HelloThere: A Corpus of Annotated Dialogues and Knowledge Bases of Time-Offset Avatars

Alberto Chierici, Nizar Habash
Computational Approaches for Modeling Languages (CAMeL) Lab
New York University Abu Dhabi
{alberto.chierici, nizar.habash}@nyu.edu

Abstract

A Time-Offset Interaction Application (TOIA) is a software system that allows people to engage in face-to-face dialogue with previously recorded videos of other people. There are two TOIA usage modes: (a) creation mode, where users pre-record video snippets of themselves representing their answers to possible questions someone may ask them, and (b) interaction mode, where other users of the system can choose to interact with created avatars. This paper presents the HelloThere corpus that has been collected from two user studies involving several people who recorded avatars and many more who engaged in dialogues with them. The interactions with avatars are annotated by people asking them questions through three modes (card selection, text search, and voice input) and rating the appropriateness of their answers on a 1 to 5 scale. The corpus, made available to the research community, comprises 26 avatars' knowledge bases and 317 dialogues between 64 interrogators and the avatars in text format.

1 Introduction

Time-Offset Interaction Applications (TOIAs) have evolved as an innovative dialogue system, bridging the interaction between individuals and prerecorded video representations of others, hence enabling users to hold conversations outside realtime constraints (Artstein et al., 2015; Traum et al., 2015; Abu Ali et al., 2018). We built on an open-source project's application, offering a dual interface targeting two distinct user groups: (a) avatar creators, individuals interested in generating their time-offset personas, and (b) interactors, those who engage with these avatars.

However, designing a robust TOIA is a challenging endeavor. The goal is to mirror human-to-human interactions as authentically as possible. This demands seamless integration from an engineering standpoint, such as flawless video clip transitions and numerous linguistic and dialogue-turns complexities that intrigue dialogue system researchers. Central to a TOIA's functionality are the avatar's Knowledge Bases (KBs), repositories of questions paired with corresponding video responses and their transcriptions. One of the inherent challenges is devising an optimal strategy for populating this KB. Should it be intuition-driven, or should it stem from authentic dialogue transcripts? Furthermore, what data sets can be useful for training models to retrieve the right answer for an interrogator interacting with the avatars? While we explored such questions in other research (Chierici et al., 2020; Chierici and Habash, 2021, 2023), here we focus on building on such body of work and present the language resources generated in the process. We explored KBs created in three distinct ways: intuition-guided (brainstormed), led by automatic suggestions (generated by GPT-3), and led by human suggestions. We used GPT-3 because our software and study were designed and set up between 2022 and 2023 before newer versions were available. The **HelloThere Corpus** offers a unique resource for dialogue researchers, enabling studies on multi-modal interactions, user engagement patterns, and the effectiveness of time-offset avatar responses. By providing annotated dialogues across different interaction modes, this corpus supports research into natural language understanding, response retrieval and generation, and user experience in asynchronous communication systems.

2 Related Work

We categorize pertinent literature on Time-Offset Interaction Applications (TOIA) into three primary areas: System Approaches, Data Sources, and Evaluation Methodologies.

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 139–148
September 18–20, 2024. ©2024 Association for Computational Linguistics

2.1 System Approaches

Our work builds upon the foundations laid in Chierici et al. (2020); Chierici and Habash (2021); Chierici et al. (2021), whose initial inspiration stemmed from the work of Traum et al. (2015) in their New Dimensions in Testimony project. While Traum et al. created a time-offset interaction with Holocaust survivor Pinchas Gutter, we extend their approach to different contexts and focus on system scalability. The TOIA open-sourced in (Chierici et al., 2021) aims to operate with fewer recorded statements, adapt to multiple users, and facilitate getting to know a stranger in a 10- to 15-minute interaction.

Following the taxonomy we proposed in Chierici et al. (2020), we work on a novel subcategory of 'self-narrative bots,' which can be seen as an intermediate between social and task-driven bots, leveraging both structured and unstructured training data (Gao et al., 2019). Retrieving the appropriate video from a TOIA Knowledge Base (KB) shares similarities with FAQ retrieval, a dichotomous problem. While its single-turn question-answer (q-a) mechanism may seem rudimentary, tasks like search and Retrieve-And-Generate – where a model retrieves relevant information and generates a response based on it – introduce complexities due to the dynamic nature of dialogue (Mass et al., 2020; Yehudai et al., 2023).

