

**SCALABLE, HUMAN-LIKE  
ASYNCHRONOUS COMMUNICATION**

---

**DISSERTATION**

Submitted in Partial Fulfillment of  
the Requirements for  
the Degree of

**DOCTOR OF PHILOSOPHY (Computer Science)**

at the

**NEW YORK UNIVERSITY  
TANDON SCHOOL OF ENGINEERING**

by

**Alberto Maria Chierici**

**May 2023**

**SCALABLE, HUMAN-LIKE  
ASYNCHRONOUS COMMUNICATION**

---

**DISSERTATION**

**Submitted in Partial Fulfillment of**

**the Requirements for**

**the Degree of**

**DOCTOR OF PHILOSOPHY (Computer Science)**

**at the**

**NEW YORK UNIVERSITY  
TANDON SCHOOL OF ENGINEERING**

**by**

**Alberto Maria Chierici**

**May 2023**

Approved:

  
Department Chair Signature

May 10, 2023

Date

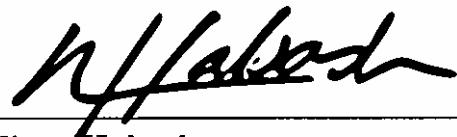
Approved by the Guidance Committee:

Major: Computer Science



**Rachel Greenstadt**  
Associate Professor,  
NYU Tandon School of Engineering

May 5, 2023  
Date



**Nizar Habash**  
Professor,  
NYU Abu Dhabi

May 5, 2023  
Date



**Torsten Suel**  
Professor,  
NYU Abu Dhabi School of Engineering

May 5, 2023  
Date



**David Traum**  
Research Professor,  
USC Viterbi School of Engineering

May 9, 2023  
Date

Microfilm or other copies of this dissertation are obtainable from

UMI Dissertation Publishing  
ProQuest CSA  
789 E. Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

## Vita

**Alberto M. Chierici** was born in Milan, Italy, in 1985. He obtained his B.Sc. in Physics from Università degli Studi di Milano (2008), Milan, Italy, and M.Sc. in Actuarial, Financial and Economic Statistics from Università degli Studi di Milano Bicocca (2011), Milan, Italy. After four years working as an actuary in the London insurance market, he became interested in machine learning. He pivoted to data science consulting work, eventually co-founding an award-winning chatbot technology company. Since July 2018, he has been pursuing his Ph.D. in the Computer Science department at New York University, Tandon School of Engineering, Brooklyn, NY, USA, under the supervision of Prof. Nizar Habash. His Ph.D. was in part funded by the Global Ph.D. Student Fellowship program at New York University Abu Dhabi, UAE. He worked as a Senior Data Scientist at Tesla in Berlin, Germany (remote) between 2021 and 2022, where he contributed to developing Safety Score Beta and the Tesla Insurance product offering in the US market. In August 2021, he published a book on artificial intelligence (AI) and ethics with New Degree Press. His research focuses mainly on Human-Computer Interaction (HCI), particularly dialogue systems, chatbots, multimedia software applications, and engineering AI software features.

## Acknowledgements

I am immensely grateful to my supervisor, Prof. Nizar Habash, for his invaluable guidance, support, and encouragement throughout my Ph.D. journey. His expertise, insight, mastery of NLP methods, linguistics, scientific methods, and personal sensitivity and understanding have shaped my research and pushed me to achieve my goals.

I would like to extend my heartfelt appreciation to my wife, Laura, my inlaws Lina and Marco, and Aunt Maria, who have been a constant source of strength and practical help during the trying times of doing a Ph.D. late in life with a family. Especially for helping out with the house chores and looking after our kids during the time I spent far from home. I thank my children Marco, Caterina, and Anastasia for enduring difficulties and showing patience towards their dad during an overly busy life span that simultaneously included scientific research, work, writing a book, and a pandemic. Their unwavering love, kindred souls, and understanding have been true blessing.

I am also deeply grateful to the many students who have helped me in various ways over the years: Bishnu Dev, Wahib Kamran, Soojin Lee, Muhammed Ali, Gautham Dinesh, Kertu Koss, Abdul Samad, Tyeece Hensley, Armaan Agrawal. Their involvement in developing the TOIA software, expanding its use cases through their fantastic capstone projects, helping in recruiting subjects for experiments, debugging, and several hours spent hacking and building together have been invaluable. I am truly grateful for their willingness to share their time and skills with me. I also want to give particular thanks to Prof. Goffredo Puccetti and Erin Collins

for their fantastic support in improving the software design and usability. Finally, I want to thank Dana Abu Ali, Muaz Ahmad, Hayat Al Hassan, Paula Dozsa, Ming Hu, and Jose Varias for their capstone project, which was the source of inspiration and the basis upon which I built my research.

I would like to express my gratitude to everyone who has made this Ph.D. possible. Although there have been many challenges and obstacles along the way, I have emerged from this experience with a sense of humility, gratitude, and confidence in the result achieved. Thank you to my dissertation guidance committee, family, grandparents, colleagues, and everyone else who has played a role in this journey. Your support and encouragement have been essential to my success, and I am grateful beyond words.

Alberto M. Chierici

May 2023

To steadfast seekers blending love and lore,  
Uniting family, knowledge, passion, and more.

## **ABSTRACT**

---

### **SCALABLE, HUMAN-LIKE ASYNCHRONOUS COMMUNICATION**

by

**Alberto Maria Chierici**

**Advisor: Prof. Nizar Habash**

**Submitted in Partial Fulfillment of the Requirements for  
the Degree of Doctor of Philosophy (Computer Science)**

**May 2023**

Time-offset interaction applications (TOIAs) are new narrative-sharing that simulate face-to-face interactions between humans and previously-recorded videos of other humans. In this dissertation, we explore the next evolution of TOIAs, designing and developing a system ideal for mass adoption while catering to individuals with differing interests, languages, and cultures. Our research expands on previous work by addressing the lack of a streamlined methodology for creating and evaluating TOIAs.

Here we show that a framework and smart software features for high-quality

prerecordings, seamless interactions between users and TOIAs, and an evaluation framework for TOIAs can effectively create good-quality interactions with limited timeframes and resources. Our contributions include an openly sourced TOIA, a library of prerecordings and corpora, a methodology for developing the dialogue manager, evaluation methodologies, two extensive Human-Computing Interaction (HCI) user studies, and guidelines for developing TOIA's interfaces. The potential of TOIAs to bridge the gap between recorded narratives and face-to-face interactions has implications for fields beyond narrative-sharing technology, such as education, therapy, and cross-cultural communication. As such, the research presented in this dissertation contributes to a growing body of work exploring the intersection of technology, storytelling, and human interaction.

The development of TOIAs can revolutionize how we interact with recorded narratives and stories. Our methodology and tools can be applied in various contexts, making this research significant for preserving cultural heritage and developing personalized and interactive storytelling. TOIAs will provide a platform for preserving cross-cultural narratives while fostering a greater understanding and appreciation of diverse perspectives.

# Table of Contents

Vita . . . . .	iv
Acknowledgements . . . . .	v
Dedication . . . . .	vii
Abstract . . . . .	viii
List of Figures . . . . .	xvii
List of Tables . . . . .	xxiii
<b>1 Introduction</b>	<b>1</b>
1.1 Contributions . . . . .	5
1.2 Thesis Organization . . . . .	7
1.3 Published Papers . . . . .	9
1.4 Definition of Terms . . . . .	10
<b>2 Literature Review</b>	<b>12</b>
2.1 Sixty Years of Dialogue Systems . . . . .	12
2.2 Techniques and Data For Developing TOIAs . . . . .	18
2.3 Creating TOIA Avatars . . . . .	21
2.4 Interacting with TOIA Avatars . . . . .	24
2.5 Evaluation Methods . . . . .	26

	xi
2.6 Outlook . . . . .	28
<b>3 Data Collection</b>	<b>31</b>
3.1 The Margarita Dialogue Corpus . . . . .	33
3.2 Dialogue Manager Baseline Models . . . . .	38
3.3 Discussion and Lessons Learned . . . . .	47
<b>4 Human Evaluation</b>	<b>52</b>
4.1 Methods . . . . .	54
4.2 Human Evaluation Results and Analysis . . . . .	58
4.3 Discussion on Human Evaluation . . . . .	67
4.4 GPT-3-based Retrieval Methodology & Evaluation . . . . .	69
4.5 Summary . . . . .	75
<b>5 TOIA Design and Development</b>	<b>77</b>
5.1 Design Principles . . . . .	80
5.2 System Interface and Backend . . . . .	82
5.3 Dialogue Manager . . . . .	87
5.4 Question Suggestion Personalization to Streamline TOIA Creation and Interaction . . . . .	88
5.5 Interaction: Creation of, and Engagement with, TOIA Avatars . . .	92
5.6 Summary . . . . .	93
<b>6 Creating TOIA Avatars</b>	<b>95</b>
6.1 User Study Methods . . . . .	98
6.2 Experimental Results and Discussion . . . . .	103
6.3 Limitations . . . . .	112

	xii
6.4    Summary . . . . .	114
<b>7    Interacting with TOIA Avatars</b>	<b>116</b>
7.1    User Study Methods . . . . .	117
7.2    Experimental Results . . . . .	125
7.3    Discussion . . . . .	148
7.4    Summary . . . . .	151
<b>8    Conclusion and Future Directions</b>	<b>152</b>
8.1    Summary and Contributions . . . . .	152
8.2    Future Directions . . . . .	154
<b>A    Supplementary Materials</b>	<b>167</b>

# List of Figures

4.1	Initial ten turns of a casual dialogue between an interrogator and a TOIA avatar using our best-performing retrieval model. . . . .	55
4.2	Annotation interface displayed to the human annotators. . . . .	56
4.3	Examples in the dev set where the BERT q-A 1:100 system choice differs from the Avatar Maker’s. . . . .	64
4.4	On the Player interface, a pop-up appears after every answer is played. The interactor has to rate the answer before going ahead with asking the next question. . . . .	71
5.1	Screenshots of the user interface. (a) is the user account page, which shows the questions generated by QS in white backgrounds and previously recorded questions (and videos) shaded in blue; (b) shows the recorder page; (c) shows suggestions appearing in a pop-up window once the user completes a recording on the Recorder page; and (d) shows the player page. . . . .	81
5.2	A view of a number of public TOIA streams that are ready for interactors to converse with. . . . .	82

5.3	System architecture. In bold and starred, the components we improved from the system proposed in [11]. . . . .	85
5.4	A view of the page showed at the first login of a user. The white cards underlined with a red line are mandatory and the user cannot record other videos before finishing this initial set of recordings. . .	93
6.1	Meaningful correlations between personality traits and user experience dimensions. Only correlations with p-values < 0.05 are shown. Left: correlations are computed using data from all 3 conditions. Right: correlations are computed using data corresponding to the GPT-3-QS and Human-QS cohorts. . . . .	110
7.1	Likert scale distributions across all the survey answers that required an answer on a scale from 1 (low, poor) to 5 (high, great). This chart excludes personality scores that are discussed later. . . . .	127
7.2	Spearman $\rho$ correlations between all quantitative (and categorical transformed to quantitative) answers to the post-interaction survey questions. We only show correlations where they are statistically significant (p-value < 0.05). . . . .	128
7.3	Likert scale distributions split by the different question suggestion setups. The ‘gpt3’, ‘manual’ and ‘off’ setups indicate groups of <i>TOIA avatars</i> that were created with the Question Suggester automated with the GPT-3 implementation, manually run by a researcher, and disabled, respectively. . . . .	131

7.4	Likert scale distributions split by the groups of variants with a different number of videos in their databases. The ‘30’, ‘60’, and ‘71+’ labels indicate groups of <i>TOIA avatar</i> variants that were created by selecting the first 30, 60, and all the videos recorded by the <i>TOIA avatar</i> -makers, respectively. . . . .	132
7.5	Likert scale distributions split by the groups of variants with different fillers played between interactions. ‘Inattentive’ and ‘Attentive’ are self-explanatory labels indicating the groups of <i>TOIA avatars</i> variants that were created by including only inattentive or attentive fillers, respectively. . . . .	133
7.6	Average perceived and self-reported Big-5 personality dimensions (OCEAN), displayed in the following order: overall, by QS setup, by <i>TOIA avatar</i> variants based on video database size, and by filler type variants. . . . .	143
7.7	Spearman $\rho$ correlations (displayed only for p-value>0.05) between self-disclosed and perceived <i>TOIA avatars</i> ’ Big-5 personality traits. Panels (a) to (c) present all data and data split by inattentive and attentive filler variants, respectively. Correlations appear consistent across cohorts, suggesting that the experimental condition regarding filler videos does not significantly influence the differences between self-disclosed and perceived personality dimensions. . . . .	144

7.8 Spearman $\rho$ correlations (displayed only for p-value>0.05) between self-disclosed and perceived <i>TOIA avatars</i> ' Big-5 personality traits. Panels (a) to (c) present data split by cohorts with GPT-3 QS, manual QS, and QS turned off, respectively. The variation in correlations may indicate that the experimental condition related to QS could potentially influence differences between self-disclosed and perceived personality dimensions. . . . .	145
7.9 Spearman $\rho$ correlations (displayed only for p-value>0.05) between self-disclosed and perceived <i>TOIA avatars</i> ' Big-5 personality traits. Panels (a) to (c) present data split by variants with all videos, the first 60 videos, and the first 30 videos, respectively. The variation in correlations may indicate that the experimental condition based on the number of videos could impact personality perception, particularly for the 30-video cohort. . . . .	146
8.1 A researcher using our TOIA for interacting with Warren Buffet (left) and Elon Musk (right). . . . .	157
8.2 The “Elephant in the Room:” a multimedia art installation using our TOIA and a custom-made user interface for building interactive experiences about sensitive topics: on the left, the initial screen presenting the topics available; on the right, an interaction is engaging with <i>TOIA avatars</i> talking about death; on the bottom, the physical installation at an exposition space in the Art Center of NYU Abu Dhabi. . . . .	159

8.3 Screenshots of new developments using our TOIA for building multilingual interaction: on the left, an interactor asks a question in Arabic (see the Arabic automatic transcription on the bottom); on the right, the <i>TOIA avatar</i> answers in English (the language setting shows on the top-left corner an American flag indicating the US-English language as default) with a subtitle in Arabic. . . . .	163
A.1 Distribution of trigram prefixes of <b>QUESTIONS</b> in the Margarita Dialogue Corpus' <b>Knowledge Base</b> . . . . .	168
A.2 Distribution of trigram prefixes of <b>questions</b> in the Margarita Dialogue Corpus' <b>dialogue DEV sets</b> . . . . .	169
A.3 Distribution of trigram prefixes of <b>questions</b> in the Margarita Dialogue Corpus' <b>dialogue TEST sets</b> . . . . .	170
A.4 Distribution of trigram prefixes of <b>ANSWERS</b> in the Margarita Dialogue Corpus' <b>Knowledge Base</b> . . . . .	171

# List of Tables

3.1	Three examples of pairs in the Margarita Dialogue Corpus KB data resource. . . . .	32
3.2	Summary of the categories in the Margarita Dialogue Corpus Knowledge Base (KB) defined by the <i>TOIA avatar</i> -maker. . . . .	32
3.3	Summary statistics on the two main data sets in the Margarita Dialogue Corpus: Knowledge Base (KB) and Dialogues (Dial). Statistics for the dialogues are also shown for the training portion vs. test and university mode (EDU) vs. personal mode (PER). . . . .	33
3.4	Summary statistics after sampling non-answers for the training set: down-sampled KB (KB') and up-sampled training dialogues. The training set remained the same with minor changes due to answers no longer being present in the KB'. . . . .	41
3.5	Recall@1 statistics for each baseline evaluated on the training set. The metrics corresponding to the thresholds that were automatically selected are bolded. . . . .	42

3.6 Thresholding considerations for the BERT model. The table shows, for each threshold level imposed to cosine similarities, the number of correct answers predicted, the number of correct non-answers predicted by the model, hence the true positive rate for answers (TPR-ans), the true positive rate for non-answers (TPR-non-ans) and the Recall@1 metric. The automatic choice of threshold for this setup is highlighted in boldface.	42
3.7 Results for each baseline on the test set. For each model and threshold selection, the Recall@k metric is shown for different levels of $k$ on the test set.	43
3.8 Recall@k only for answers (i.e., ignoring questions in the test set that did not have an answer in the KB) for each baseline. For each model, the Recall@k metric is shown for different levels of $k$ on the test set.	44
3.9 Top ranked question similarity for some meaningful examples in the test set for all three baselines.	45
4.1 Inter-annotator agreement computed using Cohen’s kappa score ( $\kappa$ ) for the dev set and the test set.	59
4.2 Average ratings assigned to the gold and top retrieval model choices in dev and test sets. %Gold specifies the ratio of model average rating to gold average rating. Rank specifies the performance rank of the retrieval model.	60
4.3 Spearman’s Rank Correlation Coefficient ( $\rho$ ) between each retrieval model and the human ratings for the dev set and the test set.	61

4.4	Information Retrieval metrics on the <b>dev</b> dialogues set for all the models, including a random selection model and using the crowd ratings as a retrieval model. The models are assessed against the original annotations made by the avatar maker. On the two right-most columns, the models are assessed against the annotations from the crowd and the newest retrieval implementation using GPT-3, respectively. We keep the best results of the older models in bold, as the GPT-3 implementation was added later than the analysis presented here.	62
4.5	Same metrics as Table 4.4, but here the models are assessed against the annotations from the crowd. We keep the best results of the older models in bold, as the GPT-3 implementation was added later than the analysis presented here.	62
4.6	SR@ <i>k</i> metrics on the <b>test</b> set only for the best performing models on SR@1 and SR@10. We keep the best results of the older models in bold, as the GPT-3 implementation was added later than the analysis presented here.	65
4.7	Distribution of interactors' ratings by mode of interaction from the conversation log data of our TOIA.	72
4.8	Inter-annotator agreement computed using coefficients of agreement that are all relevant in our scenarios where we have multiple raters using ordinal ratings.	75
6.1	Participants demographics by cohort.	102

6.2	Summary of highlighted experimental results. All metrics exclude onboarding videos (fillers and mandatory question-answer pairs). The number of questions may be larger than the number of videos linked to such questions because users could record one video as an answer to multiple questions (e.g., answering ‘yes’ or ‘no’). We highlight in bold noteworthy results that we discuss more in detail in section 6.2 and mark with ** the statistically significant results when tested against the QS-Off cohort pairwise at a significance level of 0.05 and with * at a significance level of 0.1. . . . .	104
6.3	Participants average ‘Big-5’ personality scores by cohort. . . . .	105
7.1	Interaction participants demographics. . . . .	119
7.2	Summary statistics, group and pairwise statistical tests testing the difference between groups based on the <b>suggestion setup</b> . Reporting only the survey questions where we find p-values below the significance level of 0.05. . . . .	130
7.3	Summary statistics, group and pairwise statistical tests testing the difference between groups based on the <b>no. of videos in the database</b> . Reporting only the survey questions where we find p-values below the significance level of 0.05. . . . .	134
7.4	A mixed-effect linear model for the enjoyment rating and its dependence on all three experimental conditions. . . . .	140

7.5	A mixed-effect linear model for the enjoyment rating and its dependence on the two experimental conditions that show statistical significance. The combined cohort (QS-on) of GPT-3-QS and Human-QS. The complementary group is QS-off. The length group consists of <i>TOIA avatar</i> variants with either 60 or 71+ videos in the database (high) vs. the variants with only 30 videos (low). . . . .	141
7.6	A mixed-effect linear model illustrating the interaction effect between perceived conscientiousness and the combined cohort (QS-on) of GPT-3-QS and Human-QS. The complementary group is QS-off. The length group consists of <i>TOIA avatar</i> variants with either 60 or 71+ videos in the database (high) vs. the variants with only 30 videos (low). C <sub>-</sub> , E <sub>-</sub> , and A_perceived stand for perceived Conscientiousness, Extraversion, and Agreeableness, respectively. The equation at the top of the table specifies the full model. . . . .	141
7.7	Distribution statistics of the Root Mean Square Difference between perceived and self-reported Big-5 personality dimensions (RMSEpd) by <i>TOIA avatar</i> -variant cohort. From left to right, the number of observations, mean, standard deviation, minimum value, 25th percentile, median, 75th percentile, and maximum value. . . . .	147
A.1	Summary statistics on the data sets collected in the user study on the avatar creation (CFR, Chapter 6). Statistics for the various <i>TOIA avatars'</i> knowledge bases are also shown for each video-type and by the experimental condition co-hort (Question Suggester powered by GPT-3, by a human, and switched off). . . . .	172

A.2 Summary statistics on the dialogues collected from the interaction user study's chat logs (CFR, Chapter 7). Statistics are also shown for each type of interaction with the player interface (CARD, SEARCH, VOICE).	173
A.3 Questions and type of answer collected in the post-experiment survey issued to <i>TOIA avatar</i> -makers. The full experiment is described in Chapter 6.	174
A.4 (1/2) Questions and type of answer collected in the post-experiment survey issued to participants who interacted with the <i>TOIA avatars</i> . The full experiment is described in Chapter 7.	175
A.5 (2/2) from Table A.4.	176
A.6 Ten-item personality inventory questionnaire to establish the Big-5 personality dimensions (OCEAN).	177

# Chapter 1

## Introduction

In the rapidly evolving field of computer science, the quest to create a world where computer programs can seamlessly replicate human interactions and effectively retrieve information to further the interaction has become an increasingly vital pursuit. As we strive to share narratives and authentic experiences on a global scale, the potential for such advancements extends far beyond the realms of academic research, with numerous commercial applications, such as reducing the need for expert workers (e.g., general medical practitioners, counselors) and developing highly efficient chatbots (e.g., booking a flight within a few conversational turns).

The central idea guiding this thesis is ambitious: what if we could design computer programs capable of replacing cognitively engaging conversations, effectively bridging the gap between artificial intelligence (or artificial avatars) and human interaction?

People love to share stories about themselves. Stories remain the most perva-

sive medium people use to make sense of themselves and the surrounding world. In the past twenty years, social media development has enabled people to document narratives of their daily experiences online, in public and restricted domains, through tweets, blogs, and Instagram stories, in unprecedented measure [52]. These modern-day methods are only the latest incarnation in our long history, evolving from pictures on cave walls to orally preserved poetry epics to journals, diaries, and scrapbooks.

In this dissertation, we want to explore the next evolution of narrative-sharing devices exploiting advances in Artificial Intelligence (AI) and peoples' increasing comfort with virtual communication methods such as Zoom to simulate face-to-face interactions between humans and previously-recorded videos of other humans.

We call such systems ‘Time-Offset Interaction Applications’ (TOIAs), as the interaction happens after the actual time of video creation. A TOIA is a relatively new interface that allows interactive conversations with participants who are not currently available for interaction but have previously recorded relevant utterances. Time-offset interaction enhances the experience of viewing essentially unidirectional communication by introducing the elements of familiarity, presence, and interactiveness of one-to-one communication. TOIAs lie between Question Answering (QA) and Information Retrieval (IR). They differ from QA in that a TOIA’s task is not about demonstrating comprehension of a text span but selecting a single (one-shot) appropriate answer from a restricted set of answers, a problem also known as Answer Retrieval (AR) or retrieval-based dialogue.

To contextualize this endeavor, we will delve into the historical development of human-computer interaction, with a particular focus on the emergence of the groundbreaking New Dimensions in Testimony project. Artstein et al. (2014)[6] developed the earliest forms of a TOIA at the University of Southern California’s Institute for Creative Technologies (USC’s ICT). They implemented a proof-of-concept using the likeness of Pinchas Gutter, a Holocaust survivor, and subsequently installed the final product in a museum (Traum et al., 2015 [75]). The time-offset interaction of Mr. Gutter enables short conversations about his family, his religious views, and life in a museum setting.

Building on top of the work at USC, at New York University Abu Dhabi (NYUAD), Abu Ali et al. (2018)[1] built a bilingual English and Arabic prototype. Its two main components were a recorder and a player. Using this system, some students who wanted to preserve a digital version of themselves, or act like a different character, could do so by first recording several short video clips of themselves answering various questions about their life. Once done, other individuals could converse with the pre-recorded video databases by asking natural questions. The best video clip answering the interrogator’s question would play back, giving a sense of face-to-face interaction without the other person being present on the other side. This initial prototype focused on providing bilingual English-Arabic support to the system. It provided an interface for creators and interactors to test the making of different characters: a fictional, historical persona, a university kiosk answering questions about the campus, and a person freely talking about herself.

A number of relatively new technology companies developed similar systems

for commercial use. Typeform's videoask,<sup>1</sup> Digital Humans,<sup>2</sup> and Soul Machines<sup>3</sup> are some examples. TOIAs may also be reminiscent of virtual assistants like Siri and Alexa and digitally animated characters like SGT Blackwell (Leuski et al., 2006)[32]; however, these are not authentic representations of human beings which is TOIAs' goal. The concept has also appeared in many science fiction and fantasy movies, including *Superman*,<sup>4</sup> and its following sequels, where Superman can ask questions and receive instructions from his long-dead father; *I, Robot*,<sup>5</sup> in which detective Del Spooner can question a holographic representation of the recently deceased Dr. Alfred Lanning; and *Batman & Robin*,<sup>6</sup> where Barbara Wilson (aka Batgirl) can speak with a video recording of her uncle Alfred while he has entered a coma.

TOIAs can be employed on various fronts, including but not limited to story-sharing that preserves cross-generational and cross-cultural histories, online teaching that simulates a classroom environment, simulating interviews, conducting museum tours, and similar contexts. Recent advancements in AI, cloud computing, storage capacity, and video streaming make it possible to develop such systems in an AI-enabled conversational style on a much larger scale than ever done before—perhaps thousands of such conversations per minute. We explore more of these potential use cases in the conclusive remarks of this work (Chapter 8).

However, as we move forward in our quest for more sophisticated human-

---

<sup>1</sup>[videoask.com](http://videoask.com)

<sup>2</sup>[digitalhumans.com](http://digitalhumans.com)

<sup>3</sup>[soulmachines.com](http://soulmachines.com)

<sup>4</sup>Donner, Richard. *Superman*. Warner Bros., 1978

<sup>5</sup>Proyas, Alex. *I, Robot*. Twentieth Century, 2004

<sup>6</sup>Schumacher, Joel. *Batman & Robin*. Warner Bros., 1997

computer interactions, there is a critical need to address the resource-intensive nature of existing solutions. This thesis addresses this gap by underscoring the importance of developing less resource-intensive programs for such interactions. The research scope of this thesis encompasses exploring the components and critical challenges involved in creating a program that can achieve this ambitious goal.

Our work’s importance extends beyond potential academic impacts. We aim to make time-offset interaction technology more accessible by developing a framework and methodology, addressing potential issues, and laying the groundwork for user-friendly applications. This thesis seeks to identify optimal configurations and process generalizations to create time-offset interaction experiences that accommodate diverse interests, languages, and cultures. Consequently, our contributions have broader implications for the tech industry and society as a whole.

## 1.1 Contributions

Making a TOIA fit for mass adoption raises many questions about computational linguistic modeling, human-computer interfaces, dialogue systems, AI ethics, and governance. Building such systems is not trivial because interaction with TOIAs should ideally be as close as possible to the natural, human-to-human conversation. This critical feature poses many challenges from an engineering perspective (e.g., connecting the video clips flawlessly) and from many other points of view related to answer retrieval and natural language processing. There is not yet an ideal methodology for the process of creating a time-offset interaction experience. Likewise, there is no established way to evaluate the different components of a

TOIA and their user satisfaction.

This thesis aims to identify ideal setups and generalizations of the processes to build time-offset interaction experiences catering to individuals with differing interests, languages, and cultures in a controlled context. To make the ambitious goal of this work more attainable and manageable in the context of a Doctoral dissertation, we contribute towards building foundational work (open source software, data sets, data analyses, human subject studies, and evaluations) one hand. On the other hand, we restrict ourselves to studying the particular interaction of meeting a stranger for the first time and entertaining a conversation for 10-15 minutes to learn more about this person while also making the interactions seem as natural as possible. Working on a dialogue system that is not only text-based like traditional chatbots (or answer retrieval tasks) adds other dimensions and complexities to the interactions. Specifically, this thesis contributes to these four research questions:

1. How can we optimize the experience of creating a TOIA's video recordings database? In particular,
  - (a) Shall we ask the user to brainstorm questions or find ways to elicit stories by letting the system propose meaningful questions to ask the *TOIA avatar* makers?
  - (b) How many videos are enough to record for a single *TOIA avatar*?
2. How can we make the interaction seamless? In particular,
  - (a) How to retrieve the best video answers to an interrogator's utterance;
  - (b) What elements in the interface can aid the interrogator in having a meaningful experience?

3. How do we evaluate a TOIA? In particular,
  - (a) The answer retrieval component;
  - (b) The user experience for both the TOIA recorder and player.
4. To what extent can the user know and perceive the real person behind the database of pre-recorded videos?

The contributions of this thesis, therefore, include:

- An openly-sourced TOIA, available for the research community.
- A library of video databases and corpora.
- A methodology for developing interactions on the player interface, including answer retrieval and smart question-suggester components.
- An evaluation methodology for the answer retrieval in the context of TOIAs and its different perceptions by humans.
- Two extensive user studies evaluating the experience and effect of personality traits on creating and engaging with time-offset interaction applications.

## 1.2 Thesis Organization

This thesis aims to provide a comprehensive and compelling exploration of the critical aspects of designing computer programs that can seamlessly replicate human interactions and lay down the use cases and future uses of such advancements. This work is structured across eight chapters, including this Introduction.

Chapter 2, the literature review, provides a historical perspective on dialogue systems spanning over sixty years. We delve into techniques and data for developing TOIAs, the process of creating *TOIA avatars*, their interactions, and methods for evaluation.

Chapter 3 focuses on data collection as initial means for developing a good process for creating time-offset interaction without the need to develop the entire applications. We introduce the Margarita Dialogue Corpus (MDC) and discuss the development of answer retrieval baseline models. We then engage in a thorough discussion of learned lessons and limitations.

In Chapter 4, we examine a notional evaluation with a proof of concept study based on the MDC and conversation snippets in text format rated by Amazon Mechanical Turk annotators. We detail the methods employed and present the results and analysis. We discuss the current retrieval methodology and its evaluation before summarizing the chapter’s key findings.

Chapter 5 delves into TOIA design and development, presenting the guiding design principles, system interface, back-end, and dialogue manager of the TOIA built as part of this thesis contribution. The design and software development choices are built upon the learnings from the previous two chapters. For example, we detail the final retrieved methodology we decided to implement and the development of a new critical feature: question suggestion personalization to streamline the *TOIA avatar* creation and interaction.

Chapter 6 discusses the process of creating a *TOIA avatar*, detailing the user study that evaluates the overall experience. We examine the evaluation of the feature that suggests questions, providing insights into its effectiveness and potential improvements.

Chapter 7 examines interacting with *TOIA avatars*, outlining user study methods, experimental results, and discussing the learnings. Here, we close the loop with some experimental conditions created as part of the first user study presented in the previous chapter and combine data on personality traits to analyze the personality differences between the real person behind the video databases and their perception by the people who interacted with them.

Finally, Chapter 8 concludes the thesis, revisiting the research questions, summarizing the contributions made, and discussing the implications of our work for the broader field of computer science and the future of human-computer interaction. We explore the exceptional experiences time offset interaction can unlock. We discuss the benefits of time offset interaction and the research implications and potential. We also briefly address privacy and ethical considerations.

### 1.3 Published Papers

The work developed for this dissertation resulted in the publications listed below, with more under preparation and review.

- Chierici, Alberto, Nizar Habash, and Margarita Bicec. “The Margarita

Dialogue Corpus: A Data Set for Time-Offset Interactions and Unstructured Dialogue Systems.” In Proceedings of The 12th Language Resources and Evaluation Conference, pp. 476-484. 2020.

- Chierici, Alberto, and Nizar Habash. “A View From The Crowd: Evaluation Challenges for Time-Offset Interaction Applications.” In Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pp. 75-85. 2021.
- Chierici, Alberto, Tyeece Hensley, Wahib Kamran, Kertu Koss, Armaan Agrawal, Erin Meekhof, Goffredo Puccetti, and Nizar Habash . “A Cloud-based User-Centered Time-Offset Interaction Application.” In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 265–268, 2021.

## 1.4 Definition of Terms

**TOIA** A Time-offset Interaction Application refers to software that makes real-time synchronous conversational interaction possible with an individual who is physically absent and not simultaneously engaged with the other interaction party [1, 10, 73, 74].

**TOIA Avatar** The term *avatar* strongly connects to CGI-generated synthetic characters. We use the term *TOIA avatar* to differentiate it. We indicate with this term, the individual (persona) who is physically absent in the interaction.

**TOIA Avatar Maker** A *TOIA avatar* maker is a user of a TOIA who creates a *TOIA avatar* by recording videos of themselves, transcribing questions, and the answers such videos reply to.

**Interactor** We call “interactor” or “interrogator” a user who interacts, speaks, and asks questions to a *TOIA avatar*.

# Chapter 2

## Literature Review

### 2.1 Sixty Years of Dialogue Systems

When drawing a history of dialogue systems, the most common milestone often described is the ELIZA chatbot (Weizenbaum, 1966) [80]. ELIZA simulated a Rogerian psychologist whose methods involve drawing the patient out by playing the patient’s statements back at them. This use case demonstrated how a simple chatbot can seem quite complex. This is because Rogerian psychology uses an uncommon type of dialogue in which, as Weizenbaum describes, one can “assume the pose of knowing almost nothing of the real world.” For example, when the patient would say something like, “I like riding bikes,” and the psychiatrist replies, “Tell me about bikes,” you do not assume the analyst did not know what a bike is but instead assume she had some conversational objective. Most chatbots in the early days chose a domain with similar properties for trying to pass the Turing test (Turing, 1950) [77].

ELIZA defined the first wave of dialogue systems, known as rule-based chatbots. ELIZA had a lexicon of keywords that were assigned ranks. More specific words got a higher rank vs. more generic words. For each utterance emitted by an ELIZA user, the chatbot selected the keyword with the highest rank and played back a sentence.

ELIZA's reply was defined according to the user utterance pattern recognized by regular expressions. For example, a pattern like "<\*> YOU <word> ME" was associated with the transformation (rule) "WHAT MAKES YOU THINK I <word> YOU." Something like "Sometimes you tease me" would get ELIZA playing back, "What makes you think I tease you." If there were no keyword matches, the chatbot could select a non-committal answer like "PLEASE GO ON," "THAT'S VERY INTERESTING," or "I SEE." Finally, ELIZA had an astute memory ability: whenever the word 'my' was the highest-ranked keyword, it would randomly pick a transform on the MEMORY list, apply it to the utterance, and save it. Later in the conversation, if no keyword matches a user utterance, ELIZA randomly picks a non-committal answer or returns to a topic using the top of the MEMORY queue. So, if at some point in the chat, one would say to ELIZA, "My cat keeps me company," later on, if ELIZA did not have any answers to playback, would either say, "I SEE," or "EARLIER YOU SAID YOUR CAT KEEPS YOU COMPANY."

As recounted by Jurafsky and Martin (2019) [28], "People became deeply emotionally involved with the ELIZA program. Weizenbaum tells the story of one of his staff who would ask Weizenbaum to leave the room when she talked with ELIZA. When Weizenbaum suggested that he might want to store all the ELIZA conversa-

tions for later analysis, people immediately pointed out the privacy implications, which suggested that they were having quite private conversations with ELIZA, despite knowing that it was just software.”

Modern chatbots like ALICE work thanks to updated versions of ELIZA’s pattern/transform rule set design.

Fast forward thirty years, between the ’90s and early 2000s, dialogue systems moved from rule-based architectures to corpus-based training models. Researchers and companies started collecting and annotating large conversation corpora so that they could implement text-mining or statistical learning algorithms to recognize and generate text patterns automatically. One such dataset is the Switchboard corpus of American English telephone conversations (Godfrey et al., 1992) [21]. Relatively recent systems train on movie dialogues (Danescu et al., 2011 [15], Lison and Tiedemann, 2016 [35], among others), which mirror natural communication in several forms (Forchini, 2013) [19].

Interacting with computers using spoken language has always been a desire cultivated by both science fiction novels and technologists. Steve Jobs’ infamous presentation of the Macintosh in 1984 marked an emotional moment in the relationship between humans and computers. Jobs said, “Now, we’ve done a lot of talking about Macintosh recently, but today for the first time ever, I’d like to let Macintosh speak for itself.” On his cue, Macintosh spoke in a digitized voice using a speech-generating program that had been built for the Apple II: “Hello, I am Macintosh. It sure is great to get out of that bag. Unaccustomed as I am to public

speaking, I'd like to share with you a maxim I thought of the first time I met an IBM mainframe: Never trust a computer you can't lift. Obviously, I can talk right now, but I'd like to sit back and listen. So, it is with considerable pride that I introduce a man who has been like a father to me: Steve Jobs." (Hertzfeld, 2004) [24]

Speech recognition software made so much progress that it became ubiquitously used with voice assistants. Apple's *Siri*, *Google Assistant*, Microsoft's *Cortana*, and Amazon's *Alexa* are the most well-known used assistants on commercial devices. At the time of writing, the most popular, general-purpose dialogue system is ChatGPT<sup>1</sup>, released in November 2022 by software company OpenAI.

Dialogue systems fall along two central axis pairs: modular versus end-to-end and structured (or task-driven) versus unstructured (or social bots). Modular systems, like the above-mentioned voice assistants, train multiple models to support a set of tasks (Guo and Seltzer, 2012) [23], while end-to-end systems train a single learning algorithm on dialogue data (Serban et al., 2016 [68], Shum et al., 2018 [70]). Similarly, structured systems assume a logical representation for the information exchanged in conversation – for example, slot-filling techniques (Williams et al., 2013 [83], Fast et al., 2018 [18]) – which unstructured systems do not require (Gao et al., 2019 [20]).

Thanks to increasing computing and processing power, videos or CGI-generated characters could be added to the system for a richer and more emotional connection with the users. About fifteen years ago, we started seeing the dialogue systems

---

<sup>1</sup><https://openai.com/blog/chatgpt>

that are known as ‘Time-Offset Interaction Applications’ (TOIAs).

### 2.1.1 Notable Time-Offset Systems

People utilize stories more than any other form of media to make sense of who they are and the world around them. Over the past twenty years, social media growth has allowed people to share narratives of their everyday lives online in previously unheard-of quantities. TOIAs investigates the subsequent development of narrative-sharing gadgets that mimic face-to-face interactions between people and previously captured movies of other people by utilizing breakthroughs in artificial intelligence and growing comfort with virtual communication techniques.

