5.2	Discussion and Further Research	35

# 1 Introduction

## 1.1 Problem Statement

The intersection of healthcare and technology is evolving rapidly, driven by the need to enhance the quality and accessibility of medical care in the face of growing demands. Technological innovations are increasingly recognized for their potential to transform healthcare delivery, particularly by optimizing operational efficiency and expanding the reach of care services. In this context, the focus of this thesis is on the application of technology to support elderly individuals, a demographic whose unique needs have become increasingly prominent as the global population ages.

Older adults frequently encounter a range of challenges, including chronic health conditions, reduced mobility and social isolation. These factors contribute to a decline in their overall quality of life, making it imperative to develop interventions that can provide both practical assistance and emotional support. Technology offers promising approaches for addressing these issues, particularly through systems designed to monitor health status and provide companionship, thereby alleviating some of the burdens faced by elderly individuals and their caregivers.

The Welzijn.AI program was established with these objectives in mind. It is designed to monitor the well-being of elderly individuals continuously, using the EQ-5D questionnaire as a key assessment tool. The EQ-5D, a widely recognized instrument, evaluates five critical dimensions of health: mobility, self-care, daily activities, pain/discomfort, and anxiety/depression, in addition to an overall health rating. The program's innovative approach involves integrating these assessments into a user-friendly, AI-powered companion, such as an interactive teddy bear or furry cat, that engages with elderly users on a daily basis.

This technological companion not only administers health assessments but also engages in casual conversation, offering a blend of monitoring and companionship that is crucial for the well-being of older adults. However, the development of such a system presents notable challenges, particularly related to the use of large language models (LLMs). A significant concern with LLMs is their tendency to produce inaccurate or irrelevant content, a problem known as hallucinations. These hallucinations can take various forms, each posing a threat to the coherence and reliability of interactions.

The research question guiding this study is: ***How can a questionnaire-driven approach effectively guide dialogue generation in Large Language Models to ensure coherence and mitigate the risk of hallucinations in conversational AI systems designed for elderly well-being monitoring?*** This question is addressed through a systematic literature review to examine current methods for mitigating hallucinations and enhancing the personalization of AI-driven conversations. The goal of this review is to provide other researchers with an overview that can be used in further developing the Welzijn.AI program.

Furthermore, this thesis aims to develop a foundational chat-based system for Welzijn.AI. This system will incorporate essential features such as question-asking, small talk capabilities, and the ability to connect open-ended responses to predefined multiple-choice options. By focusing on the structure provided by the questionnaire, the study aims to create a reliable base program that

minimizes the risk of hallucinations while maintaining meaningful and coherent interactions.

In summary, this thesis explores the potential of technology to enhance the care and quality of life for elderly individuals by providing continuous health monitoring and companionship. The development of a reliable base program for the Welzijn.AI program marks a significant step toward achieving this goal, laying the groundwork for future enhancements that can further support the well-being of older adults.

## **1.2 Thesis overview**

The structure of this thesis is organized as follows: Section 2 presents the background of the study, focusing on the concepts of technology in healthcare and the issue of hallucinations in LLMs. In Section 3 the methodologies used for conducting the systematic literature review, developing the base program and evaluating its effectiveness are detailed. Section 4 discusses the findings from the systematic literature review and the evaluation results of the base program. Finally, Section 5 concludes this thesis by summarizing the key findings and discussing potential areas for improvements and further research.

This thesis is part of the bachelor program Data Science and Artificial Intelligence at the Leiden Institute of Advanced Computer Science (LIACS) at Leiden University and was written under the supervision of Prof.dr. M.R. Spruit and Dr. A.E.J.L. Lefebvre.

## 2 Background Knowledge

### 2.1 Welzijn.AI

The application of technology in healthcare is a rapidly growing field, with extensive research focused on its potential to improve efficiency and effectiveness. This interest is unsurprising, given that technology can streamline processes and enhance the speed of medical services. However, it is crucial to identify where and how technology can be most effectively implemented. This study specifically examines the use of technology in supporting vulnerable elderly populations.

To gain insights into the most beneficial applications of technology for elderly, a document was created by the Leiden University Medical Center (LUMC) following interviews with multiple stakeholders. The document highlights that technology can play a significant role as a companion to combat loneliness among the elderly but should not replace human interaction (van Dijk et al., [In review](#)). Therefore, one primary goal of technological integration is to provide companionship to the elderly. Another critical goal is utilizing technology for monitoring and signaling purposes, allowing for the detection of unusual behaviour that can be reported to healthcare professionals, who can then provide appropriate assistance.

With these objectives in mind, the Welzijn.AI program has been established. The aim of Welzijn.AI is to continuously monitor the well-being of the elderly by administering the EQ-5D questionnaire. The EQ-5D is a standardized instrument used to assess five dimensions of health: mobility, self-care, daily activities, pain/discomfort and anxiety/depression. Additionally, patients rate their overall health on a scale from 0 to 100 (The EuroQol Group, [1990](#)).

Although the ultimate goal of Welzijn.AI is to create a teddy bear or furry cat that can communicate with elderly individuals, this study focuses on developing a chat-based product. One of the primary objectives for this study is to build a foundational program for Welzijn.AI which will serve as the basis for the overall initiative. This foundational program is essential as it involves creating the core structure and basic conversational framework of Welzijn.AI in a chat-based format. This foundational program can then be used for further development of Welzijn.AI.

The requirements for the foundation program, outlined below, have been directly derived from the objectives set forth by van Dijk et al. ([In review](#)) based on stakeholder interviews. These requirements ensure that the base program is equipped with essential functionalities, providing a robust foundation for Welzijn.AI.

1. **Chat function:** The program should have a robust chat function to facilitate seamless communication with users. This function forms the core of the interactive experience, enabling users to engage in conversations.
2. **Question-asking capability:** The program must be able to ask questions to users. This feature is essential as the conversations revolve around a questionnaire.
3. **Small talk functionality:** The program should be able to engage in small talk. This ability to carry on casual conversations helps build a personal connection with users and makes the conversation more natural and enjoyable.

4. **Connection of answers to multiple-choice options:** The questionnaire used in this program consists of multiple choice questions. However, to facilitate open conversations with users, the program needs a function that can connect open-ended answers to one of the multiple-choice options. This feature is also crucial for structured data collection.
5. **Answer score saving:** Each multiple choice answer has an associated score. The program should be able to save these scores for a user each time they complete the questionnaire and should also be able to save a total score. This functionality allows for tracking and analysis user responses over time, providing a means for continuous wellness monitoring.

These requirements ensure that the base program is equipped with essential functionalities, providing a robust foundation for the Welzijn.AI initiative. This allows the program to facilitate meaningful and effective user interactions, enabling it to serve as both a companion and an alert mechanism, as discussed above.



Figure 1: Three types of hallucinations

## 2.2 Hallucinations

Given the objectives of the Welzijn.AI program, an LLM is an ideal tool to facilitate interactions between elderly users and the program. LLMs are well-suited for this task due to their advanced natural language processing capabilities, adaptability and scalability, making them ideal for the conversational dynamics required by Welzijn.AI. However, LLMs are prone to a significant issue known as hallucinations, where they generate content that is nonsensical or not faithful to the provided source. Hallucinations can be categorized into three types: input-conflicting, context-conflicting, fact-conflicting hallucinations (Zhang et al., 2023).

1. **Input-conflicting hallucinations** occur when the generated content deviates from the user’s input. For instance, as illustrated in Figure 1, a user requests a dinner recipe but receives a lunch recipe instead.
2. **Context-conflicting hallucinations** arise when LLMs produce content that contradicts information they have previously generated. An example in Figure 1 shows the model providing a recipe for fried chicken and mashed potatoes, yet concluding with ”Enjoy this steak”.
3. **Fact-conflicting hallucinations** occur when LLMs present information that is factually incorrect. In Figure 1, for example, the model incorrectly claims that tomatoes are rich in calcium.

Overall, hallucinations significantly undermine the reliability of LLMs and disrupt the fluency and coherency of a conversation. This study aims to leverage the fact that the conversation centers around a questionnaire to mitigate hallucinations. This approach ensures that the conversation maintains a natural flow, remains coherent and avoids the types of hallucinations discussed above.

### 3 Method

This research comprises two main components: a systematic literature review and a practical application involving the development and testing of a program. This chapter outlines the methodologies used for both aspects. Section 3.1 discusses the approach used for the systematic literature review. Section 3.2 discusses the code for the base program of the Welzijn.AI program. Section 3.3 discusses the evaluation of the base program.

#### 3.1 Literature Review

For this thesis, a systematic literature review, as recommended by van Haastrecht et al. (2021), was conducted. Although their approach was not strictly adhered to, their approach served as both inspiration and a foundation for the review. In the original SYMBALS methodology, a backward snowballing stage follows the initial screening. However, in this study, the backward snowballing step has been omitted. Since this research addresses a novel problem and focuses on papers from the ACL conference, which typically contain the latest insights in this field, it is anticipated that backward snowballing would not yield additional relevant information. Therefore, this step has been excluded to maintain efficiency in the screening process.

The approach used in this study is summarized in a flowchart in figure 2. The focus for this literature review was exclusively on papers presented at the Association for Computational Linguistics (ACL) conferences in 2022 and 2023, as these represent the latest research contributions in this field.

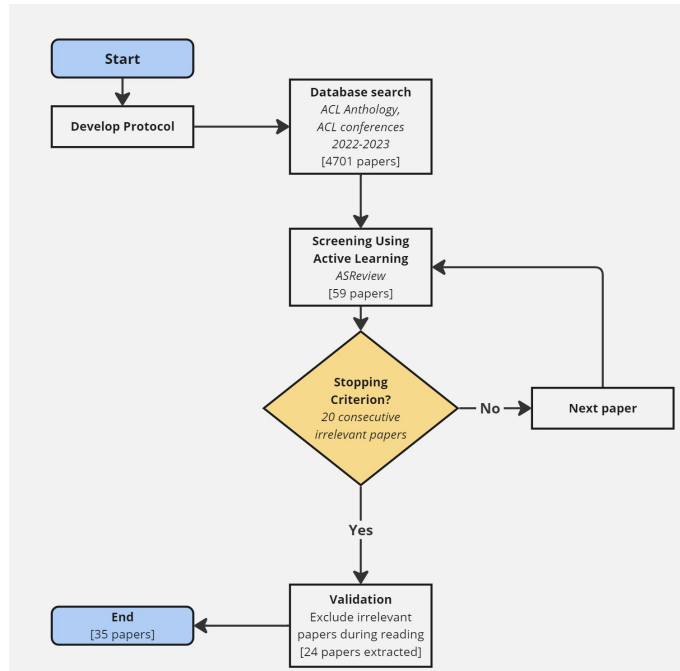


Figure 2: Adapted SYMBALS Method Flowchart



The systematic literature review’s initial step involves a database search, primarily relying on the ACL Anthology website. This site provides a comprehensive list of all papers presented at each ACL conference, sorted by year. Given the focus on ACL conferences in 2022 and 2023, the search included all papers from these specific events. Titles and abstracts were retrieved from the website using a custom Python script and exported to a CSV file, resulting in a total of 4,701 papers as the basis for this literature review.

