

Response Quality in Human-chatbot Collaborative Systems

Naman Ahuja

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Jiepu Jiang, Chair

Anuj Karpatne

Calvin J. Ribbens

May 12, 2020

Blacksburg, Virginia

Keywords: Conversational systems, chatbot, human-AI collaboration, dialogue systems

Copyright 2020, Naman Ahuja

Response Quality in Human-chatbot Collaborative Systems

Naman Ahuja

(ABSTRACT)

We study human-chatbot collaborative conversation systems that enable humans to leverage AI chatbot outputs during an online conversation with others. We evaluate response quality in two collaborative systems and compare them with human-only and chatbot-only settings. Both collaborative systems present AI chatbot results as suggestions but encourage the synthesis of human and chatbot responses to different extents. We also examine the influence of chatbot choices, including both retrieval-based and generation-based methods, and the number of suggestions on collaborative systems. Experimental results show that our collaborative systems can significantly improve the efficiency to formulate a response and improve its quality compared with a human-only system while sacrificing the fluency and humanness of the messages. Compared with a chatbot, collaborative systems can provide answers that are more fluent, human-like, and informative. We also found that the retrieval-based chatbots perform better than the generation-based one from all aspects. The optimal number of chatbot suggestions is one, and showing more suggestions has reduced user efficiency.

Response Quality in Human-chatbot Collaborative Systems

Naman Ahuja

(GENERAL AUDIENCE ABSTRACT)

Artificial Intelligence (AI) systems have become remarkably interactive and accurate with them becoming an integral part of our life. The increasing use of personal assistants like Siri and the application of AI in important real-world tasks such as medical imaging and diagnosis show that AI can perform as good as trained human experts. Organizations today are expanding at a rapid rate and need to service millions of customers concurrently to remain competitive in the market. With the recent success of AI chatbots, the collaboration of Human and AI to augment customer service management is one of the most sought out solutions to this requirement. A service flow where virtual agents and people work together can be a boon to the industry by making the human agents smarter with a bot “whispering” in their ears. We present the design of various collaborative systems we have developed and discuss the improvements in response efficiency and quality due to them in multiple online user experiments. The results of this study can be used to improve conversational chat systems that assist human agents to improve their response time and quality and identify features of the AI agent that are most beneficial for improving the conversation.

Dedication

To my family and friends for their unrelenting faith in me.

Acknowledgments

I am greatly indebted to my advisor Dr. Jiepu Jiang, for his unwavering personal and professional guidance during my course at Virginia Tech. His constructive feedback, insightful suggestions, and patience made this work possible. He gave me the freedom to experiment with the research, steering me in the right direction when needed. I would also like to thank my committee members Dr. Cal Ribbens and Dr. Anuj Karpatne. Finally, whatever I am today is because of my family and friends. I would like to thank them for believing in me.

Contents

List of Figures	ix
List of Tables	xiii
1 Introduction	1
2 Related Work	4
2.1 Conversational Systems	4
2.2 Human-AI Interaction	6
2.3 Search Engine Query Suggestion and Auto Completion	7
3 Collaborative Conversation Systems	9
3.1 Collaborative Systems C1 and C2	9
3.2 Retrieval and Generation-Based Chatbots Used	10
3.2.1 Retrieval Transformer Memory Network	12
3.2.2 Generative Transformer Memory Network	13
3.3 Experimental Design for the Pilot Study	14
3.3.1 User Study Design	14
3.3.2 Response Quality Judgments	15

3.4	Experimental Design for the Final Study	19
3.4.1	User Study Design	20
3.4.2	Response Quality Judgments	21
4	Results from the Pilot Study	26
4.1	Response Time and Keystrokes	26
4.2	Response Length	28
4.3	Response Quality	28
4.3.1	Informativeness	29
4.3.2	Fluency and Humanness	30
4.4	Number of Chatbot Suggestions	31
4.5	C1 vs. C2	32
4.6	Conclusion and Discussion	32
5	Results from the Final User Study	34
5.1	Response Time and Keystrokes	34
5.2	Response Length and Number of Conversations	35
5.3	Response Quality	37
5.4	Retrieval vs Generative Models	41
5.5	C1 vs C2	41
5.5.1	Number of suggestions	42

