information

Article

InterviewBot: Real-Time End-to-End Dialogue System for Interviewing Students for College Admission

Zihao Wang [1,*], Nathan Keyes [2], Terry Crawford [2] and Jinho D. Choi [1,*]

[1] Department of Computer Science, Emory University, Atlanta, GA 30322, USA
[2] InitialView, Atlanta, GA 30322, USA; nathan@initialview.com (N.K.); terry@initialview.com (T.C.)
* Correspondence: zihao.wang2@emory.edu (Z.W.); jinho.choi@emory.edu (J.D.C.)

Abstract: We present the InterviewBot, which dynamically integrates conversation history and customized topics into a coherent embedding space to conduct 10 min hybrid-domain (open and closed) conversations with foreign students applying to U.S. colleges to assess their academic and cultural readiness. To build a neural-based end-to-end dialogue model, 7361 audio recordings of human-to-human interviews are automatically transcribed, where 440 are manually corrected for finetuning and evaluation. To overcome the input/output size limit of a transformer-based encoder–decoder model, two new methods are proposed, context attention and topic storing, allowing the model to make relevant and consistent interactions. Our final model is tested both statistically by comparing its responses to the interview data and dynamically by inviting professional interviewers and various students to interact with it in real-time, finding it highly satisfactory in fluency and context awareness.

Keywords: interview chatbot; conversational AI; dialogue system

check for updates

Citation: Wang, Z.; Keyes, N.; Crawford, T.; Choi, J.D. InterviewBot: Real-Time End-to-End Dialogue System for Interviewing Students for College Admission. *Information* **2023**, *14*, 460. https://doi.org/10.3390/info14080460

Academic Editor: David Martins de Matos

Received: 6 July 2023
Revised: 31 July 2023
Accepted: 8 August 2023
Published: 15 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the latest advancement of conversational AI, end-to-end dialogue systems have been extensively studied [1–3]. One critical requirement is context awareness: robust dialogue systems must consider relevant parts in conversation history to generate pertinent responses [4–8]. However, these systems still suffer from issues such as hallucination, inconsistency, or lacking common sense [9], hindering them from taking place in real applications.

Numerous admission interviews are given every year to students located in 100+ countries applying to colleges in the U.S., where the interviews are often conducted online. Those interviews are usually unscripted, with an emphasis on asking the applicants thought-provoking questions based on their interests and experiences. The main objective is to provide decision-makers (e.g., admissions officers, faculty members) with an unfiltered look at those students in a daily academic environment.

Building an interview chatbot, called InterviewBot, will save time and effort for the interviewers and provide foreign students with a cost-efficient way of practising interviews when native speakers are unavailable. Nonetheless, there are a few hurdles to developing an end-to-end InterviewBot. First, it is hard to collect a sufficient amount of data covering dialogues crossing open and closed domains (Section 3.1). Second, most transformer-based encoder–decoder models adapted by current state-of-the-art systems are not designed to handle long contexts; thus, they often repeat or forget previously discussed topics (Section 3.3). Third, it is demanding to find appropriate people to interactively test such a dialogue system with a professional objective (Section 4).

This paper presents an end-to-end dialogue system that interacts with international applicants to U.S. colleges. The system questions critical perspectives, follows up on the interviewee's responses for in-depth discussions, and makes natural transitions from one topic to another until the interview ends, which lasts about 30 turns (5 min for text-based, 10 min for spoken dialogues). To the best of our knowledge, it is the first real-time system using a neural model, completely unscripted, conducting such long conversations for admission interviews. Our technical contributions are summarized as follows:

- We have developed a contextualized neural model designed to perform diarization tasks on text transcripts alone.
- We have integrated a sliding window technique to overcome the input token limit and restore the completeness of the input in the latent space.
- We have integrated extracted topics from the conversation to address issues related to topic repetition, off-topic discussions, and premature endings in conversations.

The remaining sections are organized as follows: Section 2 reviews current dialogue models, their applications, and limitations. Section 3 describes datasets and our speaker diarization and InterviewBot model architectures in detail. Section 4 gives experiment results on diarization and InterviewBot dialogue generation. Sections 5 and 6 conduct discussions on the results and conclude the paper.

2. Related Work

Dialogue systems can be categorized into closed- and open-domain systems [10]. Closed-domain systems require efficient access to domain knowledge [11] and serve specific professions, such as education [12], healthcare [13,14], or customer service [15,16]. Open-domain systems converse across multiple domains with natural transitions [2] and conduct interactions in a broader horizon [17–19]. For admission interviews, however, the conversation is often a mixture of closed (job-related questions) and open-domain (general aspects of the applicant) dialogues, which makes it more challenging to build an end-to-end system.

Several dialogue systems have been developed to communicate with humans for information exchange or elicitation across multiple domains [20–22]. Ref. [19] built a conversational system to converse proactively on popular topics with Alexa users by providing them with the requested information as well as pre-crafted transitions. Ref. [23] established a virtual interviewer to study the effect of personality on confiding and listening to virtual agents. Ref. [24] studied the role of a chatbot in a survey setup. Although these dialogue systems have shown their effectiveness in achieving their goals, they all heavily rely on design templates. Conversational agents for interviews have been experimented with for law enforcement [25], healthcare [26], job application [27], and psychology [28], among which most are proof of concept. A few interview bots have been developed on commercial platforms such as Google Dialogflow and IBM Watson Assistant, with the limitation of pre-scripted interviews; thus, they cannot proactively follow up on the user content.

Context and memory have been studied as key factors that affect model performance in context-heavy settings. Ref. [29] proposed a memory transformer to hierarchically employ memory to improve translation performance. However, in a more complex conversation setup, dialogue flow is not only about correlations between sentences or words in the semantics but rather the proceeding of conversations with a depth of topics and transitions to other topics. Other models, such as Refs. [30,31], have proposed context and external knowledge-based models in conversation-related tasks. Although the effort was proven to improve based on specific metrics, it was still not sufficient to improve the overall dialogue flow of conversations.