The next step in the literature review is the screening process, for which ASReview, as recommended by van Haastrecht et al. (2021), is used. ASReview requires a database of all papers to be reviewed, along with at least one example of a relevant paper and one example of an irrelevant paper to initiate the process. In this study, four relevant and four irrelevant papers were chosen to ensure coverage of all important subjects in the prior knowledge base. Table 1 lists these selected papers, indicating why each paper was considered relevant and providing a brief description for each irrelevant paper. To identify relevant papers, titles in the database were scanned, followed by a review of abstracts that seemed potentially significant. The irrelevant papers were chosen using ASReview’s **random** option, which retrieves a random set of papers from the database. Among these, the ones that were deemed most irrelevant to the study’s focus were labeled as such. The literature review uses the default setup for the model, which is configured as follows:

**Feature extraction technique:** TF-IDF  
**Classifier:** Naive Bayes  
**Query strategy:** Maximum  
**Balance strategy:** Dynamic resampling (Double)

The final stage of the screening process involves identifying and filtering relevant papers. For this process a stopping criterion is employed. For this study, this means that if 20 consecutive papers are classified as not meeting the relevance criteria, the screening process is terminated. The stopping criterion is met upon screening 4.19% of the total papers, resulting in the identification of 59 relevant papers, thereby concluding the screening process. During the process of writing this thesis, 24 papers were excluded because they were either outside the scope of this study or did not fully align with its subject matter.

Paper	Relevant / Irrelevant	Reason
Z. Liu et al. (2022)	Relevant	Method for reducing hallucinations
Valizadeh and Parde (2022)	Relevant	Task-oriented dialogue systems in healthcare
Mishra et al. (2023)	Relevant	Integration of emotion and empathetic responses in conversational agents
Chiu et al. (2022)	Relevant	Switching from chit-chat to task-oriented dialogues
Salahudeen et al. (2023)	Irrelevant	Semantic analysis of African languages
Fan and Hunter (2023)	Irrelevant	Converting English recipe texts into a flow graph
W. Sun et al. (2023)	Irrelevant	Explores the security risks in neural code search models
Belani and Flanigan (2023)	Irrelevant	Develops dataset and system to identify code-switching functions in speech transcripts.

Table 1: Prior knowledge selection ASReview

### 3.2 Base Program

To build a program that meets the requirements discussed in Chapter 2, Python and OpenAI’s GPT-3.5 Turbo API are used. The program is divided into five main components, each contained in a separate Python file: `main.py`, `questionnaire.py`, `small_talk.py`, `evaluate_answer.py`, `save_results.py`. This section will discuss the structure of the base program, with two flowcharts provided to clarify the processes.

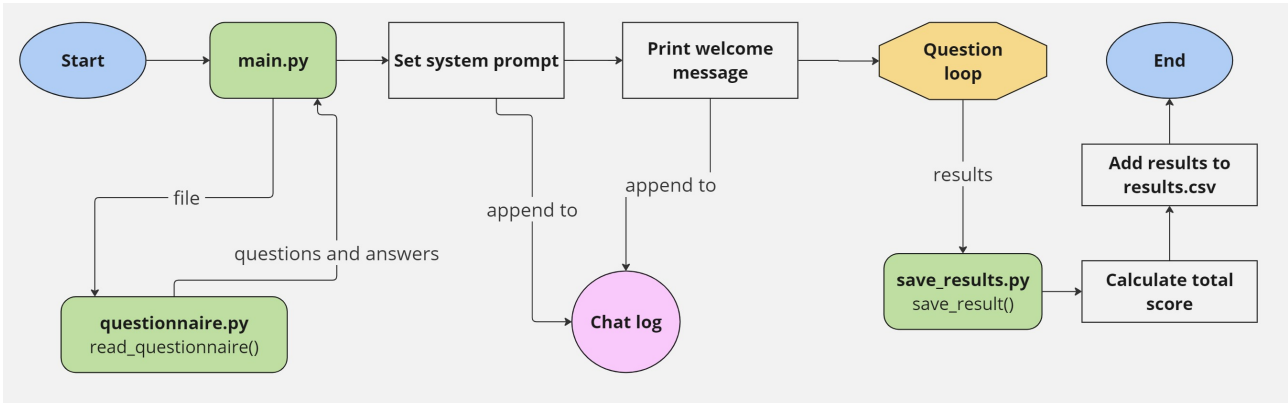


Figure 3: General overview of the main.py of the program

The first component we will discuss is `questionnaire.py`. This component of the program ensures that the questions and multiple choice answers can be extracted from a text file and saved in lists to be used in the code. As shown in Figure 3, `main.py` begins by calling `read_questionnaire()` from `questionnaire.py`. The `read_questionnaire()` function receives the path to the file containing

the questionnaire questions. It then iterates through these questions, separating the questions from the multiple-choice answers, and stores them in lists. Finally, it returns the lists of questions and answers to `main.py`

In `main.py`, the system prompt for GPT-3.5 Turbo is set. For this case, the system prompt is as follows:

Je bent Ted, een vriendelijke chatbot die reageert op de antwoorden van een vragenlijst die de gebruiker geeft. Je maakt vriendelijk een praatje met de gebruiker tussen vragen door.

English translation:

You are Ted, a friendly chatbot who responds to answers given by the user in a questionnaire. You engage in a friendly conversation with the user between questions.

This system prompt is appended to the `chatlog`, which is used to ensure that Ted can utilize the conversation history for a natural flow of the dialogue. Throughout the entire program, every output from Ted and every input from the user is appended to the `chatlog`. This leads to the next step in Figure 3, which involves printing a welcome message and appending it to the `chatlog`. After the welcome message, the question loop begins. This loop involves several processes, which are outlined in Figure 4

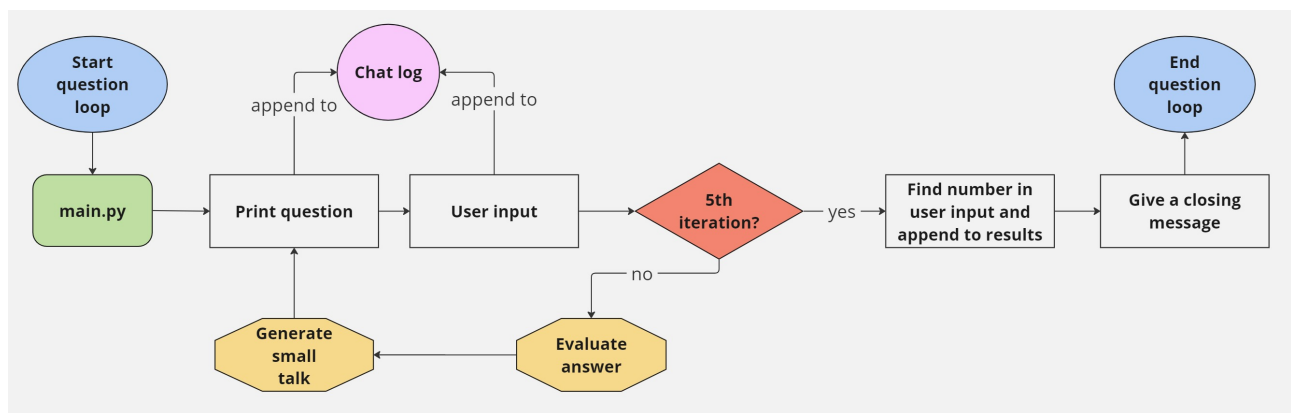


Figure 4: General overview of the question loop

The question loop begins by printing a question, after which the user provides an answer. Then the program checks whether this is the 5th iteration, indicating that it is the final question. If it is, the program extracts the user’s score from the input and adds it to the results. This is done only for the last question because it asks for a score between 0 and 100 eliminating the need to map it to a specific multiple-choice option. Ted then delivers a closing message along with a brief response to the user’s message.

If it is not the 5th iteration, the answer is evaluated using the `evaluate_answer()` function in `evaluate_answer.py`. This function maps the user’s response to one of the predefined options.

This function takes the **user\_response**, the **question**, **choices** and **results** as arguments. Here, **user\_response** is the answer provided by the user, **question** is the current question, **choices** is a list of possible answers and **results** is a list of scores.

In this function GPT-3.5 Turbo is configured with a different prompt to determine which multiple-choice answer the user response corresponds to. This is then converted into a score, where 5 is the highest score indicating the most positive answer (e.g. no trouble walking), and 1 is the lowest score indicating the most negative answer (e.g. the user cannot walk). This score is then appended to **results** and the small talk process is initiated in **small\_talk.py**.

In **small\_talk.py** the **generate\_small\_talk()** function takes **previous\_answer** and **chat\_log** as arguments. Here **previous\_answer** is the last reply given by the user and **chat\_log** is the chat history. Within **generate\_small\_talk()**, Ted initiates small talk based on the user's most recent reply. There is an automatic check for certain keywords to determine if the small talk response should transition to the next question. If so, the loop continues to the next question. Otherwise, the user can respond to the initiated small talk. A maximum number of small talk rounds per question is set to ensure that the conversation progresses to the next question after a certain point. Once this limit is reached, a closing response is given, and the user is directed to the next question.

Once the question loop ends, the **results** are passed to the **save\_result()** function in **save\_results.py**. This function calculates the total score by summing all the scores and saves the results in **results.csv** along with the date. This is the final step before reaching the end of **main.py**

The code used for the base program can be found at: <https://github.com/RuzannaBas/Thesis-QuestionnaireDrivenDialogue>

### 3.3 Evaluation

The purpose of this study is to test the hypothesis that using questionnaires as a basis for conversations can help mitigate hallucinations. To achieve this, an evaluation of the base program is required. The program is being evaluated by participants who have varying levels of experience with artificial intelligence. An overview of the evaluators is given in Table 2. Evaluators are instructed to interact with Ted, acting as if they are elderly individuals. To compensate for a limited number of evaluators, each participant is allowed to engage in multiple conversations with Ted, adopting different perspectives for each session. Following each interaction, the evaluator completes a questionnaire, which includes the following questions:

1. On a scale of 1 to 5, how would you rate the conversation overall?
2. On a scale of 1 to 5, how natural was the flow of the conversation?
3. On a scale of 1 to 5, how clear was the conversation?
4. On a scale of 1 to 5, was the conversation enjoyable/entertaining?
5. Did any hallucinations occur during the conversation?
6. Were there any other things that went wrong?

For question 5, a brief explanation with examples of the three different types of hallucinations is provided to help evaluators answer the question accurately. Additionally, the chat history from each session is saved for further analysis.

<b>Evaluator ID</b>	<b>Amount of test conversations</b>	<b>AI student? (if yes, amount of years)</b>	<b>Used AI before?</b>
<b>1</b>	1	No	Yes
<b>2</b>	5	Yes, 3	Yes
<b>3</b>	3	Yes, 3	Yes
<b>4</b>	1	No	Yes

Table 2: Evaluators overview table

## 4 Results

This chapter presents the findings from the systematic literature review and evaluates the results of the base program. The chapter is organized into multiple sections. Section 4.1 explores the outcomes of the systematic literature review, which is further divided into the following subsections: Section 4.1.1 examines knowledge-related strategies to mitigate hallucinations. Section 4.1.2 addresses methods linked to focus and control for reducing hallucinations. Section 4.1.3 concentrates on dialogue planning as a way to prevent hallucinations. Section 4.1.4 discusses an unconventional approach to hallucination mitigation. Section 4.1.5 explores techniques for integrating chit-chat with task-oriented conversations. Section 4.1.6 looks at methods for enhancing engagement in conversations through persona-based dialogue. Section 4.1.7 focuses on approaches for offering emotional support through chitchat.

