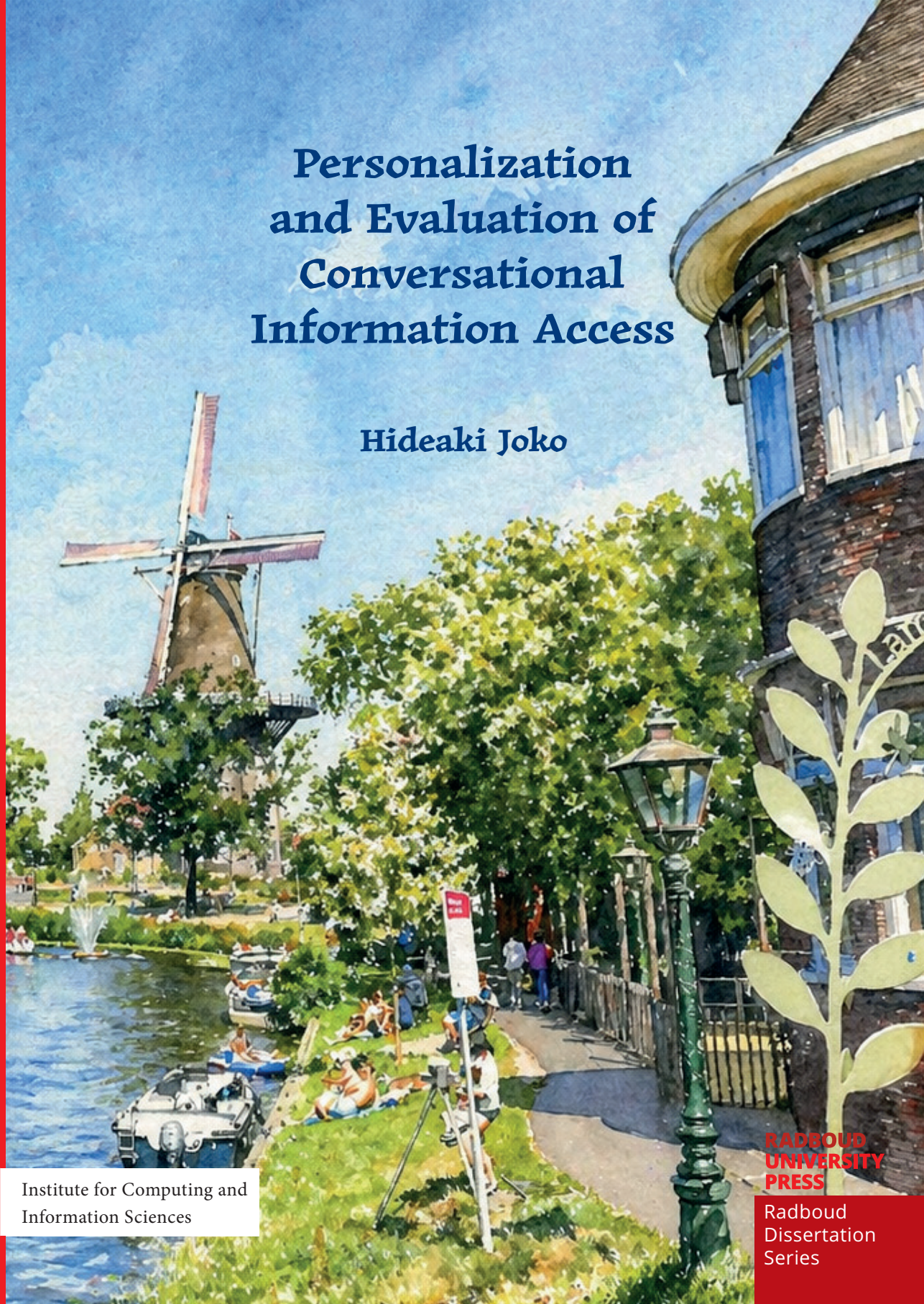
Personalization and Evaluation of Conversational Information Access

Hideaki Joko



Institute for Computing and Information Sciences

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Personalization and Evaluation of Conversational Information Access

Hideaki Joko

SIKS Dissertation Series No. 2026-41
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Personalization and Evaluation of Conversational Information Access
Hideaki Joko

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Hideaki Joko
Cover: Proefschrift AIO | Guus Gijben

ISBN: 9789465152028
DOI: 10.54195/9789465152028

RADBOUD
UNIVERSITY
PRESS

Personalization and Evaluation of Conversational Information Access

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 8 juli 2026

om 12.30 uur precies

door

Hideaki Joko

Promotor:
Prof. dr. A.P. (Arjen) de Vries

Copromotor:
Dr. F. (Faegheh) Hasibi

Manuscriptcommissie:

Prof. dr. M.A. (Martha) Larson	
Prof. dr. T. (Tetsuya) Sakai	(Waseda University, Japan)
Prof. dr. S. (Suzan) Verberne	(Universiteit Leiden)
Dr. M. (Mohammad) Alian Nejadi	(Universiteit van Amsterdam)
Dr. A. (Avishek) Anand	(Technische Universiteit Delft)

Contents

1 Introduction **1**
- 1.1 Research Objectives and Questions . . . 2
  - 1.1.1 Personal Information Extraction . . . 2
  - 1.1.2 Personalized Response Generation . . . 5
  - 1.1.3 Conversation Evaluation . . . 7
- 1.2 Contributions . . . 9
- 1.3 Thesis Overview . . . 10
- 1.4 Origins . . . 11

2 Conversational Entity Linking **13**
- 2.1 Introduction . . . 14
- 2.2 Related Work . . . 17
  - 2.2.1 Entity Linking . . . 17
  - 2.2.2 Personal Entities . . . 18
- 2.3 Dataset Selection . . . 19
- 2.4 Entity Annotation Process . . . 21
  - 2.4.1 Concepts and Named Entities . . . 22
  - 2.4.2 Personal Entities . . . 24
  - 2.4.3 Dialogue Selection and Annotation . . . 25
- 2.5 Annotation Results . . . 26
- 2.6 Conclusion . . . 29

3 Entity Linking for Personalization **31**
- 3.1 Introduction . . . 32
- 3.2 Method . . . 34
  - 3.2.1 Personal Entity Linking . . . 35
  - 3.2.2 Concept and Named Entity Linking . . . 36
- 3.3 The ConEL-2 Collection . . . 36

3.4 Results . . . 39
3.4.1 Personal Entity Linking . . . 39
3.4.2 Concept and Named Entity Linking . . . 40
3.4.3 Overall Performance . . . 42
3.5 Conclusion . . . 42

4 Personalized Response Generation 45
4.1 Introduction . . . 46
4.2 Related Work . . . 49
4.2.1 Dialogue Collection Methods . . . 49
4.2.2 Personalized Conversational Systems . . . 51
4.3 Dialogue Collection Method . . . 51
4.3.1 Task Formulation . . . 52
4.3.2 Guidance Generation . . . 54
4.3.3 Dialogue Act Classification . . . 55
4.3.4 Utterance Composition . . . 56
4.3.5 Preference Extraction . . . 57
4.3.6 Evaluation . . . 58
4.3.7 Experimental Setup . . . 61
4.4 Personal Recommendation Method . . . 62
4.4.1 Preference Extraction . . . 62
4.4.2 Personalized Recommendation . . . 63
4.4.3 Evaluation . . . 63
4.4.4 Experimental Setup . . . 65
4.5 Results . . . 65
4.5.1 Dialogue Collection Evaluation . . . 65
4.5.2 Personal Recommendation Evaluation . . . 69
4.6 Conclusion . . . 72

5 Conversation Evaluation 73
5.1 Introduction . . . 74
5.2 Related Work . . . 78
5.3 Method . . . 81
5.3.1 Conversation Particle Generation . . . 81
5.3.2 Evaluation Score Computation . . . 82
5.3.3 Instruction Optimization . . . 83
5.4 Human Annotation Collection . . . 87
5.4.1 Dialogue Collection . . . 87

5.4.2 Dialogue Annotation 88
5.4.3 Interface 90
5.4.4 Participants and Quality Control 90
5.4.5 Analysis 90
5.5 Experimental Setup 92
5.6 Results 94
5.6.1 FACE Annotation Correlation 94
5.6.2 Generalizability of FACE 96
5.6.3 FACE Interpretability 99
5.7 Analysis 101
5.8 Conclusion 102

6 Conclusion **103**
6.1 Answers to Research Questions 103
6.1.1 Personal Information Extraction 103
6.1.2 Personalized Response Generation 105
6.1.3 Conversation Evaluation 106
6.2 Positioning of the Thesis 107
6.3 Limitations 108
6.4 Future Directions 109
6.4.1 Personal Information Extraction 109
6.4.2 Personalized Response Generation 110
6.4.3 Conversation Evaluation 111

A Resources **113**
A.1 Code Repositories 113
A.2 Datasets 114

B Prompt Examples of FACE **115**
B.1 Prompts Used in FACE 115
B.2 Instructions Optimized by FACE 117

Bibliography **119**

Summary **137**

Samenvatting **139**

Acknowledgements **141**

Research Data Management **143**

Curriculum Vitae **145**

SIKS Dissertation Series **147**

Chapter 1

Introduction

The dream of having an intelligent conversation in natural language with a computer has fascinated humanity for decades. Early intelligent systems, such as IBM Watson [Ferrucci et al., 2010], advanced the field by demonstrating the capability of answering complex natural language questions. However, its expansion into real-world applications like healthcare remained limited due to the system's highly specialized nature, which hindered cross-domain generalization, and insufficient natural language understanding, which made it difficult to handle complex documents and user interactions in realistic scenarios.

Recent advancements in computational capability and neural network architectures, particularly Transformer-based models [Vaswani et al., 2017], have drastically reshaped the landscape of human-computer interaction. Large Language Models (LLMs), where Transformer meets scaling, have achieved impressive performance in generating fluent natural language, leading to excitement that human-level artificial intelligence may be within reach. Despite their impressive fluency, however, Transformer-based conversational systems have significant limitations. They often generate plausible but incorrect information [Ji et al., 2023], creating the need to ground responses in actual stored and retrieved information [Soudani et al., 2024a].

This shift towards conversational interaction, coupled with the necessity for grounding, has profoundly reshaped information retrieval systems, as users increasingly favor direct natural language answers over traditional lists of hyperlinks or information cards in search engines. This paradigm shift has reinforced the importance of conversational

information access (CIA) systems which focus on grounding system responses in retrieved knowledge to provide reliable information [Zamani et al., 2023, Gao et al., 2019, Anand et al., 2019].

Despite significant progress in conversational information access, several challenges remain, including personalization and evaluation. These systems typically lack understanding of individual users' personal contexts, limiting their ability to effectively engage users and satisfy diverse information needs [Zamani et al., 2023, Salemi et al., 2024, Yazan et al., 2025]. Additionally, evaluating conversational systems accurately and efficiently remains difficult due to the inherent complexity and subjective nature of conversational interactions [Mehri and Eskenazi, 2020, Liu et al., 2023b]. Addressing these challenges is crucial for ensuring conversational interactions are both personally relevant [Salemi et al., 2024, Gerritse et al., 2020b] and reliably evaluated [Zamani et al., 2023, Gao et al., 2019, Dietz et al., 2025], thereby enhancing user satisfaction and trust in conversational systems.

This thesis addresses these challenges by focusing on personalization in CIA systems and evaluating these systems effectively. Specifically, the thesis explores conversational information access systems by researching methods for extracting users' personal context, developing personalized CIA systems, and evaluating these systems.

1.1 Research Objectives and Questions

The main research question of this thesis is:

> **RQ Main:** *How can we develop personalized CIA systems and evaluate them effectively?*

To answer this main research question, this thesis investigates personalization in CIA systems by addressing challenges related to (1) extracting personal context, (2) generating personalized responses, and (3) evaluating these systems effectively, as illustrated in Figure 1.1.

1.1.1 Personal Information Extraction

To personalize CIA systems, understanding the user's preferences and personal context is crucial. This section provides background on

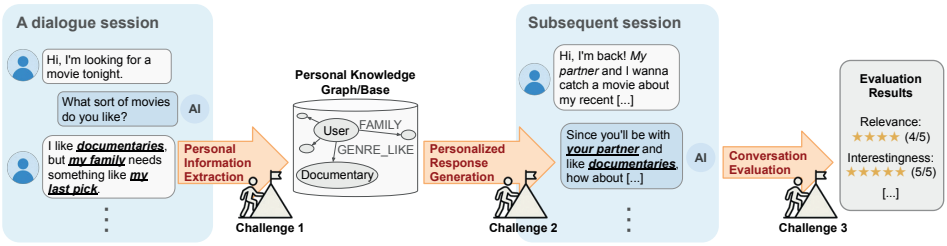


Figure 1.1: An illustration of a personalized CIA system, highlighting the three core challenges addressed in this thesis: (1) **Personal Information Extraction**, which focuses on extracting users' personal contexts by accurately identifying personal entities from conversational texts; (2) **Personalized Response Generation**, which involves developing a CIA model that utilizes the extracted context to generate personalized responses; and (3) **Conversation Evaluation**, which concerns evaluating CIA systems by considering whole conversation context, diverse interaction trajectories, and provide fine-grained performance insights.

extracting and structuring personal information from conversational interactions, as well as the associated research questions.

Background

Extracting and storing personal information from conversations is essential for enabling personalization in conversational systems. For instance, a Personal Knowledge Graph (PKG) is used to structure and represent personal entities and their relationships in a graph-based format centered around the user [Balog and Kenter, 2019]. Unlike general-purpose knowledge graphs, PKGs capture entities of personal significance, such as the user's possessions, acquaintances, or preferences; thereby enabling the system to understand and utilize the user's context in conversations, leading to better personalized responses.

A key challenge in creating personal knowledge repositories, such as PKGs, is accurately extracting personal entities from conversational texts and linking them to corresponding entries in existing knowledge bases. This process is known as **Entity Linking (EL)**. Traditional EL typically involves three main steps [Balog, 2018, van Hulst et al., 2020]: (1) mention detection, identifying text spans that refer to entities, (2)

candidate selection, retrieving candidate entities from a knowledge base for each mention, and (3) entity disambiguation, selecting the correct entity from the candidate set.

Traditional EL methods are primarily designed for general documents or short texts and tend to focus on named entities [Carmel et al., 2014, Hoffart et al., 2011]. However, different types of entities contribute to machine understanding in a conversational setting. For example, in a conversation where a user says "I love playing jazz on my guitar, a Gibson Les Paul," we can identify three types of entities: (i) *named entities*, i.e., proper nouns such as "Gibson Les Paul"; (ii) *concepts*, i.e., general categories such as "jazz"; and (iii) *personal entities*, i.e., possessive noun phrases such as "my guitar," which refer to entities of personal significance. Note that, while personal entity mentions can take many forms, as a first study of its kind in conversational EL, this thesis focuses on personal entities represented as possessive noun phrases. Although all entity types could contribute to machine understanding of conversations, what EL in conversations entails and how to develop an effective method remain open questions.

Research Questions

The first research question this thesis investigates is:

RQ1: *What does Entity Linking in conversations entail?*

We address this question by investigating the characteristics of EL in conversations and defining the task. Specifically, we collect EL annotations on conversations, dubbed ConEL, and investigate the performance of traditional EL systems on those annotations to understand the challenges of EL in conversations. We found that traditional EL tools are suboptimal for EL in conversational settings, especially struggling to handle concepts and personal entities, which are crucial for personalized conversations.

This calls for the development of a new EL method that can effectively handle conversational settings; therefore, the second research question addressed by this thesis is:

> **RQ2:** *How can we develop an Entity Linking method to accurately identify important entities for personalized conversational systems?*

We answer this question by introducing CREL, a conversational entity linking method developed to identify all types of important entities in conversations, including personal entities, concepts, and named entities. CREL leverages coreference resolution to effectively identify personal references (e.g., "my dog"), which is crucial for constructing PKGs. To train and evaluate CREL, we also construct the ConEL-2 dataset, extending ConEL with significantly more dialogues with personal entity annotations. Through extensive experiments, we found that CREL significantly outperforms traditional EL methods, particularly in handling personal entities.

1.1.2 Personalized Response Generation

Effectively utilizing extracted personal context to generate personalized responses is the next critical step. One of the first steps in achieving this goal is the availability of large-scale personalized conversational datasets that reflect user interactions and preferences in the real world. This section provides background on personalized CIA systems, discusses the need for datasets to train and evaluate personalized response generation, and outlines an associated research question.

Background

Personalized response generation in CIA concerns generating responses that satisfy the user's information needs through interactive conversations, considering the conversational context and the user's long-term preferences [Zamani et al., 2023]. While traditionally relying on straightforward rule-based approaches, recent CIA systems employ Transformer-based language models [Vaswani et al., 2017], such as BERT [Devlin et al., 2019], T5 [Raffel et al., 2020], and GPTs [Brown et al., 2020, Achiam et al., 2023], which significantly increased the performance by the improvement of natural language understanding and generation capabilities. However, despite these advancements, effectively capturing and utilizing individual user preferences for response generation remains a challenging task. Unlike conventional (ad hoc)

CIA, which focuses on addressing immediate, one-time conversations, personalized CIA requires systems to effectively elicit, store, and leverage user-specific preferences within and across multiple conversational sessions.

To achieve effective personalization, systems must first understand user preferences, typically through preference elicitation or clarifying questions, either via explicit system-initiated questions or implicit user feedback during conversations. Once elicited, these preferences need to be remembered and utilized to generate personalized responses in current or subsequent conversation sessions (preference utilization), thereby enhancing relevance and improving the overall user satisfaction.

Training models for personalized response generation requires large-scale datasets that reflect diverse real-world user interactions and preferences [Salemi et al., 2024, Salemi and Zamani, 2025]. However, existing conversational datasets often fail to capture them due to the complexity of effective preference elicitation and scalability challenges with expert-constructed dialogues. Even high-quality conversational datasets typically focus on single, relatively short conversations [Dalton et al., 2019, Dinan et al., 2019, Choi et al., 2018, Eric et al., 2020], thus lacking the necessary scale and session-level continuity for modeling personalized interactions over multiple sessions [Zamani et al., 2023]. These gaps highlight the need for a method that can efficiently collect large-scale, multi-session, and multi-domain personalized conversational datasets that reflect real-world user interactions and preferences.

Research Question

The third research question this thesis addresses is:

> **RQ3:** *How can users' personal preferences be utilized to generate personalized responses?*

To answer this question, we first propose a model to construct large-scale personalized conversational datasets, dubbed LAPS. It leverages LLMs to guide human workers in efficiently composing realistic, multi-session, and multi-domain personalized dialogues. We found that our approach significantly accelerates the dialogue construction process while preserving the diversity and quality of human-written conversations, effectively capturing diverse real user preferences.

Building upon the LAPS dataset, we then investigate how to effectively utilize users' personal preferences to generate personalized responses. We find that using semi-structured textual representations of user preferences (termed "preference memory") enables more accurate utilization of users' preferences for recommendations, improves response explanations, and mitigates LLMs' recall issues compared to using raw dialogue history. This demonstrates the effectiveness of structured personal context in enhancing personalized response generation.

1.1.3 Conversation Evaluation

Evaluation is crucial for developing CIA systems, as one cannot improve that which one cannot measure [Voorhees, 2005, Thomson, 1889]. Although human annotations are the gold standard for evaluating CIA systems, they are expensive and time-consuming; thus, automatic evaluation methods have been proposed to scale up the evaluation process. In this section, we provide background on automatic evaluation methods for CIA systems, discuss the challenges of evaluating personalized CIA systems, and outline the associated research question.

Background

Automatic conversation evaluation is crucial for efficient, scalable, and cost-effective diagnosis of problems and biases during the development of CIA systems. There are two main types of automatic evaluation methods: *reference-based* and *reference-free*. Reference-based methods use gold references to evaluate system responses, which include Recall@K, BLEU [Papineni et al., 2002], ROUGE [Lin, 2004], and BERTScore [Zhang et al., 2020]. While these methods are effective for machine translation and traditional IR tasks, they have limitations in conversation evaluation, as they overlook diverse response possibilities and various evaluation aspects. These limitations are supported by various studies showing a weak correlation between reference-based methods and human evaluations [Mehri and Eskenazi, 2020, Liu et al., 2016].

Reference-free methods have been proposed to address these limitations [Mehri and Eskenazi, 2020, Zhong et al., 2022, Liu et al.,

2023b]. These methods evaluate system responses without relying on gold references, consider multiple aspects of system quality, and allow for the assessment of various response possibilities. For instance, Mehri and Eskenazi [2020] proposed USR that leverages unsupervised methods to evaluate dialogue quality across various aspects, and Zhong et al. [2022] proposed UniEval that frames conversation evaluation as a boolean question answering (QA) task. More recent works use LLMs for evaluation [Liu et al., 2023b, Upadhyay et al., 2025, Cook et al., 2024, Thomas et al., 2024]. However, these methods primarily focus on turn-level evaluation with a fixed dialogue history, limiting their ability to assess the whole conversation and capture the diverse user-system conversation trajectories, which are crucial for evaluating system performance in real-world scenarios [Siro et al., 2022, 2023]. Therefore, there is a need for automatic, reference-free evaluation methods that can assess the entire conversation and capture diverse conversation trajectories, providing fine-grained insights into system performance.

Research Question

The fourth research question this thesis addresses is:

> **RQ4:** *How to automatically evaluate personalized CIA systems in a way that accounts for natural conversation flow and provides fine-grained insights into system performance?*

We answer this question by introducing FACE, a Fine-grained, Aspect-based Conversation Evaluation method. FACE is a reference-free evaluation method that handles diverse conversation trajectories and evaluates the entire conversation. It first decomposes system responses into contextualized conversation particles. An LLM then evaluates these particles using evaluation instructions optimized via an automatic prompt optimization technique. The resulting sub-scores are then aggregated into a single score per evaluation aspect, offering interpretable insights into system performance. Empirical evaluations show that FACE aligns closely with human judgments and provides fine-grained insights into the system's performance across multiple aspects (e.g., relevance and user understanding), thereby providing actionable feedback for system improvement.

1.2 Contributions

In this section, we summarize the main contributions of this thesis. The contributions are shown as theoretical and methodological contributions, as well as resource contributions.

Theoretical and Methodological Contributions

C1. We define the problem of EL in conversations, subdividing entities into three categories (named entities, concepts, and personal entities), and analyzing and identifying the unique challenges that make traditional EL methods suboptimal in conversational settings.

C2. We provide effective designs for collecting large-scale conversational EL annotations, used in the creation of the ConEL and ConEL-2 datasets.

C3. We propose a conversational entity linking method, CREL, which identifies personal entities and incorporates coreference resolution.

C4. We propose a scalable method for constructing large-scale personalized conversational datasets, dubbed LAPS, which uses LLMs to guide human workers, efficiently constructing multi-session, multi-domain dialogues that capture real user preferences while maintaining high quality.

C5. We investigate effective utilization of users' personal preferences for personalized response generation, showing that semi-structured preference memory improves accuracy, explanations, and mitigates LLMs' recall issues in response generation.

C6. We propose an automatic, reference-free, and fine-grained evaluation method for personalized CIA systems. The method, referred to as FACE, assesses entire conversations across diverse interaction paths using optimized instructions and particle-based scoring, providing aspect-based insights that align closely with human judgments.

Resource Contributions

C7. We release the ConEL datasets, the first-of-its-kind conversational entity linking datasets (ConEL and ConEL-2) with annotations for personal, named, and conceptual entities, enabling empirical analysis

of EL in conversations and benchmarking EL systems in conversational contexts.

C8. Alongside the ConEL datasets, we release an online resource listing around 130 conversational datasets with a detailed comparison of their characteristics.

C9. We release CREL, an open-source conversational entity linking tool that identifies personal entities, named entities, and concepts, suitable for conversational settings.

C10. We release the LAPS dataset, a large-scale, multi-session, human-written personalized dialogue dataset for research on preference extraction and personalized CIA development. The dataset contains 1,406 conversations, with 11,215 preferences extracted from the conversations, ranging across two domains: *recipe* and *movie* recommendations.

C11. We release the CRSArena-Eval dataset, a comprehensive evaluation of conversation evaluation (i.e., meta-evaluation) dataset with high-quality multi-aspect human judgments across multiple systems for benchmarking automatic evaluation methods. The dataset contains 467 conversations, with 20,962 human judgments collected across nine systems.

1.3 Thesis Overview

This section provides an overview of the thesis structure, describing each chapter and its contributions to the RQs outlined in Section 1.1. Figure 1.2 illustrates how each chapter maps to the key research areas throughout the thesis.

Chapter 2 describes EL in conversations, collecting EL annotations on conversations, dubbed ConEL, and analyzing the characteristics of EL in conversations.

Chapter 3 presents a novel EL method (CREL) for personal information extraction in conversations.

Chapter 4 discusses the construction of large-scale, multi-session human-written personalized conversational datasets, dubbed LAPS, and examines the generation of personalized responses using this data.

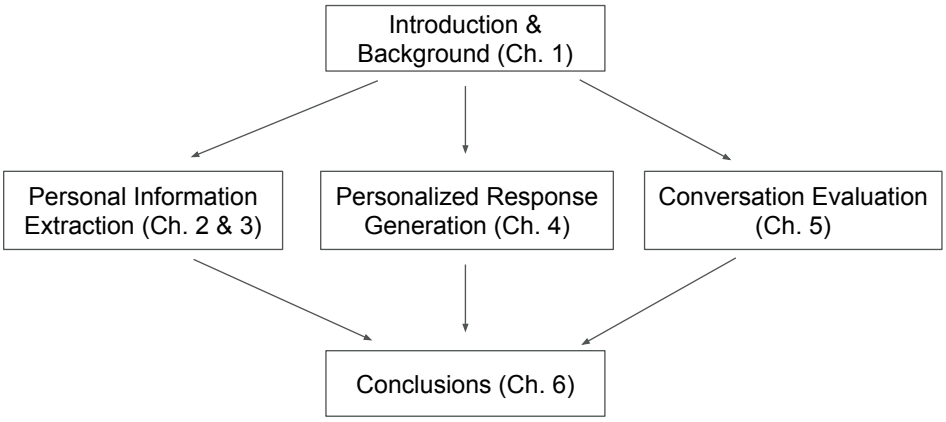


Figure 1.2: Thesis structure and connections between chapters.

Chapter 5 introduces FACE, a fine-grained, aspect-based conversation evaluation method that evaluates personalized CIA systems based on the entire conversation and captures diverse conversation trajectories.

Scope of the Thesis. This thesis focuses specifically on text-based CIA systems. While conversational systems may be multimodal, concentrating on textual conversations lets us focus on core challenges in personalization, dataset construction, and evaluation methods, which can be a basis for future multimodal extensions.

1.4 Origins

The origin of each chapter is described below.

- [Joko et al., 2021] H. Joko, F. Hasibi, K. Balog, and A. P. de Vries. Conversational Entity Linking: Problem Definition and Datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. *[Related to RQ1 and C1, C2, C7, C8; included in Chapter 2].*

- [Joko and Hasibi, 2022] H. Joko and F. Hasibi. Personal Entity, Concept, and Named Entity Linking in Conversations. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2022.
[Related to RQ2 and C3, C9; included in Chapter 3].

- [Joko et al., 2024] H. Joko, S. Chatterjee, A. Ramsay, A. P. de Vries, J. Dalton, and F. Hasibi. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.
  [Related to RQ3 and C4, C5, C10; included in Chapter 4].

- [Bernard et al., 2025] N. Bernard, H. Joko, F. Hasibi, and K. Balog. CRS Arena: Crowdsourced Benchmarking of Conversational Recommender Systems. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 2025
  [Related to RQ4 and C11; integrated in Chapter 5].

- [Joko and Hasibi, 2026] H. Joko and F. Hasibi. FACE: A Fine-Grained Reference-Free Evaluator for Conversational Information Access. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026.
  [Related to RQ4 and C6, C11; included in Chapter 5].

Other Publications. The papers listed below were also published during the course of the PhD research. While these papers indirectly contribute to the thesis, they are not direct source material for the thesis.

- [Joko et al., 2022] H. Joko, E. J. Gerritse, F. Hasibi, and A. P. de Vries. Radboud University at TREC CAsT 2021. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*, 2022.

- [Joko et al., 2026] H. Joko, S. Amirshahi, C. L. A. Clarke, and F. Hasibi. WildClaims: Information Access Conversations in the Wild(Chat). In *Proceedings of the 48th European Conference on Information Retrieval*, 2026.

Chapter 2

Conversational Entity Linking

As discussed in Chapter 1, storing and utilizing a user's personal context through personal knowledge repositories, such as personalized knowledge graphs (PKGs), is crucial for personalizing conversational systems. A key challenge in creating these repositories is accurately extracting entities from dialogue via Entity Linking (EL). However, traditional EL methods are primarily designed for general documents; their effectiveness in conversational contexts, where entities are often personal or conceptual, remains underexplored. To close this gap, this chapter addresses the research question:

RQ1: *What does Entity Linking in conversations entail?*

To answer this question, this chapter introduces the ConEL dataset, a novel conversational entity linking dataset containing crowdsourced annotations of personal entities, named entities, and concepts, to investigate the unique characteristics of EL in conversations. The chapter then analyzes these annotations to identify the main properties of conversational EL and evaluate the performance of traditional EL systems on our dataset. The findings show that existing methods are suboptimal for this task, particularly struggling to identify the personal entities and concepts that are important for personalized interactions. These results demonstrate the distinct challenges of conversational EL and highlight the importance of further research in this area.

2.1 Introduction

Conversational systems are becoming increasingly important with the proliferation of personal assistants, such as Siri, Alexa, Cortana, and the Google Assistant. In this realm, understanding user utterances plays a crucial role in holding meaningful conversations with users – this process is handled by the natural language understanding (NLU) component in traditional task-oriented dialogue systems [Gao et al., 2019]. A popular text understanding method, which has proven to be effective in various downstream tasks [Dalton et al., 2014, Hasibi et al., 2016, Xiong et al., 2017, Shang et al., 2021, Hasibi et al., 2017a, Dargahi Nobari et al., 2018], is *entity linking*: the task of recognizing mentions of entities in text and identifying their corresponding entries in a knowledge graph [Balog, 2018]. In this chapter, we aim to investigate the role of entity linking in conversational systems.

Even though large-scale neural language models (like BERT [Devlin et al., 2019] and GPT-3 [Brown et al., 2020]) have repeatedly been shown to achieve high performance in various machine understanding tasks, these do not provide replacements for explicit auxiliary information from knowledge graphs. Rather, the two should be seen as complementary efforts. Indeed, augmenting neural language models with information from knowledge graphs has been shown to be beneficial in a number of downstream tasks [Poerner et al., 2020, Peters et al., 2019, Zhang et al., 2019]. Specifically, in the context of task-based conversational systems, NLU often relies on entities for fine-grained domain classification and intent determination. This makes EL for conversational systems even more important.

Despite its importance, research on EL for conversational systems has so far been limited. Traditional EL techniques that are used for documents [Balog, 2018], queries [Hasibi et al., 2015, Li et al., 2020], or tweets [Meij et al., 2012] are suboptimal for conversational systems for a number of reasons. First, unlike documents, conversations are not "static," but are a result of human-machine interaction. Here, the user can correct the system's interpretation of their request and the system can ask clarification questions to further its understanding of the user's intent. Second, conversations are informal and it is common in a conversation to make references to entities by their pronouns; e.g., "my city," "my guitar," and "its population." A conversational system is expected

to understand and handle such mentions of personal entities [Balog and Kenter, 2019]. Third, while entity linking in documents or queries tends to focus on proper noun entities [Carmel et al., 2014, Hoffart et al., 2011], in a conversational setting all types of entities, including general concepts, can contribute to machine understanding of users' utterances. In this chapter, we aim to investigate these differences and develop resources to foster research in this area.

To investigate the main research question of this chapter, we set out to analyze entity linking annotations for existing conversational datasets. We perform a thorough analysis of existing conversational datasets and select four of these for annotation. These cover the three main categories of conversational problems [Gao et al., 2019]: question answering (QA), task-oriented systems, and social chat (cf. Sect. 2.3). We aim to annotate "natural" conversations, and therefore bias our selection of datasets towards those that are obtained using a Wizard-of-Oz setup.

Annotating dialogues, however, is an inherently complex task, where it is a challenge to keep the cognitive load sufficiently low for crowd workers. This leads us to a secondary research question:

> **RQ1a:** *What are effective designs for collecting large-scale annotations for conversational data?*

Although a large body of research exists on effective designs for collecting large-scale entity annotations [Adjali et al., 2020, Bhargava et al., 2019, Finin et al., 2010, Mayfield et al., 2011], to the best of our knowledge, there is no work on conversational data. We run a number of pilot experiments using Amazon Mechanical Turk (MTurk) to select the best design and instruction. Based on these experiments, we develop the *Conversational Entity Linking (ConEL)* dataset, consisting of 100 annotated dialogues (708 user utterances) sampled from the QuAC [Choi et al., 2018], MultiWOZ [Zang et al., 2020], WoW [Dinan et al., 2019], and TREC CAsT 2020 [Dalton et al., 2020] datasets. To enable further study in conversational EL, we also annotate a separate sample of 25 WoW dialogues (containing references to personal entities) and all 25 manually rewritten dialogues of TREC CAsT 2020.