Deep language models, such as Blenderbot [3] and Bart [32], have taken context into consideration. However, the limitation on the length of input tokens as well as conversation history has bottlenecked their applications in the real world. Recent surges of large language models, such as ChatGPT [33] and LLaMa [34], have shown strong evidence of improvement with respect to context integration. Nevertheless, there are always limitations on the input length, as well as effective ways of integrating different contexts into a language model.

3. Materials and Methods

3.1. Interview Dataset

Audio recordings of 7361 interviews were automatically transcribed with speaker identification by the online tool RevAI (https://www.rev.ai, accessed on 7 August 2023), where 440 are manually corrected on speaker ID assignment for finetuning and evaluation of our models (Table 1). Each recording contains an average of an ≈15 min long dialogue between an interviewer and an interviewee. The interviews were conducted by 67 professionals in 2018–2022. The largest age group of interviewees is 18-years-old with 59.3%, followed by 17-years-old with 29.4%. The male-to-female ratio is 1.2:1. The major country of origin is China with 81.4% followed by Belgium with 10.5%, alongside 37 other countries. Table 1 provides detailed demographics of the interviewees.

All recordings were transcribed into text and speakers were identified automatically. For speech recognition, three tools from Amazon (https://aws.amazon.com/transcribe, accessed on 7 August 2023), Google (https://cloud.google.com/speech-to-text, accessed on 7 August 2023), and RevAI (https://www.rev.ai, accessed on 7 August 2023) were assessed on 5 recordings for speaker diarization, achieving the F1-scores of 66.3%, 50.1%, and 72.7%, respectively.

Table 1. Distributions of our data. D: num of dialogues, U: avg-num of utterances per dialogue, S1/S2: avg-num of tokens per utterance by interviewer/interviewee. `TRN`/`DEV`/`TST`: training/development/evaluation (annotated) sets. `RAW`: unannotated set (auto-transcribed).

	D	**U**	**S1**	**S2**
`TRN`	140	43.8	39.3	64.0
`DEV`	150	45.0	36.2	60.3
`TST`	150	44.3	37.8	61.3
`RAW`	6921	40.4	41.5	67.6

Figure 1 shows the distribution of the ages of applicants. Most interviewees are between 17 and 19, which is an accurate reflection of the ages of high school students applying to colleges. Figure 2 shows the distribution of the applicants' countries of origin. There are 38 countries in total. The majority of applicants come from China. Other major countries are Belgium, Bangladesh, Canada, India, and Belarus. The gender distribution of applicants is shown in Figure 3. The numbers of male and female applicants are close, with the exclusion of applicants not providing gender information.

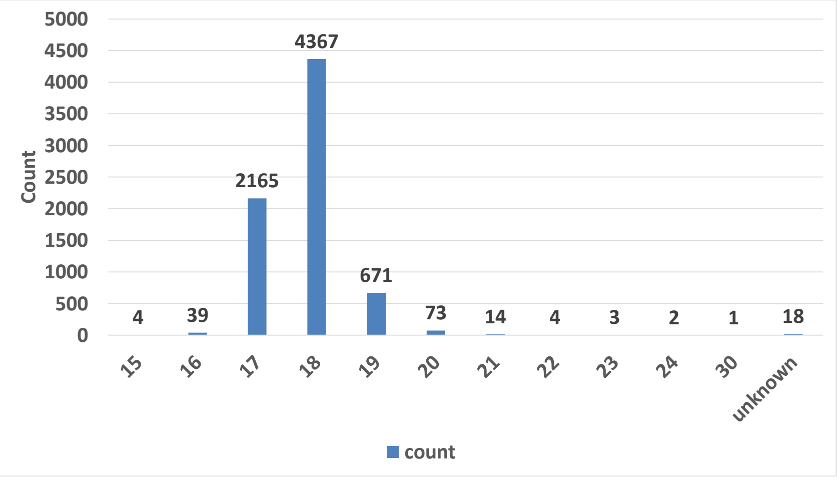

Figure 1. The interviewees' age demographics.

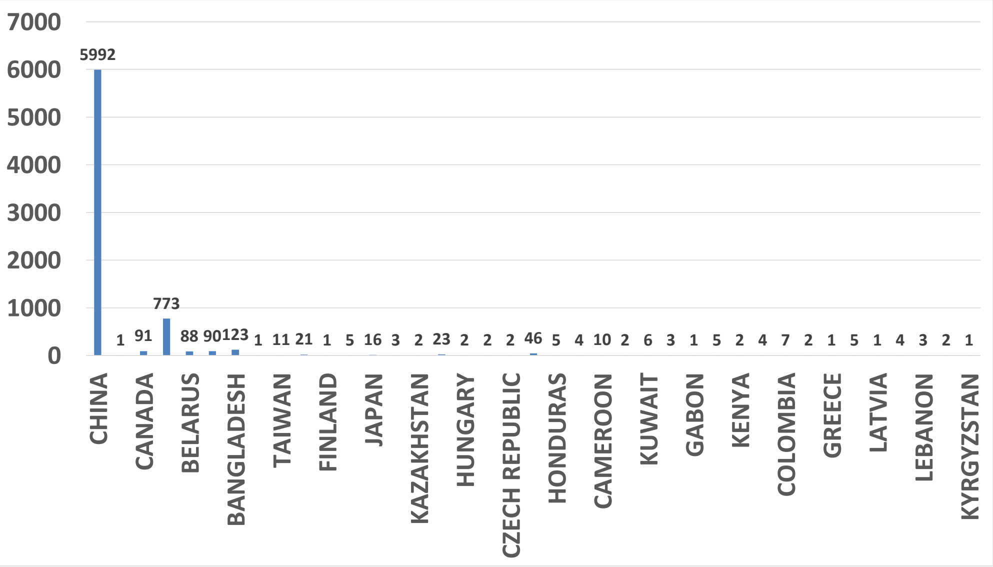

Figure 2. The interviewees' country demographics.