As the dataset scales, classification approaches may falter, highlighting the presence of long-tail problems and the challenges of chit-chat scenarios, where queries can have subtle differences (e.g., "What is your name?" vs. "What is your parent's name?"). Technologies involved range from traditional RNN models and word embeddings to newer language models like OpenAI's GPT families, Mistral, Llama and Nomic (Radford et al., 2018; Zhang et al., 2022; Touvron et al., 2023; Jiang et al., 2023; Nussbaum et al., 2024).

Recent advancements in neural architectures have led to cutting-edge performance in answer retrieval tasks, but the limited scale of our dialogue datasets–and those of similar scope–does not readily support deep learning approaches. This limitation does not preclude using pre-trained large language models for sentence similarity tasks, leveraging or not transfer and few-shot learning techniques.

While TOIAs share some similarities with recent advancements in speech and video synthesis technologies, they differ in their focus on preserving authentic human responses. Unlike synthetic systems that generate responses in real-time, TOIAs rely on pre-recorded human responses, maintaining the nuances of human communication. However, the retrieval mechanisms in TOIAs can benefit from advancements in natural language processing used in synthetic systems, particularly for improving response selection accuracy.

2.2 Data Sources

Various datasets have been employed to tackle problems related to chit-chat and question answering in dialogue systems, such as SQuAD (Rajpurkar et al., 2016), the Ubuntu dialogue corpus (Lowe et al., 2015), and bAbI (Weston et al., 2015). However, these works address tasks like time-based reasoning and logical induction, which differ from the context of TOIAs. The landscape of dialogue-focused datasets is evolving to capture complexities absent in earlier reading comprehension collections. Datasets like CoQA and HUMOD are designed with human dialogues and annotations in mind, enhancing natural conversational elements (Reddy et al., 2019; Merdivan et al., 2020). Similarly, the Douban Conversation Corpus offers insights into real-world social discussions on various topics (Wu et al., 2016).

While large-scale datasets serve various purposes, dialogue systems often operate with far smaller datasets. For instance, the Margarita Dialogue Corpus (MDC) features a Knowledge Base (KB) with only 431 answers and complete annotated dialogues (Chierici et al., 2020). The nuanced context of dialogue in TOIAsdemands different, more tailored datasets. The MDC offers a unique blend of structured and unstructured dialogues for time-offset interactions. While influential for our work, it is limited to a single avatar and real person-to-person transcripts, not mediated through a TOIA interface. This work extends the MDC by incorporating more avatars and collecting extensive real-world interactions with them, addressing identified limitations and enriching the corpus.

In previous work (Chierici and Habash, 2023; Chierici, 2023), we addressed a key challenge in TOIA development – the daunting task of creating extensive video-anchored question-answer (q-a) pair databases without overwhelming the avatar maker, improving upon Chierici et al. (2020). We introduced Question Suggester (QS), a GPT-3-based intelligent service designed to alleviate this problem by dynamically suggesting relevant follow-up questions based on the existing conversation history, significantly reducing the effort required to populate the video database and enhancing user experience.

2.3 Evaluation

We acknowledge that evaluating dialogue systems is a complex task, as traditional metrics often fail to correlate with human judgment, which itself is challenging to quantify (Li et al., 2019). The corpus we present addresses some gaps highlighted in Chierici and Habash (2021), where we performed a human evaluation study with a fictional TOIA interface and Amazon Mechanical Turk raters. We deployed the open-source software described in Chierici et al. (2021), with updates to the dialogue systems module and user interface, and built datasets using real TOIA-interactions. Participants were tasked with getting to know the avatar creator within a 10-minute interaction, evaluating each response as they interacted with the tool.