The usefulness of TOIAs has been demonstrated in many practical scenarios. For example, for keeping historical memories [75] or engaging customers visiting a website like the commercial application *videoask* proposes.<sup>2</sup> Storyfile is another commercial application for generating TOIAs.<sup>3</sup> TOIAs may also be reminiscent of virtual assistants like *Siri* and *Alexa* and digitally animated characters designed for job interview practice for young adults with developmental disabilities,<sup>4</sup> or for commercial purposes like Digital Humans<sup>5</sup> and Soul Machines;<sup>6</sup> the latter are not authentic representations of human beings, which is a TOIA goal.

The work presented in this thesis drew its initial inspiration from research at

---

<sup>2</sup>[www.videoask.com](http://www.videoask.com)

<sup>3</sup>[www.storyfile.com](http://www.storyfile.com)

<sup>4</sup><https://ict.usc.edu/prototypes/vita/>

<sup>5</sup>[www.digitalhumans.com](http://www.digitalhumans.com)

<sup>6</sup>[www.soulmachines.com](http://www.soulmachines.com)

the University of Southern California’s Institute for Creative Technologies (ICT), such as SGT Blackwell, a digitally animated army officer designed to interact with visitors at an army conference information kiosk (Leuski et al., 2006) [32]. The system recorded users’ questions via a microphone and converted speech into text through an automatic speech recognition (ASR) system. The computer then passed the text input to an answer retrieval model and selected an appropriate answer from the 83 pre-recorded lines that Blackwell could deliver.

Among the first time-offset interaction systems is a system developed by the Entertainment Technology Center of Carnegie Mellon University [43], who proposed the “Synthetic Interview” technology. The main aim of the technology was to develop a time-offset interaction with a historical persona, allowing users to interact with it as if they were interacting with the persona in real-time. The historical persona was composed using clips of an actor playing that historical character and answering questions from the user. The system was later patented and used within a few commercial projects for the creation of celebrities personas,<sup>7</sup> however, no further research was conducted on the topic. A successor time-offset interaction prototype was developed within the New Dimensions in Testimony (NDT) project, another work by ICT. The NDT uses video recordings instead of digital media. It is a technology installed at a museum allowing visitors to conduct conversations with Holocaust survivor Pinchas Gutter (Traum et al., 2015a,b [73, 75]; Artstein et al., 2015, 2016 [4, 5]). The NDT setup is remarkable regarding the number of resources that went into creating the avatar recording — hours of recording, use of top-of-the-line digital cinema cameras, etc. This work was the source of inspiration

---

<sup>7</sup><http://www.etc.cmu.edu/projects/si-studio/about.html>

for other projects [27, 51] who focused on the experience of interacting with a digital human avatar and the engineering of a fluid user experience for the player, necessarily involving high production costs.

Few commercial systems exist as well. For instance, *videoask*<sup>8</sup> is designed for sales and marketing purposes, offering a different way to interact with a website visitor. *Storyfile* is another commercial application for generating TOIAs,<sup>9</sup> which seems to focus on senior people who want to leave recorded memories to their relatives and to posterity. Such systems are developed by private institutions and are designed for narrow use cases that make them unsuitable for research purposes. Moreover, the restricted set of functionalities, the use cases they are designed for, lack of publications, and open-source repositories make it challenging to use these systems for conducting user experience research, as it is not possible to integrate a user activity tracker as well as to have complete control over the critical components needed for conducting such research.

## 2.2 Techniques and Data For Developing TOIAs

A dialogue system like a TOIA formulates a new category of chatbots that can be named ‘self-narrative bots’ (SNB). In terms of NLP tasks, they are a middle ground between social bots and task-driven bots: they may use a combination of structured and unstructured data for training. SNBs can understand an interrogator’s question to match it to a sensible answer. At the same time, they must be able to engage in a multi-turn conversation. Hence they may use reading comprehension

---

<sup>8</sup>[www.videoask.com](http://www.videoask.com)

<sup>9</sup>[www.storyfile.com](http://www.storyfile.com)

of contexts represented by a sequence of pairs rather than a standalone question. For such a system, an interesting approach would be a combination of algorithms and evaluations from the Information Retrieval (IR) context and the sequential approach (i.e., learning the sequence of a conversation).

Social bots are expected to entertain the user and hence are often evaluated by the number of turns they can make in conversations (Khatri et al., 2018) [29].<sup>10</sup> They are designed to address two primary Natural Language Processing (NLP) challenges: understanding the user question, and generating a sensible answer. From the IR standpoint, they must also search into a large enough KB to retrieve the right answer(s) across many different topics. These KBs usually come from different contexts, such as Twitter or Wikipedia pages (Ritter et al., 2010 [61], Wilcock et al., 2012 [82]).

Some dialogue management systems use question-answering techniques, which usually address a reading comprehension task. For instance, work on the SQuAD dataset (Rajpurkar et al., 2016) [57], the Ubuntu dialogue corpus (Lowe et al., 2015) [39], and bAbI<sup>11</sup> (Weston et al., 2015) [81] are designed to perform time reasoning and inductive logic (Kumar et al., 2016) [31]. Techniques include recurrent neural network models (RNN) such as sequence2sequence, word embeddings, and LSTMs. It is worth mentioning that word or sentence pre-trained embeddings alone are a simple tool to produce powerful results. Recent examples are Google’s BERT (Devlin et al., 2018) [16], Facebook’s InferSent (Conneau et al., 2017) [14], and OpenAI’s GPT (Radford et al., 2018) [56] and all their latest evolutions into ever

---

<sup>10</sup><https://developer.amazon.com/alexaprize/challenges/current-challenge/rules>

<sup>11</sup><https://research.fb.com/downloads/babi/>

advancing Large Language Models (LLMs) like LaMDA [71], OPT [87], BLOOM [66], LLaMA [72], among others.

State-of-the-art results have been achieved very recently on Answer-Retrieval tasks using deep learning architectures (Wu et al., 2016 [85], Humeau et al., 2019 [25], Roller et al., 2020 [62]). The datasets size for the TOIA we created – and for TOIAs in general – is too small for deploying deep learning techniques, but it shouldn’t restrict us from using pre-trained LLMs for sentence similarity retrieval, leveraging transfer learning or one-, two-, few-shot learning [37].

Conversational questions have challenging phenomena not present in existing reading comprehension datasets. Recent datasets that focus on free-form human dialogues and include human annotations are CoQA (Reddy et al., 2019) [60] and HUMOD (Merdivan et al., 2020) [44]. CoQA is a large-scale reading comprehension dataset that improves a dataset like SQuAD (Rajpurkar et al., 2016) [57] by including questions that depend on conversation history and by ensuring the naturalness of answers in a conversation. HUMOD instead takes inspiration from Cornell’s movie dialogue corpus (Danescu et al., 2011) [15] by adding human annotations to it. The Douban Conversation Corpus (Wu et al., 2016) [85] contains dialogues between people sampled from Douban, a popular social network in China. The dataset is public and open domain — people chat about movies, books, music, etc.

These large-scale datasets address different tasks, whereas TOIAs usually involve much smaller datasets. A system like Traum et al. (2015a) [75]’s has a Knowledge Base (KB) of about 2,000 answers. We created and use the Margarita Dialogue

Corpus (Chierici et al., 2020) [10] for initial experiments. This corpus has a KB of 431 answers, as well as a set of complete annotated dialogues.

## 2.3 Creating TOIA Avatars

Developing, streamlining, and making the process of creating TOIAs affordable are important goals. Building such systems is not easy because the interaction with the avatars should ideally be as close as possible to a real human-to-human interaction. Still, the system is retrieval-based: if there aren't answers that have been recorded, the software would play, and that might be close to answering the question, but not entirely. Or it just won't reply to certain questions. Moreover, user interface development poses challenges from an engineering point of view (for instance, connecting the video clips seamlessly) and from many other points of view related to dialogue management and natural language processing.

The New Dimensions in Testimony project inspired Abu Ali et al. (2018) [1] to build a first version of the time-offset dialogue system following the same methodology, whereby they prepared a limited (compared to the original project's vast) set of question-answer (q-a) pairs in advance. In other words, the *TOIA avatar* makers guessed what questions might be asked of them.

Traum et al. (2015b) [73] are able to use speech recognition and language understanding to select sensible responses frequently enough to enable a natural interaction flow. However, while these results are achievable with about 2,000

recorded statements, we want to build a system that allows any person to make their avatar available in a relatively short period of time (i.e., recording significantly fewer than 2,000 statements) using their laptop and webcam. Thus generalizing the process to new users and unpredictable contexts.

To the best of our knowledge, there isn't yet a convenient methodology to create avatars for a system that is ready for streamlined adoption by everyone with internet access and a webcam.

### **2.3.1 Topic/Query Suggestions for Conversation Facilitation**

Several existing works have explored methods to suggest conversation topics to facilitate conversation in different human-machine interaction scenarios. For instance, in the context of text-based human-human interaction, Shin et al. (2021) [69] designed and developed a machine learning-based chatbot agent which automatically suggests topics of mutual interest for facilitating conversations between strangers in online settings. The topic recommendation is performed by crawling social media posts of the interaction parties and measuring semantic similarities between their mentioned keywords. A user study comparing control, random topic suggester, and the proposed chatbot groups ( $N=18$ ) validate its efficiency in increasing the conversation quality and closeness of the interaction parties.

In face-to-face settings, other methods were explored to suggest conversation topics for facilitating human-human conversations between strangers via the intermediate of wearable technology. For instance, Jarusriboonchai et al. (2015) [26]

proposed the design of a wearable display that shows textual content from the wearer’s social media profiles based on the level of proximity to another user and match-making between their contents. A qualitative user study ( $N=18$ ) confirmed the utility of such devices in contexts where people have a specific purpose of being social and getting to know strangers, such as a party or conference. Similarly, the system introduced by Nguyen et al. (2015) [49] generates real-time personalized topic suggestions during a conversation using a ranking recommendation algorithm delivered via Google Glass. A user study with 38 pairs of strangers revealed the suggestions were helpful, provided they came at the right moments. While the work presented in this dissertation relates to previous work in topic/query suggestions because it is a critical component to streamline the creation of *TOIA avatars*, none of the existing works focus on integral question suggestions to elicit open-form answers and self-narratives.

### 2.3.2 Generative Models for Conversational Recommendations

Transformer networks are a variation of neural networks developed to tackle problems like sequence prediction and machine translation. The transformer architecture has exhibited an improved language understanding as well as the flexibility to work on different tasks, especially in the form of recent state-of-art tools such as Google’s Bidirectional Encoder Representations from Transformers (BERT) and OpenAI’s GPT (Generative Pre-trained Transformer) [16, 56]. Moreover, Mishra et al. (2020) [45], in their comparison of different technical approaches to text generation, showed that transformer models perform better than previous text generation courses in terms of grammatical fluency and correctness. The success of

transformer-based generative language models has led researchers to explore the utility of such models in the context of conversational recommendations.

For instance, in the context of conversational recommender systems where the system aims to recommend high-quality items to users through interactive conversations, Ma et al. (2020) [41] used GPT-2 for response generation. Zhou et al. (2020) [88] proposed a new task of topic-guided conversational recommendation, which can be decomposed into three sub-tasks: item recommendation, topic prediction, and response generation. The response generation model is implemented based on GPT-2 to produce responses for guiding users or giving a persuasive recommendation. The success of generative models in recommendation-related tasks motivated the work presented in this thesis to explore generative models for question suggestion in the context of TOIAs.

## 2.4 Interacting with TOIA Avatars

The work of Nishiyama eta al. (2016) [51] and Jones et al. (2015) [27] focused on the experience of interacting with a digital human avatar and the engineering of a fluid user experience for the player, necessarily involving high production costs. The contribution of Nishiyama et al. describes how automultiscopic 3D displays allow many viewers to experience 3D content simultaneously without wearing special glasses or headgear. Such immersive devices are ideal for the public display of time-offset interactions. They can display life-size human subjects and allow for natural personal interactions with 3D cues such as eye gaze and spatial hand gestures. Highly connected with the latter contribution, Jones et al. (2015) propose a method for synthesizing body sway to give human-like movement to

image-based avatars. In a later work, Miyauchi et al. (2017, 2019) [46, 47] focused on implementing strategies for initiating and maintaining a user’s engagement with an image-based avatar.

#### 2.4.1 The Effects of Personality

Kim et al. (2019) [30] conducted a meta-analysis to compare the self- and informant-reported means of personality traits. The study revealed moderate to high levels of agreement between the two sources for the Big Five personality traits. This finding contradicts the general belief of the self-enhancement effect, as the self-report means did not differ significantly from the informant-report means. However, when the authors compared self-reports with stranger reports, moderate mean differences were found, indicating that people tend to be more critical of unacquainted individuals. These results have significant implications for personality assessment and other fields where self-enhancement motives are relevant.

Gosling et al. (2011) [22] examined the manifestations of personality in online social networks, specifically looking at self-reported Facebook-related behaviors and observable profile information. They found that certain personality traits, such as extraversion and openness, were associated with more active and diverse online behavior, while others, such as neuroticism, were associated with less active and more self-disclosing behavior. The authors suggest that online social networks can provide valuable insights into an individual’s personality and highlight the importance of considering self-reported and observable behaviors in personality assessment.

Intriguing insights can be gleaned from studies on self-reported and observed personality traits, particularly in the context of TOIAs. These studies offer a valuable perspective on measuring personality differences between perceived and reported traits. By examining the extent to which the TOIA medium influences the perception of a *TOIA avatar*-maker’s personality, we can better understand how to design and develop more personalized and engaging TOIA experiences.

Furthermore, these findings can inform the development of tailored interactions that better meet the needs and preferences of individual users. Understanding the role of personality in TOIA user experiences is crucial for optimizing the platform to deliver more engaging and effective interactions. With these insights, we can create TOIAs that not only reflect a *TOIA avatar*-maker’s personality but also resonate with users on a more personal level, enhancing their overall experience.

## 2.5 Evaluation Methods

Traum et al. (2015b) [73] report that their TOIA gave relevant direct answers to 60-66% of user utterances, which seemed good enough from “informal impressions from current testing at a museum.” However, we don’t have to date a rigorous study about how multiple users of TOIAs evaluate such interactions.

The evaluation task in adjacent fields such as IR and QA is often criticized and remains an open problem (Liu et al., 2016) [36]. IR systems focus on the relevance of a set of documents retrieved, and ‘relevance’ itself is a notion not exempt from criticism [42]. The evaluation metrics mostly reported are the Mean

Average Precision (MAP) and the Mean Reciprocal Rank (MRR). Applying these metrics to answer retrieval makes it difficult to compare systems. In the context of TOIAs, we only care about the single retrieved answer as the most relevant. Moreover, MAP and MRR are influenced by how many relevant q-a pairs exist or are retrieved by the system. Hence, it isn't easy to compare results across different datasets or annotation methodologies of the same dataset. Other metrics like Recall@ $k$  also depend on the number of relevant q-a pairs. For example, Lowe et al. (2015) [39] report Recall@ $k$  by picking the right answer and ten randomly sampled distractors rather than computing a relevance score between a question and all the possible answers available in the knowledge base. This way, Recall@10 would always give 100%, making it difficult to judge how good the system would be from a user perspective in a practical implementation.

Question-Answering system evaluation is not necessarily relevant for TOIAs as the QA task is more about reading comprehension than the ability to retrieve an answer from a knowledge base and engage in a free-form dialogue format. Moreover, such systems often use text generation models which we didn't use in our TOIA. Text generation methodologies are usually evaluated with  $n$ -gram based metrics (Merdivan et al., 2020) [44] such as BLEU (Papineni et al., 2002) [53], ROUGE (Lin et al., 2004) [34] and METEOR (Banerjee et al., 2005) [7], which are often criticized for their poor alignment with human judgment (Chen et al., 2019) [9].

The evaluation of dialogue systems is an area of research where – to the best of our knowledge – no established or robust methodology seems to exist yet for two main inter-related reasons: automatic metrics do not correlate well with human

judgments, and human judgments are difficult to measure (Li et al., 2019) [33].

Across all the works cited above and the datasets presented later for study free-form conversations, there is a gap in addressing the question of what is a ‘good’ answer. This is important for evaluating the relevant NLP tasks and defining an annotation methodology for the TOIA we develop and present in Chapter 5. Recent work by Adiwardana et al. (2020) [2] describe Meena, a chatbot model trained end-to-end on 40 billion words from social media discussions to demonstrate the ability of a large-scale low-perplexity model to produce high-quality language outputs. Meena is trained on multi-turn dialogues and evaluated using a human evaluation metric, Sensibleness and Specificity Average (SSA), which measures the bot’s ability to make sense of and provide specific responses.

## 2.6 Outlook

The general public cannot afford current TOIA deployments due to their high production costs: creating a character (aka *TOIA avatar*) may require pre-recording about 2,000 video answers (Artstein et al., 2015) [5]). We proposed a more streamlined avatar development process in Chierici et al., 2020 [10]. However, it is still impractical for the everyday user: it involves transcribing and recording conversations based on brainstormed plausible utterances. Their work created over 400 pre-recordings and manual annotations that took several days.

Generalizing and streamlining the avatar development process for a mass-use

system is a crucial aspect of time-offset interaction research. One initial attempt in this direction was made by undergraduate students at New York University Abu Dhabi (Abu Ali et al., 2018) [1]. Their system enables users to chat with *TOIA avatars* in different languages, but its implementation is not user-friendly. It consists of two separate, non-communicating components for video recording and interaction, and it only supports local installation with no option for multiple users.

This proof-of-concept created by Abu Ali et al. (2018) [1] aims to create a TOIA capable of cross-lingual dialogue. The TOIA allowed for interactor questions in one language and avatar answers in another, with captions in the interactor’s language. The system supported both Arabic (Ar) and English (En) and all their combinations: (Ar-Ar, En-En, Ar-En, and En-Ar). The prototype supports cross-lingual avatars and lowers computational requirements. However, the system’s implementation is not without limitations. The dialogue manager relies on basic information-retrieval techniques, resulting in a less fluid conversational experience. Additionally, the *TOIA avatar* makers brainstorm questions for interactors, which leads to limitations such as running out of ideas quickly and missing obvious interactions like greetings such as “Hello” and “Goodbye.”

Our previous work proposes a *TOIA avatar* development process that involves recording and transcribing conversations to create question-answer pairs, resulting in data sources and baselines for the dialogue manager of a TOIA (Chierici et al., 2022) [10]. However, this process is impractical for everyday users. To address this limitation, we developed a proof-of-concept (PoC) cloud-based system that

introduced question suggestions using language generation models [11]. While this PoC makes the system more accessible to users online, it has several limitations, such as poor question-suggestion quality, topic clumping, slow suggestion latency, and a dialogue manager using baseline semantic-search techniques. With the event of GPT-3, we then addressed all these issues and improved our TOIA as thoroughly covered in Chapter 5. In this dissertation, we build upon the previous work by improving the cloud-based PoC with a more engaging user interface and an integrated question suggestion module to enhance the user experience. We aim to make our TOIA available to research the field further and enable anyone to create interactive self-narrative *TOIA avatars* of themselves. This work aligns with our continued efforts to advance the development of TOIAs.

# Chapter 3

## Data Collection

One of the essential components of a TOIA is the video database, which stores video recordings of individuals' utterances. These recordings are then used to facilitate the time-offset interaction. Creating video databases is crucial for successfully implementing TOIAs and their transcriptions for training and evaluating the necessary retrieval algorithms.

In this chapter, we present the first work we did for creating a reference data set for training, creating, and evaluating retrieval algorithms and providing proof of concept for creating a *TOIA avatar*. We used the TOIA built by Abu Ali et al. (2018) [1] to create the Margarita Dialogue Corpus (MDC, Chierici et al., 2020 [10]). This corpus includes a database of question-answer pairs and their corresponding video clips and several annotated dialogue transcripts (see Table 3.1). We made the corpus available for researchers interested in developing DMs for TOIAs and modeling dialogue.

Section 3.1 introduces our contribution to generating the MDC and the data

Question	Answer
“Do you miss the food, the Moldovan cuisine?”	“I used to more then, but now I got so used to just changing my diet depending on where I am at that time. And also I found some place where we can eat Russian food and the dining hall is also making Russian food from time to time... so I don’t miss it that much.”
“How do you make money with music?”	“By being good at what you do and knowing people. That’s how you get a job in the music industry and you grow from there.”
“I never asked about your major?”	“I studied music and economics. I’m a music major, economics minor and in music I do mostly composition and sound engineering.”

Table (3.1) Three examples of pairs in the Margarita Dialogue Corpus KB data resource.

annotation methodology. In Section 3.2, we present and discuss baseline models to provide an initial evaluation of a simple DM and an error analysis. Section 3.3 discusses problems encountered in analyzing the MDC, such as threshold selection and corpus expansion.

Category Class	Examples	Frequency in KB	(in %)
Meta-interactions	Hinting to users what to ask about or providing diversions.	62	7%
Pleasantries & Short Answers	Greetings, yes/no answers, compliments.	159	18%
Personal Information	Family, country, past, future, love, etc.	362	41%
New York University Abu Dhabi	University life, admissions, courses, and life in the United Arab Emirates.	309	35%
<b>Total Frequency</b>		<b>892</b>	

Table (3.2) Summary of the categories in the Margarita Dialogue Corpus Knowledge Base (KB) defined by the *TOIA avatar*-maker.

Statistics	KB	Dial (All)	Dial Training	Dial Test	Dial EDU	Dial PER
# dialogues	NA	20	10	10	10	10
# q-a pairs (in total)	892	659	340	319	296	363
# unique questions	758	NA	NA	NA	NA	NA
# unique answers	431	NA	NA	NA	NA	NA
# annotated answers	NA	888	472	416	421	467
# no-answers (in %)	NA (NA)	49 (15%)	0 (0%)	49 (15%)	25 (8%)	32 (9%)
# words (in total)	20,303	40,557	20,230	20,327	20,084	20,473
Min. # turns/dialogue	NA	22	22	26	22	24
Avg. # turns/dialogue	NA	33	34	32	30	36
Avg. # words/question	7.8	14.5	14.5	14.6	16.3	13.1
Avg. # words/answer	15.0	47.0	45.0	49.1	51.5	43.3

Table (3.3) Summary statistics on the two main data sets in the Margarita Dialogue Corpus: Knowledge Base (KB) and Dialogues (Dial). Statistics for the dialogues are also shown for the training portion vs. test and university mode (EDU) vs. personal mode (PER).

### 3.1 The Margarita Dialogue Corpus

The MDC corpus was created using two methodologically different steps. First, the *TOIA avatar*-maker can brainstorm several question-answer (q-a) pairs. Second, the *TOIA avatar*-maker records real dialogues with different people. This way, questions that the *TOIA avatar*-maker may not have brainstormed and yet do happen in real conversations can be covered. Such examples include introductions and greetings q-a pairs like “Hi–Hello”, “How are you?–I’m fine, thank you”, “Goodbye–Bye bye!”.

We proposed this methodology as a starting point to have enough q-a pairs to cover common conversation topics that the *TOIA avatar*-maker may not have considered in the prior brainstorming. Moreover, recording real dialogue gives insight into other people’s reactions and picks up on topics that depend on the

interrogator's background. From a user experience perspective, what would be the ideal experience for the interrogator? She or he should be able to question an avatar in a fluent conversation that would mirror the real experience of getting to know a stranger in a 10 to 15 minute interaction.

We selected twenty subjects as interrogators, making sure they did not know the *TOIA avatar*-maker in person, and we instructed each to engage in a 15-minute conversation with the *TOIA avatar*-maker. We instructed the *TOIA avatar*-maker to avoid asking questions back to the interrogator, although this naturally happened in a few dialogues' turns. When interrogators were unsure what to ask, they could ask the *TOIA avatar*-maker what questions or topics she could discuss. Part of these recordings (named 'training set' in later sections) was used to define the Knowledge Base (KB), and part of the recordings was used as held-out test samples (named 'test set' in later sections) to evaluate the baseline models for the DM. Moreover, we wanted data to be both 'on-topic' and 'wild' to study two different avatar interactions: making a *TOIA avatar* who can act as an information kiosk for the university she attended and an avatar who can talk about herself. We forced half of the conversation to be about one topic, called the 'university mode' (or 'EDU'): the *TOIA avatar*-maker, a New York University Abu Dhabi (NYUAD) student, could only answer questions about the campus and academic life. The second half of the dialogues did not have a set topic, and we call this the 'personal mode' (or 'PER'): we asked the interrogator to get to know the *TOIA avatar*-maker as one would do when meeting a person for the first time. We now dive into the specific data collection and annotation methodologies.

**Knowledge Base and Dialogues** We initially let the *TOIA avatar*-maker brainstorm q-a pairs ‘out of context’, meaning that the questions and answers were not part of a dialogue flow between two individuals. She defined 241 pairs of questions and answers. We then recorded 20 dialogues with real people, ten about the university (EDU mode) and ten about the *TOIA avatar*-maker (PER mode). We randomly selected ten conversations (5 in university mode and 5 in personal mode) and used them as the ‘training’ set, and we used the other ten conversations as the ‘test’ set. We then asked the *TOIA avatar*-maker to check which questions in the recorded conversations were not present in the initial q-a pairs brainstormed ‘out of context’. The new questions were then added to form the avatar’s KB, and the *TOIA avatar*-maker recorded videos answering these questions. As part of this consolidation work, some questions with similar meanings were kept in the KB and were assigned the same answer. For instance, “I like your... / I love your... / You are so... / Nice... / That’s interesting!” are paired with the same answer, “Awww thanks!” in the KB. The resulting KB has 892 q-a pairs and 431 unique answers, also available in video clips. The *TOIA avatar*-maker also recorded videos like “Can you ask me something else?”, “Can you repeat that?”, “Could you elaborate?” and so forth. The system would play them when the ranking model does not produce a result within a certain confidence threshold. We categorized these answers as ‘unsure’ in the KB.

We summarize the statistics of the two data sets, the ‘Knowledge Base’ and the ‘Dialogues’, in Table 3.3.

**Annotation** After the data was collected, the *TOIA avatar*-maker became the annotator. Hereafter, we will interchange the words ‘annotator’ and ‘*TOIA avatar*-maker’. She engaged in an annotation exercise inspired by the ‘Wizard of Oz’ setting. In a typical Wizard of Oz setup, the interrogator would query the system on a screen. Behind the scenes, a real person - the wizard - would select an appropriate answer to play to the interrogator. While the wizard operates in real-time to facilitate a dialogue with a TOIA-interactor, the annotator operates after the fact to carefully select the best answer (if available) to serve as training data. This annotation may be less prone to error because it removes the time pressure to play an answer immediately. The *TOIA avatar*-maker had all the answers recorded in the Knowledge Base at her disposal. She then paired each question appearing in the dialogues data (i.e., from the new conversations recorded) to the first-best answers available in the Knowledge Base, the second-best answer, the third, and so on until the sixth-best answer.

The purpose of this exercise is to build two data sets. The ‘training’ portion of the dialogues will have a guaranteed answer for each question because the interrogators’ recorded questions were used to consolidate the KB, as described in the paragraph above. The ‘test’ portion of the dialogues will have some questions without an answer in case none of the answers available in the KB is a good fit for a given question.

The annotator also subjectively categorized her answers, and we made them available in the data. She defined 66 categories (68, excluding the ‘unsure’ category and the ‘filler’ category, which we describe in the next paragraph). About

one-third of the categories account for 80% of the KB, showing a typical power-law behavior. The top 10 categories, accounting for more than half of the KB, are music, pleasantries, opinions, compliments, applying to NYUAD, languages, NYUAD in general, positive memories, travel, and short answers. These categories reflect well common conversation topics for the two contexts we defined (information about the university and meeting a person for the first time) as well as the personality of the annotator/avatar-maker. Table 3.2 shows high-level statistics related to these categories. Given the subjective, unstructured way these categories were defined, further work would be needed to study their applicability. For instance, they might be grouped into macro-topics in a way similar to that shown in Table 3.2. For a visual representation of the type of questions and answers in the KB and the dialogue sets, refer to Figures A.1, A.2, A.3 and A.4 in Appendix A.

It is worth pointing out that the KB contains two categories that we eliminated for training the baselines. The ‘unsure’ category represents videos saying “Sorry, I don’t have an answer for that.’ Or, “I am not sure I understand. Can you try rephrasing your question?”’. The ‘filler’ category corresponds to videos where the avatar is making gestures to fill video space between an answer and the next question the interrogator will ask. So this category is valid only from a user experience perspective when using the TOIA.

The corpus is named the ‘Margarita Dialogue Corpus’ after the avatar’s first name, and it is available to download at NYUAD CAMEL Lab’s Resource page.<sup>1</sup>

---

<sup>1</sup><http://resources.camel-lab.com/>. Go to ‘Corpora’, then to ‘Margarita Dialogue Corpus’.

## 3.2 Dialogue Manager Baseline Models

We implemented information retrieval baselines and report metrics relevant to single-turn interactions. The data set is well suited for multi-turn dialogues, and the human evaluation in Chapter 4 includes multiple turns.

We used three different models to convert sentences into sentence vectors. Then we computed the distance between an interrogator’s question vector and all the vector representations of all the questions present in the KB. The distance gives us a ranking function for every answer in the KB: the closer the question relative to a given answer in the KB for an interrogator’s question in the sentence vector space, the higher the rank of the answer in the KB as a potential reply for the interrogator’s question. Now, there are two tasks we need to get right. First, the model should be able to decide whether or not, for any question posed to the system, there exists an answer in the KB. Then it must identify what the best answer is.

In the proposed baselines, we tackled the two questions simultaneously by thresholding on the training set, i.e., we came up with a heuristic to decide if the value of the distance (or similarity) metric is high enough to ascertain whether or not an interrogator’s question is indeed similar to a question within the KB. If all the questions in the KB have their distance metric below the selected threshold, the system will output a ‘no-answer’ message. Suppose there are questions in the KB with distance metrics above the threshold. In that case, we rank their corresponding answers as candidates for the interrogator’s question using the value of the distance metric itself.

### 3.2.1 Evaluation metric

Initially inspired by Lowe et al. (2015) [39], Schatzmann et al. (2005) [67] and BLEU (Papineni, 2002) [53] (which uses multiple references), we evaluated the baseline models using a multi-reference Recall@ $k$  metric. We tasked the baseline models to select the  $k$  most likely responses and the metric accounts for the true answer being within the top  $k$  candidate responses. In practice, for real-world TOIAs, only the Recall@1 metric would be relevant. Although our database provides a rank of choices, we consider them equal and leave the ordering to further work (i.e., making models that give the ‘first-best answer’ a higher ranking than the ‘third-best answer’).

### 3.2.2 Models

**TF-IDF** The first model uses the term frequency-inverse document frequency (TF-IDF) statistics. This quantity should capture how important a given word is to some document [59], which in our case is the question. TF-IDF is a technique that is often used in document classification. The ‘term-frequency’ is the count of the number of times a word appears in a given document, and the ‘inverse document frequency’ is a multiplier that penalizes how often the word appears elsewhere in the overall collection of documents (the corpus). The statistic is defined as

$$TFIDF(w, d, D) = f(w, d) \cdot \log \frac{N}{|\{d \in D : w \in d\}|},$$

where  $f(w, d)$  is the term frequency of the word  $w$  into document  $d$ ,  $N$  is the total number of documents, and the denominator represents the number of documents in which the word  $w$  is present. Questions and answers are transformed into TF-IDF vectors, returning  $k$  answers corresponding to the top  $k$  cosine similarities between

test questions in the test set and the questions in the training set.

Given the size of this data set, yet trying to leverage state-of-the-art results achieved by deep learning in other question-answering contexts, we attempt to leverage pre-trained models that generalize well in other contexts. We chose InferSent by Conneau et al. (2017) [14], and BERT by Devlin et al. (2018) [16] because of their generalization power and versatility towards NLP tasks.

**InferSent** We investigated the InferSent pre-trained word embeddings proposed by Conneau et al. (2017) [14] as a second baseline. Similarly to TF-IDF, we developed the matching technique by checking the cosine similarity of word vectors between the validation or test sets questions and the training set answers.

**BERT** The Bidirectional Encoder Representations from Transformers (BERT) method was one of the breakthroughs in NLP in late 2018 (Devlin et al., 2018) [16]. BERT is a method of pre-training language representations, and it is helpful for transfer learning tasks. It can extract high-quality language features from any text data and fine-tune the model on a specific task like classification, entity recognition, or question answering. In particular, we use BERT to extract word and sentence embedding vectors, again calculating cosine similarities. It is worth pointing out that BERT offers an advantage over models like Word2Vec because while each word has a fixed representation under Word2Vec, regardless of the context within which the word appears, BERT produces word representations that are dynamically informed by the words around them.

Statistics	KB'	Training Dialogues	Test Dialogues
# q-a pairs (in total)	776	401	319
# unique questions	698	NA	NA
# unique answers	398	NA	NA
# no-answers (in %)	NA (NA)	61 (15%)	57 (18%)

Table (3.4) Summary statistics after sampling non-answers for the training set: down-sampled KB (KB') and up-sampled training dialogues. The training set remained the same with minor changes due to answers no longer being present in the KB'.

We implement the three vector representations of sentences described in the previous sections, TF-IDF, InferSent, and BERT, without pre-processing tokens. We then compute the similarity to questions in the KB to establish whether we have an answer for a given question and rank the answers to retrieve. We considered the similarity between a new question and every answer in the KB as an alternative model, but the results are so much weaker than question-similarity that it is not worth reporting them.

### 3.2.3 Confidence threshold selection

A particular challenge seems to be the setting of a confidence threshold for the ranking function (in all models, the cosine similarity between word vectors) to decide if, in low similarity cases, the system should give the top-ranked answers a non-answer. The presence of non-answers in the test set influences the performance metrics. 15% of the questions in the test set does not have an appropriate answer selected by the annotator. Given that we use the ‘training’ set to select the confidence threshold, we need to make some adjustments because the training set does not have any ‘non-answer’. We up-sample questions by picking questions in the

Threshold	TF-IDF	InferSent	BERT
0.05	0.277	0.234	0.269
0.1	0.277	0.234	0.269
0.15	0.277	0.234	0.269
0.2	0.277	0.234	0.269
0.25	0.277	0.234	0.269
0.3	0.277	0.234	0.269
0.35	0.282	0.234	0.269
0.4	0.287	0.234	0.269
0.45	0.307	0.234	0.269
0.5	0.317	0.234	0.269
0.55	0.342	0.234	0.269
0.6	0.352	0.239	0.269
0.65	0.357	0.242	0.269
0.7	0.362	0.252	0.269
0.75	0.362	0.267	0.272
0.8	<b>0.369</b>	0.282	0.294
0.85	0.357	0.289	0.307
0.9	0.339	<b>0.302</b>	<b>0.332</b>
0.95	0.312	0.284	0.319

Table (3.5) Recall@1 statistics for each baseline evaluated on the training set. The metrics corresponding to the thresholds that were automatically selected are bolded.

Threshold	# Correct Answer	# Correct Non-Answer	TPR-ans	TPR-non-ans	Recall @1
0.05	108	0	0.269	0.000	0.269
...	...	...	...	...	...
0.7	108	0	0.269	0.000	0.269
0.75	108	1	0.339	0.018	0.272
0.8	107	11	0.335	0.193	0.294
0.85	103	20	0.323	0.351	0.307
<b>0.9</b>	<b>91</b>	<b>42</b>	<b>0.285</b>	<b>0.737</b>	<b>0.332</b>
0.95	72	56	0.226	0.982	0.319

Table (3.6) Thresholding considerations for the BERT model. The table shows, for each threshold level imposed to cosine similarities, the number of correct answers predicted, the number of correct non-answers predicted by the model, hence the true positive rate for answers (TPR-ans), the true positive rate for non-answers (TPR-non-ans) and the Recall@1 metric. The automatic choice of threshold for this setup is highlighted in boldface.

KB whose answers were not selected by the annotator in the training portion of the dialogues. We then remove the corresponding q-a pairs from the KB. We show the resulting, ‘adjusted’ KB (KB’), training set, and test set statistics in Table 3.4.

We use a simple heuristic that accounts for the trade-off between answers’ true positives rate (TPR-ans) and non-answers true positive rate (TPR-non-ans): We give more weight to the TPR-ans because there are more answers than non-answers in the annotated data sets. Moreover, it is easier for the model to predict a non-answer by merely picking a high threshold, whereas selecting the correct answer (when there is one) is a more complex and exciting task from an NLP standpoint. A high recall - like the Recall@1 metrics seen in Table 3.5 for high threshold values - may be misinterpreted as a good result when all the model is doing is achieving maximum accuracy on the 15% of the test set’s examples that have no answers. See, for example, a look ‘under the hood’ of the Recall@1 metrics for the BERT model at different threshold levels in Table 3.6.

### 3.2.4 Results

Similarity Threshold	TF-IDF	InferSent	BERT
Recall@1	0.194	0.169	<b>0.201</b>
Recall@2	<b>0.207</b>	0.169	<b>0.207</b>
Recall@5	<b>0.210</b>	0.169	<b>0.210</b>
Recall@10	0.210	0.169	<b>0.213</b>
Recall@20	0.210	0.169	<b>0.213</b>

Table (3.7) Results for each baseline on the test set. For each model and threshold selection, the Recall@k metric is shown for different levels of k on the test set.

The results (see Table 3.7) show that these baselines are limited to achieving

an optimal user experience. Surprisingly, a pre-trained model like InferSent, which generalizes well in other NLP tasks, does not improve the more traditional TF-IDF statistics results. This level of performance might be due to the size of the problem, and perhaps further parameter tuning might lead to better results. Although BERT shows the best results, the metrics are only slightly better than the TF-IDF model.

	<b>TF-IDF</b>	<b>InferSent</b>	<b>BERT</b>
<b>Recall@1</b>	0.103	0.046	<b>0.111</b>
<b>Recall@2</b>	<b>0.168</b>	0.084	0.153
<b>Recall@5</b>	<b>0.240</b>	0.126	0.221
<b>Recall@10</b>	0.282	0.164	<b>0.302</b>
<b>Recall@20</b>	0.363	0.210	<b>0.420</b>

Table (3.8) Recall@k only for answers (i.e., ignoring questions in the test set that did not have an answer in the KB) for each baseline. For each model, the Recall@k metric is shown for different levels of  $k$  on the test set.

To gauge a better performance differentiation between the baselines, we simplified the problem by focusing only on the ability of the distance metrics to pick up the correct answer within the first—one, —two, —five, —ten, and —twenty ranked responses in the KB independently of the threshold. In other words, how well do we rank the answers by forgetting about being able to give a non-answer? Table 3.8 shows these results by considering only the Recall@k for answers rather than the former Recall@k that considers both answers and non-answers. We can see that BERT embeddings do a much better job than InferSent embeddings and, although not far from the TF-IDF model’s performance, they improve the TF-IDF performance materially when looking at the TPR in the first—twenty ranked responses.