Lastly, Section 4.2 discusses the results for the evaluation of the performance of the base program.

### 4.1 Systematic Literature Review

In this section, the discussion of the discovered papers is divided into two main subsections. The first part discusses findings on the main part of this research, namely the mitigation of hallucinations. Following that, the second subsection explores findings on how to make conversations more enjoyable. Each subsection is concluded with a table with an overview of the discussed models.

#### 4.1.1 Hallucinations and knowledge

One of the key challenges facing dialogue generation models is hallucination. This occurs when models produce responses that are factually incorrect or irrelevant to the conversation. B. Sun et al. (2023) identify two primary causes of hallucination in these models. First, the optimization objective is usually polarized by gold knowledge dialogue samples and general dialogue samples without knowledge in current knowledge-grounded dialogue datasets. Because of this models struggle with handling incomplete correct knowledge, leading them to blindly trust provided knowledge, regardless of its accuracy or relevance. Secondly, the process of generating knowledge itself can be prone to errors. Models may sometimes retrieve or create inaccurate knowledge during dialogue, leading to factually incorrect responses. In this subsection, we will discuss solutions aimed at mitigating hallucinations and improving coherence with the use of (external) knowledge.

#### **ACK-DEF reduces hallucinations by expanding training data and incorporating faulty knowledge for improved error detection and handling.**

To mitigate these hallucination problems B. Sun et al. (2023) propose an Augmentative and Contrastive Knowledge Dialogue Expansion Framework (ACK-DEF). Inspired by how humans clarify situations with incomplete knowledge, ACK-DEF tackles the problem in two ways. First, it expands the training data with variations of existing knowledge, enhancing the model’s ability to manage minor errors. Secondly, it introduces ”faulty knowledge” paired with responses acknowledging potential inaccuracies. This trains the model to recognize patterns associated with incorrect information, improving its ability to distinguish reliable from unreliable knowledge. Ultimately, by equipping models to identify and handle situations with incomplete or potentially inaccurate information, ACK-DEF aims to significantly reduce hallucinations in dialogue generation.

### **Think-Before-Speaking (TBS) improves dialogue models by integrating implicit commonsense knowledge for more accurate and coherent responses.**

Another reason dialogue models struggle with hallucinations is because they often lack the ability to consider the underlying commonsense knowledge that guides human conversation (P. Zhou et al., 2022). Otani et al. (2023) also argue that open-domain conversational tasks and datasets should include some form of situational information, as this is essential for understanding the conversation and generating coherent responses. To address this, P. Zhou et al. (2022) propose Think-Before-Speaking (TBS). TBS is a response generation framework that tackles hallucinations by first using weak supervision to identify implicit common-sense knowledge relevant to the conversation. Then, it leverages this knowledge to generate responses that are both factually accurate and follow common-sense reasoning. By enabling the model to "think" before it "speaks," TBS aims to improve the quality and informativeness of dialogue models while reducing the occurrence of hallucinations.

### **Enhancing dialogue systems with logical reasoning to reduce hallucinations.**

While the focus of Beygi et al. (2022) differs from hallucinations and coherence, it offers a potential solution for reducing hallucinations and enhancing coherence. The approach in Beygi et al. (2022) draws inspiration from human logical reasoning, a crucial aspect of human thought and communication. In task-oriented dialogues, logical reasoning is essential for understanding the user's request, asking relevant questions to address the user's task effectively, and minimizing irrelevant inquiries. However, current task-oriented dialogue systems only support very limited forms of logical reasoning (Beygi et al., 2022). Consequently, these systems often rely on external, rule-based, and domain-specific components, leading to a poor user experience. To address this, Beygi et al. (2022) propose incorporating reasoning capabilities into task-oriented dialogue systems. Specifically, it introduces a novel deep learning method to fine-tune pre-trained models for reasoning over numerical and categorical attributes within the dialogue context. The paper also presents an architecture for integrating this model. The objective is to ensure that the system employs logical reasoning to respond accurately when all required information is available. Simultaneously, it should be able to extract constraints and inform downstream components when context is partial, and refrain from providing answers when relevant information is lacking. As previously mentioned, one cause of hallucinations is models mistakenly believing they possess complete information. This proposed model would help mitigate hallucinations by distinguishing whether it has all necessary information and reasoning accordingly.

### **RHO enhances knowledge-grounded dialogue systems by integrating structured knowledge and re-ranking to reduce hallucinations.**

Knowledge-grounded dialogue systems use external knowledge to generate coherent and informative responses. However, hallucination remains a challenge in these systems (Ji et al., 2023). One contributing factor is the difference between external knowledge and textual dialogue content, making it difficult for real response generation models to learn the representations and correlations within the input sources. To address this, Ji et al. (2023) propose RHO, a method for faithful open-domain dialogue response generation with enhanced knowledge grounding and re-ranking. RHO uses structured embeddings for entities and relations within the knowledge graph (KG) and links them to the dialogue context. Within the encoder-decoder model, the representation of all the

linked entities and relations are grounded using KG embeddings. This grounding occurs both locally, focusing on individual entities, and globally, considering the entire sub-graph. These techniques enable the model to effectively encode and integrate contextual knowledge from related sub-graphs with appropriate attention. Additionally, the generated responses are re-ranked based on their degree of hallucination using conversational reasoning. Experiments indicate a 32.93% reduction in hallucinations according to human evaluations.

### **Dynamic internet-accessible models reduce hallucinations by incorporating up-to-date knowledge into dialogue responses.**

Current models are typically trained using either an encoder-decoder or decoder-only setup on large datasets of human-human conversations. The knowledge acquired during training is stored in the model’s weights, resulting in static language models that are fixed at the moment of training and do not account for the dynamic nature of the world, where new information constantly emerges. Consequently, these static language models are prone to hallucination. In contrast, Komeili et al. (2022) explore generative models capable of dynamically accessing the vast knowledge of the internet to inform their responses. Utilizing encoder-decoder architectures, the model first generates a search query based on the dialogue context. These queries are then used to retrieve relevant knowledge, which is prepended to the conversational history. Considering this encoded knowledge, a response is generated using the decoder. By accessing the internet, the model remains up-to-date. Results indicate that such models produce conversations that are less factually incorrect and more consistent.

### **Multi-Stage Dialogue Prompting (MSDP) and advanced prompting techniques enhance dialogue responses by leveraging pre-trained LMs for diverse and relevant knowledge without finetuning.**

Another issue with dialogue systems is that they often produce bland and generic outputs. One solution has been to ground dialogue responses in a source of knowledge, ensuring more informative and accurate responses while reducing hallucinations. However, these systems are limited by the size and domains of the database, preventing them from generalizing to out-of-domain topics. Additionally, substantial resources are required to retrieve information from the corpus. To overcome these limitations, Z. Liu et al. (2022) use a small database and a pretrained language model (LM) as an additional grounding source. Since an LM contains diverse knowledge, it can help dialogue systems generalize to out-of-domain topics. Z. Liu et al. (2022) propose a prompt-based approach to directly generate context-relevant knowledge from the LM. To avoid the finetuning process, which can cause overfitting and is sometimes impractical due to the API-only availability of large LMs, they suggest further prompting to generate responses based on the dialogue context and previously generated knowledge. This Multi-Stage Dialogue Prompting (MSDP) framework eliminates the need for finetuning or updating weights, allowing the system to generate relevant and factually correct knowledge while producing knowledgeable and engaging responses. Similarly, Chen et al. (2023) also recognize the problems with fine-tuned generation models for mixed-initiative dialogue generation. They propose using vanilla large pre-trained LLMs and demonstrate a systematic way of prompting, which shows strong performance.



ID	Model	Theoretical Framework	Methodology	Core Mechanism
1	Augmentative and Contrastive Knowledge Dialogue Expansion Framework (ACK-DEF)	Expansion of training data and introduction of faulty knowledge	Expansion of training data with variations and faulty knowledge for improved error detection	Fine-tuning
2	Think-Before-Speaking (TBS)	Integration om implicit commonsense knowledge	Weak supervision to identify and use implicit commonsense knowledge in response generation	Retrieval Augmented Generation
3	Logical Reasoning Enhancements	Logical reasoning inspired by human thought processes	Fine-tuning pre-trained models for reasoning over numerical and categorical attributes	Fine-tuning
4	RHO	Knowledge-grounded dialogue systems using structured knowledge embeddings	Grounding dialogue responses in knowledge graphs and re-ranking responses	Retrieval Augmented Generation
5	Internet-accessible Models	Dynamic access to up-to-date knowledge	Encoder-decoder architecture with internet query generation and knowledge retrieval	Retrieval Augmented Generation
6	Multi-Stage Dialogue Prompting (MSDP)	Multi-stage prompting without finetuning	Leveraging pre-trained LMs and multi-stage prompts for context-relevant knowledge generation	Prompting

Table 3: Overview of models and techniques to mitigate hallucinations using knowledge

#### 4.1.2 Focus and control

**Focus Learning (FocusL) enhances dialogue models by prioritizing key knowledge and adjusting word importance to reduce hallucinations and improve response quality.**

According to Deng et al. (2023), another cause of hallucination is that traditional training methods struggle to focus on the most important parts of the response. This results in dialogue models not being able to prioritize relevant knowledge during training and focusing on simpler aspects and neglecting key information instead. Deng et al. (2023) propose Focus Learning (FocusL) to address the challenge of locating and correcting the learning focus. FocusL directly influences the model’s learning process by adjusting the importance of each word in the response. It identifies key knowledge-related words and dynamically assigns them higher weight during training. By doing this, the model steers towards focusing on the important elements which reduces hallucinations and improves the quality and consistency of responses.

**Focus-Aware Response Generation (FRG) improves dialogue relevance by using multi-level encoders and focal decoders to prioritize user-specific interests and reduce hallucinations.**

Wu et al. (2023) also argue that it is crucial to ensure that responses remain focused on the user’s specific interests. Existing response generation models often struggle with this, prioritizing fluency over maintaining focus on the core topics of the inquiry. Wu et al. (2023) propose a novel approach called Focus-Aware Response Generation (FRG) to address this challenge. FRG employs a multi-level encoder, focal decoders, and a synergistic focus ranking module. The multi-level encoder enhances the learning of latent focuses from the conversation history by aggregating characteristics of the speakers and content through a speaker-level attention layer and a block-level attention layer. Each decoder in the set of focal decoders generates a candidate response guided by its corresponding focus. The focus ranking module then ranks all candidate responses generated by the focal decoders and predicts the next focus for the final input. This approach ensures that responses stay relevant to the conversation, thereby reducing hallucinations. Results demonstrate improved performance in both fluency and rationality.

**Data flow transduction ensures truthful dialogue generation by transforming actions and results into formal representations with declarative rules.**

Fang et al. (2023) propose a technique called data flow transduction to address the challenge of hallucinations. This is a method for transforming the actions and results of a dialogue agent into a formal representation suitable for generation responses. It uses declarative rules to translate the model’s actions and the resulting information flow into a specific grammar. This grammar dictates the set of grammatically correct and conversationally appropriate responses based on the conversation history and retrieved knowledge. By essentially constraining the model’s response options, data flow transduction helps ensure truthful and controllable dialogue generation.