Our findings, obtained by analyzing the annotated dialogues, are as follows.

- Mentions of personal entities are mainly present in social chat conversations.

- While named entities are deemed to be important for text understanding, specific concepts are also found useful for understanding the intents of conversational user utterances.

- Traditional EL approaches fall short in providing high precision annotations for both concepts and named entities. This calls for a methodological departure for conversational entity linking, where concepts, named entities, and personal entities are taken into consideration.

In summary, this work makes the following contributions:

- To the best of our knowledge, ours is the first study on EL in conversational systems. We subdivide entities into three categories (named entities, concepts, and personal entities), and analyze the importance of each for conversational data. We further investigate different aspects of EL for three categories of conversational tasks: QA, task-oriented, and social chat.

- We investigate effective designs for collecting large-scale EL annotations for conversational data.

- We make the annotated conversational datasets publicly available.[1] This data comes with detailed account of the procedure that was followed for collecting the annotations, which can be used for further extension of the collection.

- As an additional (online) resource, we provide a comprehensive list of around 130 conversational datasets released by different research communities with a detailed comparison of their characteristics.

The resources provided in this chapter allow for further investigation of entity linking in conversational settings, can be used for evaluation or training of conversational EL systems, and complement existing conversational datasets.

[1] `https://github.com/informagi/conversational-entity-linking`

2.2 Related Work

The related work pertinent to this chapter concerns entity linking in documents, queries, and conversational systems, as well as personal entity identification.

2.2.1 Entity Linking

Entity linking in documents. Entity linking plays an important role in understanding what a document is about [Balog, 2018]. TagMe [Ferragina and Scaiella, 2010] is one of the most popular EL tools, redesigned and improved by Piccinno and Ferragina [2014] and renamed to WAT. van Hulst et al. [2020] presented REL, which is an open source EL tool based on state-of-the-art NLP research. Other state-of-the-art EL methods include DeepType [Raiman and Raiman, 2018], Blink [Wu et al., 2020], and GENRE [Cao et al., 2021]. Although these approaches are effective for documents, it is known that EL algorithms with high performance on general documents are less effective when applied to short informal texts like queries [Cornolti et al., 2019].

Entity linking in queries. Entity linking in queries poses new challenges due to the short and noisy text of queries, their limited context, and high efficiency requirement [Hasibi et al., 2017b, Carmel et al., 2014, Cornolti et al., 2019]. Cornolti et al. [2019] tackled some of these challenges by "piggybacking" on a web search engine. Relying on external search engines, while being effective, hinders efficiency and sustainability of EL systems. Hasibi et al. [2017b] studied this challenge with a special focus on striking a balance between effectiveness and efficiency. These studies consider EL for a single query, while in conversational systems multiple consecutive user turns need to be annotated.

Entity linking in conversations. Research on conversational entity linking has been mainly focused on employing traditional entity linking and named entity recognition methods in conversational and QA systems [Kumar and Callan, 2020, Chen and Choi, 2016, Bowden et al., 2018, Chen et al., 2017, Vakulenko et al., 2018, Li et al., 2020]. Entity linking is also used in multi-party conversations to connect mentions across different parts of dialogues and mapping to their corresponding

character [Chen and Choi, 2016]. This is a subtask of entity linking, referred to as character identification. A close study to our work is [Bowden et al., 2018], where an entity linking tool, focused mainly on named entities, is developed for open-domain chitchat. In contrast to these works, we aim to understand EL for conversational systems, annotating conversations with concepts, named entities, and personal entities.

2.2.2 Personal Entities

Dealing with the mention of personal entities (e.g., "my guitar") is important for personalization of conversational systems. Consider for example the user utterance "Do you know how to fix my guitar?" To answer this question, the system has to know more about the user's guitar type; e.g., "Gibson Les Paul." This information may be available in the conversation history, previous conversations, or other sources (e.g., user's public information in social media). This information can be represented as Resource Description Framework (RDF) triples in the form of subject-predicate-object expressions $\langle e, p, e' \rangle$, e.g., $\langle$*User, guitar, Gibson Les Paul*$\rangle$. Li et al. [2014] proposed a method to detect personal entities (e) and their corresponding predefined predicates (p) in conversations. Their approach consists of three steps: (1) identifying user utterances that are related to personal entities, (2) predicting entity mentions by classifying those utterances, and (3) finding the personal entities. Tigunova et al. [2019] address the problem of identifying personal entities from implicit textual clues. They proposed a zero-shot learning method to overcome the lack of sufficient labelled training data [Tigunova et al., 2020]. All these studies focus on identifying predefined classes of predicates. Extracting RDF triples without predefined relation classes has been studied in the context of open information extraction [Yates et al., 2007, Fader et al., 2011, Angeli et al., 2015, Cui et al., 2018], but not in relation to personal entities. In this study, we annotate conversations with personal entity mentions and their corresponding entities.

2.3 Dataset Selection

There exists a large number of conversational datasets released by the natural language processing, machine learning, dialogue systems, and information retrieval communities. We made an extensive list of around 130 datasets,[2] extracted from ParlAI [Miller et al., 2017] and other dataset comparison lists [Penha et al., 2019, Choi et al., 2018, Hauff, 2020]. These datasets target three conversational problems [Gao et al., 2019]:

- *Question answering*, where users ask natural language queries and the system provides answers based on a text collection or a large-scale knowledge repository.
- *Task-oriented systems*, which assist users in completing a task, such as making a hotel reservation or booking movie tickets.
- *Social chat*, where systems are meant to be AI companions to the users and hold human-like conversations with them.

To obtain a comprehensive view of entity linking in conversational systems, we set out to analyze at least one dataset for each of the three main categories of conversational problems. To this end, we shortlisted datasets that resemble real conversations. That is, multi-domain and multi-turn datasets, collected based on actual interactions between two humans. Datasets that are extracted from web services (e.g., Reddit and Stack Exchange) or created based on templates (e.g., bAbI [Sukhbaatar et al., 2015]) were thus ignored. To ensure that the selected datasets are sizable, they were required to contain at least 100 dialogues. This list was further narrowed down by selecting relatively popular datasets based on citation counts and publication year.[3] By applying these criteria, nine datasets were shortlisted. In the final step, each dataset in our shortlist was closely examined, and at least one dataset was selected for each conversational problem; see Table 2.1 for an overview of the selected datasets. The reasoning behind our selections is detailed below.

[2]This list is publicly available at: `https://github.com/informagi/conversational-entity-linking`

[3]While admittedly this is a loose measure, it helps to identify datasets that became widely accepted by the research community.

Table 2.1: Overview of the selected conversational datasets for the entity annotation process. A sample of QuAC, MultiWOZ, and WoW, and all dialogues in TREC CAsT 2020 were used for generating the ConEL dataset.

Dataset	Task	#Convs	Avg. Turns
QuAC [Choi et al., 2018]	QA	13.6K	14.5
MultiWOZ [Zang et al., 2020]	Task-oriented	8.4K	13.5
WoW [Dinan et al., 2019]	Social chat	22.3K	9.1
TREC CAsT 2020 [Dalton et al., 2020]	QA	25	17.3

QA. Among the QuAC [Choi et al., 2018], CoQA [Reddy et al., 2019], and QReCC [Anantha et al., 2020] datasets, we selected QuAC for QA dialogues. QuAC is a widely used dataset for conversational QA and contains 13.6K dialogues between two crowd workers. CoQA, on the other hand, is a machine reading comprehension dataset with provided source texts for every dialogue. Since these source texts are not necessarily available in real conversations, CoQA was left out. QReCC is built based on questions from other datasets, including QuAC and TREC CAsT, and is focused on question rewriting. Because of the overlapping questions with other datasets, it was also ignored.

Task-oriented. The MultiWOZ [Zang et al., 2020] and KVRET [Eric et al., 2017] datasets were examined for task-oriented dialogues. MultiWOZ covers seven various goal-oriented domains: *Attraction, Hospital, Police, Restaurant, Hotel, Taxi, and Train.* KVRET, on the other hand, deals with only three domains, all of which are in-car situations. We, therefore, selected the MultiWOZ dataset, which also has more dialogues than KVRET (8.4K vs. 3K). Note that MultiWOZ has several versions [Budzianowski et al., 2018, Eric et al., 2020, Zang et al., 2020]; we used the latest version, MultiWOZ 2.2 [Zang et al., 2020].

Social chat. The Wizard of Wikipedia (WoW) [Dinan et al., 2019], Empathetic Dialogues [Rashkin et al., 2019], Persona-Chat [Zhang et al., 2018a], and TaskMaster-1 [Byrne et al., 2019] datasets were shortlisted for social chat dialogues. We excluded TaskMaster-1, as the majority of dialogues (7.7K) were collected by crowd workers who were instructed to write full conversations, i.e., played both the user and the system roles on their own. Persona-Chat and Empathetic

Dialogues are more focused on emotional and personal topics, while WoW is knowledge grounded and makes use of knowledge retrieved from Wikipedia. We therefore chose WoW as a social chat dataset.

Additionally, we also included the TREC 2020 Conversational Assistance Track (CAsT) [Dalton et al., 2020] dataset in our study. TREC CAsT [Dalton et al., 2019] is an important initiative by the IR community, and is focused on the information seeking aspect of conversations. Unlike other datasets, which represent dialogues as a sequence of user-system exchanges, TREC CAsT 2019 provides relevant passages that a system may return in response to a user utterance – therefore, a unique conversation cannot be made for a given conversational trajectory. This has been changed in TREC CAsT 2020 [Dalton et al., 2020], where a canonical response is given for each user utterance. We generated conversations for our crowdsourcing experiments using these canonical responses. In the remainder of this chapter we refer to TREC CAsT 2020 as CAsT.

2.4 Entity Annotation Process

This section describes the process of annotating dialogues from the selected conversational datasets. Our aim is to identify entities that can aid machine understanding of user utterances; this includes named entities, concepts, and mentions of personal entities. Note that our focus is on user utterances, since the system is supposedly aware of the text it generates during a conversation. The knowledge graph we use for annotations is Wikipedia (2019-07 dump).

The annotation process was performed via crowdsourcing using Amazon's Mechanical Turk (MTurk). In order to reduce the cognitive load for this complex task and to obtain the best annotation results, we ran a number of pilot experiments. In these experiments, we tested multiple task structures and interfaces using MTurk and compared the results with expert annotations of the same dialogues. The best task designs and interfaces were then used for the final annotations. Below, we describe the task design for annotating concepts, named entities, and personal entities (Sections 2.4.1 and 2.4.2), followed by the process of dialogue selection and annotation (Section 2.4.3).

Figure 2.1: Annotation interface for the entity-mention selection task (Stage 1). To keep the cognitive load low, only multiple-choice questions were used. Possible answer entities are linked to their corresponding Wikipedia article.

2.4.1 Concepts and Named Entities

We employ a two-step process for annotating explicitly mentioned entities, i.e., concepts and named entities.

Stage 1: Selecting entity-mention pairs. First, we aim to map each mention to a single entity in the knowledge graph. Workers were presented with a dialogue, a mention from the latest user utterance, and a set of candidate entities or None. They were instructed (using a concise description and a couple of examples) to find the Wikipedia article that is referred to by the mention; Figure 2.1 shows an excerpt from this task. The "None of the above" option is selected when the candidate pool does not contain the correct entity or the given mention is not appropriate. These mentions were later examined by an expert annotator and assigned the correct entity or ignored. To reduce the cognitive load on the workers, long conversations were trimmed; i.e., the middle turns in conversations with more than six turns were not presented.

Stage 2: Finding the helpful entities. The mention-entity pairs obtained in the first stage are not necessarily important for machine understanding of user utterances; consider, for instance, the entity College in utterance "I have wanted to travel to Amsterdam since college. What are the tourist attractions there?" In Stage 2, we asked workers to filter the entity-mention pairs identified in Stage 1 by selecting only those pairs that can help the system to identify the user's intent. Specifically, we provided them with a conversation history and all mention-entity pairs from a user utterance, and gave them the following instruction: *"Imagine you are an AI agent (e.g., Siri or Google Now), having a dialogue with a person. You have access to Wikipedia articles (and some other information sources) to answer the person's questions. Select the Wikipedia articles that help you to find an answer to the person's question."* We presented mention-entity pairs two times to the users (all in one assignment): once they were asked to select named entities, and the other time they were asked to select all "helpful" entities. Using this interface, we were able to identify named entities and further analyze the differences between concepts and named entities.

Generating annotation candidates. We employ a pooling approach to generate an extended set of candidate mentions and entities. Three EL tools were used to annotate the dialogues: TagMe [Ferragina and Scaiella, 2010], WAT [Piccinno and Ferragina, 2014], and REL [van Hulst et al., 2020]. Each tool was employed in two ways: (i) the *turn* method, which annotates a single turn, irrespective of the conversation history, and (ii) the *history* method, which annotates each turn given the conversation history up to that turn. For the CAsT dataset, only user utterances were given to the EL tool, while for other datasets both system and user utterances were considered as conversation history. This is due to relatively long system utterances in the CAsT dataset, which makes it infeasible for the EL tools to annotate the whole conversation history. To further improve the recall of our pool, we included the top-10 Wikipedia search results, using mentions as queries sent to the MediaWiki API.[4]

[4] `https://www.mediawiki.org/wiki/API:Main_page`

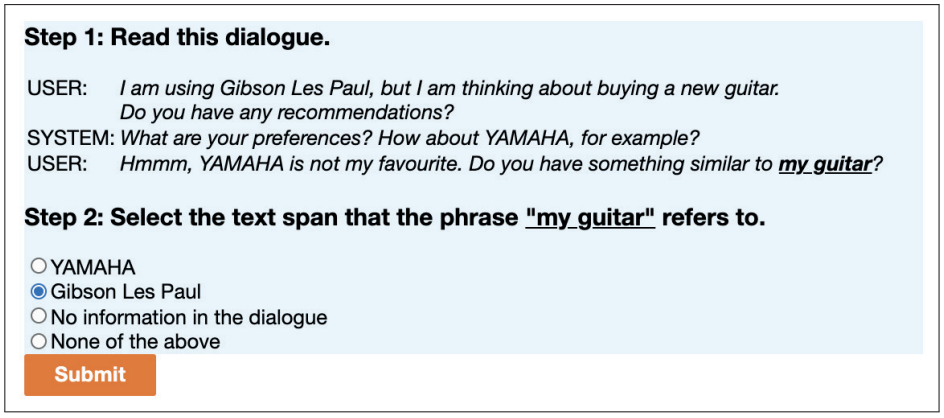


Figure 2.2: Annotation interface for mapping a personal entity mention ("my guitar") to the corresponding explicit entity mention in the conversation ("Gibson Les Paul").

2.4.2 Personal Entities

Annotating conversations with personal entities requires identifying *personal entity mentions* and mapping them to the corresponding *explicit entity mentions* in the conversation history (if exists); e.g., mapping the personal entity mention "my guitar" to the explicit entity mention "Gibson Les Paul." Once this mapping was done, mention-entity pairs can be identified as described in Stage 1 of Section 2.4.1. We note that in some cases, explicit entity mentions are not present in the conversation history and the system needs to detect them from other information sources (e.g., previous conversations or user profile data). In this study, we confined ourselves to the cases where explicit entity mentions can be found in conversation history; i.e., personal entity mentions without explicit entity mentions in the conversation history were not mapped to any entity.

We designed a crowdsourcing experiment, where workers were given a conversation history along with the personal entity mention, and their task was to select the text span in the conversation history that the given personal entity mention refers to. Figure 2.2 shows an example of this task. "None of the above" answers were resolved by an expert annotator.

Generating annotation candidates. We used a simple yet effective method to find personal entity mentions. Inspired by Li et al. [2014], we detect all text spans starting with "my" followed by one or several adjectives, common nouns, proper nouns and/or numbers (using the SpaCy[5] part-of-speech (POS) tagger). We further allowed for the word *"of"* to be part of the mention (e.g., "my favourite forms of science fiction"). For each personal entity mention, we included all the candidate mentions that were identified by our EL methods (cf. Section 2.4.1).

2.4.3 Dialogue Selection and Annotation

Annotating all dialogues in the selected datasets was infeasible for us, due to its high costs. We therefore selected a random sample of dialogues from a pool of presumably difficult dialogues from the QuAC, MultiWOZ, and WoW datasets. This pool contains dialogues with at least one *complex mention*, a personal entity mention, or a clarification question in user utterances. By complex mention we refer to cases where the same mention is linked to different entities by the EL tools (i.e., REL, WAT, and TagMe). The clarification questions were identified based on the patterns stated by Braslavski et al. [2017], and personal entity mentions were extracted as described in Section 2.4.2. A total of 100 samples were selected (25 for each dataset), amounting to 708 user utterances. Note that unlike the other datasets, CAsT contains only 25 dialogues, therefore all its dialogues were annotated. Based on this selection, we are able to analyze the differences between the three main categories of conversational tasks.

The CAsT dataset comes with manually rewritten user queries, where each rewritten query can be answered independently of the conversation history. We also annotated the manually rewritten CAsT queries to allow for a comparison between raw and rewritten queries.

To extend our analysis on personal entity linking, we annotated another sample of dialogues from the WoW dataset. To generate this sample, we randomly selected 500 dialogues that contain personal entity mentions and presented them to crowd workers to find their entity references in the dialogues (cf. Section 2.4.2). Workers agreed that, in 180 dialogues of this sample, the references to the personal

[5] `https://spacy.io/`

entity mentions are present in the dialogue. We then randomly selected 25 dialogues (containing 216 user utterances) out of these 180 dialogues and annotated their concepts, named entities, and personal entities.

To ensure high data quality, the annotation tasks were performed by top-rated MTurk workers, i.e., Mechanical Turk Masters. Since the number of Masters is small, and they mainly select tasks with a high number of human intelligence tasks (HITs), the remaining 0.4% of our annotation tasks were performed by high quality workers with a task approval rate of 99% or higher. We collected three judgments for each annotation and paid the workers ¢6 for each annotation assignment, resulting in a final cost of around $620. Fleiss' Kappa inter-annotator agreement was 0.76, 0.30, 0.61 for Stage 1, Stage 2, and personal entity annotations, respectively. Disagreements were resolved by an expert annotator.

2.5 Annotation Results

In this section, we describe our findings based on the analysis of the entity annotations obtained for the selected datasets. We also present baseline results for the entity-annotated conversations. The results are shown in Tables 2.2–2.5. In these tables, the last character of each method, "t" or "h," stands for "turn" or "history," respectively (cf. Section 2.4.1). Precision, recall, and F1 scores are micro-averaged and computed using the strong matching approach [Usbeck et al., 2015]. To understand the frequency of personal entities in conversational datasets, we applied the method described in Section 2.4.2 to identify all personal entity mentions in all the datasets. We found that WoW contains more dialogues with personal entity mentions compared to other datasets; i.e., 33% of dialogues in WoW vs. 0.3%, 11%, and 12% of dialogues in QuAC, MultiWOZ, and TREC CAsT, respectively. These results indicate that **personal entity mentions are mainly present in social chat conversations**.

Comparing concepts and named entities, we found that 43% of linked entities in the ConEL dataset are marked as named entities (NEs) by crowd workers, which implies that the remaining 57% entities are concepts. This indicates that **in addition to named entities, concepts are also found useful for understanding the intents**

Table 2.2: Entity linking results on the ConEL dataset.

	QuAC			**MultiWOZ**			**WoW**			**CAsT**			**All**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TagMe_t	29.5	37.1	32.9	18.5	23.9	20.9	33.0	41.4	**36.7**	52.7	**47.3**	49.8	35.5	39.7	37.5
TagMe_h	34.7	32.4	33.5	18.7	23.9	21.0	33.0	41.4	**36.7**	57.5	46.1	**51.2**	38.2	38.0	**38.1**
WAT_t	23.2	39.0	29.1	19.5	**36.6**	**25.5**	24.8	45.7	32.2	46.6	41.3	43.8	28.6	40.7	33.6
WAT_h	27.4	**48.6**	35.1	19.3	31.0	23.8	23.5	**55.7**	33.1	44.7	45.5	45.1	29.6	**45.5**	35.8
REL_t	23.7	21.9	22.8	**39.3**	15.5	22.2	43.8	10.0	16.3	68.1	19.2	29.9	38.8	17.7	24.3
REL_h	**37.6**	33.3	**35.4**	**39.3**	15.5	22.2	**52.9**	12.9	20.7	**70.2**	19.8	30.8	**47.6**	21.3	29.4

Table 2.3: Breakdown of entity linking results for named entities (F_{NE}) and concepts (F_C).

	QuAC		**Multi WOZ**		**WoW**		**CAsT**		**All**	
	F_{NE}	F_C	F_{NE}	F_C	F_{NE}	F_C	F_{NE}	F_C	F_{NE}	F_C
TagMe_t	32.2	6.3	17.3	9.4	**34.6**	11.8	24.4	40.7	27.5	21.6
TagMe_h	30.5	**11.3**	15.9	11.0	**34.6**	11.8	24.3	**43.3**	26.5	**24.4**
WAT_t	25.0	8.9	15.5	**15.4**	23.8	15.0	22.6	35.1	22.1	20.2
WAT_h	31.7	8.5	14.8	14.7	22.4	**16.2**	22.1	35.9	23.9	20.7
REL_t	26.1	0.0	**31.7**	3.1	18.2	8.5	63.8	2.4	35.1	2.5
REL_h	**40.7**	0.0	**31.7**	3.1	25.0	8.3	**66.0**	2.4	**43.1**	2.5

of user utterances.

Table 2.2 shows the results of different EL methods on the ConEL dataset. While TagMe achieves the highest F1 scores on WoW and CAsT, WAT and REL are the best performing tools (with respect to F1) on the MultiWOZ and QuAC datasets, respectively. Comparing the "turn" and "history" methods, we observe that conversation history improves EL results for most datasets and tools. We also find that REL has higher precision but lower recall compared to TagMe and WAT. One might argue that high precision EL is preferred in a conversational setting, as incorrect results can lead to high user dissatisfaction. This claim, however, requires further investigation, and the effect of EL on end-to-end conversational system performance is yet to be evaluated.

Table 2.3 compares EL results for named entities and concepts separately, where F1 scores are computed based on only named entities F_{NE} or concepts F_C. We observe that REL is better at linking named entities, while TagMe is better at linking concepts. This shows that although it is important to achieve high performance for both named

Table 2.4: Entity linking results on TREC CAsT raw and rewritten dialogues.

	CAsT (raw)			**CAsT (rewritten)**		
	P	R	F	P	R	F
TagMe_t	52.7	**47.3**	49.8	64.1	**66.4**	**65.2**
TagMe_h	57.5	46.1	**51.2**	65.6	64.6	65.1
WAT_t	46.6	41.3	43.8	52.9	45.3	48.8
WAT_h	44.7	45.5	45.1	54.9	50.8	52.8
REL_t	68.1	19.2	29.9	73.8	27.1	39.6
REL_h	**70.2**	19.8	30.8	**76.4**	27.9	40.8

entities and concepts, there is no single EL tool that excels at both. The results in Tables 2.3 and 2.2 suggest that ***all existing EL tools that we examined are suboptimal for EL in a conversational setting.***

Table 2.4 shows the EL results for all raw and rewritten CAsT queries. Similar to Table 2.2, we observe that there is a trade-off between precision and recall across the different EL tools. The results also show higher scores for rewritten queries compared to raw queries, which is due to resolved coreferences and richer context in the rewritten queries.

Table 2.5 shows EL results on a sample of WoW dialogues, all containing references to personal entities (cf. Section 2.4). This sample is annotated with concepts, named entities, and personal entities. The left block shows the results of different EL methods in their original form, i.e., without annotating personal entity mentions.

The right block in Table 2.5 represents a modified version of the same methods, where each method is extended to identify and link personal entity mentions, denoted with *PE*. Considering a personal entity mention m_{pe}, and an entity e, the PE method computes the cosine similarity between the word embedding of m_{pe} and the entity embedding of entity e. For every m_{pe}, we compute this similarity with all the previously linked entities in the conversation and find the most similar entity. Mention-entity pairs $\langle m_{pe}, e \rangle$ below a certain threshold τ are ignored. This threshold allows for filtering personal entity mentions that do not have the corresponding entities in the conversation history. We used Wikipedia2Vec [Yamada et al., 2020]

Table 2.5: Entity linking results on a sample of the WoW collection containing references to personal entities. The left block shows the results of the original EL methods and the right block represents the results of modified EL methods, where personal entities are also annotated.

	P	R	F		P	R	F
TagMe_t	62.2	50.2	55.6	TagMe_t+PE	61.7	51.1	55.9
TagMe_h	62.4	49.6	55.3	TagMe_h+PE	62.0	50.7	55.8
WAT_t	56.0	63.1	59.4	WAT_t+PE	56.2	64.4	60.0
WAT_h	55.6	**67.1**	**60.8**	WAT_h+PE	55.7	**68.8**	**61.6**
REL_t	64.0	18.0	28.1	REL_t+PE	63.2	18.6	28.8
REL_h	**78.0**	22.0	34.3	REL_h+PE	**77.2**	22.6	34.9

word and entity embeddings released by Gerritse et al. [2020a]. The threshold τ was set empirically by performing a sweep (on the range [0, 1] in steps of 0.1) using 5-fold cross-validation.

Comparing the left and right parts of Table 2.5, we observe a slight (albeit often negligible) performance increase for the PE method. These results show that identifying personal entity mentions and their corresponding entities is a non-trivial task and cannot be resolved with a simple extension of current approaches. This reinforces our finding that all examined EL tools are suboptimal for EL in conversations.

2.6 Conclusion

In this chapter, we studied entity linking in a broad setting of conversational systems: QA, task-oriented, and social chat. Using crowdsourcing, we analyzed existing conversational datasets and annotated them with concepts, named entities, and personal entities. We found that while both concepts and named entities are useful for understanding the intent of user utterances, personal entities are mainly important in social chats. Further, we compared the performance of different established EL methods in a conversational setting and concluded that none of the examined methods can handle this problem effectively, falling short in providing both high recall and precision, as well as annotating concepts, named entities, and personal entity mentions.

Our annotated conversational dataset (ConEL) and interface designs are made publicly available. These resources come with detailed instructions on the procedure of collecting the annotations, which can be used for further extension of the collection. Following the insights from this study, developing conversational entity linking methods and employing them in various types of conversational systems are obvious future directions.

Chapter 3

Entity Linking for Personalization

The analysis in Chapter 2 revealed that traditional Entity Linking (EL) methods are suboptimal for conversations, particularly struggling to identify the concepts and personal entities crucial for personalizing conversational systems. This finding necessitates the development of a new EL method tailored for conversational settings. Therefore, this chapter addresses the research question:

> **RQ2:** *How can we develop an Entity Linking method to accurately identify important entities for personalized conversational systems?*

To answer this question, this chapter introduces CREL, a conversational entity linking toolkit developed to identify named entities, concepts, and personal entities. Unlike existing EL methods, CREL utilizes coreference resolution to link personal references (e.g., "my dog") to explicit mentions within the conversation. CREL is trained on ConEL-2, an expansion of the original ConEL dataset (see Chapter 2) with more conversations for training. The findings show that CREL significantly outperforms state-of-the-art baselines in linking entities needed for personalized interactions.

3.1 Introduction

Understanding user utterances in conversational systems is fundamental to creating natural and informative interactions with users. Entity Linking is a powerful text understanding technique that has been extensively explored in previous studies [Shang et al., 2021] and Chapter 2, and has been shown to be effective in question-answering and retrieval tasks [Yamada et al., 2018, Gerritse et al., 2022, Dalton et al., 2014, Hasibi et al., 2016, Xiong et al., 2017, Dargahi Nobari et al., 2018]. However, the EL approaches developed for documents prove to be suboptimal for conversations, to the extent that the state-of-the-art EL models are outperformed by the outdated models when evaluated on conversational EL datasets (Chapter 2). The reasons are three-fold: **(i) Personal Entity Annotation**: Conversations often contain personal statements, referring to entities by their pronouns; e.g., "my cars." Identifying personal entities is essential for personalization, including building personal knowledge graphs (PKGs) [Balog and Kenter, 2019, Li et al., 2014, Tigunova et al., 2019, 2020] and generating personalized responses to users [Joshi et al., 2017, Luo et al., 2019]. Existing approaches for annotating documents, or even conversations [Laskar et al., 2022, Xu and Choi, 2022, Shang et al., 2021], do not identify and link personal entity mentions. **(ii) Concept Annotation**: EL models developed for documents or short texts focus mainly on annotating named entities [van Hulst et al., 2020, Cao et al., 2021]. In conversations, however, both named entities and concepts contribute to machine understanding of text and should be annotated as discussed in Chapter 2. **(iii) Training Data**: Existing EL approaches are trained on well-written texts, while conversations have multiple (informal) turns with information spread over the turns.

In this chapter, we overcome the above challenges and provide a toolkit and a training dataset as two resources for EL in conversations. The challenges are addressed in the following ways.

Personal Entity Annotation. We detect *personal entity mentions* (e.g., "my cars"), find their corresponding *explicit entity mentions* (e.g., "Life" and "Pilot"), and link them to the corresponding entities in the knowledge graph (see Figure 3.1). This task and coreference resolution both connect and stand in contrast to each other: they both aim to find mentions referring to the same entity [Singh et al., 2011,

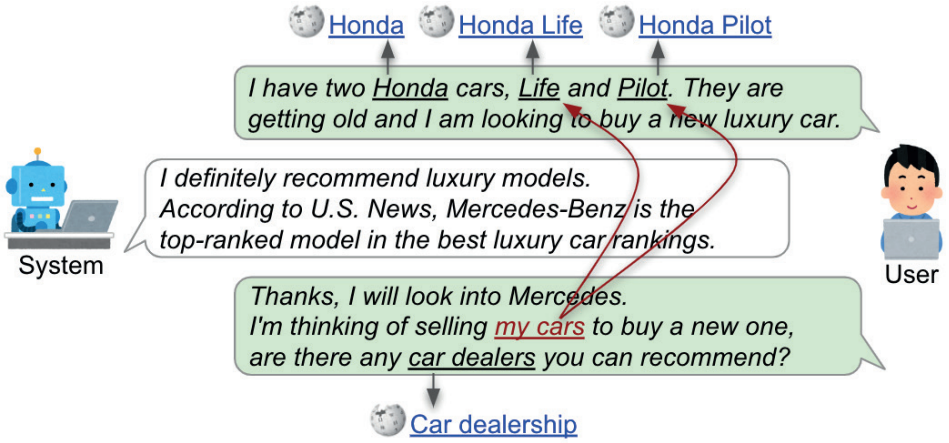


Figure 3.1: Example of entity linking in conversations.

Lee et al., 2018, Joshi et al., 2019, Kirstain et al., 2021], but personal entity mentions are sparse and different from common pronominal expressions that are resolved in coreference resolution. This implies that state-of-the-art coreference resolution methods cannot be used out-of-the-box for resolving personal entities. We, therefore, identify personal entity mentions in a separate module and modify coreference resolution techniques to map personal entities to their explicit entity mentions.

Concept Annotation. While open-domain knowledge graphs (e.g., Wikipedia) contain both named entities and concepts, conventional EL approaches are optimized for annotating only named entities, which hampers their performance for annotating conversations [van Hulst et al., 2020, Cao et al., 2021, Wu et al., 2020, Zhang et al., 2022]. Specifically, the performance of named entity and concept linking in the widely used REL entity linker [van Hulst et al., 2020] varies dramatically (F-score of 43.1 and 2.5, respectively). We mitigate this problem by training a mention detection model that can identify both concepts and named entity mentions in conversations and optimize the entity disambiguation model of REL [van Hulst et al., 2020] for all entity types. Our methods for personal entity, concept and named entity linking are integrated into our conversational EL tool, which is provided as a resource to the community. We show that this method outperforms state-of-the-art EL models.

Training Data. A key to solving the aforementioned problems is training data. Existing studies on conversational entity linking either use private proprietary datasets [Shang et al., 2021] or public datasets with an insufficient number of conversations for both training and evaluation purposes (Chapter 2). In this work, we build on the existing ConEL dataset (Chapter 2) and address its shortcomings in our new dataset: (i) In ConEL, each personal entity mention is mapped to only a single explicit entity, while in real conversations, personal entity mentions can have multiple corresponding entities (see Figure 3.1); (ii) ConEL contains a limited number of conversations with personal entities, as it synthesizes conversations mainly from question-answering and task-oriented benchmarks with a limited number of personal statements. Additionally, the ConEL-PE dataset, as in Chapter 2, generated by annotating real social chats, contains only 25 annotated conversations, which cannot be used for training purposes. Our newly developed dataset, referred to as *ConEL-2*, consists of 1,327 utterances, annotated with personal entities, concepts, and named entities. This dataset has been used for training and evaluation of our conversational EL tool.

