

# Beyond Goldfish Memory\*: Long-Term Open-Domain Conversation

Jing Xu   Arthur Szlam   Jason Weston

Facebook AI Research

New York, NY

{jingxu23, aszlam, jase}@fb.com

## Abstract

Despite recent improvements in open-domain dialogue models, state-of-the-art models are trained and evaluated on short conversations with little context. In contrast, the long-term conversation setting has hardly been studied. In this work we collect and release a human-human dataset consisting of multiple chat sessions whereby the speaking partners learn about each other’s interests and discuss the things they have learnt from past sessions. We show how existing models trained on existing datasets perform poorly in this long-term conversation setting in both automatic and human evaluations, and we study long-context models that can perform much better. In particular, we find retrieval-augmented methods and methods with an ability to summarize and recall previous conversations outperform the standard encoder-decoder architectures currently considered state-of-the-art.

## 1 Introduction

Improvements in the ability to train large neural language models, together with the availability of larger and higher quality dialogue datasets, are spurring the development of increasingly convincing open-domain dialogue models (McTear, 2020). Unfortunately, a major aspect missing from the current state of the art is that human conversations can take place over long time frames, whereas the currently used systems suffer in this setting. Commonly used training and evaluation resources – while large in terms of number of training examples – include only short conversations, typically between 2-15 turns, consisting of a single conversational session. Perhaps for that reason, the current state-of-the-art models such as Meena (Adiwardana et al., 2020) and BlenderBot (Roller et al., 2020) employ Transformers with token truncation lengths of only the 128 most recent tokens, and

are clearly incapable of incorporating long-term conversational context. Consequently, it is unclear how well these models will perform on long or multi-session open-domain conversations. In contrast, a successfully deployed bot will engage in many conversations over a length of time, as capturing organic user interest will garner continual reengagement from returning users. Long-term open-domain communication gives the opportunity for the conversation to develop and even improve with time as the model has more context and more understanding of that specific user’s interests. However current models, due to context truncation, will never use this information.

In this work we study methods for long-term open-domain conversation. As to the best of our knowledge no public domain task exists to study such methods, we collect and release<sup>1</sup> a new English dataset, entitled *Multi-Session Chat* (MSC) that consists of human-human crowdworker chats over 5 sessions, with each session consisting of up to 14 utterances, where the conversationalists reengage after a number of hours or days and continue chatting. Previous sessions are annotated with summaries of important personal points that may be useful in further conversations. When reengaging, conversationalists often address existing knowledge about their partner to continue the conversation in a way that focuses and deepens the discussions on their known shared interests, or explores new ones given what they already know.

We study the performance of two long-context conversational architectures on this task: (i) retrieval-augmented generative models (Lewis et al., 2020b; Shuster et al., 2021); and (ii) a proposed read-write memory-based model that summarizes and stores conversation on the fly. We show that both techniques outperform conventional encoder-decoder Transformers, and that training

\*We use this term colloquially, see Agranoff et al. (1965) for evidence of goldfish long-term memory.

<sup>1</sup>Dataset, model weights and code for this entire project will be made available upon acceptance.

models on our new task give long-term conversational abilities that existing state-of-the-art models lack, as shown in both automatic metrics and human evaluations. We provide extensive experiments and ablations that study the reasons behind these improvements.

## 2 Related Work

A relatively large and growing number of either natural or crowdsourced datasets have been collected and used in open-domain dialogue research. These datasets focus on the vast array of different skills required by a dialogue agent, but conversations lengths are typically short. Recent state-of-the-art open-domain dialogue agents have utilized Daily Dialogue (Li et al., 2017), PersonaChat (Zhang et al., 2018), Empathetic Dialogues (Rashkin et al., 2019), Wizard of Wikipedia (Dinan et al., 2019) and Pushshift.io Reddit (Baumgartner et al., 2020); see Huang et al. (2020) for a review of other datasets. The number of conversational turns in these datasets is in the range of 2-15 turns, we provide statistics of some of these datasets in Table 2. We note there also exist some other kinds of dialogue datasets, e.g. from fantasy role-playing (Urbanek et al., 2019; Rameshkumar and Bailey, 2020) and TV shows as well (Poria et al., 2018). Crowdsourcing long conversations is difficult due to both the expense and the difficulty of employing crowdworkers for long lengths of time due to so called Human Intelligence Tasks (HITs) being typically of a short duration – only “a few minutes” (Paolacci et al., 2010). While organic long conversations regularly transpire on the internet, e.g. on messaging platforms, these are proprietary, and privacy concerns make public release implausible.

Several existing datasets explore the use of personal knowledge used as context to dialogue, which can be seen as a short, simple memory provided to the bot. In Mazaré et al. (2018) such personas were extracted from Reddit and used to train agents. In Zhang et al. (2018) personas were first crowdsourced, and speakers were asked to play those roles. Other works have considered encoding personas into vector-based weights (Li et al., 2016).

In this work, we explore summarizing the long-term conversations that occur in order to store useful information about them. Summarization is a rich field where the vast majority of work focuses on summarizing documents (Kaikhah, 2004; Kryściński et al., 2019; Cheng and Lapata, 2016), for

example summarizing in order to predict other relevant information (West et al., 2019), while there is some work on dialogue as well (Goo and Chen, 2018; Gliwa et al., 2019; Pan et al., 2018).

Standard Transformers have a fixed context length which due to the all-vs-all self-attention mechanism becomes inefficient when it is too large. Consequently, many existing pre-trained models have short token truncation lengths, e.g. 128 tokens, as in BlenderBot (Roller et al., 2020) and Meena (Adiwardana et al., 2020), or 1024 tokens, as in BART (Lewis et al., 2020a). A number of approaches have been proposed to ameliorate this issue. Long-context Transformers consider ways to speed up the self-attention mechanism (Child et al., 2019; Kitaev et al., 2019; Beltagy et al., 2020) and retrieval-augmented methods consider ways to select the pertinent parts of the context to consider (Dinan et al., 2019; Lewis et al., 2020b; Shuster et al., 2021) which can also be related to earlier neural QA methods (Chen et al., 2017).

## 3 Multi-Session Chat

To conduct research on long-term conversations, we require data to both train on and to evaluate models. We consider the natural case where two speakers chat online in a series of sessions as is for example common on messaging platforms. Each chat session consists of 6-7 turns for each speaker. Then, after a certain amount of (simulated) time has transpired, typically hours or days, the speakers resume chatting, either continuing to talk about the previous subject, bringing up some other subject from their past shared history, or sparking up conversation on a new topic. We consider this multi-session long conversation setup, and name our dataset *Multi-Session Chat* (MSC).

**Data Collection** To build our publicly available dataset we employ crowdworkers. We provide screenshots of the task, and details of quality control via onboarding, crowdworker co-rating, and automatic evaluation procedures in Appendix B.