6 Discussion and Conclusion	47
6.1 Limitations	49
Bibliography	51
Appendices	61
Appendix A Tools Used	62
A.1 Amazon Mechanical Turk	62
A.2 ParlAI	63
Appendix B User Study Setup	66
Appendix C Judgement Task Setup	71
Appendix D User Study Instructions	73
Appendix E Results	74
Appendix F Examples from User Studies	79
Appendix G Sample Conversation and Suggestions	82

List of Figures

3.1	C1 before clicking the “Show Suggestions” button.	17
3.2	C1 after clicking the “Show Suggestions” button.	18
3.3	C2 before clicking the “Show Suggestions” button.	18
3.4	C2 after clicking the “Show Suggestions” button.	19
3.5	Judgement task to rate user responses.	20
3.6	Judgement task options to rate user responses.	22
3.7	Verification task for judgement.	23
3.8	Judgement task to rate user responses.	25
4.1	Time and keystrokes needed for completing a response in human-only (H) and collaborative (C1 and C2) settings.	28
4.2	Response length characteristic in human-only (H), collaborative (C1 and C2), and chatbot-only settings.	29
4.3	Informativeness of responses in human-only (H), collaborative (C1 and C2), and chatbot-only settings.	30
4.4	Fluency and humanness of responses in human-only (H), collaborative (C1 and C2), and chatbot-only settings.	31
5.1	Average time taken to finish a response.	35

5.2	Average number of keystrokes for a response.	35
5.3	Average time taken to finish a response in collaborative settings.	36
5.4	Average number of keystrokes to finish a response in collaborative settings.	37
5.5	Average response length.	38
5.6	Average number of conversations.	38
5.7	Average response length with Wikipedia Sentences Retriever.	39
5.8	Average response length with Memory Network Retriever with original response.	40
5.9	Average response length with Memory Network Retriever without the original response.	41
5.10	Average response length with Generative Transformer Memory Network.	42
5.11	Average number of conversations in collaborative settings.	43
5.12	Response correctness.	44
5.13	Response usefulness.	44
5.14	Response fluency.	44
5.15	Response realness.	45
5.16	Response interestingness.	45
5.17	Response logicalness.	45
5.18	Overall response quality.	46
A.1	Amazon Mechanical Turk worker task portal.	63
A.2	The observation/action dict in ParlAI.	65

B.1	User study preview page.	69
B.2	Instructions for C1 setting.	70
C.1	Basic details about the judgement task shown to the workers.	71
D.1	Instructions for setting with no suggestions.	73
D.2	Instructions for C1 setting.	73
D.3	Instructions for C2 setting.	73
E.1	Average response usefulness in Settings 2 and 3.	74
E.2	Average response usefulness in Settings 4 and 5.	74
E.3	Average response correctness in Settings 2 and 3.	75
E.4	Average response correctness in Settings 4 and 5.	75
E.5	Average response fluency in Settings 2 and 3.	75
E.6	Average response fluency in Settings 4 and 5.	76
E.7	Average response humanness in Settings 2 and 3.	76
E.8	Average response humanness in Settings 4 and 5.	76
E.9	Average response interestingness in Settings 2 and 3.	77
E.10	Average response interestingness in Settings 4 and 5.	77
E.11	Average response logicalness in Settings 2 and 3.	77
E.12	Average response logicalness in Settings 4 and 5.	78
E.13	Average overall response quality in Settings 2 and 3.	78

E.14 Average overall response quality in Settings 4 and 5. 78

List of Tables

3.1	Human-chatbot collaborative conversation systems.	10
3.2	Sample conversation contexts and suggestions from Wikipedia retriever.	16
3.3	Task settings for user study with retrieval and generative models.	20
4.1	Examples of collected responses from the user study	27
6.1	Results for Setting 2.	49
6.2	Results for Setting 3.	49
F.1	Examples of collected responses from the user study in Setting 2.	80
F.2	Examples of collected responses from the user study in Setting 3.	80
F.3	Examples of collected responses from the user study in Setting 4.	81
F.4	Examples of collected responses from the user study in Setting 5.	81
G.1	Sample conversation contexts and suggestions for Wikipedia Retrieval.	83
G.2	Sample conversation contexts and ranked candidate responses from Wikipedia (M1)	84
G.3	Sample conversation contexts and candidate responses from Retrieval Trans- former Memory Network	85