3 Data Acquisition and Annotation

Our work resulted in collecting and annotating dialogue data comprising 2.2 million words. This effort was part of a large user study involving 90 individuals, some who built the avatars, and others evaluated their interaction quality, along with testing and evaluating a few software features and related research questions discussed in (Chierici and Habash, 2023). Ethical considerations were upheld as our institution's Institutional Review Board approved the experiments, and participants consented to release data transcriptions, annotations, and video recordings for research purposes only. In both parts of the study, participants were university students recruited via an online form that included informed consent and details about the study. In the first part, 26 individuals aged 18-24 participated, with 14 females and various international provenance. They are fluent in English and major in various fields, mostly science. In the second part, 64 people participated. They were mostly between the ages of 18-23, and 35 were female. All are also fluent in English, though 80% consider it their second language. The majority were science majors, and a subset of 16 had participated in the previous part of the study. To clarify how data is collected, we describe the user interface used in the extensive user study that generated the corpora.

3.1 User Interface

The user interface (UI) components are: (see Fig. 1)

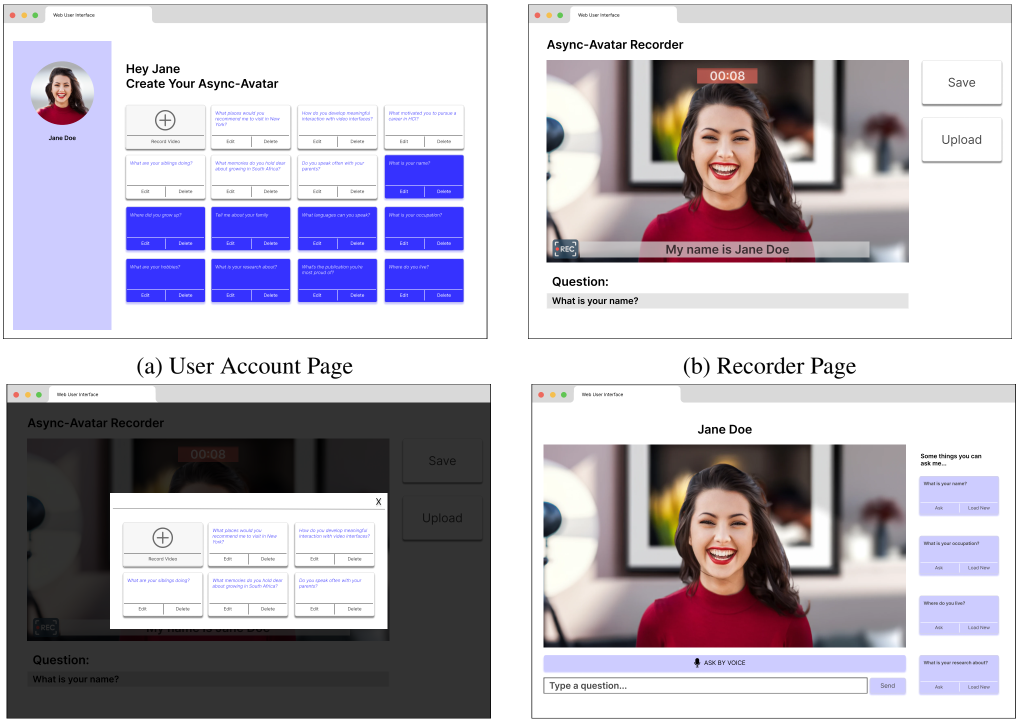
1. User Account (Fig. 1 (a)): This is the initial page that users see after creating an account. It displays a button to create new videos, suggested questions for creating new videos, and videos previously recorded by the user.

2. Recorder (Fig. 1 (b)): This page is accessed by clicking on the buttons to add a new video or edit a previously recorded video or a suggested question in the User Account page. This is where users can create new videos by 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. 1 (c)) with the command for creating a new video and follow-up question suggestions.