**Personas** Crowdworkers are asked to play a role, rather than speaking about their own personality, which helps mitigate privacy concerns, and ensures diversity even if the same crowdworker conducts multiple conversations. In addition to the crowdworkers being specifically told to play the role, they are also told not to discuss aspects of their real profiles or indeed any personally identifiable informa-

Data Type	Epsisodes	Train Utts.	Summary	Epsisodes	Valid Utts.	Summary	Epsisodes	Test Utts.	Summary
Session 1	8939	131,438	59,894	1,000	7,801	7,768	1015	6,634	6,572
Session 2	4000	46,420	46,420	500	5,897	5,897	501	5,939	5,939
Session 3	4000	47,259	26,976	500	5,890	5,890	501	5,924	5,924
Session 4	1001	11,870	-	500	5,904	5,904	501	5,940	5,940
Session 5	-	-	-	500	5,964	-	501	5,945	-
Total	-	236,987	133,290	-	31,456	25,459	-	30,382	24,375

Table 1: Data statistics of our MULTI-SESSION CHAT dataset. Speakers converse across *sessions*, each of which is a short focused conversation, with subsequent sessions picking up the conversation again hours or days later. We show the number of episodes, utterances (utts) and response summaries for each session.

Dataset	Num. Episodes	Num. Utterances	Unique Tokens	Avg. Utt. Length	Sessions per Episode	Utterances per Episode
Pushshift.io Reddit	-	1.2B	~1M	25.4	1	3.2
PersonaChat (Zhang et al., 2018)	8,939	131,438	18,688	11.9	1	14.7
Wiz. of Wikipedia (Dinan et al., 2019)	18,430	166,787	52,490	19.7	1	9.0
Daily Dialog (Li et al., 2017)	22,236	87,170	20,673	14.5	1	3.9
Empathetic Dialog (Rashkin et al., 2019)	24,850	64,636	19,458	15.3	1	2.6
MULTI-SESSION CHAT (1-3)	4,000	161,440	37,366	21.4	3	40.4
MULTI-SESSION CHAT (1-4)	1,001	53,332	23,387	23.0	4	53.3

Table 2: Comparison of the training data statistics of the MULTI-SESSION CHAT (MSC) dataset compared to other open-domain datasets. We show MSC in two categories: episodes with 3 or 4 sessions, named (1-3) or (1-4).

tion. The role is provided as a series of sentences describing characteristics, events and opinions of the character they are playing. We use the 1,155 personas crowdsourced from Zhang et al. (2018), validation and test use separate personas from the ones used in the training set.

**Session 1** For the first chat session we use the PERSONACHAT dataset (Zhang et al., 2018), which already involves short conversations where two speakers get to know each other for the first time. We note that these conversations rarely go beyond the superficial stage because speakers simply do not have enough turns to discuss any topic deeply.

**Sessions 2, 3, 4, ...** For subsequent sessions, we first select a random amount of (simulated) time that has elapsed since the previous session, chosen to be either 1-7 hours or 1-7 days, as ideally speakers would reengage within that timeframe. We ask the crowdworkers to play the same roles that were played in the previous session, acting as if that amount of time has transpired. We note these crowdworkers may not be the same ones that played those characters in previous sessions, but will be playing the same roles: this makes the task tractable in a crowdworking framework where jobs are typically short, and matching pairs over a long duration would be infeasible. We instruct the workers to “chitchat with another worker for 6

turns, as if you were *catching up* since last time you two spoke.” and that “When you expand the topic, make sure it makes sense with the personal details *already* mentioned.”, i.e. emphasizing that not only must they play their role, but also pay attention to previous interactions with the other speaker.

**Session Lengths** We collect two lengths of training conversation: 4000 episodes with 3 sessions, and 1001 episodes with 4 sessions. For the validation and test data, the sessions extend up to 5 sessions, giving us a way to measure long-context session performance that extends beyond the training set distribution.

**Conversation Summaries (Extended Personas)** We give crowdworkers access to all previous dialogues between the two conversational roles (for the role they are playing, and their partner’s role). However, as the conversation gets longer, this becomes infeasible to read and digest within a limited amount of time. Therefore, between each session, including after session 1, we run a separate crowdworker task in which conversations are summarized into important points, which are much shorter than the full dialogues themselves. We then show previous dialogues, along with these summaries, as the primary reference for subsequent session dialogues. As these summaries were collected in order to store the important points pertinent to either one or the

other speaker, they can also be seen to function as extensions of the original given personas. As the two speakers continue to converse they create more depth to those characters.

**Dataset Examples** Two dataset examples, which consist of four sessions each, along with example summary annotations, are given in Appendix C (provided in the Appendix due to their length).

**Dataset Statistics** Statistics of the multi-session chat dataset are given in Table 1 and a comparison with other standard open-domain dialogue datasets is given in Table 2. We can see that the number of training utterances per episode is larger than other datasets (last column of Table 2). Our multi-session training chats that last 4 sessions have an average of  $\sim 53$  utterances in a full conversation (over all sessions), while our validation and test chats over 5 sessions have an average of  $\sim 66$  utterances. In contrast, other standard datasets are in the range of 2.6-14.7 utterances on average. This brings challenges in open-domain dialogue modeling due to the large context size, e.g. an average of 1614 tokens as tokenized by the BlenderBot BPE dictionary (Roller et al., 2020), where the Transformer used in that work has a truncation length of 128. Further information on the dataset including analysis of its quality is given in Appendix B.

## 4 Modeling Multi-Session Chat

### 4.1 Transformer Encoder-Decoders

The most straight-forward approach for modeling dialogue using our new task is simply to use a large language model as is standard in open-domain dialogue, i.e. an encoder-decoder Transformer as in the Meena (Adiwardana et al., 2020) and BlenderBot (Roller et al., 2020) systems. We consider using the BST 2.7B parameter model from BlenderBot as an initial pre-trained model, which we then fine-tune on the Multi-Session Chat task.

**Encoder Truncation** As BST 2.7B has a truncation of 128 tokens in the encoder, we consider extending this to a larger input. To do this, we extend its available positional encodings from 128 to 512 or 1024 tokens as we fine-tune the whole network on the downstream task. We add new positional embeddings to be trained such that the existing ones (the first 128 most recent tokens) do not change from before. We then evaluate the impact of these choices in order to select the best model.

### 4.2 Retrieval-Augmentation

A popular technique when dealing with a large collection of text, only some of which is relevant, is to use a retrieval-augmented Transformer. A retrieval system is used to search over a text collection, and select some of it to be included in the final encoding which is attended to by the Transformer decoder.

**RAG** The RAG (Retrieval-Augmented Generation) approach (Lewis et al., 2020b) utilizes a neural-retriever-in-the-loop which is itself a second Transformer. Documents to be retrieved are stored in an approximate nearest-neighbor FAISS index (Johnson et al., 2019), and a DPR (Dense Passage Retrieval) (Karpukhin et al., 2020) Transformer bi-encoder model is used to score document-context pairs in order to rank them based on their match, where the base DPR model is pre-trained on QA data pairs. The DPR model is thus used to both retrieve from the FAISS index, and then score the top  $N$  candidates. The entire system is trained end-to-end so that retrieval is optimized to help improve generation. This setup was shown to work for dialogue in particular in Shuster et al. (2021).