In summary, this work makes the following resources available to the community:[1]

- An open-source tool for entity linking in conversations, which can uniquely identify personal entities, concepts, and named entities. Our developed conversational EL methods outperform state-of-the-art baselines.

- A dataset containing EL annotations for conversations, which can be used for training and evaluation purposes.

3.2 Method

Our conversational entity linking approach consists of two components; (1) personal entity linking, and (2) concept and named entity linking.

[1] `https://github.com/informagi/conversational-entity-linking-2022`

3.2.1 Personal Entity Linking

Inspired by coreference resolution approaches [Joshi et al., 2019, Kirstain et al., 2021, Lee et al., 2017, 2018], our model identifies personal entity mentions and their entity antecedents from conversation history. We disentangle mention detection from entity antecedent assignment and employ a BERT-based model (cf. Sect. 3.2.2) for detecting entity mentions and a rule-based approach [Li et al., 2014] for identifying personal entity mentions. Specifically, we follow Chapter 2 to identify personal mentions, starting with "my" or "our" tokens followed by adjectives, common nouns, proper nouns, pronouns, numbers, particles, and articles. Also, "of" and "in" are allowed to be part of the mention. While admittedly this approach cannot capture all forms of personal entity mentions, we believe that it covers a reasonably large number of cases to be used for early experimentation in this direction. We leave the comprehensive research on different forms of personal entity mention for future work.

Similar to Kirstain et al. [2021], we denote the start (s) and end token (e) representations of mentions as:

$$m^s = GeLU(\mathbf{W}^s v) \quad m^e = GeLU(\mathbf{W}^e v), \tag{3.1}$$

where $\mathbf{W}^s$ and $\mathbf{W}^e$ are trainable weights and v is contextualized representation of a token. We use LongFormer [Beltagy et al., 2020] to obtain contextualized embeddings as it can accommodate the long conversation history. The input to the LongFormer model is the current turn and conversational history. The hidden state in the output of the last layer of the model is used as v.

Assuming mention m_{bf} appears before mention m_{af}, the compatibility score between the two mentions is computed by:

$$\begin{aligned} score(m_{bf}, m_{af}) = m^s_{bf} \cdot B^{ss} \cdot m^s_{af} + m^s_{bf} \cdot B^{se} \cdot m^e_{af} \\ + m^e_{bf} \cdot B^{es} \cdot m^s_{af} + m^e_{bf} \cdot B^{ee} \cdot m^e_{af}, \end{aligned} \tag{3.2}$$

where B represents trainable weights and the superscripts s and e denote the start and end tokens of a mention. Equation 3.2 is computed for all pairs of personal entity mentions and explicit entity mentions. During inference, personal and explicit entity mention pairs with the *score* higher than threshold τ are selected.

3.2.2 Concept and Named Entity Linking

Entity linking typically involves two subtasks: (i) identifying mention boundaries and (ii) disambiguating mentions using a knowledge graph. In order to utilize existing advancements on these two subtasks, we need to adapt them to the conversational setup. For mention detection (MD), this translates to adapting MD models to the conversational domain and identifying mentions of both named entities (NEs) and concepts. We employ the vanilla BERT [Devlin et al., 2019] architecture and fine-tune it to predict BIO labels, corresponding to begin, inside, and outside of mentions. The BERT model utilized in this work is pre-trained on conversational datasets such as DailyDialogues [Li et al., 2017], OpenSubtitles [Lison and Tiedemann, 2016], Debates [Zhang et al., 2016], and Blogs [Schler et al., 2006].

For entity disambiguation (ED), we train REL [van Hulst et al., 2020] disambiguation approach on our ConEL-2 dataset. REL is one of the state-of-the-art EL tools that is optimized for document annotation, but its disambiguation approach, considering both the local and global context of mentions, can be seamlessly adapted to conversations. Our entity linking method, consisting of the aforementioned MD and ED models is named CREL, which is an extension of REL tool for conversations.

3.3 The ConEL-2 Collection

To facilitate the development of EL approaches for conversations, we construct the ConEL-2 collection. We draw the conversations from an existing conversational dataset. To select the dataset, we focus on social chat datasets, as personal entities are mainly found in social chat conversations (Chapter 2). We choose the Wizard of Wikipedia (WoW) [Dinan et al., 2019] dataset, which contains multi-domain multi-turn conversations collected from actual interactions between two people. The selected dialogues are annotated in three stages: (i) identifying personal entity mentions, (ii) pairing personal and explicit entity mentions, and (iii) entity linking. Unless stated otherwise, each human intelligence task (HIT) is annotated by three turkers, and additional turkers are added when needed based on our quality check strategies.

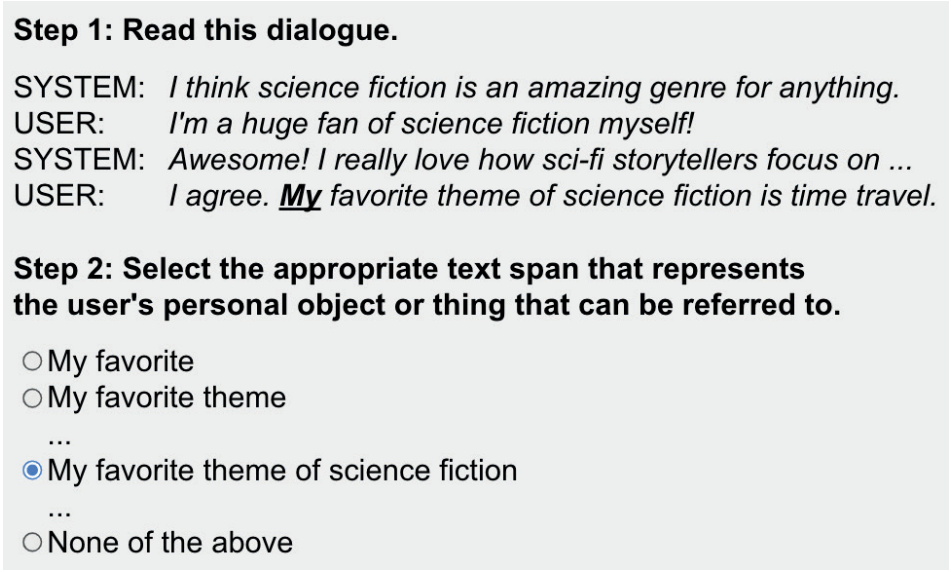


Figure 3.2: Annotation interface for identifying personal entity mentions (stage 1 of annotation process).

Stage 1: Identifying personal entity mentions. In this stage, we ask turkers to identify text spans referring to personal entities (i.e., personal entity mentions). First, we extract conversations that contain "my" or "our" in at least one user utterance, which amounts to 7,888 out of 22,311 conversations in the WoW dataset. To make the annotation process technically and financially feasible, we randomly select 1,000 conversations from the extracted dialogues. We then show the dialogue history and a set of candidate mentions to the turkers and ask them to find the best personal entity mention; the interface is shown in Figure 3.2. Candidate mentions are text spans starting with "my" or "our," followed by a maximum of 10 subsequent words. Our statistics on the collected annotations show that 98.9% of the HITs were agreed upon by two or more turkers, with a Fleiss' kappa of 0.864. We move forward with the 985 conversations, where at least two turkers agreed on all personal entity mentions of the conversation.

Stage 2: Pairing personal and explicit entity mentions. In this stage, we ask turkers to find the corresponding explicit entity mentions for a given personal entity mention. For example, in Figure 3.1, the personal entity mention "my cars" is shown to turkers, and they are asked

to find the corresponding explicit entity mentions "Life" and "Pilot." We create a pool of candidate explicit entity mentions by using existing EL tools: TagMe [Ferragina and Scaiella, 2010], WAT [Piccinno and Ferragina, 2014], REL [van Hulst et al., 2020], and GENRE [Cao et al., 2021]. The detected mentions are manually processed and meaningless duplicates such as "a restaurant" and "restaurant" are resolved. Turkers are then given the option to select one corresponding explicit entity mention or the "not in dialogue" option if the corresponding entity mention is not found in the dialogue. Fleiss' kappa for the inter-annotator agreement is 0.434, which is considered moderate. To improve the annotation quality, two extra annotations are collected for HITs where two turkers agreed on one option. We then select all conversations with equal or more than three turkers agreed on one option (either an explicit entity mention or the "not in dialogue" option). This selection amounts to 290 conversations, which are passed to the next step.

We note that one of the shortcomings of the existing ConEL dataset is identifying only one explicit entity mention for each personal mention (cf. Section 3.1). In ConEL-2, we address this issue and ask turkers to find another corresponding explicit entity mention for all extracted conversations if any.

Stage 3: Entity linking. In this stage, we ask turkers to select the corresponding Wikipedia article to the given entity mention, e.g., mapping the mention "car dealers" to the entity *Car dealership* in Figure 3.1. We first use the EL tools described in the stage 2 to create the pool of candidate mention-entity pairs. Noisy mentions such as "please" and "I am" are manually removed from the pool. We also include top-10 Wikipedia search[2] results using mentions as queries. Turkers are then asked to select one of the candidate entities for each entity mention if any.

The statistics show 98.7% of the HITs were agreed by two or more turkers (Fleiss' kappa of 0.839), which highlights the high quality of our annotation process. We therefore include all annotation results which are agreed by at least two turkers. For the remaining 1.3% of the non-agreed HITs, we recollected two extra annotations for each and select the entity that is selected by the majority of turkers. The

[2] `https://www.mediawiki.org/wiki/API:Main_page`

Table 3.1: Statistics of the ConEL-2 collection. Personal entity annotations represent the number of annotations for personal entity mentions, with or without corresponding explicit entity mentions in the dialogue.

	Train	Val	Test
Conversations	174	58	58
User utterances	800	267	260
NE and concept annotations	1428	523	452
Personal entity annotations	268	89	73

annotation results are shown in Table 3.1. We split the dataset into train, validation, and test set with the 60%-20%-20% ratio.

3.4 Results

3.4.1 Personal Entity Linking

Experimental Setup. We take the LongFormer-large model[3] [Beltagy et al., 2020] pre-trained on OntoNote Release 5.0 [Hovy et al., 2006], and fine-tune it on the training set of the ConEL-2 dataset. This model is referred to as $S2E_{conv}$. We perform training with the batch size of 8 and learning rate of 10^{-5} using AdamW optimizer, and the validation set of ConEL-2 is used for early stopping. For evaluation, we use validation and test sets of the ConEL-2 datasets as well as the ConEL-PE dataset (Chapter 2). ConEL-PE contains 25 dialogues from the WoW dataset. We only use personal entity annotations for this evaluation.

Baselines. We compare our model with the Wikipedia2Vec-based method (W2V) as in Chapter 2, where antecedent assignment is based on similarity between Wikipedia2vec [Yamada et al., 2020] embeddings of entities and personal entity mentions. We also report on the official S2E model[4] [Kirstain et al., 2021] and its variation $S2E^{-}_{conv}$, where fine-tuning is performed only on our conversational dataset (without using OntoNote dataset).

[3] https://huggingface.co/allenai/longformer-large-4096

[4] https://github.com/yuvalkirstain/s2e-coref

Table 3.2: Personal entity linking results. Micro-averaged precision, recall, and F1 scores are computed.

	ConEL-2 (Val)			ConEL-2 (Test)			ConEL-PE		
	P	R	F	P	R	F	P	R	F
S2E [Kirstain et al., 2021]	.375	.415	.394	.292	.302	.297	.317	.481	.382
W2V (Chapter 2)	.324	.508	.395	.458	.429	.443	.436	.630	.515
$S2E^{-}_{conv}$	.494	**.615**	.548	.450	**.571**	.503	.400	**.741**	.519
$S2E_{conv}$	**.714**	.538	**.614**	**.673**	.524	**.589**	**.613**	.704	**.655**

Results. We measured the performance of our rule-based personal entity mention detection approach and obtained micro-averaged precision, recall, and F1 of 0.908, 0.898, and 0.903, respectively. This indicates that identifying personal entity mentions can be handled effectively by a simple approach. Assigning entity antecedents, on the hand, is a challenging task.

Table 3.2 shows the results for personal entity linking. We observe that $S2E_{conv}$ fine-tuned on both OntoNote and conversational datasets outperforms all baselines. It also indicates that, fine-tuning on conversational data substantially improves personal entity linking performance (S2E vs. $S2E_{conv}$ results).

3.4.2 Concept and Named Entity Linking

Experimental Setup. For mention detection, we fine-tune bert-base-cased-conversational[5] on our training data with the batch size of 16, learning rate of 5×10^{-5}, and AdamW optimizer, with the validation set of ConEL-2 for early stopping. For ED, we train the REL ED model, following the same training procedure as van Hulst et al. [2020]. As input for these models, we use only current turns for efficiency; although using the entire conversation history has shown slightly better performance (Chapter 2), it doesn't change our conclusion, as it mainly benefits REL, with marginal benefits for WAT and TagMe. For evaluation, we use the ConEL-2 and ConEL datasets. ConEL consists of 100 conversations that are collected from three different conversational tasks (question answering, task-oriented, and social

[5] `https://huggingface.co/DeepPavlov/bert-base-cased-conversational`

Table 3.3: Concept and named entity linking results. F_{MD} and F_{EL} represent the micro-averaged F1 scores for mention detection and entity linking, respectively.

	ConEL-2 (Val)		ConEL-2 (Test)		ConEL	
	F_{MD}	F_{EL}	F_{MD}	F_{EL}	F_{MD}	F_{EL}
GENRE [Cao et al., 2021]	.290	.252	.320	.299	.350	.211
REL [van Hulst et al., 2020]	.304	.244	.279	.231	.462	.245
TagMe [Ferragina and Scaiella, 2010]	.559	.478	.611	.504	.510	.375
WAT [Piccinno and Ferragina, 2014]	.616	.539	.613	.519	.416	.336
REL*($BERT_{conv}$)	**.748**	**.652**	.716	.587	.551	.424
REL*($BERT_{WP\text{-}conv}$)	.697	.624	.708	.592	.508	.399
REL*($BERT_{NER\text{-}conv}$)	.723	.635	.693	.576	.548	.407
CREL	.742	.651	**.729**	**.597**	**.559**	**.429**

chat) and annotated with concepts and named entities. Only concept and named entity annotations are used for this evaluation.

Baselines. REL [van Hulst et al., 2020], GENRE [Cao et al., 2021], TagMe [Ferragina and Scaiella, 2010], and WAT [Piccinno and Ferragina, 2014] are strong publicly available EL tools that are used as our baselines. Additionally, we train three BERT-based MD models: (i) $BERT_{conv}$: The bert-base-uncased model,[6] fine-tuned on our training data, (ii) $BERT_{WP\text{-}conv}$: Similar to $BERT_{conv}$, but fine-tuned on the Wikipedia anchor links from Cao et al. [2021] before fine-tuning on ConEL-2 training data, and (iii) $BERT_{NER\text{-}conv}$: The bert-base-NER model which is originally trained for entity recognition (NER),[7] fine-tuned on our training data. The fine-tuning procedures are the same as CREL. We combine these models with the REL ED model trained on conversations (denoted as REL*) and report on their results.

Results. The results are reported in Table 3.3. F_{MD} and F_{EL} are the F1 scores for MD and EL, respectively. The results show that the REL and GENRE, while being state-of-the-art EL toolkits, perform worse than TagMe and WAT. This is because REL and GENRE are intended to detect only named entities, while TagMe and WAT are able to identify concepts as well. Comparing CREL with REL* methods, we observe that adaptation of BERT for conversational domain plays

[6] `https://huggingface.co/bert-base-uncased`

[7] `https://huggingface.co/dslim/bert-base-NER`

Table 3.4: Overall entity linking results, annotating both personal entities (PE) and concepts and named entities (EL).

		ConEL-2 (Val)			ConEL-2 (Test)			ConEL-PE		
EL	PE	P	R	F	P	R	F	P	R	F
REL	W2V	.523	.159	.244	.516	.159	.243	.488	.206	.290
REL	$S2E_{conv}$	.523	.162	.248	.522	.165	.250	.466	.206	.286
GENRE	W2V	.468	.180	.260	.569	.223	.320	.537	.256	.347
GENRE	$S2E_{conv}$	.589	.169	.263	**.675**	.217	.328	**.646**	.256	.367
TagMe	W2V	.504	.407	.451	.530	.440	.481	.528	.518	.523
TagMe	$S2E_{conv}$	.612	.395	.480	.621	.434	.511	.636	.518	.571
WAT	W2V	.506	.489	.497	.545	.464	.501	.536	.628	.579
WAT	$S2E_{conv}$	.585	.490	.534	.582	.464	.516	.571	.648	.607
CREL	W2V	.636	.589	.612	.607	.554	.579	.573	.714	.635
CREL	$S2E_{conv}$	**.712**	**.593**	**.647**	.634	**.566**	**.598**	.600	**.724**	**.656**

an important role in conversational entity linking, reflected by the highest score for ConEL-2 (test) and ConEL datasets.

3.4.3 Overall Performance

Putting all the pieces together, we report on the overall performance of our model for linking personal entities, concepts and named entities. For this experiment, we report only on the ConEL-2 and ConEL-PE datasets, as ConEL does not have annotations for personal entities. The results are shown in Table 3.4. We notice at first glance that our proposed personal entity linking extension ($S2E_{conv}$) leads to improved F-scores compared to W2V in most cases. The results also show that our EL tool (CREL with $S2E_{conv}$) achieves the highest F-score for all sets, reinforcing the importance of adapting EL approaches for conversations.

3.5 Conclusion

In this chapter, we tackled the problem of entity linking in conversational settings, where not only named entities, but also concepts and personal entities are important. To this end, we collected conversational entity annotations, which allows us to train and evaluate entity

linking in conversations. Additionally, based on the collected annotations, we develop a conversational EL tool. Our empirical results show that our EL tool achieves the highest performance over conventional EL tools for documents and short texts. The collected annotations and our EL tool with detailed instructions are publicly available as a research resource for further study in conversations. Future work includes incorporating coreference resolution methods to identify more diverse personal entity mentions.

Chapter 4

Personalized Response Generation

The previous chapters developed methods for personal information extraction, yet effectively utilizing extracted personal context for personalized response generation poses a different challenge. The central research question this chapter addresses is:

> **RQ3:** *How can users' personal preferences be utilized to generate personalized responses?*

The difficulty in addressing this question stems from the lack of realistic large-scale dialogue datasets for training and evaluating personalized response generation models. The existing datasets often lack the multi-session nature and real-world user preferences, while previous collection approaches that rely on experts are difficult to scale.

In this chapter, we first introduce LAPS, a method that leverages large language models (LLMs) to guide human workers in efficiently generating personalized dialogues to answer this research question. The LAPS approach collects large-scale, human-written, multi-session, and multi-domain conversations with real user preferences.

Building upon LAPS, to answer RQ3, we then train a preference extraction model and perform personalized response generation using a semi-structured *preference memory* constructed from user's preferences. The experiments demonstrate that preference memory improves accuracy, provides explanations, and mitigates LLMs' recall issues in response generation.

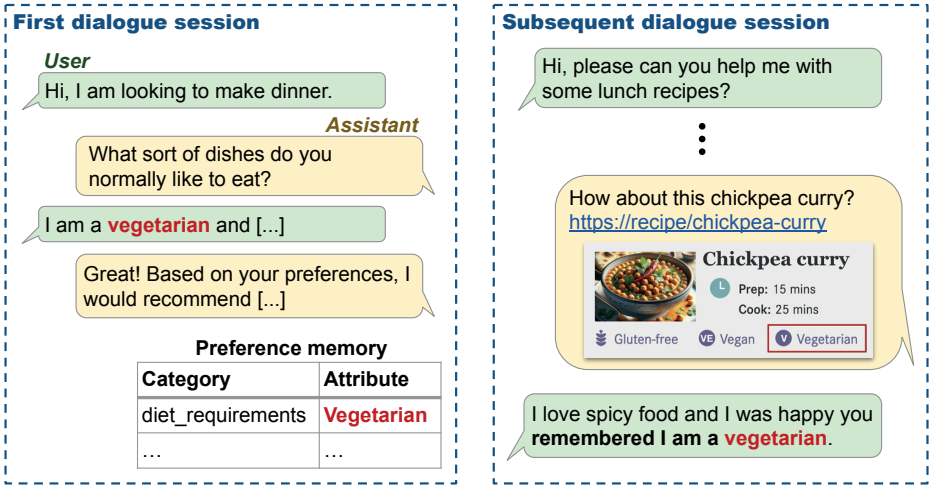


Figure 4.1: A snippet from a multi-session dialogue. User preferences are extracted and stored in memory to generate personalized recommendations in subsequent sessions.

4.1 Introduction

Personalization is paramount for conversational search and recommendation [Luo et al., 2019, Brown and Levinson, 1987]. Conversational agents need to meet users' expectations and provide them with individually tailored responses. In real-world scenarios, where users interact with dialogue agents across multiple sessions, conversational systems need to accurately understand, extract, and store user preferences and articulate personalized recommendations based on the stored user profile; see Figure 4.1. The Information Retrieval (IR) community has studied various aspects of personalization in conversational systems. For instance, the concept of Personal Knowledge Graph (PKG) [Balog and Kenter, 2019] is introduced to enable personalization of (conversational) search systems. Similarly, TREC Interactive Knowledge Assistance Track (iKAT) utilizes Personal Text Knowledge Base (PTKB) [Aliannejadi et al., 2024] for persona-based conversational search. Recent years have also witnessed tremendous progress in LLMs [Brown et al., 2020, Ouyang et al., 2022]. Yet LLM-based conversational agents do not effectively handle user preferences, delivering generalized recommendations that fail to capture the nuanced interests of individual users.

A major hindrance in developing a personal conversational system is the unavailability of *large-scale human-written* conversational datasets. These are needed for training new models and understanding user behavior, in expressing their preferences over multiple dialogue sessions [Radlinski et al., 2019, Joho et al., 2017]. However, constructing such datasets has proven a daunting task [Bernard and Balog, 2023, Radlinski et al., 2019]. Preference elicitation in conversations is complex, and crowd workers engage poorly with the task. While human experts deliver quality results, recruiting them as intermediary coaches or as human agents to generate conversations does not scale. The recently-emerged paradigm of collecting synthetic conversational data through LLMs [Gilardi et al., 2023, Yunxiang et al., 2023, Lee et al., 2022, Xu et al., 2023, Chen et al., 2023, Kim et al., 2023, 2022, Jandaghi et al., 2023, Cheng et al., 2025, Askari et al., 2025] raises concerns about dialogue diversity [Reif et al., 2023, Chung et al., 2023, Yu et al., 2023, Dell'Acqua et al., 2023, Park et al., 2023]. Crucially, LLM-generated conversations do not represent actual user preferences and interactions, undermining the credibility for developing *future* personal conversational systems.

The critical question that arises here is:

> **RQ3a:** *How can we collect cost-effective yet large-scale multi-session human-written conversational datasets for personalized response generation?*

We address this question by proposing LAPS, an **L**LM-**A**ugmented **P**ersonalized **S**elf-Dialogue method to collect large-scale personal conversations. LAPS employs an LLM to dynamically generate personal *guidance* for crowd workers, playing both user and assistant roles. The guidance is generated based on the previously elicited user preferences and the current state of the dialogue, determined by a dialogue act classifier. After each dialogue session, LAPS extracts preferences from the dialogue and stores them in a *preference memory*. This memory is a key-value store of user preferences, analogous to the PKG [Balog and Kenter, 2019] and PTKB [Aliannejadi et al., 2024] concepts. Using LAPS, we collect 1,406 multi-domain multi-session dialogues, paired with 11,215 preferences.

Our follow-up research question concerns the quality of LAPS-generated datasets:

> **RQ3b:** *How do the LLM-augmented self-dialogues compare to human- and synthetically-generated conversations?*

We compare our conversations with a wide range of widely used conversational datasets and show that LAPS-generated conversations score higher in diversity (based on Dist-n, Ent-n, and SELF-BLEU metrics) and quality (based on UniEval [Zhong et al., 2022]). We further compare LAPS- and LLM-generated dialogues using GPT-3.5 and GPT-4 and show that LLM-generated dialogues are less diverse than those involving humans, even with temperature tuning.

Although LAPS extracts preferences in a semi-structured format and stores them in a preference memory, one could wonder whether such memory is needed, given LLMs' abilities in handling long context from the previous sessions. This leads us to the second follow-up research question:

> **RQ3c:** *How can preference memory enhance the effective utilization of user preferences in recommendations?*

To address this question, we train a preference extraction model on our dataset and use it to build a preference memory from previous sessions. These preferences are then incorporated into the LLM's prompt for generating personalized recommendations. We compare this approach to the baseline prompting method, where dialogue histories of all sessions are appended to the prompt. Our experiments show that by incorporating preference memory, the model can more accurately utilize the users' disclosed preferences for recommendations than the baseline method. The notable advantage of preference memory is that it contributes to more explainable recommendations. Finally, we found that when using the baseline method, the LLM struggles with recalling user preferences; likely due to lengthy prompts.

Contributions of this work are as follows:

- We introduce LAPS method for collecting scalable multi-session personalized dialogues with actual user preferences.
- We analyze and compare various dialogue collection methods, demonstrating that LAPS collects lexically diverse and high-quality dialogues, uncovering the diversity issue of generating fully-synthetic dialogues with LLMs.

- We study the benefits of storing and using user preferences in a semi-structured format (preference memory) and show that it helps an LLM in recalling previously disclosed preferences when generating personalized recommendations.

- Enabled by LAPS, we release a unique conversational dataset that is multi-session, human-written, large-scale, and contains users' personal preferences.[1]

4.2 Related Work

4.2.1 Dialogue Collection Methods

Human-Human interactions are arguably the optimal strategy for collecting natural dialogues. MultiWOZ [Budzianowski et al., 2018] and PersonaChat [Zhang et al., 2018a] are notable examples, offering task-oriented and chit-chat datasets, respectively. These datasets, however, focus less on real user preferences, relying instead on predefined tasks or personas. Addressing this limitation, Radlinski et al. [2019] introduced CCPE-M, a dataset emphasizing actual user preferences. Similarly, Bernard and Balog [2023] introduced MG-ShopDial, an e-commerce conversational dataset with genuine preferences. Despite their quality, the small size of these datasets (502 dialogues for CCPE-M and 64 for MG-ShopDial) limits their utility for training large models, and they overlook the multi-session aspect of real-world dialogues.

A quality-quantity trade-off in dialogue collection is highlighted by Bernard and Balog [2023]. Initially attempting crowdsourcing, Bernard and Balog [2023] shifted to a volunteer-based collection due to low engagement, resulting in higher quality but fewer dialogues. This underscores the challenge in gathering large, high-quality datasets representing true user preferences.

Self-Dialogue, where a single worker simulates both roles, is an effective approach for large-scale data collection. Introduced by Krause et al. [2017], this technique has been shown to produce high-quality data, with Fainberg et al. [2018] noting its increased coherence and reduced errors compared to human-human dialogue. Byrne et al. [2019]

[1]The code and dataset is available at `https://github.com/informagi/laps`

further validated this, emphasizing the superior linguistic diversity and fewer mistakes in self-dialogue data. However, self-dialogue has limitations, especially in managing complex conversations, such as those involving preference elicitation, while acting in two distinct roles. Additionally, self-dialogue cannot authentically capture unknown user preferences as the same person plays both user and assistant roles, leading to fewer clarification scenarios than in human-human interactions [Fainberg et al., 2018]. Our approach introduces LLM-augmented personalized self-dialogues, leveraging LLMs to ease the cognitive load on workers and addressing the limitation of capturing unknown preferences by involving an LLM, which doesn't have pre-existing knowledge of user preferences.

(Semi)-Synthetic Dialogue Generation offers an alternative to relying on crowd workers [Soudani et al., 2023, Lee et al., 2022, Xu et al., 2023, Chen et al., 2023, Kim et al., 2023, 2022, Jandaghi et al., 2023, Leszczynski et al., 2023]. Lee et al. [2022] developed PERSONACHATGEN using two LLMs for dialogues between personas, requiring multiple model calls for one dialogue. Chen et al. [2023] optimized this by using a single model call with a carefully crafted prompt. Leszczynski et al. [2023] took a different approach, generating synthetic dialogues by combining user-generated content and metadata. However, concerns exist about the diversity in LLM-generated texts. Reif et al. [2023] noted lexical and syntactic repetition, developing LinguisticLens for syntactic diversity analysis. Chung et al. [2023] emphasized the need for human intervention to enhance diversity, and Yu et al. [2023] pointed out the uniformity in LLM outputs from simple prompts, suggesting diverse prompts for more varied data. The diversity issue is also evident in fields like social science and business [Dell'Acqua et al., 2023, Park et al., 2023].

To address this, semi-synthetic data collection methods like having crowd workers edit LLM-generated texts have been effective [Shah et al., 2018, Rastogi et al., 2020, Bae et al., 2022]. Shah et al. [2018] introduced M2M, a framework using templates for initial dialogue generation and crowd workers for rewriting. Similarly, Rastogi et al. [2020] used a schema-guided method. Our research improves upon these by using LLMs for providing workers with guidance for response composition, promoting diversity and reducing the influence of generated drafts.

4.2.2 Personalized Conversational Systems

Problem-driven conversational systems, especially those for search and recommendation, benefit from personalization. Shifting from traditional approaches such as collaborative filtering, recent personalization focuses on more interactive approaches, such as preference elicitation [Radlinski et al., 2019, Li et al., 2023, Christakopoulou et al., 2016, Sun and Zhang, 2018, Zhao et al., 2022, Kostric et al., 2021, 2023]. These elicited preferences can be stored as a knowledge graph [Balog and Kenter, 2019] through entity linking [van Hulst et al., 2020] or in text format [Aliannejadi et al., 2024], and later used for personalization. Storing user preferences in a (semi-)structured format enables conversational agents to better satisfy users' information needs and can be also useful to mitigate bias [Gerritse et al., 2020b]. Following this line, our method elicits user preferences and stores them in a semi-structured preference memory.

Memory and feedback also offer the promise of making systems better aligned with user needs. This approach, tracing back to ALFRED [Riesbeck, 1981], involves storing past failures to improve future interactions. Recently, Madaan et al. [2023] advanced this concept with SELF-REFINE, enabling LLMs to refine their outputs iteratively using their own feedback. For tasks requiring deeper personalization, human feedback is invaluable. Madaan et al. [2022] enhanced LLM response quality by adding a memory module that remembers information from the user's past sessions. Aligned with Madaan et al. [2022], our approach includes a memory module to store user preferences from past interactions, enabling enhanced personalization.

4.3 Dialogue Collection Method

We propose LAPS, an LLM-Augmented Personalized Self-Dialogue construction method, capable of collecting large-scale, human-written, multi-session, and multi-domain conversations, paired with extracted user preferences. The method consists of four key elements (cf. Fig. 4.2): (i) dialogue act classification, (ii) guidance generation, (iii) utterance composition, and (iv) preference extraction.

The dialogue act classifier determines the next action that the assistant should take; e.g., recommend. Based on the dialogue act,

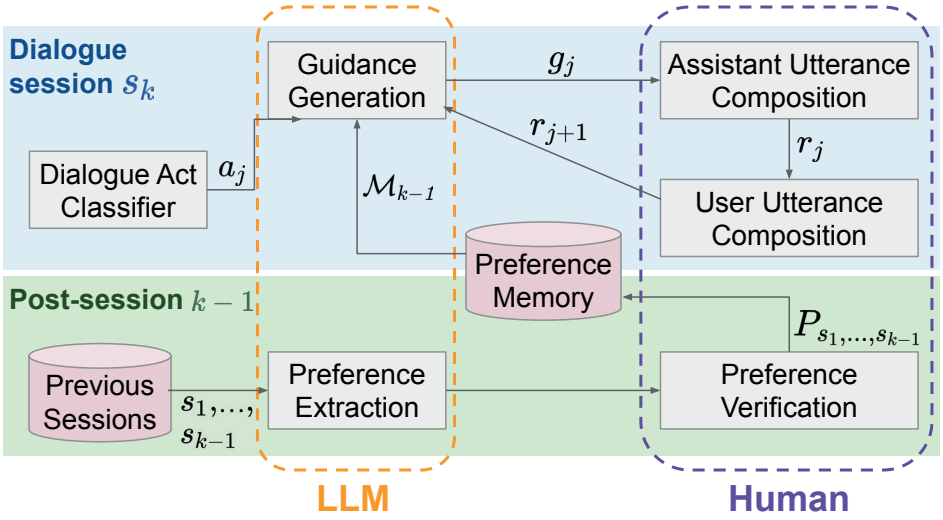


Figure 4.2: Overview of our dialogue collection method (LAPS).

the LLM generates guidance considering the dialogue history and the previously extracted preferences. The human agent then composes the assistant response based on the LLM-generated guidance, and then switches to the role of a user, providing a response to the previous utterance. The process continues until a relevant recommendation is made and the dialogue session is completed. Upon completion of a session, the preferences are extracted from the dialogue using an LLM and checked by the human agent. These preferences, once confirmed by the same human agent, are stored in the *preference memory* and used in subsequent sessions for generating personalized guidance. The human agent is then encouraged to initiate a new dialogue session for another scenario in the given domain. This process continues until the human agent exits the job or reaches the end of all pre-defined session scenarios.