3. Player (Fig. 1 (d)): Here, users can interact with previously recorded videos of public *TOIA avatars* . The player interface comprises a video looping different 'filler' videos–clips without audio, where the *TOIA avatar* does not speak. Users can click on suggested questions displayed on the right side of the video, triggering an immediate response from the *TOIA avatar* . We call this interaction type 'CARD' in our later data description. Users can also ask questions verbally using a voice input button, and they are then transcribed and matched to appropriate responses. There's a button to interact with the *TOIA avatar* by voice (marked as 'VOICE' in the data), and below that button, a text input field allows users to type in their questions, which are then matched to the most relevant pre-recorded response (interaction labeled as 'TYPE' in the data). These interaction modes offer flexibility in how users engage with the avatars, catering to different preferences and contexts. <

3.2 Creating Avatars

The first step of our user study focused on evaluating the methodology for creating avatars, using both qualitative and quantitative approaches. A key aspect of this evaluation was examining the impact of different question generation methods (for a more detailed discussion of this, we refer readers to the publication presenting the user experience study, Chierici and Habash (2023)). Metrics include the efficiency of avatar creation, the quality



Hey Jane
Create Your Async-Avatar

What places would you recommend me to visit in New York?

How do you develop meaningful interaction with video interfaces?

What motivated you to pursue a career in HCI?

What are your siblings doing?

What memories do you hold dear about growing in South Africa?

Do you speak often with your parents?

What is your name?

Where did you grow up?

Tell me about your family

What languages can you speak?

What is your occupation?

What are your hobbies?

What is your research about?

What's the publication you're most proud of?

Where do you live?

(a) User Account Page

My name is Jane Doe

Question:
What is your name?

(b) Recorder Page

What places would you recommend me to visit in New York?

How do you develop meaningful interaction with video interfaces?

What are your siblings doing?

What memories do you hold dear about growing in South Africa?

Do you speak often with your parents?

Question:
What is your name?

(c) QS Pop-up

Jane Doe

Some things you can ask me...

What is your name?

What is your occupation?

Where do you live?

What is your research about?

(d) Player Page

Figure 1: User Interface (UI) designs. These are similar to what we used when collecting data, though the actual UI has since evolved. (a) is the user account page showing the QS in white backgrounds and previously recorded questions (and videos) shaded in blue; (b) is the recorder page; (c) shows suggestions appearing in a pop-up window once the user completes a recording on the Recorder page; and (d) is the player page.

of suggested questions, and the influence of the creator's personality traits on user acceptability and interface interaction. Three experimental conditions were examined when creating *TOIA avatars* ' KBs: 1) GPT-3-based question suggestions (GPT-3 QS), 2) human-curated questions (Human-QS), and 3) a no-suggestion, brainstorming condition (QS-off). As a result, 26 avatars were crafted: 10 through GPT-3 QS, 8 via Human-QS, and 8 using the QS-off approach.

3.3 Avatar Interaction

In the second step of the user study, to investigate key interaction metrics, including the minimum number of videos needed for a satisfying experience, variants of the original 26 avatars were created. These variants were based on three conditions concerning video count (first 30, first 60, or all recorded videos) and two filler videos (attentive or inattentive) types. Thus, each original avatar spawned 6 distinct interaction variants, leading to 156 unique avatars. We aimed to collect at least two evaluations for robust statistical analysis for each, totaling 312 unique dialogue interactions (to satisfy some experimental constraints and replace participants who withdrew, we ended up with 317 dialogues in total).

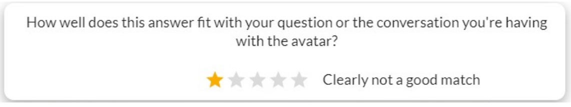
3.4 Single-turn Answer Retrieval

We employ the GPT-3 model family from OpenAI for the retrieval task, specifically geared for semantic similarity-based text search (Neelakantan et al., 2022).[1] This choice was informed by the model's superior performance tested on the Margarita Dialogue Corpus (Chierici et al., 2020). In our setup, q-a pairs are documents and converted into 1024-dimensional vector embeddings using the 'text-search-ada-doc-001' model. Incoming user queries are similarly transformed into 1024-dimensional vector embeddings through the 'text-search-ada-query-001' model. The Dialogue Manager (DM) suggests an answer when the cosine similarity between the query and document vectors exceeds a threshold of 0.29. If the similarity falls below this cutoff, the DM defaults to a predetermined set of videos intended for situations where no appropriate answer exists, such as "I haven't recorded an answer for that question." Our dia-

[1] For implementation guidelines, see https://beta.openai.com/docs/guides/embeddings.

logue data set also reports the similarity measure for each answer played out as a response to the interactors' questions.