**FiD and FiD-RAG** We also consider the Fusion-in-Decoder (FiD) (Izacard and Grave, 2020), another method that has been shown to perform well. In this approach, the pre-trained retriever is used directly: each of the top  $N$  documents returned is prepended to the context and encoded separately by the encoder, and finally all the results are concatenated. The decoder then attends to these encodings to produce a final response. We consider the pre-trained retriever to either be standard pre-trained DPR, or the RAG-trained retriever, called FiD-RAG (Shuster et al., 2021).

**Retriever and Documents** In this work the set of passages in the memory is not large enough to require a FAISS index, but it is large enough that retrieval may be useful. We thus store for every item in the memory the vector encoding by the DPR model (whereas in the FAISS approach this dense vector is approximated instead). Then given a dialogue context, we score each memory using the bi-encoder, and use the top  $N$  for generation. In our case, the memories consist of dialog utterances from the history of the conversation. We consider the chunk (document) size as a hyperparameter and try either encoding utterances as separate documents, or else whole sessions (or session summaries) as documents. The latter (whole se-

Pre-Train Model	Truncation	Sessions 1-4	Session 1	Session 2	Session 3	Session 4	Trunc% (S4)
<i>With no previous session context</i>							
BST 2.7B	128	9.23	8.76	9.45	9.31	9.40	51%
BST 2.7B	512	9.06	8.18	9.42	9.26	9.36	0%
BST 2.7B	1024	9.08	8.20	9.46	9.29	9.37	0%
<i>With previous session dialogue context</i>							
BST 2.7B	128	9.16	8.75	9.32	9.22	9.32	100%
BST 2.7B	512	8.87	8.15	9.14	9.04	9.17	100%
BST 2.7B	1024	8.89	8.17	9.18	9.05	9.16	80%
<i>With previous session summary context</i>							
BST 2.7B	128	9.09	8.77	9.24	9.12	9.24	100%
BST 2.7B	512	8.79	8.17	8.69	9.15	9.22	36%
BST 2.7B	1024	8.80	8.18	9.05	8.91	9.04	0%

Table 3: **Comparison of different context truncation lengths and context types** when training on MULTI-SESSION CHAT. We show validation perplexity for various models across different sessions, and percent of tokens truncated for session 4 (last column).

Model Context	Session				Session Openings			
	2	3	4	5	2	3	4	5
No Session History	9.46	9.29	9.37	9.30	9.96	10.99	10.69	10.46
Dialogue History	9.18	9.05	9.16	9.08	7.55	8.48	8.27	7.94
Gold Summary	9.04	8.90	9.02	8.96	6.98	7.96	7.94	7.77
Gold Summary (without time features)	9.05	8.91	9.04	8.95	6.97	7.95	7.97	7.74
Gold Summary (partner’s only)	9.14	8.99	9.11	9.03	7.66	8.49	8.49	8.07
Gold Summary (self only)	9.29	9.10	9.18	9.13	8.40	8.94	8.52	8.39
Predicted Summary	9.11	8.98	9.07	9.00	7.44	8.43	8.20	7.81

Table 4: **Summaries vs. Dialogue Context Performance** when training on MULTI-SESSION CHAT, reporting validation perplexity, using a BST 2.7B-1024 pre-trained model with MSC fine-tuning. Note that the last row in this Table corresponds to the SumMem-MSC 2.7B (truncate 1024) row in Table 15 in the Appendix.

sions) worked better, and we report those in the final results. For  $N$  we try values 3, 5 and 6, and also choose the best for each method according to the validation set.

### 4.3 Summarization Memory-Augmentation

The retrieval-augmentation models described in the previous section retrieve from the set of past dialogues. Simply storing historical text in the memory in their raw form is a simple approach that is often used elsewhere in the literature, e.g. in question answering or knowledge-grounded dialogue. However, those approaches have two potential drawbacks: (i) there is a lot of context to store, and hence retrieve from; (ii) no processing has been done on that content, so the reading, retrieving and combining operations required to generate an answer leave a lot of work for the model to do. We therefore propose instead a novel memory augmentation that first *summarizes pertinent knowledge* and only stores that in an attempt to solve both problems.

The procedure involves two main components:

1. An encoder-decoder abstractive summarizer that takes as input the dialogue history, and outputs a summary of new pertinent information contained in the last dialogue turn, or “no-summary” if there is no new information found. When found, the summarized knowledge is added to the long-term memory.
2. A memory-augmented generator that takes the dialogue context and access to the long-term memory, and generates the next response.

For (1) we can use the human annotated data from our newly collected MSC task to know what summaries to generate (see section 3 and Figure 1 in the Appendix). We thus train a supervised encoder-decoder model to produce summaries.

For (2) we can use the same systems as presented in subsection 4.2 to both retrieve from the summarization memories, and to finally generate an appropriate response. That is, we store the summaries in documents and retrieve them using either RAG, FiD or FiD-RAG.

Model Context	Session				Session Openings				Sparsity
	2	3	4	5	2	3	4	5	
Gold summary	9.04	8.90	9.02	8.96	6.98	7.96	7.94	7.77	42.0%
Predicted Summary (sampling 5%)	9.11	8.98	9.07	9.00	7.44	8.43	8.20	7.81	29.1%
Predicted Summary (sampling 25%)	9.11	8.97	9.07	9.01	7.46	8.53	8.22	7.94	41.4%
Predicted Summary (sampling 50%)	9.14	8.99	9.08	9.02	7.57	8.62	8.37	8.11	50.7%
Predicted Summary (sampling 100%)	9.14	8.99	9.10	9.03	7.68	8.69	8.56	8.25	61.8%

Table 5: **Predicted Summaries when subsampling the no-summary class** on MULTI-SESSION CHAT, reporting validation perplexity, using a BST 2.7B-1024 pre-trained model with MSC fine-tuning. The last column shows the sparsity of the summarizations (how often a summary line is generated), which can be controlled by subsampling the no-summary class at training time. Subsampling gives better results and closer sparsity levels to the original human annotated data.

Training Data	Session				
	1	2	3	4	All
Session 1	8.24	11.4	11.2	11.3	10.5
Sessions 1+2	8.21	9.21	9.09	9.24	8.94
Sessions 1+2+3	8.16	9.05	8.93	9.06	8.80
Sessions 1+2+3+4	8.16	9.02	8.89	9.02	8.77

Table 6: **Varying the Number of Training Sessions** when training on MULTI-SESSION CHAT, reporting validation perplexity, using a BST 2.7B-1024 pre-trained model with MSC using gold summaries.

## 5 Experiments

**Using session dialogue context** We compare different context types in Table 3, evaluating over sessions 1-4. We observe an improvement in perplexity when incorporating the dialogue history from previous chat sessions, compared to no session context, for all sessions after the first one, and for all context lengths – with larger context lengths giving better improvement. This shows that our human conversationalists do use previous sessions to make dialogue more salient in successive sessions as this is reflected in the collected human-human dataset – and that our models are able to utilize this information well when training on this data.