G.4 Sample conversation contexts and candidate responses from Generative Transformer Memory Network	86
---	----

List of Abbreviations

AI Artificial Intelligence

API Application Programming Interface

HIT Human Intelligence Task

MTurk Amazon Mechanical Turk

NLP Natural Language Processing

SDK Software Development Kit

UI User Interface

Chapter 1

Introduction

In the past decade, Artificial Intelligence (AI) systems have become remarkably interactive and accurate with they becoming an integral part of our life. The increasing use of personal assistants like Siri and the application of AI in important real-world tasks such as medical imaging and diagnosis show that AI can perform as good as trained human experts in an experimental setting ([McKinney et al., 2020](#)). With the establishment of consortiums like The Partnership on AI to Benefit People and Society (PAI) ([Heer, 2018](#)), these developments have emerged at a time when the traditional view of AI of replacement of humans with machines is evolving to recognize the effect of their symbiosis. This suggests that human society is entering an era where humans and AI will increasingly work together and the collaboration of human-AI teams will become a lot more commonplace. Hybrid systems that enable human-AI collaboration aims to create an ensemble by combining their complementary strengths that behave more intelligently than human and AI individually.

Humans and AI chatbots have different but complementing skills, and AI can help humans expand their abilities in various ways. Since AI systems are wizards at assembling data from gigabytes of past data and topics, such a collaborative workflow helps free humans from the time-consuming effort of scanning through large repositories to find the correct information quickly to make informed decisions. By broadening the information to different areas, it can help us make creative actions. A system that enables us to collaborate with AI can provide us with new ways to interact with the world. Consider the AI agent Cortana

which can transcribe a meeting and provide a voice searchable version of the same to all the attendees. Similarly, such agents can parallelly interact with several customers to provide routine service to them.

In this thesis, we study a representative human-AI collaboration paradigm where both humans and AI can perform the same task with their advantages and drawbacks, while a collaboration between the two may complement each other. We examine a particular application area for human-AI collaboration that is of interest to the Information Retrieval (IR) and Natural Language Processing (NLP) communities—online text-based conversation systems, where a human user works with a chatbot to address other people’s requests conversationally. This application problem has many potential use cases in both commercial and personal settings. For example, e-commerce companies can use collaborative conversation systems to assist a large volume of customers online on Black Friday to augment customer service management.

Today’s fast-paced and competitive businesses need to keep their organization running at a record pace. One of the key challenging areas for a large enterprise is supporting customer queries. Organizations need solutions that can scale and manage millions of users and their queries concurrently. In a commercial customer service setting, a human agent in collaboration with AI chatbot can enable a far higher level of productivity, greater accuracy, and delivering more value within the customer experience. An individual user can also benefit from the system we have studied to reduce the effort of personal communications through messaging apps and emails. AI can take the whole conversation context into account and suggest humans the next responses to continue the conversation. It can provide ideas that are related to the conversation to open new avenues by increasing the breadth of the topics discussed.

The rest of the thesis introduces the design and results of multiple studies for human-chatbot

collaborative conversation systems. We have designed two collaborative systems each with a variety of configurations allowing human users to take advantage of chatbot responses to formulate their answers in a conversation. We used multiple retrieval-based and generative machine learning models to provide suggestions to users based on the conversation context in crowdsourced user studies. We finally compared the responses from collaborative systems in different settings and with human-only and chatbot-only baselines in terms of productivity, efficiency, and various response quality attribute through another crowdsourcing study.

We find that our collaborative systems can significantly improve the speed and reduce the effort to formulate a response and improve its quality compared with a human-only system, though they might hurt the fluency and humanness of the responses. Compared with a chatbot-only system, the collaborative systems can provide responses that are more fluent, human-like, and informative.

Chapter 2

Related Work

2.1 Conversational Systems

Advances in NLP and AI has driven the proliferation of a variety of conversational systems (Luger and Sellen, 2016; Myers et al., 2018). While task completion chatbot systems suggest the next actions based on the current conversation context derived from message history (Zhao and Eskenazi, 2016), non-task-oriented chatbot systems aim to converse with humans in a natural and meaningful way on a variety of open domain topics (Ritter et al., 2011). To build such non-task-oriented chatbots, various retrieval and generation-based methods have been proposed.