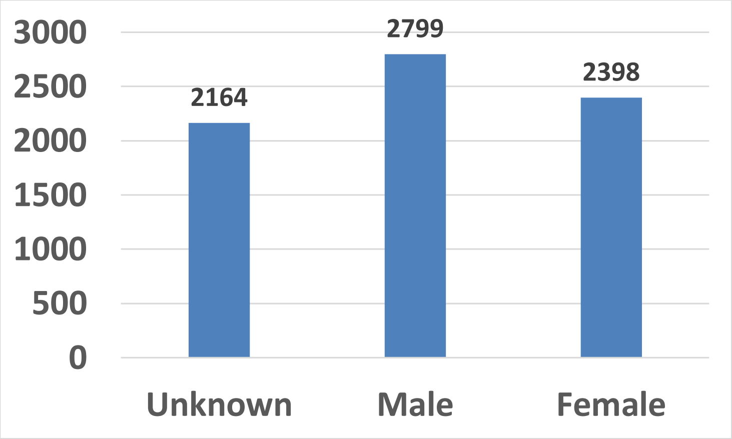

Figure 3. The interviewees' gender demographics.

3.2. Speaker Diarization

Speaker diarization is the task of segmenting an audio stream into utterances according to the speaker's identity and is considered critical in automatic transcription [35]. Conversation data with diarization errors can lead to a major failure in building robust dialogue models. Our most accurate transcriber, RevAI, still gives 27.3% errors for speaker diarization (Section 3.1). The main reason is that audios from the interviewer (`S1`) and the interviewee (`S2`) are recorded in one channel, so that they are saved in a single waveform, while no clear pauses exist between `S1` and `S2`'s speeches or their speeches often overlap. The following example illustrates when the speech of `S2` (underlined) is not recognized as a separate utterance:

`S1`: Hi, it 's nice to meet you. Nice to meet you.
`S2`: Um, can you tell me what is a topic that um, you cannot stop talking about?

Thus, speaker diarization models are developed to provide clean data to our dialogue model (Section 3.3). Figure 4 depicts the distributions of different types of diarization errors found in 100 dialogues. Most errors are caused by filler words and arbitrary concatenation (joining multiple utterances as one with no apparent patterns, not caused by filler words).

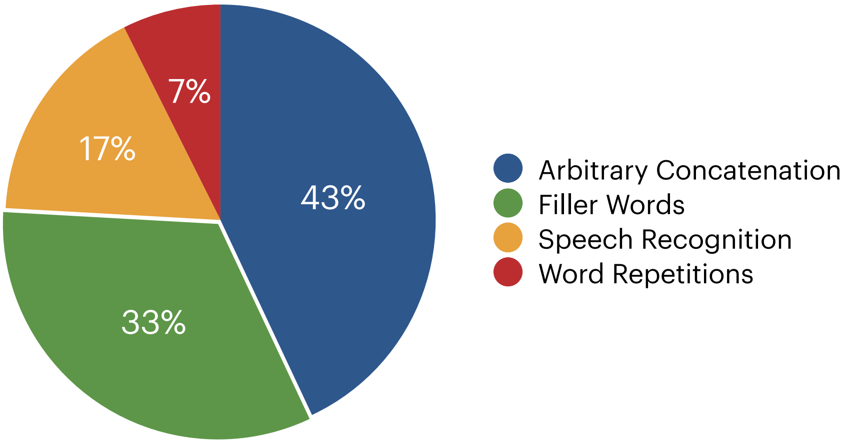

Figure 4. Distributions of the diarization error types. Section 5.2 provides examples of each error type.

3.2.1. Manual Annotation

A total of 440 dialogues were sampled, in which every token is annotated either `1` if it is one of the last two tokens of an utterance before the speaker is switched, and `0` otherwise. For the above example, the 8–9th tokens are the last two tokens of the utterance before it switches to `S2` and so are the 13–14th tokens before switching to `S1`; thus, they are annotated `1` (We also annotated only the last token as `1`, or annotated all words from `S0` as `0` and from `S1` as `1`, which yielded worse results in terms of the end performance):

Hi	,	it	's	nice	to	meet	you	.	Nice	to	meet	you	.
`0`	`0`	`0`	`0`	`0`	`0`	`0`	`1`	`1`	`0`	`0`	`0`	`1`	`1`

Doccano was used as the annotation tool [36], and ELIT was used for the tokenization [37]. To measure the inter-annotator agreement, ten dialogues were double-annotated that show a high kappa score of 84.4%.

3.2.2. Pseudo Annotation

Because our annotated data are relatively small, a larger dataset was pseudo-created for this task using 2400 dialogues in the Switchboard [38] and 6808 dialogues in the BlendedSkillTalk [39] datasets (thus, a total of 9208 dialogues). These two datasets were chosen because their dialogues sound more speech-originated than others, having an adequate amount of filler words. Among the four types of diarization errors (Figure 4), the ones caused by filler words (33%) can be simulated on dialogues that do not contain such errors using statistical heuristics (filler words are inferred by the outputs of the part-of-speech tagger and the dependency parser in ELIT).

The errors associated with filler words were pseudo-inserted into dialogues from the two datasets by finding an utterance either beginning or ending with a filler word and concatenating it with an utterance before or next to it. Global search was made to the entire dialogues for finding such utterances to mimic the distributions in Table 2 such that about 40.4% of the dialogues in the pseudo-created data would contain two utterances with diarization errors, where 46.7% of them are caused by the filler word *okay*, and so on. It is possible that more than two utterances get joined; in our case, up to eight utterances were concatenated. Table 3 includes the statistics of our pseudo-created dataset for transfer learning.