**FutureTOD enhances task-oriented dialogue by using self-training with future utterances to improve context understanding and coherence, reducing reliance on contrastive learning.**

Pre-trained language models (PLMs) trained on large general text corpora often struggle in task-oriented dialogues due to the inherent differences in language patterns between open-ended text and goal-driven conversations. Existing dialogue pre-training methods primarily rely on contrastive learning, which requires identifying positive and negative examples to guide the model. However, this approach faces challenges in selecting high-quality training pairs. To address these limitations, Zeng, He, et al. (2023) propose a novel pre-training framework called FutureTOD. FutureTOD distills future knowledge into the representation of the previous dialogue context using future utterances. Instead of relying on contrastive learning objectives, FutureTOD employs a self-training framework, eliminating the need for contrastive pairs. This approach allows the model to learn both local context information and the overall flow of task-oriented conversations, which can help reduce hallucinations and ensure coherence.

## **Gradient-guided Controllable Retrieval (GRACE) improves attribute-based text generation by integrating natural language patterns and guiding inference to enhance relevance and fluency.**

One approach to ensuring the generation of relevant responses is by controlling text generation and steering it toward a specific direction. This can be achieved through attribute-based text generation, which requires the generated text to satisfy a given attribute, essentially a control code for a specific topic, sentiment, or style. PLMs can generate fluent text by learning from large corpora; however, controlling them is challenging because they do not inherently learn to adapt to control signals. Solutions for this include re-training a PLM with supervised control signals or fine-tuning it on domain-specific data, which ensures high controllability and fluency. However, attribute-based supervised datasets usually originate from specific domains and lack generalization across different domains. Additionally, re-training large PLMs is very expensive. To address these issues, an inference-based approach can be used, where the PLM is frozen and generation preferences are influenced during the inference stage. Wen et al. (2023) argue that because auto regressive inference decomposes text generation into multiple steps, an inference-based approach will always push word-level probabilities toward the target attribute. This can disrupt the natural inference process, resulting in less fluent sentences. Therefore, Wen et al. (2023) suggest that inference-based methods require guiding information that satisfies both attribute and common language patterns to achieve attribute-based text generation. Natural language patterns ensure fluency and grammaticality, so it is preferable for the controlling information to come from a natural text span. Wen et al. (2023) propose augmenting attribute-based generation through Gradient-guided Controllable Retrieval (GRACE) considering the target attributes. In this framework, a discriminator is trained to compute the attribute distribution of a given context. A retrieval repository is built, storing natural text with its semantic and attribute information distilled from unlabeled data. The generation model then extracts attribute-related information with similar semantics through controllable retrieval. Irrelevant attributes are disentangled from the retrieval results, and the PLM representations are integrated into the generation process. Simultaneously, an algorithm iteratively revises the stepwise generation based on gradients. By optimizing toward the target attribute, the algorithm retrieves information with stronger attribute intensity, thus improving the attribute relevance of the generated text.

ID	Model	Theoretical Framework	Methodology	Core Mechanism
7	Focus Learning (FocusL)	Prioritization of key knowledge and adjustment of word importance	Dynamic weighting of key knowledge-related words during training	Fine-tuning
8	Focus-Aware Response Generation (FRG)	User-specific focus prioritization	Multi-level encodes and focal decoders with a focus ranking module	Retrieval Augmented Generation
9	Data Flow Transduction	Transformation of dialogue actions and results into formal representations	Use of declarative rules to generate grammatically correct and contextually appropriate responses	Fine-tuning
10	FutureTOD	Self-training with future utterances for context understanding	Self-training framework using future utterances to improve dialogue coherence	Fine-tuning
11	Gradient-guided Controllable Retrieval (GRACE)	Attribute-based text generation with gradient-based control	Inference-based approach with natural language patterns for attribute-guided generation	Retrieval Augmented Generation

Table 4: Overview of models and techniques to mitigate hallucinations using focus and control

### 4.1.3 Dialogue planning

For goal-directed dialogue systems, the aim is also to proactively steer conversations towards a designated target. Achieving this requires effective dialogue planning, which involves taking appropriate actions and smoothly guiding dialogue topics toward the target. This process should be both coherent and natural.

**COLOR uses Brownian bridge stochastic processes for global dialogue planning, enhancing coherence and target alignment through latent space mapping and real-time user feedback.**

J. Wang et al. (2023) propose a global planning method to connect the current context to the target at each dialogue turn. The dialogue path should balance coherence with smooth transitions towards the target. J. Wang et al. (2023) conceptualize goal-directed dialogue as a Brownian bridge stochastic problem, where path trajectories without a target are modeled as a Brownian motion, implying that neighboring trajectory points are similar while distant ones are dissimilar. They introduce a coherent dialogue planning approach using the Brownian bridge (COLOR) stochastic process. This method maps dialogue path points into a latent space conditioned on the current context and the designated target. User utterances are also mapped into the latent space in real-time to ensure goal-directed behaviour, with user feedback used to adjust the density and uncertainty of the Brownian bridge. The training process employs a contrastive objective to ensure global coherence, and PLMs are fine-tuned using the derived latent trajectories to plan

dialogue paths. These paths provide step-by-step explanations for reaching the target and serve as natural language prompts for generating system utterances. Experiments demonstrate that this approach produces more coherent utterances and achieves the target at a significantly higher success rate, although directing dialogue in the context of open-domain chit-chat remains challenging.

### **PLANET enhances long-form text generation by integrating dynamic content planning and latent representations to improve coherence and overall quality.**

Hu et al. (2022) identify the problem of incoherence in long-form text generation models. These models excel at producing fluent text, but struggle to generate coherent and structured long-form content. To address this issue, Hu et al. (2022) propose PLANET, a framework that integrates dynamic content planning within the generation process. PLANET introduces latent representations within the transformer decoder. These representations serve as a semantic blueprint, guiding the model to generate sentences that align with a cohesive overall plan. For each sentence, an autoregressive decoder produces a latent representation that captures the core semantic content. This latent representation then attends to both previous latent representations and the preceding context to dynamically plan the overall semantic structure of the sentence. Following this planning stage, each position within the sentence attends to the corresponding latent representation and previously generated words to determine the specific wording. To ensure the latent representations effectively guide content planning, PLANET incorporates a sentence-level auxiliary task. This task involves predicting a "bag-of-words" representation, essentially providing supervision signals for the lexical semantics of the corresponding sentence. Additionally, a contrastive learning objective is introduced to differentiate between coherent and incoherent sentence sequences. This objective utilizes various strategies to construct negative samples, essentially teaching the model to identify and avoid incoherent patterns. Evaluations demonstrate that PLANET significantly outperforms strong baseline models. The generated texts show superior coherence, richer content, and improved overall quality when compared to baselines.

### **MTGP uses a global path generator to guide multi-turn dialogue, improving coherence and fluency in responses through structured planning.**

A. Liu et al. (2023) also propose using planning to guide dialogue and ensure coherent responses. They introduce the Multi-turn Target-oriented Dialogue Guided by Generative Global Path (MTGP) model, which generates a global path using a generative model to guide multi-turn dialogue. The MTGP model comprises two components: a path generator and a next-turn response generator. The path generator first creates a global path in the form of a natural language sentence, linking concepts in the source context with the global target. Then, a response is generated to guide the dialogue. These generated responses are continuously added to the context as dialogue history. Experiments show that this approach results in more coherent and fluent responses.

### **DCG model enhances multi-attribute controllable dialogue by disentangling planning from generation, allowing for generalization to unseen attribute combinations.**

Building upon the concept of dialogue planning for goal-oriented conversations, Zeng, Zhao, et al. (2023) address a challenge in multi-attribute controllable dialogue generation. Their work, titled "Seen to unseen: exploring compositional generalization of multi-attribute controllable dialogue generation," investigates how to create dialogues that consider various controllable attributes, even when encountering unseen combinations of those attributes. They propose a prompt-based

disentangled controllable dialogue generation model (DCG) to achieve this. This model separates the dialogue planning process from the dialogue generation itself. Zeng, Zhao, et al. (2023) use prompts to condition the dialogue planning on the desired attributes, allowing the model to generalize to unseen combinations. This disentanglement promotes both controllability and coherence in the generated dialogues. Their work highlights the potential of utilizing planning mechanisms to improve the effectiveness of controllable dialogue systems.

ID	Model	Theoretical Framework	Methodology	Core Mechanism
12	COLOR (Brownian Bridge Stochastic Processes)	Brownian bridge stochastic processes for global dialogue planning	Latent space mapping and real-time user feedback	Fine-tuning
13	PLANET	Dynamic content planning and latent representations	Integrates dynamic content planning within generation process	Fine-tuning
14	MTGP (Multi-turn Target-oriented Dialogue Guided by Generative Global Path)	Transformation of dialogue actions and results into formal representations	Generates global path using a generative model	Prompting
15	DCG (Disentangles Controllable Dialogue Generation)	Self-training with future utterances for context understanding	Prompt-based disentangled controllable dialogue generation	Prompting

Table 5: Overview of models and techniques to mitigate hallucinations using dialogue planning

#### 4.1.4 Different approach

**SimOAP enhances dialogue coherence and consistency by over-sampling diverse responses and using a two-stage post-evaluation process with TF-IDF and NLI models.**

J. Zhou, Pang, et al. (2023) argue that existing dialogue models possess strong capabilities, enabling them to generate consistent and coherent responses. However, the challenge lies in effectively using these capabilities. J. Zhou, Pang, et al. (2023) address the issue of reducing hallucinations and improving coherency by critically examining the classical approaches of existing dialogue models. One of the characteristics of current dialogue models is the assumption that the response with the highest probability is the best. However, J. Zhou, Pang, et al. (2023) argue that the proportion of high-quality responses is greater among low-probability responses compared to high-probability ones. Based on this insight, they propose a simple two-stage method: over-sampling and post-evaluation (SimOAP), to enhance coherence and consistency in persona-based dialogues. In the first stage of this method, diverse responses are generated instead of focusing solely on high-probability responses. In the second stage, a coherent response is selected. To make the post-evaluation stage time-efficient, the TF-IDF algorithm calculates the semantic similarity between candidate responses and historical utterances. This helps assess the coherence of candidate responses and filter out those that exhibit poor coherence. Subsequently, a consistency evaluation with a Natural Language Inference (NLI) Model is employed to select the final response. By using TF-IDF first, the number of candidate responses is reduced, thereby decreasing the time cost of the subsequent evaluation. Evaluations conclude that this method improves the backbone models and outperforms baselines.

ID	Model	Theoretical Framework	Methodology	Core Mechanism
16	SimOAP	Enhances dialogue coherence and consistency by over-sampling diverse responses	Oversampling and two-stage post-evaluation with TF-IDF and NLI models	Prompting

Table 6: Overview of models and techniques to mitigate hallucinations using a complete different approach

#### 4.1.5 Unifying task and chitchat skills

Another important aspect of this research is ensuring that the conversation remains engaging and enjoyable. Unlike task-oriented dialogue systems, which are necessary for conducting questionnaires, chit-chat dialogue systems are needed to make conversations entertaining. These two types of dialogue systems differ fundamentally. As X. Zhao et al. (2022) discuss, chit-chat dialogue systems are characterized by longer turns and higher diversity compared to task-oriented systems. Also the purpose of these systems is different as a chit-chat dialogue system aims to entertain while a task-oriented dialogue system focuses on completing specific tasks. Most models focus on only one of these two types of dialogue systems, while it can be expected that a single conversation needs both the skills of chatting and task-oriented dialogue.