**Using session summary context** We also show performance of using gold session summary contexts, as annotated by crowdworkers, in Table 3. As the summaries include salient points, they are potentially more informative than dialogue context for a generative model. We find perplexities improve when using summaries compared to using dialogue context (or no context at all) over all sessions after the first one, and for all context lengths, although the improvements are not large. This shows that conversation summaries are potentially useful for dialogue generation in the long-context case.

### Comparing performance on session openings

Session openings in the MSC dataset look quite different to other dialogue datasets that do not have a session format. This is because they involve an opening message that is intended to reengage the other speaker after a period of time, using known information that has been exchanged between speakers. In Table 4 we compare models that use different context types on only these opening responses. In this case we find much more pronounced perplexity differences between no session context history, dialogue history or summary context history. For example, we see around around 2 perplexity points difference between using or not using previous session context. We show examples of opening session generations in Appendix C. We observe that opening messages are categorically different to other conversation turns, typically involving a statement or question given knowledge of shared interests contained in the long-context. This explains why collection of our new dataset is so important for this goal, as reflected in perplexity improvements. That is, they indicate that our new task will likely help improve multi-session conversational engagement with users compared to existing training schemes.

### Comparing different context lengths

As shown in Table 3 changing the context length of a Transformer can impact the performance in our task. With no previous session context, improvements are minimal for sessions 2 onwards. However, using session dialogue or summary contexts we do see improvements with larger lengths of 512 or 1024 tokens, compared to 128. The last column of Table 3 shows the percentage of responses where the input to the Transformer is truncated for session 4, for each truncation length. One can see that using summaries can be beneficial as they are shorter,

Model	Session 1	Session 2	Session 3	Session 4	Session 5	Session Openings
BST 2.7B (Roller et al., 2020)	8.97	9.98	10.26	10.40	10.50	12.92
MSC 2.7B (truncate 128)	8.87	8.89	9.10	9.21	9.27	8.95
MSC 2.7B (truncate 1024)	8.25	8.76	8.93	9.07	9.16	8.09
MSC 2.7B (RAG)	8.22	8.78	8.97	9.11	9.17	8.10
MSC 2.7B (FiD)	8.22	8.75	8.92	9.05	9.11	8.06
MSC 2.7B (FiD-RAG)	8.23	8.75	8.93	9.04	9.11	8.03
SumMem-MSC 2.7B (truncate 1024)	8.25	8.71	8.89	9.01	9.09	8.04
SumMem-MSC 2.7B (RAG)	8.24	8.81	9.00	9.10	9.17	8.05
SumMem-MSC 2.7B (FiD)	8.20	8.71	8.89	9.00	9.07	7.91
SumMem-MSC 2.7B (FiD-RAG)	8.22	8.70	8.89	9.00	9.07	7.87

Table 7: **Test perplexity across sessions** for our retrieval- and memory-augmented models (bottom two blocks) compared to several encoder-decoder baselines (top three rows).

meaning they are truncated less often, which can thus also help performance.

**Summary context performance** We can ablate the summary model training data to understand its impact further, results of which are given in Table 4. We see that removing the time feature (indicating how long ago the previous session occurred) only has minimal effect. Removing either the partner or self summary (and keeping the other one), on the other hand, has a larger effect in both cases, where keeping the self summary is slightly more important. Keeping both features is best. These differences, as before, are magnified when looking at session opening performance.

**Predicted summary models** We train models to predict dialogue summaries, and use predicted summaries of previous sessions as context (instead of the full dialogue history or the gold summary). The training data for predicting summaries consists of, for each turn, either a summarizing sentence or the *no\_summary* label. As 42% of turns have the *no\_summary* label, this can be overexpressed in the model at beam decoding time<sup>2</sup>, we therefore experiment with sampling this label only  $K\%$  of the time during training in Table 5. Example predictions (for the 5% sampling model) are shown in Figure 1. We find that subsampling gives better results and closer sparsity levels to the original human annotated data (e.g., with  $K = 25\%$ ). We compare predicted summaries with  $K = 5\%$  sampling to other methods of modeling long-context in Table 4. We observe results that are between using a standard dialogue history (predicted summaries are slightly better), and using gold summaries (predicted summaries are not as good).

<sup>2</sup>We use a beam size of 3 and minimum beam length 10 with no context blocking.

**Varying the number of training sessions** We vary the amount of available training sessions from 1-4, with results reported in Table 6. We observe large gains when using more than one training session compared to only one (around 1.5 perplexity points), again justifying the construction of our MSC training data. The gains however decrease with the number of available sessions, e.g. between having 1-3 training sessions vs. 1-4 only gives a 0.03 perplexity gain averaged across sessions. The gain even on session 4 is not that large despite the 1-4 training data being in-distribution, whereas 1-3 is not, in addition to 1-4 having more training data.

**Retrieval-augmentation model** Comparison of our retrieval-augmented methods are given in Table 7, training on MSC using the BST 2.7B model as pre-training, hence called MSC 2.7B (RAG), (FiD) or (FiD-RAG), depending on the augmentation method. These methods are compared to the existing BlenderBot model (BST 2.7B), or training with MSC with no augmentation (MSC 2.7B with different dialogue history context truncation lengths). We find that all three retrieval augmentation methods, when using the session level-document size as retrieval documents, can effectively use retrieval to extend the conversation history length. We see a large performance improvement over the existing BlenderBot model or a truncation of 128 of the MSC 2.7B model. Performance improvements over MSC 2.7B with a truncation length of 1024 are minimal, but the retrieval-augmented models are guaranteed to have a memory that essentially never forgets the conversation, no matter how long it gets, whereas the truncation model does not.

**Summary memory model variants** We next compare the summary memory models, whereby

Model	Reference own topic	Reference other’s topic	New topic	Engaging Response	Final Rating	# Annotated Responses
BST 2.7B (Roller et al., 2020)	19.9%	14.5%	69.0%	53.0%	3.14	668
MSC 2.7B (truncate 128)	15.8%	21.8%	75.8%	56.5%	3.29	673
MSC 2.7B (truncate 1024)	15.0%	22.5%	74.4%	54.2%	3.47	653
SumMem-MSC 2.7B (RAG)	19.6%	33.8%	72.7%	<b>62.1%</b>	<b>3.65</b>	668
SumMem-MSC 2.7B (FiD)	22.1%	30.7%	76.4%	<b>58.9%</b>	<b>3.62</b>	662
SumMem-MSC 2.7B (FiD-RAG)	24.2%	26.4%	78.3%	<b>59.3%</b>	<b>3.68</b>	649

Table 8: **Human Evaluation Results.** Performance of various models measured during conversations with crowdworkers. Engaging response and final rating numbers in bold are statistically significant compared to BST 2.7B ( $p$ -value  $< 0.05$ ) using a  $t$ -test. See subsection 5.1 and Appendix B for more details.

previous dialogue history is summarized before being stored in the model’s long-term memory, called SumMem-MSC 2.7B. We use the RAG, FiD, or RAG-FiD methods to retrieve from that memory, or we compare to a fixed memory of 1024 tokens that is truncated, resulting in four different methods that we compare. Results are given in Table 7. While improvements are small, we see the same patterns as for the retrieval-augmented methods that SumMem-MSC 2.7B FiD-RAG is better than FiD which is in turn better than RAG, with FiD and FiD-RAG better than truncation at session openings. Moreover, all SumMem-MSC models outperform their retrieval-augmented model counterparts MSC 2.7B (RAG/FiD/FiD-RAG). SumMem-MSC 2.7B (FiD-RAG) thus provides the best results out of all methods tested in this work.