Test Question	Model: Top Ranked Similar Question in the KB'	Corresponding Answer (Similarity Score)
“So is this your natural hair color?”	TF-IDF: “Is this your natural hair color?” (0.972) InferSent: same question (0.847) BERT: same question (0.965)	“I started dying my hair red since I was about 15. I just like that it stands out since I like attention.”
“So where were you raised?”	TF-IDF: “Where are you from?” (0.458) InferSent: same question (0.825) BERT: “When did you graduate?” (0.761)	TF-IDF: “I’m from Moldova, a small country in Eastern Europe.” InferSent: same answer BERT: “I graduated in May 2019.”
“Oh my God. That could have been. That’s wild. And do you miss home when you’re here in Abu Dhabi?”	TF-IDF: “Do you miss home a lot?” (0.424) InferSent: “What can I do for fun in Abu Dhabi?” (0.815) BERT: “Do you have siblings who study in New York or Abu Dhabi?” (0.803)	TF-IDF: “I’ve lived abroad before so I’m used to being away, so no, I don’t miss home too much. Plus, I skype my family often.” InferSent: “You can go to the beach, bars or visit the Louvre museum!” BERT: “I don’t.”
“When are you heading back home?”	TF-IDF: “Are most people back home Orthodox?” (0.566) InferSent: “How far are you guys from the city?” (0.822) BERT: “Where are you from?” (0.774)	TF-IDF: “They’re mostly Christian Orthodox, but there’s also Protestants and Seventh Day Adventists and some versions of those.” InferSent: “We are 30 minutes away from the city center by car or bus.” BERT: “I’m from Moldova, a small country in Eastern Europe.”

Table (3.9) Top ranked question similarity for some meaningful examples in the test set for all three baselines.

### 3.2.5 Error analysis

Table 3.9 shows a few meaningful examples from the error analysis, where we can make some hypotheses on how the baselines work and point out the challenges associated with the threshold selections.

In the first example, all baselines pick up the right question in the KB (hence the right answer). However, the cosine similarity for InferSent sentence embedding falls below the model’s selected threshold (0.9), so for that instance, the system would return a non-answer. In the second example, the test example’s question is ‘So where were you raised?’ and both TF-IDF and InferSent models pick up the right question in the KB: ‘Where are you from?’. However, for all baselines, we are below the threshold so that the system would return a non-answer. The system would be correct for the InferSent case as a similarity score of 0.741 is not high enough to be confident it corresponds to the right answer, but it is not correct to state that there is no answer in the KB. The examples point to the direction of separating the two problems (‘Do we have an answer in the KB?’ and ‘Which one?’).

The third example seems to show what meanings different sentence embeddings give importance. Though all the answers do not exceed the thresholds, TF-IDF correctly identifies the best matching question in the KB. TF-IDF spots the ‘miss home’ keywords, whereas InferSent and BERT seem to pick up on a location (‘Abu Dhabi’). BERT might also give importance to a particular meaning of the word ‘home’ in the test question, matching something related to ‘siblings’ in the KB.

The fourth example shows that TF-IDF again functions by keyword matching

(‘back home’), whereas InferSent and BERT embeddings seem to capture different meanings of the words ‘back home’. InferSent picks up provenience, and BERT looks like identifying geographic distance.

These examples show how different sentence embeddings weigh keywords or try to capture meanings or other elements within a sentence. They also show that the vector space model they infer to sentences is incomparable and may mean completely different things. Cosine similarities in different vector spaces have different scales too: we could also notice it in Table 3.5 where TF-IDF’s results tend to be more evenly spread across thresholds, whereas InferSent and BERT have values more clumped towards high cosine similarities.

### 3.3 Discussion and Lessons Learned

Building the Margarita Dialogue Corpus and working on the DM baselines has exposed the following challenges.

**Threshold selection** We performed a confidence threshold selection with a simplistic heuristic. We are looking into separating the problems of establishing if an answer exists in the KB and what is the best answer. Regarding the confidence threshold to select for establishing if the system can answer a question, two directions for improvement include expanding the KB building upon the work of (Traum et al., 2015) [73], and automating the scoring of chatbot responses as in work by Yuwono et al. (2019) [86].

**Word tagging and entity recognition** As pointed out in the error analysis, we could already improve some baseline results by exploring and implementing models for word tagging, entity recognition, or semantic parsing. In this way, a TOIA’s dialogue manager should be able to differentiate between sentences with the same meaning and words, apart from just one word (usually the predicate’s object) that points to a completely different answer. An example of this would be “Do you have any siblings?” vs. “Do you have any pets?”.

**Accounting for annotated ordering** The annotator indicated the order of answers from the most plausible at a given point of the conversation to the least plausible. We could develop a more sophisticated evaluation metric for taking into account this information. To better assess a language model, we can compare the answer ranking provided by the model with the ranking provided by the annotator. An idea would be to modify the Recall@ $k$  metric to give weight to ordering the first  $k$  ranked answers rather than counting the mere presence or absence of correct answers within the first  $k$  ranked answers.

**Human evaluation** Different or additional evaluation methodologies could be drawn from the HCI community. For example, recent work by Amershi et al. (2019) [3] proposes eighteen generally applicable design guidelines for human-AI interaction practitioners working on applications and features that use AI and to researchers interested in the development of guidelines for human-AI interaction design. Moreover, in the context of unstructured multi-turn dialogue modeling, the most used automatic evaluation metrics are biased and correlate poorly with human judgments of response quality (Loewe et al., 2017) [38]. Improving the human evaluation framework is an important research direction for the Margarita

Dialogue Corpus and dialogue systems in general, and the work by Li et al. (2019) [33] proposes a novel, exciting perspective. We'll study this problem in more detail in the next Chapter.

**Expand corpus creation** One point emphasized throughout this work is the importance of the data size. More data shall be recorded and annotated for a single avatar, as suggested by Artstein et al. (2015) [5]. To understand how much data is enough for a time-offset conversation, one approach could be to create synthetic avatars by borrowing dialogues from other data sets, such as movie scripts. Another exciting approach would be to expand the corpus by creating more avatars and collecting many user interactions as we will do with the extensive user studies described in chapters 6 and 7. The resulting corpora's summary and descriptive statistics are shown in Appendix A, tables A.1 and A.2.

**Using richer features** The baselines presented here do not use further information added by the category annotations, the two different modes (university and personal) of conversation, or the multi-turn nature of the dialogue data sets. Machine Learning approaches can leverage the additional information for classifying a question as having an existing answer in the KB and better-ranking answers. Moreover, structuring the training and test data by combining more dialogue turns can be used to fine-tune deep learning models such as BERT in a broader context. For example, such a structure might address the best following answer for a context like a -question triplet. To improve the task of selecting the correct answer, we could leverage the sequential dialogues, pair every question with plausible answers, and sample implausible answers from the KB, building on the approach introduced by Lowe et al. (2015) [39]. The hypothesis is that a system that tracks a dialogue

sequentially might be better suited for selecting the correct answer and engaging the user in a social conversation. We considered implementing deep learning models, especially RNNs and LSTMs, but setting up a model comparable to one that produces state-of-the-art results led us to abandon this approach because the number of parameters far exceeds the size of the data sets examined in this work.

**Self-narrative long recordings** An alternative, perhaps more challenging, route for developing time-offset interaction with avatars is for the *TOIA avatar*-maker to record a long, self-narrative video. The system would then operate with reading comprehension algorithms to play only the video clip snippet corresponding to the best answer to an interrogator’s question.

**Context transferability** One other area of investigation is using the same avatar for a different context. For instance, the Margarita Dialogue Corpus *TOIA avatar*-maker created two corpora for two different contexts: providing information about New York University Abu Dhabi and speaking about herself when introduced to a stranger. Other contexts of interest could be self-narrative for a curriculum. The exciting aspect will be to study if there are parts of the dialogue that we can transfer between contexts, e.g., the avatar’s talking style, jargon, or vocabulary.

\*\*\*

We proposed an original approach to collect and annotate data for training and evaluating the DM of a TOIA. We created two types of data sets: an intuition-based, single-turn knowledge base, and in-context, multi-turn dialogues annotated by the *TOIA avatar*-maker. We make the Margarita Dialogue Corpus available to the research community, including the recorded video clips of the *TOIA avatar*-maker’s

answers.<sup>2</sup> We implemented three baselines for laying down the basis to improve the answer selection for a TOIA that allows anyone to create self-made avatars in a relatively short time frame and low cost though we realize further work is needed for streamlining the process – which we will present in chapters 5 and 6.

While the baselines report on single-turn metrics, the MDC could also be suited to research unstructured, multi-turn dialogues with low-resource data and transfer learning tasks. Some of the results observed point to interesting research paths, such as improving the best answer confidence threshold, defining the correct setup for human evaluation, or understanding how much data is enough for a TOIA.

---

<sup>2</sup><http://resources.camel-lab.com/>

# Chapter 4

## Human Evaluation

Ideal TOIA interactions are expected to mirror a dialogue with a real person, including all the possible directions it may take — which naturally has great ramifications on evaluation: when two people meet and engage in casual conversation, questions may range over different topics and depending on the answer to a specific question, different conversational question-answer threads may unravel. Also, unexpected or different answers to a specific question can be acceptable and not cause a change in the overall conversational flow.

So, how can we answer the question *what is a ‘good’ (i.e., ‘right’, ‘correct’ or ‘relevant’) answer?* As this study was performed as an initial evaluation process prototype and before the update and development of our TOIA software, we explore the question using the dataset that was manually annotated by its avatar maker – the Margarita Dialogue Corpus (MDC, see Chierici et al., 2020 [10] or the previous chapter). The best-performing answer-retrieval model we could produce had a low Recall@1 – 24% on the development (dev) set and just below 10% on the test set.

When trying to interact with the MDC knowledge base, one would expect to get a correct answer about 1/4th of the time or less. However, when chatting with the avatar using the best performing model (Figure 4.1), we could see that the system wasn't so bad at entertaining and holding a conversation. Hence we asked a 'crowd' of human annotators to give their opinion, and we learned that the task of defining the correct answers isn't straightforward, primarily due to misaligned expectations about answer relevance. In this chapter, we also explore a number of metrics and single out Success Rate @ $k$  (SR@ $k$ ) as the most pertinent metric for optimizing TOIAs.

We explain the human annotation methodology, the model, and the metrics used to address the comparison between the avatar maker's annotated answers and what the 'crowd' believes are good answers in Section 4.1. We analyze the dataset we annotated in Section 4.2 and discuss the results in Section 4.3. We also introduce the results of the latest answer-retrieval model implemented in the software we developed in our Lab (see Chapter 5) after the study presented here. The findings documented in this chapter led us to the updated retrieval methodology and to add additional interaction options to the software interface. In Section 4.4, we describe the methodology learned in this study applied to the real-world, TOIA interface we have developed and used for the extensive study about interacting with *TOIA avatars* documented in Chapter 7. In Section 4.5, we share how to apply this work's conclusions to develop better avatars.

## 4.1 Methods

Evaluating dialogue systems like chatbots and tasks like question-answering (QA) remains challenging. Reasons include the great variety in contexts and use cases for these systems and the high cost of human evaluation.

TOIA lie between Question Answering (QA) and Information Retrieval (IR). They differ from QA in that a TOIA’s task is **not** about demonstrating comprehension of a text span (Rajpurkar et al., 2016)[57], Reddy et al., 20019)[60] but selecting a single (one-shot) appropriate answer from a restricted set of answers, a problem also known as Answer Retrieval (AR) or retrieval-based dialogue (Boussaha et al., 2019)[8].

### 4.1.1 Initial Retrieval Models Explored

We used five models to retrieve answers to the MDC dialogue dataset questions and shortlist the top candidate responses for the ‘crowd’ annotation task. When analyzing the final retrieval results, we’ll also introduce an updated model introduced one year after the study was performed.

**(1) TF-IDF q-Q:** Let  $q$  be a query from a user (in our case, a question in the MDC dialogue dataset) and  $Q$  a question annotated in the MDC KB. We vectorized  $q$  and  $Q$  using a TF-IDF vectorizer trained on the KB, and computed the shortest distance between  $q$  and  $Q$  with cosine similarity. We used the sci-kit learn Python library for the TF-IDF vectorizer (Pedregosa et al., 2011)[55].

```

Interrogator: Hi
Avatar: Hey!
Interrogator: What's your name
Avatar: My name is Margarita.
Interrogator: Nice to meet you
Avatar: Nice to see you, too!
Interrogator: How old are you?
Avatar: I am 22.
Interrogator: Do you study or work?
Avatar: I just graduated from New York University Abu Dhabi!
Interrogator: Cool, and what do you plan to do next?
Avatar: I will be working in Tax, Transfer Pricing, at PwC starting August 2019!
Interrogator: How long ave you been in Abu Dhabi?
Avatar: So I came here four years ago, but because of all the traveling probably two years in total.
Interrogator: Wow where have you been?
Avatar: I've been to Buenos Aires and New York for a semester each and I've done to January terms in Shanghai and Paris.
Interrogator: Which one did you prefer
Avatar: Music completes this artistic and techie side of me because I studied composition and sound engineering, but economics fulfilled my social and analytical side, because I wanted to learn about how to make an impact in the world and I love math.
Interrogator: Did you prefer New York or Buenos Aires?
Avatar: My favorite place ever is New York. I hope to get back there at some point. I just love how vibrant the city is and there's always something to do and people are just the right amount of polite and minding their own business.

```

Figure (4.1) Initial ten turns of a casual dialogue between an interrogator and a TOIA avatar using our best-performing retrieval model.

**(2) Okapi BM25 q-Q:** Okapi BM25 (Trotman et al., 2014)[76] is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document. We used the Rank-BM25 implementation in Python.<sup>1</sup> Since BM25 was the worst-performing approach, we do not report on it further due to limited space.

**(3) BERT q-Q:** BERT is a large deep learning model architecture and one of 2018’s breakthroughs in NLP (Devlin et al., 2018)[16]. We computed the sentence embedding for each  $q$  and  $Q$  by taking the mean of BERT pre-trained layers. The cosine similarity between embeddings gives us the ranking function for computing how close a query in the dialogues is to a question in the KB.

---

<sup>1</sup><https://pypi.org/project/rank-bm25/>

**(4 and 5) Fine-tuned BERT q-A:** We fine-tuned BERT on answer selection as a classification task. Let A be an answer in the KB. For every Q-A pair in the KB, we labeled them as 1's to indicate a relevant match. We then sampled some irrelevant (or ‘wrong’) matches for every question and labeled them as 0's. We tried different sampling ratios, namely drawing one wrong match for every correct one (1:1), ten wrong ones (1:10), a hundred (1:100), and using all the available utterances (1:All). To increase the data size further and better generalize for questions phrased differently, we augmented the train data by sampling synthetic questions using the methodology proposed by Wei et al. (2019)[79] and their Python implementation.<sup>2</sup> We fine-tuned BERT for three more epochs (we chose a few epochs as advised by Dodge et al., (2020)[17]) using Wolf et al. (2019)[84]’s Transformers library. We only report on **BERT q-A 1:100** and **BERT q-A 1:All** as they perform best.

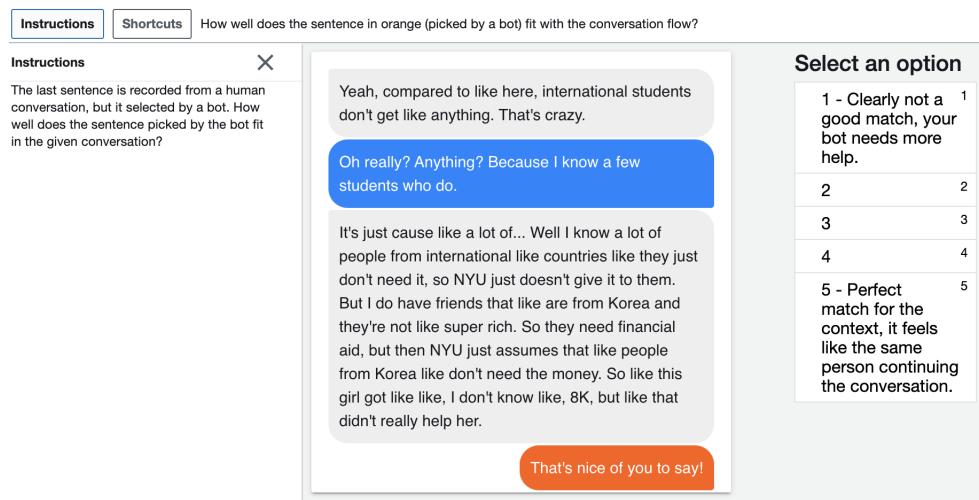


Figure (4.2) Annotation interface displayed to the human annotators.

<sup>2</sup>[https://github.com/jasonwei20/eda\\_nlp](https://github.com/jasonwei20/eda_nlp)

**(Post-study addition) GPT-3 q-QA:** Since the time of the prototype study on human evaluation presented here, a variety of Natural Language Processing tasks, including answer retrieval, have evolved from retrieval-based models to large language models (LLMs). Our research shifted from retrieval shortcomings to more efficient interactions influenced by software engineering, production, and user evaluations. LLM API improvements and the user preference for convenient interactions through clicking and typing (Chapter 7) made up for gaps still present in an answer-retrieval, voice-only interaction. We implemented OpenAI’s GPT-3 models for text search using semantic similarity (Neelakantan et. al, 2022)[48] between sentence embedding of a new query,  $q$ , and sentence embedding of a question and answer pair,  $QA$ , in the *TOIA avatar*’s knowledge base. We further document the implementation details and API parameters for the retrieval implementation in the next Chapter.

#### 4.1.2 Crowd Annotations

We developed a web interface (Figure 4.2) for collecting the annotations from the ‘crowd’ using the crowdsourcing platform Amazon Mechanical Turk (AMT). Full anonymity of the users was maintained, and the host institution’s ERB review didn’t raise ethical concerns.

For each question in the MDC dialogue dataset, we took the union of the top 10 answers retrieved by the five different retrieval techniques described above. On average, each question has about 24 selected answers. Using a sliding window on all the dialogue questions, we selected three conversation turns and appended the prediction as a fourth turn (interrogator-avatar-interrogator-predicted avatar

response) without specifying who was whom, and always starting the dialogue snippet from an interrogator’s question and ending with the avatar’s answer. We chose four turns because it seems to give an optimal context size by looking at the annotations performed on the HUMOD dataset (Merdivan et al., 2020)[44].

So we have 339 dialogue snippets for the dev set and 341 for the test set. Each human annotator could rate as many snippets as they wanted in one task. On average, they rated 23 sampled dialogue conversations. They were asked to rate the last reply of the dialogue snippet on a 1–5 scale according to the dialogue context (where 1: Clearly, not a good match; 5: Perfect match for the context). For each dialogue—three different annotators rate the reply pair. For each dialogue context, there are, on average, 72 annotated answers (24 times 3), resulting in a total dataset size of 24,291 annotations for the dev dialogues and 24,555 for the test dialogues. To maintain high-quality responses in the data, we defined a blacklist of annotators who gave poor-quality annotations. We forced each annotator to give a rating for the ‘gold answer’ the *TOIA* *avatar*-maker gave in the dialogues data. If the annotator gave a rating lower than 4 (i.e., 1, 2, or 3) to the gold answer, we removed them from the annotations. While this blacklisting methodology is quite restrictive (we lose about 36% annotations), we have a large enough number left for our purposes.

## 4.2 Human Evaluation Results and Analysis

We analyze our annotations regarding inter-annotator agreement and the relationship between the crowd’s opinion and the best-retrieved answers by the models.

<b>Rater 1 vs Rater 2</b>	$\kappa$ (dev)	$\kappa$ (test)
Closest two ratings	0.51	0.50
Lowest two ratings	0.23	0.20
Highest two ratings	0.07	0.11
Random two ratings	0.10	0.04

Table (4.1) Inter-annotator agreement computed using Cohen’s kappa score ( $\kappa$ ) for the dev set and the test set.

We then report the IR metrics on the models we decided to study.

#### 4.2.1 Inter-annotator Agreement

We computed the weighted Cohen’s kappa score (Cohen, 1968)[13] between human ratings to compute inter-annotator agreement excluding the blacklisted annotations. Following the approach of Merdivan et al. (2020)[44], we calculated the weighted kappa score for different configurations of three ratings for each different context-predicted answer pair. We calculated the weighted kappa score for the closest two (as a majority voting) ratings, the highest two ratings, and the lowest two ratings, and on a random selection of two ratings from the three ratings of each predicted answer. For example, if a dialogue snippet is rated 1, 2, and 5, we keep the closest two (1 and 2) and randomly assign them to **Rater 1** and **Rater 2**. Table 4.1 shows the results of each combination for the MDC’s dialogue dev set and test set. The inter-annotator agreement results on the test and dev set seem consistent, with the exception of the random two ratings scenarios. Though representing moderate agreement, the highest is between the closest two ratings (expected by design). Moreover, the crowd shows higher agreement on the lowest two ratings than on the highest two. As our analysis points out, the poor level of agreement between annotators isn’t about the quality of the annotations. It instead shows the difficulty of defining a good answer for a TOIA’s interrogator.

	<b>Gold</b>	<b>TF-IDF</b>	<b>BERT</b>	<b>BERT</b>	<b>BERT</b>
	<b>q-Q</b>	<b>q-Q</b>	<b>q-A</b>	<b>q-A</b>	
			<b>1:100</b>	<b>1:All</b>	
<i>dev</i>					
<b>Average</b>	4.53	4.03	3.99	4.17	4.01
<b>%Gold</b>		89.0	88.2	92.1	88.5
<b>Rank</b>	2	4	1	3	
<i>test</i>					
<b>Average</b>	4.59	3.01	2.98	3.47	3.25
<b>%Gold</b>		65.6	64.9	75.7	70.9
<b>Rank</b>	3	4	1	2	

Table (4.2) Average ratings assigned to the gold and top retrieval model choices in dev and test sets. %Gold specifies the ratio of model average rating to gold average rating. Rank specifies the performance rank of the retrieval model.

#### 4.2.2 Crowd Ratings of Retrieval Top Choices

Next, we consider the average rating given by the AMT workers to the gold answer and to the top retrieved reply by our four models. We include the ratings for all snippets excluding blacklisted annotations for both dev and test. We drop 35% of the annotations for the top retrieved answers due to blacklisting, consistently with the overall drop reported above. See Table 4.2 for the averages, percentage of the gold answer (i.e., how close to the gold answer is a model), and model ranking.

The standard deviation of the average ratings for the gold answer is 0.35 in the dev and 0.39 in the test (because of blacklisting, we only keep ratings 4 and 5 for the gold answer).<sup>3</sup> The standard deviation of the retrieval models ranges from 1.17 to 1.33 in dev and 1.20 to 1.38 in the test.

---

<sup>3</sup>For reference, the average of all ratings of the gold answers (i.e., without blacklisting) is 3.96 for dev and 3.76 for the test, with a corresponding standard deviation of 0.71 and 0.74, respectively.

Retrieval	$\rho$ (dev)	$\rho$ (test)
<b>TF-IDF q-Q</b>	0.25	0.10
<b>BERT q-Q</b>	0.16	0.08
<b>BERT q-A 1:100</b>	0.30	0.13
<b>BERT q-A 1:All</b>	0.29	0.15

Table (4.3) Spearman’s Rank Correlation Coefficient ( $\rho$ ) between each retrieval model and the human ratings for the dev set and the test set.

The results indicate that, although the crowd disagrees, they generally give high ratings to the best-retrieved answers. So, annotators may disagree in many instances, but when the models retrieve sensible answers, these are recognized by the annotators. For this reason, we decide not to resolve the annotator’s disagreements, and in the analysis that follows, we use the average rating between the three (or less because of blacklisting) scores given by the crowd for each dialogue context-predicted reply pair. According to the crowd, the model with the best top choice is BERT q-A 1:100, and the model with the worst top choice is BERT q-Q.

#### 4.2.2.1 Correlations Between Models and Annotations

We also computed the Spearman’s Rank Correlation Coefficients between the rankings produced by four of the models used for answer retrieval and the annotators’ ratings (always excluding the blacklisted annotations). The results are displayed in Table 4.3. While the correlations are weak (yet statistically significant as all the p-values approached 0), we can notice a mixed behavior. The models performing better (See Tables 4.4 and 4.5) do not necessarily correlate more with human ratings. This is a ranking correlation. So the crowd may rank differently than the models’ answers but agree on the top-ranked replies, as we have seen earlier. Furthermore, on the 24 answers for each dialogue snippet, on average, few are the top-ranked by the models. The majority are ‘negative’ examples, where it’s easier

to disagree or rank differently.

Metric	Avatar Maker's Annotations						
	Rand	TF-IDF	BERT	BERT	BERT	Crowd	GPT-3
	q-Q	q-Q	q-A	q-A			
			1:100	1:All			
R@1	0.0	23.4	22.9	<b>24.0</b>	21.0	9.1	42.8
R@2	0.2	29.5	28.0	<b>33.3</b>	29.5	15.5	64.2
R@10	1.1	38.2	42.0	<b>55.2</b>	43.1	34.0	100.0
MRR	2.6	69.1	60.5	70.1	<b>81.8</b>	48.0	57.7
MAP	2.4	66.5	69.1	<b>90.3</b>	59.6	18.8	100.0
SR@1	0.0	32.4	31.8	<b>33.2</b>	29.1	12.6	51.8
SR@2	0.3	40.6	38.2	<b>45.0</b>	40.9	20.9	60.2
SR@10	1.5	49.1	51.5	<b>66.8</b>	57.1	43.5	84.3

Table (4.4) Information Retrieval metrics on the **dev** dialogues set for all the models, including a random selection model and using the crowd ratings as a retrieval model. The models are assessed against the original annotations made by the avatar maker. On the two rightmost columns, the models are assessed against the annotations from the crowd and the newest retrieval implementation using GPT-3, respectively. We keep the best results of the older models in bold, as the GPT-3 implementation was added later than the analysis presented here.

Metric	Crowd's Annotations						
	Rand	TF-IDF	BERT	BERT	BERT	Crowd	GPT-3
	q-Q	q-Q	q-A	q-A			
			1:100	1:All			
R@1	0.1	6.5	3.9	<b>7.5</b>	6.6	20.9	13.4
R@2	0.2	9.6	6.3	<b>10.8</b>	9.6	40.4	20.7
R@10	1.3	21.5	<b>24.7</b>	24.0	18.7	98.4	62.0
MRR	7.0	49.4	49.1	53.6	<b>63.4</b>	97.3	44.1
MAP	4.5	33.6	<b>35.7</b>	28.8	17.6	97.3	48.2
SR@1	0.3	36.6	36.0	<b>42.3</b>	36.9	97.3	65.9
SR@2	0.9	50.6	48.8	<b>55.4</b>	52.4	97.3	74.1
SR@10	5.6	75.3	80.7	<b>82.4</b>	79.5	97.3	100.0

Table (4.5) Same metrics as Table 4.4, but here the models are assessed against the annotations from the crowd. We keep the best results of the older models in bold, as the GPT-3 implementation was added later than the analysis presented here.

#### 4.2.2.2 Versatile Questions and Answers

Excluding random noise or poor-quality annotations, one hypothesis is that the more volatile (or, the higher the disagreement in) the ratings for a given q-a pair, the more difficult it is to assign a ‘ground truth’ value to an annotation. To validate this hypothesis, we computed a more practical proxy of disagreement. The Coefficient of Variation (CoV) is the standard deviation of the three ratings given on the same q-a pairs divided by their average. The CoV quantifies the variability of the ratings with respect to the average rating for a given q-a pair.

Let A be the set of questions with a CoV higher than the 75th percentile (0.50) and B the set of questions with a CoV lower than the median (0.25). A has 167 utterances, B has 239 and their intersection has 133. Set A less the intersection defines the ‘versatile’ questions, i.e. utterances that go well with many answers and generate high disagreement. Set B less the intersection represents ‘one-sided’ questions, i.e. questions that don’t go well with many answers, hence generating low disagreement. To confirm this expected behavior, we re-computed the Weighted Cohen’s kappa on the two versatile and one-sided questions. The uplift in agreement or disagreement confirmed our interpretation. E.g., for the one-sided questions, the inter-annotator agreement doubles on the highest two ratings, it improves by a few points for the closest two ratings and the lowest two ratings, while  $\kappa$  becomes negative for the one-sided questions. It seems that there are less versatile questions (34) than one-sided ones (106) in the dev dialogues. Repeating the same analysis on the answers, we have 152 versatile, 237 one-sided, and 118 in the intersection.

---

**Question** (*redacted*) I think it's somewhere in California. Basically, people are no longer afraid of being caught with meth. So like the whole city is like a capital of meth. So what do you feel about that? Does that change what you previously said?

**Avatar Maker** No. **BERT** Kind of.

---

**Question** So would you feel the same way about like meth, heroin, coke because technically those...?

**Avatar Maker** (Can you ask me some- **BERT** Sure.  
thing else?, I'd rather talk about some-  
thing else, sorry.)

---

**Question** (*redacted*) How about like your perception of like religion, like are you religious or?

**Avatar Maker** I am sure something big- **BERT** I started questioning my beliefs  
ger than us exists, but I don't follow any when I was a teenager after watching lots  
specific book or rituals. of documentaries and talking to agnostics  
alike.

---

**Question** (*redacted*) How is your romantic life.

**Avatar Maker** I like to keep things like **BERT** Relationships are harder to main-  
this private. tain because of all the study-away. Also,  
the campus is a bit isolated from the city  
so it's not as lively.

---

**Question** Yeah, that's fascinating. So no toilet in your granny house?

**Avatar Maker** No. **BERT** Sure.

---

**Question** Bye!

**Avatar Maker** Goodbye! **BERT** See you later!

Figure (4.3) Examples in the dev set where the BERT q-A 1:100 system choice differs from the Avatar Maker's.

Metric	Avatar Maker			Crowd		
	BERT q-Q	BERT q-A	GPT-3 1:100	BERT q-Q	BERT q-A	GPT-3 1:100
SR@1	7.8	<b>11.9</b>	32.9	13.5	<b>25.2</b>	69.4
SR@2	12.2	<b>19.7</b>	41.3	23.3	<b>35.5</b>	74.2
SR@10	22.6	<b>36.1</b>	64.3	<b>75.8</b>	68.2	100.0

Table (4.6) SR@ $k$  metrics on the **test** set only for the best performing models on SR@1 and SR@10. We keep the best results of the older models in bold, as the GPT-3 implementation was added later than the analysis presented here.

#### 4.2.2.3 A Note on IR Metrics

Given that a TOIA’s ability to engage in a conversation depends critically on retrieving a correct answer from the knowledge base, we focus on models (and hence metrics) used for Information Retrieval.

The IR metrics we use to evaluate the answer retrieval ability of our TOIA are Recall@ $k$  (Lowe et al., 2015)[39], Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and we include the Success Rate (SR@ $k$ ) inspired by the work of Sakata et al. (2019)[63]. Success Rate is the fraction of questions for which at least one correct answer is ranked among the top  $k$ .

Tables 4.4 and 4.5 summarize the retrieval metrics for four of the models we worked with. We added two benchmarks: a model that randomly picks any answer from the KB and used the average crowd ratings’ as a ‘retrieval’ model. When we compare the avatar maker’s annotations vs. the crowd’s annotations, we label an answer as ‘correct’ for the crowd if it received an average rating  $\geq 3.5$ . The metrics consistently show that the best retrieval model looks like the fine-tuned BERT on the q-A relevance task with a training set ratio of 1:100 between correct and

wrong Q-A pairs. We can now notice why Recall@ $k$ , MRR, and MAP may not be adequate for the context of TOIA (and for the wider task of q-a retrieval from a KB).

One would expect that AMT raters would be more generous in classifying answers as ‘relevant’ for a given question. In fact, there are often cases where a sensible answer gets retrieved by a model (Figure 4.3), but the avatar maker did not deem it appropriate. Other utterances like yes/no, sure, and OK answers are relevant for many questions, but, as expected, the avatar maker would be more selective in choosing which one between a yes or a no is an appropriate answer. However, the Recall@ $k$ , MAP, and MRR look lower when the models are assessed against the crowd-annotated data. This is partly due to the models trained on the data annotated by the avatar maker but mostly because the crowd is indeed more generous, and the examples of relevant q-a pairs increased vs. the avatar maker’s annotations. Moreover, MRR is highly influenced by the number of documents retrieved by a model. In fact, the trivial model retrieving all possible answers in the KB would give a 100% MRR. MRR is the only metric for which it seems that the BERT q-A model with the 1:All sampling ratio performs better than the 1:100 ratio, but, in reality, this is due to the model with the 1:All ratio retrieving more documents.

Including the SR@ $k$  metrics makes things easier to assess. SR@ $k$ ’s for models evaluated on the crowd’s annotations are consistently higher than the respective models assessed on the avatar maker’s annotations, except for the best model. BERT q-A (1:100) gives SR@ $k$ ’s that are higher than R@ $k$ ’s when evaluated on the data annotated by the avatar maker. The difference is even steeper on the data

annotated by the crowd.

We evaluated the retrieval models versus the data with combined annotations, i.e. both by the AMT workers and the avatar maker. The results have negligible differences with respect to the assessment against the data annotated by the crowd, suggesting that the avatar-maker annotations are mostly included in the crowd annotations.

### 4.3 Discussion on Human Evaluation

Under the constraint that a TOIA is a single output system interacting with users with different expectations, we identify two challenges: first, how do we define a ‘good’ answer? and second, what’s an appropriate metric to use?

**Accurate vs Plausible Answers** Models that produce state-of-the-art results in other domains do not perform as well in a TOIA. While a model like BERT q-A 1:100 retrieves *plausible* answers, the avatar maker’s *accurate* answers differ (Figure 4.3, Table 4.2). This is also shown by the weak correlation between the IR models and the human annotators and by the poor results the ‘crowd model’ generates on the data annotated by the avatar maker (Tables 4.4 and 4.5). For some answers, it seems that the avatar maker missed them when annotating the dialogues set due to human error. In fact, she had to manually go through 431 answers for 659 questions for a total of 284,029 look-ups. On the other hand, many questions require affirmative or negative answers, which make sense when evaluating a dialogue snippet, but only one type of answer is correct for the avatar.

**Viable Metrics for TOIA Evaluation** A TOIA is a single-output system where the best answer should be retrieved as the top-ranked document, and there may be more than one answer that suits perfectly a conversation turn. This makes traditional IR metrics unsuitable for optimizing time-offset systems, so we identified SR@1 as the metric that gives the best indication of the ability of the system to retrieve a ‘good’ answer. For  $k > 1$ , SR@ $k$  gives us more insight into how to improve a model. For instance, the best-performing model can retrieve a good answer in the top 10 ranked retrieved utterances in more than 80% of the cases. This information can be used to improve the system, e.g. by retrieving the top 10 answers using BERT q-A 1:100 and fine-tuning a re-ranking methodology that pushes on top the best answers. Table 4.6 shows the SR@ $k$  metrics for the test set, and it’s interesting to notice that BERT q-Q yields a better SR@10 on the crowd’s annotations than the BERT q-A 1:100 model.

**Limitations** We limited the study to a retrieval problem and did not leverage the conversational format of the dialogues set. There are some turns where we can observe co-reference (a few examples can be seen in Figures 4.1 and 4.3). We manually annotated co-references in a sample of 100 dialogue turns, which happen in about 5% of the dialogues. So while the IR techniques produce some errors due to the conversational structure, this is not as material as to invalidate this study.

Regarding the annotation methodology, a ‘fairer’ way to annotate the answers might have been to ask the AMT workers to rate every question in the KB paired with every dialogue snippet in the dialogue sets. So when we use the crowd ratings as a model (Tables 4.4 and 4.5), we are limited to the answers that were rated

by the human annotators. Rating all answers for every single question would be unpractical, and picking the union of the top 10 retrieved answers from our models makes sure that the human raters could see an answer annotated by the avatar-maker for at least 66.8% of the questions (SR@10 of the best-performing model, Tables 4.4 and 4.5).

## 4.4 GPT-3-based Retrieval Methodology & Evaluation

In the next Chapter (5), we provide a technical description of the latest development of the dialogue manager (DM) at the time of designing and deploying this dissertation’s latest experiment. As the chapter explains, we have adjusted our performance expectations for the DM. Initially, we expected the DM to provide the correct answer every time. We thought we had to optimize the DM to achieve above 90% Success Rate @1 (SR@1). Moreover, while developing the software for allowing a human evaluation that covers the whole experience of creating *TOIA avatars* as well as interacting with them, we were limited to testing DM models with the MDC, a corpus that idealizes an interaction with a *TOIA avatar* but it’s not real-world data coming from actually interacting with a TOIA.

We could only achieve SR@1 in the region of 25%–42% with fine-tuned, state-of-the-art LLMs over the MDC. We now consider a success rate of 50% to 60% acceptable performance, as we have introduced two additional interaction modes: typing and clicking on suggested questions. These modes are often preferred by users, as they are faster, help to overcome conversation blocks such as not knowing

what to ask, and reduce speech-to-text errors caused by users speaking English as a second language, including accent and grammar errors.

To evaluate the performance of the new DM, we included in the human subject experiment described in Chapter 7 the same question we used for the Amazon Mechanical Turk crowd annotation exercise described in previous sections. In this section, we present how we designed the human evaluation of the new DM. In particular, we will limit our focus to the evaluation setup and its results.

#### 4.4.1 Evaluation Setup

In Chapter 5.3, we will describe our current retrieval implementation using GPT-3. Here, we underline that this implementation’s decision is based on the superior results it demonstrates in the MDC and AMT evaluations, as discussed in previous sections (refer to Tables 4.4, 4.5, and 4.6). Once we completed the building and deployment of our TOIA, we evaluated its ability to retrieve meaningful answers in a real-world context. This evaluation involved interacting with the TOIA, rather than relying solely on proofs of concept using text data and a notional database like the MDC.