4.3.1 Task Formulation

The objective of this task is twofold: firstly, to create a large-scale collection of multi-session dialogues written by humans, focusing on user preferences; and secondly, to extract and compile the specific user preferences mentioned within these dialogue sessions. Formally, the

dialogue collection method F is defined as a mapping from a set of task descriptions $\mathbb{T}$ and human agents $\mathbb{H}$ to a set of dialogue sessions and their extracted preferences $\mathbb{S}$:

$$F : \mathbb{T}, \mathbb{H} \rightarrow \mathbb{S}$$
$$F(t, h) = [(s_1, P_{s_1}), \ldots, (s_n, P_{s_n})],$$

where t is a task description for a topic, h is a human agent with identical preferences for a given topic, and n represents the total number of sessions. The dialogue session s_i and its corresponding preference set P_{s_i} are defined as:

$$s_i = [u_1^i, u_2^i, \ldots, u_m^i],$$
$$P_{s_i} = \{(c, \wp_j) \,|\, c \in C, j \in \mathbb{N}\}, \tag{4.1}$$

where the dialogue session s_i composed of a sequence of utterances u, and the preference set P_{s_i} is a set of category-preference pairs $(c, \wp_j)$, where the category c belongs to the set of categories C; e.g., {`allergy`, `cuisine`, `diet`}. We note that the preference set P_{s_i} is generated only after the completion of session s_i, by extracting user preferences from the dialogue post-session and validating them with the same worker. The extracted preferences of a human agent are stored in a memory component $\mathcal{M}$, defined as:

$$\mathcal{M}_k = \bigcup_{\substack{1<i\leq k \\ k\leq n}} P_{s_i},$$

where session s_k is the last completed session by the human agent.

Here we draw an analogy between the preference memory $\mathcal{M}$ and PKGs [Balog and Kenter, 2019], where the personal information of users is stored according to an ontology. The `<subject, predicate, object>` triplets in PKGs correspond to human (h), category (c), and preference ($\wp$) in preference memory, respectively. We note that unlike a PKG that is built based on a pre-defined ontology, preference memory uses a more relaxed version of categories that are extracted on-the-fly from user utterances. Preference memory can be also viewed as a semi-structured form of PTKB, where free-form sentences about a user's persona are transformed in a key-value format.

4.3.2 Guidance Generation

Generating guidance for human agents is central to collecting large-scale, high-quality, human-written utterances in LAPS. Large-scale construction of conversational data requires recruiting crowd workers. However, due to the complexity of the task and high cognitive load of generating conversational utterances, crowd workers show poor engagement [Bernard and Balog, 2023]. The challenge is even more intense for the preference elicitation task, where we need to coach the human agent simulating the system to ask engaging questions to reveal user preferences [Radlinski et al., 2019]. A remedy could be utilizing LLMs to generate system utterances. This, however, results in less diverse conversations and is not in line with our aim of generating training data for *future* personalized conversational search and recommendation systems. Even by instructing crowd workers to re-write LLM-generated utterances, we observed (in our pilot studies) that crowd workers become less creative in generating their own utterances and tend to replicate pre-generated utterances.

To alleviate the aforementioned problems, we propose to coach crowd workers throughout the conversation using automatically generated personalized guidance, and let the workers compose their own utterances via dialogue self-play. Using this approach, we reduce the cognitive load of a highly complex task to a minimum, allowing workers to focus on a simple sub-task at a time and generate high-quality and engaging conversations. Formally, to compose the assistance utterances u_j, the human agent receives personalized guidance g_j generated by an LLM. The guidance generation function $\mathcal{G}$ is defined as:

$$\mathcal{G}\left(u_1, \ldots, u_{j-1}, a_j, \mathcal{M}_k\right) = g_j,$$

where $\mathcal{M}_k$ denotes the preference memory extracted from sessions $(s_1, \ldots, s_k)$, utterances $u_1, \ldots, u_{j-1}$ represents conversations history up until turn j, and a_j is the action that needs to be taken for turn j, obtained from the dialogue act classifier (cf. Subsection 4.3.3).

Instantiation. As an instantiation of this function, we prompt GPT-3.5 `turbo` to generate personalized guidance. The guidance prompt template takes the dialogue history $u_1, \ldots, u_{j-1}$, preference memory $\mathcal{M}_k$, and instructions for the current dialogue act a_j as inputs. The prompts include detailed step-by-step instructions for chain-of-thought

prompting [Wei et al., 2022]. For instance, the prompt for the *preference elicitation* act in a given `DOMAIN` is as follows:

> *You are an advisor, who supports* `$DOMAIN` *recommendation assistants to compose responses to users. In this step, the assistant* ***collects information about user's preferences*** ($[a_j]$).
>
> Preference memory: $[\mathcal{M}_k]$
> Dialogue history: $[u_1, \ldots, u_{j-1}]$
>
> *Step 1: Identify the last user turn.*
> *Step 2: Explain the intent of the last user utterance.*
> *Step 3:* ***Which preference(s) should the assistant ask next*** *given the dialogue history?*
> *Step 4:* ***Compose very short guidance for the human assistant on how to write a response to the user.***
> *Step 5: Output in JSON format following [...]*
> *Ensure that you distinctly label and delineate Steps 1, 2, 3, 4, and 5.*
> *Let's think step by step:*

The guidance prompt for *recommendation* act is similar to the *preference elicitation* act, except that the assistant is instructed to recommend an item based on the user's personal preferences with URLs. The guidance also includes "*When making recommendations, if necessary, effectively utilize the user's preferences, such as [...]*" to encourage the assistant to use the user's preferences disclosed in the previous sessions if necessary.

4.3.3 Dialogue Act Classification

Dialogue act is an action that can change the (mental) state of conversation and guide the system to generate the next utterance [Gao et al., 2019]. Dialogue act is used as an input to the guidance generation function and is obtained by the dialogue act classifier $\mathcal{A}$:

$$\mathcal{A}(u_1, \ldots, u_{j-1}, a_{j-1}) = a_j$$

which determines action a_j based on the dialogue history $u_1, \ldots, u_{j-1}$ and the previous dialogue act a_{j-1}.

Instantiation. We instantiate the act classifier function by prompting GPT-3.5 `Turbo`. A series of dialogue acts are defined, outlining the

specific actions to be taken sequentially. The primary dialogue acts in our setup are: (1) *greeting*, (2) *preference elicitation*, (3) *recommendation*, (4) *follow-up questions*, and (5) *goodbye*. A dialogue act is selected upon completion of the previous act, which is determined by the LLM using detailed chain-of-thought instruction prompts; e.g., the *recommendation* act is selected when the LLM produces the response "true" for the instruction prompt "*[...] Has the user shared any of the preferences listed above? Has the assistant collected why the user has the preference? If both are true, return true.*"

When collecting preferences for the first session, the *preference elicitation* act is further divided into three sub-actions for collecting (i) *must-have*, (ii) *should-have*, and (iii) *could-have* preferences. Once the *must-have* and *should-have* preferences are collected, the subsequent sessions only collect *could-have* preferences, using the preference memory for other preferences.

4.3.4 Utterance Composition

Conversation utterances are composed by the human agent for both user and assistant roles via dialogue self-play. For the assistant role, the composition process is supported by the LLM-generated *guidance*, which enables generating diverse human-written system utterances. For the user role, the human agent mainly needs to elaborate his preferences and state his opinion about the recommendations. The system and user utterance generation functions are defined as:

$$\mathcal{W}_s\left(S_k, g_j, u_1, \ldots, u_{j-1}\right) = u_j,$$
$$\mathcal{W}_u\left(S_k, u_1, \ldots, u_j\right) = u_{j+1},$$

where $\mathcal{W}_s$ and $\mathcal{W}_u$ represent the function for writing system and user responses, respectively. Here, g_j denotes the guidance for generating utterance u_j, and $S_k = \{s_1, \ldots, s_k\}$ represents the session history, comprising all previous dialogue sessions up until, but not including, the current session.

Instantiation. These functions are instantiated by recruiting crowd workers and instructing them to write a self-dialogue for our task: *"Your task is to chat with yourself both as a user (seeking a cooking recipe or movie) and an assistant (offering recipe or movie recommendations).*

You need to discuss preferences and receive suggestions while playing both roles." This instruction is followed by brief descriptions of roles, session settings, and chat interface instructions, as well as general information about the payment and rejection policy.

For the user role, workers are instructed to be themselves, provide their preferences, review the recommendations, and offer feedback. In contrast, the assistant role entails more tasks. Assistants must elicit preferences to inform their recommendations and provide URLs for the recommended items. Following user feedback, assistants confirm why the user likes or dislikes the recommendation to obtain a reusable preference for the next session.

4.3.5 Preference Extraction

Upon completion of a dialogue session, user preferences are extracted from the dialogue. Due to the inherent ambiguity and complexity of human-written conversations, we collect these preferences from the same human agent that generates the conversation. Note that here we only collect user preferences that are mentioned in the course of previous conversation sessions and not general user preferences. Formally, given the dialogue session s_i, the preference extraction function $\mathcal{E}$ is defined as:

$$\mathcal{E}(s_i, C) = P_{s_i}, \tag{4.2}$$

where C denotes the set of preference categories, and P_{s_i} is the set of category-preference pairs extracted from the dialogue session.

Instantiation. Extraction of preferences is performed using a semi-automated approach, where the initial set of preferences is extracted by an LLM and then validated by the human agent. We prompt GPT4 with chain-of-thought instructions to generate preference attributes for each preference category $c \in C$, given the dialogue session s_i and preference categories C.

Once an initial set of preferences is extracted, the worker is instructed to confirm the correctness of each extracted preference and verify all disclosed preferences during the conversations are extracted. The worker is then directed to start a new conversation session or terminate the task.

4.3.6 Evaluation

Baseline Datasets. We evaluate our LAPS method by comparing existing conversational datasets with the conversations generated by LAPS. We carefully select the baseline datasets by examining 170 datasets listed in Chapter 2, and narrowing down our selection by the following criteria: (i) inclusion of preference elicitation, (ii) utilization of preferences, and (iii) focus on task-oriented dialogues. We prioritize datasets that are both published in peer-reviewed venues and publicly available. The selected datasets are summarized in Table 4.1. For a fair comparison, when possible, we select a single domain from each baseline dataset that overlaps with the domains of the dataset created by LAPS, namely recipe and movie. For open-domain datasets, e.g., PersonChatGen [Lee et al., 2022], we select dialogues with food-related personas, categorized under "Food" and "Drink."

Baseline Methods. We tried three other human-based dialogue collection methods: *Human-Human*, *self-dialogue*, and *LLM-human*. These methods do not scale and cannot generate quality conversations. In the *Human-Human* method, we encountered difficulties in pairing up two workers simultaneously due to the high dropout rates, caused by the complex nature of multi-session preference elicitation. For the *self-dialogue* method, we followed Speggiorin et al. [2022] and simplified the assistant role by providing users with a predefined response. This, however, led to homogeneous dialogues, even though the workers were instructed to rewrite the response. In the *LLM-human* method, we generated assistant responses using GPT-3.5 `turbo` and asked workers to rewrite the utterances. However, even after experimenting with multiple temperature settings, the dialogue remained homogeneous (a common issue with LLMs reported also [Reif et al., 2023, Chung et al., 2023, Yu et al., 2023, Dell'Acqua et al., 2023, Park et al., 2023]) and workers often accepted the grammatically correct responses without re-writing them.

While we focus on collecting dialogues with actual user preferences, one might wonder whether an LLM could also play the role of a human agent. To address this question, we prompt the LLM to act both as a user and an agent, using the same input and output as human agents in LAPS (cf.§ 4.3.4). We report on the results obtained from GPT-3.5 `turbo` and GPT-4-1106 `preview` with a temperature parameter of 1.0.

Table 4.1: Overview of the selected baseline datasets and ours. EXP and CW denote an expert and a crowd worker, respectively.

Dataset	**Collection Method**	**#Dial**	**Tasks**	**Domains**	**Scalability**	**Actual User Preferences**	**Preference Elicitation**	**Preference Extraction**	**Multi-Session**
SGD	Semi-synthetic	16,142	Booking, rec., etc.	20 domains, inc. restaurants	✓	×	×	×	×
M2M	Semi-synthetic	3,008	Booking	Restaurants, movies	✓	×	×	×	×
PersonaChatGen	Fully-synthetic	1,649	Personal chit-chat	Open domain	✓	×	×	×	×
Taskmaster-1	Self-dialogue	7,708†	Booking, ordering	6 domains, inc. restaurants	✓	×	×	×	×
MultiWOZ	Human-human (CW)	8,438	Booking, rec., etc.	7 domains, inc. restaurants	✓	×	×	×	×
CCPE-M	Human-human (EXP)	502	Rec.	Movies	×	✓	✓	△	×
MG-ShopDial	Human-human (EXP)	64	Rec., QA, etc.	E-commerce	×	✓	✓	×	×
LAPS	LAPS	1,406	Rec.	Recipes, movies	✓	✓	✓	✓	✓

† Includes only self-dialogues.
△ Annotations are provided but no entity-relation pair extraction with like/dislike distinction.

These dialogues are 3 sessions long.

Dialogue Diversity. We use three metrics to measure the lexical diversity of the collected conversations: (i) Dist-n [Li et al., 2016], (ii) Ent-n [Zhang et al., 2018b], and (iii) Self-BLEU [Zhu et al., 2018]. ***Dist-n*** measures the response diversity by computing the ratio of distinct n-grams to all n-grams in the given collection. Following Li et al. [2016], we use Dist-1 and -2 to measure the lexical diversity of conversation utterances. ***Ent-n*** aims to enhance the Dist-n measure by incorporating the frequency differences of n-grams into consideration, leveraging the entropy of the n-gram distribution. We report on Ent-4, following Zhang et al. [2018b]. ***Self-BLEU*** considers one utterance as a hypothesis and the rest of the utterances in the collection as references and calculates the BLEU score for each utterance. Following the NLTK's default setting [Bird et al., 2009] and Zhu et al. [2018], we compute the BLEU scores for $n = 1, 2, 3, 4$ for each hypothesis-reference pair and take the mean over all computed BLEU scores.

Normalization. A frequently overlooked pitfall of diversity metrics is their dependency on the total number of words in the dialogues. For instance, Dist-n scores are typically higher for datasets with fewer total number of words, as they are less likely to have repeated words. To address this, we set a word cutoff for all datasets and randomly sample dialogues until the total word count reaches this cutoff. The cutoff is set at 7,012 words, which is the minimum number of total words for user/system utterances across all datasets. We perform 100 random sampling per dataset and report the average scores, as well as two-tailed independent t-test results. Additionally, since M2M [Shah et al., 2018] dataset is lowercased, we lowercase all other datasets for a fair comparison.

Dialogue Quality. Dialogue quality evaluation is inherently challenging due to its subjective nature. Human evaluation, often considered the gold standard, is known to be highly sensitive to task design and instructions. Even with much care, it still suffers from differing bias and high variance per annotator, especially in crowdsourced environments [Deriu et al., 2021, Smith et al., 2022, Finch et al., 2023, Li et al., 2019, Zhang et al., 2023]. Automatic evaluation, while capable of mitigating the aforementioned issues, is also known to have its own limitations, including a bias towards machine-generated responses [Liu

et al., 2023b]. Aware of these limitations, we opt for automatic evaluation for two reasons: (1) our aim is to ensure our dialogue quality aligns with other high-quality datasets, rather than attaining state of the art, and (2) we use human workers to compose responses in their own words, which is less likely to be overestimated by metrics biased towards machine-generated responses.

After examining four reference-free automatic evaluation methods [Mehri and Eskenazi, 2020, Lin and Chen, 2023, Liu et al., 2023b, Zhong et al., 2022], we select **UniEval** [Zhong et al., 2022] as our automatic evaluation metric considering availability, cost, and performance. For evaluation aspects, we choose naturalness, understandability, and coherence from Zhong et al. [2022]. The aspects that require the conditioning fact as an input are not used, as the factuality of user preferences falls outside the scope of our study. We use the official implementation of UniEval[2] and its default settings. To ensure that the evaluation is computationally feasible using our available computational resources, we randomly select 100 dialogues (consisting of 1.8K responses on average) from each dataset. For significance testing, we use two-tailed independent t-tests ($p < 0.05$).

4.3.7 Experimental Setup

Domains. Our domains are *recipe* and *movie*. The recipe domain involves planning for the next dinner (session 1), breakfast (session 2), and lunch (session 3). The movie domain involves planning to watch a movie with family, friends, or alone (session 1), exploring another movie by the same director or actress/actor as the previous recommendation (session 2), and watching with different people or a different occasion (session 3). For each domain, we curated a list of categories that are relevant to the task and categorized them into categories of *must-have*, *should-have*, and *could-have*. These categories are detailed in our online repository.

Participants and Quality Control. We recruited Prolific[3] workers from English-speaking countries, having $\geq$ 98% approval rate and $\geq$ 1000 previous submissions. For the movie domain, we only invited

[2] `https://github.com/maszhongming/UniEval`

[3] `https://www.prolific.com/`

workers that accurately performed that task for the recipe domain. Throughout the experiments, we actively communicated with workers, answering over 250 questions and incorporating their feedback to clarify instructions. £14 was paid for completing three sessions, which took ~65 minutes. Workers were also allowed to terminate the task at an earlier session and receive partial payment. This allowed us to collect high-quality multi-session dialogues, as workers completing all sessions tend to be more engaged in the task. After crowdsourcing, we manually reviewed the dialogues to ensure data quality, making corrections or deletions as necessary. Common errors (aside from malicious behavior of not providing meaningful responses) include failing to include URLs in recommendations and misunderstanding their current role in the conversation.

Chat Interface. We collect conversations using TaskMAD [Speggiorin et al., 2022] and further develop it to support new features for our task. The human agent interacts with two chat interfaces for system and user roles, and each interface consists of a text box for composing responses and an instruction box to clarify role responsibilities.

4.4 Personal Recommendation Method

The end goal for large-scale preference elicitation conversational datasets is to enable personal conversational search and recommendation. Based on LAPS, we collect a large-scale dataset for the movie and recipe domains and use it to train a model for extracting personal preferences from conversations. The preferences are then used to generate personalized recommendations.

4.4.1 Preference Extraction

In this task, we aim to automatically extract user preferences from a dialogue session (cf. Eq. 4.2). We cast this task as a seq-to-seq QA [Chung et al., 2022] and fine-tune an LLM with instruction prompts eliciting user preference regarding category and conversation session. Formally, our method decomposes Eq. 4.2 as $\mathcal{E}(s, \mathcal{C}) = \bigcup_{c \in \mathcal{C}} \mathcal{E}'(s, c)$, where $\mathcal{E}'$ is the preference extraction function for individual category c.

We use FlanT5 [Chung et al., 2022] as the base model and use instruct prompts to read the given session and answer questions about the user's preferences, such as *"What cuisine does the user like?"* For the recipe domain, we use FlanT5 Small (80M parameters), Base (250M), and Large (780M) models and fine-tune them on our recipe dataset. For the movie domain, we explore a domain adaptation approach, where the initially fine-tuned model on the recipe domain (FlanT5-Large) is further fine-tuned on the movie domain.

4.4.2 Personalized Recommendation

With the personalized recommendation task, we aim to generate a recommendation response based on user's personal preferences. Personal preferences are stated in the conversation history h and all previously completed sessions S_k. By utilizing the preference extraction method (cf. § 4.4.1), we construct the preference memory $\mathcal{M}_k$ based on session history S_k and use it for recommendation. Formally, recommendation utterance u^{pred} is generated by the recommendation generation function R, defined as $R(h, \mathcal{M}_k) = u^{pred}$. Recommendation responses are generated by an LLM (LLaMA-2-7B [Touvron et al., 2023]) using zero-shot prompting. The prompt contains instructions to generate a personalized recommendation appended with the preference memory and history of the current conversation.

4.4.3 Evaluation

Recommendation Baseline. For our personalized recommendation method, we consider a baseline method, where the preference memory $\mathcal{M}_k$ is replaced with raw utterances from session history S_K. This baseline allows us to measure the effectiveness of using semi-structured fine-grained preferences over raw conversational utterances.

Recommendation Human Evaluation. Two human experts are instructed to evaluate the *rationale* and *relevance* aspects of recommendations: "Which assistant's rationale is more in line with the user's preferences?", and "Which assistant's recommendation item meets user preferences?" Following Li et al. [2019], we conduct pairwise comparisons between responses from our method and a baseline.

For each aspect, the annotators choose win, lose, or tie options, and disagreements are resolved through discussion. For evaluation, we randomly select 50 and 10 samples from the recipe and movie domains, respectively.

Recommendation Automatic Evaluation. Automatic machine translation measures such as ROUGE and BLEU assume significant overlap between ground truth and valid responses. This strong assumption does not hold for dialogue systems and in particular for personal recommendations, where valid responses represent high level of diversity. The lack of correlation between human evaluation is reported in previous study [Liu et al., 2016] and has been observed in our study as well. Addressing this challenge, we introduce an automatic reference-based evaluation metric, **Preference Utilization (PU)**, which aims to measure the utilization of user preferences in the recommendation response. Let $\mathcal{P} = \{\wp_1, \wp_2, \ldots\}$ denote the set of preferences (without their corresponding categories as defined in Eq. 4.1). The subsets $\mathcal{P}^{pred} \subseteq \mathcal{P}$ and $\mathcal{P}^{ref} \subseteq \mathcal{P}$ denote preferences $\wp_i \in \mathcal{P}$ that appear in predicted and reference recommendation responses, respectively. Preference utilization precision P_{PU} and recall R_{PU} are computed as:

$$P_{PU} = \frac{|\mathcal{P}^{pred} \cap \mathcal{P}^{ref}|}{|\mathcal{P}^{pred}|}, \quad R_{PU} = \frac{|\mathcal{P}^{pred} \cap \mathcal{P}^{ref}|}{|\mathcal{P}^{ref}|}.$$

In our experiments, we perform exact string matching to generate $\mathcal{P}^{pred}$ and $\mathcal{P}^{ref}$. When reference utterance u^{ref} includes URLs, we extract the content from the corresponding web page and use it in string matching. For our ground truth, we only use the assistant responses about recommendations which are accepted by the user.

Preference Extraction Evaluation. We evaluate preference extraction using exact match and BERTScore [Zhang et al., 2020]. Exact match assesses the case-insensitive string match of preferences between predictions and ground truth, while BERTScore accounts for different expressions of identical preferences. Here, preferences within each category are flattened into a single string with commas as delimiters and used for computation of BERTScore.

Table 4.2: Statistics of LAPS dataset.

Domain	Split	#Dialogue Sets			#Pref	#Utt	#Dial
		Single-Session	Two-Session	Three-Session			
Recipe	Train	163	24	160	5,538	9,342	691
	Val	24	5	21	772	1,333	97
	Test	48	10	41	1,600	2,610	191
	Total	**235**	**39**	**222**	**7,910**	**13,285**	**979**
Movie	Train	46	14	72	2,225	3,974	290
	Val	5	1	13	351	642	46
	Test	11	4	24	729	1,220	91
	Total	**62**	**19**	**109**	**3,305**	**5,836**	**427**

4.4.4 Experimental Setup

For preference extraction, we fine-tuned Flan-T5 models on our training set using the HuggingFace Transformers library [Wolf et al., 2020]. For both domains, the batch size is set to 8, the learning rate to 5×10^{-5}, and the AdamW optimizer is used. Training is conducted for 10 epochs, with the best checkpoint selected based on the validation set. We ran all our experiments on a single GPU (NVIDIA A100 40GB). For a personalized recommendation, we use Llama-2-7B [Touvron et al., 2023] due to its ability to handle a sufficiently long context, while being capable of running on a single GPU. We run the model (from the HuggingFace model hub) 10 times for each dialogue and report the average score.

4.5 Results

This section presents the results of the proposed methods for dialogue collection (§4.5.1) and personal recommendation (§4.5.2)

4.5.1 Dialogue Collection Evaluation

Using LAPS, we collect a large-scale multi-session personalized dataset, as shown in Table 4.2. The number of preferences is the total number of (s, p, o) triples, where s is the user, p is the preference category, and

Table 4.3: Lexical diversity scores. Significance against all baselines is marked by $^{+}$.

Dataset	Dist-1/2	Ent-4	Self-BLEU$^{\blacktriangledown}$
SGD	0.179 / 0.538	8.311	0.964
M2M	0.057 / 0.290	7.922	0.955
PersonaChatGen	0.165 / 0.523	8.261	0.970
Taskmaster-1	0.207 / 0.644	8.384	<u>0.949</u>
MultiWOZ	0.158 / 0.505	8.345	0.966
CCPE-M	0.175 / 0.571	8.414	0.961
MG-ShopDial	**0.234** / <u>0.653</u>	8.199	**0.935**
LAPS-Recipe	0.207 / 0.650	<u>8.563</u>$^{+}$	0.955
LAPS-Movie	<u>0.222</u>$^{+}$ / **0.666**$^{+}$	**8.593**$^{+}$	0.954

$^{\blacktriangledown}$ Lower is better.

o is the preference attribute. Using human verification, we identified 4.5% error rate in preferences extracted by GPT-4, highlighting the need for human involvement for accurate preference extraction.

Lexical Diversity. Table 4.3 shows the results of lexical diversity evaluation, indicating that LAPS-Movie and -Recipe achieve the highest diversity scores with respect to Dist-2 and Ent-4. Considering all metrics, LAPS-Movie is on par with MG-ShopDial, a dataset of human-human dialogues involving trained volunteers as an assistant role. These results demonstrate that LAPS collects lexically diverse dialogues as effectively as human-human dialogue collection methods.

Comparing the lexical diversity of user and system utterances in Figure 4.3, we observe that our method achieves high lexical diversity for both user and assistant utterances. This suggests that LLM's guidance can help workers compose diverse responses. Notably, we observe that assistant utterances in CCPE-M exhibit less diversity than those of the user. This could be attributed to the small number of participants acting as assistants in CCPE-M and their often short and direct responses. This highlights that achieving diversity is non-trivial, even with trained experts. LAPS's success in achieving diversity further demonstrates the effectiveness of our method.

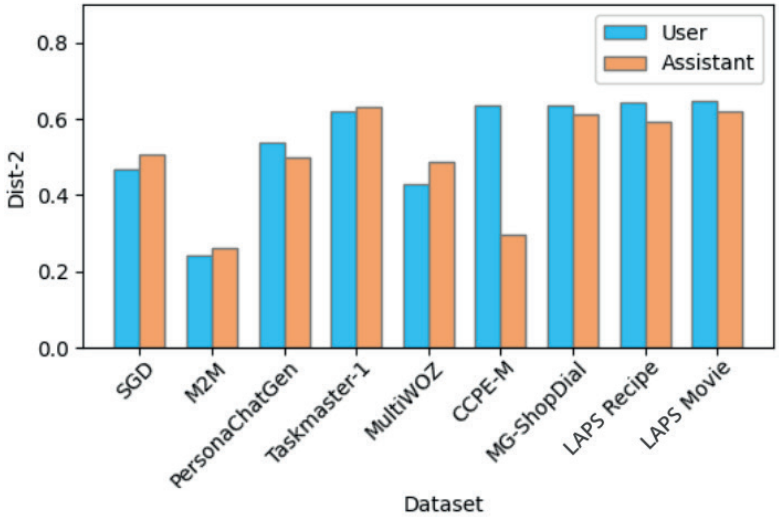


Figure 4.3: Lexical Diversity of user and assistant utterances.

Table 4.4: Lexical diversity scores of synthetic dialogue generation. Significance against all baselines is marked by $^{+}$.

Domain	Method	Dist-1/2	Ent-4	Self-BLEU$^{\blacktriangledown}$
Recipe	Synthetic GPT-3.5	0.127 / 0.430	8.308	0.981
	Synthetic GPT-4	0.183 / 0.597	**8.601**	0.976
	LAPS	**0.207**$^{+}$ / **0.65**$^{+}$	8.563	**0.955**$^{+}$
Movie	Synthetic GPT-3.5	0.138 / 0.444	8.331	0.977
	Synthetic GPT-4	0.178 / 0.559	8.481	0.979
	LAPS	**0.222**$^{+}$ / **0.666**$^{+}$	**8.593**$^{+}$	**0.954**$^{+}$

$^{\blacktriangledown}$ Lower is better.

Table 4.4 compares LAPS- and LLM-generated conversations. The results show that synthetic dialogues are less diverse than LAPS, even with GPT-4 temperature tuning. This suggests potential diversity pitfalls in synthetic personalized dialogue generation using LLMs. One way to mitigate this issue is using a synthetic persona, as in PersonaChatGen [Lee et al., 2022]. However, as Table 4.3 shows, it still falls short of human-involved LAPS, suggesting the diverse nature of human preferences.

Table 4.5: UniEval scores. Significance against all baselines is marked by $^{+}$.

Dataset	NAT	UND	COH	Avg.
SGD	0.794	0.781	0.758	0.778
M2M	0.634	0.616	0.701	0.650
Taskmaster-1	0.792	0.779	0.782	0.784
MultiWOZ	0.870	0.860	0.848	0.859
CCPE-M	0.716	0.708	0.689	0.704
MG-ShopDial	0.743	0.730	0.687	0.720
LAPS-Recipe	0.867	0.860	0.891^{+}	0.872^{+}
LAPS-Movie	**0.874**	**0.868**	$\mathbf{0.897}^{+}$	$\mathbf{0.880}^{+}$
PersonaChatGen†	0.894	0.887	0.738	0.839

† Scores provided as a reference, but do not represent fair comparison.

Dialogue Quality. Table 4.5 shows the results of the dialogue quality evaluation. For coherence, LAPS outperforms the other datasets. For naturalness and understandability, PersonaChatGen, which is an LLM-based fully-synthetic dataset, outperforms the other datasets. This is consistent with the findings from Liu et al. [2023b], which shows that LLM-based synthetic dialogues tend to have higher scores for language-model-based automatic evaluation metrics. Excluding fully-synthetic datasets, LAPS's performance is among the best, demonstrating that our method collects high-quality dialogues as effectively as human-human dialogue collection methods. On average, LAPS outperforms the other datasets, highlighting the effectiveness of our method.

Discussion. Based on these results we can answer our first and second research questions: ***RQ3a:*** *Using LAPS, we collect large-scale multi-session human-written conversations that contain actual user preferences.* and ***RQ3b:*** *LAPS-collected dialogues show high diversity and quality, on par with expert-involved human-human dialogues, as highlighted in Figure 4.4.*

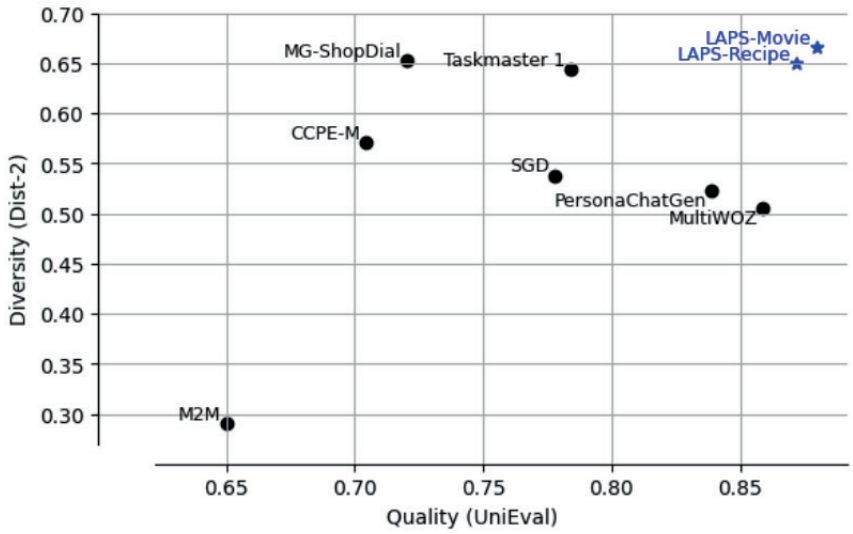


Figure 4.4: LAPS collects diverse and high-quality dialogues compared to other dialogue collection methods.

Table 4.6: Preference extraction results. Domain adapt. represents the model is first fine-tuned on the recipe domain and then further fine-tuned on the movie domain.