**Further Detailed Automatic Metrics** Our analysis so far measured perplexity. We report more automatic metrics (F1 and BLEU) in Appendix A, which yield similar conclusions.

## 5.1 Human Evaluation

We perform a human evaluation using crowdworkers. The conversations begin with two randomly chosen personas from the validation set, and one is assigned to the crowdworker who is asked to play that role. We select the conversation to be the 5<sup>th</sup> session that these two speakers will converse, and make available the summary of the previous 4 sessions. We ask the crowdworkers to have a natural conversation, where they will also evaluate their partner’s responses for conversational attributes, in particular whether they reference knowledge of their own or the other speaker’s persona (or topics they discussed) from previous sessions, from the current session, or neither. On each turn of the conversation the crowdworker is asked to check all attribute boxes that apply. A screenshot can be found in Figure 6 in the Appendix showing the UI.

Each conversation consists of 15 messages (7 from the human, 8 from the bot). At the end of the conversation, an additional question collects an overall engagingness score (out of 5) for their speaking partner.

The results are given in Table 8. We find that MSC-trained models outperform BlenderBot (BST 2.7B) in terms of both per-turn engaging responses and final ratings. Further, our summarization memory models (all three variants RAG, FiD and FiD-RAG) outperform encoder-decoders with different levels of truncation of the dialogue history (MSC 2.7B with truncate 128 and 1024). For example, SumMem-MSC 2.7B (RAG) achieves an engaging response rate of 62.1% and final rating of 3.65, compared to BlenderBot’s 53.0% and 3.14 and MSC 2.7B (truncate 1024)’s 54.2% and 3.47. For all MSC models, while rates of referencing their own topics are not particularly increased, we do observe increased rates of referencing partner topics from previous sessions, with higher rates for the summarization memory models. For example, 33.8% for SumMem-MSC 2.7B (RAG) compared to BlenderBot’s 14.5%. This is likely an important reason why human raters feel the summarization memory models are more engaging.

## 6 Conclusion

We have shown that existing dialogue models, both in terms of training data and models trained, fail to conduct long-term conversations adequately. Our work investigates recent model architectures to ameliorate this issue, and collects a new crowd-sourced task, *Multi-Session Chat* to both train and evaluate these models. We show, in terms of both automatic metrics and human evaluations, that these long-context dialogue modeling approaches outperform the previous systems. Future work should investigate further improvements to architectures for the long-context dialogue setting.

## 7 Ethical Considerations

The dialogue models we use in this work utilize large language models, and therefore have similar concerns as in other work, in particular concerns about toxic language, bias and other issues during language generation (Bender et al., 2021). For open-domain dialogue in particular, see Xu et al. (2020); Dinan et al. (2021) for reviews of the literature and evaluation of recent methods that try to mitigate these safety issues.

Our work focuses on models with long-term memory and open-domain conversations wherein speakers may divulge personal interests. We remark that, during data collection, crowdworkers were specifically playing roles with given personality traits, not talking about themselves, and hence not identifying any personal information. During conversations with our trained models, the models will store information they learn from the exchange. In contrast to current standard language models, our models have the capability of storing this in the long-term. This information is stored in the memory of the model, private to the individual’s conversation, and hence is not shared with anyone else.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Bernard W Agranoff, Roger E Davis, and John J Brink. 1965. Memory fixation in the goldfish. *Proceedings of the National Academy of Sciences of the United States of America*, 54(3):788.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435*.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1870–1879. Association for Computational Linguistics.
- Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
- Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. *arXiv preprint arXiv:1905.05471*.
- Richard Csaky and Gábor Recski. 2020. The gutenber dialogue dataset. *arXiv preprint arXiv:2004.12752*.
- Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsun corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.
- Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.
- Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.
- Khosrow Kaikhah. 2004. Automatic text summarization with neural networks. In *2004 2nd International IEEE Conference on Intelligent Systems’*. *Proceedings (IEEE Cat. No. 04EX791)*, volume 1, pages 40–44. IEEE.

- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.
- Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.
- Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raïson, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.
- Michael McTear. 2020. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251.
- Haojie Pan, Junpei Zhou, Zhou Zhao, Yan Liu, Deng Cai, and Min Yang. 2018. Dial2desc: end-to-end dialogue description generation. *arXiv preprint arXiv:1811.00185*.
- Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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 673–683, Hong Kong, China. Association for Computational Linguistics.
- Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. *arXiv preprint arXiv:1909.07405*.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213. ACL.

## A Extra Results

**Further Test Set Metrics** We show the test BLEU-1 in Table 11, test BLEU-2 in Table 12, test BLEU-4 in Table 13 and test F1 in Table 14.

**Main Validation Results** We show the validation perplexity in Table 15 (corresponding to the test perplexity in Table 7).

## B Data Collection & Data Quality

### B.1 Data Collection & Quality Control

**Crowdsourced Data Collection** The data collection lasted for around 6 months and in total over 1000 crowdworkers who are English-speaking annotators located in the United States were recruited and compensated through the Amazon Mechanical Turk platform. Before the data collection starts, all crowdworkers are informed that any message they send may be publicly disclosed for research purposes, and are instructed not to send any personal identifiable information (for example, name, address, email, or phone number etc.) in their messages.

**Quality Control** To optimize the quality of collected data, we implement a list of quality controls in both the conversation summarization task and the multi-session chat task. All crowdworkers must achieve high scores on the onboarding task that resembles the actual crowdsourcing tasks before they are eligible to work on the Human Intelligence Task (i.e. HIT, the term used by Amazon’s Mechanical Turk to refer to a single instance of a crowdworker task). During the actual multi-session chat, crowdworkers are instructed to report at each conversational turn if the message from the other speaker is of poor quality or has BAD behaviors, for example, contradicting to or repeating what has been mentioned; changing topics too often, etc. A final rating is also collected at the end of each chat indicating how much they enjoy talking to their conversational partner. Crowdworkers that are frequently reported as producing messages of low quality or

receiving low ratings are blocked from working on any future HITs. Moreover, dialogues that fail the acceptability checks such as minimum average message length or have too many dialogue turns rated as low quality are also filtered out from the final dataset.