We recruited 64 participants. They were required to chat with the *TOIA avatar* for at least 10 minutes. We asked the participant to record their screen and themselves using the laptop’s webcam, interacting with the *TOIA avatar*. We introduced, in the Player interface, a pop-up (see Figure 4.4) asking the same 5-level rating introduced in the AMT experiment every time our TOIA played a video answer to a user question: “On a 1–5 scale according to the dialogue context,

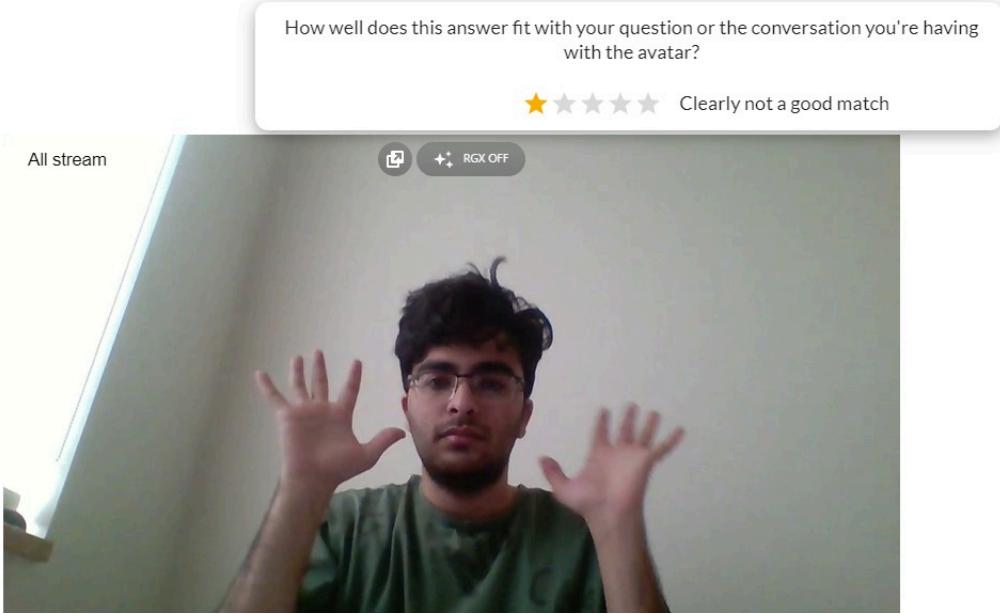


Figure (4.4) On the Player interface, a pop-up appears after every answer is played. The interactor has to rate the answer before going ahead with asking the next question.

how well does this answer fit with your question or the conversation you are having with the avatar?” (where 1: Clearly, not a good match; 5: Perfect match for the context). Our *TOIA* saves all the chat logs and the interactor ratings in its database.

Every participant interacted with at least four different *TOIA avatars* except for one who interacted with three and one with two. Participants interacted with a ‘variant’ of an *TOIA avatar* but never with the same person. Some variants were built by selecting fewer videos in their databases than others, some with different filler videos (attentive or inattentive). Why we built such variants is not relevant at this point. We encourage the curious reader to jump to Chapter 7 to appreciate the full scope of that experiment. The significant experiment design consequences for a DM evaluation standpoint are that:

Mode	#	%	Mean	StDev	Min	25%	50%	75%	Max
CARD	2,851	31.3	4.6	0.9	1.0	5.0	5.0	5.0	5.0
SEARCH	2,459	27.0	3.9	1.6	1.0	3.0	5.0	5.0	5.0
VOICE	3,790	41.7	3.5	1.6	1.0	2.0	4.0	5.0	5.0
Total	9,100	100.0	4.0	1.5	1.0	3.0	5.0	5.0	5.0

Table (4.7) Distribution of interactors' ratings by mode of interaction from the conversation log data of our TOIA.

- Every *TOIA avatar* variant was tested twice by two different interactors;
- The same interactor never interacted more than once with the same person (*TOIA avatar*);
- Different variants of the same *TOIA avatar* have to be evaluated separately regarding answer retrieval because they may have a different number of videos available to retrieve.
- Variants are not different from the point of view of the condition on the filler video, so we can group those as different annotators for the same *TOIA avatar*, making it four opinions for each avatar.

#### 4.4.2 Latest DM's Evaluation Results

The experiment generated 9,100 pairs of questions and answers. Table 4.7 displays the summary statistics and how the answers' ratings are distributed based on the method of interaction used. The data reveals that when we consider each mode of interaction individually, users opted for the voice modality the most (41.6%), followed by clicking on suggested questions (31.3%) and typing (27.0%). It is worth noting that while the voice modality was the most commonly used mode of interaction, the data does not necessarily indicate that it was the preferred method.

Overall, people tended to use quicker modalities, such as clicking on suggested questions or typing (58.4% combined) instead of speaking.

Looking at Table 4.7, it is interesting to see that the CARD mode, which should be deterministic and retrieve only correct answers, still shows some noise. As expected, the average rating is the highest, and the standard deviation is the lowest. Still, some users rated poorly for answers that were necessarily retrieved correctly. Observing several recorded interactions, the most common reason for that behavior is clicking on the wrong rating or the participant distractingly skimming quickly to the next question. The SEARCH mode of interaction involves typing an original question or selecting an auto-completed suggestion. It shows the second-highest average rating but a similar standard deviation to the VOICE mode. As we cannot differentiate between the original questions vs. autocompletion of a suggested question in our TOIA's log, we limit ourselves to analyzing the retrieval SR@1 for the interaction by VOICE. Qualitative assessment of observed user experiences points out that the participants switched between the three modalities in no particular order during a conversation, but almost all started by voice. There is, of course, a natural tendency to switch to CARD or SEARCH when the speech recognition over VOICE fails.

Similarly to the AMT experiment with the MDC described previously, we define the success rate @1 by counting how often the interactor rated an answer as correct vs. incorrect. We analyze two options to define the correct answer: including rating 3 (neutral) into the definition of correct and only categorizing the 4 and 5 ratings as correct. In the former case, the SR@1 would be 68.2%, and in the second

case, 54.5%. Both are substantial improvements over the results of the previously best retrieval model, described in the AMT crowd ratings. Moreover, the results align with the findings on applying GPT-3 over the MDC corpus and its crowd annotation – see the SR@1 result of 69.4% in Table 4.6.

#### 4.4.3 TOIA Interaction Raters Agreement

Table 4.8 shows the inter-rater agreement. We can only measure the agreement on a small sample of questions asked. To define the instances rated by multiple interactors, we need to group together different ways of paraphrasing the same question and group together variants. For example, if an interactor asked “What do you do for a living?” and another interactor asked “What do you do?” to the same variant, we have one instance and two raters of the same instance.

We group instances together by using the cosine similarity between sentence embeddings of the questions asked by the interactors. When the similarity between questions’ embeddings is above  $0.87 +/ - 0.003$  for the same variant, we define the questions belonging to the same instance. We manually inspected the clusters to ensure that we were examining the same questions, as we can accurately measure agreement when it is certain that we are evaluating identical questions. We established the threshold by trial and error, manually running many examples, and added a slight randomization (the  $+/- 0.003$ ) because no threshold picked is a clear-cut between similar and dissimilar questions.

This procedure left us with only 86 comparable instances with different numbers of raters. We chose four different coefficients of agreement that are all relevant in

Coefficient	Value	C.I.	p-value
Gwet's AC1	0.82	(0.64, 1.00)	$1.66 \times 10^{-13}$
Fleiss Kappa	0.79	(0.61, 0.97)	$1.85 \times 10^{-13}$
Brennan-Prediger	0.81	(0.63, 1.00)	$8.35 \times 10^{-14}$
Conger's kappa	0.76	(0.57, 0.94)	$6.26 \times 10^{-12}$

Table (4.8) Inter-annotator agreement computed using coefficients of agreement that are all relevant in our scenarios where we have multiple raters using ordinal ratings.

our scenarios where we have multiple raters using ordinal ratings: Gwet's AC1, Fleiss Kappa, Brennan-Prediger, and Conger's kappa. All coefficients indicate intense levels of agreement and are statistically significant (see Table 4.8 for their numerical results and p-values).

Finally, in line with what we did for the AMT crowd annotations exercise on the MDC, we measured the correlation coefficient between the retrieval model result of the current DM and the interactors' ratings, and the result is a correlation coefficient of 0.44 (p-value:  $1.03 \times 10^{-153}$ ). The stronger correlation vs. previous models shows higher agreement between the model results and the interactors' opinions.

## 4.5 Summary

We explored the challenge of defining a ‘good’ answer in the context of a TOIA by evaluating human-*TOIA avatar* dialogues. We learned that the perceived ‘right’ answer for avatar interrogators differs from the avatar maker expectations partly because some questions and answers are too versatile, i.e., they go well with many answers and questions, respectively. Additionally, yes and no answers are

equally perceived as relevant by users interrogating an avatar but would be right or wrong for a given avatar maker. We make all the human annotations we collected available to the research community.

We challenged classical retrieval metrics and proposed that TOIA’s dialogue managers should optimize Success Rate @1. Success Rate @ $k$  for different levels of  $k$  can help identify how to improve retrieval techniques. In particular, SR@1 results informed our upgrade decisions in the answer retrieval methodology for the final software implementation presented in the next Chapter.

Finally, while the technical approach taken in the most recent answer-retrieval development for our TOIA is described in detail in the next Chapter (5), we presented how the prototype human evaluation methodology studied with the AMT raters was carried across our TOIA. We, therefore, showed the results of an evaluation performed by human subjects interacting with real *TOIA avatars*, which achieve a satisfactory level for real-world, functioning time-offset interactions.

# Chapter 5

## TOIA Design and Development

There are challenging limitations in the experience of creating time-offset interactions and self-narrative interactive systems in general, particularly the ones involving multimodal interaction like video, voice, text, and clicks (or taps on a mobile interface). Long hours, intense efforts, and cumbersome user experiences for the *TOIA avatar*-maker make it hard for such tools to become mainstream. They also make it expensive and creatively frustrating to create corpora and *TOIA avatars* available for research into dialogue systems and new modes of interaction. In this chapter, we describe the Time-Offset Interaction Application (TOIA) we developed in NYUAD’s Computational Approaches to Modeling Language (CAMEL) Lab as a collaborative effort led by the author. Software components have been developed and maintained by the author, a few undergraduate students working on the software as part of their capstone project, and some students working as research assistants for the lab. The latest release is open-sourced on GitHub.<sup>1</sup>

---

<sup>1</sup><https://github.com/CAMEL-Lab/TOIA-2.0>

While a similar system was released in recent time for commercial use,<sup>2</sup> the overarching goal of our research is to democratize and streamline TOIAs. Towards democratization, we make our system open-source for researchers and individuals interested in building interactivity experiments, conducting user-experience studies, and improving the technology collaboratively. Furthermore, along describing our implementation, we present a number of tactics we developed to streamline the experience of creating *TOIA avatars* and interacting with them.

Our recent work [1] and proofs-of-concept (PoCs) [10, 11] have concentrated on studying critical components of the *TOIA avatar* creation process in an attempt to make a TOIA more usable and accessible without incurring massive production costs and time frames. Systems like the one proposed by Artstein et al. 2015 [5] require recording about 2,000 video answers, sophisticated technical equipment, and post-editing CGI effects. In Chierici et al. 2020 [10], we proposed a faster *TOIA avatar* development process, but it is still impractical for the everyday user: it involves transcribing and recording conversations based on brainstormed plausible utterances, a process that results in creating more than 400 video recordings and manual annotations that take several days to complete. Abu Ali et al. 2018 [1] focused on the possibility to chat with the avatars in different languages, but their system implementation is not simple to use because it has two separate, non-communicating components for recording videos and interacting with them. It requires local installation and does not support multiple users. Moreover, the dialogue manager was too limited and couldn't deliver a meaningful interaction with *TOIA avatars*. In this chapter, we propose a TOIA implementation that

---

<sup>2</sup>[www.storyfile.com](http://www.storyfile.com)

streamlines the creation of impressive and original interaction experiences with a pleasant, quick, and simple user experience.

Our contributions are as follows:

- (1) a TOIA recording interface that is enriched with a Large Language Model (LLM) for facilitating the *TOIA avatar* creation process by incorporating personalized user suggestions,
- (2) a TOIA player interface that enhances interaction with *TOIA avatars* by providing smart follow-up suggestions for dialogue and by implementing multi-modal interactions, i.e., speech, typing, or suggested questions, and
- (3) open-source software for deploying and creating TOIAs that are flexible enough for running user studies, adding and turning on-and-off different features. We welcome collaborators to improve the technology and develop new interactions via our GitHub repository.

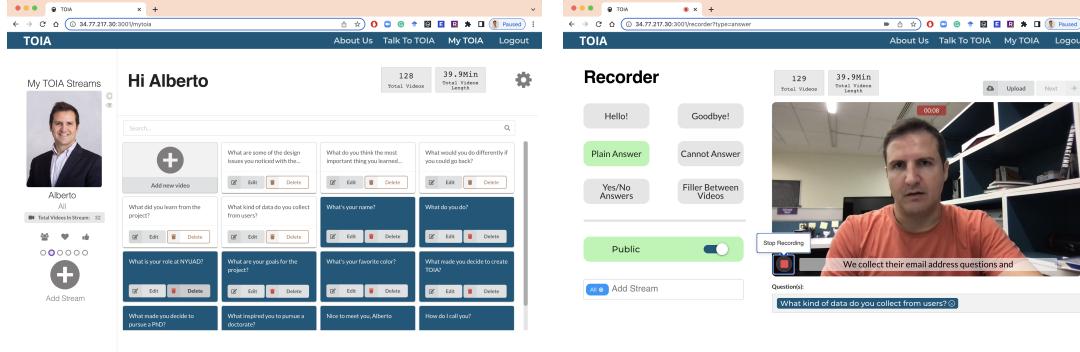
In the following section, we present the design principles we referred to for the creation of our TOIA’s interfaces. In Section 5.2, we show the system’s interface and describe the backend’s design and implementation. In Section 5.4 we discuss our question suggestion personalization tactics for streamlining *TOIA avatar* creation and engagement. In Section 5.5, we bring to life the interactivity experience for potential users.

## 5.1 Design Principles

We designed our TOIA so that the time-offset interaction feels like a natural interaction to all users: the user who interacts with other people’s pre-recordings (henceforth, *interactor*), and the user who records their own narratives (henceforth *TOIA avatar-maker*). As per best practices in UX design, each step of interaction with the interface should minimize the user’s cognitive load and be psychologically satisfying. We followed the human-computer interface heuristics established by Nielsen et al. (1994)[50]. Some design decisions to reach this goal included: the creation of a common visual framework across the whole system for all types of users; the use of familiar layouts and vocabulary; the creation of affordances by auto-generating suggested questions to answer and to suggest particular kinds of interactions; providing psychological satisfaction with profiles that emphasize social interaction; giving users agency and autonomy over aspects of their TOIA experience; and providing reassurance with repeated language and layouts.

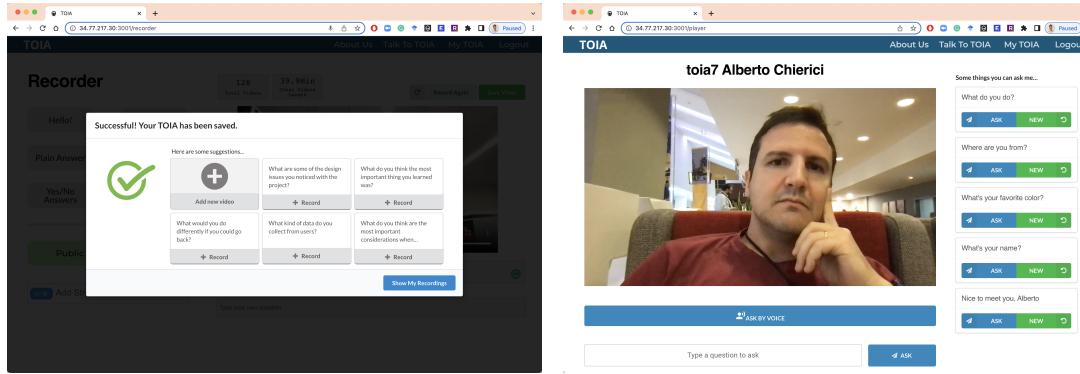
The designed experience also avoids indicators that this is a digital system (e.g., removing frozen video frames, creating seamless video stitching, continuing to play video while the user types/speaks), while adding in some of the aspects of natural human conversation (such as “filler” videos for attentive nodding). These design principles give our TOIA a natural and intuitive user experience designed for universal usability.

An important extension to the previous work by Abu Ali et al. (2018) [1] is the introduction of a social network aspect. This does not just serve the goal of creating a community of *TOIA avatar* makers but also provides helpful feedback to



(a) User Account Page

(b) Recorder Page



(c) QS Pop-up

(d) Player Page

Figure (5.1) Screenshots of the user interface. (a) is the user account page, which shows the questions generated by QS in white backgrounds and previously recorded questions (and videos) shaded in blue; (b) shows the recorder page; (c) shows suggestions appearing in a pop-up window once the user completes a recording on the Recorder page; and (d) shows the player page.

them including what additional questions are asked, and what answers are liked or disliked. Another extension is giving the *TOIA avatar-maker* control over which recordings are playable individually and as part of *streams* (i.e. collections or albums) that define different contexts and intentions for the interactions.

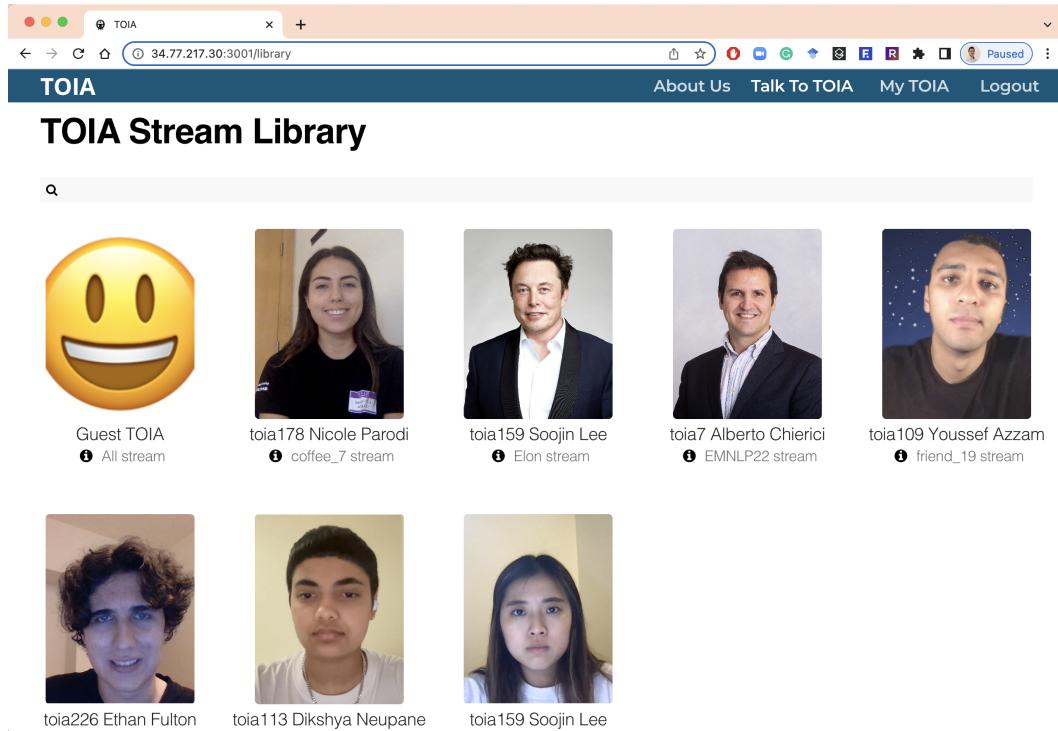


Figure (5.2) A view of a number of public TOIA streams that are ready for interactors to converse with.

## 5.2 System Interface and Backend

This section provides a description of the user interface and the software components of the system. We highlight updates, refinements, improvements, and the resolution of previously identified issues from earlier PoCs [10, 11].

### 5.2.1 Interface

The user interface (UI) main components are: the user account page, the recorder page, and the player page (Fig. 5.1).

1. **User Account** (Fig. 5.1 (a)): This is the initial page that users see after creating an account. It displays a persistent menu at the top of the screen, a

button to create new videos, suggested questions for creating new videos, and videos previously recorded by the user. It also includes a user-settings menu and statistics such as the number of recorded videos and the total duration of the recordings.

*Streams* To the mid-right of the figure, there is a link to a management page for the streams associated by the *TOIA avatar*-maker. Each stream can get its own profile that specifies the functional purpose of the stream, e.g., a *business* stream may target job interviews, while a *family stream* can focus on sharing family histories. Streams can be made public or be only shared with specific users. All viewable streams can be accessed on the *Talk to TOIA* page (see Figure 5.2).

*Videos* The bottom half of Figure 5.1 (a) shows a collection of recorded videos (blue-shaded cards), and some entries with questions and no recordings (white cards). New questions are presented as videos that have no content and are waiting to be filled. The *TOIA avatar*-maker can delete existing video entries or suggested questions, or click or *Edit* to record a response or change an existing response. The *TOIA avatar*-maker can also create a completely new video by clicking on the big plus sign to the top left (preceding the first empty card) of the figure. Edit and Add actions will take the *TOIA avatar*-maker to the recorder view. All videos in the *TOIA avatar*-maker's account can be filtered for display using keyword search.

2. **Recorder** (Fig. 5.1 (b)): This page is accessed from the persistent menu or by clicking the button to add a new video on a previously recorded video or a

suggested question in the User Account page. This is where users can create new videos by selecting the type of video they want to record (e.g., plain answer, yes or no answer, filler), typing a question, and hitting the record button. The system automatically transcribes what the user says, and the user can edit the transcriptions before saving the video. Once a video is saved, the user interface shows a pop-up menu (Fig. 5.1 (c)) with the command for creating a new video and follow-up question suggestions.

3. **Player** (Fig. 5.1 (d)): The player interface can be accessed through the *Talk to TOIA* page, which lists all the publicly viewable streams which the interactor user has access to (see Figure 5.2). The player interface is intentionally as simple as a *Facetime*, *Zoom* or *Skype* interface. Here, users can interact with previously recorded videos of public *TOIA avatars*. The player interface comprises a video looping the different fillers recorded by the avatar. It displays on the right side of the video some suggestions of questions that the user can click and get asked directly to the *TOIA avatar*. At the bottom, there's a button to interact with the *TOIA avatar* by voice, and below that button, a text input where the user can type questions.

### 5.2.2 Backend

The backend system hosted in the cloud is composed of four components (Fig. 5.3).

1. **Database:** We use a relational database management system to store user information, language preferences, the links to, and meta-data of, all the user video recordings corresponding to the suggested questions. We also track the

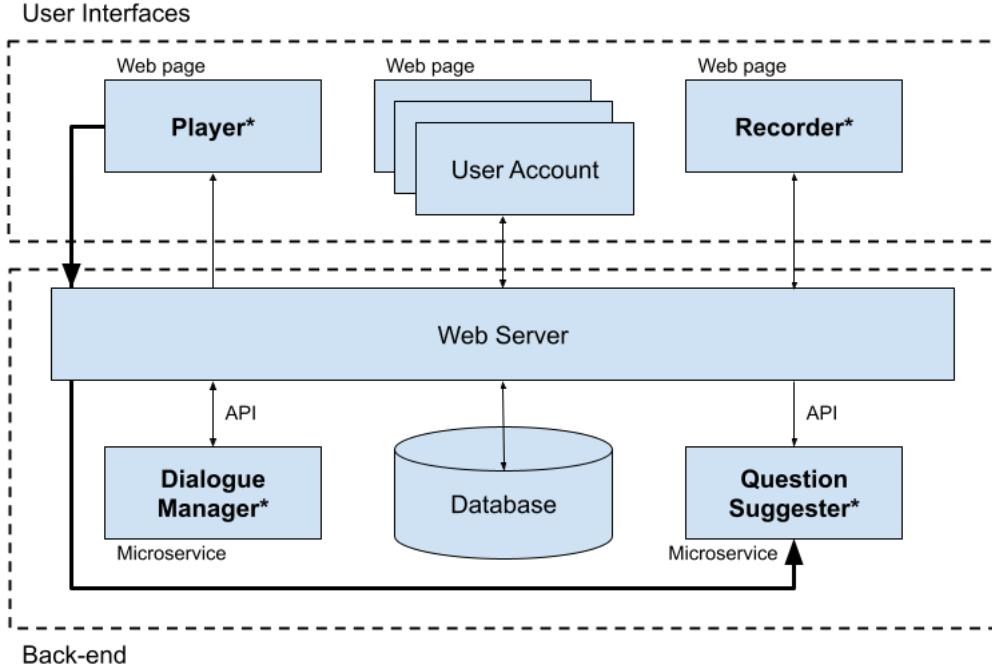


Figure (5.3) System architecture. In bold and starred, the components we improved from the system proposed in [11].

conversation logs with the player and user-tracker statistics such as how long a user spends on the recorder, user session timestamps, and so on.

2. **Dialogue Manager (DM)**: This microservice returns the video ID of the video recording containing the response to a question asked by a TOIA interactor. We provide more details in Section 5.3.
3. **Question Suggester (QS)**: This microservice provides personalized suggested questions that can assist the *TOIA avatar* creation and interaction processes. As such it is integrated with both the recorder and the player interfaces. We built the personalized suggestion suggester using GPT-3 models due to their impressive results in generating sensible and consistent text, their API availability for rapid implementation and testing, and their suitability

for an application that is intended to scale to many users. We discuss the QS further in Section 5.4.

4. **Web Server:** This is the system backend orchestrator. It is implemented in NodeJS and connects all the API calls to and from the different components. To minimize the effort of the *TOIA avatar* makers, we include ASR software for transcribing their answers in real-time. We opted for the Google Speech-To-Text API due to its performance and ability to include punctuation in the transcription. This component doesn't show in Fig. 5.3 because it's part of the Web Server, and we didn't need to develop a separate microservice ourselves.

**Improvements and changes over initial PoCs** Inspired by the original system proposed by Abu Ali et al. (2018)[1], we completely redesigned and developed from scratch the front-end components for our TOIA. With the aid of user experience mapping, affinity diagrams, and iterating design components with wireframes, we reached an end product that adheres to functionality and usability and facilitates a seamless user experience.

While we described the most important interfaces above, we also developed a home page, login and sign-up screens, and various menu and navigation aid items that we don't find necessary to describe in detail. All users (*TOIA avatar*-makers and Interactors) must register and have an account to be able to make their own *TOIA avatars* or converse with public *TOIA avatars*.

We have made a number of changes and improvements over the version of the

software reported in recently published work [11]. We extended the player interface by adding typing and clicking on suggested cards to speed up interaction and make up for limits in the Dialogue Manager. We upgraded the DM with a more high-performing language model to retrieve the videos. We redesigned the QS using a better LLM that could suggest coherent utterances with faster performance and introduce topic de-clumping strategies.

### 5.3 Dialogue Manager

In the context of a Time-Offset Interaction Application (TOIA), The Dialogue Manager (DM) retrieves the best video answer to a question. It is a crucial component as it directly impacts the experience of interacting with a *TOIA avatar* when the dialogue is executed by voice. Small mistakes in such a task would drive users away out of frustration. Advancements in NLP have evolved from retrieval-based models to large language models (LLMs). LLMs allow retrieval methodologies to leverage sentence embeddings derived by training generative models.

Since the initial explorations of retrieval methodology for developing the DM, we pivoted the research angle from retrieval shortcomings to creating richer and more efficient ways of interactions. Software production and the user evaluation presented in Chapter 4 influenced our choice for two main reasons. First, LLMs' API availability and performance improved exponentially and reduced the need to develop a custom-made solution. Second, user experience tests, PoCs, and more extensive studies (see Chapter 7) showed that users preferred a faster interaction by clicking and typing (with auto-completion) on the user interface. These modalities

used just below 60% of the time (again, CFR Chapter 7) vs. interaction by speech (north of 40%), do not require a DM activation as the questions displayed are already in the database. The system looks up and retrieves a video answer without machine learning or statistical algorithms.

For the retrieval task, we implement the OpenAI’s GPT-3 models family designed for text search using semantic similarity (Neelakantan et. al, 2022)[48],<sup>3</sup> due to their outperformance on the Margarita Dialogue Corpus analyzed in Chapter 4.

Documents are question-answer pairs combined and transformed into 1024-dimensional vector embeddings using the ‘text-search-ada-doc-001’ model, and a new question (or query) coming from the interactor is transformed into a 1024-dimensional vector embedding using the ‘text-search-ada-query-001’ model. The DM proposes an answer if the cosine similarity between the query and document vectors is above 0.29. Below this threshold, the DM outputs a random selection of videos we require to record for lack of response scenarios (e.g., “I haven’t recorded an answer for that question”).

## 5.4 Question Suggestion Personalization to Stream-line TOIA Creation and Interaction

Based on an in-house pilot study using the previous version of the TOIA [11] Question Suggester, we observed that the text generated was not always coherent and it included too many questions on the same topic (‘topic clumping’), resulting

---

<sup>3</sup>Implementation guidelines are available at <https://beta.openai.com/docs/guides/embeddings/what-are-embeddings>

in confused or overwhelmed users. We describe below how we addressed these issues in this version.

### 5.4.1 Personalizing the Recorder’s Suggestions while Mitigating Topic Clumping

Topic clumping occurs when the text generator produces text on repetitive topics or focuses too narrowly on one topic. This happens when the model’s input is representative of the same topic or too similar topics. For example, consider the following model input from our experimental data: “*Q: What is your favorite food? A: I love Italian and Japanese food.*” “*Q: What about Japanese food do you like the most? A: The quality and freshness of ingredients.*” The model output was questions like: “*What is your favorite ingredient? What are some of your favorite healthy foods? Do you like sushi?*”. The more the user kept answering those questions, the more suggestions tended to repeat themselves or converge only on one topic.

One way of avoiding or reducing the risk of topic clumping is to include several dialogue turns in the GPT-3 prompt.<sup>4</sup> However, this would result in slower performance, which would not be suitable for real-time interactions with the interface. Consequently, to achieve a faster performance in real-time, we opted to design a brief prompt including the four most recent conversation turns recorded by the user (past 2 question-answer pairs), by which we request the model generator

---

<sup>4</sup>To work with the GPT-3, one needs to feed it with a text prompt that explains the task to fulfill. GPT-3 then completes the prompt by generating further text. For example, an example of a prompt is: “Write a tagline for an ice cream shop.” A possible output of GPT-3 is “We serve up smiles with every scoop!” GPT-3 is a statistical model, which means that it can generate different outputs as a response to the same prompt. For example, the following are two different outputs in response to the same prompt in the ice cream shop tagline example: (1) “The best ice cream in town!” (2) “A place where you can indulge in your sweet tooth without breaking the bank.”

to suggest five plausible questions for following up on the conversation. At the time  $t$ , the prompt would look like:

```
Suggest five plausible questions
for following up the conversation.
```

Q: <question<sub>(t-1)</sub>>

A: <answer<sub>(t-1)</sub>>

Q: <question<sub>(t)</sub>>

A: <answer<sub>(t)</sub>>

**Possible questions:**

We randomly pick 3 out of the 5 suggestions and display a maximum of 5 at a time to the user. In practice, whenever a user records a question-answer pair, QS gives that pair and the previous one to the prompt. QS then takes the GPT-3 output, parses the 5 suggestions, and randomly picks only 3. The user is shown the 3 suggestions, plus any remaining questions from the previous QS activation (one API call to the QS microservice). This design helps us avoid topic clumping by displaying to the user suggestions on topics covering two consecutive QS activations, which correspond to a sliding window of six conversation turns.

#### 5.4.2 Personalizing the Player’s Suggestions

As we observed users interacting with *TOIA avatars* using the Player, we realized that typing and clicking on questions provided a faster user experience, primarily preferred by users instead of speaking (which is further complicated by transcription errors). We included typing (described earlier) and cards to click on questions available in the database. However, we didn’t want to clutter the interface

with too many questions. We also wanted to show personalized questions: i.e., questions relevant to the current conversation a user is having with a *TOIA avatar*. This is achieved by integrating QS with the Player interface and developing a different API endpoint that generates a prompt specifically for suggesting questions to ask (rather than questions to answer as it happens on the Recorder page). The instruction is to understand the conversation and generate personalized follow-up questions that we have in our database, that are relevant to the conversation and that will help to continue the conversation naturally and coherently. At a time  $t$ , the prompt would be:

**Understand the following conversation:**

**Q: <question<sub>(t-1)</sub>>**

**A: <answer<sub>(t-1)</sub>>**

Select the 5 best follow-up questions from the following:

<question-retrieved-1>

<question-retrieved-2>

<question-retrieved-3>

...

<question-retrieved-50>

Q represents a question the user asks, and A represents the avatar's answer. The objective we give to the model is to select the five best follow-up questions from a list of 50 suggestions (question-retrieved-1, question-retrieved-2, etc.) based on the context of the previous conversation. The 50 suggestions are the top 50 retrieved utterances using the dialogue manager.

## 5.5 Interaction: Creation of, and Engagement with, TOIA Avatars

It's now the time to bring to life what the actual experience of creating *TOIA avatars* and interacting with them looks like using our TOIA. The application is accessible from any laptop (at the time of writing, the design isn't responsive yet and would be hard to use on a mobile device or tablet). Users can sign up and create an account by giving their first name, last name, email address, the language they intend to use for recording videos, and optionally, upload a user profile picture.

Once logged in, or once created an account, we invite users to familiarize themselves with the *TOIA avatar*-creation process first. The initial page they land on is the User Account page which prompts them to complete a list of mandatory video recordings (see Figure 5.4), which constitute basic utterances like “Hello”, “Goodbye”, “My name is...”, and filler videos.<sup>5</sup> It takes about 2 minutes to record them. After that, the user will receive automatic suggestions for follow-up questions to keep recording. Users can keep recording as long as they want.

At any point, users can choose whether they want to create their *TOIA avatar* or to interact with one. To interact with *TOIA avatars*, users first select the ‘Talk to TOIA’ menu, which leads them on the *TOIA avatars* library page, which shows them all the publicly available *TOIA avatars* they can chat with. Once selected one of them, users land on the Player page, where they can ask questions to the

---

<sup>5</sup>Filler videos play between interactions to give a sense of presence, e.g., the person adjusting her hair or nodding attentively.

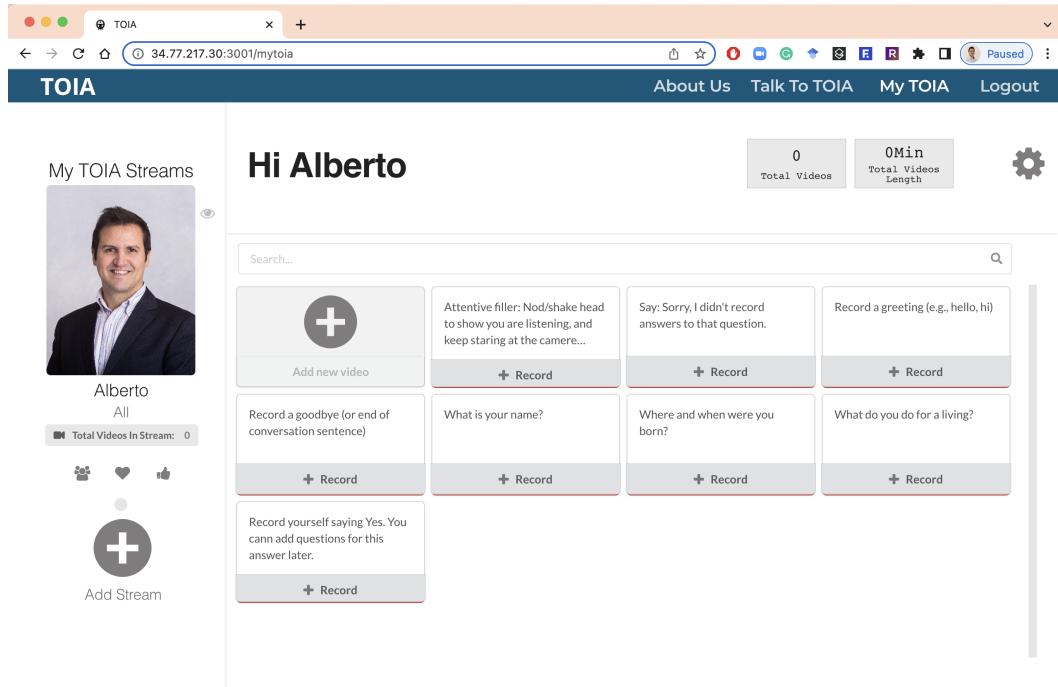


Figure (5.4) A view of the page showed at the first login of a user. The white cards underlined with a red line are mandatory and the user cannot record other videos before finishing this initial set of recordings.

*TOIA avatar* by voice, typing them on a bar and pressing Enter (or clicking on the button next to the typing bar, see Figure 5.1 (d)), or by clicking on the suggested questions popping up on the right side of the video feed.

## 5.6 Summary

We propose a TOIA that streamlines the creation of original interaction experiences with a pleasant, quick, and simple user experience. The system includes a Question Suggester that helps *TOIA avatar*-makers in the process of creating their self-narrative personas as well as enhancing the interaction with *TOIA avatars*. The system is intended to be an open-source tool put at the disposition of HCI and

dialogue systems researchers, to enable further developments and research projects using time-offset interaction, which can have some advantages over CGI-generated synthetic characters in some contexts.

# Chapter 6

## Creating TOIA Avatars

TOIAs potential applications span different areas such as digitally preserving personal narratives and historical memories [40, 75], interactive online teaching, simulated interviews, coaching, cognitive rehabilitation or simulated therapy sessions in the context of mental health, and digital copies of humans, among others. Despite the great potential of time-offset interaction in a wide range of Human-Machine Interaction (HMI) applications, these promising interfaces are still under-explored within the HMI community. Existing systems are either commercial<sup>1</sup> or designed for narrow use cases such as the creation of interactive copies of celebrities [6, 43, 74].

The development of such systems requires many resources (hours of recording, use of top-of-the-line digital cinema cameras, etc.). The implementations we proposed in recent years have concentrated on democratizing TOIAs in an attempt to make them available for researchers to advance research on the topic, as well as to allow any person to use the system for creating interactive self-narrative

---

<sup>1</sup>[www.storyfile.com](http://www.storyfile.com)

*TOIA avatars* of themselves [1, 10, 11].

One of the most crucial steps in creating a *TOIA avatar* is building a knowledge base (KB) consisting of questions and their associated answers (in the form of pre-recorded videos). The most immediate option would be to brainstorm questions and record oneself answering them. This task is challenging as it is difficult to brainstorm potential questions to record while trying to predict the hypothetical paths of future conversations, especially since the eventual interactor is unknown. In Chapter 3, we presented the process of creating the Margarita Dialogue Corpus (MDC) as an alternative to creating a *TOIA avatar* by recording conversations with strangers to solicit enough meaningful questions to record in the *TOIA avatar*'s KB – that process is not streamlined enough. It is highly impractical to ask any *TOIA avatar*-maker to undergo the same steps. However, the experience of making the MDC showed us that if a *TOIA avatar*-maker engages in dialogue with others, it generates a KB with meaningful questions that reflect real-life conversations. So, what if we could automatically generate questions so that the experience of recording a *TOIA avatar* looks like having a conversation with someone?

Several existing works have explored methods to suggest conversation topics to facilitate conversation in different human-machine interaction scenarios [26, 49, 69]. However, they mostly aim to enhance face-to-face conversation between strangers (e.g., ice-breakers). Moreover, none of the existing works focused on integral questions or phrase suggestions in TOIAs. Recent advances in Large Language Models (LLMs) have achieved tremendous success in several natural language generation tasks making it possible to auto-complete previously generated text. Specifically,

the transformer architecture has exhibited an improved language understanding as well as the flexibility to work on different tasks in terms of grammatical fluency and correctness [45], especially in the form of recent state-of-art tools such as OpenAI’s GPT (Generative Pre-trained Transformer) [56]. The latest (at the time of writing) GPT-3 language model has proved to be a great fit for open-ended text generation. Its decoder-based underlying architecture allows learning the task of predicting the next word, making it suitable for the task of question suggestions generation.