Domain	Model Size	Domain Adpt.	Exact Match			BERTScore		
			P	R	F	P	R	F
Recipe	Small	–	0.454	0.408	0.412	0.590	0.589	0.589
	Base		0.464	0.476	0.454	0.622	0.623	0.622
	Large		**0.532**	**0.493**	**0.494**	**0.651**	**0.650**	**0.650**
Movie	Large	×	0.453	0.415	0.425	0.598	0.592	0.595
		✓	**0.470**	**0.424**	**0.432**	**0.618**	**0.614**	**0.616**

4.5.2 Personal Recommendation Evaluation

Preference Extraction. Table 4.6 shows the preference extraction performance of different FlanT5 pre-trained model sizes for the recipe topic. The results show that the performance for the Recipe domain improves as the model size increases. Using the recipe domain as the source domain, we further fine-tune the FlanT5-Large fine-tuned model on the target movie domain. The results show that domain adapta-

Table 4.7: Human evaluation results of recommendations.

Domain	Prompting Method	Rationale			Relevance		
		Win	Lose	Tie	Win	Lose	Tie
Recipe	Standard	17	29	4	18	22	10
	Memory	**29**	**17**	4	**22**	**18**	10
Movie	Standard	3	6	1	**4**	**3**	3
	Memory	**6**	**3**	1	3	4	3

Table 4.8: Recommendation results. Significance against Standard is marked by $^{+}$.

Domain	Prompting Method	#Prompt Tokens	Preference Utilization		
			P_{PU}	R_{PU}	F_{PU}
Recipe	Standard	880	**0.554**	0.311	0.398
	Memory	**308**	0.470	**0.411**$^{+}$	**0.438**$^{+}$
Movie	Standard	957	**0.508**	0.364	**0.424**
	Memory	**311**	0.443	**0.397**$^{+}$	0.419

tion consistently outperforms direct fine-tuning. This demonstrates the adaptability of our dataset to other domains through domain adaptation.

Recommendation Human Evaluation. Table 4.7 presents the human evaluation results for recommendation quality. In the recipe domain, using preference memory outperforms the baseline, for both recommendation quality and rationale. For the movie domain, the Memory method excels in rationale but not in recommendation quality. Given that improvements are consistently found in the rationale aspect, using preference memory is effective in improving the rationale for recommendations, a critical factor for transparent and explainable recommendations.

Recommendation Automatic Evaluation. To evaluate the effectiveness of our preference-based prompting, we compare it with the baseline standard prompting method. The downstream recommendation task results are depicted in Table 4.8. The Preference Utilization results show a similar trend to the human evaluation results, demonstrating the overall win of Memory over the Standard method. In the

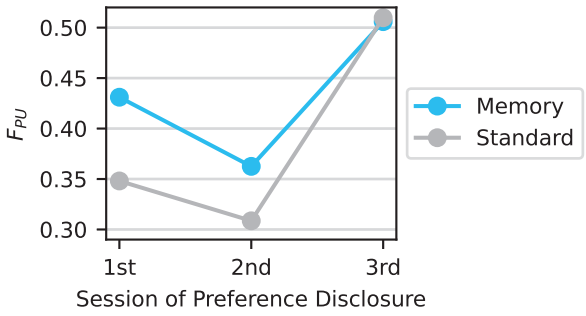


Figure 4.5: Breakdown of Preference Utilization (F_{PU}, recipe domain).

recipe domain, Memory outperforms Standard in F_{PU} score (0.438 vs. 0.398), whereas in the movie domain, their scores show no statistical significance (0.419 vs. 0.424). Error analysis indicates that the preference memory in the movie domain is sometimes insufficient for making recommendations, suggesting a need for improving the preference extraction method. Nevertheless, preference-based prompting for movies reduces the prompt length threefold while maintaining a comparable F_{PU} score. This threefold reduction has significant implications for real-world applications where inference cost is a critical factor.

Preference Utilization by Session. Figure 4.5 shows the Preference Utilization scores by session in recipe recommendations. The x-axis represents the session where preferences are disclosed, as detailed in Section 4.4.3. The analysis focuses on the recommendations in the third session to examine Preference Utilization across different sessions. The graph shows methods' struggle with utilizing preferences from earlier sessions (1st and 2nd) than those from the ongoing (3rd) session. A similar phenomenon is reported by Liu et al. [2023a] in a retrieval augmentation setting, referred to as *LLMs' recall issue in long prompt inputs*. The graph also demonstrates that preference-based prompting more effectively utilizes earlier session preferences than standard prompting, suggesting that preference memory can mitigate long prompt recall issues. We observe that this pattern (recall issues in earlier sessions and the effectiveness of preference-based prompting in addressing them) is generally consistent across the movie domain and different sessions, though not always statistically significant. Overall, our analysis shows the effectiveness of preference memory in long

prompts.

Discussion. Based on these results we can positively answer our last research questions: ***RQ3c:*** *Preference memory enhances effective utilization of user preferences in recommendations, improves the rationale of recommendations, and mitigates long prompt recall issues.*

4.6 Conclusion

In this research, we proposed a method to collect large-scale multi-session personalized conversations reflecting actual user preferences. Our method, LAPS, employs LLMs to generate personalized guidance for human workers, reducing the cognitive load for a highly complex task. Extensive experiments demonstrate, while being a scalable and high-quality data collection method, LAPS collects utterances as diverse as the expert-involved methods. We further showed that utilizing extracted user preferences results in more effective personal recommendations compared to using raw user utterances of previous sessions. In our experiment, fully-synthetic LLM-based methods do not yield diverse conversations. Using actual user preferences from LAPS as personas is a promising avenue to explore for future.

Chapter 5

Conversation Evaluation

The previous chapters established methods for personal information extraction (Chapters 2 and 3) and personalized response generation through large-scale personalized dataset construction (Chapter 4), enabling personalized conversational information access (CIA) system development. However, systematic and reliable evaluation of these systems remains an open challenge. Existing automatic evaluation methods are often insufficient for assessing the dynamic nature of conversations. This necessitates an automatic, reference-free method that provides fine-grained performance insights. Therefore, this chapter addresses the research question:

> **RQ4:** *How to automatically evaluate personalized CIA systems in a way that accounts for natural conversation flow and provides fine-grained insights into system performance?*

To answer this question, this chapter introduces **FACE**: a **F**ine-grained, **A**spect-based **C**onversation **E**valuation method. FACE is reference-free and provides evaluation scores for diverse turn and dialogue-level qualities. To enable meta-evaluation, this chapter also introduces the **CRSArena-Eval**, a dataset with high-quality multi-aspect human judgments on human-system conversations. The findings show that FACE achieves a strong correlation with these human judgments, outperforming state-of-the-art methods. Unlike existing large language model (LLM)-based methods that provide single uninterpretable scores, FACE provides insights into system performance, enabling the identification and location of problems within conversations.

5.1 Introduction

Recent advances in generative LLMs have reshaped information access systems, enabling rich human-system *conversational* interactions. In parallel, offline evaluation of information access systems is also evolving with the increasing adoption of LLM–based approaches [Trippas et al., 2025]. In industrial settings, it is common practice to design evaluation prompts for specific aspects of a system, to reduce the high cost of human evaluation, and to help early identification of system issues during the development phase [Dey and Desarkar, 2023, Dubois et al., 2024]. The research community, while maintaining skepticism, has valued the potential of LLM-based evaluation, and various LLM-based evaluation methods have been proposed for information access systems [Dietz et al., 2025, Balog et al., 2025, Dietz, 2024, Upadhyay et al., 2024b, Farzi and Dietz, 2024, Thakur et al., 2025, Aliannejadi et al., 2025, Upadhyay et al., 2024a].

Research gap. Most research on LLM-based evaluation for information access systems revolves around assessing the relevance of retrieved passages [Upadhyay et al., 2024a,b, Dietz, 2024] or comparing generated texts against gold nuggets [Pradeep et al., 2025, Thakur et al., 2025]. Similar efforts have been made for the offline evaluation of CIA, where gold nuggets are generated for predefined trajectories of conversations [Abbasiantaeb et al., 2025, Aliannejadi et al., 2025, 2024]. However, evaluation of conversational systems requires going beyond measuring factual accuracy over constrained conversational trajectories. Sakai [2023b] identified twenty evaluation criteria (or aspects) beyond factual correctness for conversational systems (e.g., modesty and conciseness), along with a set of requirements that a conversational evaluation method should satisfy. To date, this broader perspective has not been fully realized. As highlighted in the latest IR strategic meeting (SWIRL), developing "reliable and informative methods for evaluating dynamic trajectories of conversations at scale" is still an open question [Trippas et al., 2025].

Research goal. The central research question that arises here is:

> **RQ4:** *How to automatically evaluate personalized CIA systems in a way that accounts for natural conversation flow and provides fine-grained insights into system performance?*

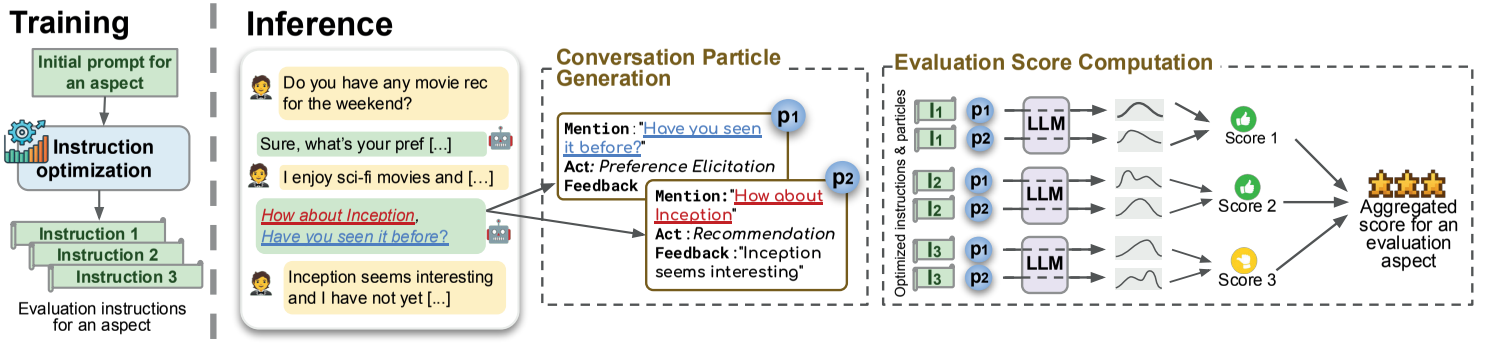


Figure 5.1: Illustration of FACE for a turn-level aspect. Instruction optimization generates a set of diverse evaluation instructions for the given aspect (e.g., relevance) based on an initial prompt. For evaluation, each conversation is decomposed into particles containing a dialogue act, a mention (text span from a system turn), and user feedback from the following turn. A response distribution is created for each instruction-particle pair and the weighted summation of the scores is computed. The final score is obtained by aggregating the scores across all instructions and particles. For turn-level aspects, aggregation is performed per turn, while for dialogue-level aspects, scores of particles across the entire dialogue are aggregated.

Specifically, our focus is on "*beyond factual accuracy*" as it is an essential yet underexplored dimension of CIA evaluation. Developing such an evaluation method poses several challenges. Reference-based evaluation methods restrict the evaluation process to predefined conversational trajectories. Even recent simulation-based approaches, such as those used in TREC iKAT 2025, remain constrained by predefined rubric questions [Aliannejadi et al., 2025]. Reference-free conversational evaluation methods, on the other hand, mostly rely on LLM-based prompting and produce a single, uninterpretable score [Mehri and Eskenazi, 2020, Zhong et al., 2022, Liu et al., 2023b], making it difficult to trace the score back to the underlying contributing factors. This violates the *Specificity* requirement of conversational evaluation methods, which requires evaluation methods to "be specific when locating problem(s) within conversations" [Sakai, 2023b].

Method. We propose FACE, a Fine-grained Aspect-based Conversation Evaluation method, which is a reference-free approach for evaluating conversations across multiple evaluation aspects. We make the following departures from existing LLM-based conversation evaluation methods. First, we mimic the human evaluation process by modeling the diverse thought processes of multiple human assessors through a pool of optimized evaluation instructions. These optimized instructions, obtained using beam search and bandit optimization, reduce the arbitrary nature of prompt engineering and make the evaluation process robust. Second, we introduce the concept of "*conversation particle*," which encompasses a composite nugget design consisting of an atomic statement, a dialogue act, and user feedback. These conversation particles enable evaluating an atomic statement, while providing a broader view of the overall conversational state. By evaluating these conversation particles using a set of optimized instructions, our method achieves specificity. Additionally, by aggregating these scores at the turn or dialogue level, FACE supports both turn-level and dialogue-level evaluation aspects. An overall overview of FACE is illustrated in Figure 5.1.

The introduction of FACE leads us to the following sub-research questions:

- **RQ4a:** *How does FACE correlate with human judgments?*

- **RQ4b:** *How generalizable are FACE-optimized instructions across different LLMs and domains?*

- **RQ4c:** *Can fine-grained evaluation scores of FACE provide insights about system's issues?*

Meta-Evaluation. To evaluate FACE, we collect 20,962 human annotations for 467 human-system conversations of the CRSArena-Dial dataset [Bernard et al., 2025]. These conversations are obtained from interaction of real users with nine diverse conversational recommender systems. Human annotation of these conversations cover evaluation scores for seven evaluation aspects that are shown to be predictive of response quality and user satisfaction [Siro et al., 2023, 2022]. These include two turn-level aspects: *relevance* and *interestingness*, and five dialogue-level aspects: *understanding*, *task completion*, *conversation efficiency*, *interest arousal*, and *overall impression*. FACE is then applied to these evaluation aspects and the obtained evaluation scores are compared to human evaluation scores.

Experiments. Our experiments demonstrate that FACE outperforms state-of-the-art evaluation methods by a large margin, achieving system and turn/dialogue-level Spearman of 0.9 and 0.5, respectively, without observing any system-human conversations for instruction optimization. We demonstrate that FACE provides a reusable instruction pool that can be robustly used across different LLMs and conversation types. Remarkably, FACE outperforms strong baselines that utilize GPT-4 as their backbone on two chit- chat datasets, despite using an 8B-parameter Llama 3.1 as its backbone. Lastly, and importantly, we present a case study of two competitive conversational systems and demonstrate how FACE scores can diagnose issues and pinpoint problems within a conversation.

Key contributions of this chapter are summarized as follows:

- We propose FACE, a conversation evaluation method that evaluates dynamic trajectories of conversations on diverse evaluation aspects and significantly outperforms state-of-the-art evaluation methods (Sec. 5.6.1).

- We introduce the notion of conversation particles and demonstrate their effectiveness in reducing bias (Sec. 5.7) and improving evaluation sample efficiency (Sec. 5.7).

- We show that utilizing a pool of optimized instructions mitigates the arbitrary nature of prompt engineering for automatic evaluation and is robust across different LLMs and conversation types (Sec. 5.6.2).

- We demonstrate that FACE scores are beyond a single uninterpretable score and can be used by humans to identify and locate potential system issues (Sec. 5.6.3).

- We develop the CRSArena-Eval meta-evaluation dataset, containing 20,962 human annotations over 467 conversations, for evaluating conversational evaluators. The dataset and a meta-evaluation interface for evaluating an evaluator against CRSArena-Eval are available in our repository.

5.2 Related Work

Automatic Conversation Evaluation. Although human annotations are the gold standard for evaluating CIA systems, they are expensive and time-consuming. Therefore, automatic evaluation methods have been proposed to scale up the evaluation process. There are two main types of automatic evaluation methods: *reference-based* and *reference-free* [Soudani et al., 2024b].

Reference-based methods use gold references to evaluate system responses, which include Recall@K, BLEU [Papineni et al., 2002], ROUGE [Lin, 2004], and BERTScore [Zhang et al., 2020]. While these methods are effective for machine translation and question-answering, they have limitations in conversation evaluation, as they overlook diverse response possibilities and various evaluation aspects. These limitations are supported by various studies showing a weak correlation between reference-based methods and human evaluations [Mehri and Eskenazi, 2020, Bernard et al., 2025, Liu et al., 2016, Bernard and Balog, 2025].

Reference-free methods have been proposed to address these limitations [Mehri and Eskenazi, 2020, Zhong et al., 2022, Liu et al., 2023b]. These methods evaluate system responses without relying on gold references, consider multiple aspects of system quality, and allow for the assessment of various response possibilities. However, these methods primarily focus on turn-level evaluation, limiting their ability to assess the whole conversation, which are crucial for evaluating system performance in real-world scenarios [Siro et al., 2022, 2023]. FACE addresses these limitations by evaluating the system based on the whole conversation and capturing multiple conversation trajectories from diverse user-system interactions.

LLMs for Evaluation. Recent advancements in LLMs have led to various automatic evaluation methods, ranging from simple prompting to user simulation, that show strong performance [Upadhyay et al., 2024b, Dubois et al., 2024, Liu et al., 2023b, Lin and Chen, 2023, Zheng et al., 2023, Dey and Desarkar, 2023]. For instance, G-Eval [Liu et al., 2023b] uses an LLM to generate an evaluation prompt, which is then used to assess the system's response, demonstrating a strong correlation with human evaluations in chit-chat conversations. However, improved performance of LLM-based methods comes with challenges. *Evaluation bias* is one of them; LLMs tend to favor longer responses (length bias) [Wang et al., 2023c, Dubois et al., 2024] and are biased toward texts from similar models (self-bias) [Xu et al., 2024, Liu et al., 2023b, Balog et al., 2025]. *Generalizability* is another challenge: LLMs are sensitive to handcrafted, arbitrary prompts, which are not reusable for different models [Sclar et al., 2024, Razavi et al., 2025]. FACE mitigates these biases with conversation particles and instruction optimization (see Sec. 5.7).

Instruction optimization. To address the arbitrary nature of handcrafted prompts, various instruction optimization (or prompt optimization) methods have been proposed [Chen et al., 2024a, Kong et al., 2024, Pryzant et al., 2023, Ye et al., 2024, Chen et al., 2024b, Fernando et al., 2024, Zhou et al., 2023b, Yang et al., 2024, Yuksekgonul et al., 2024]. Zhou et al. [2023b] introduced APE, an automatic prompt engineering method that uses LLMs to generate prompt candidates, and perform a Monte Carlo search to find the optimal prompt. Kong et al. [2024] proposed PRewrite, where prompts are optimized using proximal pol-

icy optimization (PPO) [Schulman et al., 2017]. Pryzant et al. [2023] proposed an instruction optimization method using "textual gradient," which provides natural language feedback for an LLM to optimize prompts. These studies focus on common tasks like question answering (QA) and classification, leaving conversation evaluation unexplored. More importantly, optimized prompts are not generalizable between LLMs [Zhou et al., 2023b], which is crucial for evaluation tasks, where reusability is essential. In this chapter, we present a method for applying a textual gradient approach to enhance conversation evaluation alongside effective strategies for transferring optimized prompts across different settings.

Nugget-based Evaluation. To assign granular evaluation scores, multiple nugget-based evaluation approaches are proposed [Voorhees, 2003, Mayfield et al., 2024, Pradeep et al., 2025, Lin and Demner-Fushman, 2005, Ekstrand-Abueg et al., 2013, Takehi et al., 2023, Rajput et al., 2011, Dietz, 2024, Sakai, 2023a]. Nugget-based evaluation was proposed [Voorhees, 2003] to assign granular scores to system responses for non-binary queries, wherein a nugget, which is an atomic piece of information, serves as the unit of evaluation, enabling a more traceable assessment. Although the original nugget-based evaluation was intended as a manual method, many efforts have aimed to automate or semi-automate it [Mayfield et al., 2024, Pradeep et al., 2025, Lin and Demner-Fushman, 2005, Ekstrand-Abueg et al., 2013, Takehi et al., 2023, Rajput et al., 2011, Dietz, 2024, Abbasiantaeb et al., 2025]. One of the earlier methods is POURPRE [Lin and Demner-Fushman, 2005], an automatic nugget-based evaluation method that uses n-gram co-occurrences to assess nugget presence in system responses. More recent research employs LLMs for nugget matching; Pradeep et al. [2025] introduced the AutoNuggetizer framework in TREC RAG 2024, where LLMs automatically create nuggets and assign them to system responses. However, these works are reference-based and/or focus on individual responses, and more importantly, they do not target information-access conversations.

SWAN [Sakai, 2023b] proposes a conceptual framework for evaluating CIA systems by decomposing user-system interactions into units and evaluating them in various aspects. While SWAN is promising, its execution remains an open question, specifically how to automatically create and assess nuggets while ensuring the method's generalizability.

This work addresses these challenges.

5.3 Method

Our Fine-Grained Aspect-based Conversation Evaluation (FACE) approach handles the one-to-many nature of conversations and provides detailed scores for each turn and dialogue-level evaluation aspect. Figure 5.1 illustrates the FACE method. Using beam search and bandit algorithms, FACE first optimizes a set of instructions for a given evaluation aspect. During evaluation, a dialogue is decomposed into conversation particles, each containing a dialogue act (e.g., "Recommendation"), mention (e.g., "How about Inception"), and corresponding user feedback from the user response (e.g., "Inception seems interesting"). Each particle is independently evaluated with optimized instructions via an LLM, generating score distributions and resulting in turn/dialogue-level scores for a given aspect.

We note, without detailed elaboration, that FACE is applicable to a broad range of evaluation aspects [Sakai, 2023b] and conversation types (e.g., task-oriented and chit-chat dialogues). The primary focus of this chapter, however, is on conversational recommender systems (CRSs) and seven widely recognized turn- and dialogue-level evaluation aspects, following Siro et al. [2022]; see Section 5.4.2 for description of these aspects.

This section describes the evaluation steps of FACE: particle generation (Sec. 5.3.1) and evaluation score computation (Sec. 5.3.2), followed by the instruction optimization process (Sec. 5.3.3).

5.3.1 Conversation Particle Generation

FACE sets two goals: (1) enable reference-free evaluations to address state explosion in natural conversation evaluation, and (2) provide fine-grained scores at both turn- and dialogue-level to locate undesired system behavior within the conversation [Sakai, 2023b]. To achieve these, we introduce *conversation particle*, a self-contained information unit decomposed from conversations. Each particle is composed of three parts: (i) Dialogue `act` is the system's action associated with the particle, such as "recommendation" or "preference elicitation;" (ii) `Mention` denotes the particle text within the system's response, like

"How about the movie A?"; and (iii) `Feedback` is the user's evaluative reply, for instance, "The movie A seems interesting." Following Chapter 4, we use 5 dialogue acts: *greetings*, *preference elicitation*, *recommendation*, *goodbye*, and *others*.

We instruct an LLM to decompose system responses into particles, denoted as *decomposer* in the rest of the chapter. Let u_t be the target system response at turn t, h the dialogue history preceding u_t, and u_{t+1} the user's turn following u_t. The decomposer $\mathcal{D}$ maps (h, u_t, u_{t+1}) to conversation particles $\mathbf{P}_u$:

$$\mathbf{P}_u = \mathcal{D}(h, u_t, u_{t+1}),$$

where each particle $p \in \mathbf{P}_u$ is a triplet (`act`, `mention`, `feedback`). The full particle list for a dialogue, $\mathbf{P}_d$, is the union of particles from all responses with the dialogue, i.e., $\mathbf{P}_d = \bigcup_{u \in d} \mathbf{P}_u$.

Appendix B details prompts used for particle generation. It is shown that LLMs are more effective than traditional methods like dependency parsing and information extraction for decomposing texts into atomic units [Pradeep et al., 2024, Alaofi et al., 2024].

5.3.2 Evaluation Score Computation

FACE utilizes optimized instructions to generate scores for each conversation particle. Formally, given a particle p and the evaluation instruction I^α for the aspect α, an LLM generates a response r_p^α. To address known issues with LLM-generated scores, such as low variance and their noise [Liu et al., 2023b], we obtain a response distribution $\{r_{p,i}^\alpha\}_{i=1}^n$ and compute a weighted sum over the response set:

$$\mathcal{S}_{particle}(I^\alpha, p) = \sum_{i=1}^{n} r_{p,i}^\alpha P(r_{p,i}^\alpha | I^\alpha, p; \theta), \tag{5.1}$$

where $P(.)$ is a probability of $r_{p,i}^\alpha$ from an LLM parameterized by θ, and n is the number of sampled evaluation responses.

These particles scores are then aggregated per turn or conversation by taking their mean,

$$\mathcal{S}(I^\alpha, \mathbf{P}_x) = \frac{1}{|\mathbf{P}_x|} \sum_{p \in \mathbf{P}_x} \mathcal{S}_{particle}(I^\alpha, p),$$

where $\mathbf{P}_x$ is the set of particles for a given turn or dialogue, depending on evaluation aspect α; e.g., for *relevance*, aggregation is performed over particles of a turn, and for *task completion* aggregation is done for all particles of the dialogue.

A unique feature of FACE is utilizing diverse reasoning paths for each evaluation aspect, which is obtained by selecting optimized chain-of-thought (CoT) instructions. The intuition is that evaluation requires complex reasoning, and an optimal answer can be obtained by marginalizing various thought paths [Wang et al., 2023b]. Here, a set of top-performing optimized instructions with various CoT instructions, $\mathbf{I}^\alpha$, are applied to particles, and the resulting scores are aggregated to obtain the final score s^α:

$$s^\alpha = FACE(\mathbf{I}^\alpha, \mathbf{P}_x) = \frac{1}{|\mathbf{I}^\alpha|} \sum_{I^\alpha \in \mathbf{I}^\alpha} \mathcal{S}(I^\alpha, \mathbf{P}_x).$$

5.3.3 Instruction Optimization

Instruction optimizer generates diverse optimized CoT instructions for a given evaluation aspect. The goal is to obtain a representative set of thought processes (via CoT) for an evaluation aspect, and leverage them to evaluate unseen human-system conversations. The optimization process is performed on annotated human-human conversations to capture human thinking process and reasoning when assessing dialogue quality. We note, at the outset, that all the optimization and selection algorithms are performed independently for each aspect. For notational simplicity, we shall drop superscript α from aspect-related instruction and evaluation scores in this section.

To optimize instructions, we assume access to human evaluated dialogues, $\mathbf{H} = \{(x_i, l_i)\}_{i=1}^m$, where x_i is either a turn or an entire dialogue depending on the aspect, and l_i is its label. Similarly, we assume access to an LLM $\mathcal{L}_1$ that generates evaluation scores, $\mathbf{S_I} = \{(x_i, \text{FACE}(\mathbf{I}, \mathbf{P}_{x_i}))\}_{i=1}^m$, where $\mathbf{P_{x_i}}$ corresponds to particles of a specific turn or conversation and $\mathbf{P} = \bigcup_{x_i} \mathbf{P_{x_i}}$ is all particles in the dialogue collection.

The optimization objective is to identify a set of optimal instructions $\mathbf{I}^*$ that maximizes the correlation between human labels and the scores generated by the automatic evaluator:

$$\arg\max_{\mathbf{I}^*} \mathcal{C}(\mathbf{H}, \mathbf{S}_{\mathbf{I}^*}),$$

Algorithm 1 Instruction optimization of FACE

Require: Human evaluations $\mathbf{H}$, initial prompt I, iterations K, beam width b, candidate size b', gradient samples α

Initialize $\mathbf{I}^{\text{pool}} \leftarrow \emptyset$, $\mathbf{I}_1 \leftarrow \{I\}$
for $k = 1, ..., K$ **do**
  for each instruction $I_{k_j} \in \mathbf{I}_k$ **do**
    for each particle $p \in \mathbf{P}$ **do**
      // Get score
      $s_{p,k_j} \leftarrow \mathcal{S}_{particle}(I_{k_j}, p)$ ▷ Eq. 5.1
      // Textual Gradient Generation
      $\mathbf{G}_{p,k_j} \leftarrow \mathcal{G}_{\nabla}(I_{k_j}, s_{p,k_j}, l_p; \alpha)$ ▷ Eq. 5.2
      // Instruction Rewriting
      $\mathbf{I}'_{p,k_j} \leftarrow \mathcal{R}_{\delta}(I_{k_j}, \mathbf{G}_{p,k_j})$ ▷ Eq. 5.3
    end for
  end for
  $\mathbf{I}'_k = \bigcup_{p \in \mathbf{P}} \bigcup_j \mathbf{I}'_{p,k_j}$ ▷ Collect all rewrites
  // Instruction Selection
  $\mathbf{I}_k^{\text{cand}} = Select_{b'}^{UCB}(\mathbf{I}'_k)$ ▷ Eq. 5.4 and Algorithm 2
  $\mathbf{I}^{\text{pool}} \leftarrow \mathbf{I}_k^{\text{cand}} \cup \mathbf{I}^{\text{pool}}$ ▷ Update pool (Eq. 5.5)
  // Select top-b instructions
  $\mathbf{I}_{k+1} \leftarrow \arg\max_{\mathbf{I}_k \subseteq \mathbf{I}^{\text{pool}}, |\mathbf{I}_k| = b} \mathcal{C}(\mathbf{H}, \mathbf{S}_{\mathbf{I}_k})$
end for
$\mathbf{I}^* \leftarrow \arg\max_{\mathbf{I}^* \subseteq \mathbf{I}^{\text{pool}}} \mathcal{C}(\mathbf{H}', \mathbf{S}_{\mathbf{I}^*})$
return $\mathbf{I}^*$

where $\mathcal{C}(.)$ represents the correlation function.

The optimization process employs an LLM $\mathcal{L}_2$ to refine instructions based on the scores generated by the evaluator LLM $\mathcal{L}_1$. FACE employs a non-parametric optimization algorithm using textual gradients [Pryzant et al., 2023]. Here, natural language "gradients" (as opposed to numerical gradients) are generated to describe the shortcomings of instructions. The gradients are used to rewrite original instructions in the opposite semantic direction. The best instructions are iteratively selected using beam search and Upper Confidence Bound (UCB) bandits, based on correlations with human judgments. The process consists of three stages.

(1) Textual Gradient Generation. This is an iterative process, where a static prompt ∇ is used for generating textual gradients (lines

7-8 of Algorithm 1). At each iteration k, the prompt ∇ takes an evaluation instruction I_{k_j} from the current set of instructions $\mathbf{I}_k$, its prediction score $s_{p,k_j} = \mathcal{S}_{particle}(I_{k_j}, p)$, and the corresponding human label l_p for a given particle p. A set of textual gradients $\mathbf{G}_{p,k_j}$ is then generated by:

$$\mathbf{G}_{p,k_j} = \mathcal{G}_{\nabla}(I_{k_j}, s_{p,k_j}, l_p; \alpha), \tag{5.2}$$

where $\mathcal{G}_{\nabla}(.)$ is the gradient generation function, with parameter α denoting the number of gradients generated per instruction-score pair. Since human annotations are provided at the turn- or dialogue-level, l_p represents human annotation for the turn or dialogue that contains the particle p.

For the prompt ∇, we employ reasoning templates [Ye et al., 2024], which provide a set of items to be considered by $\mathcal{G}_{\nabla}$. Our items include identifying inconsistencies between the predicted and human annotations, evaluating the correctness of the current task and CoT instructions, and suggesting edits to these instructions, if necessary.

(2) Instruction Rewriting. This step updates each instruction using textual gradients (lines 9-10 of Algorithm 1). For each particle p at iteration k, we use the rewriting function $\mathcal{R}_{\delta}$ with the prompt δ, which takes the current instruction I_{k_j} and gradients $\mathbf{G}_{p,k_j}$ to obtain updated instructions $\mathbf{I}'_{p,k_j}$:

$$\mathbf{I}'_{p,k_j} = \mathcal{R}_{\delta}(I_{k_j}, \mathbf{G}_{p,k_j}). \tag{5.3}$$

An LLM $\mathcal{L}_3$ is used for the rewriting function $\mathcal{R}_{\delta}$ with the prompt δ guiding it to revise the current instruction, considering the provided feedback.

We note that, while theoretically two distinct LLMs $\mathcal{L}_2$ and $\mathcal{L}_3$ are used for gradient generation and instruction rewriting, the two steps can be merged into a single LLM call by concatenating ∇ and δ. This halves the number of LLM calls, resulting in a significant speedup of the optimization process.

(3) Instruction Selection. The step identifies the most promising instructions for the next iteration in two stages of selecting candidate instructions using UCB bandit, and identifying the top beam based on evaluation scores on the training data. This step corresponds to lines 14-18 of Algorithm 1.