3.5 Annotations



How well does this answer fit with your question or the conversation you're having with the avatar?

★☆☆☆☆ Clearly not a good match

Figure 2: 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.

We have three kinds of annotations. First, the Knowledge Base (KB) of each avatar who linked a question with an answer. Second, we collect the questions the automated and human QS suggested and mark them as selected or rejected by the avatar maker when creating their video recordings. Third, we have 64 human subjects who conversed with an avatar variant for a minimum of 10 minutes. We employed a 5-point rating scale, triggered by a pop-up after each video-based answer, to collect user assessments (Figure 2). Participants interacted with at least four different avatars (barring a few exceptions, who interacted with eight and one person just with one avatar).

Key conditions for the experimental design include:

- Each avatar variant received evaluations from at least two different participants.
- Participants never interacted with the same avatar more than once.
- Variants with different numbers of videos require separate evaluations.
- Filler video types were not considered separate conditions, allowing for collective evaluations.
- Interaction methods were flexible: participants could ask questions through text, voice, or preset options shown on the right of the player page by clicking on them (Fig 1d (d)).

4 Data Description and Exploration

Data for this study is accessible on NYUAD CAMeL Lab's Resource page.[2] We present the summary statistics of the two main language resources, 'Knowledge Base' and 'Dialogues', in Tables 1 and 2. We then discuss the agreement between annotations, a baseline retrieval evaluation, and a qualitative assessment of the topics covered in the corpora.

4.1 Avatar Knowledge Bases

In the first part of the human subject study (Table 1), the data generated encompasses 26 distinct avatars, each with a unique set of q-a pairs and dialogues. The data is structured into three cohorts: GPT-3-QS, Human-QS, and QS-Off, providing us with a rich platform to compare avatar behavior and performance across different conditions. The choice to create 26 distinct avatars was made to balance depth and breadth in our corpus. This number allows for a diverse range of personalities and interaction styles while remaining manageable for detailed analysis and within budget and time constraints. The distribution across different question suggestion methods (10 GPT-3 QS, 8 Human-QS, and 8 QS-off) enables comparative studies on the effectiveness of these approaches in creating engaging and comprehensive avatar knowledge bases. Here, we describe general insights and patterns observed across the three cohorts.

The corpus comprises 3,548 q-a pairs across all 26 subjects, with an average of 136.5 per subject. The data set encompasses 606,458 words, with an average of 43.1 words per question and longer answers (127.9 words on average).

The 'answer' category is overwhelmingly prevalent, constituting 2,407 of the q-a pairs—averaging about 92.6 per subject. This dominance underscores the avatars' primary role: to deliver informative and substantive responses. The Human-QS cohort exhibits the highest word count per answer, indicative of more elaborate and nuanced responses.

The Human-QS cohort answers are the longest, followed closely by those of the QS-Off cohort. Categories like 'exit,' 'greeting,' 'no-answer,' and 'y/n-answer' are relatively (and obviously) rare across all cohorts. However, they exhibit diversity in terms of average word count. These categories might be infrequent but serve specific roles within the dialogic interaction and should not be overlooked.

4.2 Dialogues

Dialogues offer a more dynamic measure of conversational capabilities and limitations, allowing

[2] http://resources.camel-lab.com/

		By video-type					
	Total	**answer**	**exit**	**filler**	**greeting**	**no-answer**	**y/n-answer**
All (N=26 Subjects)							
# 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
GPT-3-QS Cohort (N=10 Subjects)							
# 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
Human-QS Cohort (N=8 Subjects)							
# 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
QS-Off Cohort (N=8 Subjects)							
# 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

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

		By Interaction Type		
	Tot	**CARD**	**SEARCH**	**VOICE**
# dialogues	317			
# q-a pairs	9,684	2,955	2,579	4,150
# no-answers	792	17	182	593
(in %)	8.2%	0.6%	7.1%	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 2: Summary statistics on the dialogues collected from the interaction user study's chat logs. Statistics are also shown for each type of interaction with the player interface (CARD, SEARCH, VOICE).

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 3: Distribution of interactors' ratings by mode of interaction from the conversation log data of our TOIA.

for deeper understanding beyond individual, single-turn questions and answers. The data on dialogues is grouped into two key tables: Table 2 captures metrics by interaction type, while Table 3 focuses on annotations results (retrieval ratings) by mode.