Research in behavioral sciences and human-machine interaction (HMI) has underlined the effect of personal factors such as the user’s gender, age, personality, or culture on the user’s experience in HMI [54, 64]. From the system’s perspective, specifically in embodied HMI (interaction with machine interface having physical or virtual embodiment like a TOIA’s interface), the system’s personality was also found to affect the user’s experience or engagement state directly [64]. This has motivated research towards personalized human-machine interaction, which considers the user’s personal factors in the design of human-machine interaction interfaces and the development of human-centered automatic detection models of behavior [65, 78].

These intuitions and literature review prompted us to develop **Question Suggester (QS)** – the GPT-3-based service that generates suggested questions to answer to the *TOIA avatar*-maker. We described its development and workings in Chapter 5. The aim of the QS is to facilitate the creation of *TOIA avatar* maker databases without exhausting their creators’ creativity, energy, and interest, a foreseeable obstacle in the user experience of TOIAs. In this Chapter, we present the user study we conducted to evaluate the value of QS against human-based

suggestions and a no-suggestion setup. To account for personal factors in the design of future TOIA systems, we investigate the effect of personality traits on the *TOIA avatar* maker’s user experience. Our ultimate goal is to design a user-aware QS process that works together with automatic personality identification to create the optimal user experience for *TOIA avatar* makers.

## 6.1 User Study Methods

We designed an experiment that involves qualitative and quantitative assessments of the QS feature’s user experience, time-saving power, and ability to suggest meaningful questions. The experiment was approved by our institution’s *Institutional Review Board*. In the following, we present the user experience study we designed to validate the utility of the QS. We describe the interaction scenario and the data collection procedure.

### 6.1.1 Experimental Conditions & Interaction Scenario

**Experimental Conditions** In order to study the effect of the QS feature on user experience, we created three experimental conditions: GPT-3 QS, Human-QS, and QS-off. The main aim is to understand how these three conditions impact the user’s rating of the experience, the time taken to create their set of video recordings, and whether there are significant differences between suggesting questions via GPT-3 generation or a human.

1. GPT-3-QS cohort: users who record their video utterances with a QS implemented using GPT-3;

2. Human-QS cohort: users who receive suggestions manually from the same *Wizard of Oz* researcher behind the scene but without knowing they come from a human;
3. QS-Off cohort: users who do not receive suggestions have to brainstorm the whole set of question-answer pairs.

We had to accommodate participants using the software on their schedules to preserve their creative space and work in different time zones. So, there is a time delay for the Human-QS cohort between using the system and the researcher manually writing suggestions into the system. The human researcher aimed to replicate the GPT-3 prompt method closely by reading a sliding window of four dialogue turns and brainstorming follow-up questions for each four-turn snippet. These were then entered into the Question Suggester's queue. However, the researcher faced difficulty generating five distinct suggestions for each dialogue snippet. They could only quickly produce one or two follow-up questions at a time.

**Interaction Scenario** The participants signed up for the experiment with an online form, including informed consent, that explained the study's overall goal and how the interactions and recordings would proceed. The form also included a section explaining the allowance the selected participants would receive after participating in the study: 300 AED (about \$USD 82) was paid to those selected to participate in the study. Although we are aware that paying participants may introduce information bias, we could not avoid it in our study. However, we experimented in a controlled manner and paid the participants a minimal amount to minimize their vested interest. Moreover, their qualitative feedback gave us strong indications of their overall appreciation of the experience and genuine interest in using our TOIA.

We restricted the participant pool to university students. This choice is because we wanted an easy-to-access and (to a certain extent) homogeneous pool of users. 104 individuals signed up to participate in the experiment. Among those who signed up, we performed stratified random sampling in a way to obtain a consistent sample in terms of gender, age, year, and field of study across the three experimental condition cohorts. Based on the stratified sampling, 73 participants were invited to take part in the study. Among these, only 35 accepted the invitation, of which 9 withdrew, which resulted in a total of 26 participants (14 female). The age range of the participants was 18 – 24. All participants were fluent in English, with 9 having English as their mother tongue. The participants majored in either Humanities (Arts, Interactive Media), Engineering, Science (Mathematics, Biology, Computer Science), or Social Science (Psychology, Economics, Political Science), with the majority coming from science (12 participants). Table 6.1 summarizes the statistics on participants' demographics per cohort.

The subjects were instructed to use the TOIAinterface presented in section 5.2 in order to create a video database for their own interactive *TOIA avatar*. To help the users familiarize themselves with the interface, the interface prompts users to complete a list of mandatory video recordings which constitute basic utterances like “Hello”, “Goodbye”, “My name is...”, as well as filler videos. Filler videos play between interactions to give a sense of presence, e.g., the person adjusting her hair or nodding attentively. We asked participants to record attentive (looking at the camera, showing attention to the speaker) and inattentive fillers (looking elsewhere, distracted) as an additional experimental condition we want test at the interaction stage – see Chapter 7. We exclude the mandatory videos (a total of 29) from our

analysis. So, for participants who recorded the minimum amount of 100 videos, we count 71 videos only.

The subjects conducted the user-interface interactions remotely using their laptops at their own convenience and convenience. They had to participate in a mandatory webinar where one researcher demonstrated the system to them and sent all participants written, detailed instructions and troubleshooting options by email. The participants were instructed to screen-record their user-interface interaction sessions in order to allow researchers to observe them.

In order to ensure consistency among the different user interactions and to avoid potential user fatigue, the following experimental parameters were controlled.

- The experiment duration was limited to 10 consecutive days for each user.
- We instructed participants to record a minimum of 100 video utterances (including the mandatory ones). No constraints were put on the allowed maximum number of recorded utterances.
- We instructed participants to limit the duration of a single user-interface interaction to a maximum of 1 hour.
- We instructed participants to limit their user-interface interaction to a maximum of 3 sessions per day.

**Collected Data** We collected the video recordings of the *TOIA avatar* makers, the recording timestamps, the user session timestamps, text transcriptions of

	Participants	QS Mode			All
		GPT-3	Human	Off	
<b>Age</b>	18	1		2	3
	19	3	3	4	10
	20	1	2	2	5
	21	2	3		5
	22	2			2
	24	1			1
<b>Gender</b>	Female	5	4	5	14
	Male	5	4	3	12
<b>Mother Tongue</b>	English	3	3	3	9
<b>School</b>	Other	7	5	5	17
	Humanities	1			1
	Engineering	1	2	1	4
	Science	5	2	5	12
	Social Science	3	4	2	9

Table (6.1) Participants demographics by cohort.

questions and answers, and zoom recordings of user-*TOIA avatar*-interactions. We show summary, and descriptive statistics of this corpus in Appendix A, Table A.1.

**Post Interaction Survey** After completing their *TOIA avatar* recordings, the participants completed a follow-up survey probing for qualitative and quantitative feedback on the experience. The full survey questions are shown in Table A.3. The user experience questionnaire aimed to gauge users' perceptions of and experience of the interface. Qualitative items sought to gather descriptive feedback on strengths and weaknesses, while quantitative items measured satisfaction with the software's performance, such as the frequency of transcript retyping.

To investigate the effect of personality traits on the *TOIA avatar* maker's user experience, the survey also included a 10-item short version of the Big Five Inventory personality questionnaire [58]. The aim is to check for possible personality

biases—for instance, a more sociable individual may be inclined to record more videos and provide longer answers).

## 6.2 Experimental Results and Discussion

In this section, we present the results of the study. We describe qualitative and quantitative results supporting the case for adopting QS in creating time-offset interactions. We evaluate and compare the *TOIA* *avatar*-maker interaction with the interface across the three conditions: (1) No QS, (2) GPT-3 QS, and (3) Human QS. We look at the effect of the proposed QS on different aspects: the total number of recorded videos, the recording experience, suggestions acceptability by the users, and topic clumping. We also investigate the correlation between personality traits and user experience. Table 6.2 presents the descriptive statistics of the experiment’s output.

### 6.2.1 QS Effect on the Total Number of Recorded Videos

On average, the QS-Off cohort recorded 82.1 (StDev: 7.0) videos per user, the GPT-3-QS cohort recorded 93.7 (StDev: 31.1) videos per user and the Human-QS Cohort 91.5 (StDev: 13.0). As expected, the average user recorded more videos with the suggestions on but, by conducting t-tests for testing the null-hypothesis that the means between pairwise cohorts are the same, we find out that there isn’t enough evidence to reject the null-hypothesis when comparing the QS-Off cohort vs. the GPT-3-QS cohort. That is perhaps due to the high variance in the GPT-3-QS cohort. However, we can reject the null hypothesis when comparing the QS-Off cohort vs. the Human-QS cohort’s manually made suggestions (p-value: 0.007).

	GPT-3-QS	Human-QS	QS-Off	Overall 104
#Users	10	8	8	26
#Videos	937	732	657	2326
min	71	78	71	71
avg	<b>93.7</b>	<b>91.5**</b>	82.1	89.5
max	<b>172</b>	<b>110</b>	93	172
StDev	31.1	13.0	7.0	20.8
% from QS	<b>51.6%</b>	<b>54.0%</b>	0.0%	na
#Generated Suggestions	2491	532	0	na
#Accepted Suggestions	690	460	na	na
%	<b>28%</b>	<b>86%</b>	na	na
Ratings on a 1 to 5 scale: Average (StDev)				
Overall Experience	<b>4.6**</b> (0.52)	<b>4.4*</b> (0.52)	4.0 (0.76)	4.4 (0.63)
Recording Videos	4.3 (0.82)	4.3 (0.71)	4.6 (0.52)	4.4 (0.70)
Deleting Videos	4.4 (0.84)	4.0 (0.76)	4.0 (1.2)	4.2 (0.92)
Editing Editing	3.3 (0.82)	3.4 (1.30)	2.6 (0.74)	3.1 (0.99)
Quality of ASR Transcriptions	2.7 (0.95)	2.1 (0.64)	3.0 (1.07)	2.6 (0.94)
Overall Software	4.0 (0.67)	3.8 (0.46)	4.38 (0.52)	4.0 (0.60)
Tot videos duration (m)	249	230	139	619
Min (s)	1	1	1	1
Avg (s)	<b>16</b>	<b>19</b>	13	16
Max (s)	118	239	81	239
StDev (s)	14	25	12	18
Perceived average experiment duration				
per user (h)	<b>15.5</b>	<b>11.6</b>	8.2	12.0
per video (m)	<b>10</b>	<b>8</b>	6	8

Table (6.2) Summary of highlighted experimental results. All metrics exclude onboarding videos (fillers and mandatory question-answer pairs). The number of questions may be larger than the number of videos linked to such questions because users could record one video as an answer to multiple questions (e.g., answering ‘yes’ or ‘no’). We highlight in bold noteworthy results that we discuss more in detail in section 6.2 and mark with \*\* the statistically significant results when tested against the QS-Off cohort pairwise at a significance level of 0.05 and with \* at a significance level of 0.1.

We have enough evidence to show that suggestions indeed aid the creative process of recording the interactions typical to TOIAs. Looking at the max number of videos recorded, we can see that outliers like 110 and 172 were observed only in

<b>Personality Dimension</b>	<b>QS Mode</b>			
	<b>GPT-3</b>	<b>Human</b>	<b>Off</b>	<b>All</b>
Extraversion	3.4	2.6	3.1	3.1
Conscientiousness	3.6	3.3	3.8	3.5
Neuroticism	2.5	2.8	2.9	2.7
Openness	3.3	3.6	3.5	3.4
Agreeableness	4.0	3.6	3.6	3.7

Table (6.3) Participants average ‘Big-5’ personality scores by cohort.

the cohorts with a version of QS. The suggestions make it possible to go well above the minimum of 71 videos requested by users (recalling that for results reporting purposes, the actual minimum is 71 because we exclude mandatory and filler videos) and not deteriorate user experience.

### 6.2.2 QS Effect on the Recording Experience

When asked *‘Describe your experience with coming up with new videos to record. Did you struggle to come up with inspiration, and why?’*, all participants in the QS-Off cohort answered affirmatively, often mentioning going blank rather quickly and having to search on Google typical questions to ask a stranger. One user in this cohort reported: ‘I sometimes struggled and I think the problem was a one-sided conversation. I did not have someone who would guide the flow of the conversation so I got the impression that I am talking with the wall’. Interestingly, many participants in the GPT-3-QS and Human-QS cohorts mentioned the suggestions were helpful, but after a while, they stopped accepting suggestions because they were either repetitive or focusing too narrowly on a topic (topic clumping). However, when asked *‘Describe briefly your experience using the software. What did you find working well, and what areas we shall improve?’*, the participants in the GPT-3-QS and Human-QS cohorts praised the suggestions as a way to help them creatively

come up with questions. So, it looks like there are some limitations but users welcomed the suggestions.

### 6.2.3 QS Effect on Overall User Satisfaction

Regarding satisfaction with the overall experience (*‘Rate your experience with the technology [Did you enjoy your experience creating your avatar?’, 1–5 Likert scale]*), the GPT-3-QS cohort gave an average rating of 4.6, the Human-QS gave 4.4, and the QS-off cohort gave 4.0. As expected, the difference between the ratings given by the Human-QS cohort and the GPT-3-QS cohort is not statistically significant (p-value: 0.158). Curiously, while the difference between the GPT-3-QS and the QS-off cohorts is statistically significant (p-value: 0.044) at  $\alpha = 0.05$  significance level, the difference between the Human-QS and the QS-off is statistically significant (p-value: 0.099) by increasing the significance level to  $\alpha = 0.10$ . This is likely due to the high variance in ratings given by the QS-off cohort, so the steeper difference between the averages of the GPT-3-QS cohort and the QS-off cohort may give more of a signal compared to the smaller difference between the averages of the Human-QS cohort and the QS-off cohort.

While we encouraged and welcomed feedback on aspects we were not considered during the experiment, we highlight that participants did not raise any worrisome issues. We only received praise for the ability of the system to work on self-reflection. Three users stated: “While making the videos, I was able to learn a few new things about myself by answering the questions I had come up with, so it really did double as a tool for introspection,” “I felt like I got to know myself better creating this avatar and I would love to get to know other people through their avatars,” and

“It was also a fun way to get to know myself.”

#### 6.2.4 ASR Correction Effect

According to the participants, most problems with the software were caused by the low performance of the ASR implemented, which got poor ratings across all cohorts (Table 6.2). As highlighted both by quantitative and qualitative user responses to the survey, ASR errors led to re-typing many answers, hence making the recording process more time-consuming and frustrating. Unfortunately, the under-performance of the ASR seems to bias the time measurements. The GPT-3-QS cohort reported an average perceived time of 10 minutes per video (estimated from the survey question ‘*How long did it take you to create your avatar?*’), the Human-QS cohort 8 minutes, and the QS-Off cohort 6 minutes. These times reflect the overall use of the software, including thinking about what to do, typing, editing the ASR transcriptions, etc. The QS suggestions may have triggered more thoughtful and unexpected questions, leading *TOIA* *avatar*-makers to take more time to think about what to answer, eventually recording longer answers, therefore taking a long time to edit the ASR transcriptions. In fact, the average single video duration in the GPT-3-QS is 16 seconds, 19 seconds in the Human-QS cohort, and 13 seconds in the QS-Off cohort.

#### 6.2.5 Suggestion Acceptability

On average, of all recorded videos in the cohort with GPT-3-QS, 51.6% are questions coming from the recommendations, while the remaining 48.5% were brainstormed. The proportions for the cohort with Human-QS are 54.0% and 45.0%, respectively. There is no significant difference in the proportions of questions

suggested by the automatic QS (GPT-3) and the Human-QS when comparing the two groups ( $p$ -value: 0.439). However, a different pattern emerges when considering the acceptance rate of suggestions. The acceptance rate refers to the proportion of suggestions made by the system that was actually selected by the user. Notably, the acceptance rate of suggestions generated by the GPT-3-QS was only 28%, compared to an acceptance rate of 86% for suggestions generated by the Human-QS. In this analysis, we can't make conclusions about causation, but the results show that GPT-3 could be as good as a human suggester in helping the user formulate questions rather than brainstorming them all and that GPT-3 generates many more suggestions that eventually get ignored by users vs. a human who can suggest fewer, yet more relevant questions.

**GPT-3 QS Suggestion Edits** In order to validate the quality of the GPT-3 QS generated questions, we tracked if users edited the suggested questions and ran a grammar-and-spell checker over them. The results show that no single question was edited by the user. And in fact, no grammatical or spelling errors were found. This confirmed that GPT-3 is able to generate grammatically correct and fluent suggestions.

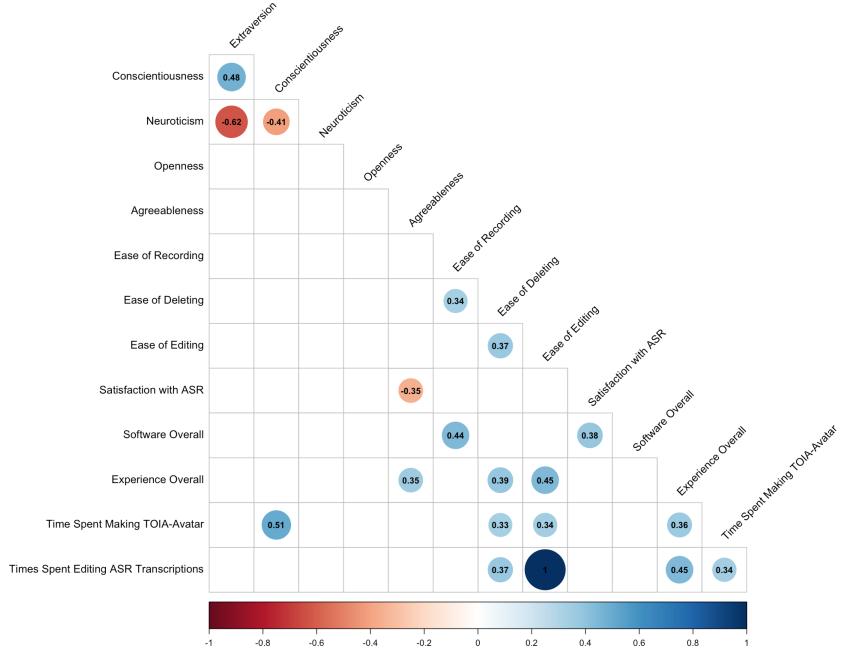
### 6.2.6 Topic Clumping

In order to evaluate the efficiency of the implemented topic clumping mitigation strategy, we asked the participants to report whether they experienced topic clumping in the suggested questions to them. Participants stated that topic clumping happened after recording some videos, in a range spanning between 30 and 80 recordings. That is partly due to the participants forgetting to delete suggestions

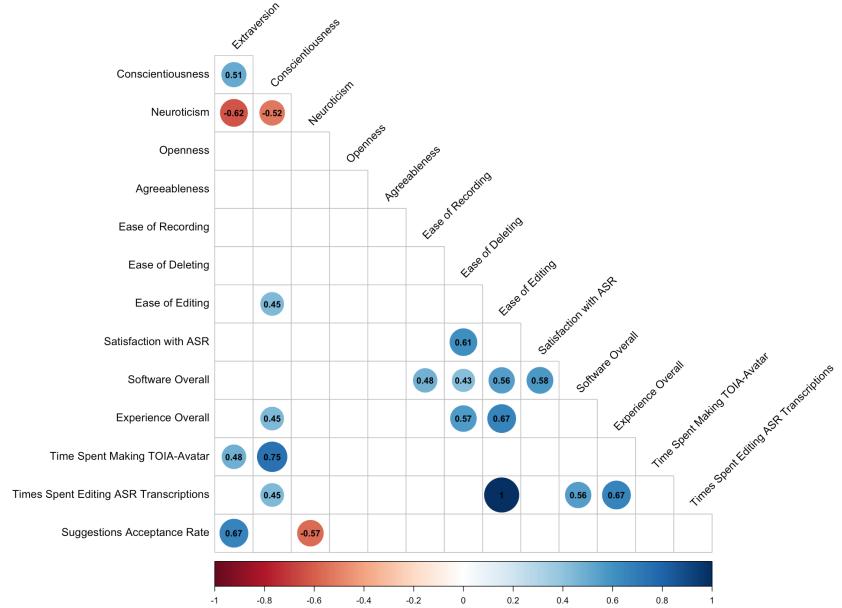
as they were instructed and partly to the sliding window over conversation turns between consecutive QS activation making the GPT-3 model see the same conversation turn twice. Though mitigated by randomly selecting 3 out of the 5 questions suggested, two consecutive QS API calls may end up proposing three follow-up questions to the same conversation turn. Again, this would be mitigated by deleting suggestions and moving on, but not all users understood that immediately. In the future, we plan to consider other strategies that navigate between staying on topic and avoiding topic clumping.

### 6.2.7 Effect of Personality Traits on User Experience

Figure 6.1 presents the meaningful correlation analysis between the Big Five Personality scores and user experience dimensions. We analyze correlations using data from all conditions (Figure 6.1 (a)) and using data corresponding to the GPT-3-QS and Human-QS cohorts (Figure 6.1 (b)). We only report correlations with significant p-values ( $< 0.05$ ). Regarding the overall experience, the results show that personality traits are poorly or not significantly correlated with the overall software and experience ratings. The only exception is a mild and statistically significant positive correlation of 0.35 between agreeableness and overall experience, indicating that people with higher agreeableness tend to give higher experience satisfaction ratings. The result differs when filtering only the cohorts who had suggestions: while the overall software rating does not correlate with personality traits, the overall experience ratings positively correlate (0.45) with conscientiousness. This result may partly validate that personality types do not have a material impact on the results of this experiment and partly that we did an excellent job in assigning the participants to their cohorts. Average personality scores are pretty consistent



(a) All data



(b) Question suggestions only: GPT-3-QS and Human-QS cohorts

Figure (6.1) Meaningful correlations between personality traits and user experience dimensions. Only correlations with  $p$ -values  $< 0.05$  are shown. Left: correlations are computed using data from all 3 conditions. Right: correlations are computed using data corresponding to the GPT-3-QS and Human-QS cohorts.

across cohorts, especially when comparing cohorts who had suggestions (aggregating QS-GPT3 and QS-Human data) vs. cohorts without (Table 6.3).

Interestingly, when we filtered the data on participants who used suggestions, we found significant correlations of personality dimensions with the suggestions' acceptance rates. We can observe an acceptance rate highly and positively correlated (0.67) with extraversion and a negative correlation with neuroticism (-0.57). It seems that participants with a more substantial degree of extraversion tend to accept more suggestions vs. participants with a higher neuroticism dimension, who tended to accept fewer suggestions.

Regarding other meaningful correlations with personality scores across all cohorts, we observe a mild, statistically significant correlation between agreeableness and satisfaction with ASR (-0.35). More agreeable people may be more disappointed with ASR transcriptions that tend to be wrong. Looking only at the cohorts with suggestions, we can observe some interesting correlations: conscientiousness with ease of editing (0.45), experience overall (0.45), time spent making *TOIA avatar* (a steep 0.75), and times spent editing ASR transcriptions (0.45). A high degree of conscientiousness relates to doing a thorough job. Hence, these participants spent longer making videos and made more edits to the ASR transcriptions. In summary, they engaged more with the experience of making their *TOIA avatar*, eventually resulting in higher satisfaction scores.

Finally, it is worth pointing out other compelling correlations that are not associated with personality traits. The overall *TOIA avatar*-making experience positively correlates with ease of deleting (0.39) and editing (0.45) videos, time spent making the *TOIA avatar* (0.36), and time spent editing ASR transcriptions (0.45), hinting that these interactions with the software and the time spent engaging

with it had a material, positive impact on the user experience. Counter-intuitively, the editing transcription positively correlates with overall satisfaction with the experience instead of negatively correlated. A confounding effect could be a participant's active interest in spending longer engaged with the process. The two main components that seem to have a material impact on software overall are ease of recording and satisfaction with the ASR transcriptions, positively correlated with coefficients 0.44 and 0.38, respectively, for all cohorts (6.1 (a)). Curiously, when looking at the cohort with suggestions (6.1 (b)), deleting and editing videos have high correlations, too, with the overall software rating (0.43 and 0.56, respectively). One possible interpretation is that participants who had suggestions had more often engaged with actions like deleting videos or edited answers because they may have had second thoughts about the suggestions or edited how to answer on the spot. Participants who brainstormed questions, having already prepared a question and an answer in their mind, may engage less with deleting or editing videos.

### 6.3 Limitations

Although the present study provided some insights into the experience of an avatar-maker using the TOIA platform, it also presents some limitations, which we discuss in the following.

**Statistical Power** We know that the data size's statistical power is low. However, we were limited by the participant pool, the stratified sampling of participants to invite, and the budget (cost was about \$80 per participant) as described in Section 7.1. The low statistical power did not impair the experiment findings as we draw meaningful conclusions by selecting a significance level  $\alpha = 0.05$  on all the

t-tests. And the findings are further corroborated by the qualitative information we collected and documented extensively.

**Discarding Suggestions** Early testing of the system showed that users preferred to see a limited number of suggestions and felt overwhelmed if they covered too much screen space. So we reduced the number of suggestions shown to the user at any time and instructed participants that they had to delete suggestions they did not want to use to see the next ones. Not all participants understood this immediately. As a result, some participants in the cohort receiving suggestions may have spent some time brainstorming answers instead of using suggestions. All the experimental sessions were recorded, and when the researchers noticed this behavior, they intervened, reminding participants to delete unwanted suggestions.

**Human vs. GPT-3** A few experimental design choices limit the comparison between the Human-QS implementation and the GPT-3-QS. While GPT-3 suggestions were real-time, the human ones most often were not. Due to the nature of the user experience study, we had to accommodate participants to use the software on their schedules. When the researcher could be available, he inserted suggestions in real-time, but most often, there was a delay between user sessions and the creation of follow-up suggestions. This was mitigated by creating enough suggestions to keep the user seeing them, always five at a time, during an entire recording session. We also instructed users to limit their recording sessions to one hour so the researcher had enough time to read their conversation turns and make suggestions. Not everybody respected this requirement, as two users crammed everything last minute to meet the experiment deadline. On the other hand, human suggestions were naturally free from repetitions or topic clumping, as confirmed by the significantly

higher acceptance rate than the GPT-3 suggestions.

**Language** The experiment is conducted in English. We cannot assess if its findings can be transferred to other languages with complex morphology and no GPT-3 implementations.

**Ethical Considerations** GPT-3 models could come up with unpleasant or triggering questions that touch on open nerves. The creation of these systems could have harmful ways to be used and need to implement robust privacy processes that respect the user’s right to be forgotten and preserve their privacy according to their preferences. For instance, two suggestions in Figure ?? appeared to one researcher, making him uncomfortable. Speaking of his kids, the ‘Do you get to see them often?’ suggestion gave him the impression that the system assumes the kids are either old and away or the parents are separated. The question ‘What would you do if (name redacted) was ever sick?’ may trigger topics that are uncomfortable or memories that someone would not want to face at that moment. Moreover, it raised the concern that if a user shares private details like their kids’ names, the system would know that information and perhaps use it inappropriately.

## 6.4 Summary

There are important limitations in the experience of creating time-offset interactions. Long hours, efforts, and cumbersome user experiences for the *TOIA* avatar-maker make it hard for such tools to become mainstream. They also make it expensive and creatively frustrating to create corpora and avatars available for research into dialogue systems and new modes of interaction. In our TOIA, we

developed *Question Suggester* – an essential, intelligent element to add to TOIA systems. The QS helps *TOIA avatar* makers in the process of creating their self-narrative personas. We conducted a user experience study to evaluate the potential of QS against human-based suggestions and a no-suggestion setup. We included in the study an investigation of the effect of personality traits on the *TOIA avatar* maker’s user experience to account for personal factors in the design of future TOIAs.

Overall, the results show that the QS feature improves the *TOIA avatar* makers’ experience and the creative process. Personality traits were also found to be correlated with some user experience dimensions such as suggestions acceptability.

As we move towards addressing the research question about how many videos are necessary to record, in the next chapter, we describe the experimental setup and second user study that helps us assess the quality, breadth, and depth of material recorded by *TOIA avatar*-makers. We invited a large number of human subjects to interact with different variants of the *TOIA avatars* we created under controlled conditions, and we test the effectiveness of the QS technology implementation in the Player page to aid interactors with self-narrative chatbots by suggesting questions to ask.

## Chapter 7

### Interacting with TOIA Avatars

Thus far, our research has primarily focused on the experience of creating *TOIA avatars*. While this is a more crucial and demanding aspect of creating time-offset interactions, the ultimate goal is to engage with created *TOIA avatars* that are interesting and engaging. As we successfully determined a streamlined, relatively quick, and creativity-aided way for creating time-offset interactions, we need to close the loop and measure if *TOIA avatars* created that way also offer a great interaction experience. Moreover, important open questions about interacting with time-offset avatars still need to be addressed.

As it involves both the creation and interaction experience, one key research question is to determine the minimum number of videos needed to create an engaging enough *TOIA avatar* and how this number is influenced by the creation process (i.e. with creative aid or through brainstorming). A streamlined creation process that requires fewer videos may be more convenient for the creator but could result in a poorer interaction experience for the interactor due to a limited number

of answers (therefore, range of topics) covered by the *TOIA avatar*.

Additionally, it is important to examine the impact of the avatar's personality on engagement with the interactor and whether the original personality of the *TOIA avatar*-maker is conveyed effectively through the time-offset interaction medium. Finally, we aim to study the influence of the filler videos on the interactor's experience.

In this chapter, we present the user study experiment designed to address the questions above and to evaluate the impact of the avatar creation process, outlined in Chapter 6, on the interaction stage.

## 7.1 User Study Methods

The experiment was approved by our institution's *Institutional Review Board*. The following presents the user experience study we designed to answer the above research questions. We describe the interaction scenario and the data collection procedure.

### 7.1.1 Interaction Scenario & Experimental Conditions

**Experimental Conditions** In the previous experiment described in Chapter 6, we created 26 *TOIA avatars* using three setups: 10 with GPT-3 QS, 8 with Human-QS, and 8 with QS-off. To compare hypotheses on the minimum required number of videos in the knowledge base of each *TOIA avatar*, we created variants based on three video count conditions: using the first 30, the first 60, or all recorded

videos. Our choice of video count was motivated by a few considerations. A similar commercial time-offset interaction builder ([storyfile.com](http://storyfile.com)) offers users the ability to record up to 30 video answers for free (at the time of writing). While there is no published research on why 30 videos were chosen, we hypothesized that 30 videos might provide an overview of a time-offset interaction but not be sufficient to create a meaningful *TOIA avatar*. For the 60-video cut, we picked 60 because the minimum number of videos recorded in our previous experiment was 71. We selected a cut-off point far enough from 30, but not too close to the minimum of 71 to ensure meaningful differences between the variants.

We also introduced variants based on the type of filler videos to display during the interaction: only attentive fillers or only inattentive ones. As such, we created 6 variants of the original *TOIA avatars* (3 conditions on the number of videos X 2 conditions on the filler videos). In total, we have  $26 \times 6 = 156$  different variants available for interaction. To have more robust ratings and more statistical power, we wanted to collect at least two different opinions for each variant, or, in other words, let each variant be tested by two different people. That amounts to  $156 \times 2 = 312$  total interactions.

As we recruited participants in the same pool as the *TOIA avatars* for the same reasons as the previous experiment (see Chapter 6), we had an additional condition to avoid: that the interactor knew the *TOIA avatar*-maker already. The participants' terms of engagement included a financial allowance for interacting with at least four *TOIA avatars*. To satisfy the terms, the constraints above, and that the same interactor would not interact with different variants of the same

*TOIA avatar*, we had a pool of 64 participants. Most interacted with 4 variants each, and 8 of them interacted with 8 variants. The total number of interactions recorded was 317.

	Participants	#	%
<b>Age</b>	18	8	12.5
	19	13	20.3
	20	20	31.3
	22	8	12.5
	23	5	7.8
	37	1	1.5
<b>Gender</b>	Female	35	54.7
	Male	25	45.3
<b>Mother Tongue</b>	English	13	20.3
	Other	51	79.7
<b>School</b>	Humanities	5	7.8
	Engineering	11	17.2
	Science	24	37.5
	Social Science	20	31.2
	Undecided	4	6.3
<b>Is TOIA Avatar</b>	No	48	75.0
	Yes	16	25.0
<b>TOT</b>		64	100.0

Table (7.1) Interaction participants demographics.

**Interaction Scenario** The participants signed up for the experiment with an online form, including informed consent that explained the study's overall goal and how the interactions and recordings would proceed. The form also included a section explaining the allowance the selected participants would receive after participating in the study: 80 AED (about \$USD 22) was paid to those selected to participate in the study. Although we are aware that paying participants may introduce information bias, we could not avoid it in our study. However, we experimented in a controlled manner and paid the participants a minimal amount to

minimize their vested interest. Moreover, their qualitative feedback gave us strong indications of their overall appreciation of the experience and genuine interest in using our TOIA.

One hundred eighteen individuals signed up to participate in the experiment. Among those who signed up, we performed stratified random sampling in a way to obtain a consistent sample in terms of gender, age, year, and field of study across the three experimental condition cohorts. Based on the stratified sampling, 90 participants were invited to participate in the study. Among these, only 78 accepted the invitation, of which 14 withdrew, which resulted in a total of 64 participants (35, or 54.7%, female). The age range of the participants was 18 – 23 (with a 37-year-old outlier). All participants were fluent in English, with 51 ( 80%) having English as a second language. The participants majored in either Humanities (Arts, Interactive Media), Engineering, Science (Mathematics, Biology, Computer Science), or Social Science (Psychology, Economics, Political Science), with the majority coming from Science (24 participants). Sixteen (25%) subjects also participated in the previous experiment by creating their *TOIA avatar*. Table 7.1 summarizes the statistics on participants' demographics per cohort.

The subjects were instructed to use the TOIA interface presented in Chapter 5 to interact with *TOIA avatars*. The users were invited to a room to interact with a maximum of four avatars. Some participants interacted with a different number – eight, five, or one – to satisfy the constraints mentioned earlier. Before the interaction, one researcher demonstrated the system to them and sent all participants written, detailed instructions and troubleshooting options by email.

The participants were instructed to screen-record their user-interface interaction sessions to allow researchers to observe them and collect useful data on how they reacted to the interface.

The following experimental parameters were controlled to ensure consistency among the different user interactions and avoid potential user fatigue:

- We instructed participants to interact with each variant for a minimum of ten minutes and keep interacting as long as they found the conversation interesting or engaging;
- We instructed participants to complete a post-interaction survey immediately after interacting with each variant;
- We double-checked the participants did not know the *TOIA avatars* before interacting with them;
- Participants who interacted with more than four avatars did so on two different days.

**Collected Data** We collected the video recordings of the interactors, the recording timestamps, the user session timestamps, text transcriptions of questions and answers, and Zoom recordings of interactions. We show summary, descriptive statistics of this corpus in Appendix A, Table A.2.

**Post Interaction Survey** After completing their *TOIA avatar* recordings, the participants completed a follow-up survey probing for qualitative and quanti-

tative feedback on the experience. The full survey questions are shown in Tables A.4 and A.5: the user experience questionnaire aimed to gauge users' perceptions of the interface, their experience using it, their personality traits, and the *TOIA avatars* perceived personalities, along with qualitative items sought to gather descriptive feedback, and possible uses of such software, what they learned about the *TOIA avatar* they spoke with. Quantitative items measured satisfaction with the software's performance and ratings on the user experience enjoyment. Participants were also required to nominate a *TOIA avatar* for a "best *TOIA avatar*" contest.

To investigate the effect of personality traits of the *TOIA avatar* interaction experience, the survey also included a 10-item short version of the Big Five Inventory personality questionnaire [58] (see tables A.4 and A.6 in the Appendix). Moreover, we asked participants to answer the same personality questionnaire about the *TOIA avatar* they interacted with. That way, we can compare them against the *TOIA avatar*-maker answers and ascertain if personality comes across consistently or differently through this medium of communication.

### 7.1.2 Statistical Methods

We implemented several statistical techniques to test for differences and correlations between the post-interaction survey results. We chose the techniques based on the type of data collected and to what extent we can assume underlying assumptions like normality and independence. Below is a brief overview of the techniques and their relevance to the data. It is not easy to determine the normality of the distribution of our data. While some survey responses pass tests for normality, the granularity of a scale from one to five makes these tests too susceptible to

outliers. Moreover, a graphical exploration (see Figure 7.1) already shows that most ratings do not follow a symmetric distribution.

**Spearman Rho Assessment** Using the Spearman rho assessment, we examined the relationships between quantitative variables, such as the Likert scale ratings, and survey responses that can be transformed into an ordinal variable. This technique is appropriate when the data is not normally distributed, or there is evidence of non-linearity in the relationship between the variables. Spearman rho provides a rank-based measure of the strength and direction of the relationship between two variables, with values ranging from -1 to +1.

**Kruskal-Wallis Statistic** We chose the Kruskal-Wallis statistic to test for differences in Likert scale ratings and survey responses between multiple groups, such as the different experimental setups. This technique is appropriate when the data is not normally distributed, with more than two groups. The Kruskal-Wallis test provides a rank-based measure of the difference between groups without assuming equal variances or normality. To double-check the consistency of results and apparent differences between pairs of groups (e.g., QS-off vs. GPT-3-QS), we also checked other non-parametric statistical tests that do not assume normality, namely the Pairwise Mann-Whitney tests with Bonferroni correction and the Welch ANOVA test.

**Pairwise Mann-Whitney Tests (with Bonferroni Correction)** Pairwise Mann-Whitney tests without and with Bonferroni correction were used to compare the differences in Likert scale ratings and survey responses between two groups at a time, such as the GPT-3 generated suggestion setup and the manually suggested

setup. This technique is appropriate when the data is not normally distributed and two groups are being compared. The Mann-Whitney test provides a rank-based measure of the difference between groups, and the Bonferroni correction adjusts for multiple comparisons.

**Welch ANOVA Test** The Welch ANOVA is a parametric test used to compare two or more independent groups. It is an extension of the traditional one-way ANOVA test that can handle unequal variances between the groups. Welch ANOVA tests the null hypothesis that the means of all groups are equal against the alternative hypothesis that at least one group mean is different. Welch ANOVA assumes that the data are normally distributed, and the assumption of equal variances is unnecessary.

**Mixed-Effects Model** A mixed-effects model was used to examine the relationships between the Likert scale ratings or other survey responses and the experimental setups while controlling for potential confounding variables, such as participant's mother tongue or personality traits. This technique is appropriate when the data is not normally distributed and there are multiple independent and dependent variables and potentially confounding variables to be controlled for. The mixed-effects model estimates the fixed effects (e.g., the effect of the experimental setups) and the random effects (e.g., the variability between participants). For instance, when studying the personality effects, we have one subject giving the same rating on their personality questionnaire while giving different ratings to other questions in the same survey because they refer to different *TOIA avatar* variants they tested.