Let $\mathbf{I}'_k = \bigcup_{p \in \mathbf{P}} \bigcup_j \mathbf{I}'_{p,k_j}$ be the set of all rewritten instructions at the iteration k. The first stage identifies $b' \geq b$ promising candidate

instructions $\mathbf{I}_k^{\text{cand}}$ from the rewritten instructions $\mathbf{I}'_k$ and stores them in an *instruction pool*:

$$\mathbf{I}_k^{\text{cand}} = Select_{b'}^{UCB}(\mathbf{I}'_k), \tag{5.4}$$

$$\mathbf{I}^{\text{pool}} \leftarrow \mathbf{I}_k^{\text{cand}} \cup \mathbf{I}^{\text{pool}}. \tag{5.5}$$

The second stage then evaluates these candidate instructions by computing the correlation of the generated scores with human labels using the training set. Therefore, the instructions for the next iteration $k+1$ are generated by:

$$\mathbf{I}_{k+1} = \underset{\mathbf{I}_k \subseteq \mathbf{I}^{\text{pool}}, |\mathbf{I}_k| = b}{\arg\max} \mathcal{C}(\mathbf{H}, \mathbf{S}_{\mathbf{I}_k}), \tag{5.6}$$

where $\mathbf{S}_{\mathbf{I}_k}$ denotes all predicted scores using instructions $\mathbf{I}_k$. This process follows a beam search approach, where at each iteration we create a pool of candidates, assess their performance, and select the top ones to form the new beam for continued exploration. Once all iterations are complete, the final optimal instructions $\mathbf{I}^*$ are selected by their correlation scores on a validation set $\mathbf{H}'$ (line 19 of Algorithm 1).

The UCB selection algorithm, denoted as $Select_{b'}^{UCB}(\cdot)$, is presented in Algorithm 2. For each iteration t, $N_t(I)$ denotes the number of evaluations of instruction I on sampled particles and $Q_t(I)$ denotes its estimated correlation. Following Pryzant et al. [2023], it samples a subset of particles and their corresponding human annotations, then selects the instruction that maximizes the UCB criterion $Q_t(I) + \beta\sqrt{\log t / N_t(I)}$, where β is an exploration constant. The selected instruction is evaluated on the sampled particles, and its estimated effectiveness is updated based on the correlation with human annotations. After T iterations, the algorithm returns the candidate instructions $\mathbf{I}_k^{\text{cand}}$ containing the top b' instructions according to their final estimated effectiveness Q_T with the $SelectTopInstructions_{b'}(Q_T)$ function. This forms a set of promising candidates for the next stage of the selection process. We note that, for efficient execution of the UCB process, we approximate it by dividing T into multiple small batches and processing each batch in parallel.

Algorithm 2 $Select_{b'}^{UCB}(\cdot)$ - Candidate Selection with UCB Bandits

Require: All rewritten instructions $\mathbf{I}'_k$, particles $\mathbf{P}$, human annotations $\mathbf{H}$, number of UCB iterations T, and the number of selected instructions b'.

Initialize $N_t(I) \leftarrow 0, Q_t(I) \leftarrow 0, \forall I \in \mathbf{I}'_k$
for $t = 1, ..., T$ **do**
  // Sample particles and corresponding labels uniformly
  $\mathbf{P}_{smp} \subset \mathbf{P}$, $\mathbf{H}_{smp} \subset \mathbf{H}$
  // Select the instruction with the highest UCB criterion
  $I \leftarrow \arg\max_{I \in \mathbf{I}'_k} \{Q_t(I) + \beta \sqrt{\frac{\log t}{N_t(I)}}\}$
  // Evaluate the instruction I on the sampled particles
  $\mathbf{S}_{smp} \leftarrow \{\mathcal{S}_{particle}(I, p)\}_{p \in \mathbf{P}_{smp}}$
  // Compute correlation and update UCB criterion
  Observe reward $r \leftarrow \mathcal{C}(\mathbf{S}_{smp}, \mathbf{H}_{smp})$
  $N_t(I) \leftarrow N_t(I) + |\mathbf{P}_{smp}|$
  $Q_t(I) \leftarrow Q_t(I) + \frac{r - Q_t(I)}{N_t(I)}$
end for
return $\mathbf{I}_k^{\text{cand}} \leftarrow SelectTopInstructions_{b'}(Q_T)$

5.4 Human Annotation Collection

To assess the correlation of automatic conversation evaluation methods with human judgments, we need a dataset and crowdsource human annotations on a set of human-system conversations; we thus create the CRSArena-Eval dataset to enable this meta-evaluation, focusing on recommendation conversations as a case study, which we describe in this section. To achieve this, we first collect human-system conversations (Section 5.4.1), then crowdsource human annotations on these conversations (Section 5.4.2).

5.4.1 Dialogue Collection

To create our meta-evaluation dataset, we first collect multi-turn conversations between users and various systems, on which we later collect human annotations as described in Section 5.4.2. Similarly to Mehri and Eskenazi [2020], our goal in creating such a dataset is not to train and benchmark the best-performing systems, but rather to collect responses with varying quality of systems and obtain reliable human judgments of these responses. This section describes the first

human-system dialogue collection process.

Dialogue Collection Platform. To collect human-system conversations, inspired by Chatbot Arena [Chiang et al., 2024], we develop a platform named *CRS Arena*. Our CRS Arena displays two CRSs side-by-side in a battle-style interface where users interact with them one after the other, resulting in a collection of conversations and user feedback. While users can declare a winner, these declarations cannot directly serve as evaluation data due to the lack of evaluation aspects and granularity, which we will further collect annotations on in Section 5.4.2.

CRSs. The resulting dataset, called *CRSArena-Dial*, consists of dialogues with multi-turn interactions between users and nine state-of-the-art CRSs trained on OpenDialKG [Moon et al., 2019] and ReDial [Li et al., 2018] datasets. Specifically, the CRSArena-Dial dataset consists of human conversation with nine state-of-the-art CRSs, including KBRD [Chen et al., 2019], BARCOR [Wang et al., 2022a], UniCRS [Wang et al., 2022b], ChatGPT [Wang et al., 2023a], and CRB-CRS [Manzoor and Jannach, 2022], each developed based on OpenDialKG [Moon et al., 2019] and ReDial [Li et al., 2018] datasets, except CRB-CRS, which is solely on the ReDial dataset.

Collected Dialogues. The CRSArena-Dial dataset is collected through two distinct crowdsourcing environments: an open setup with public access and a closed setup with access restricted to a selected group of crowd workers. A key feature of this dataset is that it contains authentic conversations between real users and fully automated CRSs, as opposed to the Wizard-of-Oz setting commonly used in dialogue corpora creation. While the minimally regulated environment may introduce some noise, the dataset is highly representative of how regular users naturally express preferences and seek recommendations.

5.4.2 Dialogue Annotation

In this section, we describe the process of collecting human annotations on the CRSArena-Dial dataset described in Section 5.4.1.

Dataset. We collect human annotations on the CRSArena-Dial dataset as described in Section 5.4.1. To ensure the quality of dialogues, we excluded seven dialogues that were unsuitable for our annotation, such

as those with only a single user utterance, resulting in 467 dialogues with a total of 2,235 system responses for our annotations.

Evaluation Aspects. Over 50 aspects have been proposed for evaluating conversational systems in the literature. We follow Siro et al. [2022] and collect annotations for seven evaluation aspects, whose effectiveness for CRS evaluation has been rigorously validated [Siro et al., 2022, 2023]. These aspects cover both system and user-centric features of a conversation. The aspects and their descriptions (used as instructions to annotators) are as follows:

Turn-level Aspects:

- *Relevance (0–3):* Does the assistant's response make sense and meet the user's interests?
- *Interestingness (0–2):* Does the response make the user want to continue the conversation?

Dialogue-level Aspects:

- *Understanding (0–2):* Does the assistant understand the user's request and try to fulfill it?
- *Task Completion (0–2):* Does the assistant make suggestions that the user finally accepts?
- *Efficiency (0–1):* Does the assistant suggest items matching the user's interests within the first three interactions?
- *Interest Arousal (0–2):* Does the assistant try to spark the user's interest in something new?
- *Overall Impression (0–4):* What is the overall impression of the assistant's performance?

These instructions and scales are based on Siro et al. [2022], with minor adjustments from Sakai [2023b] for clarity. Turn-level aspects are evaluated for each system turn, while dialogue-level aspects are evaluated for the entire dialogue.

Although we select seven thoroughly assessed aspects supported by the existing literature [Siro et al., 2023], we note that FACE can be

applied to a broad range of aspects and the choice of aspects is often context-dependent. Therefore, FACE algorithm allows researchers and practitioners to explore diverse aspects relevant to their own systems.

5.4.3 Interface

To crowdsource high-quality annotations, we built an interface, through multiple pilot experiments, to overcome the widely reported challenges in the literature [Bernard and Balog, 2023, Eickhoff and de Vries, 2011]. Those challenges include workers (1) skipping reading context, (2) misunderstanding aspect definitions, (3) getting distracted by overwhelming information on the annotation page, and (4) annotating randomly without focus. Our interface requires users to pass a quiz on aspect definitions and enforces the annotation of each turn before evaluating the entire dialogue. It further includes hidden tests with expert-verified answers, and workers who fail to meet the required agreement are excluded.

5.4.4 Participants and Quality Control

We recruited Prolific[1] workers from English-speaking countries with a 100% approval rate and ≥ 1000 previous submissions. Considering some workers exhibit behavior aimed at just maximizing financial gain [Eickhoff and de Vries, 2011], we filtered out those with a history of subpar submissions to ensure quality. Furthermore, workers with $<30\%$ agreement with experts on hidden tests were excluded. Each batch of work contained annotations for 20 dialogues, taking around 40 minutes to complete, at the cost of £6. Three annotations per annotation task were collected. In case of disagreement, additional annotations were collected until ties were resolved.

5.4.5 Analysis

We now analyze our collected annotations, referred to as *CRSArena-Eval*.

Statistics. A total of 20,962 annotations were collected from 109 workers, spanning 467 dialogues and 2,235 system turns. On average,

[1] `https://www.prolific.co/`

Table 5.1: Inter-annotator agreement of CRSArena-Eval and AB-ReDial [Siro et al., 2023] based on Pearson's r, Spearman's ρ, and Krippendorff's α correlations.

Aspect	CRSArena-Eval			AB-ReDial		
	r	ρ	α	r	ρ	α
Turn-level						
Relevance	**0.613**	**0.611**	**0.612**	0.527	0.502	0.526
Interestingness	**0.386**	**0.386**	**0.375**	0.209	0.217	0.209
Dialogue-level						
Understanding	**0.505**	**0.477**	**0.496**	0.321	0.313	0.318
Task Completion	**0.481**	**0.440**	**0.482**	0.345	0.321	0.346
Efficiency	**0.297**	**0.297**	**0.289**	0.225	0.225	0.226
Interest Arousal	0.242	0.241	0.226	**0.254**	**0.291**	**0.247**
Overall Impression	**0.573**	**0.526**	**0.572**	0.321	0.300	0.321

each dialogue has 14.6 annotation tasks, each annotated by three workers. Noteworthy, our annotation interface made the process highly efficient, requiring only 8 seconds per annotation. For 92% of the tasks, an agreement was achieved by the first three annotators and for the remaining 8% of tasks additional annotations were collected to resolve ties. In total, these processes produced 6,805 final labels.

Inter-annotator Agreement. Given the ordinal nature of judgments, we report inter-annotator agreement using Pearson's r and Spearman's ρ, following Mehri and Eskenazi [2020], along with Krippendorff's α. Average scores across all aspects are $r = 0.443$, $\rho = 0.425$, and $\alpha = 0.436$, indicating moderate agreement. For comparison, the same aspects on the AB-ReDial dataset [Siro et al., 2022, 2023] yield average agreements of $r = 0.328$, $\rho = 0.325$, and $\alpha = 0.327$, showing higher agreement of our CRSArena-Eval annotations.

Table 5.1 presents the breakdown of inter-annotator agreement for each aspect, showing that CRSArena-Eval consistently achieves higher agreement than AB-ReDial in all aspects except interest arousal. These results demonstrate that CRSArena-Eval provides higher quality annotations even compared to existing high-quality annotations of the AB-ReDial dataset, showing both the difficulty of the task and high quality annotations of our dataset.

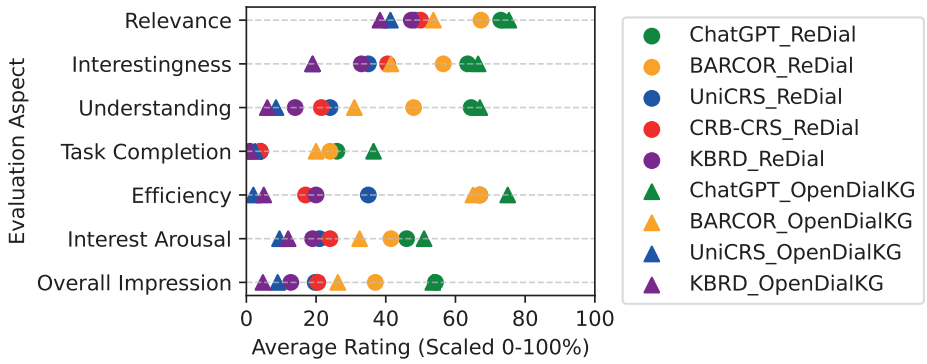


Figure 5.2: Distribution of human annotation scores for seven aspects across nine systems in CRSArena-Eval.

System Score Distribution. Figure 5.2 shows the distribution of collected scores for the nine CRSs in CRSArena-Eval. It shows no system reaches the high end of the scale, indicating that existing CRSs do not fully satisfy users. This aside, the scores cover a broad range, reflecting differing system quality, which is crucial to assess the ability of automatic evaluators to distinguish system performance [Mehri and Eskenazi, 2020].

5.5 Experimental Setup

Datasets. For instruction optimization, we use the *AB-ReDial* [Siro et al., 2022, 2023] dataset, which contains annotations of human-human conversations from the ReDial dataset (cf. Sec. 5.4.2). The annotations are obtained for the seven evaluation aspects and are per turn/dialogue. This ensures no human-system conversations are involved in the optimization process. We use 60% of AB-ReDial for training and the rest for validation. For evaluation, we use the *CRSArena-Eval* dataset (cf. Sec. 5.4). CRSArena-Eval (RD)/(KG) denote the subset of the dataset for systems developed using ReDial [Li et al., 2018] and OpenDialKG [Moon et al., 2019], respectively.

Settings. Unless indicated otherwise Llama-3.1-8B-Instruct [Dubey et al., 2024] is used for FACE. All experiments were performed using the SGLang [Zheng et al., 2024] library for its prediction efficiency. The

temperature of 0.6 is set across all experiments unless otherwise stated. Following Pryzant et al. [2023], we run the instruction optimization for $K = 6$ iterations and stored $b' = 16$ instructions in the instruction pool, resulting in 96 instructions. We set parameters $\alpha = 2$, $c = 1$, $b = 4$, and use the batch size of $B = |\mathbf{I}'_k|/2$ and $T = 5B$ iterations. A sampling size of $n = 5$ is used to create a score distribution (Sec. 5.3.2). All hyperparameters are obtained using the validation set or following Pryzant et al. [2023]. The final selection of the optimal instruction set $\mathbf{I}^*$ (line 18 of Algorithm 1) is done using the validation set and results in a set of 16 instructions.

Prompts. Inspired by TREC RAG 2024 [Pradeep et al., 2025], the decomposer prompt starts with: *"Your task is to extract conversation nuggets [...]"* followed by CoT prompts and their format. The textual gradient prompt ∇ begins with *"Examine the original instructions, predicted nugget score, and gold score."* and then identifies inconsistency between predicted and gold scores, followed by suggestions to the instruction if necessary. The instruction rewriting prompt δ starts with *"Propose new instructions of $\tilde{5}0$ words based on [...]"* followed by the guidance on how to rewrite based on the textual gradient, inspired by Ye et al. [2024]. The seed evaluation instruction I is as follows *"Given the dialogue, evaluate the quality of the target nugget based on the given aspect. Step 1: [...]"*. Examples of prompts are provided in Appendix B.

Baselines. Multiple automatic evaluators are used as our baselines: LLM^{Direct} directly prompts the LLM to annotate the given turn/dialogue using the same instructions as human annotations for CRSArena-Eval; $LLM^{CoT+ICL}$ adds CoT [Kojima et al., 2022, Wei et al., 2022] and in-context learning (ICL) [Brown et al., 2020] with two examples to the prompt of $\text{LLM}^{\text{Direct}}$; *UniEval* [Zhong et al., 2022] and *G-Eval* [Liu et al., 2023b] are state-of-the-art reference-free conversation evaluation methods. Since UniEval covers limited aspects, we only report those overlapping with ours. For fair comparison, we use Llama-3.1-8B-Instruct as the backbone for LLM-based methods $\text{LLM}^{\text{Direct}}$, $\text{LLM}^{\text{CoT+ICL}}$, and G-Eval in Section 5.6.1. To show generalizability to different LLMs in Section 5.6.2, other LLMs are used for our experiments.

Metrics. For correlation metrics, we use Pearson's and Spearman's to appropriately handle the annotation scales. Following Mehri and Eskenazi [2020], correlation significance is computed by p-value derived from t-distribution using Python's SciPy library [Virtanen et al., 2020].

5.6 Results

In this section, we address the research questions outlined in Section 5.1 through a series of experiments: **RQ4a:** *How does FACE correlate with human judgments?* **RQ4b:** *How generalizable are FACE-optimized instructions across different LLMs and domains?* **RQ4c:** *Can fine-grained evaluation scores of FACE provide insights about system's issues?*

5.6.1 FACE Annotation Correlation

Annotation Correlation. Table 5.2 shows the annotation correlation results, demonstrating that FACE, on average, outperforms all baselines by a large margin. We note that FACE is not optimized on any human-system conversations (cf. Sec. 5.5), highlighting its strong generalization to unseen systems. Although one can argue that FACE can capture some information from the human-human ReDial dataset during the instruction optimization process, results on CRSArena-Eval (KG) shows generalization and robustness of FACE to unseen recommendation datasets.

The results also demonstrate that even without instruction optimization (FACE w/o train), FACE remains competitive with the state-of-the-art method, G-Eval, suggesting the effectiveness of our particle-based approach. The benefit of training is more pronounced for challenging aspects such as *Interestingness*, *Efficiency*, and *Interest Arousal*, which indicates that FACE effectively optimizes instructions for aspects that LLMs struggle to capture using a single thought process.

Ranking Correlation. Table 5.3 shows the correlation of system rankings created by different automatic evaluation methods. System ranking correlations are calculated by averaging each system's score, ranking systems, and measuring correlation with system rankings

Table 5.2: Annotation correlations based on Pearson's r and Spearman's ρ. All columns show the correlations averaged over all aspects. All FACE correlations are statistically significant with $p < 0.01$.

Methods	Turn-level				Dialogue-level										All	
	Rel.		Int.		Und.		Task		Eff.		Int.		Overall			
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
CRSArena-Eval (RD)																
$\text{LLM}^{\text{Direct}}$	0.464	0.455	0.248	0.260	0.522	0.482	0.405	0.363	0.101	0.101	0.217	0.203	0.564	0.522	0.360	0.341
$\text{LLM}^{\text{CoT+ICL}}$	0.453	0.446	0.175	0.177	0.481	0.457	0.425	0.400	0.174	0.174	0.188	0.174	0.498	0.472	0.342	0.329
UniEval	0.311	0.288	0.182	0.242	0.246	0.225	–	–	–	–	–	–	0.395	0.387	–	–
G-Eval	0.490	0.471	0.302	0.289	0.490	0.444	0.351	0.364	**0.488**	0.482	0.332	0.325	0.577	0.577	0.433	0.422
FACE w/o train	0.468	0.462	0.279	0.290	0.605	0.574	0.482	0.392	0.339	0.423	0.235	0.255	0.617	0.555	0.432	0.422
FACE	**0.549**	**0.550**	**0.443**	**0.437**	**0.650**	**0.635**	**0.570**	**0.453**	0.484	**0.534**	**0.447**	**0.430**	**0.712**	**0.668**	**0.551**	**0.530**
CRSArena-Eval (KG)																
$\text{LLM}^{\text{Direct}}$	0.452	0.452	0.238	0.231	0.599	0.546	0.538	0.481	0.137	0.137	0.425	0.378	0.655	0.557	0.435	0.397
$\text{LLM}^{\text{CoT+ICL}}$	0.419	0.408	0.203	0.190	0.562	0.520	0.475	0.434	0.114	0.114	0.309	0.279	0.599	0.521	0.383	0.352
UniEval	0.416	0.428	0.262	0.401	0.563	0.541	–	–	–	–	–	–	0.618	0.659	–	–
G-Eval	0.533	0.505	0.334	0.316	0.535	0.475	0.430	0.422	0.485	0.463	0.424	0.403	0.656	0.642	0.485	0.461
FACE w/o train	0.492	0.486	0.308	0.324	0.664	0.611	0.426	0.411	0.240	0.419	0.297	0.322	0.672	0.557	0.443	0.447
FACE	**0.543**	**0.527**	**0.471**	**0.453**	**0.719**	**0.677**	**0.593**	**0.484**	**0.518**	**0.543**	**0.449**	**0.404**	**0.766**	**0.679**	**0.580**	**0.538**

Methods	Turn-level		Dial-level		All	
	r	ρ	r	ρ	r	ρ
R@1	-0.197	0.060	-0.120	0.081	-0.142	0.075
R@10	-0.192	0.048	-0.111	0.071	-0.134	0.064
iEvaLM$^{\text{free}}$	0.191	0.191	0.232	0.232	0.257	0.184
iEvaLM$^{\text{attr}}$	0.527	0.527	0.553	0.553	0.575	0.517
Distinct-3	0.716	0.841	0.665	0.780	0.680	0.798
Distinct-4	0.654	0.800	0.609	0.760	0.622	0.771
LLM$^{\text{Direct}}$	0.860	0.822	0.872	0.799	0.868	0.806
G-Eval	0.740	0.840	0.893	0.830	0.850	0.833
FACE	**0.930**	**0.842**	**0.913**	**0.837**	**0.918**	**0.838**

Table 5.3: System ranking correlations on CRSArena-Eval, averaged over corresponding aspects. All FACE correlations are statistically significant with $p < 0.05$.

based on human judgments. We obtain correlation for reference-based metrics by computing system rankings from reported scores [Wang et al., 2023a]. As baselines, we report Recall, and Distinct-n [Li et al., 2016] as reference-based metrics, and LLM$^{\text{Direct}}$ and G-Eval as top-performing reference-free metrics from Table 5.2.

From the results, we notice that recall metrics are insufficient, which is in line with the literature [Bernard et al., 2025]. Surprisingly, iEvaLM baselines struggle to achieve high ranking correlations; this could be due to their reliance on a recall-based approach to evaluate the simulation results. While the results show that FACE outperforms all baselines, we note that system ranking correlations should be used with caution, as they are less informative than annotation correlation, especially for competitive systems [Faggioli et al., 2023, Dietz et al., 2025].

Overall, we can answer **(RQ4a)**: FACE achieves high annotation and system ranking correlations with human judgments, outperforming state-of-the-art methods by a large margin.

5.6.2 Generalizability of FACE

We hypothesize that the pool of optimized instructions by FACE can be reused for different LLMs and domains. To assess this hypothesis,

Methods	LLM	Size	CRS-RD		CRS-KG		Avg.	
			r	ρ	r	ρ	r	ρ
$\text{LLM}^{\text{Direct}}$	Llama	8B	0.564	0.522	0.655	0.557	0.610	0.540
G-Eval	Llama	8B	0.577	0.577	0.656	0.642	0.617	0.610
FACE	Llama	8B	**0.712**	0.668	**0.766**	0.679	**0.739**	0.674
FACE*	Gemma	9B	0.689	**0.687**	0.718	**0.703**	0.704	**0.695**
FACE*	Gemma	2B	0.647	0.603	0.728	0.646	0.688	0.625
FACE*	Qwen	7B	0.698	0.664	0.764	0.693	0.731	0.679
FACE*	Qwen	3B	0.643	0.632	0.725	0.674	0.684	0.653
FACE*	Qwen	1.5B	0.557	0.606	0.605	0.635	0.581	0.621

Table 5.4: Results on generalizability of FACE to other LLMs. CRS-RD and -KG represent CRSArena-Eval (RD) and (KG), respectively. All FACE annotation correlations are statistically significant with $p < 0.01$.

we take the instruction pool and re-select the top instructions (line 18 of Algorithm 1) for different LLMs and datasets. We denote this adapted FACE as *FACE**.

Generalization to other LLMs. To assess generalizability of FACE to other LLMs, we use the AB-ReDial validation set and re-select instructions for five LLMs: Gemma 2 (9B and 2B) and Qwen 2.5 (7B, 3B, 1.5B). Table 5.4 shows annotation correlation results for top-performing baselines of Table 5.2. The results indicate that adapting to Gemma 9B and Qwen 7B achieves performance comparable to FACE. Interestingly, our method is highly effective for small models: FACE adapted to Gemma 2B outperforms G-Eval, which uses an LLM with 4x more parameters. A similar observation can be made for Qwen 3B and 1.5B.

Generalization to chitchat conversations. To examine generalizability of FACE to another type of conversations, we evaluate FACE on the existing chitchat datasets: (1) *USR-Persona* [Mehri and Eskenazi, 2020] (based on PersonaChat [Zhang et al., 2018a]), containing personalized chit-chats, and (2) *USR-Topical* [Mehri and Eskenazi, 2020] (based on Topical-Chat [Gopalakrishnan et al., 2019]), containing knowledge-grounded conversations. These datasets provide annotations for six evaluation aspects, of which "maintains context" is

Methods	USR-Persona		USR-Topical		Avg.	
	r	ρ	r	ρ	r	ρ
ROUGE-L	0.114	0.091	0.193	0.203	0.154	0.147
BLEU-4	0.147	0.151	0.131	0.235	0.139	0.193
METEOR	0.250	0.256	0.250	0.302	0.250	0.279
BERTScore	0.188	0.157	0.214	0.233	0.201	0.195
Dial-M	0.400	0.390	0.370	0.400	0.385	0.395
USR	0.607	0.528	0.416	0.377	0.512	0.453
UniEval	0.616	0.580	**0.595**	**0.613**	0.605	0.597
G-Eval$^{\text{GPT-3.5}}$	0.441	0.458	0.519	0.544	0.480	0.501
G-Eval$^{\text{GPT-4}}$	0.607	0.670	0.594	0.605	0.601	0.638
FACE	0.473	0.544	0.498	0.506	0.486	0.525
FACE$^{*}_{\text{72B}}$	**0.681**	**0.697**	0.570	0.582	**0.625**	**0.639**

Table 5.5: Results on generalizability of FACE to chitchat conversations. All FACE correlations are statistically significant with $p < 0.01$.

the only aspect that is similar to ours. We use the instruction pool for the relevance aspect and re-select optimal instructions using the validation set of USR-Persona for USR-Topical evaluation and vice versa, to ensure that the test set is completely unseen.

Table 5.5 presents the results of FACE generalizability to chitchat conversations, with G-Eval results obtained using GPT-3.5 and 4. On average, FACE outperforms all baselines except G-Eval with GPT-4, while FACE$^{*}_{\text{72B}}$ outperforms all baselines. This is especially striking, considering that FACE is completely blind to category of conversations and uses an LLM with a lower number of parameters than GPT. Additionally, using an open model for evaluation has the added value of reproducibility.

Based on the results of Tables 5.4 and 5.5, we answer our second research question **(RQ4b)**: FACE-optimized instructions are highly generalizable to different LLMs and domains, by performing a simple adaptation of FACE to a new LLM/domain. The adaptation to larger models can even surpass the state-of-the-art method with GPT-4 as a backbone on chit-chat conversations.

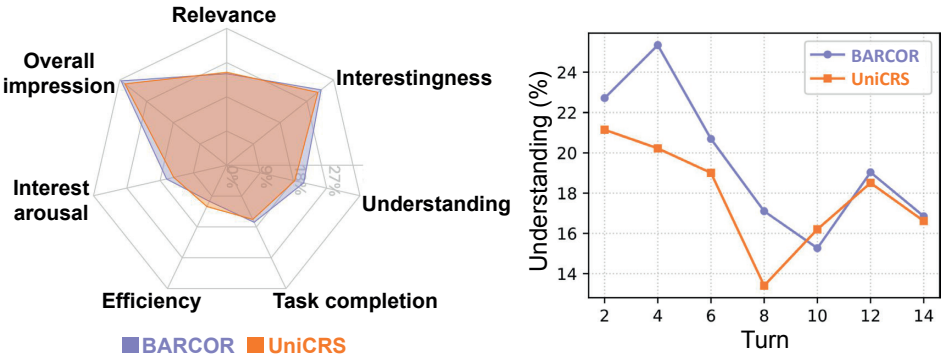


Figure 5.3: Breakdown analysis comparing BARCOR and UNICRS. Left: FACE evaluation results for each aspect. Right: Scores of Understanding aspect per system turn.

5.6.3 FACE Interpretability

We present a preliminary small-scale case study to demonstrate how FACE fine-grained scores can assist humans in identifying issues in CIA systems. Specifically, we compare FACE scores for two systems that are difficult to diagnose and receive contradictory evaluations from human and existing metrics: BARCOR [Wang et al., 2022a] and UniCRS [Wang et al., 2022b]. Humans and FACE prefer BARCOR (cf. Fig. 5.2 and [Bernard et al., 2025]), while recall-based metrics favor UniCRS [Wang et al., 2023a]. The left radar chart in Figure 5.3 shows FACE analysis for BARCOR and UniCRS for seven evaluation aspects. Although the overall impression indicates similar performance, FACE scores show that UniCRS excels in relevance and efficiency, while BARCOR is better in user understanding and keeping users interested. The right graph shows the detailed analysis of the user understanding aspect, where particle scores are aggregated for each turn. It is evident from the plot that BARCOR obtains higher scores in earlier turns. This indicates that BARCOR understands user preferences early on, which may explain its higher human preference. These insights suggest that, while UniCRS excels in recommendations, overall performance can be improved by focusing on user understanding.

To further understand each particle's contribution, we analyze particle scores per dialogue act and assess how they influence the overall impression of the systems. We found that particles associated with

Table 5.6: Example conversations from UniCRS and BARCOR, with particles scored in the top quartile shown in green and those in the bottom quartile shown in red.

Role	Utterance
User	Give me a title for the best fantasy based film
Asst	I would recommend the following items 1 The Lord of the Rings 2 Pan's Labyrinth (2006) 3 Avatar (2009)
User	which of the three do you think has the overall most comprehensive plotline
Asst	yes it does
User	Can you choose one that is most comprehensive
Asst	Yes, it does.
User	Do you not understand what I am asking?
...	...

UniCRS

Role	Utterance
User	Hello! Can you give me a movie recommendation, please?
Asst	What kind of movies do you like?
User	Sci-fi, for instance
Asst	I would recommend the following items: 1: Blade Runner 2049 (2017) 2: Star Wars: Episode VIII – The Last Jedi (2017) 3: Arrival (2016)
User	Thanks, those are great recommendations. Could you give me any comedy movies?
Asst	Have you seen The Hangover (2009)?
User	I haven't, but I have heard of it. How many Hangover movies are there?
...	...

BARCOR

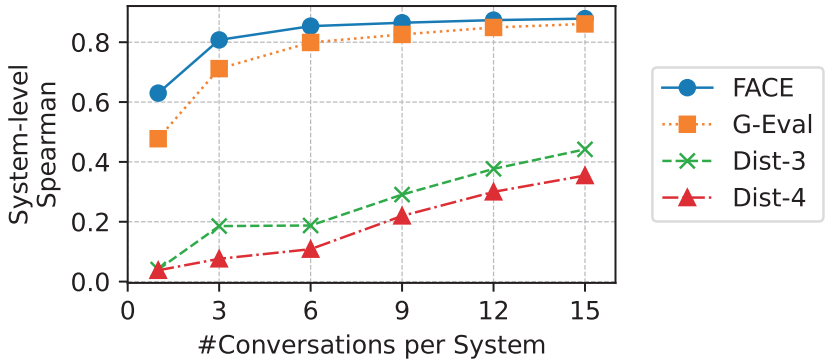


Figure 5.4: Sample efficiency of different evaluation methods for the overall aspect on CRSArena-Eval.

"preference elicitation" scored higher for BARCOR than UniCRS (45.5 vs. 39.0; scaled 0-100%), which shows BARCOR's superior user understanding stems from more effective preference elicitation. Table 5.6 shows example conversations from both CRSs, illustrating the general phenomenon captured by FACE scores. The examples demonstrate that while UniCRS provides good initial recommendations, it fails to understand the user and elicit further preferences, whereas BARCOR successfully elicits preferences and recommends items matching user interests. This difference can be overlooked when using item-level recall metrics, and it demonstrates the strength of FACE's interpretability.

Overall, we answer **(RQ4c)** positively: FACE can provide valuable insights into systems' behavior, which are useful for system improvement.

5.7 Analysis

Sample Efficiency. To determine the systems' score and create a system ranking, the evaluation method needs to be fed with a sample of user-system conversations. To measure how many samples are needed to find a system ranking with a high correlation with human judgments, we plot system ranking correlations for various conversation counts per system in Figure 5.4. The results indicate that FACE has strong sample efficiency; it achieves a Spearman correlation of 0.8 with gold

rankings using only 3 dialogues per system, making it twice as efficient as the best-performing existing method, G-Eval. Given that collecting human-system conversations require cost and effort, FACE's sample efficiency significantly enhances actual usability.