The data set encompasses 317 dialogues, unfolding over 9,684 q-a pairs. These pairs are distributed across CARD (2,955), SEARCH (2,579), and VOICE (4,150) interactions. The 'No-Answers' account for 792 pairs or 8.2% of the total interactions. The dialogues encompass just over 1.6 million words, with an average of 30.5 turns per dialogue, 32.5 words per question, and 133 words per answer. The average of 30.5 turns per dialogue implies that the conversations are not just transactional but likely complex and multilayered.

VOICE-based interactions comprise the bulk of the dataset with the highest number of q-a pairs and a 14.3% 'No-Answers' rate. This suggests that voice interactions are frequent and more susceptible to information gaps or misunderstandings. The exceptionally low 'No-Answers' rate in CARD interactions (0.6%) is a consequence of the more scripted or straightforward engagement due to a deterministic retrieval (it is not 100% deterministic because the suggested cards are retrieved using prompting GPT-3 text completion and not always the underlying questions are reproduced verbatim).

CARD interactions have the highest average words per answer at 158.1, indicating a propensity for asking questions with more detailed responses in this particular mode of interaction –perhaps these are less trivial or less mundane questions that users wouldn't ask if they didn't see the suggestion on the card.

Looking at Table 3, the mean rating stands at 4.0 across all interactions with a standard deviation of 1.5. The scores range from a minimum of 1.0 to a maximum of 5.0. While VOICE accounts for 41.7% of all interactions, it has the lowest mean score of 3.5 and the same standard deviation as SEARCH. This follows from VOICE being the interaction that mostly depends on answer retrieval algorithms to provide answers. In contrast, CARD interactions have the highest mean score of 4.6 and a low standard deviation of 0.9. SEARCH interactions yield a mean of 3.9 and a slightly higher standard deviation, indicating a middle ground between VOICE and CARD. A mean score of 4.0 suggests that while the system performs reasonably well, raters may be particularly generous, and there remains scope for targeted improvements. Given the high volume but variable quality, the VOICE category could benefit from refined natural language understanding algorithms to reduce 'No-Answers' and improve consistency.

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

Table 4: Inter-annotator agreement computed using coefficients of agreement that are all relevant in our scenarios where we have multiple raters using ordinal ratings.

4.3 Retrieval Evaluation Results

The interaction experiment yielded a total of 9,100 q-a pairs, with the summary statistics and answer ratings across different interaction modalities presented in Table 3. The data show that the voice modality was the most frequently utilized method of interaction, accounting for 41.6% of the cases. This was followed by clicking on suggested questions (31.3%) and typing (27.0%). However, frequency of use does not necessarily indicate user preference. Collectively, quicker interaction modalities like clicking and typing were used more often, comprising 58.4% of the interactions.

Anomalies in the CARD mode were observed despite its deterministic nature. Although it garnered the highest average rating, some users still rated answers poorly. Closer observation revealed that misclicks and inattentiveness during ratings were the primary causes of these anomalies. The SEARCH mode revealed similar variability in user ratings, echoing the patterns observed in the VOICE mode. Due to limitations in our log data, we restricted our analysis to the SR@1 performance in VOICE interactions. Qualitative insights suggest that participants often switched between the three modalities during a conversation, primarily initiating voice interactions.

We measured retrieval success with Success Rate@1 (SR@1) based on two scenarios: including neutral ratings (3, 4, and 5), which resulted in an SR@1 of 68.2%, and only considering high ratings (4 and 5), which yielded an SR@1 of 54.5%.