Overall, these statistical techniques allow us to examine the relationships and differences between variables while taking into account the non-normality of the

data, potential confounding variables, and differing variances between and within different groups.

## 7.2 Experimental Results

We evaluate and compare the interactions with *TOIA avatars* across different experimental conditions: (1) *TOIA avatars* created without the aid of Question Suggester (referred to as “QS off” or simply “off” condition), with the aid of the GPT-3 based QS (“gpt3” or “GPT-3-QS”), and with the aid of the human QS (“manual” or “human”); (2) *TOIA avatars* with a database of 30 answers, 60 answers and above 71 (“71+”); (3) *TOIA avatars* playing attentive filler videos between interactions, and inattentive.

We are investigating how the different experimental conditions affect various aspects of the interaction experience. Specifically, we are examining user satisfaction ratings, the length of the experience, answering capabilities, the relevance of answers, understanding, speed, overall pace, and flow of the conversation. In addition, we are studying ratings related to ASR (automatic speech recognition) performance and overall enjoyment. We are also exploring how the rating of the interactor’s personality dimensions and the perceived and self-disclosed personality dimensions of the *TOIA avatars* impact the ratings and how the experimental conditions may impact the perceived personality of the *TOIA avatars*. Finally, we are analyzing qualitative feedback received in response to the following survey questions: “Any other comments on your experience with the avatar?”, “Would you like to interact with other avatars?”, “Would you create an avatar of yourself?” and “How did it

feel talking to the avatar?"

### 7.2.1 Quantitative Results Overall & By Experimental Condition

We transformed some of the multiple-choice or free-text answers to the post-interaction survey into numerical ones. Sometimes, we had to first create manual annotations for a free text answer, making it categorical, then transform it to numerical. We show the Spearman  $\rho$  correlation coefficients between the quantitative variables in Figure 7.2. We implement the aforementioned transformations as follows:

- Gender: 1 for Female, 0 for Male (nobody answered “Other”).
- Language: First, we group all non-English mother tongues as Non-English. Then we assign 1 if it’s non-English, and 0 otherwise.
- Duration: 10 if <10 min, 15 if 10-15 min, 20 if 15-20 min, 25 otherwise.
- Rephrase: 0, 2, 4, 6 for 0, 1-2, 3-4, 5+, respectively, answered to “Overall, how many times did you have to rephrase your questions before finding the answer you were looking for? (excluding questions you were informed had no prerecorded answers to).”
- Repeat: the same as above for answers to “Overall, how many times did you have to repeat the same question/comment for the avatar to respond relevantly?”
- No answer: 1, 2, 3 for Once, Twice, More, respectively, answered to “For

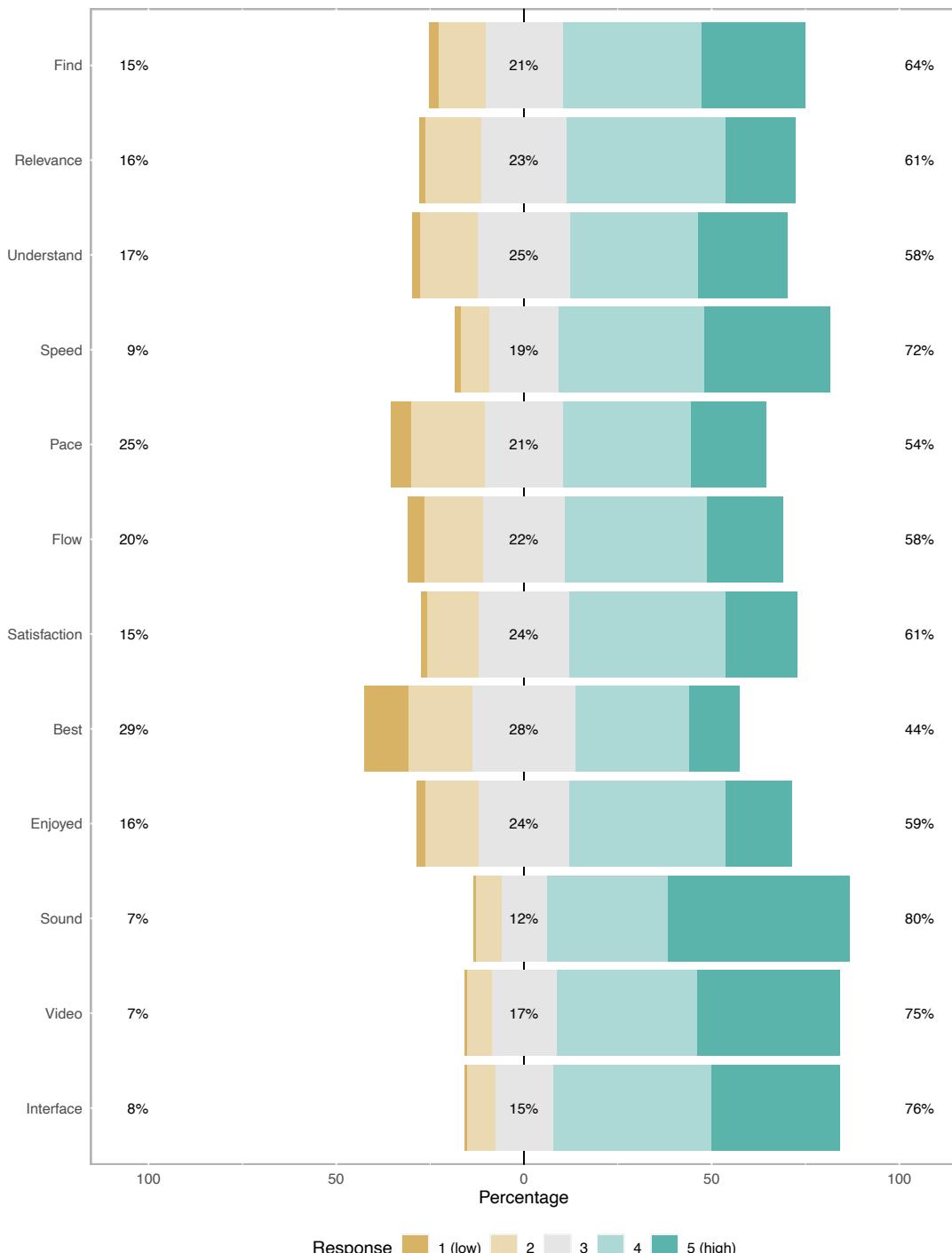


Figure (7.1) Likert scale distributions across all the survey answers that required an answer on a scale from 1 (low, poor) to 5 (high, great). This chart excludes personality scores that are discussed later.

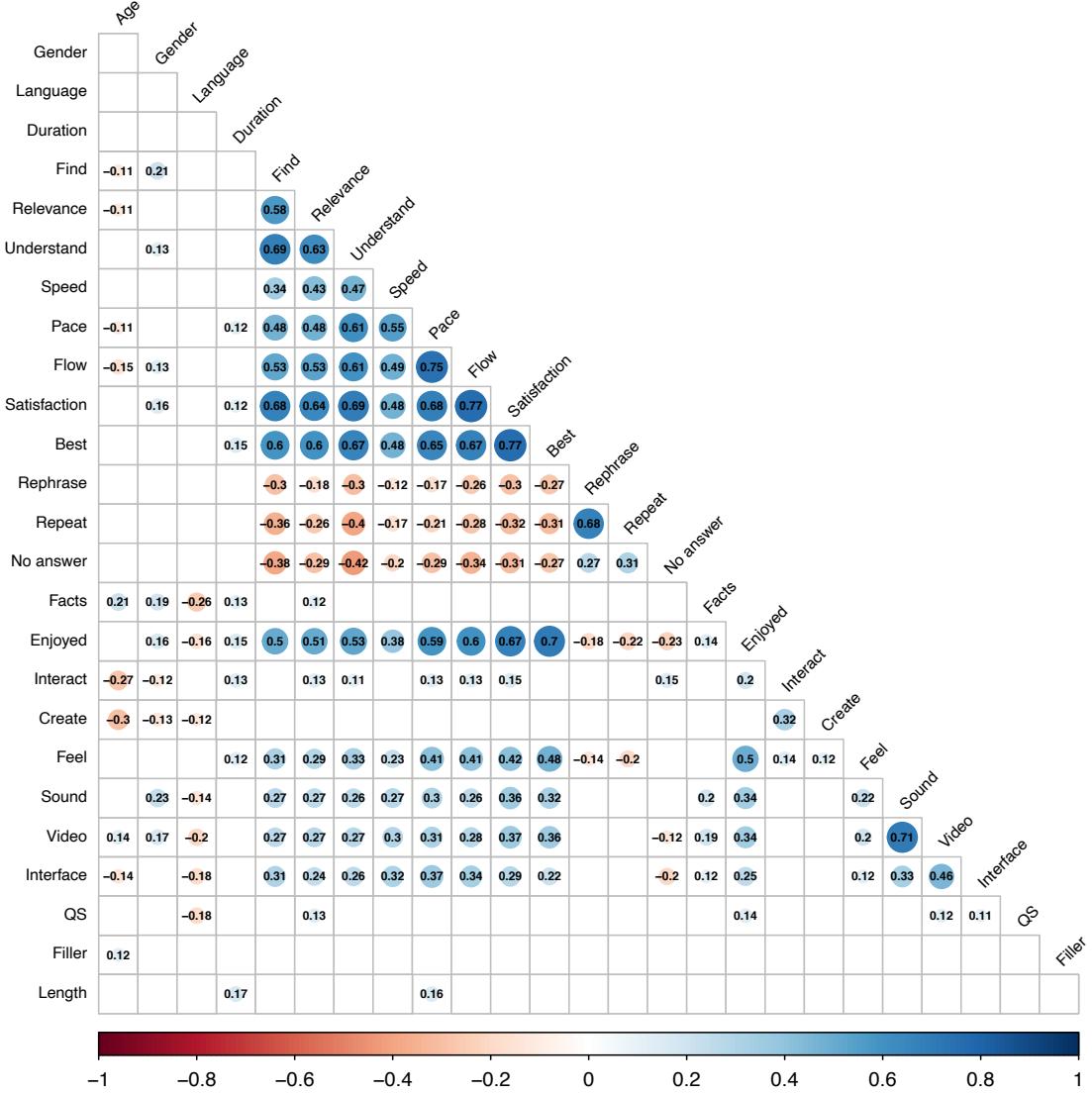


Figure (7.2) Spearman  $\rho$  correlations between all quantitative (and categorical transformed to quantitative) answers to the post-interaction survey questions. We only show correlations where they are statistically significant ( $p\text{-value} < 0.05$ ).

questions with no answers, how many times did you have to ask a question before you realized there probably is no recorded answer to your question?"

- Interact: -1, 0, 1 for Yes, Maybe, No answered to "Would you like to interact with other avatars?"

- Create: -1, 0, 1 for Yes, Maybe, No answered to “Would you create an avatar of yourself?”
- Feel: Most of the free-text comments to this question focused on the words “enjoyed” / “great”, “felt natural” / “real”, “neutral” / mix between good and bad feelings, and negative comments. We manually annotate answers with such categories and transform them into an ordinal variable taking the values 2, 1, 0, and -1 for “enjoyed”, “natural”, “neutral”, “not great”, respectively. By reading the answers, it looks apparent that “enjoyed” carried a stronger positive connotation than “natural”. A summary of these answers is given in Figure [ref needed].
- QS: 3, 2, and 1 for GPT-3, manual and off, respectively.
- Filler: 1 and 0 for Attentive and Inattentive, respectively.
- Length: 1, 2, and 3 for 30, 60, and 71+ accordingly.

All other questions are already quantitative, on Likert scales from 1 to 5. We center and standardize the data before computing correlations and running mixed-effect regressions.

Figure 7.1 displays the distributions of Likert scales across all questions that required an answer on a scale from 1 (low, poor) to 5 (high, great), excluding the Big-5 personality questionnaires. The experimental results indicate that the participants’ overall experience was positive, as the average Likert scale ratings were consistently above 3.5 and medians towards 4. We show the numerical results in tables 7.2 and 7.3 only for the questions that showed statistically significant differences between groups.

The survey questions revealed that avatars with GPT-3 generated suggestions scored the highest, followed by manually suggested setups and brainstormed ones. In contrast, the filler setup did not have any impact on the survey questions. The number of video answers in the database varied between the low setup (30 videos) and high setups (60 or all videos). No significant differences were found between the two high-video setups.

Group	#	Mean	StDev	Median	Mann	Bonferroni Correction
<b>Language</b>						
(Kruskal: <b>0.000281</b> , Welch ANOVA: <b>0.000535</b> )						
<i>gpt3</i>	124	0.69	0.46	na	vs. manual: 0.000327	0.000980
<i>manual</i>	96	0.90	0.31	na	vs. off: 0.400356	0.01478
<i>off</i>	97	0.86	0.35	na	vs. gpt3: 0.004927	1.000000
<b>Relevance</b>						
(Kruskal: <b>0.045890</b> , Welch ANOVA: <b>0.01595</b> )						
<i>gpt3</i>	124	3.73	1.00	4.00	vs. manual: 0.596439	1.000000
<i>manual</i>	96	3.65	1.03	4.00	vs. off: 0.069561	0.208683
<i>off</i>	97	3.43	0.97	3.00	vs. gpt3: 0.017819	0.053456
<b>Best</b>						
(Kruskal: <b>0.048850</b> , Welch ANOVA: <b>0.033165</b> )						
<i>gpt3</i>	124	3.24	1.24	3.00	vs. manual: 0.705834	1.000000
<i>manual</i>	96	3.32	1.16	3.00	vs. off: 0.021567	0.145302
<i>off</i>	97	2.91	1.17	3.00	vs. gpt3: 0.048434	0.064702
<b>Enjoyed</b>						
(Kruskal: <b>0.028659</b> , Welch ANOVA: <b>0.028269</b> )						
<i>gpt3</i>	124	3.70	1.01	4.00	vs. manual: 0.745529	1.000000
<i>manual</i>	96	3.67	0.94	4.00	vs. off: 0.031859	0.042695
<i>off</i>	97	3.35	1.04	3.00	vs. gpt3: 0.014232	0.095578

Table (7.2) Summary statistics, group and pairwise statistical tests testing the difference between groups based on the **suggestion setup**. Reporting only the survey questions where we find p-values below the significance level of 0.05.

However, the statistical test results highlighted in tables 7.2 and 7.3 reveal some

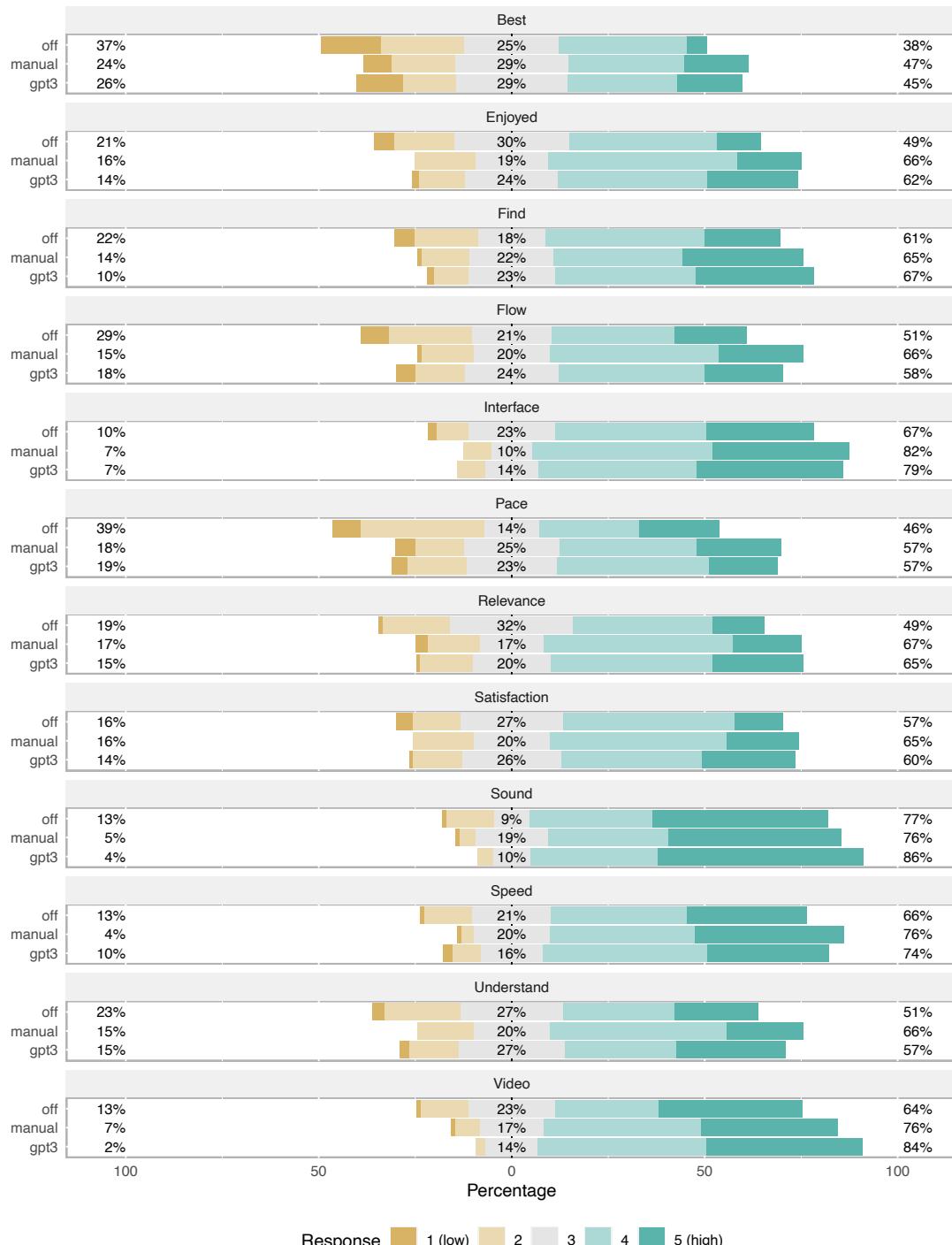


Figure (7.3) Likert scale distributions split by the different question suggestion setups. The ‘gpt3’, ‘manual’ and ‘off’ setups indicate groups of *TOIA avatars* that were created with the Question Suggester automated with the GPT-3 implementation, manually run by a researcher, and disabled, respectively.

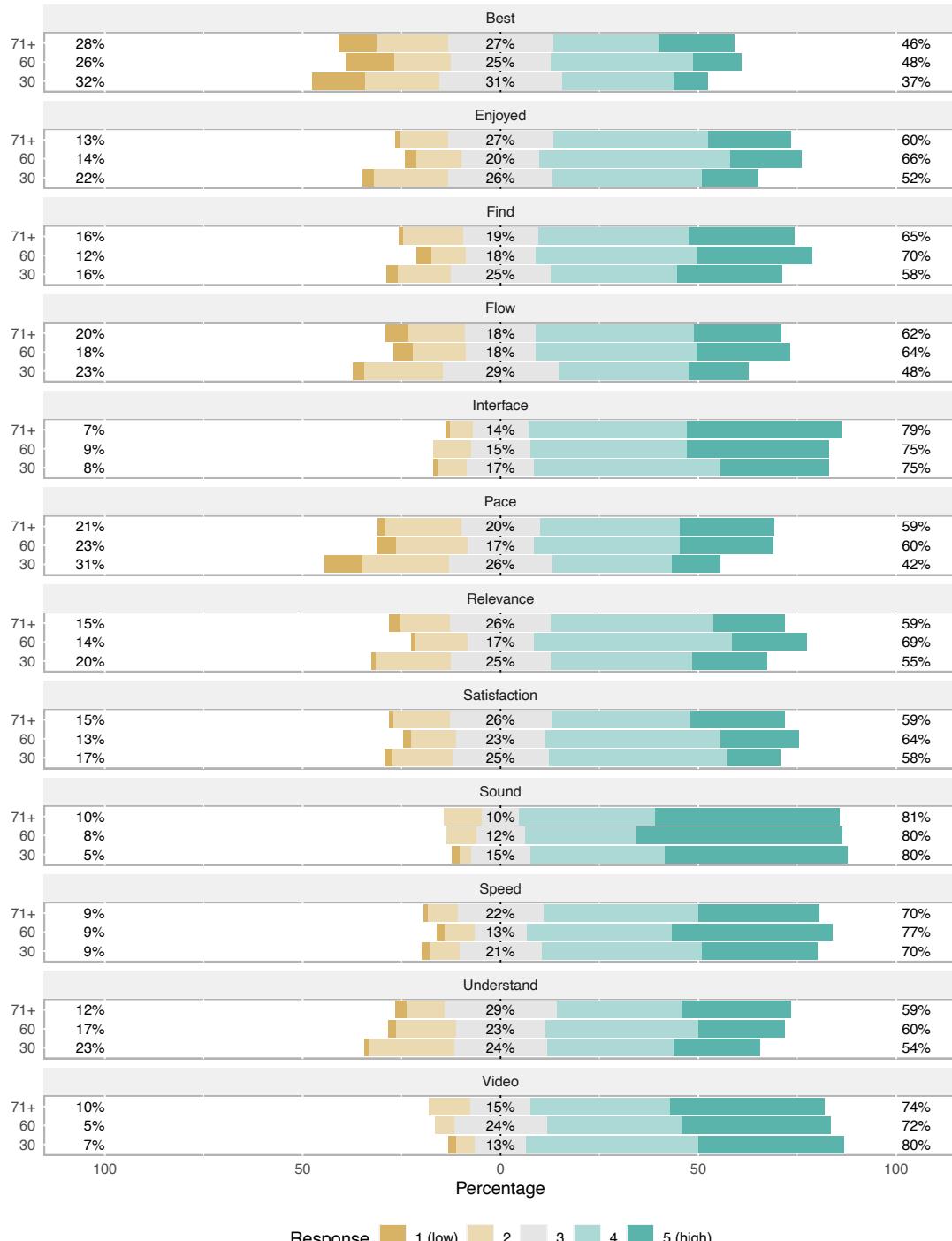


Figure (7.4) Likert scale distributions split by the groups of variants with a different number of videos in their databases. The '30', '60', and '71+' labels indicate groups of *TOIA avatar* variants that were created by selecting the first 30, 60, and all the videos recorded by the *TOIA avatar*-makers, respectively.

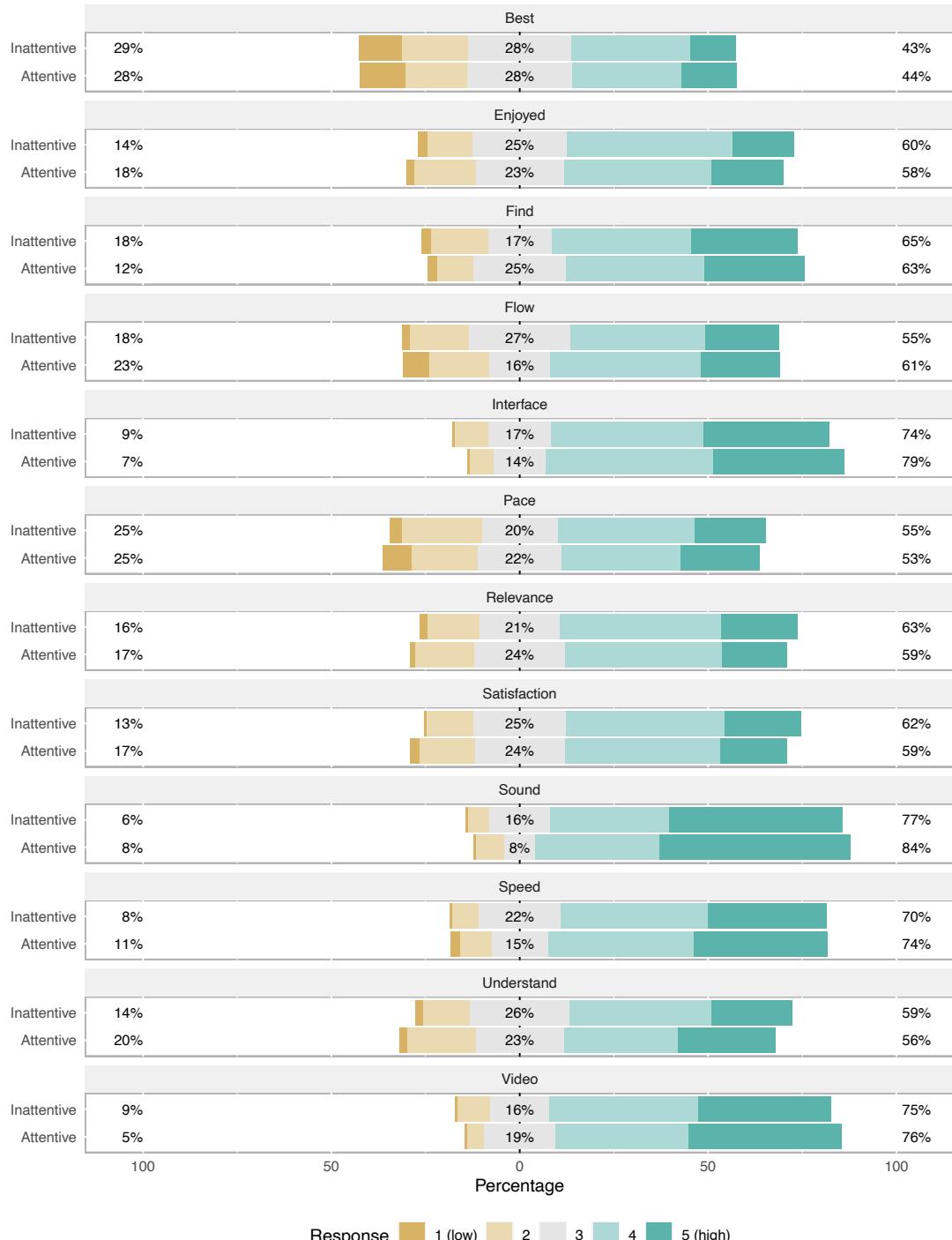


Figure (7.5) Likert scale distributions split by the groups of variants with different fillers played between interactions. 'Inattentive' and 'Attentive' are self-explanatory labels indicating the groups of *TOIA avatars* variants that were created by including only inattentive or attentive fillers, respectively.

Group	#	Mean	StDev	Median	Mann	Bonferroni Correction
<b>Duration</b>						
(Kruskal: <b>0.011295</b> , Welch ANOVA: <b>0.014946</b> )						
30	106	15.00	3.72	15.00	vs. 60: 0.172179	0.516536
60	106	15.66	3.78	15.00	vs. 71+: 0.095638	0.286914
71+	105	16.57	4.06	15.00	vs. 30: 0.003137	0.00941
<b>Pace</b>						
(Kruskal: <b>0.006826</b> , Welch ANOVA: <b>0.00676</b> )						
30	106	3.14	1.17	3.00	vs. 60: 0.007772	0.023317
60	106	3.57	1.17	4.00	vs. 71+: 0.933743	1.000000
71+	105	3.60	1.11	4.00	vs. 30: 0.005279	0.015838
<b>Interact</b>						
(Kruskal: <b>0.004254</b> , Welch ANOVA: <b>0.012972</b> )						
30	106	0.75	0.58	1.00	vs. 60: 0.040828	0.122484
60	106	0.59	0.69	1.00	vs. 71+: 0.001547	0.004641
71+	105	0.84	0.48	1.00	vs. 30: 0.246876	0.740629

Table (7.3) Summary statistics, group and pairwise statistical tests testing the difference between groups based on the **no. of videos in the database**. Reporting only the survey questions where we find p-values below the significance level of 0.05.

differences between groups in survey responses that are statistically significant. We list the main discoveries below. It is worth noting that the Bonferroni correction to the Mann-Whitney U test can be overly conservative and may result in decreased power to detect true effects. In our case, where we have at most three groups to compare, an uncorrected p-value is enough evidence for the significance of an effect.

- The mother tongue proportions differ between the group of subjects who interacted with *TOIA avatars* built with question suggestions automated by GPT-3 (gpt3) vs. brainstormed (off) or with human suggestions (manual). 69% of the subjects interacting with the gpt3 cohort have mother tongues other than English, vs. 90% and 86% for the manual and off cohorts, respectively. However, the impact of the interactor's mother tongue on other ratings does

not come out as significant when tested via regression and mixed effects analysis.

- The survey question “Rate your experience with the Avatar: Relevance of the avatar’s responses to your questions or comments (including those you were informed had no answers to)” has higher average and median ratings for the gpt3 and manual cohorts (3.73, 4.00 and 3.65, 4.00 respectively) than the off cohort (3.43, 3.00). The differences are especially statistically significant between the gpt3 and off cohorts without correction and borderline with Bonferroni correction (corrected p-value: 0.053456).
- The survey question asking to rate from 1 to 5 if the *TOIA avatar* should have been nominated as the best avatar attracted less generous ratings overall. The gpt3 and manual cohorts have higher averages (3.24 and 3.32, respectively) than the off cohort (2.91). The differences are especially statistically significant between the gpt3 and off cohorts and between the manual and off cohorts without correction (p-values: 0.021567 and 0.048434, respectively).
- Arguably the most important question, “Did you enjoy your experience with the avatars?” has higher average and median ratings for the gpt3 and manual cohorts (3.70, 4.00 and 3.67, 4.00 respectively) than the off cohort (3.35, 3.00). The differences are especially statistically significant between the manual and off cohorts with and without Bonferroni correction (p-values: 0.031859 and 0.042695, respectively). They are also significant by looking at the uncorrected p-value for the off vs. gpt3 cohorts comparison (p-value: 0.014232).
- We transformed the question “How long was your interaction with the avatar?” to an ordinal value that assumes the values 10, 15, 20, 25 for the “< 10 min,’

“10-15 min,” “15-20 min” and “20+ min” answers respectively. The average duration so computed differs between the group of subjects who interacted with *TOIA avatars* variants built by taking the first thirty (30), sixty (60), and all videos (71+) recorded by the makers. As expected, the average duration increases from the 30 to the 71+ cohorts: the more answers are available, the longer the subjects interact. Post-hoc statistical tests reveal that only the difference between the 71+ and the 30 cohorts is significant (p-value: 0.003137, corrected: 0.00941).

- The same as the above holds for the ratings given to the question “Rate your experience with the Avatar: The pace of the overall conversation felt natural.” The average rating is 3.60 for the 71+ cohort and 3.14 for the 30 cohort (p-value: 0.005279, corrected: 0.015838).
- “Would you like to interact with other avatars?” was interestingly answered positively in higher proportions for the 30 and 71+ cohorts (75% and 84%, respectively) than the 60 cohort (59%). Significant differences appear in the 30 vs. 60 cohorts (uncorrected p-value: 0.040828) and the 60 vs. 71+ (p-value: 0.001547, corrected: 0.004641). This might be due to different reasons, though. Intuitively, the interaction with the 30 cohort might be quick enough to show a propensity for more interactions. The interaction with the 71+ cohort might be longer on average but more interesting, motivating the subject to interact with other avatars.

## 7.2.2 Personality Effects

We investigate the relationships between Big-5 personality dimensions, between personalities and the other post-interaction survey answers overall, and compare different groups. There are three groups of personality dimensions to account for:

1. The *TOIA avatar*-makers' answers to the personality questionnaire about themselves in the post-experiment survey described in Chapter 6; we call this the “self-disclosed” personality.
2. The interactors' answers to the personality questionnaires about the *TOIA avatar* they chatted with; we refer to this as the “perceived” personality.
3. The interactors' answers to the personality questionnaires about themselves; we interchange the terms “interactor's” or “subject's” personality here.

**Is personality a confounder?** The first step to determine if personality is potentially a confounder factor over the results analyzed in the previous subsection, hence challenging possible conclusions, is to explore if there are statistically significant relationships between personality ratings alone on an outcome quantity (i.e., a survey answer). We use mixed-effect models to account for group effects. We run one model for each quantitative outcome from the survey responses, taken as the dependent variable. Then we use the five personality dimensions as independent variables. When using the perceived or the self-reported personality variables, the mixed effects group variable is the *TOIA avatar* ID – self-reported personality is fixed for the same *TOIA avatars*, regardless of the variants tested; perceived personality has a group effect even if the perceived personality ratings may vary for the same *TOIA avatar*. When using the interactors' personality dimensions,

the variable for the group effect is the interactor ID. Results show that:

- When analyzing Gender composition, female *TOIA avatars* have higher perceived conscientiousness than males.
- Perceived agreeableness and conscientiousness have positive, significant effects on the rating given to “Ease in finding answers to your questions (excluding questions that you were informed had no prerecorded answers to),” and to “Ease at which the avatar seemed to understand your questions and/or comments.”
- Perceived conscientiousness has a positive, significant effect on the rating given to “Relevance of the avatar’s responses to your questions or comments (including those you were informed had no answers to).”
- Perceived extraversion and conscientiousness have positive, significant effects on the rating given to “The pace of the overall conversation felt natural” and to “The overall flow of interacting with the avatar.”
- Perceived agreeableness, extraversion, and conscientiousness have positive, significant effects on the rating of “Satisfaction with the overall performance of the avatar.”
- Perceived extraversion and conscientiousness positively and significantly affect the best avatar recommendation. Perceived neuroticism has a negative, significant effect.
- Perceived agreeableness has a negative and significant effect on the answer to “Overall, how many times did you have to rephrase your questions before

finding the answer you were looking for? (excluding questions that you were informed had no prerecorded answers to).”

- Perceived agreeableness has a negative and significant effect on the answer to “Overall, how many times did you have to repeat the same question/comment for the avatar to respond relevantly?”
- Perceived extraversion has a significant, negative impact on the answer to “For questions with no answers, how many times did you have to ask a question before you realized there probably is no recorded answer to your question?”
- Perceived extraversion and conscientiousness positively impact the enjoyment rating (“Did you enjoy your experience with the avatars?”)
- High perceived conscientiousness is likely to influence a more positive answer to the question “Would you like to interact with other avatars?”
- Perceived extraversion, conscientiousness, and openness significantly positively affect the optimistic comments written as an answer to “How did it feel talking to the avatar?”
- Perceived conscientiousness impacts positively, while openness negatively impacts the software’s sound quality ratings.
- Perceived conscientiousness has a positive, and neuroticism has a negative effect on ratings given to the software’s video quality.
- Self-disclosed dimensions have no meaningful effects on any survey answer, except for a demographics effect: interactors with English as mother tongues

were significantly more likely to interact with less extroverted (according to self-disclosed personality) *TOIA avatars* and more open *TOIA avatars*.

- The interactor's (or subject's) personality dimensions have only a few noteworthy effects: extraversion and agreeableness are related to higher durations of the interaction; extraversion has a negative effect on how many times the subject had to repeat the same question; more conscientious and less agreeable subjects tended to respond more positively to the question asking if they would like to create their *TOIA avatars*; more extraverted interactors were more likely to describe more positive experiences in answer to "How did it feel talking to the avatar?"

<code>enjoyed ~ C(suggestion_setup) + C.filler_setup) + C(length_setup)</code>							
Model:	MixedLM	Dependent Variable:	enjoyed				
No. Observations:	317	Method:	REML				
No. Groups:	64	Scale:	0.7311				
Min. group size:	1	Log-Likelihood:	-436.3618				
Max. group size:	9	Converged:	Yes				
Mean group size:	5						
	Coef.	Std.Err.	z	$P >  z $	[0.025	0.975]	
Intercept	3.479	0.138	25.276	0.000	3.209	3.749	
$C(suggestion\_setup)[T.\text{manual}]$	0.071	0.123	0.581	0.561	-0.169	0.312	
$C(suggestion\_setup)[T.\text{off}]$	-0.264	0.124	-2.122	0.034	-0.507	-0.02	
$C(fill\_setup)[T.\text{Inattentive}]$	0.012	0.106	0.115	0.908	-0.195	0.22	
$C(length\_setup)[T.\text{60}]$	0.23	0.125	1.842	0.065	-0.015	0.476	
$C(length\_setup)[T.\text{71+}]$	0.254	0.128	1.986	0.047	0.003	0.506	
Group Var	0.257	0.096					

Table (7.4) A mixed-effect linear model for the enjoyment rating and its dependence on all three experimental conditions.

The second step to analyze confounding effects is to combine personality dimensions and experimental conditions in a mixed effect regression and analyze the impact on the coefficients and their significance by iteratively removing one

---

`enjoyed ~ C(suggestion_grouped) + C(length_grouped)`

---

Model:	MixedLM	Dependent Variable:	enjoyed			
No. Observations:	317	Method:	REML			
No. Groups:	64	Scale:	0.7260			
Min. group size:	1	Log-Likelihood:	-432.8732			
Max. group size:	9	Converged:	Yes			
Mean group size:	5					
	Coef.	Std.Err.	z	$P >  z $	[0.025	0.975]
Intercept	3.464	0.117	29.550	0.000	3.234	3.694
C(suggestion_grouped)[T.QS-on]	0.295	0.111	2.647	0.008	0.077	0.513
C(length_grouped)[T.low]	-0.241	0.110	-2.191	0.028	-0.456	-0.0025
Group Var	0.252	0.094				

---

Table (7.5) A mixed-effect linear model for the enjoyment rating and its dependence on the two experimental conditions that show statistical significance. The combined cohort (QS-on) of GPT-3-QS and Human-QS. The complementary group is QS-off. The length group consists of *TOIA avatar* variants with either 60 or 71+ videos in the database (high) vs. the variants with only 30 videos (low).

---

`enjoyed ~ C(suggestion_grouped) + C(length_grouped) + C_perceived + E_perceived  
+ E_perceived + A_perceived + C_perceived:C(suggestion_grouped)`

---

Model:	MixedLM	Dependent Variable:	enjoyed			
No. Observations:	317	Method:	REML			
No. Groups:	26	Scale:	0.7280			
Min. group size:	12	Log-Likelihood:	-415.9842			
Max. group size:	14	Converged:	Yes			
Mean group size:	12.2					
	Coef.	Std.Err.	z	$P >  z $	[0.025	0.975]
Intercept	-0.071	0.503	-0.142	0.887	-1.057	0.914
C(suggestion_grouped)[T.QS-on]	1.399	0.562	2.490	0.013	0.298	2.500
C(length_grouped)[T.low]	-0.213	0.102	-2.088	0.037	-0.412	-0.013
C_perceived	0.177	0.061	2.882	0.004	0.057	0.297
E_perceived	0.681	0.128	5.320	0.000	0.430	0.933
A_perceived	0.167	0.078	2.127	0.033	0.013	0.321
C_perceived:C(suggestion_grouped)[T.QS-on]	-0.327	0.153	-2.136	0.033	-0.628	-0.027
Group Var	0.066	0.045				

---

Table (7.6) A mixed-effect linear model illustrating the interaction effect between perceived conscientiousness and the combined cohort (QS-on) of GPT-3-QS and Human-QS. The complementary group is QS-off. The length group consists of *TOIA avatar* variants with either 60 or 71+ videos in the database (high) vs. the variants with only 30 videos (low). C-, E-, and A\_perceived stand for perceived Conscientiousness, Extraversion, and Agreeableness, respectively. The equation at the top of the table specifies the full model.

covariate from the regression and observing how much the coefficients and the p-values change. Although there is not enough evidence for the linearity of the relationship between these variables, the technique still helps us make inferences on the possible effects between covariates. The results, shown in Tables 7.4, 7.5 and 7.6, point out that self-described TOIA-avatar personality and interactor personality do not appear to be confounding variables, but perceived personality – especially perceived conscientiousness – could be. To draw a conclusive statement on confounders would require structural changes to the experiment design. Still, the statistical analysis highlights that the effect we observed of the suggestions and length setups on some ratings, especially on the extent the interaction experience is enjoyable, still holds regardless of the possible confounding effect of personality. Moreover, further observations may point to the discovery that the QS setup itself could have influenced the perceived levels of the personality dimensions.