Table 2. Distributions of filler words with regard to diarization errors. Dist: percentage of dialogues containing # number of utterances with errors caused by the filler words. *filler_word*: percentage of the filler word appearing in the corresponding dialogue group.

#	**Dist**	***okay***	***yeah***	***right***	***um***	***so***	***uh***	***well***	***like***	***oh***
2	40.4	46.7	16.0	8.0	8.5	8.0	4.4	4.7	0.2	3.1
3	35.9	33.3	29.8	3.9	8.5	11.1	6.2	2.3	0.6	4.1
4	8.6	33.7	24.5	5.6	9.7	11.2	5.1	3.6	1.0	5.1
5	7.3	28.9	30.7	6.6	4.2	15.1	6.6	4.8	0.6	2.4

Table 3. Distributions of the pseudo-created datasets (Switchboard, BST) and our interview data (before and after diarization). D: number of dialogues, U: avg-number of utterances, S1/S2: avg-number of tokens per utterance by `S1`/`S2`. `TRN`/`DEV`/`TST`: training/development/evaluation (annotated) sets. `RAW`: unannotated set. Note that we follow the same splits suggested by the original papers of the Switchboard and BST datasets for comparability.

	Switchboard				**BlendedSkillTalk (BST)**				**Interview Data (Before)**				**(After)**		
	D	**U**	**S1**	**S2**	**D**	**U**	**S1**	**S2**	**D**	**U**	**S1**	**S2**	**U**	**S1**	**S2**
`TRN`	1115	42.2	31.6	31.3	4819	6.2	25.0	25.2	140	42.2	37.4	73.1	43.8	39.3	64.0
`DEV`	21	16.6	49.4	43.7	1009	6.1	25.8	25.3	150	44.1	34.9	67.3	45.0	36.2	60.3
`TST`	19	32.7	32.9	32.9	980	6.2	26.2	26.2	150	44.2	34.2	69.0	44.3	37.8	61.3
`RAW`	N/A				N/A				6921	39.6	38.2	75.1	40.4	41.5	67.6

3.2.3. Joint Model

The joint model consists of two parts. First, we established a binary classification task that enforces the model to learn to differentiate utterances that have diarization errors. The second part is a diarization model for tackling the problem specifically. The intention behind this design is that the binary classification task could enhance the embedding representation on a higher level to perform the diarization task better.

Figure 5 shows an overview of our speaker diarization model. Let $U_i = \{w_i^\circ, w_{i1}, \ldots, w_{in}\}$ be the i'th utterance to be handled, where w_i° is the special token representing U_i and w_{ij} is the j'th token in U_i. U_i is fed into the encoder $\mathcal{E}$ that generates the embeddings $\{e_i^\circ, e_{i1}, \ldots, e_{in}\}$. The previous utterances $\{U_{i-k}, \ldots, U_{i-1}\}$ are also fed into $\mathcal{E}$, which generates $\{e_{i-k}^\circ, \ldots, e_{i-1}^\circ\}$ (in our case, $k = 5$ is the context window). These embeddings are fed into a transformer layer for utterance-level weighting, which creates the context embedding e_c. Finally, $e_c \oplus e_i^\circ$ is fed into a softmax layer that outputs o_u to make a binary decision of whether or not U_i includes any error. Jointly, each $e_c \oplus e_{ij}$ is fed into another softmax that outputs o_j to decide whether or not w_{ij} is one of the last two tokens of an utterance.

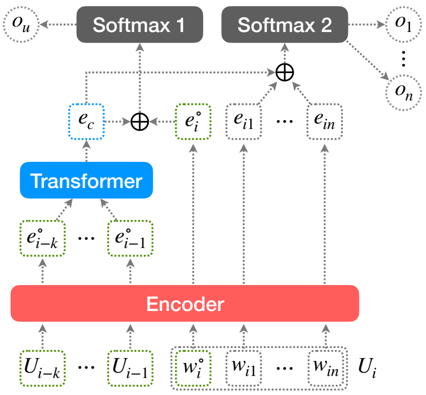

Figure 5. The overview of our diarization model.

3.3. Dialogue Generation

Figure 6 depicts an overview of our dialogue generation model. Since inputs to the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ are limited by the total number of tokens that the pretrained language model accepts, *sliding window* (Section 3.3.1) and *context attention* (Section 3.3.2) are proposed to handle long utterances and contexts in the previous utterances, respectively. In addition, *topic storing* is used to remember user-oriented topics brought up during the interview (Section 3.3.3). The input to $\mathcal{E}$ and output of $\mathcal{D}$ include the speaker ID `S1`, `S2` or special tokens to indicate the beginning `B`, the ending `E`, and topic questions `Q` as the first token followed by an utterance from the interviewer or interviewee, respectively. Hyperparameters were finetuned by cross-validations.

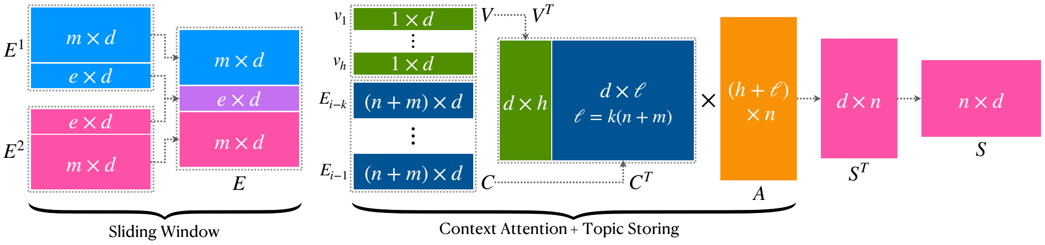

Figure 6. The overview of our dialogue generation model.