Bias Analysis. We conduct analyses to see whether FACE exhibits length bias and self-bias (cf. Sec. 5.2). For length bias [Dubois et al., 2024, Wang et al., 2023c], using the CRSArena-Eval dataset, we examine the correlation between a system's average word count in conversations and the overall score. We find that Pearson's correlations are 0.824 and 0.868 for FACE and humans, respectively. This indicates no sign of length bias compared to humans, which is in line with existing work [Chiang et al., 2024, Dubois et al., 2024] that report humans also favour longer responses, highlighting the nuanced nature of the LLM length bias.

For self-bias, where LLMs prefer system responses over human ones, we use FACE to evaluate pairs of system- and human-generated responses and see if they show any preferences compared to gold human annotators. We examine two conversation types: USR-Persona for chit-chat, and a combination of CRSArena-Eval and AB-ReDial for a recommendation. We could not find evidence for self-bias in FACE; e.g., for USR-Persona, FACE aligns with human preferences 77.8% of the time when humans prefer human-generated responses and 71.4% of the time when they prefer system-generated responses.

5.8 Conclusion

We present FACE, a fine-grained, aspect-based evaluation method for CIA systems. It addresses the shortcomings of existing metrics, such as focusing on fixed dialogue history with reference-based metrics, limited generalizability of LLM-based metrics, and relying on non-granular scores with limited insights. FACE is shown to strongly correlate with human judgments, generalize across LLMs and domains, and provide insights for system improvement. Future work needs to address current limitations (see Section 6.3) by further examining evaluation biases, assessing effectiveness across broad domains, and exploring how FACE can help expert evaluators.

Chapter 6

Conclusion

This chapter begins by describing our answers to the research questions laid out in Chapter 1, followed by future work in the area of personalized conversational information access (CIA) systems.

6.1 Answers to Research Questions

The main research question that we set out to address, as discussed in Chapter 1, was:

> **RQ Main:** *How can we develop personalized CIA systems and evaluate them effectively?*

To answer this question, we focused on the four sub-questions from three key areas: personal information extraction, personalized response generation, and conversation evaluation.

6.1.1 Personal Information Extraction

To personalize CIA systems, understanding the user's context is crucial. One of the promising techniques for extracting personal information from conversations is Entity Linking (EL), which is proven to be effective in documents and short queries. However, while EL for those use cases has been extensively studied, research on its application in conversational settings has been limited. Therefore, we began investigating the characteristics of EL in conversations by addressing

the first research question:

> **RQ1:** *What does Entity Linking in conversations entail?*

To address this question, we introduced ConEL in Chapter 2, a dataset of conversations annotated with named entities, general concepts, and personal entities. Through our analysis of this dataset, we discovered that human annotators identified not only named entities but also general concepts as important entities in conversations, with personal entities showing a particularly high presence in social chat conversations. Based on these findings, we concluded that *entity linking in conversations entails linking entities that include general concepts, named entities, and personal entities to enhance machine understanding of users' utterances, in contrast to traditional approaches that focus mainly on named entities.*

Building upon this dataset, we subsequently analyzed the performance of traditional EL systems. Our evaluation revealed that traditional EL tools were suboptimal for conversational contexts, as they particularly struggled to handle concepts and personal entities, which proved crucial for enabling personalized conversations.

This finding led us to the second research question:

> **RQ2:** *How can we develop an Entity Linking method to accurately identify important entities for personalized conversational systems?*

To address this question, we proposed CREL in Chapter 3, a conversational EL method designed for personalized conversational systems. CREL employs coreference resolution to recognize and link personal entity mentions to their corresponding entries in a knowledge graph, thereby addressing the challenges of personal contextual understanding. To enable the development and evaluation of CREL, we created ConEL-2, an extension of the ConEL dataset that contains 11 times more dialogues with personal entity annotations. Our evaluation results demonstrate that CREL significantly outperforms traditional EL systems, particularly in handling personal entities, which is crucial for personalized CIA systems.

Overall, we defined the task of EL in conversations and demonstrated that existing EL systems were insufficient for conversational

contexts; we then proposed CREL, an EL method for conversations that effectively addressed personal contextual challenges and significantly outperformed traditional EL systems.

6.1.2 Personalized Response Generation

In the previous section, we addressed the challenges of EL in conversational settings, which is a crucial step for personalized response generation. However, even with an effective EL method, generating responses that effectively reflect a user's personal preferences remains a key challenge.

> **RQ3:** *How can users' personal preferences be utilized to generate personalized responses?*

To investigate this research question, we must first address a significant bottleneck: the limited availability of realistic large-scale dialogue datasets for training and evaluating personalized response generation models. Without datasets that accurately reflect real-world user interactions and preferences, it becomes impossible to effectively investigate how personal preferences can be leveraged in response generation.

In the first half of Chapter 4, we proposed LAPS, a framework that leverages large language models (LLMs) to efficiently guide human workers in composing multi-session and multi-domain conversations that capture diverse user preferences. Our experiments confirmed that this approach significantly accelerates dataset construction while maintaining the diversity and quality of human-written dialogues, effectively addressing the scalability constraints of existing approaches. The results demonstrated that LAPS-produced conversations exhibit the same level of naturalness and diversity as expert-created ones, highlighting the framework's effectiveness in collecting real-world personalized conversational data.

In the second half of Chapter 4, we addressed RQ3 by training a preference extraction model and performing personalized response generation using semi-structured user preferences (termed *preference memory*), which are constructed from the preferences extracted by the model. The experiments demonstrated that utilizing preference memory enhances effective preference utilization, provides explanations for recommendations, and mitigates LLMs' recall limitations during

response generation. This suggests the effectiveness of using extracted personal preferences for generating personalized responses.

6.1.3 Conversation Evaluation

The evaluation of personalized CIA systems is crucial, yet existing methods often fail to assess the whole conversation across diverse interaction trajectories. This highlights the necessity for more effective, automatic evaluation, which motivated our fourth research question:

> **RQ4:** *How to automatically evaluate personalized CIA systems in a way that accounts for natural conversation flow and provides fine-grained insights into system performance?*

To answer this question, we introduced FACE in Chapter 5. FACE is a fine-grained, aspect-based evaluation method for CIA systems that captures the natural conversation flow and provides detailed insights into system performance. FACE employs an algorithm that first decomposes system responses into fine-grained "conversation particles." It then evaluates these particles using an LLM guided by a diverse set of instructions. These instructions are iteratively refined through a prompt optimization process that leverages "textual gradients," which is a form of natural language feedback, and uses beam search to select the instructions that best correlate with human judgments.

To enable the evaluation of automatic evaluators (meta-evaluation), we created CRSArena-Eval, a new meta-evaluation dataset containing 20,962 human annotations on 467 conversations with nine diverse conversational recommender systems (CRSs). The annotations cover seven key turn- and dialogue-level aspects, allowing for a robust meta-evaluation of how well automatic evaluators correlate with human judgments.

Evaluation of FACE on CRSArena-Eval demonstrates that FACE closely aligns with human judgments, significantly outperforming traditional automatic evaluation methods. In addition, we demonstrated FACE's generalizability, showing its optimized instructions can be effectively adapted to other LLMs and to a different conversational domain, namely chit-chat. Moreover, we showed that FACE's fine-grained scores can be used to identify the issue within the conversation, and provide insights for system improvement.

6.2 Positioning of the Thesis

This section positions the contributions of this thesis in relation to existing work in the CIA field.

Personal Information Extraction. Chapters 2 and 3 advance personal information extraction by addressing EL challenges in conversations, overcoming limitations of existing EL methods for well-written documents [van Hulst et al., 2020, Cao et al., 2021] that perform suboptimally in conversational contexts. Recent advances by Li et al. [2023], Aliannejadi et al. [2024] demonstrate the growing importance of personal preference extractions, positioning this work as foundational for extracting preferences to populate such personal knowledge bases/graphs from conversational text. The methods developed here can serve as a basis for further exploration into more context-aware personalized CIA systems.

Personalized Response Generation. Chapter 4 contributes a scalable, LLM-augmented method for collecting personalized dialogue data, addressing the scalability challenge of prior human-written collections from studies [Radlinski et al., 2019, Bernard and Balog, 2023]. This work is positioned to complement the recent LLM post-training paradigm [Ouyang et al., 2022, Zhou et al., 2023a] by providing a method for collecting the authentic preference-grounded data that can be used for model alignment to further enhance personalized CIA systems.

Conversation Evaluation. Chapter 5 introduces a fine-grained automatic evaluation method that addresses the limitations of existing approaches [Zhang et al., 2020, Mehri and Eskenazi, 2020, Liu et al., 2023b, Zhong et al., 2022], which produce uninterpretable scores and fail to assess the whole conversation or diverse interaction trajectories. This work is positioned to strengthen the LLM-as-a-judge paradigm through automated instruction optimization, addressing LLM biases identified by multiple studies [Wang et al., 2023c, Dubois et al., 2024, Xu et al., 2024, Liu et al., 2023b, Dietz et al., 2025, Faggioli et al., 2023, Clarke and Dietz, 2024], while providing interpretable scores for developing advanced user simulators and meta-evaluation benchmarks.

6.3 Limitations

This section discusses the limitations of this thesis.

General Limitations. As we have discussed in Chapter 1, one of the limitations of this thesis is that our focus has been restricted to text-based interactions. However, the growing popularity of voice assistants and visual chatbots signals a shift toward multi-modal interactions [Zamani et al., 2023]. To solve this limitation, exploration of multi-modal extensions incorporating voice or visual inputs is needed.

Ethical Risks of Social Chat. Although this thesis focuses on information access, several chapters involve social chat as a conversational setting for analysis and evaluation. We acknowledge that such interaction is not entirely innocuous; social chat systems can foster over-reliance and emotional dependency, posing risks to user autonomy and privacy [Akbulut et al., 2024]. While the mitigation of these risks fall outside the scope of this thesis, we acknowledge that the importance of recognizing these potential harms is fundamental to the ethical design of conversational systems.

Personal Information Extraction. In Chapter 3, we proposed CREL, a conversational EL method for conversations. However, the current method only focuses on personal entity mentions starting with possessive pronouns such as "my" or "our". Therefore, it cannot capture personal entities expressed through complex references, such as "the car I owned." The mitigation strategy might involve training a personal entity mention detection model using the same approach as mention detection in traditional EL systems [van Hulst et al., 2020], but with a dataset specifically targeting personal mentions that can be built upon the same annotation strategy we developed in ConEL (Chapter 2) and ConEL-2 (Chapter 3). Another limitation is that our method may not generalize to domain-specific knowledge bases with long-tail entities [Hoveyda et al., 2024], and requires domain adaptation.

Personalized Response Generation. In Chapter 4, we introduced LAPS, a scalable approach for creating personalized conversational datasets by utilizing LLMs to guide human annotators. A limitation, however, is that creating task descriptions and specific LLM prompts still requires manual effort for each new task setting or domain. Future research should aim to moderate this process to allow researchers to

more conveniently adapt frameworks to various settings, thus improving scalability and usability further.

Conversation Evaluation. In Chapter 5, we introduced FACE, a fine-grained, aspect-based conversation evaluation method. A few limitations are as follows: (1) While FACE is a general method applicable to various conversations, this chapter evaluated it only on CRSs and one aspect for chit-chat; further exploration in other domains/aspects is needed. (2) Similarly, the proposed meta-evaluation dataset, CRSArena-Eval, specifically targets CRSs, which is a subset of the broader field of information-access conversations which often involve more complex user needs and mixed initiatives [Zamani et al., 2023]; thus, incorporating broader types of information-access conversations is needed. (3) Although FACE did not exhibit bias in our analysis (Sec. 5.7), evaluating unknown or emerging biases [Fang et al., 2025] remains underexplored. (4) Knowing the limitations of LLM-based evaluation [Dietz et al., 2025, Clarke and Dietz, 2024, Faggioli et al., 2023, Takehi et al., 2024], we emphasize that FACE may not replace expert human evaluations. Instead, it facilitates research and development on conversational systems by offering a scalable and efficient method for evaluating some potential known issues of CRSs [Dubois et al., 2023].

6.4 Future Directions

This section discusses several promising future research directions in the area of personalized CIA systems.

6.4.1 Personal Information Extraction

Representation of Personal Information. Chapters 2 and 3 demonstrated the extraction of personal information from conversations using EL methods. However, determining *how* to store personal information for CIA systems remains an open question. While knowledge graphs (KGs), as employed in Chapters 2 and 3, represent one possible approach, alternative storage methods are being explored, including personalized textual knowledge bases (PTKBs) [Aliannejadi et al., 2024] and storing personal information within the parametric knowledge of

neural networks [Sukhbaatar et al., 2015, Zhang et al., 2018a]. Textual representations, including KGs and PTKBs, offer higher transparency, control, and scrutability [Radlinski et al., 2022], while neural representations could provide easier optimization. Each approach has its advantages and disadvantages, and therefore further research is needed to determine the optimal approach for storing personal information for effective personalization in CIA systems.

Privacy-Aware Personal Information Extraction. Another important future direction involves addressing privacy concerns related to personal information. With the growing popularity of agentic information-access systems [Trippas et al., 2025, Hoveyda et al., 2025], which autonomously perform tasks on behalf of users, these systems are increasingly exposed to sensitive privacy information. For instance, when a system accesses a user's medical records, it may inadvertently expose and store sensitive information about that user. Therefore, future research should explore strategies for determining when personal information should *not* be stored while still maximizing the benefits of personalization.

6.4.2 Personalized Response Generation

Optimizing Personal Information Utilization. While Chapters 2, 3, and 4 have explored approaches to extract and utilize personal information, such as personal knowledge graphs (PKGs), a significant challenge remains in determining *how* and *when* to effectively retrieve and apply this PKG information within conversational systems. In the current LLM paradigm, conversational systems typically retrieve relevant PKG details and naively insert them into the model's input context [Aliannejadi et al., 2024]. Future research could investigate more dynamic methods, potentially employing techniques similar to reinforcement learning from human feedback (RLHF) [Ouyang et al., 2022] to train models that selectively retrieve and utilize PKG information based on conversational context and user preferences, thereby enhancing responses to better align with user needs.

Factuality and Transparency. While CIA systems may generate responses that users find appealing, the prevalence of factual fabrications [Zamani et al., 2023, Joko et al., 2026] and sycophantic

behavior [Amirshahi et al., 2025, Sharma et al., 2024] in LLMs poses significant challenges to ensuring factual accuracy and transparency. For instance, a system that generates an incorrect cure for serious diseases may receive positive user feedback despite being potentially harmful. Similarly, a system that produces biased responses based on users' personal information might elicit favorable reactions while undermining fairness principles. One potential solution involves implementing uncertainty estimation techniques [Kadavath et al., 2022], which assign confidence scores to generated responses, thereby enabling users to assess the reliability of the information they receive. However, current uncertainty estimation methods have proven unreliable when external knowledge is incorporated into LLM contexts [Soudani et al., 2025], a common scenario in CIA systems. Consequently, future research should prioritize developing more robust uncertainty estimation methods that enable CIA systems to provide confidence scores alongside their responses, thus enhancing factuality and transparency while empowering users to make informed decisions.

6.4.3 Conversation Evaluation

Automatic Evaluation of Personalized CIA Systems. In Chapter 5, we proposed FACE, a fine-grained, aspect-based evaluation method for conversational systems. However, as mentioned in the limitation section (Section 6.3), the evaluation was limited to CRSs and only one aspect for chit-chat, therefore the expansion to cover more general CIA systems is needed. When expanding to general CIA systems, however, several challenges arise. (1) *Differing knowledge levels among users* present a significant challenge, as users possess varying degrees of background knowledge that result in different information gains from identical responses [Trippas et al., 2025]. For example, a novice user may require comprehensive explanations, whereas an expert may prefer concise yet highly technical responses. (2) *Varying user preferences and priorities*, as different users may prioritize different aspects of the response. One might prioritize comprehensiveness, while another might value conciseness or creativity. These challenges necessitate *personalized evaluation criteria*, requiring the evaluation process itself to be also personalized.

Realistic User Simulations for Evaluation. In Chapter 5, we utilized human-system dialogues to evaluate conversational systems, yet a key challenge is that collecting such dialogues, while easier than collecting expert annotations, could still require efforts. An alternative approach could involve the use of user simulations to generate user-system interactions [Bernard, 2025, Trippas et al., 2025, Verberne et al., 2015]. Existing user simulation methods, however, struggle to produce conversations that are naturally diverse, frequently suffer from biases, and have difficulty accurately reflecting real user behaviors [Balog and Zhai, 2023]. For example, as discussed in Chapter 4, LLM-simulated users tend to have lower lexical diversity compared to real humans. Therefore, future research should focus on developing realistic user simulators, with key areas to explore including: (1) ensuring a broad range of simulated conversations to capture a realistic distribution of user behaviors and preferences, while (2) mitigating biases in simulations, such as over/under-representation of certain cultures, demographics, genders, or medical conditions, and (3) maximizing the effectiveness of simulations in terms of providing better feedback for system development and evaluation.

Appendix A

Resources

This appendix provides the links to the code repositories and datasets used in this thesis. See also the Research Data Management section for more details on the research data management. Note that some repositories include both code and datasets from their respective chapters.

A.1 Code Repositories

The code repositories of the methods proposed in this thesis are publicly available at:

- **CREL** (Chapter 3): The implementation of our conversational entity linking method. The repository also includes the ConEL-2 dataset.
  - `informagi/conversational-entity-linking-2022`
- **LAPS** (Chapter 4): The implementation of our method for constructing large-scale personalized conversational datasets. The repository also includes the LAPS dataset.
  - `informagi/laps`
- **FACE** (Chapter 5): The implementation of our fine-grained, aspect-based conversation evaluation method. The repository also includes the CRSArena-Eval dataset.
  - `informagi/face`

A.2 Datasets

The datasets released as part of this thesis are publicly available at:

- **ConEL dataset** (Chapter 2): The initial conversational entity linking dataset, annotated with named entities, personal entities, and concepts, comprises a total of 125 dialogues and 821 user utterances. The repository also contains **online resources listing around 130 conversational datasets** with a detailed comparison of their characteristics.
  - `informagi/conversational-entity-linking`
- **ConEL-2 dataset** (Chapter 3): A larger-scale resource for training and evaluating conversational EL systems. This dataset extends original ConEL with 290 multi-turn conversations, comprising 1,327 total utterances, and is annotated with a total of 2,833 entities, including personal entities, concepts, and named entities.
  - `informagi/conversational-entity-linking-2022`
- **LAPS dataset** (Chapter 4): A large-scale collection of multi-session, human-written dialogues focused on personalization. The dataset spans the recipe and movie domains, containing 1,406 conversations and 11,215 extracted user preferences.
  - `informagi/laps`
- **CRSArena-Eval** (Chapter 5): A meta-evaluation dataset for automatic evaluation methods. It consists of 467 human-system dialogues containing 2,235 system turns across nine conversational recommender systems, annotated with 20,962 high-quality human judgments on seven different evaluation aspects.
  - `informagi/face`

Appendix B

Prompt Examples of FACE

This appendix provides the prompts used in our FACE method in Chapter 5.[1]

B.1 Prompts Used in FACE

Conversation Particle Generation Prompt

Dialogue History: `{dialogue_history}`
Target Assistant Turn: `{target_turn}`
User's Response: `{user_response}`

Your task is to extract conversation nuggets, which are minimal, atomic units of information or facts from the target assistant turn.

Each nugget consists of:

- `"dialogue_act"`*: one of the following labels:* `"greeting,"` `"preference elicitation,"` `"recommendation,"` `"goodbye,"` *or* `"others."`
- `"nugget_mention"`*: the atomic unit of information from the target assistant turn. [...]*
- `"user_feedback"`*: the excerpt of user feedback against the given nugget. [...]*

[1]Note that in all prompts, the term "nugget" refers to "conversation particle" as defined in the paper. All used and optimized prompts can be found in our GitHub repository.

The output must be a JSON list of nuggets. [...]

Must think step by step:

1. *Explain the dialogue history, the target assistant turn, and the user feedback.*
2. *How many conversation nuggets are found in the target assistant turn?*
3. *For each nugget, discuss the meaning of the user feedback.*
4. *Output in JSON format.*

Textual Gradient Prompt ∇

Examine the original instructions, predicted nugget score, and gold dialogue (or turn) score.

- *Based on the gold dialogue (or turn) score, is the predicted nugget score reasonable?*
- *Does original instructions describe how to use the nugget's information correctly?*
- *Necessary to edit the original instructions?*

Instruction Rewriting Prompt δ

Propose new instructions of ~50 words based on the feedback.

- *Note that the full dialogue can be changed, thus your new instructions must be general enough to handle different contexts.*
- *Note that the task is "nugget" evaluation, not "turn" or "dialogue" evaluation; thus, the new instructions should focus on how to use the nugget.*
- *Must provide "task description" and explicit "step-by-step instructions" for the nugget evaluation; in step-by-step instructions labeling each step as "Step 1," "Step 2," and so on.*
- *Break down the evaluation into smaller steps and provide a checklist ("Does the nugget...?" or "Is this nugget...?") for each step.*

...

Initial Prompt (Before Optimization)

Task description: *Given the dialogue, evaluate the quality of the target nugget based on the* `{evaluation_aspect}`*.*

Step-by-step instructions:

- ***Step 1:*** *Read the dialogue history, target nugget, and user's response.*
  - *What does the target nugget convey?*
- ***Step 2:*** *Carefully read the grading criteria.*
  - *What are the grading criteria?*
- ***Step 3:*** *Evaluate the target nugget.*
  - *Which grade should be assigned to the target nugget?*

B.2 Instructions Optimized by FACE

Here, we provide the optimized instruction examples for the dialogue-level overall impression aspect and the turn-level relevance aspect. Please note that, in the actual process, FACE optimizes **multiple** instructions for each aspect, as shown in Figure 5.1.

Optimized Instructions for Overall Impression Aspect (Dialogue-level)

Task description: *Evaluate the nugget based on its relevance, accuracy, and usefulness.*

Step-by-step instructions:

- ***Step 1:*** *Check if the nugget is relevant to the conversation.*
  - *Does the nugget relate to the dialogue context?*
  - *Is the nugget a direct response to the user's question or concern?*
  - *Is the nugget related to the user's preferences or interests?*
- ***Step 2:*** *Evaluate the nugget's accuracy.*
  - *Is the information in the nugget accurate based on the dialogue?*
  - *Does the nugget correctly represent the conversation?*

- ***Step 3:*** *Assess the nugget's usefulness.*
  - *Does the nugget provide a helpful or relevant suggestion?*
  - *Does the nugget address the user's needs or concerns?*
  - *Does the nugget facilitate a meaningful continuation of the conversation?*

Optimized Instructions for Relevance Aspect (Turn-level Aspect)

Task description: *Evaluate the quality of the target nugget based on its relevance to the user's request.*

Step-by-step instructions:

- ***Step 1:*** *Identify the user's request and the nugget's suggestion.*
  - *Step 1.1: Does the nugget's suggestion directly address the user's request?*
  - *Step 1.2: Is the nugget's genre or category aligned with the user's interest?*
- ***Step 2:*** *Assess the nugget's relevance.*
  - *Step 2.1: Does the nugget's information accurately address the user's need?*
  - *Step 2.2: Is the nugget's suggestion consistent with the user's preferences or interests?*

Bibliography

Z. Abbasiantaeb, S. Lupart, L. Azzopardi, J. Dalton, and M. Aliannejadi. Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

O. Adjali, R. Besançon, O. Ferret, H. Le Borgne, and B. Grau. Building a Multimodal Entity Linking Dataset From Tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

C. Akbulut, L. Weidinger, A. Manzini, I. Gabriel, and V. Rieser. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

M. Alaofi, N. Arabzadeh, C. L. Clarke, and M. Sanderson. Generative Information Retrieval Evaluation. *arXiv preprint arXiv:2404.08137*, 2024.

M. Aliannejadi, Z. Abbasiantaeb, S. Chatterjee, J. Dalton, and L. Azzopardi. TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview. *arXiv preprint arXiv:2401.01330*, 2024.

M. Aliannejadi, S. Lupart, M. Gohsen, Z. Abbasiantaeb, N. Mirzakhmedova, J. Kiesel, and J. Dalton. TREC iKAT 2025: The Interactive Knowledge Assistance Track Overview. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*, 2025.

S. Amirshahi, A. Bigdeli, C. L. Clarke, and A. Ghenai. Evaluating the Robustness of Retrieval-Augmented Generation to Adversarial Evidence in the Health Domain. *arXiv preprint arXiv:2509.03787*, 2025.

A. Anand, L. Cavedon, H. Joho, M. Sanderson, and B. Stein. Conversational Search. *Dagstuhl Reports*, 2019.

R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, and S. Chappidi. Open-Domain Question Answering Goes Conversational via Question Rewriting. *arXiv preprint arXiv:2010.04898*, 2020.

G. Angeli, M. J. Johnson Premkumar, and C. D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015.

A. Askari, R. Petcu, C. Meng, M. Aliannejadi, A. Abolghasemi, E. Kanoulas, and S. Verberne. SOLID: Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking Dialogs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

S. Bae, D. Kwak, S. Kim, D. Ham, S. Kang, S.-W. Lee, and W. Park. Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

K. Balog. *Entity-Oriented Search*, volume 39. 2018. ISBN 978-3-319-93933-9.

K. Balog and T. Kenter. Personal Knowledge Graphs: A Research Agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, 2019.

K. Balog and C. Zhai. User Simulation for Evaluating Information Access Systems. *arXiv preprint arXiv:2306.08550*, 2023.

K. Balog, D. Metzler, and Z. Qin. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer. *arXiv:2004.05150*, 2020.

N. Bernard and K. Balog. MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 23–27, 2023, Taipei, Taiwan*, 2023.

N. Bernard and K. Balog. Limitations of Current Evaluation Practices for Conversational Recommender Systems and the Potential of User Simulation. In *Proceedings of the 2025 International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2025.

N. Bernard, H. Joko, F. Hasibi, and K. Balog. CRS Arena: Crowdsourced Benchmarking of Conversational Recommender Systems. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 2025.

N. M. E. Bernard. *User Simulation for the Development and Evaluation of Conversational Information Access Agents*. PhD thesis, University of Stavanger, 2025.

P. Bhargava, N. Spasojevic, S. Ellinger, A. Rao, A. Menon, S. Fuhrmann, and G. Hu. Learning to Map Wikidata Entities To Predefined Topics. 2019.

S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc.", 2009.

K. Bowden, J. Wu, S. Oraby, A. Misra, and M. Walker. SlugNERDS: A Named Entity Recognition Tool for Open Domain Dialogue Systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

P. Braslavski, D. Savenkov, E. Agichtein, and A. Dubatovka. What Do You Mean Exactly? Analyzing Clarification Questions in CQA. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, 2017.

P. Brown and S. C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge university press, 1987.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020.

P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, B. Goodrich, D. Duckworth, S. Yavuz, A. Dubey, K.-Y. Kim, and A. Cedilnik. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

N. D. Cao, G. Izacard, S. Riedel, and F. Petroni. Autoregressive Entity Retrieval. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

D. Carmel, M.-W. Chang, E. Gabrilovich, B.-j. P. P. Hsu, and K. Wang. ERD' 14 : Entity Recognition and Disambiguation Challenge. *SIGIR Forum*, 48, 2014.

D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

L. Chen, J. Chen, T. Goldstein, H. Huang, and T. Zhou. InstructZero: Efficient Instruction Optimization for Black-Box Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

M. Chen, A. Papangelis, C. Tao, S. Kim, A. Rosenbaum, Y. Liu, Z. Yu, and D. Hakkani-Tur. PLACES: Prompting Language Models for Social Conversation Synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, 2023.

Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. Towards Knowledge-Based Recommender Dialog System. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

W. Chen, S. Koenig, and B. Dilkina. RePrompt: Planning by Automatic Prompt Engineering for Large Language Models Agents. *arXiv preprint arXiv:2406.11132*, 2024b.

Y.-H. Chen and J. D. Choi. Character Identification on Multiparty Conversation: Identifying Mentions of Characters in TV Shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016.

Y. Cheng, K. Mao, Z. Zhao, G. Dong, H. Qian, Y. Wu, T. Sakai, J.-R. Wen, and Z. Dou. CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proceedings of the 2024 International Conference on Machine Learning*, 2024.

E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

K. Christakopoulou, F. Radlinski, and K. Hofmann. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*, 2022.

J. Chung, E. Kamar, and S. Amershi. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

C. L. A. Clarke and L. Dietz. LLM-Based Relevance Assessment Still Can't Replace Human Relevance Assessment. *arXiv preprint arXiv:2412.17156*, 2024.

J. Cook, T. Rocktäschel, J. Foerster, D. Aumiller, and A. Wang. TICKing All the Boxes: Generated Checklists Improve LLM Evaluation and Generation. In *Workshop on Language Gamification at NeurIPS*, 2024.

M. Cornolti, P. Ferragina, M. Ciaramita, S. Rüd, and H. Schütze. SMAPH: A Piggyback Approach for Entity-Linking in Web Queries. *ACM Trans. Inf. Syst.*, 37, 2019.

L. Cui, F. Wei, and M. Zhou. Neural Open Information Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

J. Dalton, L. Dietz, and J. Allan. Entity Query Feature Expansion Using Knowledge Base Links. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014.

J. Dalton, C. Xiong, and J. Callan. TREC CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of TREC '19*, 2019.

J. Dalton, C. Xiong, and J. Callan. TREC Conversational Assistance Track (CAsT). `https://github.com/daltonj/treccastweb`, 2020.

A. Dargahi Nobari, A. Askari, F. Hasibi, and M. Neshati. Query Understanding via Entity Attribute Identification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

F. Dell'Acqua, E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*, September 2023.

J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, 54, 2021.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

S. Dey and M. S. Desarkar. Dial-M: A Masking-based Framework for Dialogue Evaluation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2023.

L. Dietz. A Workbench for Autograding Retrieve/Generate Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

L. Dietz, O. Zendel, P. Bailey, C. Clarke, E. Cotterill, J. Dalton, F. Hasibi, M. Sanderson, and N. Craswell. LLM-Evaluation Tropes: Perspectives on the Validity of LLM-Evaluations. *arXiv preprint arXiv:2504.19076*, 2025.

E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations (ICLR)*, 2019.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Y. Dubois, P. Liang, and T. B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. In *Conference on Language Modeling (COLM)*, 2024.

C. Eickhoff and A. P. de Vries. How Crowdsourcable is Your Task? In *Proc. of CSDM '11*, pages 11–14, 2011.

M. Ekstrand-Abueg, V. Pavlu, M. Kato, T. Sakai, T. Yamamoto, and M. Iwata. Exploring Semi-Automatic Nugget Extraction for Japanese One Click Access Evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013.

M. Eric, L. Krishnan, F. Charette, and C. D. Manning. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017.

M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, A. Kumar, A. Goyal, P. Ku, and D. Hakkani-Tur. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

A. Fader, S. Soderland, and O. Etzioni. Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 2023.

J. Fainberg, B. Krause, M. Dobre, M. Damonte, E. Kahembwe, D. Duma, B. Webber, and F. Fancellu. Talking to Myself: Self-Dialogues as Data for Conversational Agents. *arXiv preprint arXiv:1809.06641*, 2018.

H. Fang, S. Tao, N. Chen, K.-X. Chang, and T. Sakai. Do Large Language Models Favor Recent Content? A Study on Recency Bias in LLM-Based Reranking. *arXiv preprint arXiv:2509.11353*, 2025.

N. Farzi and L. Dietz. Exam++: LLM-based Answerability Metrics for IR Evaluation. In *Proceedings of LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval*, 2024.

C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

P. Ferragina and U. Scaiella. TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010.

D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 2010.

S. E. Finch, J. D. Finch, and J. D. Choi. Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

J. Gao, M. Galley, and L. Li. Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval*, 13, 2019.

E. Gerritse, F. Hasibi, and A. De Vries. Graph-Embedding Empowered Entity Retrieval. In *Proceedings of the 42nd European Conference on Information Retrieval (ECIR)*, 2020a.

E. J. Gerritse, F. Hasibi, and A. P. de Vries. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, 2020b.

E. J. Gerritse, F. Hasibi, and A. P. de Vries. Entity-Aware Transformers for Entity Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

F. Gilardi, M. Alizadeh, and M. Kubli. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 2023.

K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, and D. Hakkani-Tur. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proceedings of Interspeech 2019*, 2019.

F. Hasibi, K. Balog, and S. E. Bratsberg. Entity Linking in Queries: Tasks and Evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, 2015.

F. Hasibi, K. Balog, and S. E. Bratsberg. Exploiting Entity Linking in Queries for Entity Retrieval. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*, 2016.