We show screenshots of the crowdsourced Multi-Session Chat task in Figure 5 as well as the crowdsourced human evaluation task in Figure 6.

### B.2 Data Quality Analysis

Following analysis of previous datasets such as DailyDialogue that exposed significant overlap between train and test (Csaky and Recski, 2020; Csáky et al., 2019), we measure various overlap statistics on our newly collected dataset.

Table 9 gives statistics for the number of unique utterances in a given data split (comparing to all other utterances in the same data split). We see there are very few duplicated messages across crowdworkers. For example in the validation set there are only 4, these are: “*What kind of dogs do you have ?*”, “*How old are your children ?*”, “*What kind of dancing do you do ?*” and “*What kind of dog is he ?*”, i.e. very common questions that just happened to be asked twice each across different conversations.

Table 10 gives statistics for the number of unique utterances across data splits (comparing either valid or test utterances to all utterances in the train data split). We again see there are very few duplicated messages across data splits, looking at these specific messages we again see they are typical things that might be commonly asked.

## C Dataset Examples

**MSC Dataset Examples** We show two MSC dialogue examples in Figure 2 and Figure 3 each consist of four sessions. We also show example summary annotations in Figure 1.

**Session Opening Examples** We show example session opening predictions of a model trained on gold summaries in Figure 4.

Split	Total	Unique	Unique%
Train	105429	105549	99.88%
Valid	17687	17691	99.97%
Test	17798	17803	99.97%

Table 9: Utterance overlap by data split.

Split vs. Split	Overlap	Overlap%
Train vs. Valid	24	0.135%
Train vs. Test	45	0.252%

Table 10: Utterance overlap across data splits.

Model	Session 1	Session 2	Session 3	Session 4	Session 5	Session Openings
BST 2.7B (Roller et al., 2020)	15.1	14.7	14.1	13.8	13.8	10.9
MSC 2.7B (truncate 1024)	15.5	15.5	15.2	14.7	14.7	11.3
SumMem-MSC 2.7B (RAG)	15.3	15.4	15.1	14.9	14.8	11.5
SumMem-MSC 2.7B (FiD-RAG)	15.5	15.7	15.4	15.3	15.1	11.6

Table 11: **Test BLEU-1 across sessions** for our memory-augmented models compared to several encoder-decoder baselines (top two rows).

Model	Session 1	Session 2	Session 3	Session 4	Session 5	Session Openings
BST 2.7B (Roller et al., 2020)	5.43	4.84	4.39	4.19	4.31	1.79
MSC 2.7B (truncate 1024)	5.53	5.26	4.91	4.69	4.75	2.18
SumMem-MSC 2.7B (RAG)	5.37	5.26	4.96	4.89	4.85	2.45
SumMem-MSC 2.7B (FiD-RAG)	5.42	5.26	5.13	5.02	5.00	2.42

Table 12: **Test BLEU-2 across sessions** for our memory-augmented models compared to several encoder-decoder baselines (top two rows).

## D Model Training Settings

We use the openly available [ParlAI](#) framework for all training runs, as well as for evaluations, where metrics are measured using default settings. All the fine-tuned models are trained with a maximum of eight 32GB GPUs (NVIDIA V100), optimized with Adam using  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1e - 08$ . Models are trained up to 4000 updates with batch size up to 128. The typical fine-tuning time for standard transformer encoder-decoder is 8 hrs before it early stops, and for retrieval-based model is 16 hrs.

Model	Session 1	Session 2	Session 3	Session 4	Session 5	Session Openings
BST 2.7B (Roller et al., 2020)	0.858	0.625	0.535	0.551	0.570	0.107
MSC 2.7B (truncate 1024)	0.837	0.745	0.629	0.595	0.631	0.139
SumMem-MSC 2.7B (RAG)	0.919	0.699	0.677	0.664	0.656	0.228
SumMem-MSC 2.7B (FiD-RAG)	0.853	0.765	0.673	0.679	0.678	0.222

Table 13: **Test BLEU-4 across sessions** for our memory-augmented models compared to several encoder-decoder baselines (top two rows).

Model	Session 1	Session 2	Session 3	Session 4	Session 5	Session Openings
BST 2.7B (Roller et al., 2020)	18.3	19.4	19.0	19.0	19.4	13.7
MSC 2.7B (truncate 1024)	18.6	20.0	19.9	19.8	20.0	14.1
SumMem-MSC 2.7B (RAG)	18.3	19.9	19.9	20.0	20.1	14.4
SumMem-MSC 2.7B (FiD-RAG)	18.6	20.1	20.0	20.2	20.2	14.5

Table 14: **Test F1 across sessions** for our memory-augmented models compared to several encoder-decoder baselines (top two rows).

Model	Session 1	Session 2	Session 3	Session 4	Session 5	Session Openings
BST 2.7B (Roller et al., 2020)	8.84	10.56	10.44	10.51	10.44	13.04
MSC 2.7B (truncate 128)	8.75	9.32	9.22	9.32	9.23	8.95
MSC 2.7B (truncate 1024)	8.17	9.18	9.05	9.16	9.08	8.06
MSC 2.7B (RAG)	8.14	9.16	9.06	9.18	9.10	8.04
MSC 2.7B (FiD)	8.16	9.14	9.02	9.10	9.04	7.97
MSC 2.7B (FiD-RAG)	8.16	9.13	9.02	9.10	9.04	7.96
SumMem-MSC 2.7B (truncate 1024)	8.18	9.11	8.98	9.07	9.00	7.97
SumMem-MSC 2.7B (RAG)	8.16	9.19	9.07	9.17	9.09	7.95
SumMem-MSC 2.7B (FiD)	8.16	9.09	8.97	9.07	8.99	7.82
SumMem-MSC 2.7B (FiD-RAG)	8.16	9.08	8.96	9.07	8.99	7.78

Table 15: **Valid perplexity across sessions** for our retrieval- and memory-augmented models (bottom two blocks) compared to several encoder-decoder baselines (top three rows).