3.3.1. Sliding Window

The sliding window technique aims to overcome the limitation of input length by separating a long sentence into multiple sections within. The mathematical formulations are described below. Let $n = m + e$ be the max-number of tokens that $\mathcal{E}$ and $\mathcal{D}$ accept ($e < m < n$). Every utterance U whose length is greater than n is split into U^1 and U^2 as follows (w_i is the i'th token in U):

$$\begin{aligned} U^1 &= \{w_1, \ldots, w_m, w_{m+1}, \ldots, w_n\} \\ U^2 &= \{w_{m+1}, \ldots, w_n, w_{n+1}, \ldots, w_{n+m}\} \end{aligned} \tag{1}$$

In our case, n = 128, m = 100, and e = 28 such that $n + m = 228$ is sufficiently long enough to handle most utterances based on our stats. $\mathcal{E}$ takes U^1 and U^2 then produces $E^1 = \{e^1_1, \ldots, e^1_n\}$ and $E^2 = \{e^2_{m+1}, \ldots, e^2_{n+m}\}$, where $e^*_i \in \mathbb{R}^{1 \times d}$ is the embedding of w_i. Finally, the embedding matrix $E \in \mathbb{R}^{(n+m) \times d}$ of U is created by stacking all of the following embeddings:

$$\{e^1_1, \ldots, \frac{1}{2}\sum_{i=1}^{2}(e^i_{m+1}), \ldots, \frac{1}{2}\sum_{i=1}^{2}(e^i_n), \ldots, e^2_{n+m}\}$$

For utterances whose lengths are less than or equal to n, zero-padding is used to transform $\mathcal{E}$'s output from $\mathbb{R}^{n \times d}$ to $\mathbb{R}^{(n+m) \times d}$.

3.3.2. Context Attention

Let U_i be the i'th utterance to be generated as output. Let $C \in \mathbb{R}^{\ell \times d}$ be the context matrix stacking the embedding matrices of the previous utterances $\{E_{i-k}, \ldots, E_{i-1}\}$, where k is the number of previous utterances to be considered and $\ell = k(n+m)$. The transpose of C is multiplied by the attention matrix $A \in \mathbb{R}^{\ell \times n}$ such that $C^T \cdot A \rightarrow S^T \in \mathbb{R}^{d \times n}$. Thus, $S \in \mathbb{R}^{n \times d}$ represents the context summary of $U_{i-k}, \ldots, U_{i-1}$, which is fed into the decoder $\mathcal{D}$.

3.3.3. Topic Storing

Even with the context attention, the model still has no memory of contexts prior to U_{i-k}, leading it to repeat the same topics that it has already initiated. To overcome this issue, topic storage is introduced to remember key topics derived by the interviewer. Every interview in our data came with 8–16 questions by the interviewer annotated after each interview by the data provider, who used those questions during the interview and thought that they led to assessing crucial aspects of the interviewee. Our final model considers these questions the "key topics" and dynamically stores them as the dialogue progresses. During training, these questions are converted into embeddings and stored dynamically as a list of topics discussed in previous turns. During decoding, the model generates such topical questions with a specific flag and stores them in the same way.

Let $Q = \{q_1, \ldots, q_h\}$ be the topical question set. During training, $\mathcal{D}$ learns to generate `Q` instead of `S1` as the first token of the interviewer's utterance that contains any $q_i \in Q$. In addition, it generates `B/E` if the interviewer begins/ends the current dialogue with that utterance (Table 4). Any utterance starting with `Q` is encoded by $\mathcal{E}$ and feed-forward layers that create abstract utterance embedding $v_i \in \mathbb{R}^{1\times d}$ to represent topics. These embeddings get stacked as the interview goes on to create the topic matrix $V \in \mathbb{R}^{h\times d}$. If $|Q| < h$, then zero-padding is used to create V (in our case, $h = 16$). Finally, V is stacked with the context matrix C (Section 3.3.2), and $(V \oplus C)^T \in \mathbb{R}^{d\times(h+\ell)}$ is multiplied by the attention matrix $A \in \mathbb{R}^{(h+\ell)\times n}$ to create the transpose of the context summary matrix $S \in \mathbb{R}^{n\times d}$.

Table 4. An interview dialogue conducted by our best model (CT in Section 4). `S1/S2`: interviewer/interviewee (chatbot/human), `B/E`: beginning/ending utterance (chatbot), `Q`: topical question (chatbot).

Flag	Utterance
`B`	Hello, and thank you so much for coming today. uh, Let us start by having you just to state your full legal name and also the school that you are currently attending?
`S2`	Sure, my name is David, and I go to Shanghai high school in China.
`Q`	Great. um, yeah, let's talk about leadership experiences. Can you give me some specific examples?
`S2`	Sure, when I was a junior, I was elected to be the Vice President in the Student Union and I organized many activities to help students to enrich their life on campus.
`S1`	Okay, great. Thank you so much for sharing that. so, um, tell me a little bit about your leadership style. What do you think you bring to the student union?
`S2`	Sure, I am very open-minded, and with some of my colleagues, we together established a lot of new programs on top of our existing ones. Many of those programs became very popular.
`S1`	Ok, great. and so, um, what, uh, what would you say has been the most challenging aspect of being a leader?
…	
`Q`	Great. uh, um, hmm, let' talk about your courses. What courses did you take in high school and which courses do you like most?
`S2`	Sure, I took Math, English, PE, Chemistry, Art, and so on. I like Math most.
`Q`	Okay. so, um, tell me a little bit about your, your interest in mathematics.
…	
`E`	Okay, great. well, thank you very much for sharing that with me. um, and thank you for taking the time out of your day to interview with us.