F. Hasibi, K. Balog, and S. E. Bratsberg. Dynamic Factual Summaries for Entity Cards. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017a.

F. Hasibi, K. Balog, and S. E. Bratsberg. Entity Linking in Queries: Efficiency vs. Effectiveness. In *Proceedings of 39th European Conference on Information Retrieval*, 2017b.

C. Hauff. Conversational IR. `https://github.com/chauff/conversationalIR`, 2020. Online; accessed October 2020].

J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

M. Hoveyda, A. Vries, F. Hasibi, and M. de Rijke. Real World Conversational Entity Linking Requires More Than Zero-Shots. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

M. Hoveyda, H. Oosterhuis, A. P. de Vries, M. de Rijke, and F. Hasibi. Adaptive Orchestration of Modular Generative Information Access Systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL*, 2006.

P. Jandaghi, X. Sheng, X. Bai, J. Pujara, and H. Sidahmed. Faithful Persona-based Conversational Dataset Generation with Large Language Models. *arXiv preprint arXiv:2312.10007*, 2023.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 2023.

H. Joho, L. Cavedon, J. Arguello, M. Shokouhi, and F. Radlinski. First International Workshop on Conversational Approaches to Information Retrieval (CAIR'17). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

H. Joko and F. Hasibi. Personal Entity, Concept, and Named Entity Linking in Conversations. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2022.

H. Joko and F. Hasibi. FACE: A Fine-Grained Reference-Free Evaluator for Conversational Information Access. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026.

H. Joko, F. Hasibi, K. Balog, and A. P. de Vries. Conversational Entity Linking: Problem Definition and Datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

H. Joko, E. J. Gerritse, F. Hasibi, and A. P. de Vries. Radboud University at TREC CAsT 2021. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*, 2022.

H. Joko, S. Chatterjee, A. Ramsay, A. P. de Vries, J. Dalton, and F. Hasibi. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

H. Joko, S. Amirshahi, C. L. A. Clarke, and F. Hasibi. WildClaims: Information Access Conversations in the Wild(Chat). In *Proceedings of the 48th European Conference on Information Retrieval*, 2026.

C. K. Joshi, F. Mi, and B. Faltings. Personalization in Goal-Oriented Dialog. In *Proceedings of the NeurIPS'17 Workshop on Conversational AI*, 2017.

M. Joshi, O. Levy, L. Zettlemoyer, and D. Weld. BERT for Coreference Resolution: Baselines and Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022.

H. Kim, J. Hessel, L. Jiang, P. West, X. Lu, Y. Yu, P. Zhou, R. Le Bras, M. Alikhani, G. Kim, M. Sap, and Y. Choi. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. *arXiv preprint arXiv:2212.10465*, 2023.

M. Kim, C. Kim, Y. H. Song, S.-w. Hwang, and J. Yeo. BotsTalk: Machine-sourced Framework for Automatic Curation of Large-scale Multi-skill Dialogue Datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Y. Kirstain, O. Ram, and O. Levy. Coreference Resolution without Span Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

W. Kong, S. Hombaiah, M. Zhang, Q. Mei, and M. Bendersky. PRewrite: Prompt Rewriting with Reinforcement Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

I. Kostric, K. Balog, and F. Radlinski. Soliciting User Preferences in Conversational Recommender Systems via Usage-related Questions. In *Proceedings of the Fifteenth ACM Conference on Recommender Systems*, 2021.

I. Kostric, K. Balog, and F. Radlinski. Generating Usage-related Questions for Preference Elicitation in Conversational Recommender Systems. *ACM Trans. Recomm. Syst.*, 2023.

B. Krause, M. Damonte, M. Dobre, D. Duma, J. Fainberg, F. Fancellu, E. Kahembwe, J. Cheng, and B. Webber. Edina: Building an Open Domain Socialbot with Self-Dialogues. *arXiv preprint arXiv:1709.09816*, 2017.

V. Kumar and J. Callan. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

M. T. R. Laskar, C. Chen, A. Martsinovich, J. Johnston, X.-Y. Fu, S. B. Tn, and S. Corston-Oliver. BLINK with Elasticsearch for Efficient Entity Linking in Business Conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

K. Lee, L. He, and L. Zettlemoyer. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

Y.-J. Lee, C.-G. Lim, Y. Choi, J.-H. Lm, and H.-J. Choi. PERSONACHATGEN: Generating Personalized Dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, 2022.

M. Leszczynski, R. Ganti, S. Zhang, K. Balog, F. Radlinski, F. Pereira, and A. T. Chaganty. Talk the Walk: Synthetic Data Generation for Conversational Music Recommendation. *arXiv preprint arXiv:2301.11489*, 2023.

B. Z. Li, S. Min, S. Iyer, Y. Mehdad, and W.-t. Yih. Efficient One-Pass End-to-End Entity Linking for Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

B. Z. Li, A. Tamkin, N. Goodman, and J. Andreas. Eliciting Human Preferences with Language Models. *arXiv preprint arXiv:2310.11589*, 2023.

J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

M. Li, J. Weston, and S. Roller. Acute-Eval: Improved Dialogue Evaluation with Optimized Questions and Multi-Turn Comparisons. *arXiv preprint arXiv:1909.03087*, 2019.

R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. Towards Deep Conversational Recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

X. Li, G. Tur, D. Hakkani-Tür, and Q. Li. Personal Knowledge Graph Population from User Utterances in Conversational Understanding. In *2014 IEEE Spoken Language Technology Workshop*, 2014.

Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 2017.

C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 2004.

J. Lin and D. Demner-Fushman. Automatically Evaluating Answers to Definition Questions. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

Y.-T. Lin and Y.-N. Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI*, 2023.

P. Lison and J. Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the Middle: How Language Models Use Long Contexts. *arXiv preprint arXiv:2307.03172*, 2023a.

Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

L. Luo, W. Huang, Q. Zeng, Z. Nie, and X. Sun. Learning Personalized End-to-End Goal-Oriented Dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

A. Madaan, N. Tandon, P. Clark, and Y. Yang. Memory-Assisted Prompt Editing to Improve GPT-3 after Deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint arXiv:2303.17651*, 2023.

A. Manzoor and D. Jannach. Towards Retrieval-Based Conversational Recommendation. *Information Systems*, 2022.

J. Mayfield, D. Lawrie, P. McNamee, and D. W. Oard. Building a Cross-Language Entity Linking Collection in Twenty-One Languages. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation*, 2011.

J. Mayfield, E. Yang, D. Lawrie, S. MacAvaney, P. McNamee, D. W. Oard, L. Soldaini, I. Soboroff, O. Weller, E. Kayi, K. Sanders, M. Mason, and N. Hibbler. On the Evaluation of Machine-Generated Reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

S. Mehri and M. Eskenazi. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

E. Meij, W. Weerkamp, and M. De Rijke. Adding Semantics to Microblog Posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 2012.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476*, 2017.

S. Moon, P. Shah, A. Kumar, and R. Subba. OpenDialKG: Explainable Conversational Reasoning with Attention-Based Walks over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 2022.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

P. S. Park, P. Schoenegger, and C. Zhu. Diminished Diversity-of-Thought in a Standard Large Language Model. *arXiv preprint arXiv:2302.07267*, 2023.

G. Penha, A. Balan, and C. Hauff. Introducing MANtIS: A Novel Multi-Domain Information Seeking Dialogues Dataset. *arXiv preprint arXiv:1912.04639*, 2019.

M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

F. Piccinno and P. Ferragina. From TagME to WAT: A New Entity Annotator. In *Proceedings of the First International Workshop on Entity Recognition and Disambiguation*, 2014.

N. Poerner, U. Waltinger, and H. Schütze. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, and J. Lin. Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. *arXiv preprint arXiv:2411.09607*, 2024.

R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, and J. Lin. The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, 2025.

R. Pryzant, D. Iter, J. Li, Y. Lee, C. Zhu, and M. Zeng. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019.

F. Radlinski, K. Balog, F. Diaz, L. Dixon, and B. Wedin. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 2020.

J. Raiman and O. Raiman. DeepType: Multilingual Entity Linking by Neural Type System Evolution. In *AAAI*, 2018.

S. Rajput, V. Pavlu, P. B. Golbus, and J. A. Aslam. A Nugget-Based Test Collection Construction Paradigm. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2011.

H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 2020.

A. Razavi, M. Soltangheis, N. Arabzadeh, S. Salamat, M. Zihayat, and E. Bagheri. Benchmarking Prompt Sensitivity in Large Language Models. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III*, 2025.

S. Reddy, D. Chen, and C. D. Manning. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7, 2019.

E. Reif, M. Kahng, and S. Petridis. Visualizing Linguistic Diversity of Text Datasets Synthesized by Large Language Models. *arXiv preprint arXiv:2305.11364*, 2023.

C. Riesbeck. Failure-Driven Reminding for Incremental Learning. In *IJCAI*, 1981.

T. Sakai. Fairness-based Evaluation of Conversational Search: A Pilot Study. In *Proceedings of the Tenth International Workshop on Evaluating Information Access (EVIA 2023), a Satellite Workshop of the NTCIR-17 Conference*, 2023a.

T. Sakai. SWAN: A Generic Framework for Auditing Textual Conversational Systems. *arXiv preprint arXiv:2305.08290*, 2023b.

A. Salemi and H. Zamani. Learning from Natural Language Feedback for Personalized Question Answering. *arXiv preprint arXiv:2508.10695*, 2025.

A. Salemi, S. Mysore, M. Bendersky, and H. Zamani. LaMP: When Large Language Models Meet Personalization. *arXiv preprint arXiv:2304.11406*, 2024.

J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of Age and Gender on Blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

P. Shah, D. Hakkani-Tür, B. Liu, and G. Tür. Bootstrapping a Neural Conversational Agent with Dialogue Self-Play, Crowdsourcing and On-Line Reinforcement Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

M. Shang, T. Wang, M. Eric, J. Chen, J. Wang, M. Welch, T. Deng, A. Grewal, H. Wang, Y. Liu, Y. Liu, and D. Hakkani-Tur. Entity Resolution in Open-domain Conversations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. Bowman, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, et al. Towards Understanding Sycophancy in Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

S. Singh, A. Subramanya, F. Pereira, and A. McCallum. Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

C. Siro, M. Aliannejadi, and M. de Rijke. Understanding User Satisfaction with Task-oriented Dialogue Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

C. Siro, M. Aliannejadi, and M. De Rijke. Understanding and Predicting User Satisfaction with Conversational Recommender Systems. *ACM Trans. Inf. Syst.*, 2023.

E. Smith, O. Hsu, R. Qian, S. Roller, Y.-L. Boureau, and J. Weston. Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, 2022.

H. Soudani, E. Kanoulas, and F. Hasibi. Data Augmentation for Conversational AI. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023.

H. Soudani, E. Kanoulas, and F. Hasibi. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2024a.

H. Soudani, R. Petcu, E. Kanoulas, and F. Hasibi. A Survey on Recent Advances in Conversational Data Generation. *arXiv preprint arXiv:2405.13003*, 2024b.

H. Soudani, E. Kanoulas, and F. Hasibi. Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

A. Speggiorin, J. Dalton, and A. Leuski. TaskMAD: A Platform for Multimodal Task-Centric Knowledge-Grounded Conversational Experimentation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Y. Sun and Y. Zhang. Conversational Recommender System. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

R. Takehi, A. Watanabe, and T. Sakai. Open-Domain Dialogue Quality Evaluation: Deriving Nugget-level Scores from Turn-level Scores. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2023.

R. Takehi, E. M. Voorhees, and T. Sakai. LLM-Assisted Relevance Assessments: When Should We Ask LLMs for Help? *arXiv preprint arXiv:2411.06877*, 2024.

N. Thakur, R. Pradeep, S. Upadhyay, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, and J. Lin. Assessing Support for the TREC 2024 RAG Track: A Large-Scale Comparative Study of LLM and Human Evaluations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, 2025.

P. Thomas, S. Spielman, N. Craswell, and B. Mitra. Large Language Models Can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

W. Thomson. *Popular Lectures and Addresses*, volume 1 of *Nature Series*. Macmillan and Co., 1889.

A. Tigunova, A. Yates, P. Mirza, and G. Weikum. Listening Between the Lines: Learning Personal Attributes from Conversations. In *Proceedings of the World Wide Web Conference*, 2019.

A. Tigunova, A. Yates, P. Mirza, and G. Weikum. CHARM: Inferring Personal Attributes from Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

J. R. Trippas, J. S. Culpepper, et al. Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *SIGIR Forum*, 2025.

S. Upadhyay, E. Kamalloo, and J. Lin. LLMs Can Patch Up Missing Relevance Judgments in Evaluation. *arXiv preprint arXiv:2405.04727*, 2024a.

S. Upadhyay, R. Pradeep, N. Thakur, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, and J. Lin. A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look. *arXiv preprint arXiv:2411.08275*, 2024b.

S. Upadhyay, R. Pradeep, N. Thakur, D. Campos, N. Craswell, I. Soboroff, and J. Lin. A Large-Scale Study of Relevance Assessments with Large Language Models Using UMBRELA. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, 2025.

R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL: General Entity Annotator Benchmarking Framework. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

S. Vakulenko, M. de Rijke, M. Cochez, V. Savenkov, and A. Polleres. Measuring Semantic Coherence of a Conversation. In *The Semantic Web – ISWC 2018*, 2018.

J. M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. P. de Vries. REL: An Entity Linker Standing on the Shoulders of Giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All You Need. *Advances in neural information processing systems*, 2017.

S. Verberne, M. Sappelli, K. Järvelin, and W. Kraaij. User Simulations for Interactive Search: Evaluating Personalized Query Suggestion. In *Advances in Information Retrieval*, 2015.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text REtrieval Conference*, 2003.

E. M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

T.-C. Wang, S.-Y. Su, and Y.-N. Chen. Barcor: Towards a Unified Framework for Conversational Recommendation Systems. *arXiv preprint arXiv:2203.14257*, 2022a.

X. Wang, K. Zhou, J.-R. Wen, and W. X. Zhao. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022b.

X. Wang, X. Tang, X. Zhao, J. Wang, and J.-R. Wen. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023a.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023b.

Y. Wang, H. Ivison, P. Dasigi, J. Hessel, T. Khot, K. R. Chandu, D. Wadden, K. MacMillan, N. A. Smith, I. Beltagy, and H. Hajishirzi. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023c.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

C. Xiong, J. Callan, and T.-Y. Liu. Word-Entity Duet Representations for Document Ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

C. Xu, D. Guo, N. Duan, and J. McAuley. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. *arXiv preprint arXiv:2304.01196*, 2023.

L. Xu and J. D. Choi. Online Coreference Resolution for Dialogue Processing: Improving Mention-Linking on Real-Time Conversations. *arXiv preprint arXiv:2205.10670*, 2022.

W. Xu, G. Zhu, X. Zhao, L. Pan, L. Li, and W. Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

I. Yamada, R. Tamaki, H. Shindo, and Y. Takefuji. Studio Ousia's Quiz Bowl Question Answering System. In *Proceedings of the NIPS '17 Competition: Building Intelligent Systems*, 2018.

I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen. Large Language Models as Optimizers. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

A. Yates, M. Banko, M. Broadhead, M. Cafarella, O. Etzioni, and S. Soderland. TextRunner: Open Information Extraction on the Web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2007.

M. Yazan, S. Verberne, and F. Situmeang. Improving RAG for Personalization with Author Features and Contrastive Examples. In *Advances in Information Retrieval*, 2025.

Q. Ye, M. Ahmed, R. Pryzant, and F. Khani. Prompt Engineering a Prompt Engineer. In *Findings of the Association for Computational Linguistics ACL 2024*, 2024.

Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. Ratner, R. Krishna, J. Shen, and C. Zhang. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. *arXiv preprint arXiv:2306.15895*, 2023.

M. Yuksekgonul, F. Bianchi, J. Boen, S. Liu, Z. Huang, C. Guestrin, and J. Zou. Textgrad: Automatic "Differentiation" via Text. *arXiv preprint arXiv:2406.07496*, 2024.

L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, and Z. You. Chatdoctor: A Medical Chat Model Fine-Tuned on Llama Model Using Medical Domain Knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

H. Zamani, J. R. Trippas, J. Dalton, and F. Radlinski. Conversational Information Seeking. *Foundations and Trends® in Information Retrieval*, 2023.

X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, and J. Chen. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020.

J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018a.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto. Benchmarking Large Language Models for News Summarization. *arXiv preprint arXiv:2301.13848*, 2023.

W. Zhang, W. Hua, and K. Stratos. EntQA: Entity Linking as Question Answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. In *Advances in Neural Information Processing Systems*, 2018b.

Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

C. Zhao, T. Yu, Z. Xie, and S. Li. Knowledge-aware Conversational Preference Elicitation with Bandit Feedback. In *Proceedings of the ACM Web Conference 2022*, 2022.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

L. Zheng, L. Yin, Z. Xie, C. L. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. Gonzalez, C. Barrett, and Y. Sheng. SGLang: Efficient Execution of Structured Language Model Programs. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. LIMA: Less Is More for Alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2023a.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large Language Models are Human-Level Prompt Engineers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023b.

Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

Summary

Conversational interactions have reshaped information retrieval systems, as users increasingly favor direct answers over traditional hyperlinks. This paradigm shift has reinforced the importance of conversational information access (CIA) systems that ground responses in retrieved knowledge to provide reliable information. To build more personalized and reliable CIA systems, this thesis addresses key challenges by focusing on three areas: (1) personal context extraction, (2) personalized response generation, and (3) effective and interpretable system evaluation.

First, we tackle personal context extraction. We introduce the ConEL dataset and define the problem of Entity Linking (EL) in conversations. We find that traditional tools struggle with the unique challenge of identifying personal entities and concepts. To address this, we propose CREL, a novel EL method tailored for this setting. We find that CREL significantly outperforms these tools, accurately identifying all entity types vital for personalization.

Second, we focus on personalized response generation. Effectively utilizing extracted personal context to generate personalized responses remains challenging without large-scale dialogue datasets capturing realistic user preferences across sessions. We introduce LAPS, a scalable method using large language models (LLMs) to guide human workers in composing high-quality dialogues that capture diverse, real user preferences. Using the dataset from LAPS, we investigate how personal preferences can be utilized for personalizing responses. We show that using semi-structured preference memory, compared to appending raw conversation history, improves the accuracy of preference utilization and explanations in responses while mitigating LLMs' recall issues.

Finally, we address the need for effective and interpretable system evaluation. We introduce FACE, an automatic, reference-free evalua-

tion method that provides a fine-grained, aspect-based assessment of entire conversations. It decomposes system responses into conversation particles and uses an LLM with optimized instructions to generate interpretable scores. We find that FACE's evaluations align closely with human judgments, offering actionable insights while outperforming existing evaluation methods.

In this thesis, alongside methodological contributions, we place a significant focus on resource contributions. Specifically, we release the ConEL, LAPS, and CRSArena-Eval datasets for conversational entity linking, training personalized dialogue systems, and the meta-evaluation of conversation evaluators, respectively. These resource contributions, coupled with the method and theoretical advancements the thesis introduces, further the field of personalized CIA research.

Samenvatting

Conversationele interacties hebben informatie-retrieval-systemen opnieuw vormgegeven, aangezien gebruikers in toenemende mate de voorkeur geven aan directe antwoorden boven traditionele hyperlinks. Deze paradigmaverschuiving heeft het belang versterkt van conversationele informatietoegangssystemen (CIA) die hun antwoorden baseren op opgehaalde kennis om betrouwbare informatie te bieden. Om meer gepersonaliseerde en betrouwbare CIA-systemen te bouwen, behandelt dit proefschrift belangrijke uitdagingen door zich te richten op drie gebieden: (1) extractie van persoonlijke context, (2) gepersonaliseerde responsgeneratie, en (3) effectieve en interpreteerbare systeemevaluatie.

Ten eerste pakken we de extractie van persoonlijke context aan. We introduceren de ConEL-dataset en definiëren het probleem van Entity Linking (EL) in conversaties. We constateren dat traditionele tools moeite hebben met de unieke uitdaging van het identificeren van persoonlijke entiteiten en concepten. Om dit aan te pakken, stellen we CREL voor, een nieuwe EL-methode die is toegesneden op deze setting. We constateren dat CREL deze tools significant overtreft en alle entiteitstypen die vitaal zijn voor personalisatie nauwkeurig identificeert.

Ten tweede richten we ons op gepersonaliseerde responsgeneratie. Het effectief benutten van geëxtraheerde persoonlijke context om gepersonaliseerde antwoorden te genereren blijft uitdagend zonder grootschalige dialoogdatasets die realistische gebruikersvoorkeuren over sessies heen vastleggen. We introduceren LAPS, een schaalbare methode die grote taalmodellen (LLMs) gebruikt om menselijke werkers te begeleiden bij het samenstellen van hoogwaardige dialogen die diverse, echte gebruikersvoorkeuren vastleggen. Met de dataset van LAPS onderzoeken we hoe persoonlijke voorkeuren kunnen worden benut voor het personaliseren van antwoorden. We tonen aan dat het ge-

bruik van semi-gestructureerd voorkeursgeheugen, vergeleken met het toevoegen van ruwe conversatiegeschiedenis, de nauwkeurigheid van voorkeursbenutting en verklaringen in antwoorden verbetert, terwijl het de recall-problemen van LLMs mitigeert.

Ten slotte behandelen we de behoefte aan effectieve en interpreteerbare systeemevaluatie. We introduceren FACE, een automatische, referentievrije evaluatiemethode die een fijnmazige, aspect-gebaseerde beoordeling van gehele conversaties biedt. Het ontleedt systeemresponsen in conversatiedeeltjes en gebruikt een LLM met geoptimaliseerde instructies om interpreteerbare scores te genereren. We constateren dat de evaluaties van FACE nauw overeenkomen met menselijke oordelen, waarbij het actiegerichte inzichten biedt en bestaande evaluatiemethoden overtreft.

In dit proefschrift leggen we, naast methodologische bijdragen, een significante nadruk op bijdragen aan hulpbronnen. Specifiek publiceren we de datasets ConEL, LAPS en CRSArena-Eval voor respectievelijk conversationele entity linking, het trainen van gepersonaliseerde dialoogsystemen, en de meta-evaluatie van conversatie-evaluatoren. Deze bijdragen aan hulpbronnen, gekoppeld aan de methode en theoretische vorderingen die het proefschrift introduceert, bevorderen het vakgebied van gepersonaliseerd CIA-onderzoek.

Acknowledgements

Looking back, leaving my stable job and established life to start over abroad, so soon after the birth of my child, was a bold decision. First and foremost, I want to thank my wife, who supported me at every step of this journey with love and care. My son traveled the world because of me, yet he remained energetic and brought me joy, inspiration, and many opportunities to learn together; thank you for sharing this journey with me.

Academic life: Faegheh, you are the one who made this journey possible. I came to Radboud with great ambition but little knowledge of how to write a world-class paper, and more importantly, of the true meaning of contributing to the scientific community. You have always been kind and supportive throughout my PhD journey, and from you, I have learned how to conduct research, write a strong paper, and embrace the mindset of passion and perseverance. I am truly fortunate to have you as my supervisor, and I would like to express a huge thank you for your support and guidance.

Arjen, thank you for always supporting me behind the scenes, offering guidance whenever I was stuck, and helping me when I needed it most. I clearly remember how, during the Glasgow collaboration, your nudge helped us push through difficult challenges, resulting in an impactful SIGIR paper. You supported me in many other moments as well, always in such an understated way; thank you.

Lastly, I would like to acknowledge the valuable feedback from the thesis committee. Thank you for taking the time to read my thesis and provide your comments. After incorporating your feedback, I am confident that the quality of the thesis has been greatly improved. I am very grateful for your help and support.

Collaborations: I was fortunate to have multiple amazing cross-country collaborations, including with Krisztian, Jeff, Charlie, and many others. These experiences taught me the power of collaboration, and I am deeply grateful for them. Furthermore, I was fortunate to visit and live in other countries, which broadened my perspectives and deepened my understanding of different cultures. Thank you, Dyaa, Jiyin, and Raymond, for accepting me as a Visiting Researcher at Signal AI, where I learned about R&D in a startup as well as life in London, and thank you, Charlie, for hosting me as a Visiting Scholar at the University of Waterloo.

Colleagues and friends: My PhD journey would not have been the same without the good friends and colleagues around me. Whenever I came to the office, I always found someone for a casual chat that sparked my curiosity and motivated my research. The opportunity to nerd out about NLP and IR, and to share my interests and findings with you all, has always been a pleasure. I cannot name everyone here (as it would be a long list), but thank you for always being with me and motivating my research.

Netherlands: Throughout the years, I met and interacted with many people in the Netherlands. Having lived there with my family for nearly five years, we naturally faced many challenges, however, we were very fortunate to receive support from many people. I already feel nostalgic about the Netherlands, its amazing people, its beautiful nature (e.g., Keukenhof and Scheveningen Beach), and more importantly the atmosphere of the country. I feel that the Netherlands is my second home, and I would love to visit again whenever I can.

Research Data Management

The research datasets produced during this PhD research have been managed according to the research data management policy of the Institute for Computing and Information Sciences of Radboud University, the Netherlands,[2] and are listed as follows:

- Resources for Chapter 2:
  - ConEL: Conversational Entity Linking Datasets:
    informagi/conel: Zenodo. 10.5281/zenodo.15315077
- Resources for Chapter 3:
  - CREL and ConEL-2: Tool and Dataset for Conversational Entity Linking:
    informagi/crel: Zenodo. 10.5281/zenodo.15320433
- Resources for Chapter 4:
  - LAPS: LLM-Augmented Personalized Self-Dialogue:
    informagi/laps: Zenodo. 10.5281/zenodo.15320611
- Resources for Chapter 5:
  - FACE and CRSArena-Eval: Automatic Evaluator and Meta-Evaluation Dataset for Conversational Information Access:
    informagi/face: Zenodo. 10.5281/zenodo.17231386

[2] `https://www.ru.nl/rdm/`, last accessed January 28th, 2025

Curriculum Vitae

Hideaki Joko, born in Yokohama, Kanagawa, Japan, has maintained a deep interest in understanding how humans process complex natural language information and how these mechanisms can be applied to computational systems. To pursue this interest, he obtained a Bachelor's degree in Applied Mathematics from Waseda University in 2014 and a Master's degree from the University of Tokyo in 2016, through his research on Natural Language Processing (NLP). Following his studies, he served as a Research Engineer at Mitsubishi Electric in Japan from 2016 to 2020, where he researched and developed NLP and Information Retrieval (IR) methods for industrial applications, resulting in multiple patents and awards for his contributions.

In September 2020, he embarked on his PhD journey in the Data Science group within the Institute for Computing and Information Sciences (iCIS) at Radboud University in the Netherlands, where he focused on the research area of conversational information-access systems. During his PhD, he published his research at multiple international conferences such as SIGIR, CIKM, WSDM, and ECIR. From June to October 2022, he served as a Visiting Researcher at Signal AI in London, UK, where he developed an NLP method to enhance risk intelligence through sentiment analysis. He also conducted multiple international collaborations with the University of Glasgow and the University of Stavanger. Towards the end of his PhD, he served as a Visiting Scholar at the University of Waterloo, where he worked with Professor Charles Clarke on conversational information access in the real-world setting.

He currently works as an Applied Scientist at Thomson Reuters Labs, where he focuses on translating academic research in NLP and IR into industrial applications to deliver conversational information-access systems in the real world.

SIKS Dissertation Series

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT
30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TU/e), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OU), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TU/e), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
31 Gongjin Lan (VUA), Learning better – From Baby to Better
32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

---

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

---

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

---

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies

15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions

48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
50 Lea Krause (VUA), Contextualising Conversational AI
51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models
52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems
54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law
57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply
59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support
60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement
61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications
62 Khaleel Asyraaf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills
63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data
64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search
65 Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities

---

2026 01 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support
02 Hezha Hassan Mohammedkhan (TiU), Estimating Body Measurements of Children from 2D Images: Towards the Automatic Detection of Malnutrition
03 Kyriakos Psarakis (TUD), Democratizing Scalable Cloud Applications: Transactional Stateful Functions on Streaming Dataflows
04 Boyu Xu (UU), Exploring Indirect Relations Between Topics in Neuroscience Literature Using Augmented Reality to Inform Experimental Design
05 Koen Minartz (TU/e), Stochastic Simulation with Geometric Deep Generative Models
06 Azim Afroozeh (CWI, VUA), FastLanes: A Next-Gen File Format
07 Inès Blin (VUA), Narrative Understanding with Knowledge Graphs
08 Paul van Vulpen (UU), Debating Digital Dominance: Decentralized Technology Governance For Strategic Autonomy
09 Afrizal Doewes (TU/e), Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback
10 Nikolaos Delapaschos Kondylidis (VUA), Establishing Task-Oriented Understanding between Agents
11 Işıl Baysal Erez (UT), Handling Missing Data with Meta-Learning and Large Language Models
12 Xue Li (UvA), From Fine-tuning to Prompting: A Paradigm Shift in Knowledge Graph Construction
13 Isaac da Silva Torres (VUA), Guidelines To Flux Between Conceptual Models: Understanding Complex Digital Business Ecosystems
14 Philip Lippmann (TUD), Synthetic Data for Robust Language Modelling

15 Rashmi Khazanchi (OU), Artificial Intelligence in Education: Impact of AI-Based Systems on Mathematics Achievement
16 Carolina Ferreira Gomes Centeio Jorge (TUD), Modelling Artificial Trust for Effective Human-AI Teamwork
17 Maria Tsfasman (TUD), Towards Predicting Memory in Multimodal Group Interactions
18 Riccardo Lo Bianco (TU/e), Deep Reinforcement Learning for Automated Decision-Making in Process Management Systems
19 Israel Campero Jurado (TU/e), Innovations in Optimization and Applications in Healthcare
20 Iftitahu Ni'mah (TU/e), Contrastive Learning and Evaluation in Low Resource Scenario of Natural Language Processing
21 Francisco N.F.Q. Simoes (UU), Causality, Information, and Decision-Making
22 Ruben Verhagen (TUD), Transparent and Explainable Agents for Human-Agent Teaming
23 Eduardo Calò (UU), Automatically Expressing the Meaning of Logical Formulae in Natural Language
24 Shuai Han (UU), Improving Sample Efficiency of Reinforcement Learning: Exploiting Structural Knowledge for Decision Making
25 Francis Saa-Dittoh (VUA), From Radio to AI: African Community-Driven Development of Sustainable Information Systems
26 Mohammed Al Owayyed (TUD), Interactive Simulation-Based Learning Tools for Training Children's Helpline Counsellors
27 Bram Grooten (TU/e), Adaptive Reinforcement Learning: Lean and Dynamic Agents for Robust Generalization
28 Anouck Braggaar (TiU), Does This Answer Your question? Evaluating, Replicating, and Improving Task-Oriented Dialogue Systems
29 Afsana Khan (UM), Unlocking the Value of Data with Vertical Federated Learning
30 Yucheng Yang (TU/e), Generalization in Reinforcement Learning: Theories and Algorithms for Downstream Task and Multi-objective Trade-off Adaptation
31 Luis Pedro Silvestrin (VUA), Efficient Machine Learning for Time-Varying Data: Contributions to Time-Series Machine Learning and Transfer Learning
32 Giovanni Varricchione (UU), Declarative Specifications for Efficient and Safe Reinforcement Learning
33 Markus Funke (VUA), Carving Sustainability into Software Architecture
34 Melika Ayoughi (UvA), External and Internal Semantics for Visual Understanding
35 Clinton Cao (TUD), Modeling Behavior Patterns in Microservice Applications: Practical Applications Using Finite State Machines
36 Jonathan Kamp (VUA), Interpreting the Methods that Interpret Language Models
37 Emre Erdogan (UU), Computational Theory of Mind for Effective Hybrid Intelligence
38 Mahmoud Shokrollahi-Far (TiU), Quran Mining: Computational Stylometry of Quranic Texts
39 Aleksandr Chebykin (CWI, TUD), Efficient Evolutionary Hyperparameter Optimization for Deep Learning
40 Kateryna Ihorivna Holubinka (OU), Holistic Integration of Desktop Virtual Reality Technology in Higher Education: A Design-Based Research Approach
41 Hideaki Joko (RUN), Personalization and Evaluation of Conversational Information Access
42 Nedo Alexander Bartels (VUA), Conceptualizing Platform Revenue Models: A Taxonomy-Driven Design Approach
43 Maksim Gladyshev (UU), "Who Is to Blame?" and "What Is to Be Done?" A Formal Study of Counterfactual Approaches to Responsibility and Causation in (Multi-Agent) AI Systems
44 Iffat Fatima (VUA), Software Architecture Evaluation for Sustainability
45 Tristan Tomilin (TU/e), Scalable Benchmarking of Continual and Safe Embodied Reinforcement Learning

9 789465 152028 >

Radboud Universiteit