4.4 TOIA Interaction Rater Agreement

Inter-rater agreement was assessed on a small sample and is reported in Table 4. To identify equal instances rated by multiple interactors, paraphrased

Theme (Short Name) & *Sample Question*
Opinion and personal beliefs (Opinion) *Do you believe in second chances?*
Reflection, Self-Awareness, Goals (Reflection) *If you were to die this evening with no opportunity to communicate with...*
Student Life, Major, Education (Education) *What are you studying right now?*
Food (Food) *What is your favorite dish at Circle Cafe?*
Preferences, Interests, and Lifestyle (Lifestyle) *What is the most crucial element in a balanced life?*
Cities, Countries and Travel (Travel) *Are you interested in traveling to Australia?*
Music (Music) *Can you recommend some songs you like?*
Books, Movies, TV (Media) *What was the last tv series you binge watched?*
Personal experiences, opinions, and advice (Advice) *When was your first kiss?*
Name, Age, Birthplace, Location (Identity) *How old are you?*
Family (Family) *What is your family like?*
Hobbies, Pastimes (Hobbies) *What's your favorite way to spend a day off?*
Animals (Animals) *If you could have an animal sidekick, what would it be and why?*
Abu Dhabi (AbuDhabi) *How is living in Abu Dhabi?*
Sports (Sports) *Are you involved in sports?*
Job, Career Aspirations, Plans After Graduation (Career) *What do you want to do after graduation?*
People Qualities and Characteristics (Traits) *What do you value in people?*
Greetings (Greetings) *Hello!*
Missing Home (Home) *Do you miss home?*
Time (Time) *What time do you...*
Miscellaneous, Trivia (Trivia) *Morgan supporting in the World Cup...*
Language (Language) *How many languages do you speak?*

Table 5: Summary of the topic clustering for questions asked by voice.

questions were grouped using cosine similarity of their sentence embeddings and checked manually to identify groups of the same question asked. A heuristically inspected threshold of 0.87 +/- 0.003 was used to cluster similar questions, leaving us with 86 comparable instances.

We computed four coefficients, namely Gwet's AC1, Fleiss Kappa, Brennan-Prediger, and Conger's kappa, to measure the agreement level. All coefficients indicated significant levels of agreement (see Table 4 for numerical results and p-values).

Lastly, we observed a correlation coefficient 0.44 (p-value: 1.03×10^{-153}) between the retrieval results and the interactors' ratings. This stronger correlation compared with the work of (Chierici and Habash, 2021) underscores a higher agreement between the retrieved responses and human opinions in our setup.

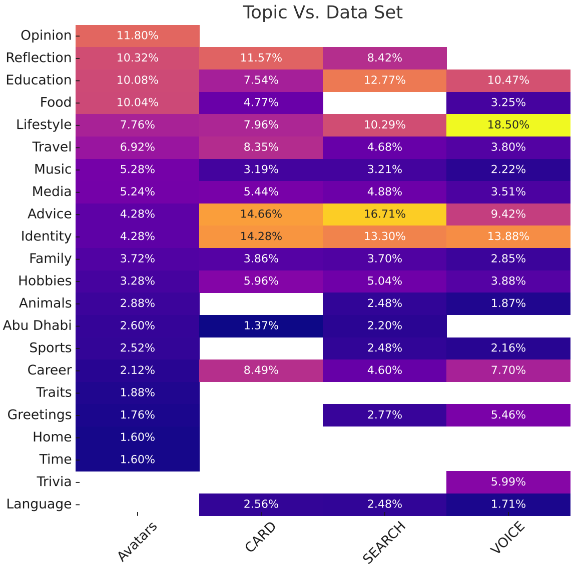


Figure 3: Heatmap of Topic Groups vs. Corpus Subset: The heatmap visualizes the distribution of questions across various topic groups ("Topic") and a subset of the HelloThere Corpus ("Data Set")—Avatars (the KBs of the recorded avatars), and (dialogue interactions by) CARD, SEARCH, and VOICE. The color intensity represents the proportion of questions, with brighter shades indicating higher proportions. Topics are ordered by higher coverage in the avatars' KBs.

4.5 What do People Ask?

We carried out topic clustering by leveraging the embeddings generated from GPT-3.5 Turbo. Specifically, we utilized the k-means clustering algorithm to group similar questions and tune the number of clusters until we identified recurring themes and could group them together sensibly. While we acknowledge this is a subjective labeling process, the clustering helped identify common themes across the avatars' KBs and the dialogues, providing insights into the types of questions present in the corpus. We describe the topics in Table 5 and map their occurrence in the corpus in Figure 3. The heatmap visualization allows us to identify and quantify the prevalence of different topic clusters across the corpus subsets. The color intensity represents the proportion of questions in each topic-subset combination, offering an intuitive view of user interests and avatar knowledge distribution. This visualization helps identify potential gaps in avatar knowledge bases (*Avatars* on the X-Axis) and areas of high user engagement, informing future improvements in TOIA system design.