4. Results

4.1. Speaker Diarization Results

Table 3 shows the distributions of the pseudo-created data (Section 3.2.2), as well as our interview data (Section 3.1) before and after the diarization where errors in the train/dev/test sets are manually annotated (Section 3.2.1) and errors in the raw set are automatically corrected by the joint model (Section 3.2.3). For the encoder, the RoBERTa large model is used [40] (several transformer encoders including BERT [41] were evaluated and RoBERTa yielded the best results). After diarization, `S2`'s utterances with diarization errors get split such that the average length of `S2`'s utterances decreases while the average length of dialogues slightly increases. Meanwhile, some parts of `S2`'s utterances, incorrectly separated from `S1`'s utterances by the transcriber, are recovered back to `S1`; thus, the average length of `S1`'s utterances increases.

Table 5 shows results of three models: the *baseline* model taking U_i and producing $O_w = \{o_1, \ldots, o_n\}$, the *context* model taking $\mathcal{U}_c = \{U_{i-k}, \ldots, U_i\}$ and producing O_u, and the *joint* model taking $\mathcal{U}_c$ and producing O_u and O_w (Figure 5). The baseline model does not create e_c, so e_{i*} is directly fed to Softmax 2. Also, the baseline and context models do not use e_i°, so only Softmax 2 is used to produce the outputs. For evaluation, the F1-scores of the label `1` on the last two tokens are used. All models are developed three times and their average scores and standard deviations are reported.

Table 5. Diarization model performance. Ours: trained on `TRN` of our interview data (after) in Table 3. Transferred: trained first on the `TRN` mixture of Switchboard and BST, then finetuned on `TRN` of our data.

	Baseline	**Context**	**Joint**
Ours only	92.9 ± 0.4	92.9 ± 0.3	92.9 ± 0.2
Transferred	93.2 ± 0.3	93.4 ± 0.3	**93.6 ± 0.3**

When trained on only our data, all models perform similarly. The joint model slightly outperforms the others when transfer learning is applied. Although the improvement is marginal, the joint model has the benefit of identifying utterances with diarization errors, showing an F1-score of 93.6% for this task, while the transferred models generally show a much higher performance on the other datasets than the non-transferred models. Thus, the joint transferred model is used to auto-correct all dialogues in `RAW`.

4.2. Dialogue Generation Results

For our experiments, on the diarized data from the diarization model, the encoder and the decoder in BlenderBot 1.0 [3] are used. (There have been updated versions of BlenderBot introduced [8,42]. However, we chose the first version for our experiments because we found it to be as effective yet much more efficient than the newer versions since the newer models focus on improvement on different perspectives, such as privacy and external knowledge incorporation.) Three models are developed as follows:

- BB: Blenderbot Baseline Model;
- SW: Blenderbot with Sliding Window;
- CT: Blenderbot with Sliding Window and Concatenation of Topic Storing.

All models are first trained on `raw` and finetuned on `TRN` in Table 1). We followed the setups on the training parameters in the original Blenderbot paper. To assess real-life performance, ten interviews are conducted per model, where each interview consists of exactly 30 turns. Qualitative analysis is performed on the top three most frequently occurring errors as follows:

- Repetitions: how often it repeats topics already covered in the previous utterances.
- Early ending (EE): implies ending the interview without covering a sufficient amount of topics.
- Off topic (OT): how often it makes utterances that are not relevant to the current topic.

Table 6 shows the error analysis results. The repetition rates are significantly reduced as the model gets more advanced. Compared to the baseline, the CT model conducts 3.5 times longer conversations before it attempts to end the interview while generating twice fewer off-topic utterances, which is very promising. Examples of these error types are provided in Section 5.3.

Table 6. The error analysis of all generation models. R: avg-% of repeated topics, EE: avg-% of the interview conducted before the model attempts to end (higher is better), OT: avg-% of off-topic utterances.

Model	Repetitions (R)	Early Ending (EE)	Off Topic (OT)
BB	30.0	13.3	20.0
SW	16.7	23.3	26.7
CT	**6.7**	**46.7**	**10.0**

4.2.1. Static Evaluation

Following previous work [43], static evaluation is performed on the CT model, where the input is every batch of k-utterances and prior topics per interview, and its output is compared to the corresponding human response in `TST` (Table 1). The average BLEU score is 0.08 and cosine similarity is 0.19, which are low. However, such static evaluation assesses each output independently and obstructs dialogue fluency by artificially inserting human utterances into the model, and thus does not reveal its capability in conducting long contextualized interviews.

4.2.2. Real-Time Evaluation

The CT model is deployed to an online text-based platform in a public cloud. For real-time evaluation, five professional interviewers and ten students are invited to have conversations with our InterviewBot and give ratings from 1 to 5 to indicate their overall satisfaction. The average dialogue duration is 256 s. Almost half of the evaluators are satisfied (Scores 4 and 5) and another 40% indicate a positive attitude on the coverage of topics and discussions (Score 3), implying that it performs reasonably well for this realistic setting (Table 7). Overall, with an average score of 3.5, the InterviewBot has shown great potential in applying to practical applications.

Table 7. The rating distribution of the InterviewBot conversations for real-time evaluation. 5: very satisfied, 4: satisfied, 3: neutral, 2: unsatisfied, 1: very unsatisfied.

Score	5	4	3	2	1	Average Score
Interviewer (Count)	1	1	2	1	0	3.4
Student (Count)	2	3	4	0	1	3.5
Total (Count)	3	4	6	1	1	3.5

5. Discussion

5.1. Chatbot Demonstration

Table 4 presents an example dialogue conducted by our chatbot showcasing the utilization of sliding window and topic storing (CT) techniques. Overall, the chatbot demonstrates its ability to conduct a comprehensive interview by asking relevant follow-up questions, adapting to various conversation topics, and providing meaningful responses accordingly.

5.2. Examples of Diarization Errors

The following are examples used to illustrate the sources of diarization errors (underlined). In many cases, interviewers and interviewees overlap in speeches or thinking out loud with or without filler words, which concatenates the two utterances. A small portion of diarization errors are from speech recognition and word repetition errors.