**Perceived vs. Self-disclosed personalities.** As one can observe on the radar (or “spider”) plots in Figure 7.6, the perceived and self-disclosed personality dimensions of the *TOIA avatars* do not seem to differ too much overall except for Agreeableness and Openness, where self-disclosed look more pronounced than perceived. By analyzing the different experimental conditions investigated, we can observe that: varying the number of videos doesn’t impact too much the perceived personality; the inattentive fillers seem to make the perceived Agreeableness and Neuroticism slightly more pronounced, and the attentive fillers increase a bit the perceived Conscientiousness; the QS setups appear to show more and more considerable differences.

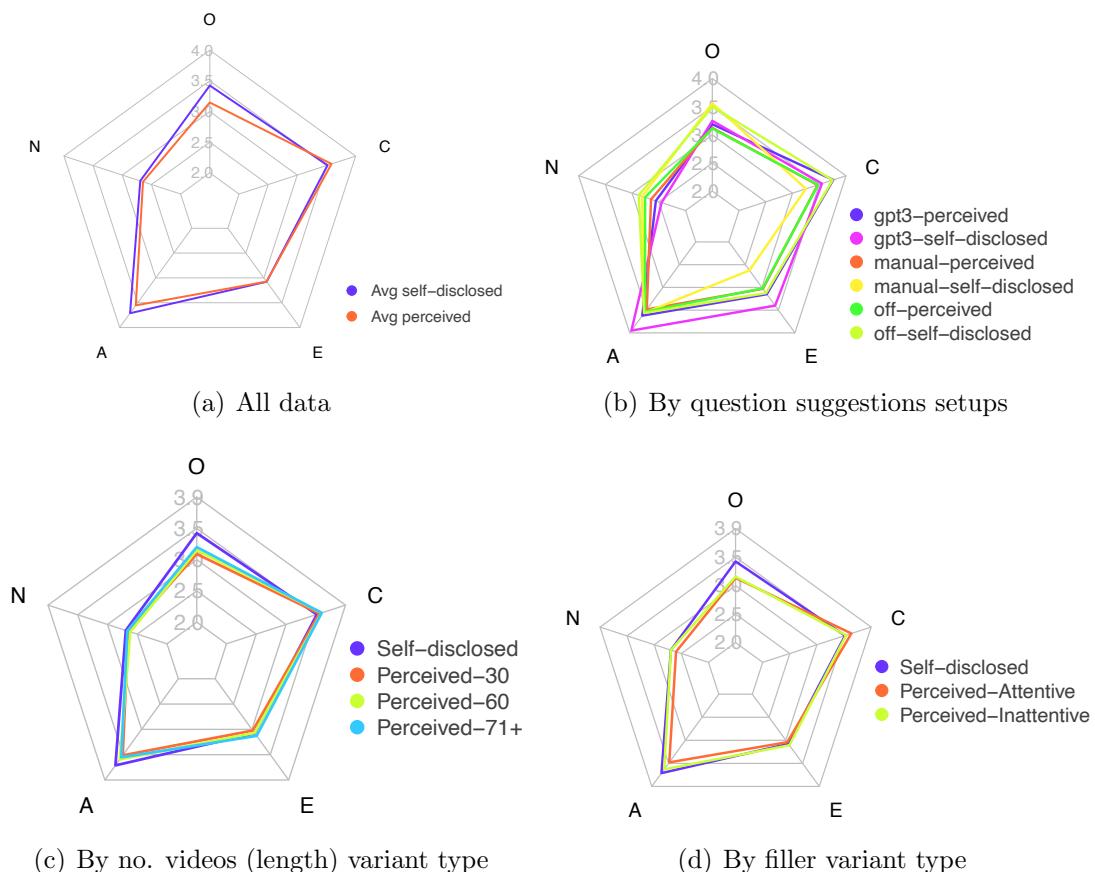
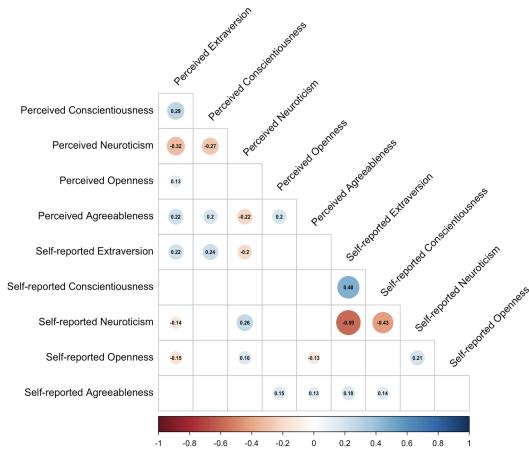
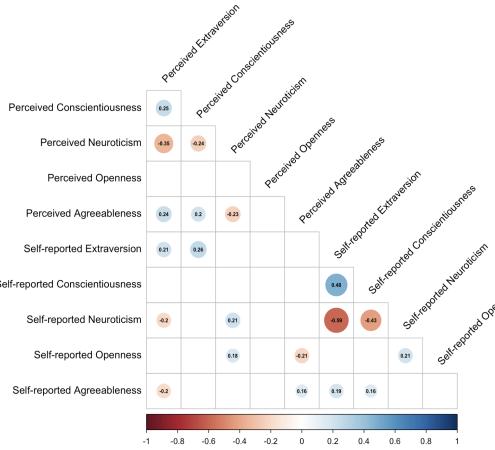


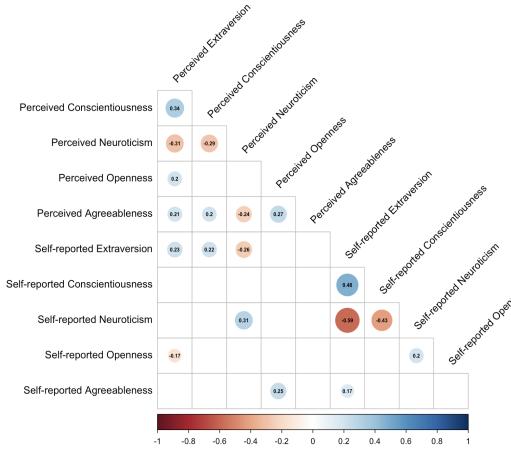
Figure (7.6) Average perceived and self-reported Big-5 personality dimensions (OCEAN), displayed in the following order: overall, by QS setup, by *TOIA avatar* variants based on video database size, and by filler type variants.



(a) All data



(b) Inattentive Fillers



(c) Anattentive Fillers

Figure (7.7) Spearman  $\rho$  correlations (displayed only for  $p\text{-value} > 0.05$ ) between self-disclosed and perceived TOIA avatars' Big-5 personality traits. Panels (a) to (c) present all data and data split by inattentive and attentive filler variants, respectively. Correlations appear consistent across cohorts, suggesting that the experimental condition regarding filler videos does not significantly influence the differences between self-disclosed and perceived personality dimensions.

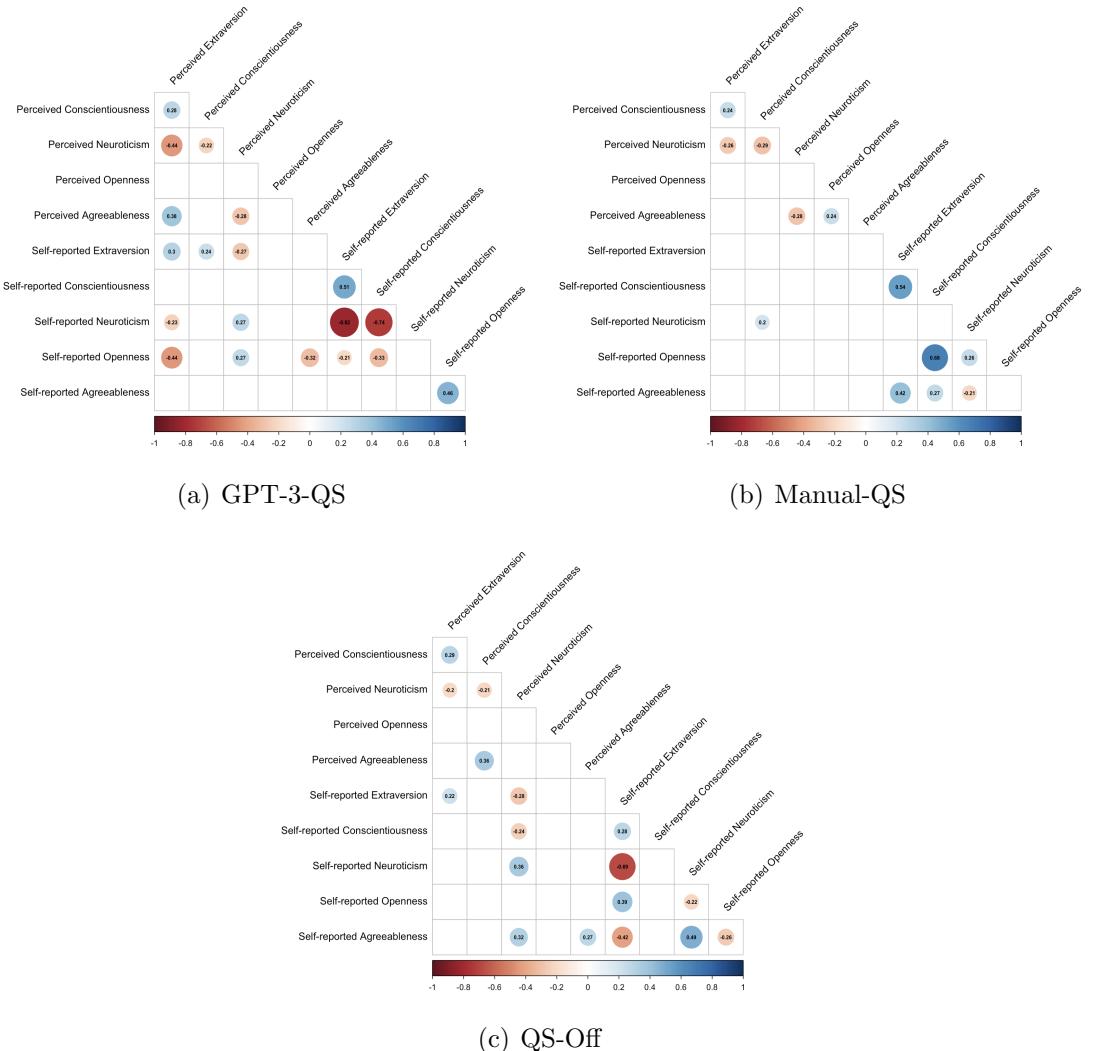


Figure (7.8) Spearman  $\rho$  correlations (displayed only for  $p\text{-value} > 0.05$ ) between self-disclosed and perceived TOIA avatars' Big-5 personality traits. Panels (a) to (c) present data split by cohorts with GPT-3 QS, manual QS, and QS turned off, respectively. The variation in correlations may indicate that the experimental condition related to QS could potentially influence differences between self-disclosed and perceived personality dimensions.

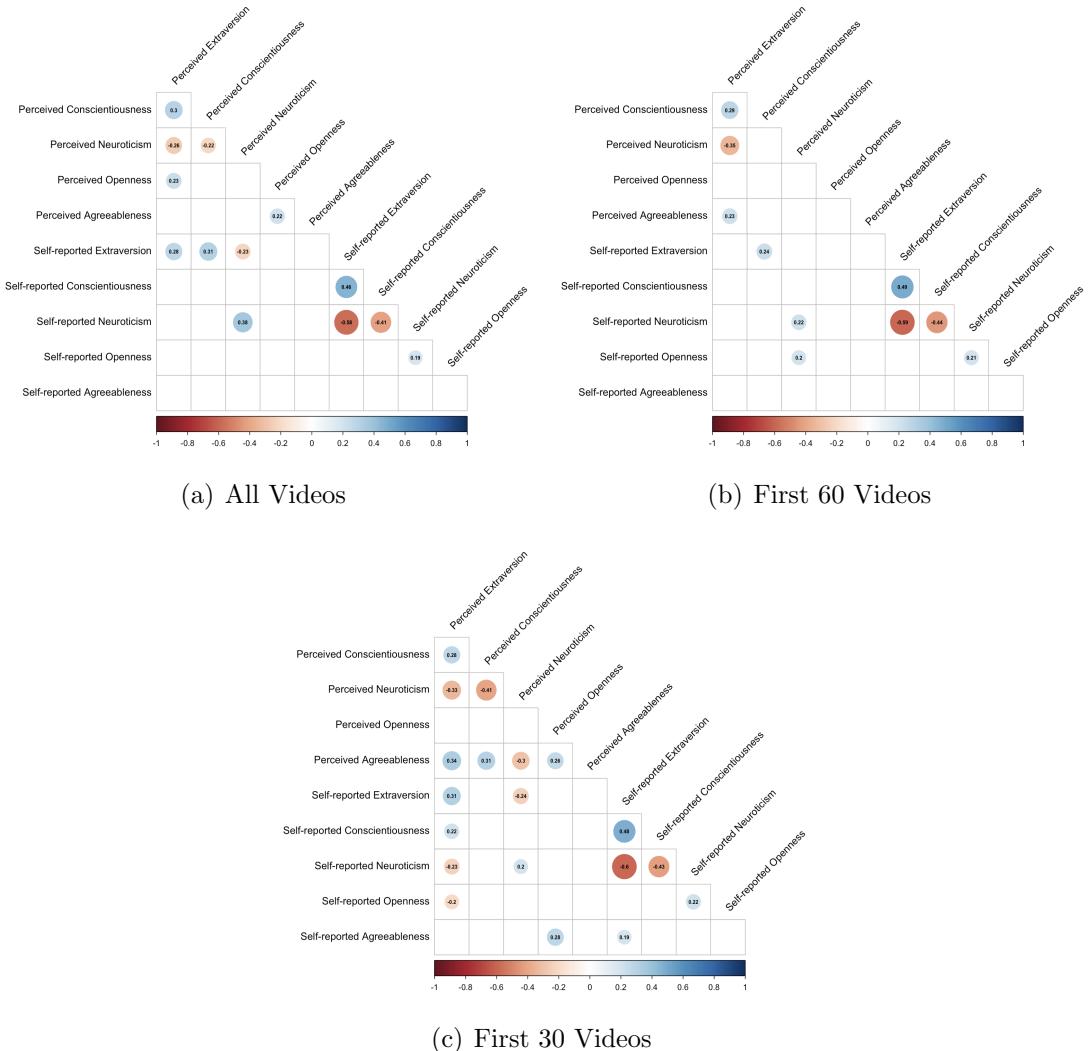


Figure (7.9) Spearman  $\rho$  correlations (displayed only for  $p\text{-value} > 0.05$ ) between self-disclosed and perceived TOIA avatars' Big-5 personality traits. Panels (a) to (c) present data split by variants with all videos, the first 60 videos, and the first 30 videos, respectively. The variation in correlations may indicate that the experimental condition based on the number of videos could impact personality perception, particularly for the 30-video cohort.

To see that the QS setup could have influenced the perceived levels of the personality dimensions, we first observe that the perceived traits are closer, whereas the self-disclosed vary more. This result makes us focus on the gaps between perceived and self-disclosed, which occurs in different personality dimensions across the groups. Perceived agreeableness, conscientiousness, and extraversion display material gaps for the gpt3 setup, where self-disclosed are higher than perceived. The manual setup shows significant gaps across all dimensions but agreeableness. The off-setup shows significant gaps in conscientiousness, neuroticism, and openness.

Cohort	#	Mean	StDev	Min	25%	50%	75%	Max
<b>QS</b>								
<i>gpt3</i>	124	0.48	0.16	0.14	0.37	0.47	0.56	1.01
<i>manual</i>	96	0.42	0.14	0.10	0.33	0.39	0.50	0.80
<i>off</i>	96	0.37	0.14	0.14	0.26	0.35	0.46	0.81
<b>Length</b>								
<i>30</i>	106	0.43	0.16	0.14	0.32	0.42	0.53	1.00
<i>60</i>	106	0.43	0.17	0.10	0.29	0.39	0.53	1.01
<i>71+</i>	106	0.42	0.14	0.14	0.32	0.42	0.51	0.87
<b>Filler</b>								
<i>Attentive</i>	158	0.43	0.15	0.10	0.32	0.42	0.52	0.86
<i>Inattentive</i>	158	0.42	0.16	0.14	0.31	0.40	0.52	1.01

Table (7.7) Distribution statistics of the Root Mean Square Difference between perceived and self-reported Big-5 personality dimensions (RMSEpd) by *TOIA avatar*-variant cohort. From left to right, the number of observations, mean, standard deviation, minimum value, 25th percentile, median, 75th percentile, and maximum value.

To dig into these differences further, we define a metric to summarize the difference between the five dimensions in one number. Inspired by the Root Mean Squared Error used in regression models, we take the square root of the average squared difference between the five personality dimensions. Let this quantity be called the RMSEpd (RMSE-personality-dimensions). By looking at the distributions

of the RMSEpd by the experimental condition in Table 7.7, we can see that in the QS setups, the difference between self-disclosed and real personality increases from a slight difference on the QS-off setting to a more significant gap on the manual-QS and even a wider one on the gpt3 setting. Now, that could be because the perceived traits don't vary a lot in general, and the gap is a direct consequence of the self-disclosed differences between those groups. However, was that the case, we would observe consistent differences across the five personality traits. Instead, we observe some traits differentiated more than others.

### 7.3 Discussion

The experimental results provide insight into using *TOIA avatars* in interactive settings. Overall, the results suggest that GPT-3 generated suggestions may lead to a more positive user experience and assertive perceived personality for avatars. This finding is consistent with previous research showing the potential benefits of using AI-generated content in various applications, including chatbots and virtual assistants. The experimental condition on fillers, playing only inattentive or attentive frames between interactions, seems not to impact any of the experience and software performance dimensions captured in the post-experiment survey. The video length setup shows a significant impact, and the results suggest that 60 videos look like a good cutoff between good (above 60) and poor experiences (below 60). It is worth pointing out that our study indicates that recording 60 video utterances is sufficient for creating a comprehensive and satisfactory avatar for the specific purpose of a 10-15-minute interaction to get to know a stranger. However, our evidence does not support the use of fewer than 60 videos, and we do not claim

that this number is enough for more complex use cases or ‘peak experience’ if such a thing can even be defined. Moreover, our study did not test the effectiveness of *TOIA avatars* with substantially more than 200 videos.

The findings also suggest that the *TOIA avatar*’s perceived personality differs from the self-described personality. In fact, by probing into that direction, we find out that is often the case with the GPT-3 generated suggestion setup, indicating that the suggestion setup influenced the perceived personality. Perceived conscientiousness is the dimension that results in having a significant impact across all the outcomes explored. According to the Big Five Personality Inventory, a conscientious person tends to be organized, responsible, reliable, and goal-oriented. They are diligent in their work, detail-oriented, and take their commitments seriously. They are often seen as dependable and trustworthy, following their promises and being accountable for their actions.

Conscientious people are typically self-disciplined and are motivated to achieve their goals. They have a strong sense of duty and are committed to fulfilling their obligations. They are often high achievers who strive for excellence in their work and personal lives. They are also typically very orderly, neat, and punctual. Interestingly, these traits look amplified by the TOIA medium when creating a *TOIA avatar* following automatically generated questions (both in the case of GPT-3- and human-powered QS). In contrast, they look diminished when the *TOIA avatar*-maker brainstorms questions.

We can think of a few possible explanations for this finding, but we must design

a new experiment to test such hypotheses. One possible explanation could be that QS brings someone out of their comfort zone, displaying a slightly different personality than their perceived one. Whereas brainstorming questions just reinforces the *TOIA avatar*'s self-perceived personality. This would align with findings that perceived traits like openness and extraversion show significant effects in the statistical analysis, but that is not true for self-disclosed personality dimensions. Brainstorming only allows for self-reflection, which was also reported in the qualitative answers of the QS-off cohort describing the experience of creating their *TOIA avatars*. Thus, the *TOIA avatar* personality may be closer to the perceived personality when the *TOIA avatar* was designed by brainstorming because only answers derived from the *TOIA avatar*'s personality are available.

Second, GPT-3 generated suggestions may be more shallow, “depersonalizing,” or less attractive, resulting in a more superficial version of the person behind the *TOIA avatar*. In Chapter 6, we learned how makers received many more suggestions from the GPT-3-QS vs. the Human-QS but accepted a much higher proportion of suggestions coming from a human. Yet, the breadth of topics, and the number of video answers GPT-3-QS makes available, may increase the perception of conscientiousness of the *TOIA avatar*-maker.

While our statistical results have sizably supported our intuitions and expectations, it is necessary to acknowledge that they have limitations. One potential area for improvement is the size of our data set. While we ensured that the data size was sufficient to draw meaningful conclusions, we recognize that reducing the data set to smaller groups to analyze a subset of experimental conditions may further hinder the ability to detect significant patterns or correlations. Another limitation is the

potential for confounding variables or biases that may have affected our results, despite our efforts to control for them. Furthermore, statistical results alone may not fully capture the nuances and complexities of the phenomena we are studying. Therefore, while our results are encouraging and suggest that our intuitions were largely accurate, it is essential to interpret them cautiously and consider them alongside other sources of evidence. Overall, our statistical results are valuable and contribute to a better understanding of interacting with *TOIA avatars*, but we remain open to alternative explanations or perspectives.

## 7.4 Summary

The experimental results suggest that avatars with GPT-3 generated suggestions are perceived more positively and have more assertive perceived personalities than avatars with manually suggested or brainstormed setups. The findings also indicate that the number of videos in the database may affect user perceptions and preferences. Piecing together the experimental results we explored in both Chapter 6 and here, we can conclude that GPT-3 generated suggestions achieve better ratings when creating *TOIA avatar* and interacting with them. However, they influence the avatars' perceived personality in interactive settings. Further research is needed to understand this result's underlying mechanisms, potential limitations, and how to address them. Future research should explore the benefits and limitations of GPT-3 generated suggestions in interactive settings and other AI-based approaches. It is also helpful to investigate the impact of different factors, such as the type and quality of the suggestions, the user's familiarity with AI-generated content, and the context and purpose of the interaction.

# Chapter 8

## Conclusion and Future Directions

### 8.1 Summary and Contributions

This thesis explores the challenging task of creating Time-Offset Interaction Applications (TOIAs) that can offer engaging experiences similar to meeting a stranger and having a stimulating conversation with them within a 15-30 minute timeframe while providing natural and engaging experiences for users with varying interests, languages, and cultures. We made foundational contributions to four research directions outlined in the introduction, which were: (1) optimizing the experience of creating a TOIA's video recordings database, (2) making the interaction seamless, (3) evaluating a TOIA, and (4) determining to what extent the user can know and perceive the real person behind the database of pre-recorded videos

We optimized the process for creating a TOIA video recordings database by devising a system that elicits answers from avatar makers in a user-friendly manner. We investigated user-driven brainstorming of questions and a system-driven ap-

proach to determine the most effective means of capturing user stories. Additionally, we found out that recording sixty five or more video utterances is a good starting point for creating an *TOIA avatar* that can handle an interesting conversation with a stranger for about half an hour.

While more work is needed before interactions with *TOIA avatars* result seamless, we engineered different software components that go in that direction. We evaluated their user acceptance and satisfaction. We developed a robust methodology for retrieving the best video answers to an interrogator's utterance at least 60% of the time. We proposed different means of interaction to compensate the 40% of times the answer cannot be retrieved correctly or simply, for breaking the ice. We still face a limitation: the interaction by typing (SEARCH) was selected less frequently than both clicking on CARD and VOICE, despite CARD being evidently faster than VOICE. Consequently, we lack a quantitative method to discern users' preferences for what they deem as seamless interactions or to understand why they favor one mode of interaction over another. We identified Question Suggester, onboarding questions, and the ability to type, speak and click on recommended questions as key elements in the user interface that facilitate meaningful experiences for both the avatar maker and the user engaging with the *TOIA avatar*.

We established a human evaluation methodology and singled out one retrieval metric for optimizing TOIA experiences, which assesses the answer retrieval component and the overall user experience. This methodology allows us to learn and validate the effectiveness of our approach and user satisfaction in creating *TOIA avatars* and engaging with them.

Finally, we contributed to the research community by providing an openly-sourced TOIA platform, a library of video databases and corpora, a methodology for developing interactions on the player interface, and two extensive user studies that evaluate the software, the experience, and the effect of personality traits on time-offset interaction applications.

The results and contributions of this thesis demonstrate the feasibility and potential of TOIAs for a number of applications and highlight possible future directions for this field. In particular, developing multilingual and multicultural TOIAs and exploring novel machine learning and generative AI techniques could further enhance the naturalness and effectiveness of the interactions between users and TOIAs.

Overall, this thesis made significant strides in addressing the challenges of creating Time-Offset Interaction Applications that cater to individuals with differing interests, languages, and cultures, and laid the foundation for future development and adoption of TOIAs.

## 8.2 Future Directions

Working on our TOIA and the journey to democratize access to such applications opened up various research streams and topics. Many are worth pursuing and would make this technology advance. This dissertation focused on areas we consider both essential to the success of TOIAs and attainable in a reasonable time frame.

Time Offset Interaction (TOI) is an innovative concept that enables communication across different timelines. With TOI, individuals can interact with others from different historical periods, interact with their future selves, and explore possibilities that could otherwise remain unexplored. TOI could be a powerful tool for unlocking modern experiences and providing benefits in various areas.

With different time intervals, users not only have access to the ability to receive real-time communication but also can engage in conversations that allow for delayed responses as well as post-dated messages and requests to simulate future discussions.

The technology behind TOI has the potential to revolutionize communication and open up new opportunities in HCI and NLP research, business, and personal life. This chapter will outline the potential benefits of TOI in various areas, including communication with future generations, reflective introspection, mental health coaching, and more. We will discuss how TOI can enhance these experiences and provide a new level of engagement.

### 8.2.1 Benefits of Time Offset Interaction

TOI could be beneficial in many areas, including:

- Societal and human development applications.
  - Time capsule for talking with future generations. By using TOI, individuals can have conversations with future generations, giving them a glimpse of the past. It can provide a unique opportunity to learn from

the past and gain better knowledge of history.

- Art installations. TOI technology can create interactive art installations that allow visitors to engage in dialogue with the artist.
- Interacting with celebrities. TOI can allow individuals to interact with famous people they admire in a safe and private setting.
- Introspection, therapy and mental health.
  - Reflective introspection. By using TOI, individuals can engage in thoughtful conversations with themselves, allowing them to explore their thoughts and feelings in a safe environment.
  - Safe space for sensitive topics. TOI provides individuals with a secure, confidential, and protected way of confidently discussing taboo topics.
  - Mental health coaching. TOI technology can help people with mental health issues by offering counseling and support on a regular or one-time basis at scale.
- Education.
  - Preparing for job interviews. TOI can help job seekers by allowing them to record themselves answering questions that might be asked in job interviews. It can also be helpful for employers wanting to gain insight into the candidates.
  - Coaching, upskilling. TOI can provide a self-paced environment to improve new skills, such as public speaking, presenting, story-telling.

**Time capsule** TOI allows people to communicate with future generations in a previously impossible way. It is a tool that can preserve memories, thoughts, and emotions for future generations, like the first use-case that inspired this thesis work, the New Dimensions in Testimony (Traum et al., 2015) [74]. By creating a repository of TOI video-messages, individuals can leave behind a legacy that can be cherished for generations. This can be a powerful tool for families to pass down their values and traditions to future generations. Furthermore, TOI messages can educate future generations about historical events, cultural practices, and social issues.

**Art Installations** TOI can provide a unique opportunity for visitors of museums and art installations to engage in dialogue with the artist without the artist being present in multiple locations and hours. It also enables individuals to interact with artists from different historical periods, gain insights into their creative processes, and understand the meaning behind their art installations. With TOI, visitors can engage in meaningful conversations with the artist and gain a deeper appreciation for their work. They can feel even more inspired by learning about their creative process.

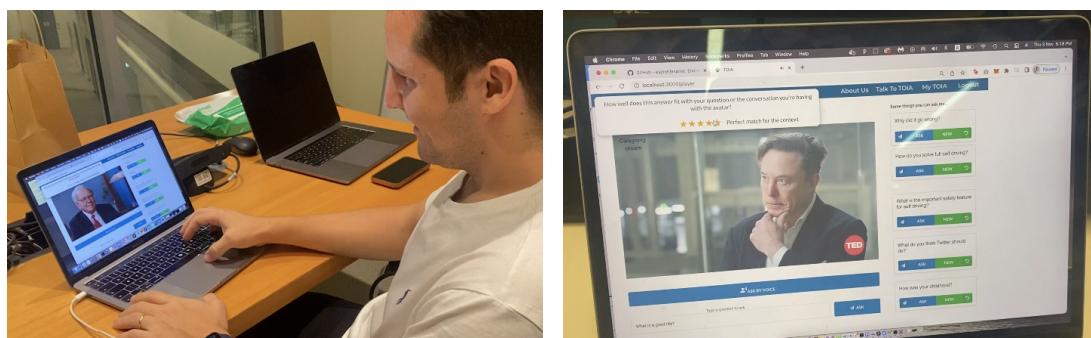


Figure (8.1) A researcher using our TOIA for interacting with Warren Buffet (left) and Elon Musk (right).

**Interacting with Celebrities** TOI provides a unique opportunity for individuals to interact with celebrities from the present or different historical periods. Today's social media often give the illusion of knowing them by being mere 'followers' on Twitter, Instagram, or other platforms. Often these platforms lack authenticity and are managed by the celebrity's PR team. Imagine instead having the possibility to have a more personal connection by speaking with them. With TOI, individuals can engage in meaningful conversations with celebrities, learn from their experiences, gain insights into their lives, and understand their creative processes. An undergraduate student working on her capstone in collaboration with our lab is already working on a feature for the TOIA recording interface that allows users to upload videos, segment them, and transcribe their audio to create more engaging interactive experiences with publicly available video content. This example is shown as the 'Elon stream' in Figure 5.2, and in Figure 8.1 where one person can interact with Warren Buffet or Elon Musk.

**Reflective Introspection** TOI enables individuals to engage in reflective introspection, a critical tool for self-improvement. It allows individuals to interact with their past selves, understand their mistakes, and learn from their experiences. With TOI, individuals can gain insights into their personal growth and development, identify areas that need improvement, and develop strategies to achieve their goals.

TOI is a powerful tool for conversing with oneself. It enables individuals to interact with themselves from different life periods and gain valuable insights into their thoughts and behaviors. With TOI, individuals can engage in meaningful conversations with themselves, understand their inner workings, and develop strategies for self-improvement.

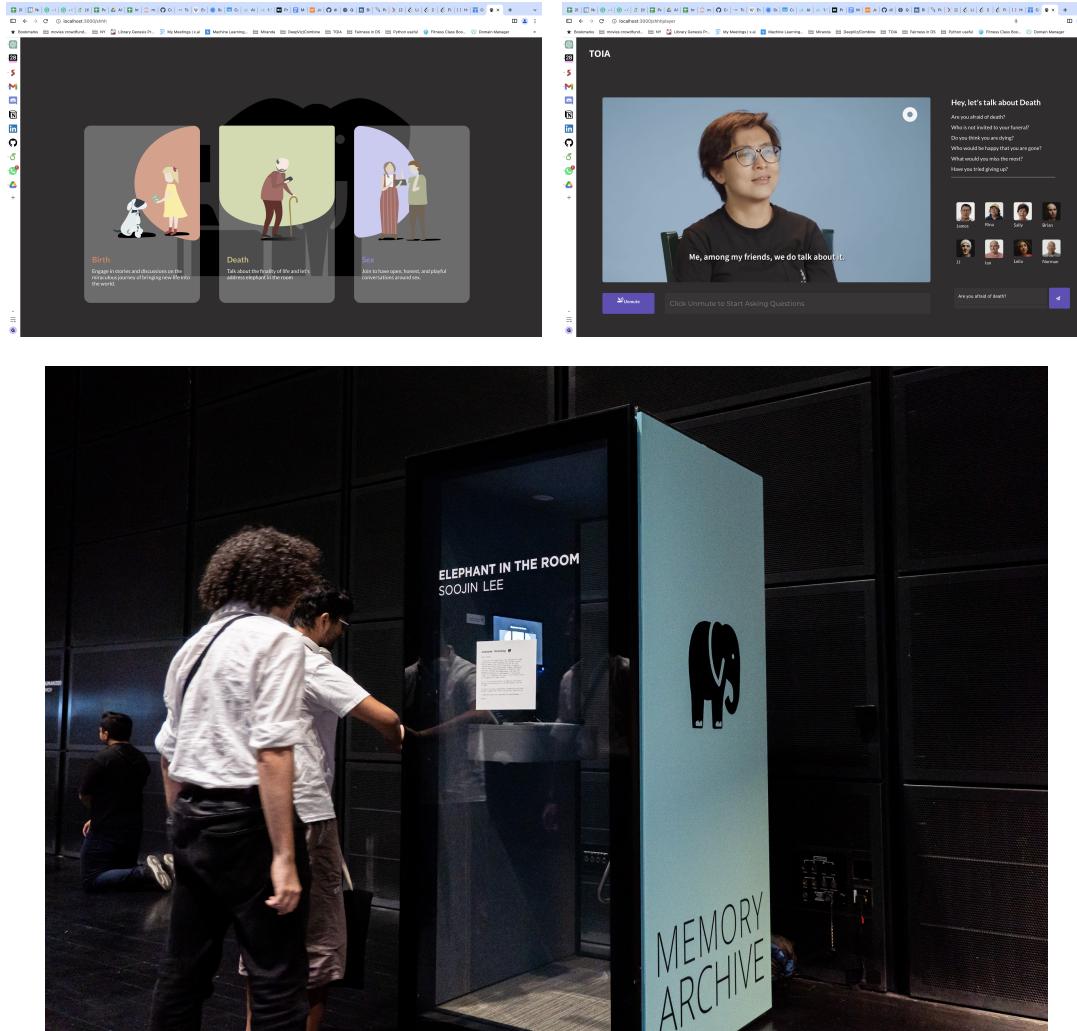


Figure (8.2) The “Elephant in the Room:” a multimedia art installation using our TOIA and a custom-made user interface for building interactive experiences about sensitive topics: on the left, the initial screen presenting the topics available; on the right, an interaction is engaging with *TOIA avatars* talking about death; on the bottom, the physical installation at an exposition space in the Art Center of NYU Abu Dhabi.

**Safe Space For Sensitive Topics** TOI can be a powerful tool for discussing taboo topics in complete privacy and confidence. It enables individuals to interact with others from different historical periods and engage in meaningful conversations about taboo topics. With TOI, individuals can share their thoughts and ideas without fear of judgment and gain valuable insights into different perspectives. For example, an undergraduate student collaborating with our lab expanded our TOIA by designing the “Elephant in the Room,” a multimedia art installation for interacting with topics like childbirth, death, and sex (Figure 8.2). The installation is accepting visitors for a few days at NYU Abu Dhabi at the time of writing. Early feedback from people engaging with the installation included:

- “The experience feels so therapeutic.”
- “The phone booth is very cozy, and I love the setting. I really enjoyed having a private talk in the booth.”
- “I want to come back tomorrow when it is less busy and sit down to talk for hours.”
- “I love the concept and the topics. I especially enjoyed the conversation about death.”
- “Did you interview these individuals? How did you meet them?” (People assumed that the artist who made this project had met these individuals in person.)
- “My favorite part was when I asked questions, and you [the artist, ndr] responded, ‘There is no video for this question. Please try another question.’”

Maybe it's because I know you personally, but it really added a cherry on top to make the interaction feel more human and less artificial."

- "Visitor: 'Can you hear what I'm asking?' Artist: 'No, I can't.' Visitor: 'Good.'—Hinting that visitors may have held private conversations in the booth.

**Mental Health Coaching** Building on findings from our user study where 26 individuals created their *TOIA avatars* (Chapter 6, TOI could be an excellent tool for mental health coaching. Qualitative feedback collected in the post-experiment survey consistently indicated the *TOIA avatar*-makers appreciated the tool as a medium for introspection. Moreover, TOI could enable individuals to interact with mental health professionals from different historical periods and gain valuable insights into mental health issues. With TOI, individuals can engage in therapy sessions with mental health professionals from different backgrounds, learn about other treatment options, and develop strategies to improve their mental health.

**Preparing for Job Interviews** TOI could be an excellent tool for preparing for job interviews. It enables individuals to interact with themselves in the future and learn from their experiences. In fact, we did run a few prototypes in this direction, and early user feedback indicates that shy individuals could appreciate such a solution. Past recordings can be used to prepare for future job interviews, including identifying potential challenges and developing strategies to overcome them. With TOI, individuals can gain valuable insights into their evolving thoughts about career paths and develop plans to achieve their goals. Companies, coaches, or mentors can access a library of interactive, pre-recorded candidates to screen or coach, regardless of time zones and location.

**Coaching, upskilling** TOI can be a powerful tool for fostering self-paced learning environments, enabling individuals to hone professional skills like public speaking, presenting, and storytelling. The unique approach to communication and interaction allows users to engage with the content at their own pace, mitigating the pressures and anxieties typically associated with real-time, face-to-face interactions and senior professionals' busy schedules. By providing an environment in which individuals can practice, receive feedback, and refine their techniques, TOI ensures a more immersive and personalized learning experience. Furthermore, TOI could encourage iterative improvement and cultivates an atmosphere of continuous growth. As users progress through the stages of skill development, they can leverage the flexibility and adaptability that time-offset interactions offer. This enables them to pinpoint specific areas for improvement, work on them at their leisure, and then apply the acquired knowledge to subsequent interactions.

### 8.2.2 Research Implications

The concept of Time Offset Interaction (TOI) has tremendous potential for various fields, including psychology, communication, and computer science. There is an opportunity to conduct further research to improve the seamless interaction of TOI, design mobile-responsive applications, handle multilingual experiments, develop TOI as an API, and record long videos to parse the content and make an avatar out of it.