#### **UniDS integrates chat and task-oriented dialogue capabilities into a unified framework using a two-stage training method and data schema to handle both types effectively.**

To address this, X. Zhao et al. (2022) propose UniDS, an auto-regressive language model-based system capable of handling both types within a unified framework. However, the differences in these systems, such as the absence of belief states in chit-chat dialogues pose challenges for unification. X. Zhao et al. (2022) tackle these challenges by designing a data schema that incorporates belief states, database result representation, and system acts for chit-chat dialogues. To address the challenge of the chit-chat dialogue system needing more data and training due to its high diversity, a two-stage training method is used. First, a chit-chat dialogue model is trained with a large dataset of chit-chat dialogues, then UniDS is further trained with mixed dialogue data consisting of both chit-chat and task-oriented dialogue data. Experimental results show that UniDS performs comparably with state-of-the-art chit-chat systems. Furthermore, UniDS demonstrates good switch ability between the two different types of dialogues.

#### **SalesBot combines chit-chat and task-oriented dialogue by using a simulated user-salesperson framework to generate training data for seamless transitions between casual and task-focused conversations.**

Similarly, Chiu et al. (2022) address the challenge of combining chit-chat dialogue systems and task-oriented dialogue systems by drawing inspiration from a salesperson’s capabilities. A salesperson must be able to engage in casual conversation while also identifying opportunities to switch to a sales pitch. For a dialogue system this translates to the system engaging in a casual chat while simultaneously detecting implicit task-oriented intents. To achieve this, Chiu et al. (2022) propose a new framework with a simulated user and a simulated salesperson to automatically generate

dialogues that transition from social chit-chats to task-oriented dialogues. This data can then be used to train models in a semi-supervised and unsupervised usage. While the primary goal in Chiu et al. (2022) is to generate data, their approach provides valuable insights into how such a system should function. The proposed framework consists of three main parts: open-domain dialogue generation, chit-chat to task-oriented transition, and task-oriented dialogue generation. The framework starts with open-domain dialogue generation. During this process, a task-oriented intent detector identifies potential transitions to task-oriented dialogues. Once an intent is detected, a transition turn is generated using a template followed by response regeneration to ensure fluency.

ID	Model	Theoretical Framework	Methodology	Core Mechanism
17	UniDS	Integrates chat and task-oriented dialogue capabilities into a unified framework	Two-stage training method with a data schema for handling both types of dialogue	Fine-tuning
18	Salesbot	Combines chit-chat and task-oriented dialogue using a simulated user-salesperson framework	Generates training data for transitions between casual and task-focused conversations	Prompting

Table 7: Overview of models and techniques to combine chit-chat and task-oriented conversation

#### 4.1.6 Persona-based dialogue

An essential aspect of chit-chat is ensuring the conversation remains engaging and entertaining. One effective way to achieve this is through persona-based dialogue systems. Xu et al. (2022) highlight that incorporating a persona increases engagement and fosters long-term user intimacy. Kwon et al. (2023) note that personalized responses based on long-term user memory lead to more trustworthy interactions. Furthermore, S. Liu et al. (2023) mention that personalized dialogue models also produce more consistent dialogues. However, maintaining relevance and consistency with personas presents challenges in generating personalized dialogues.

#### Addressing the WWH problem with a Multi-Session Personalized Conversation dataset and adjustable blending weights to enhance persona-grounding for natural, personalized responses

To address these challenges, Kwon et al. (2023) tackle the WWH (what, when, how) problem in open-domain dialogue systems. This problem consists of three questions: 1) What personal information should be grounded given the conversation context? 2) When should responses be generated using personal information? and 3) How can natural and human-like personalized responses be created? Kwon et al. (2023) propose a method that controls the inclination of models to generate personalized responses. This technique blends persona-augmented datasets to construct a personalized dialogue system enabling human-like natural conversation.

First, Kwon et al. (2023) create a Multi-Session Personalized Conversation dataset, which is used to train the model to ground the provided persona information effectively for personalized responses. The model’s persona-grounding level can be controlled by adjusting the blending weights



of the conversational datasets. These datasets are also enriched with negative samples of persona subsets. Furthermore, to enhance both generation quality and the controllability as well as the interpretability of persona-grounded generation, a turn label is used that indicates whether a turn is personalized or casual. Experimental results show the effectiveness of the proposed methods in both subjective and objective evaluations.

### **MPC tackles data scarcity and training costs in persona-based and open-domain chatbots by using prompt-based techniques for long-term coherence and flexibility.**

One of the major challenges with persona-based dialogue systems, and open-domain chat bots in general, is the scarcity of training data and the high costs associated with training models. As noted by Cao et al. (2022) most existing generation models for this task rely on training with sufficient persona-based dialogues. However, availability of such data is limited due to the expensive nature of data collection. Besides the limited data scale, persona-based dialogues are inherently more complex to learn compared to conventional dialogues. This data scarcity also applies to open-domain chat bots, where fine-tuning large models can be difficult. Lee et al. (2023) address this challenge with their Modular Prompted Chatbot (MPC) approach. MPC utilizes prompt-based techniques to achieve long-term coherence and flexibility in open-domain conversations, potentially enabling more natural and engaging chitchat without the need for extensive fine-tuning.

### **Addressing limitations in persona-based models by tackling issues with data availability, scalability, and the need for explicit persona descriptions.”**

Additionally, these models often require explicit persona descriptions during both the training and inference stages, which severely limits their application in real-world scenarios. According to W. Zhou et al. (2023), detailed persona descriptions for speakers are usually unavailable in most real-world situations. Furthermore, Xu et al. (2022) highlight that retrieval-augmented generative models also face data challenges as they require long-session conversations for training, which are both costly and difficult to annotate. As S. Liu et al. (2023) mention, methods typically depend on dialogue data paired with user traits, profiles, or persona description sentences, which are also difficult to gather. Moreover, explicit personas usually include only a few user traits or profile sentences, limiting the amount of information they convey and thereby restricting the model’s ability to capture and express nuanced personalization. Finally, Xu et al. (2023) show that in addition to the high costs of manually collecting persona dialogue datasets, there are scalability issues when dealing with unseen persona information in fine-tuned datasets.

### **In-context prompting learning enables zero-shot persona-based dialogue by integrating preset persona information into the dialogue context, reducing the need for labeled data.**

Recent advancements propose innovative solutions to these challenges. For instance Xu et al. (2023) propose an in-context prompting learning mechanism to address zero-shot persona-based dialogue generation. This is done by formatting the preset persona information into multiple turns of a dialogue and placing it at the beginning of the original context. This approach enables zero-shot persona customization without the need for annotating specific persona data, leveraging pre-trained dialogue foundation models to generate persona-based dialogues without costly labeled data.

**PLATO-LTM framework enhances long-term persona consistency by extracting, storing, and integrating user and chatbot personas in real time, eliminating the need for extensive long-session data.**

Another approach to avoid the use of expensive data is presented by Xu et al. (2022). This work also tackles the issue of dialogue models forgetting user personas between interactions. They introduce the PLATO-LTM framework which extracts and remembers both user and chat bot personas in real time. Because of this the chat bot can have long-term persona dialogue without training on long-session data. This framework consists of three modules: a Persona Extractor(PE), which filters irrelevant information and extracts persona sentences, a Long-Term Memory(LTM) for storing persona information and a Generation Module that incorporates retrieved persona sentences directly into the dialogue context using a large-scale model. Automated and human evaluations demonstrate that these methods significantly enhance consistency in long conversations, resulting in more engaging and coherent interactions.

**RECAP improves persona-based dialogue by retrieving relevant information and fusing it with conversation context through hierarchical transformers and advanced attention mechanisms.**

While S. Liu et al. (2023) also focus on retrieving persona descriptions implicitly, it specifically addresses weaknesses in both the retrieval and fusion phases. The authors propose RECAP (Retrieval-Enhanced Context Aware Prefix), a retrieval-fusion approach where persona-relevant information is retrieved first, then fused with the conversation context during decoding. To enhance persona communication, RECAP employs a hierarchical transformer retriever specifically trained to identify information that effectively reflects the user’s persona. Furthermore, the model utilizes a continuous pre-layer prefix approach along with a two-step cross-attention mechanism, both of which have been shown to be beneficial for dialogue generation. These advancements contribute to RECAP’s ability to generate more engaging and user-specific dialogue responses.

**Automatically inferring user personas from conversation history for flexible and scalable personalization, using a Persona Approximator and Persona Generator.**

Building on the idea of implicit persona for dialogue personalization, W. Zhou et al. (2023) propose a method to automatically infer the user’s persona directly from conversation history. This approach eliminates the need for pre-defined persona descriptions during inference, allowing for more flexible and scalable personalization. The core of their work lies in two persona detection approaches: a Persona Approximator that analyzes conversation history to mimic the output of a separate persona encoder trained on explicit descriptions, and a Persona Generator that treats persona detection as a text generation task, directly creating a persona description for the speaker based solely on the conversation history. Experiments on dialogue datasets with and without persona description demonstrate the effectiveness of this approach.

**Enhancing persona-based dialogue models through data distillation, diversification, and curriculum training for improved robustness and performance.**

Lastly, Cao et al. (2022) also address the challenges associated with limited persona-based dialogue data and the complexity of such data. The approach involves manipulating the training data to make it easier for the model to learn from. This method is model-agnostic and can be packed with any base model to improve robustness and consistency. This method tackles the challenges by

first simplifying the training data through a process called data distillation. This step focuses on extracting essential information from the original data, creating simpler examples for the model to learn from. Following distillation, the method employs various techniques for data diversification. Diversification helps the model learn from a wider range of scenarios and prevents over fitting on the simplified data. Finally, the method utilizes a data curriculum, where the model is trained first on the easier, augmented, distilled data and then on the more complex original data. This staged approach allows the model to progressively learn from the complexities of persona-based dialogue data. Experimental results show that this method effectively improves the performance of the dialogue models used in this study.