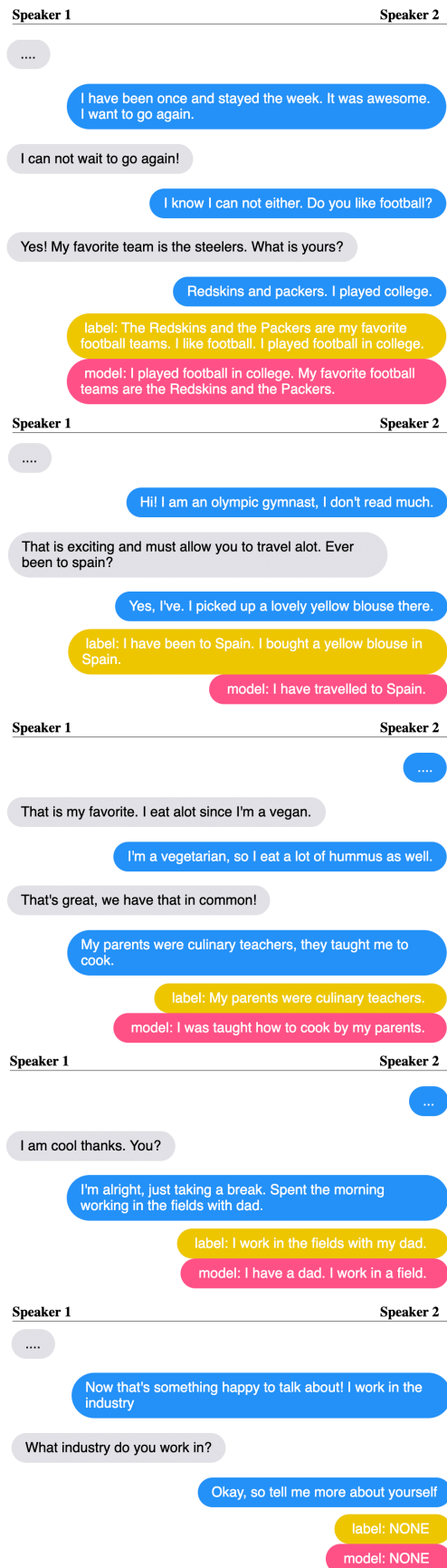


Figure 1: Example summary annotations and predictions on the validation set. We show the gold human annotation (label) and our model prediction (model).



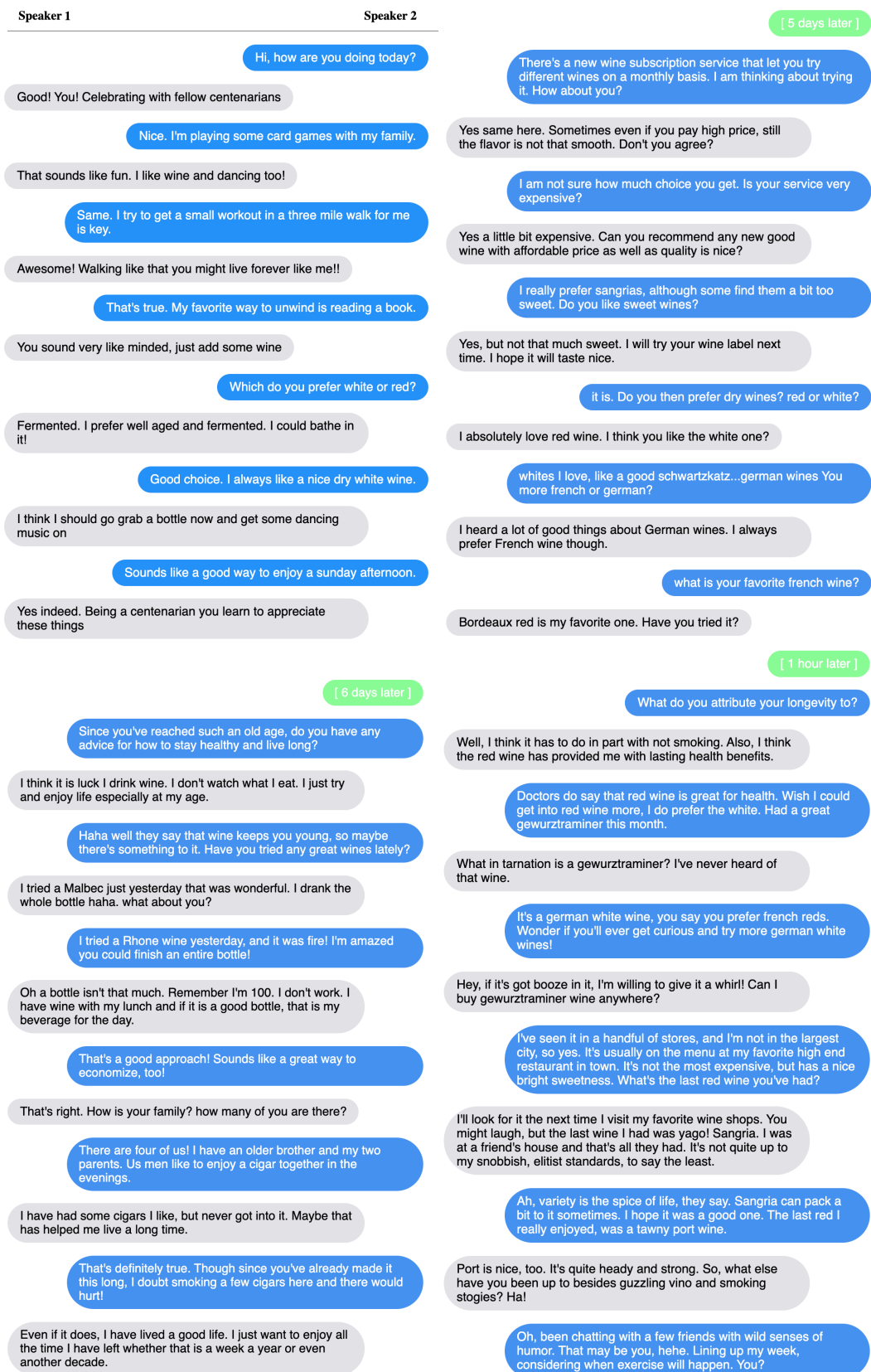


Figure 3: Example four session conversation from the newly collected Multi-Session Chat dataset. New sessions refer back to previous subjects, explore them in depth, or spark up conversation on new topics.