The heatmap presents several key observations about how different topics fare across the HelloThere Corpus subsets. For instance, 'Identity' and 'Advice' are standout topics in the dialogues. The 'Lifestyle' topic is the most common in the VOICE channel, suggesting a focus on personal and day-to-day queries in voice-based (free-form) interactions. Interestingly, 'Education' and 'Reflection' topics are pretty evenly distributed across all modalities but VOICE and the avatars' KBs, signifying their universal appeal to users. Contrarily, the localized topic of 'Abu Dhabi' seems less prevalent than in previous sub-sets. Some topics, such as 'Home' and 'Time,' lag in user engagement across all sets. Furthermore, a newly added 'Trivia' category shows particular traction in the VOICE channel, hinting at various questions that don't necessarily slot into the existing categories. Lastly, it's worth noting that there are visible data gaps in topics like 'Opinion' and 'Traits,' which appear exclusively in the Avatars channel. This could signify a lack of user engagement for these topics in the dialogues.

5 Conclusion and Further Work

In this paper, we presented the HelloThere corpus, which includes two main categories of datasets: 26 single-turn knowledge bases and multi-turn dialogue corpora featuring annotated chat logs. To ensure consistency, we have standardized our terminology throughout, using "q-a pairs" to refer to question-answer pairs in the knowledge bases and dialogues. All q-a pairs are rated by Human interactors and benchmarked for answer retrieval.

The HelloThere Corpus offers a multifaceted resource for the SIGDial community. It is beneficial for benchmarking conversational agents, studying user behavior, and conducting multimodal analysis. It allows for focused studies on dialogue complexity, retrieval failures, and localized or general user interests, providing a comprehensive foundation for future research in natural language interactions.

The key future directions we plan to work on include: (a) expanding the corpus with more data to support diverse research applications; (b) refining models to enhance answer retrieval efficiency and engagement in multi-turn dialogues; and (c) providing and evaluating model performance under multilingual conditions.

References

Dana Abu Ali, Muaz Ahmad, Hayat Al Hassan, Paula Dozsa, Ming Hu, Jose Varias, and Nizar Habash. 2018. A bilingual interactive human avatar dialogue system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 241–244.

Ron Artstein, Anton Leuski, Heather Maio, Tomer Mor-Barak, Carla Gordon, and David Traum. 2015. How many utterances are needed to support time-offset interaction? In *The Twenty-Eighth International Flairs Conference*.

Alberto Chierici and Nizar Habash. 2021. A view from the crowd: Evaluation challenges for time-offset interaction applications. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 75–85.

Alberto Chierici and Nizar Habash. 2023. Tell me more, tell me more: Ai-generated question suggestions for the creation of interactive video recordings. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1725–1730. IEEE.

Alberto Chierici, Nizar Habash, and Margarita Bicec. 2020. The margarita dialogue corpus: A data set for time-offset interactions and unstructured dialogue systems. In *Proc. of Language Resources and Evaluation Conference*.

Alberto Chierici, Tyeece Kiana Fredorcia Hensley, Wahib Kamran, Kertu Koss, Armaan Agrawal, Erin Meekhof, Goffredo Puccetti, and Nizar Habash. 2021. 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.

Alberto Maria Chierici. 2023. *Scalable, Human-Like Asynchronous Communication*. Ph.D. thesis, New York University Tandon School of Engineering.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. 2020. Unsupervised faq retrieval with question generation and bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, et al. 2015. New Dimensions in Testimony: Digitally preserving a Holocaust survivor'??s interactive storytelling. In *Proceedings of the International Conference on Interactive Digital Storytelling*, pages 269–281.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Asaf Yehudai, Matan Vetzler, Yosi Mass, Koren Lazar, Doron Cohen, and Boaz Carmeli. 2023. Qaid: Question answering inspired few-shot intent detection. *arXiv preprint arXiv:2303.01593*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.