<b>ID</b>	<b>Model</b>	<b>Theoretical Framework</b>	<b>Methodology</b>	<b>Core Mechanism</b>
19	Persona-Grounded Dialogue System	Addresses WWH problem in open-domain dialogue	Multi-session Personalized Conversation dataset and adjustable blending weights	Fine-tuning
20	Modular Prompted Chatbot (MPC)	Tackles data scarcity and training costs in persona-based chatbots	Utilize prompt-based techniques like few-shot in-context learning and Chain-of-Thought for long-term coherence	Prompting
21	In-Context Prompting Learning	Improves zero-shot persona consistency	Pre-train a persona-augmented dialogue generation model and utilize in-context prompting	Prompting
22	PLATO-LTM	Enhances long-term persona consistency	Extracts, stores, and integrates user and chatbot personas in real time	Prompting
23	RECAP	Improves persona-based dialogue with retrieval and fusion	Retrieval-enhanced context-aware prefix with hierarchical transformers	Retrieval Augmented Generation
24	Automatic Persona Inference	Infers user personas from conversation history	Persona Approximator and Persona Generator for flexible personalization	Prompting
25	Data Distillation Approach	Enhances persona-based dialogue robustness	Data distillation, diversification and curriculum training	Fine-tuning

Table 8: Overview of models and techniques to make conversations persona-based

#### 4.1.7 Emotional support through chit-chat

This section delves into existing research on emotional support conversation (ESC) systems, which can be particularly valuable for elderly users. Jiang et al. (2023) highlight that studies have shown that chit-chat can help elderly with cognitive restoration. Additionally, studies have demonstrated the positive impact of emotional support on maintaining and enhancing cognitive function in older adults. Integrating chit-chat capabilities into the dialogue system creates an opportunity to provide emotional support, potentially contributing to the cognitive well-being of elderly users.

##### **Advanced empathetic response generation using DiffusEMP framework for nuanced empathy expression.**

Empathy is a crucial factor in emotionally engaging conversation. However existing emotional response generation models often struggle to express nuanced empathy beyond generic phrases. Bi et al. (2023) address this limitation with DiffusEMP, a framework that leverages diffusion models for empathetic response generations. DiffusEMP incorporates multi-grained control signals to better model multi-dimensional empathy and improve its expression. These multi-grained control signals are used at the utterance, sentence and token levels and are extracted from the golden responses during training. Response candidates obtained from an emotion enhanced matching method are used as the signal source. Additionally, a control range masking strategy is employed in the diffusion language model to ensure precise control over the generation of response tokens. This allows DiffusEMP to produce responses that are not only relevant to the conversation but also express empathy tailored to the specific situation.

##### **SUPPORTER: A reinforcement learning-based system for balancing empathy and elicitation in multi-turn emotional support conversations.**

Another challenge in ESC is that models often struggle to lay out an entire ESC process and effectively improve a user’s mental state. This is a difficult task because the agent must handle complex situations with fluctuating emotions and respond appropriately. Additionally, it is crucial to balance empathy and elicitation when aiming to improve someone’s mental state. The agent must also maintain general conversational goals, such as coherence. To address these challenges, J. Zhou, Chen, et al. (2023) introduce SUPPORTER a system designed to facilitate multi-turn emotional support conversations with positive emotion elicitation using a mixture-of-experts (MoE) based reinforcement learning approach. SUPPORTER employs a model that receives rewards based on how effectively it improves the user’s emotional state during multi-turn interactions and how coherent the conversation is. This approach enables the system to learn and adapt its response strategies, promoting a more positive and supportive conversation flow while maintaining coherence and balancing empathy with elicitation.

##### **Dynamic commonsense knowledge selection for improved empathy and coherence in conversation models.**

Incorporating commonsense knowledge into conversation models offers another path to generating empathetic responses. However, a significant challenge lies in merging this knowledge effectively. Straightforward merging methods can confuse the system, leading to a decline in response consistency. Cai et al. (2023) propose a solution that addresses this issue by focusing on selecting the most appropriate commonsense knowledge based on the user’s emotional state. This approach

ensures that the generated responses maintain consistency with the conversation history while expressing a greater degree of empathy. By dynamically selecting relevant knowledge, Cai et al. (2023) avoid overwhelming the system and enhance the overall coherence of the generated responses.

### **Cognitive stimulation dialogue system with multi-source knowledge fusion for elderly users.**

Focusing on the specific needs of elderly users, Jiang et al. (2023) propose a cognitive stimulation dialogue (CSD) system that leverages multi-source knowledge fusion. This approach integrates knowledge about cognitive health principles and user needs to enrich conversation content and generate responses that benefit mental well-being. To address the challenge of data sparsity, the system employs a progressive masking method during training. This method involves masking random tokens with external knowledge, enabling the model to learn effective encoders for classifying dialogue responses based on embedded cognitive stimulation principles and emotional support strategies.

### **PAL: A reinforcement learning-based dialogue system for generating polite and empathetic responses in counseling conversations.**

As Mishra et al. (2023) highlight, it is important for a counseling conversational agent to adapt its support to the users' mental state whilst being polite and empathetic. Active listening combined with accurate interpretation of the user's emotions and meaning is crucial for effective counseling. This approach demonstrates care and understanding, fostering a sense of being heard and valued. By aligning responses with the client's emotions, reassurance and compassion can be effectively offered. In order to do this, Mishra et al. (2023) proposed the PAL (Polite and empAthetic counSELing) dialogue system, which creates a comfortable and engaging atmosphere during counseling. To achieve this, two counseling conversation datasets with labels for emotions, politeness and empathy are developed and annotated. The PAL system is built by training a language model on these datasets and then fine-tuning it using reinforcement learning to ensure responses are polite, empathetic, consistent and adaptive. This approach allows PAL to generate engaging and fluent responses that adapt to the client's emotions.

### **PAL: A persona-augmented conversational model for personalized and emotionally supportive responses.**

In addition to being adaptive to the user's emotions, taking their personality into account can significantly enhance emotional support. As highlighted by Cheng et al. (2023), effective emotional support benefits from a thorough understanding of the user's personality. Using more specific and persona-related language fosters long-term connection with the user. Consequently, Cheng et al. (2023) propose PAL, which stands for Persona-Augmented Emotional Support. PAL is a conversational model designed to dynamically utilize the seeker's persona to generate more informative and personalized responses for effective emotional support. PAL achieves this by extracting persona information from earlier conversations and employing attention mechanism to deepen the understanding of the user. Additionally, it uses a strategy-based controllable generation method to actively incorporate persona information into responses, ensuring that the support provided is both relevant and personalized.

**TransESC: Enhancing flow and naturalness in emotional support conversations through smooth state transitions.** Another crucial aspect is ensuring the flow and naturalness of emotional support conversations, which can significantly improve the user experience. To achieve this, W. Zhao et al. (2023) propose the Transitions of ESC (TransESC) system. This system emphasized smooth transitions between dialogue turns by considering three types of state transitions: 1) semantics, referring to the topic and meaning, 2) strategy, the approach used to provide emotional support and 3) emotion, the user’s emotional state. TransESC leverages a model that explicitly considers and integrates these state transitions during the generation of emotional support responses.

ID	Model	Theoretical Framework	Methodology	Core Mechanism
26	DiffusEMP	Addresses limitations in nuances empathy expression	Uses diffusion models and multi-grained control signals for response generation	Prompting
27	SUPPORTER	Balances empathy and elicitation in emotional support conversations	Reinforcement learning with mixture-of-experts for multi-turn interactions	Fine-tuning
28	Dynamic Commonsense Knowledge Selection	Enhances empathy and coherence by selecting relevant knowledge	Selects commonsense knowledge based on emotional state for consistency	Retrieval Augmented Generation
29	Cognitive Stimulation Dialogue System	Focuses on cognitive stimulation and data sparsity for elderly users	Multi-source knowledge fusion and progressive masking during training	Retrieval Augmented Generation
30	PAL (Polite and Empathetic Counselor)	Adapts responses to user’s mental state and emotional needs	Reinforcement learning with emotion, politeness and empathy labels	Fine-tuning
31	PAL (Persona-Augmented Emotional Support)	Enhances emotional support by integrating user persona	Utilizes persona information and controllable generation methods	Fine-tuning
32	TransESC	Improves flow and naturalness in emotional support conversations	Considers semantic, strategy and emotional state transition	Fine-tuning

Table 9: Overview of models and techniques to provide emotional support through chat

## 4.2 Base Program Evaluation

This section evaluates the performance of the base program developed for this thesis. The evaluation results are presented in two tables. During the evaluation, a total of 10 test conversations were conducted by 4 evaluators with varying levels of experience with AI. Table 10 summarizes the feedback from evaluators on the first four questions of the questionnaire. The results indicate a favorable assessment of the program, with conversations being rated as clear, entertaining and natural. These positive ratings suggest that the program performs well in natural terms of user

experience and interaction quality.

<b>Score</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>General score conversation</b>	0%	0%	0%	70%	30%
<b>Natural flow conversation</b>	0%	0%	40%	50%	10%
<b>Clearness conversation</b>	0%	0%	10%	20%	70%
<b>Entertainment level conversation</b>	0%	0%	10%	30%	60%

Table 10: Distribution of ratings for various aspects of the base program

Table 11 provides an analysis of hallucinations in the dialogue responses. It shows that 100% of the conversations were free from any hallucinations. This indicates that the program maintains a high level of accuracy and reliability in its responses with minimal errors.

	<b>Occurrence</b>
<b>No hallucination</b>	100%
<b>Input-related hallucination</b>	0%
<b>Context-related hallucination</b>	0%
<b>Fact-related hallucination</b>	0%

Table 11: Distribution of hallucination types in dialogue responses

Evaluators also had the chance to report any other problems with the conversation. A few things can be concluded from those responses.

1. Ted occasionally attempts to prematurely conclude the small talk, leading it to say something like "Let's go to the next question" while still engaged in the small talk loop. This results in an awkward situation, as the next question has not been asked yet, and the program is waiting for a user response. Table 13 presents an instance of this in a conversation with Evaluator 3.
2. Sometimes, Ted may ask if the user wants to proceed to the next question or pose a small talk-related question, even after the small talk loop has ended. As a result, the user is presented with a choice or question but is unable to respond. Table 12 provides an example of this situation occurring in a conversation with Evaluator 4.
3. Ted often repeats part of the user response in its reply, which can become annoying over time and makes the conversation feel unnatural. Table 13 illustrates a subtle example. Ted's response, "It's great to hear that the pills you're taking can help with the pain!", may seem like a sincere reaction. However, such replies can sometimes come across as repetitive since similar responses are given after nearly every question.