- **Arbitrary Concatenation**
  What do you think the benefits might be of this kind of technology? If we develop it, I think this technology will eventually replace, um, human delivery.

- **Filler Words**
  Oh, no, I'm going to make majoring mathematics. Okay. Okay. Now why, why do you think receiving an education is important?
- **Speech Recognition**
  Um, okay. My name is <inaudible>. I'm a senior year student come from Hunger-Free. Which high school are you from?
- **Word Repetition**
  I heard it said, so it's kind of like a DIY community community. Are there community activities?

We conduct speaker diarization errors by annotating 100 conversations with an average of 39 turns. The type of errors with their stats are shown in Table 8. The major errors are caused by filler words and arbitrary concatenation. Counts of single errors are counted. Table 9 lists examples underlined for each type of error, respectively.

From the table, there are different occasions for ASR errors, such as words not transcribed, words transcribed incorrectly, etc. Word repetitions happen rarely when non-filler words repeat in a sentence. Concatenations of two adjacent utterances take the largest portion of all errors.

Table 8. Speaker diarization error analysis in the text transcription from interview recordings. Notions are Misc: miscellaneous; ASR: auto speech recognition; WR: word repetition; FW: filler words; AC: adjacent concatenation.

Count	Type		
212	Misc	ASR: 147	WR: 65
289	FW		
378	AC		
# of Convs	100		
Avg # of turns	39		

Table 9. Error type examples.

Error Type	Examples
`ASR`	Um, okay. My name is <inaudible>. I'm a senior year student come from Hunger-Free language school.
`WR`	I heard it said, so it's kind of like a DIYcommunity community.
`FW`	Oh, no, I'm going to make majoring mathematics. Okay. Okay. Now why, why do you think receiving an education is important?
`AC`	What do you think the benefits might be of this kind of technology? If we develop it, I think this technology will eventually replace, um, human delivery.

5.3. Examples of Generation Limitations

The following are examples for each type of generation error (underlined). An early ending prevents the InterviewBot from covering critical aspects of applicants and further discussion. The InterviewBot cannot name mentions properly, either skipping over or generating the wrong entities in the conversations. Random generation generates random words and punctuation and concatenates them incorrectly. Not following up on previous topics hinders the InterviewBot from becoming proactively engaged with interviewees.

- **Early Ending**
  InterviewBot: Could you please state your name?
  Interviewee: Sure. My name is David.
  ... 4 more turns ...
  InterviewBot: Thank you and best wishes to your applications.

- **Repetition** InterviewBot: Now why, why do you think receiving an education is important?
  ... more turns ...
  InterviewBot: Why do you think people value education so much?
- **Off Topic** Interviewee: Sure, I really like Mathemathcs.
  InterviewBot: What do you do for fun?

6. Conclusions

Our InterviewBot is a model-based dialogue system equipped with contextual awareness and topic sensitivity that conducts college admission interviews. Questions covering diverse topics and discussions in extended follow-ups are carried along the conversations, which have been assessed by professional interviewers and student volunteers. The average satisfaction score of 3.5 projects the prevailing deployment of the InterviewBot for thousands of college applicants, especially for international students.

However, though with promising future applications, the current version of the InterviewBot has two major limitations. First, the early ending in Table 6 still happens, where an ending utterance gets generated after an insufficient amount of turns, in which case the interview may not cover critical aspects of the applicant. Second, the bot makes good follow-ups to various topics; however, it needs to derive deeper discussions with more details.

In future work, the main focus is to enrich the follow-up discussions on topics or sub-topics during interactions by training the InterviewBot with more structured topic-flow materials. This task would indirectly alleviate the early ending issue by deepening the discussions on certain topics.

Author Contributions: Conceptualization, Z.W. and J.D.C.; methodology, Z.W. and J.D.C.; software: Z.W.; formal analysis, Z.W. and J.D.C.; investigation, Z.W. and N.K.; resources, Z.W., N.K. and T.C.; data curation, Z.W.; writing—original draft preparation, Z.W. and J.D.C.; supervision, J.D.C.; project administration, J.D.C.; funding acquisition, J.D.C. and T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by InitialView.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our annotatedcorpus and annotation guidelines are publicly available under the Apache 2.0 license on https://github.com/emorynlp/MLCG accessed on 7 August 2023.

Acknowledgments: We gratefully acknowledge the support of the InitialView grant.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; pp. 270–278. [CrossRef]
2. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a human-like open-domain chatbot. *arXiv* **2020**, arXiv:2001.09977.
3. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.L.; et al. Recipes for Building an Open-Domain Chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 July 2021; pp. 300–325. [CrossRef]
4. Serban, I.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; Bengio, Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.