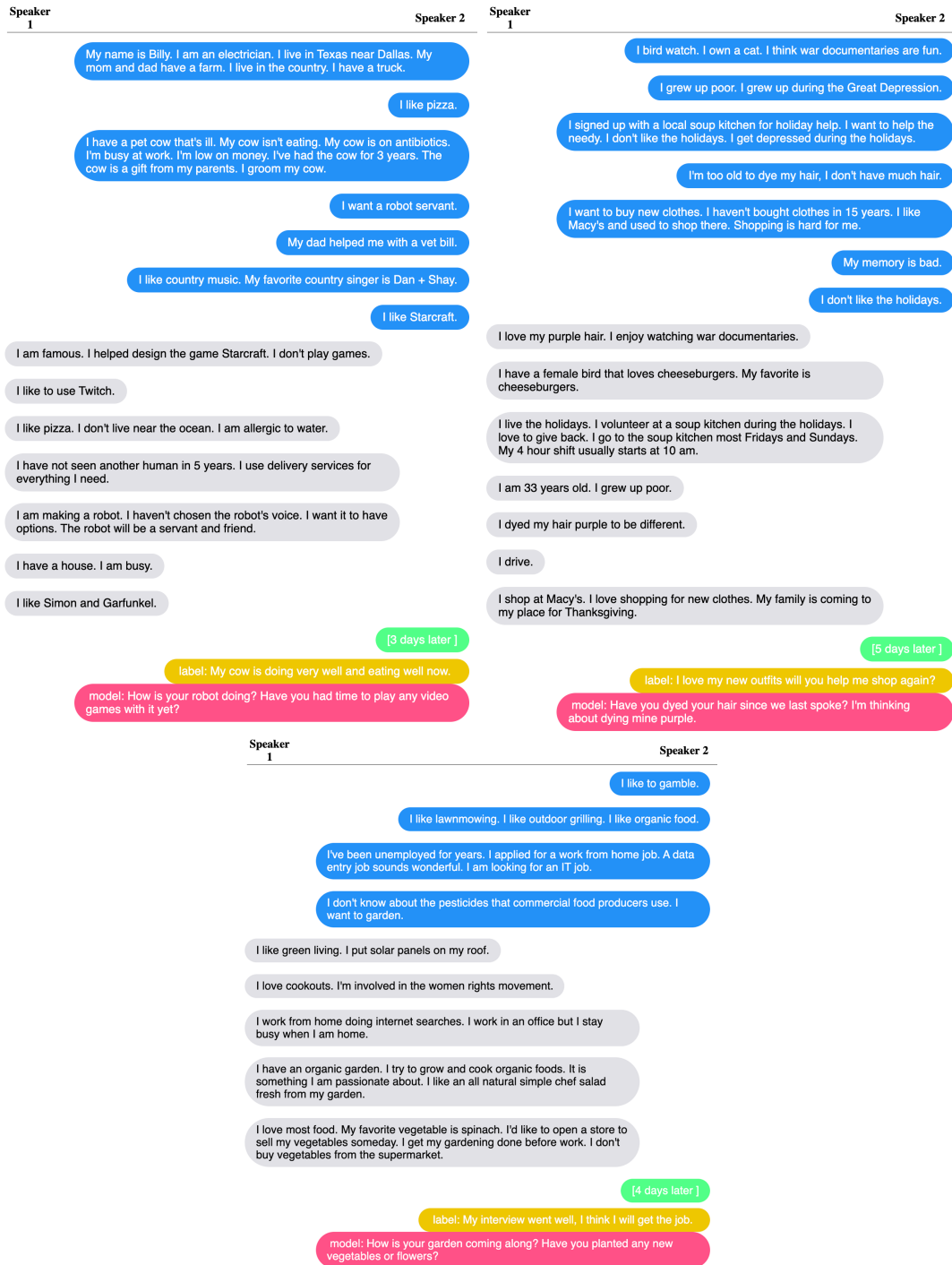


Figure 4: Example opening annotations and predictions given gold summaries on the validation set. We show the gold human annotation (label) and our model prediction (model).

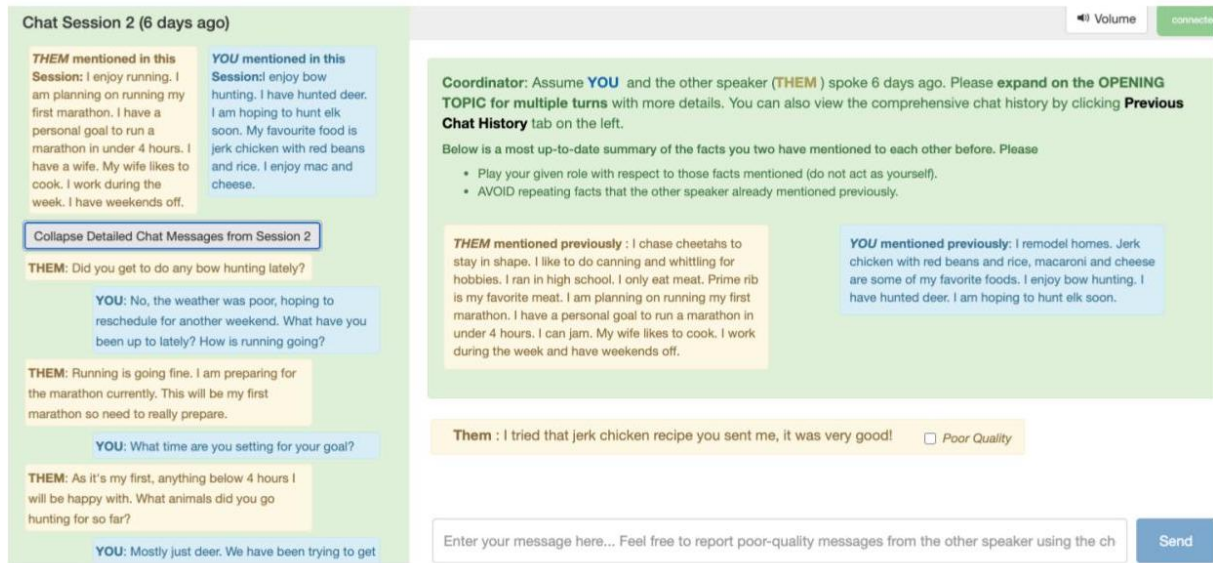


Figure 5: Screenshot from the crowdforker multi-session chat. The left panel shows the instructions as well as all dialogue history from previous sessions, and the right panel contains the conversation for the current session.

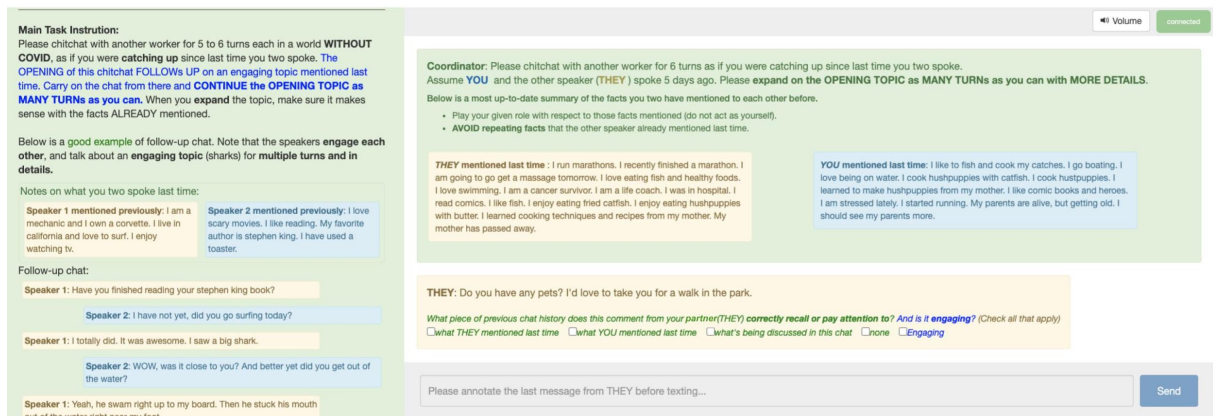


Figure 6: Crowdforker evaluation task screenshots. The left panel shows the instructions, and the right panel contains the conversation. In the human evaluation results in the main paper the "what THEY mentioned last time" binary checkbox is converted to a percentage over all annotated responses, and termed "Reference own topic" in Table 8. Similarly, "what YOU mentioned last time" is termed "Reference other's topic" in Table 8.