	<b>Original</b>	<b>English translation</b>
Ted	<p>De volgende vraag gaat over dagelijkse activiteiten (bijv. werk, studie, huishouden, gezins- of vrijetijdsactiviteiten). Zou u zeggen dat:</p> <p>U geen problemen heeft met het uitvoeren van uw dagelijkse activiteiten?</p> <p>U een beetje problemen heeft met het uitvoeren van uw dagelijkse activiteiten?</p> <p>U matige problemen heeft met het uitvoeren van uw dagelijkse activiteiten?</p> <p>U ernstige problemen heeft met het uitvoeren van uw dagelijkse activiteiten?</p> <p>U niet in staat bent om uw dagelijkse activiteiten uit te voeren?</p>	<p>The following question is about daily activities (e.g., work, study, household chores, family or leisure activities). Would you say that:</p> <p>You have no problems performing your daily activities?</p> <p>You have a few problems performing your daily activities?</p> <p>You have moderate problems performing your daily activities?</p> <p>You have severe problems performing your daily activities?</p> <p>You are unable to perform your daily activities?</p>
Evaluator 4	<p>Ik heb een hulp voor het huishouden, ik hou me dus vooral bezig met vrijetijdsactiviteiten. Die gaan best prima. Gisteren heb ik een leuke bridgemiddag gehad met mijn vrienden.</p>	<p>I have help with household chores, so I mostly focus on leisure activities. Those are going quite well. Yesterday, I had a nice afternoon of bridge with my friends.</p>
Ted	<p>Wat fijn dat je hulp hebt voor het huishouden, dat kan zeker wat extra ruimte geven om te genieten van je vrijetijdsactiviteiten! Een bridgemiddag klinkt als een gezellige en leuke manier om je tijd door te brengen. Het is belangrijk om momenten te hebben waarop je kunt ontspannen en plezier kunt hebben met vrienden. Bridge is een geweldige sociale activiteit waarbij je je geest scherp houdt. Heb je altijd al bridge gespeeld of is het iets nieuws wat je hebt ontdekt? Ik ben benieuwd naar je ervaringen! Laten we doorgaan met de volgende vraag. Nu wil ik je vragen over pijn of ongemak. Zou u zeggen dat:</p> <p>U geen pijn of ongemak heeft?</p> <p>U een beetje pijn of ongemak heeft?</p> <p>U matige pijn of ongemak heeft?</p> <p>U ernstige pijn of ongemak heeft?</p> <p>U extreme pijn of ongemak heeft?</p>	<p>It's great that you have help with household chores; it definitely gives you some extra space to enjoy your leisure activities! An afternoon of bridge sounds like a pleasant and enjoyable way to spend your time. It's important to have moments where you can relax and have fun with friends. Bridge is a fantastic social activity that keeps your mind sharp. Have you always played bridge, or is it something new you've discovered? I'm curious about your experiences!</p> <p>Let's move on to the next question.</p> <p>Now, I'd like to ask you about pain or discomfort. Would you say that:</p> <p>You have no pain or discomfort?</p> <p>You have a little pain or discomfort?</p> <p>You have moderate pain or discomfort?</p> <p>You have severe pain or discomfort?</p> <p>You have extreme pain or discomfort?</p>

Table 12: Ted and evaluator 4 example conversation



	Original	English translation
Ted	De volgende vraag gaat over pijn of ongemak. Zou u zeggen dat:  U geen pijn of ongemak heeft? U een beetje pijn of ongemak heeft? U matige pijn of ongemak heeft? U ernstige pijn of ongemak heeft? U extreme pijn of ongemak heeft?	The following question is about pain or discomfort. Would you say that:  You have no pain or discomfort? You have a little pain or discomfort? You have moderate pain or discomfort? You have severe pain or discomfort? You have extreme pain or discomfort?
Evaluator 3	Ja ik heb wel wat pillen die ik dagelijks moet slikken, die helpen wel tegen de pijn, dus dankzij die pillen zou ik zeggen b of c	Yes, I do have some pills that I take daily, which help with the pain, so thanks to those pills, I would say b or c
Ted	Fijn om te horen dat de pillen die je slikt kunnen helpen tegen de pijn! Het is belangrijk om de juiste zorg en medicatie te krijgen om je welzijn te ondersteunen. Het is goed dat je iets hebt gevonden wat voor jou werkt. Dankbaar voor moderne geneeskunde, toch? Laten we doorgaan met de volgende vraag.	It's great to hear that the pills you're taking can help with the pain! It's important to get the right care and medication to support your well-being. It's good that you've found something that works for you. Grateful for modern medicine, right? Let's move on to the next question.
User	Wat is de volgende vraag?	What is the next question?

Table 13: Ted and evaluator 3 example conversation

### 4.3 Combining Literature and Software

The findings from the systematic literature review provide valuable insights for enhancing the Welzijn.AI base program. By integrating the strategies and models discussed in the review, Welzijn.AI can be further refined to reduce hallucinations, improve the naturalness and engagement of conversations, and offer better emotional support. This section discusses how specific papers and methodologies from the literature can be applied to the future development of Welzijn.AI

Reducing hallucinations in dialogue systems is crucial for ensuring the reliability and trustworthiness of Welzijn.AI. The model used for this study has not been fine-tuned specifically for Welzijn.AI and its purposes. If this is done in the future, ACK-DEF can be used to incorporate variations of existing knowledge. The model can then be trained to better detect and manage incomplete and inaccurate information. Furthermore, the TBS framework can be implemented to ensure that the model is able to incorporate commonsense knowledge into its responses, which will result in more accurate and coherent responses, thereby minimizing the risk of hallucinations.

As discussed above, the base program still struggles with generating engaging interactions without repeating user answers. To provide meaningful and engaging interactions, for example DiffusEMP can be used to generate more nuanced and empathetic responses. Furthermore, the PLATO-LTM Framework can be used to remember user and chat bot personas in real-time, ensuring that the conversation remains coherent and engaging over multiple sessions, which is particularly beneficial when the same questions are asked daily. The PAL (Persona-Augmented Emotional Support) model can also be used to adapt to the user's emotions and personality, ensuring that the conversation is centered around the user and is not generic.

The evaluation of the base program provided valuable insights into areas for improvement. In addition to the need to reduce repetition in responses, the premature conclusion of small talk and

repetitive responses were identified as issues. To address these, techniques like SimOAP method and UniDS framework can be integrated to allow the system to seamlessly switch between entertaining conversations and more focused, task-oriented dialogue.

By systematically integrating these literature based techniques into Welzijn.AI, the software can be significantly improved in terms of reducing hallucinations, enhancing conversational engagement and providing better emotional support. These advancements will not only improve the user experience but also ensure that Welzijn.AI serves as a reliable and supportive tool for its users.

## 5 Conclusions and Discussion

### 5.1 Conclusion

This study aimed to address the research question: *How can a questionnaire-driven approach effectively guide dialogue generation in Large Language Models to ensure coherence and mitigate the risk of hallucinations in conversational AI systems designed for elderly well-being monitoring?*

To answer this research question, the study focuses on two main activities. First, a systematic literature review was conducted to gain an overview of current methods for mitigating hallucinations and improving coherence in AI-generated dialogues. This review also explored ways to make conversation engaging and personalized for users. The findings provide a foundation for further research and development of Welzijn.AI. The literature review highlighted various approaches and strategies that can be implemented to ensure that dialogue systems remain coherent and meaningful, while also reducing the likelihood of producing inaccurate or misleading information, commonly referred to as hallucinations.

Figure 5 and 6 present a summary of the methods discussed in Chapter 4 together with the IDs of the papers. The categorization of approaches in the provided diagrams was based on the primary focus and methodology used in mitigating hallucinations and engaging users in dialogue systems. Figure 5 organizes the measures for hallucination mitigation into three main categories: knowledge-based approach, planning-based approach and focus- and control-based approach. The knowledge-based approach includes methods that rely on integrating and manipulating external knowledge to enhance response accuracy and reduce hallucinations. The planning-based approach focuses on strategies that involve structured dialogue planning to ensure coherence and target alignment throughout the conversation. Finally, the focus- and control-based approach centers on refining the dialogue flow through global and dynamic content planning and prioritizing user-specific interests.

In Figure 6, the focus shifts to engaging the user, which is split into two major categories: persona-based dialogue and emotional support. The persona-based dialogue group contains methods aimed at enhancing user engagement through personalized interactions. This group also includes methods to address challenges in managing explicit persona descriptions and diversifying training data. The emotional support category includes techniques designed to provide empathetic and emotionally resonant responses.

These categorizations help in systematically applying different strategies depending on the primary objective, whether it is mitigating hallucinations or enhancing user engagement in conversations.

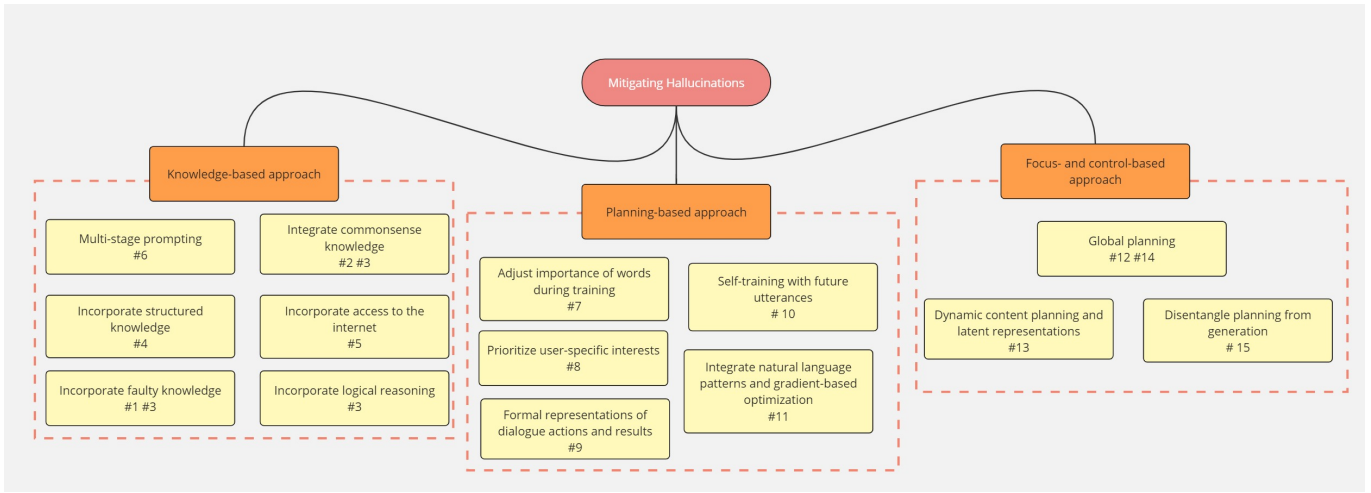


Figure 5: Overview of discussed measures for hallucination mitigation

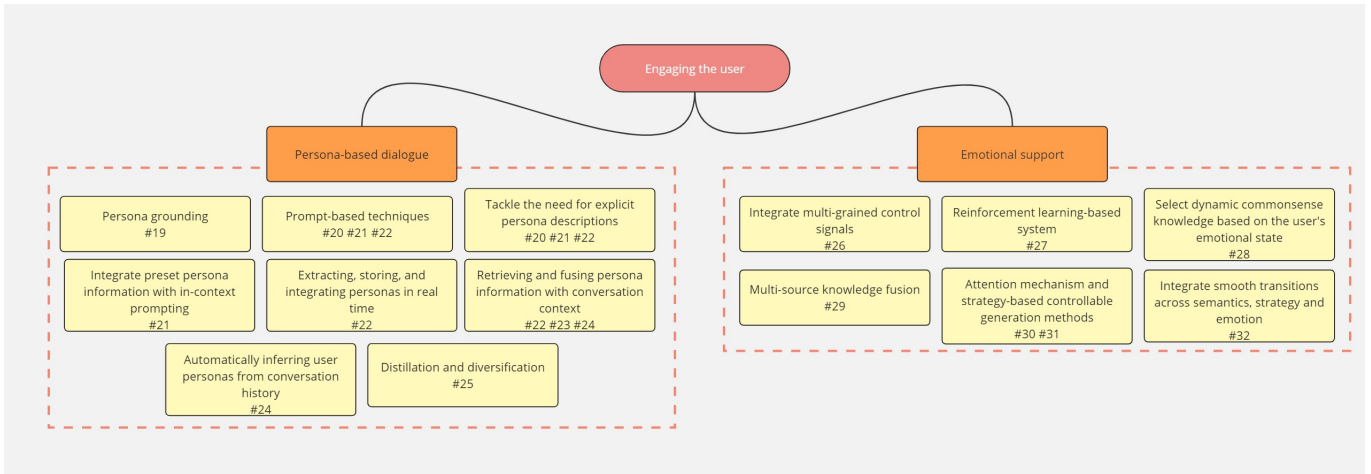


Figure 6: Overview of discussed measures for user engagement

The second part of the study focused on developing a base program for Welzijn.AI, which is intended as a foundation for further development. This program helps to answer the research question as the conversations center around a questionnaire. The evaluation of the program indicated that a questionnaire-driven approach can indeed ensure coherence and reduce the risk of hallucinations in conversational AI systems designed for elderly well-being monitoring. By limiting the conversation primarily to the questionnaire, with only a few diversions for small talk, hallucinations are effectively mitigated. Furthermore, the developed program exhibits a natural flow of conversation that is both engaging and clear for users. These results demonstrate that using a structured approach, such as a questionnaire, can provide the necessary guidance to maintain high-quality interactions with minimal errors.