5. Mehri, S.; Razumovskaia, E.; Zhao, T.; Eskenazi, M. Pretraining methods for dialog context representation learning. *arXiv* **2019**, arXiv:1906.00414.
6. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 85–96. [CrossRef]
7. Zhou, H.; Ke, P.; Zhang, Z.; Gu, Y.; Zheng, Y.; Zheng, C.; Wang, Y.; Wu, C.H.; Sun, H.; Yang, X.; et al. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv* **2021**, arXiv:2108.01547.
8. Xu, J.; Szlam, A.; Weston, J. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 5180–5197. [CrossRef]
9. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Wu, Z.; Guo, Z.; Lu, H.; Huang, X.; et al. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv* **2021**, arXiv:2109.09519.
10. Ilievski, V.; Musat, C.; Hossmann, A.; Baeriswyl, M. Goal-oriented chatbot dialog management bootstrapping with transfer learning. *arXiv* **2018**, arXiv:1802.00500.
11. Lian, R.; Xie, M.; Wang, F.; Peng, J.; Wu, H. Learning to select knowledge for response generation in dialog systems. *arXiv* **2019**, arXiv:1902.04911.
12. Cunningham-Nelson, S.; Boles, W.; Trouton, L.; Margerison, E. A review of chatbots in education: Practical steps forward. In Proceedings of the 30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators Becoming Agents of Change: Innovate, Integrate, Motivate, Brisbane, Australia, 8–11 December 2019; pp. 299–306.
13. Fan, X.; Chao, D.; Zhang, Z.; Wang, D.; Li, X.; Tian, F. Utilization of self-diagnosis health chatbots in real-world settings: Case study. *J. Med. Internet Res.* **2021**, *23*, e19928. [CrossRef]
14. Amiri, P.; Karahanna, E. Chatbot use cases in the Covid-19 public health response. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1000–1010. [CrossRef]
15. Baier, D.; Rese, A.; Röglinger, M.; Baier, D.; Rese, A.; Röglinger, M. Conversational User Interfaces for Online Shops? A Categorization of Use Cases. In Proceedings of the International Conference on Information Systems, Libertad City, Ecuador, 10–12 January 2018.
16. Nichifor, E.; Trifan, A.; Nechifor, E.M. Artificial intelligence in electronic commerce: Basic chatbots and the consumer journey. *Amfiteatru Econ.* **2021**, *23*, 87–101. [CrossRef]
17. Ahmadvand, A.; Choi, I.; Sahijwani, H.; Schmidt, J.; Sun, M.; Volokhin, S.; Wang, Z.; Agichtein, E. Emory irisbot: An open-domain conversational bot for personalized information access. *Alexa Prize. Proc.* **2018**.
18. Wang, Z.; Ahmadvand, A.; Choi, J.I.; Karisani, P.; Agichtein, E. Emersonbot: Information-focused conversational AI Emory university at the Alexa Prize 2017 challenge. *1st Proceeding Alexa Prize.* **2017**.
19. Finch, S.E.; Finch, J.D.; Ahmadvand, A.; Dong, X.; Qi, R.; Sahijwani, H.; Volokhin, S.; Wang, Z.; Wang, Z.; Choi, J.D.; et al. Emora: An inquisitive social chatbot who cares for you. *arXiv* **2020**, arXiv:2009.04617.
20. Safi, Z.; Abd-Alrazaq, A.; Khalifa, M.; Househ, M. Technical aspects of developing chatbots for medical applications: scoping review. *J. Med. Internet Res.* **2020**, *22*, e19127. [CrossRef] [PubMed]
21. Khoa, B.T. The Impact of Chatbots on the Relationship between Integrated Marketing Communication and Online Purchasing Behavior in The Frontier Market. *J. Messenger* **2021**, *13*, 19–32. [CrossRef]
22. Okonkwo, C.W.; Ade-Ibijola, A. Chatbots applications in education: A systematic review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100033. [CrossRef]
23. Li, J.; Zhou, M.X.; Yang, H.; Mark, G. Confiding in and listening to virtual agents: The effect of personality. In Proceedings of the 22nd International Conference on Intelligent User Interfaces, Limassol, Cyprus, 13–16 March 2017; pp. 275–286.
24. Kim, S.; Lee, J.; Gweon, G. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In Proceedings of the 2019 CHI Conference, Scotland, UK, 4–9 May 2019; pp. 1–12. [CrossRef]
25. Minhas, R.; Elphick, C.; Shaw, J. Protecting victim and witness statement: Examining the effectiveness of a chatbot that uses artificial intelligence and a cognitive interview. *AI Soc.* **2022**, *37*, 265–281. [CrossRef]
26. Ni, L.; Lu, C.; Liu, N.; Liu, J. Mandy: Towards a smart primary care chatbot application. In *Proceedings of the International Symposium on Knowledge and Systems Sciences*; Springer: Singapore, 2017; pp. 38–52.
27. Xiao, Z.; Zhou, M.X.; Fu, W.T. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In Proceedings of the 24th International Conference on Intelligent User Interfaces, Marina del Ray, CA, USA, 16–20 March 2019; pp. 437–447.
28. Siddig, A.; Hines, A. A Psychologist Chatbot Developing Experience. In Proceedings of the AICS, Wuhan, China, 12–13 July 2019; pp. 200–211.
29. Al Adel, A.; Burtsev, M.S. Memory transformer with hierarchical attention for long document processing. In Proceedings of the 2021 International Conference Engineering and Telecommunication (En&T), Dolgoprudny, Russian, 24–25 November 2021; pp. 1–7. [CrossRef]
30. Raheja, V.; Tetreault, J. Dialogue Act Classification with Context-Aware Self-Attention. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 3727–3733. [CrossRef]

31. Ghosh, S.; Varshney, D.; Ekbal, A.; Bhattacharyya, P. Context and Knowledge Enriched Transformer Framework for Emotion Recognition in Conversations. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8. [CrossRef]
32. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461.
33. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
34. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
35. Anguera, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; Vinyals, O. Speaker diarization: A review of recent research. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 356–370. [CrossRef]
36. Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. doccano: Text Annotation Tool for Human. 2018. Available online: https://github.com/doccano/doccano (accessed on 7 August 2023).
37. He, H.; Xu, L.; Choi, J.D. ELIT: Emory Language and Information Toolkit. *arXiv* **2021**, arXiv:2109.03903.
38. Stolcke, A.; Ries, K.; Coccaro, N.; Shriberg, E.; Bates, R.; Jurafsky, D.; Taylor, P.; Martin, R.; Van Ess-Dykema, C.; Meteer, M. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.* **2000**, *26*, 339–374. [CrossRef]
39. Smith, E.M.; Williamson, M.; Shuster, K.; Weston, J.; Boureau, Y.L. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020; pp. 2021–2030. [CrossRef]
40. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
42. Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E.M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage. *arXiv* **2022**, arXiv:2208.03188.
43. Montahaei, E.; Alihosseini, D.; Baghshah, M.S. Jointly Measuring Diversity and Quality in Text Generation Models. *arXiv* **2019**, arXiv:1904.03971.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.