**Seamless Interaction** One of the critical challenges of TOI is to make the interaction seamless. Currently, the technology for TOI is limited, and there are often errors in speech-to-text, which can hinder the conversation. Future research

could focus on developing technology similar to the deep learning models powering deep fakes to add a sense of continuous presence. The goal is to make TOI feel like real-time, synchronous communication over a video call, indistinguishable from face-to-face conversations.

**Design** Another area for further research is designing TOI applications that are mobile-responsive and user-friendly. The current technology for TOI is primarily desktop-based and not optimized for mobile devices. Research into user interface and experience could focus on developing mobile-responsive designs which enable users to have time-offset interactions from their smartphones.

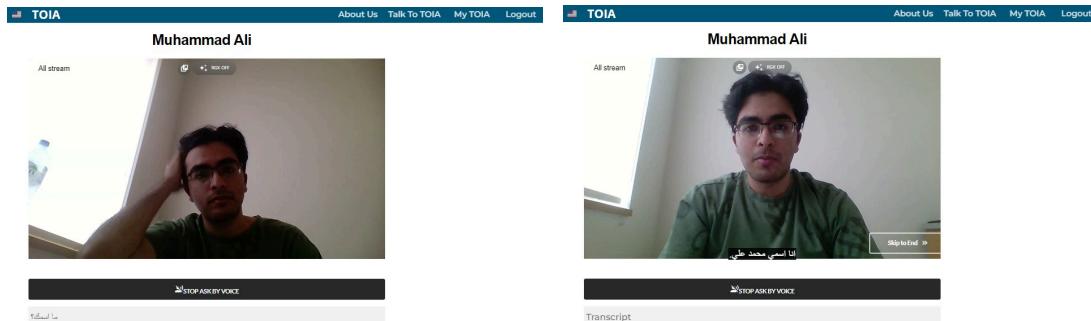


Figure (8.3) Screenshots of new developments using our TOIA for building multilingual interaction: on the left, an interactor asks a question in Arabic (see the Arabic automatic transcription on the bottom); on the right, the *TOIA avatar* answers in English (the language setting shows on the top-left corner an American flag indicating the US-English language as default) with a subtitle in Arabic.

**Multilanguage Experiment** Another exciting area for future research is exploring TOI in different languages. Multilingual BERT models could be used for modeling other languages, but more work may need to be done to ensure that the technology is effective for TOI in multiple languages. Some undergraduate students working on their capstone project in our lab, are already using our TOIA

for building interactions from many-to-many languages (e.g., one can interact in Arabic or French with Italian- or Spanish-speaking *TOIA avatars*). Figure 8.3 shows a demo of their product.

**API Integration** TOI could be developed as an API that integrates with social media or any website. An API would enable developers to integrate TOI into their applications, increasing the reach of the technology and allowing it to be more widely used. Future research could focus on developing a TOI API that is easy to use and accessible for developers.

**Long-Video Recording** Finally, a promising area for future research is recording long videos to parse the content and make an avatar out of it. By registering a long video of a person, it may be possible to parse the content and create an avatar that can interact with users in the future. This technology would enable people to communicate with others from different timelines, even if the person is no longer alive.

### 8.2.3 Privacy and Ethical Considerations

Before the widespread use of TOI, it is essential to consider the privacy and ethical implications. Currently, TOI technology is limited, so there are few risks and concerns associated with its use. But as the technology develops, so will its risks, privacy, and ethical concerns. Privacy is a significant concern as the time-offset interactions occur in real-time, and the recordings could be misused. Ethical considerations include the impact of time-offset communication on people's lives, mental well-being, and self-expression. For instance, Chapter 6 reveals that a

medium can alter some of a person's personality traits, making them appear more or less pronounced.

In his book about the ethics of AI, Chierici (2021) [12] discusses the difference between human progress and human development. Progress is defined as the movement toward a destination, while development refers to the process of growing, maturing, and innovating. The author argues that technological advancements like AI are considered developments rather than progress and defining what progress means and the endgame is essential. The practical reality of today's AI technology goals is that they are aligned with utilitarian and profit-making objectives, and the historical development of AI has been driven mainly to satisfy economic goals. The same book discusses the commercial success of AI, how it drove research and investment in R&D, and how it was built mainly to satisfy economic goals. The issue is whether or not the economic goal is morally viable, and the book explores the anthropological view of the classical economy to understand this.

When exploring ethical consequences and analyzing the risks of TOI, it is essential to explore and have open conversations about different and potentially conflicting anthropological views. In this way, a sense of moral virtue and consensus is built. Such conversations can provide guidance on best practices and ensure that the technology is developed in an ethical manner.

\*\*\*

TOIAs offer a unique opportunity for individuals to engage in meaningful conversations with each other despite time and distance. Time-offset interaction can

revolutionize communication and open up new possibilities in various areas of life, such as talking with future generations, art installations, interacting with celebrities, reflective introspection, discussing taboo topics in complete privacy and confidence, mental health coaching, preparing for job interviews and education. TOIAs should be further investigated to realize their full potential because they are crucial tools for enabling contemporary experiences.

The concept of time-offset interaction has significant potential for various research fields, including psychology, communication, and computer science. Future research could focus on improving the seamless interaction, designing mobile-responsive applications, handling multilingual experiments, developing TOIAs as an API (Application Programming Interface), and recording long videos to parse the content and make an avatar out of it. These research areas could unlock the full potential of time-offset interaction and bring about new and exciting possibilities for the future.

## Appendix A

### Supplementary Materials

This appendix includes supplementary materials, such as tables and figures, which serve to enrich and complement the information presented throughout the thesis. We provide these additional resources for the reader's convenience. They serve as an informative reference to further enhance their understanding of the subject matter.

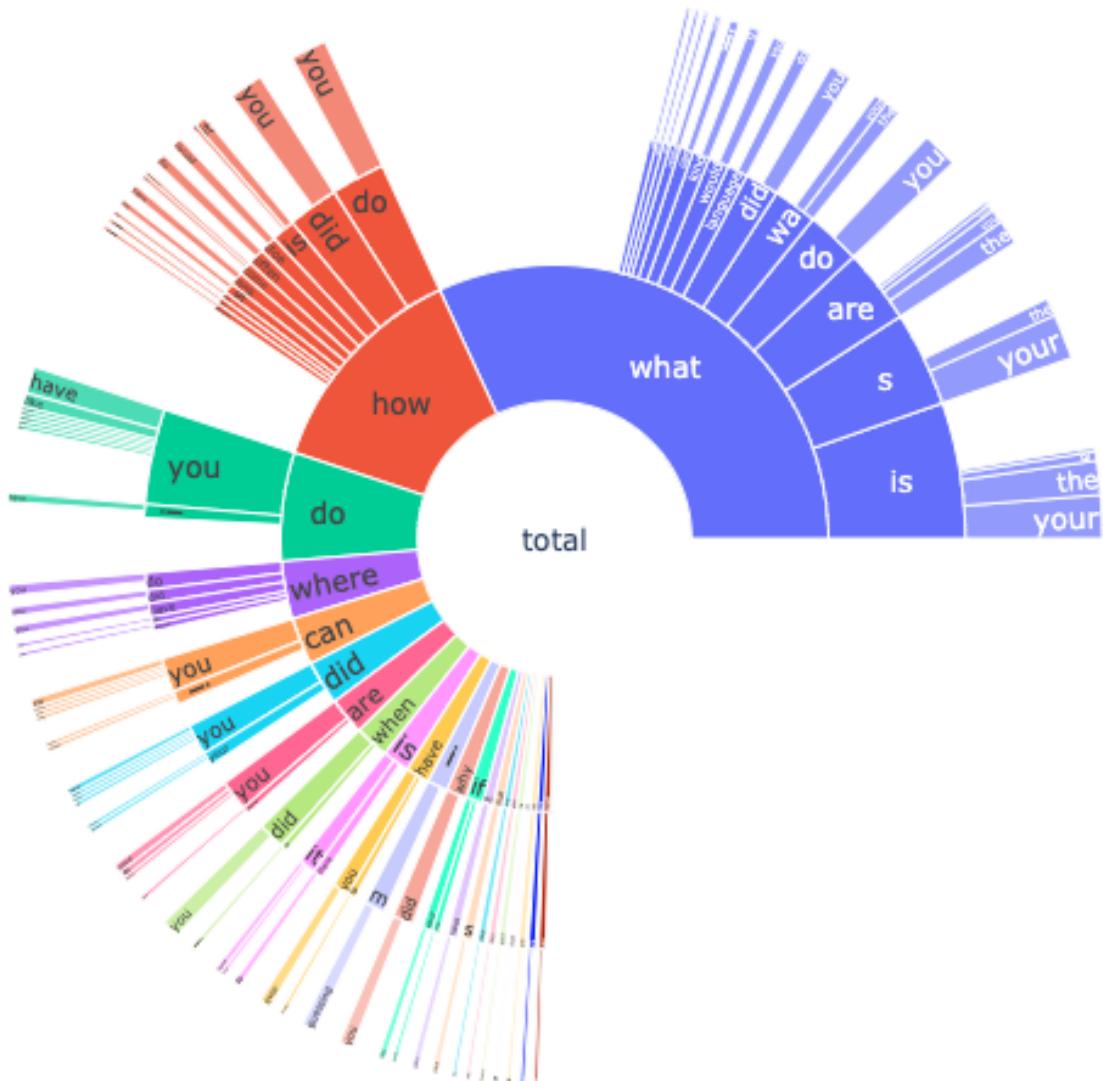


Figure (A.1) Distribution of trigram prefixes of **QUESTIONS** in the Margarita Dialogue Corpus' **Knowledge Base**.

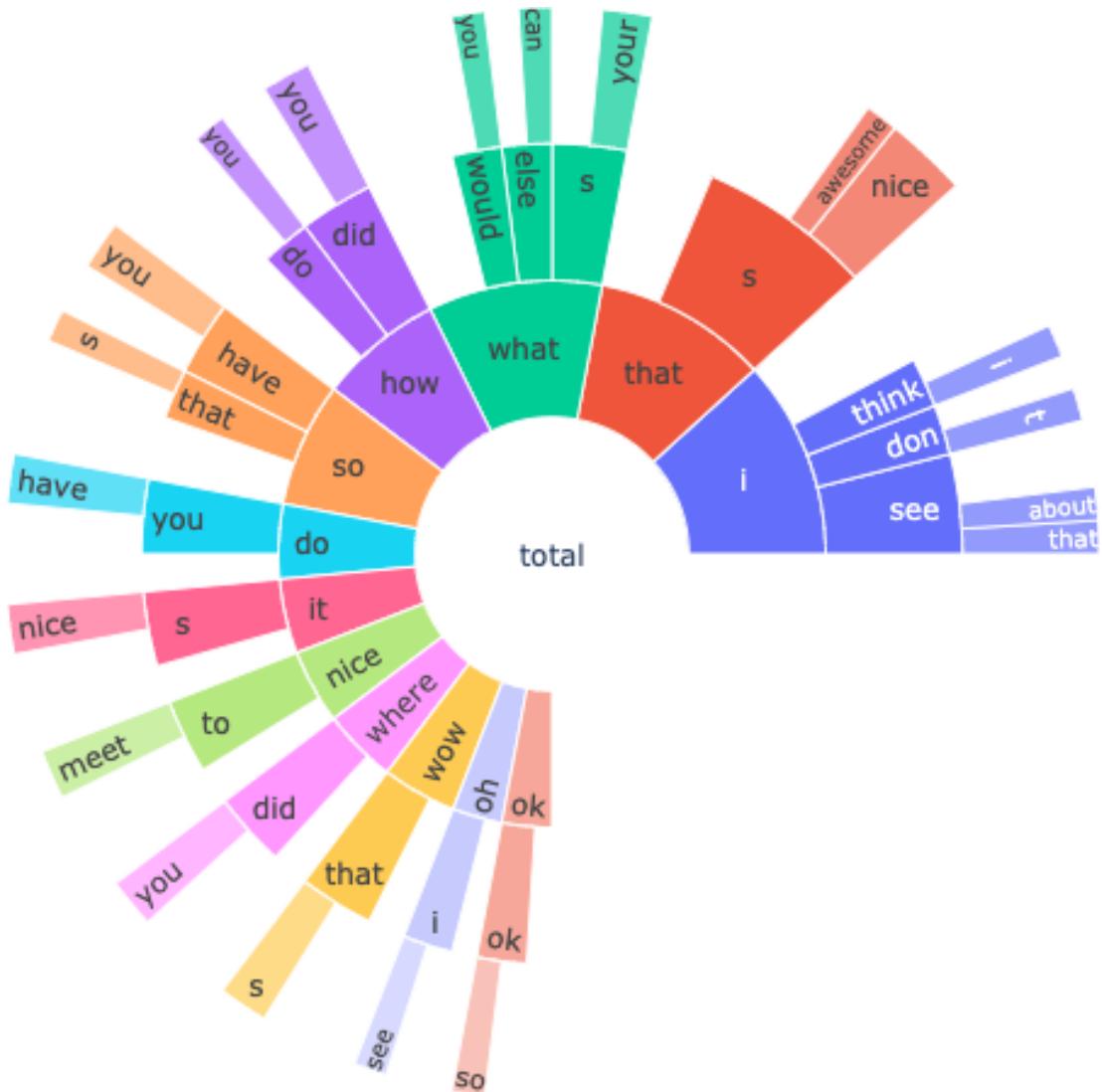


Figure (A.2) Distribution of trigram prefixes of **questions** in the Margarita Dialogue Corpus' dialogue **DEV** sets.

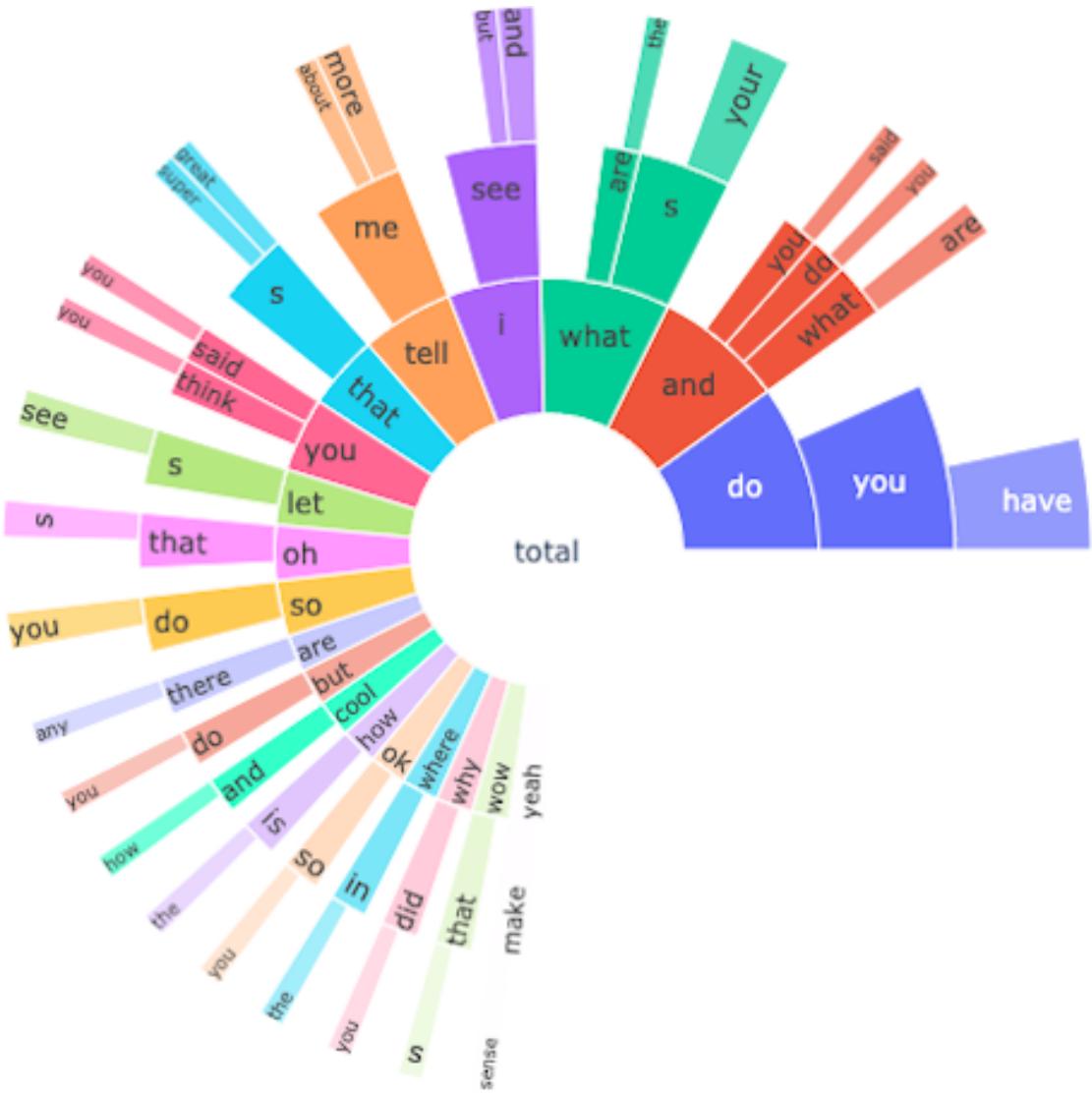


Figure (A.3) Distribution of trigram prefixes of **questions** in the Margarita Dialogue Corpus' **dialogue TEST** sets.

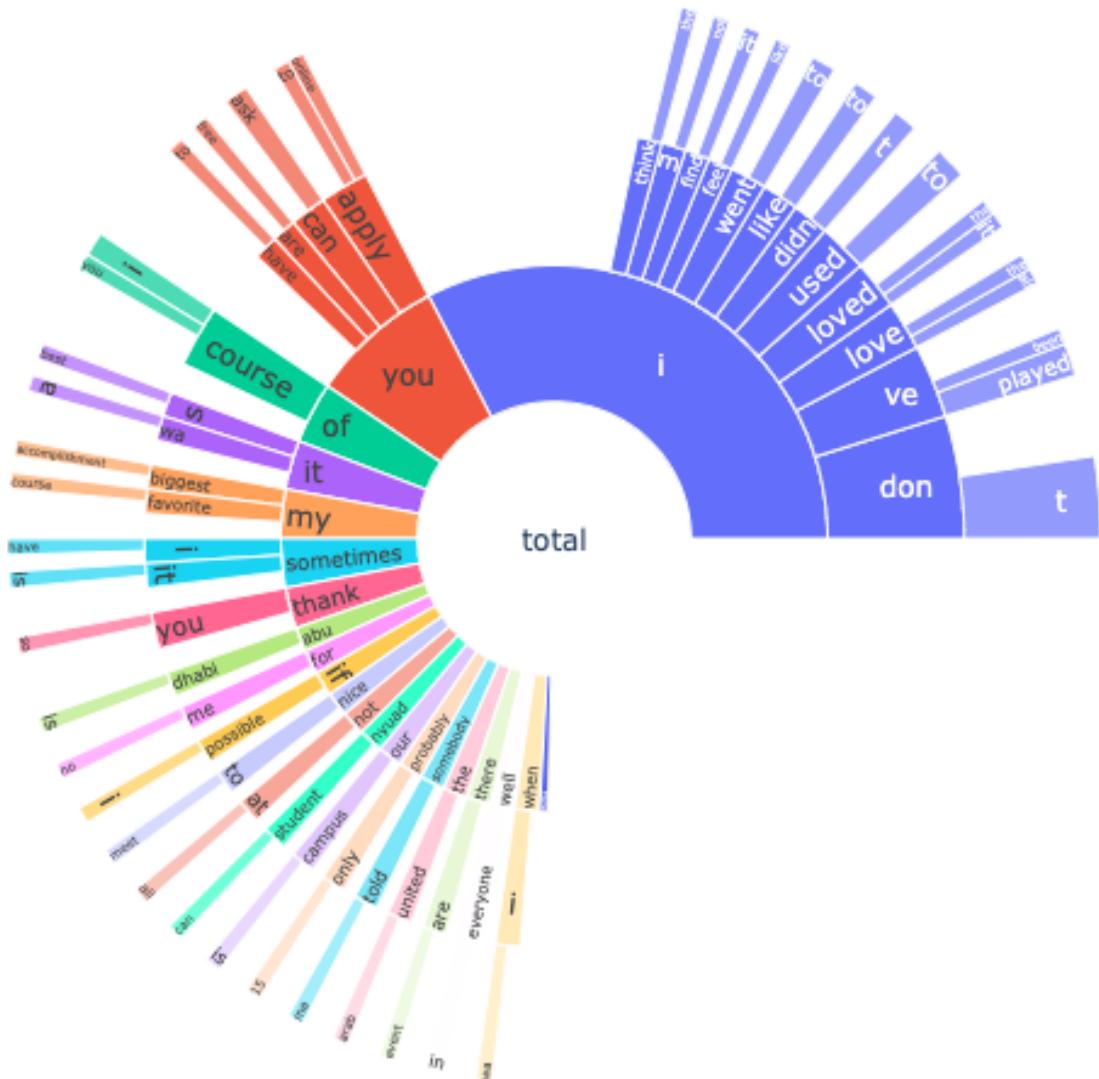


Figure (A.4) Distribution of trigram prefixes of **ANSWERS** in the Margarita Dialogue Corpus' **Knowledge Base**.

	Tot	By video-type					
		answer	exit	filler	greeting	no-answer	y/n-answer
<i>All (N=26 Subjects)</i>							
# q-a pairs	3,548	2,407	47	696	49	157	192
(Avg./subject)	136.5	92.6	1.8	26.8	1.9	6.0	7.4
# words	606,458	536,318	2,600	43,784	1,645	12,560	9,551
Avg. # words/question	43.1	40.1	31.4	59.3	22.8	32.7	38.2
Avg. # words/answer	127.9	182.8	23.9	3.6	10.7	47.3	11.5
Avg. Video duration (s)	12.9	16.5	3.3	6.6	2.5	5.1	2.7
<i>GPT-3-QS Cohort (N=10 Subjects)</i>							
# q-a pairs	1,538	1,067	20	284	18	70	79
(Avg./subject)	153.8	106.7	2.0	28.4	1.8	7.0	7.9
# words	251,522	223,504	1,127	17,518	561	4,815	3,997
Avg. # words/question	43.0	40.9	31.2	58.8	22.9	28.7	35.9
Avg. # words/answer	120.5	168.6	25.2	2.9	8.3	40.1	14.7
Avg. Video duration (s)	13.1	16.4	3.7	6.9	2.3	5.1	3.0
<i>Human-QS Cohort (N=8 Subjects)</i>							
# q-a pairs	1,094	791	12	198	16	41	36
(Avg./subject)	136.8	98.9	1.5	24.8	2.0	5.1	4.5
# words	218,935	197,552	739	13,555	641	4,269	2,179
Avg. # words/question	45.2	41.0	36.5	64.2	25.3	41.0	50.9
Avg. # words/answer	154.9	208.8	25.1	4.3	14.8	63.1	9.6
Avg. Video duration (s)	15.0	18.2	3.3	7.8	2.9	5.7	2.7
<i>QS-Off Cohort (N=8 Subjects)</i>							
# q-a pairs	916	549	15	214	15	46	77
(Avg./subject)	114.5	68.6	1.9	26.8	1.9	5.8	9.6
# words	136,001	115,262	734	12,711	443	3,476	3,375
Avg. # words/question	40.5	37.1	27.7	55.4	20.1	31.3	34.7
Avg. # words/answer	108.0	172.8	21.2	4.0	9.4	44.2	9.2
Avg. Video duration (s)	10.2	14.3	2.9	5.0	2.2	4.6	2.5

Table (A.1) Summary statistics on the data sets collected in the user study on the avatar creation (CFR, Chapter 6). Statistics for the various *TOIA avatars*' knowledge bases are also shown for each video-type and by the experimental condition co-hort (Question Suggester powered by GPT-3, by a human, and switched off).

	Tot	By Interaction Type		
		CARD	SEARCH	VOICE
# dialogues	317			
# q-a pairs	9,684	2,955	2,579	4,150
# no-answers (in %)	792 8.2%	17 0.6%	182 7.1%	593 14.3%
# words	1,602,582	581,826	426,964	593,792
Avg. # turns/dialogue	30.5	9.3	8.1	13.1
Avg. # words/question	32.5	38.8	31.9	28.3
Avg. # words/answer	133.0	158.1	133.7	114.8

Table (A.2) Summary statistics on the dialogues collected from the interaction user study's chat logs (CFR, Chapter 7). Statistics are also shown for each type of interaction with the player interface (CARD, SEARCH, VOICE).

Question	Input Type
First Name	Text
Last Name	Text
Age	Number
Gender	Male Female Other
Mother Tongue	Text
Did you enjoy your experience creating your avatar?	1-5 Scale
Describe briefly your experience using the software. <i>What did you find working well, and what areas we shall improve?</i>	Text
	< 3 hours
How long did it take you to create your avatar?	3-10 hours 10-30 hours 30+ hours
Rate your experience with the technology	
[Ease to record a new video]	1-5 Scale
[Ease to delete an existing video]	1-5 Scale
[Ease to edit a previously recorded video]	1-5 Scale
[Satisfaction with automatic speech transcriptions <i>(think how much you needed to edit the text of your answers)</i>	1-5 Scale
[Satisfaction with the overall performance of the software]	1-5 Scale
Overall, how many times did you have to re-type your video transcriptions?	1-5, 6-10, 10-20, 20+
Describe your experience with coming up with new videos to record.	
<i>Did you struggle to come up with inspiration, and why?</i>	Text
Do you want to keep using this software? <i>Please elaborate on why and for what uses, or why not.</i>	Text
Any other comments on your experience with recording your avatar?	Text
Would you like to interact with other avatars?	Text
Who would you like to see an avatar of?	Text
How do you see yourself? [I am reserved.]	1-5 Scale
...	...
See Table A.6	

Table (A.3) Questions and type of answer collected in the post-experiment survey issued to *TOIA* *avatar*-makers. The full experiment is described in Chapter 6.

Question	Short Form	Input Type
First Name		Text
Last Name		Text
Age		Number
Gender	Gender	Male Female Other
Mother Tongue	Language	Text
<b>How do you see yourself?</b>	[I am reserved.] ... See Table A.6	1-5 Scale ... Text
<b>Please confirm the name of the avatar you interacted with.</b>		
<b>How long was your interaction with the avatar?</b>	Duration	< 0 min 10-15 min 15-20 min 20+ min
<b>Rate your experience with the Avatar</b>		
[Ease in finding answers to your questions ( <i>excluding questions that you were informed had no prerecorded answers to</i> )]	Find	1-5 Scale
[Relevance of the avatar's responses to your questions or comments ( <i>including those you were informed had no answers to</i> )]	Relevance	1-5 Scale
[Ease at which the avatar seemed to understand your questions and/or comments]	Understand	1-5 Scale
[Speed at which the avatar responded to your questions ]	Speed	1-5 Scale
[The pace of the overall conversation felt natural]	Pace	1-5 Scale
[The overall flow of interacting with the avatar]	Flow	1-5 Scale
[Satisfaction with the overall performance of the avatar]	Satisfaction	1-5 Scale
<b>We are running a little contest for the best avatar: based on your interaction, do you recommend this avatar for the best avatar prize?</b>	Best	1-5 Scale

Table (A.4) (1/2) Questions and type of answer collected in the post-experiment survey issued to participants who interacted with the *TOIA avatars*. The full experiment is described in Chapter 7.

Question	Short Form	Input Type
Overall, how many times did you have to rephrase your questions before finding the answer you were looking for? ( <i>excluding questions that you were informed had no prerecorded answers to</i> )	Rephrase	0, 1-2 3-4, 4+
Overall, how many times did you have to repeat the same question/comment for the avatar to respond relevantly?	Repeat	0, 1-2 3-4, 4+
For questions with no answers, how many times did you have to ask a question before you realized there probably is no recorded answer to your question?	No answer	Once Twice More
Name three-five facts that come up immediately in your mind about the avatar you interacted with.	Facts	Text
Describe the personality of the avatar you interacted with. [They are reserved.]		1-5 Scale
...		...
See Table <a href="#">A.6</a>		
Any other comments on your experience with the avatar?		Text
Did you enjoy your experience with the avatar?	Enjoyed	1-5 Scale
Would you like to interact with other avatars?	Interact	Yes No Not Sure
Would you create an avatar of yourself?	Create	Yes No Not Sure
Who would you like to see an avatar of?		Text
How did it feel talking to the avatar?	Feel	Text
Rate your experience with the quality of technology		
[Quality of sound]	Sound	1-5 Scale
[Quality video]	Video	1-5 Scale
[Ease in using the interface]	Interface	1-5 Scale

Table (A.5) (2/2) from Table [A.4](#).

<b>Question</b>	<b>Input Type</b>
<b>How do you see yourself?</b>	
[I am reserved.]	1-5 Scale
[I tend to be lazy.]	1-5 Scale
[I am relaxed and can handle stress well.]	1-5 Scale
[I have few artistic interests.]	1-5 Scale
[I tend to find fault with others.]	1-5 Scale
[I am outgoing, sociable.]	1-5 Scale
[I do a thorough job.]	1-5 Scale
[I get nervous easily.]	1-5 Scale
[I have active imagination]	1-5 Scale
[I am generally trusting to others.]	1-5 Scale

Table (A.6) Ten-item personality inventory questionnaire to establish the Big-5 personality dimensions (OCEAN).

# Bibliography

- [1] D. Abu Ali, M. Ahmad, H. Al Hassan, P. Dozsa, M. Hu, J. Varias, and N. Habash. A bilingual interactive human avatar dialogue system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 241–244, 2018.
- [2] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.
- [3] S. Amershi, D. Weld, M. Vorvoreanu, A. Journey, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. Guidelines for human-ai interaction. 2019.
- [4] R. Artstein, A. Gainer, K. Georgila, A. Leuski, A. Shapiro, and D. Traum. New dimensions in testimony demonstration. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 32–36, 2016.
- [5] R. Artstein, A. Leuski, H. Maio, T. Mor-Barak, C. Gordon, and D. Traum. How many utterances are needed to support time-offset interaction? In *The Twenty-Eighth International Flairs Conference*. Citeseer, 2015.

- [6] R. Artstein, D. Traum, O. Alexander, A. Leuski, A. Jones, K. Georgila, P. Debevec, W. Swartout, H. Maio, and S. Smith. Time-offset interaction with a holocaust survivor. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 163–168, 2014.
- [7] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [8] B. E. A. Boussaha, N. Hernandez, C. Jacquin, and E. Morin. Deep retrieval-based dialogue systems: a short review. *arXiv preprint arXiv:1907.12878*, 2019.
- [9] A. Chen, G. Stanovsky, S. Singh, and M. Gardner. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, 2019.
- [10] A. Chierici, N. Habash, and M. Bicec. The margarita dialogue corpus: A data set for time-offset interactions and unstructured dialogue systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 476–484, 2020.
- [11] A. Chierici, T. K. F. Hensley, W. Kamran, K. Koss, A. Agrawal, E. Meekhof, G. Puccetti, and N. Habash. A cloud-based user-centered time-offset interaction application. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 265–268, 2021.

- [12] A. M. Chierici. *The ethics Of AI: Facts, Fictions and Forecasts*. New Degree Press, 2021.
- [13] J. Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.
- [14] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.
- [15] C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*, 2011.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [17] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.
- [18] E. Fast, B. Chen, J. Mendelsohn, J. Bassan, and M. S. Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 473. ACM, 2018.
- [19] P. Forchini. Using movie corpora to explore spoken american english. *Variation and Change in Spoken and Written Discourse: Perspectives from corpus linguistics*, 21:123, 2013.

- [20] J. Gao, M. Galley, L. Li, et al. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298, 2019.
- [21] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.
- [22] S. D. Gosling, A. A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14(9):483–488, 2011.
- [23] P. J. Guo and M. I. Seltzer. Burrito: Wrapping your lab notebook in computational infrastructure. 2012.
- [24] A. Hertzfeld. *Revolution in The Valley [Paperback]: The Insanely Great Story of How the Mac Was Made.* ” O'Reilly Media, Inc.”, 2004.
- [25] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*, 2019.
- [26] P. Jarusriboonchai, T. Olsson, V. Prabhu, and K. Väänänen-Vainio-Mattila. Cuesense: A wearable proximity-aware display enhancing encounters. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2127–2132, 2015.
- [27] A. Jones, J. Unger, K. Nagano, J. Busch, X. Yu, H.-Y. Peng, O. Alexander, M. Bolas, and P. Debevec. An automultiscopic projector array for interactive

- digital humans. In *ACM SIGGRAPH 2015 Emerging Technologies*, page 6. ACM, 2015.
- [28] D. Jurafsky and J. H. Martin. Speech and language processing (3rd draft ed.), 2019.
- [29] C. Khatri, B. Hedayatnia, A. Venkatesh, J. Nunn, Y. Pan, Q. Liu, H. Song, A. Gottardi, S. Kwatra, S. Pancholi, et al. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*, 2018.
- [30] H. Kim, S. I. Di Domenico, and B. S. Connally. Self-other agreement in personality reports: A meta-analytic comparison of self-and informant-report means. *Psychological science*, 30(1):129–138, 2019.
- [31] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387, 2016.
- [32] A. Leuski, R. Patel, D. Traum, and B. Kennedy. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, 2006.
- [33] 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.
- [34] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

- [35] P. Lison and J. Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.
- [36] 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*, pages 2122–2132, Austin, Texas, Nov. 2016. Association for Computational Linguistics.
- [37] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.
- [38] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*, 2017.
- [39] R. Lowe, N. Pow, I. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015.
- [40] M. Ma, S. Coward, and C. Walker. Question-answering virtual humans based on pre-recorded testimonies for holocaust education. In *Serious games and edutainment applications*, pages 391–409. Springer, 2017.
- [41] W. Ma, R. Takanobu, M. Tu, and M. Huang. Bridging the gap between conversational reasoning and interactive recommendation. *arXiv preprint arXiv:2010.10333*, 2020.

- [42] C. Manning, P. Raghavan, and H. Schutze. Introduction to information retrieval. 2008.
- [43] D. Marinelli and S. Stevens. Synthetic interviews: the art of creating a 'dyad' between humans and machine-based characters. In *Proceedings 1998 IEEE 4th Workshop Interactive Voice Technology for Telecommunications Applications. IVTTA '98 (Cat. No. 98TH8376)*, pages 43–48. IEEE, 1998.
- [44] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, and M. Geist. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762, 2020.
- [45] S. K. Mishra, P. Goel, A. Sharma, A. Jagannatha, D. Jacobs, and H. Daume. Towards automatic generation of questions from long answers. *arXiv:2004.05109*, 2020.
- [46] T. Miyauchi, M. Nishiyama, and Y. Iwai. Directing a target person among multiple users using the motion effects of an image-based avatar. In *International Conference on Human-Computer Interaction*, pages 341–352. Springer, 2019.
- [47] T. Miyauchi, A. Ono, H. Yoshimura, M. Nishiyama, and Y. Iwai. Embedding the awareness state and response state in an image-based avatar to start natural user interaction. *IEICE TRANSACTIONS on Information and Systems*, 100(12):3045–3049, 2017.
- [48] A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

- [49] T. T. Nguyen, D. T. Nguyen, S. T. Iqbal, and E. Ofek. The known stranger: Supporting conversations between strangers with personalized topic suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 555–564, 2015.
- [50] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 1994.
- [51] M. Nishiyama, T. Miyauchi, H. Yoshimura, and Y. Iwai. Synthesizing realistic image-based avatars by body sway analysis. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 155–162. ACM, 2016.
- [52] R. E. Page. *Stories and social media: Identities and interaction*. Routledge, 2013.
- [53] 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*, pages 311–318, 2002.
- [54] E. Park, D. Jin, and A. P. del Pobil. The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 9, 2012.
- [55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- [56] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf*, 2018.
- [57] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- [58] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.
- [59] J. Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ, 2003.
- [60] S. Reddy, D. Chen, and C. D. Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.
- [61] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics, 2010.
- [62] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

- [63] W. Sakata, T. Shibata, R. Tanaka, and S. Kurohashi. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116, 2019.
- [64] H. Salam, O. Celiktutan, H. Gunes, and M. Chetouani. Automatic context-driven inference of engagement in hmi: A survey. *arXiv preprint arXiv:2209.15370*, 2022.
- [65] H. Salam, V. Manoranjan, J. Jiang, and O. Celiktutan. Learning personalised models for automatic self-reported personality recognition. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 53–73. PMLR, 2022.
- [66] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [67] J. Schatzmann, K. Georgila, and S. Young. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*, 2005.
- [68] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [69] D. Shin, S. Yoon, S. Kim, and J. Lee. Blahblahbot: Facilitating conversation between strangers using a chatbot with ml-infused personalized topic suggestion.

In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

- [70] H.-Y. Shum, X.-d. He, and D. Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, 2018.
- [71] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [72] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambrø, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [73] D. Traum, K. Georgila, R. Artstein, and A. Leuski. Evaluating spoken dialogue processing for time-offset interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 199–208, 2015b.
- [74] D. Traum, A. Jones, K. Hays, H. Maio, O. Alexander, R. Artstein, P. Debevec, A. Gainer, K. Georgila, K. Haase, et al. New dimensions in testimony: Digitally preserving a holocaust survivor’s interactive storytelling. In *Proc. of International Conference on Interactive Digital Storytelling*, 2015.
- [75] D. Traum, A. Jones, K. Hays, H. Maio, O. Alexander, R. Artstein, P. Debevec, A. Gainer, K. Georgila, K. Haase, et al. New dimensions in testimony: Digitally preserving a holocaust survivor’s interactive storytelling. In *International Conference on Interactive Digital Storytelling*, pages 269–281. Springer, 2015a.

- [76] A. Trotman, A. Puurula, and B. Burgess. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65, 2014.
- [77] A. M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 10 1950.
- [78] V. Vikashini, H. Salam, J. Nasir, B. Bruno, and O. Celiktutan. Personalized productive engagement recognition in robot-mediated collaborative learning. In *24th ACM International Conference on Multimodal Interaction (ICMI)*, number CONF, 2022.
- [79] J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6383–6389, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [80] J. Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [81] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.
- [82] G. Wilcock. Wikitalk: A spoken wikipedia-based open-domain knowledge

- access system. In *Proceedings of the workshop on question answering for complex domains*, pages 57–70, 2012.
- [83] J. Williams, A. Raux, D. Ramachandran, and A. Black. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, 2013.
- [84] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.
- [85] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*, 2016.
- [86] S. K. Yuwono, B. Wu, and L. F. D’Haro. Automated scoring of chatbot responses in conversational dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 357–369. Springer, 2019.
- [87] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [88] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*, 2020.

ProQuest Number: 30491705

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality  
and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2023).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license  
or other rights statement, as indicated in the copyright statement or in the metadata  
associated with this work. Unless otherwise specified in the copyright statement  
or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,  
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization  
of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 - 1346 USA