## 5.2 Discussion and Further Research

While the results are promising, there are several areas for improvement in this study, as highlighted in Chapter 4. One key area is the tendency of the program to repeat parts of the user's messages, which can become annoying over time and diminish the conversational experience. Enhancing the diversity and creativity of responses could make interactions more engaging and feel more natural to users. Additionally, the program occasionally closes small talk rounds prematurely, leading to abrupt transitions that can disrupt the conversation flow. Improving the logic governing conversation transitions could enhance user experience by making interactions smoother and more intuitive.

Another area of improvement is the scoring system. Currently, GPT-3.5 Turbo is used to match user messages to predefined multiple-choice answers. While this approach provides a basic framework for understanding user inputs, it could be enhanced by incorporating findings from R. Wang et al. (2024) to improve accuracy and relevance in matching user inputs to responses.

Furthermore, it is important to note that the sample size of the evaluation conducted in this study was relatively small. Although the results provide valuable insights, the limited number of test conversations may not fully capture the range of potential user interactions or the robustness of the program's performance across diverse scenarios. Expanding the sample size in future evaluations will be crucial to validate the findings, uncover additional areas of improvement and ensure that the enhancements to Welzijn.AI are well-tested and reliable across a broader spectrum of users.

Overall, these insights will guide the ongoing development and refinement of Welzijn.AI, ensuring it better meets the needs of elderly users while maintaining coherence and reducing the risk of hallucinations.

## References

- Belani, R., & Flanigan, J. (2023, July). Automatic identification of code-switching functions in speech transcripts. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 7438–7448). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.469>
- Beygi, S., Fazel-Zarandi, M., Cervone, A., Krishnan, P., & Jonnalagadda, S. (2022, May). Logical reasoning for task oriented dialogue systems. In S. Malmasi, O. Rokhlenko, N. Ueffing, I. Guy, E. Agichtein, & S. Kallumadi (Eds.), *Proceedings of the fifth workshop on e-commerce and nlp (ecnlp 5)* (pp. 68–79). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.ecnlp-1.10>
- Bi, G., Shen, L., Cao, Y., Chen, M., Xie, Y., Lin, Z., & He, X. (2023, July). DiffusEmp: A diffusion model-based framework with multi-grained control for empathetic response generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2812–2831). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.158>
- Cai, H., Shen, X., Xu, Q., Shen, W., Wang, X., Ge, W., Zheng, X., & Xue, X. (2023, July). Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 7858–7873). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.498>
- Cao, Y., Bi, W., Fang, M., Shi, S., & Tao, D. (2022, May). A model-agnostic data manipulation method for persona-based dialogue generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7984–8002). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.550>
- Chen, M., Yu, X., Shi, W., Awasthi, U., & Yu, Z. (2023, July). Controllable mixed-initiative dialogue generation through prompting. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 951–966). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-short.82>
- Cheng, J., Sabour, S., Sun, H., Chen, Z., & Huang, M. (2023, July). PAL: Persona-augmented emotional support conversation generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 535–554). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.34>
- Chiu, S., Li, M., Lin, Y.-T., & Chen, Y.-N. (2022, May). SalesBot: Transitioning from chit-chat to task-oriented dialogues. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6143–6158). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.425>
- Deng, Y., Zhang, X., Huang, H., & Hu, Y. (2023, July). Towards faithful dialogues via focus learning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4554–4566). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.250>

- Fan, Y., & Hunter, A. (2023, July). Understanding the cooking process with English recipe text. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 4244–4264). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.261>
- Fang, H., Balakrishnan, A., Jhamtani, H., Bufe, J., Crawford, J., Krishnamurthy, J., Pauls, A., Eisner, J., Andreas, J., & Klein, D. (2023, July). The whole truth and nothing but the truth: Faithful and controllable dialogue response generation with dataflow transduction and constrained decoding. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 5682–5700). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.351>
- Hu, Z., Chan, H. P., Liu, J., Xiao, X., Wu, H., & Huang, L. (2022, May). PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2288–2305). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.163>
- Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M., & Fung, P. (2023, July). RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 4504–4522). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.275>
- Jiang, J., Wang, S., Li, Q., Kong, L., & Wu, C. (2023, July). A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 10628–10640). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.593>
- Komeili, M., Shuster, K., & Weston, J. (2022, May). Internet-augmented dialogue generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8460–8478). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.579>
- Kwon, D., Lee, S., Kim, K. H., Lee, S., Kim, T., & Davis, E. (2023, July). What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In S. Sitaram, B. Beigman Klebanov, & J. D. Williams (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 5: Industry track)* (pp. 707–719). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-industry.68>
- Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., & Lee, K. (2023, July). Prompted LLMs as chatbot modules for long open-domain conversation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 4536–4554). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.277>
- Liu, A., Wang, B., Tan, Y., Zhao, D., Huang, K., He, R., & Hou, Y. (2023, July). MTGP: Multi-turn target-oriented dialogue guided by generative global path with flexible turns. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 259–271). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.18>

- Liu, S., Cho, H., Freedman, M., Ma, X., & May, J. (2023, July). RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8404–8419). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.468>
- Liu, Z., Patwary, M., Prenger, R., Prabhumoye, S., Ping, W., Shoeybi, M., & Catanzaro, B. (2022, May). Multi-stage prompting for knowledgeable dialogue generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the association for computational linguistics: Acl 2022* (pp. 1317–1337). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-acl.104>
- Mishra, K., Priya, P., & Ekbal, A. (2023, July). PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 12254–12271). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.685>
- Otani, N., Araki, J., Kim, H., & Hovy, E. (2023, July). On the underspecification of situations in open-domain conversational datasets. In Y.-N. Chen & A. Rastogi (Eds.), *Proceedings of the 5th workshop on nlp for conversational ai (nlp4convai 2023)* (pp. 12–28). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.nlp4convai-1.2>
- Salahudeen, S. A., Lawan, F. I., Wali, A., Imam, A. A., Shuaibu, A. R., Yusuf, A., Rabi, N. B., Bello, M., Adamu, S. U., & Aliyu, S. M. (2023, July). HausaNLP at SemEval-2023 task 12: Leveraging African low resource TweetData for sentiment analysis. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), *Proceedings of the 17th international workshop on semantic evaluation (semeval-2023)* (pp. 50–57). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.semeval-1.6>
- Sun, B., Li, Y., Mi, F., Bie, F., Li, Y., & Li, K. (2023, July). Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 1741–1750). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-short.148>
- Sun, W., Chen, Y., Tao, G., Fang, C., Zhang, X., Zhang, Q., & Luo, B. (2023, July). Backdooring neural code search. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9692–9708). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.540>
- The EuroQol Group. (1990). Euroqol 5d. <https://meetinstrumentenzorg.nl/instrumenten/euroqol-5d/>
- Valizadeh, M., & Parde, N. (2022, May). The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6638–6660). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.458>
- van Dijk, B., et al. (In review). Kan technologie de vereenzaming van ouderen tegengaan?



- van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., & Spruit, M. (2021). Symbols: A systematic review methodology blending active learning and snowballing. *Frontiers in research metrics and analytics*, 6, 685591.
- Wang, J., Lin, D., & Li, W. (2023, July). Dialogue planning via brownian bridge stochastic process for goal-directed proactive dialogue. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 370–387). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.25>
- Wang, R., Verberne, S., & Spruit, M. (2024). Attend all options at once: Full context input for multi-choice reading comprehension. In N. Goharian, N. Tonello, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval* (pp. 387–402). Springer Nature Switzerland.
- Wen, Z., Tian, Z., Huang, Z., Yang, Y., Jian, Z., Wang, C., & Li, D. (2023, July). GRACE: Gradient-guided controllable retrieval for augmenting attribute-based text generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 8377–8398). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.530>
- Wu, Y., Lu, W., Zhang, Y., Jatowt, A., Feng, J., Sun, C., Wu, F., & Kuang, K. (2023, July). Focus-aware response generation in inquiry conversation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 12585–12599). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.797>
- Xu, X., Gou, Z., Wu, W., Niu, Z.-Y., Wu, H., Wang, H., & Wang, S. (2022, May). Long time no see! open-domain conversation with long-term persona memory. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the association for computational linguistics: Acl 2022* (pp. 2639–2650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-acl.207>
- Xu, X., Lei, Z., Wu, W., Niu, Z.-Y., Wu, H., & Wang, H. (2023, July). Towards zero-shot persona dialogue generation with in-context learning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 1387–1398). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.90>
- Zeng, W., He, K., Wang, Y., Zeng, C., Wang, J., Xian, Y., & Xu, W. (2023, July). FutureTOD: Teaching future knowledge to pre-trained language model for task-oriented dialogue. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6532–6546). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.360>
- Zeng, W., Zhao, L., He, K., Geng, R., Wang, J., Wu, W., & Xu, W. (2023, July). Seen to unseen: Exploring compositional generalization of multi-attribute controllable dialogue generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14179–14196). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.793>
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren’s song in the ai ocean: A survey on hallucination in large language models. <https://arxiv.org/abs/2309.01219>
- Zhao, W., Zhao, Y., Wang, S., & Qin, B. (2023, July). TransESC: Smoothing emotional support conversation via turn-level state transition. In A. Rogers, J. Boyd-Graber, & N. Okazaki

- (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 6725–6739). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.420>
- Zhao, X., He, B., Wang, Y., Li, Y., Mi, F., Liu, Y., Jiang, X., Liu, Q., & Chen, H. (2022, May). UniDS: A unified dialogue system for chit-chat and task-oriented dialogues. In S. Feng, H. Wan, C. Yuan, & H. Yu (Eds.), *Proceedings of the second dialdoc workshop on document-grounded dialogue and conversational question answering* (pp. 13–22). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.dialdoc-1.2>
- Zhou, J., Chen, Z., Wang, B., & Huang, M. (2023, July). Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1714–1729). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.96>
- Zhou, J., Pang, L., Shen, H., & Cheng, X. (2023, July). SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9945–9959). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.553>
- Zhou, P., Gopalakrishnan, K., Hedayatnia, B., Kim, S., Pujara, J., Ren, X., Liu, Y., & Hakkani-Tur, D. (2022, May). Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1237–1252). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.88>
- Zhou, W., Li, Q., & Li, C. (2023, July). Learning to predict persona information for dialogue personalization without explicit persona description. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 2979–2991). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.186>