Retrieval-based chatbots use some type heuristic to pick a candidate response from a repository of responses based on the current conversation context. These systems just pick a response from a fixed set and do not generate any new text. Selection from a pool of human responses leads them to generally produce informative and fluent responses. While work in the past focused on building retrieval-based methods employing only the last input message to suggest responses for single-turn conversations (Wang et al., 2013), recent studies have focused on multi-turn response selection that takes in account and process all the previous utterances to provide a response relevant to the context of the conversation. Lowe et al. (2015) matched a response with the literal concatenation of context utterances. Yan et al. (2016) used a deep neural network to conduct matching after concatenating utterances with

the input message as reformulated queries. [Wu et al. \(2019\)](#) proposed a new context-based matching model for selection of responses in multi-turn conversation.

Unlike retrieval-based methods that reuse old messages and can suggest multiple responses, generation-based methods construct new responses from scratch. The LSTM sequence-to-sequence model ([Sutskever et al., 2014](#)) is a neural generation model that given the previous dialogue context maximizes the probability of generating a response. [Manning and Eric \(2017\)](#) proposed a copy-augmented sequence-to-sequence architecture. The decoding is often quite brittle, as errors may accumulate over time. To address this issue, [Ranzato et al. \(2015\)](#) proposed a sequence level training algorithm that directly optimizes metric used at test time (such as BLEU or ROUGE). [Li et al. \(2016\)](#) introduced a neural reinforcement learning (RL) generation method that simulates a conversation between virtual agents and reward sequences that are informative, coherent, and are easy to answer.

Since the set of response candidates provided by retrieval-based models are based on a fixed repository of past utterances, they might behave less flexible than generation-based models. In contrast, generation-based methods sometimes produce universal and non-relevant responses such as “I don’t know” and may contain grammatical mistakes. The human-AI collaboration paradigm we are studying in this thesis can be looked at from the lens of both response retrieval and generation problems. To suggest response candidates for a conversation, either human responses from past conversations can be picked or the suggestion can be generated from scratch. We conduct user study experiments to evaluate the collaborative systems employing generation-based and a variety of retrieval-based models based on their ability to improve human productivity and response quality using various metrics.

2.2 Human-AI Interaction

For decades, research has focused on various guidelines and recommendations to improve the interaction between humans and AI systems for various scenarios. Norman (1994) and Höök (2000) provided guidelines to restrict unwanted actions by intelligent systems by keeping a check on their levels of autonomy. To prevent users from getting frustrated by unpredictable AI agents Jameson (2002) suggested managing the expectations at earlier stages. Recently works have focused on increasing the transparency of AI systems (Kulesza et al., 2015; Lim and Dey, 2009; Rader et al., 2018) and offer personalized Human-AI interfaces (Deuschel, 2018; Findlater and Mcgreneire, 2004). Amershi et al. (2019) synthesized and validated the relevance and applicability of a unified set of design guidelines in AI systems. Considering the growth of advanced AI technologies collaborating with humans in a variety of applications, there is a need for innovation in system designs to make these applications effective.

Hybrid Intelligence aims to create an ensemble by combining the complementary strengths of human intelligence and AI that behaves more intelligent than they both can in separation (Kamar, 2016; Lasecki et al.). There is a need for interdisciplinary research on the role of humans in the AI conversational systems pipeline. Dellermann et al. (2019) conceptualizes hybrid intelligence systems and the corresponding dimensions for the design of such systems and provides guidelines for the implementation of such applications. Many human facing applications in industry today have an AI system embedded in them which has led to significant growth in designing UI to interact with these systems (Amershi et al., 2014; Doshi-Velez and Kim, 2017). Researchers have proposed many novel interfaces for these interactions but many of them tend to require specific experts (Amershi et al., 2015; Krause et al., 2016). Thus, there is a need for an interactive human-AI collaboration system that is seamless, natural, and efficient for a diverse set of applications.

To enable this collaboration research has been focusing on different aspects like the interactive impact of actions performed in a shared context (Liang et al., 2019). Also, there have been studies to understand the trade-offs between performance gains and compatibility with human mental models (Bansal et al., 2019). Studies show that understanding and adapting to human intentions have a positive impact on users making the interactions effective (Dragan and Srinivasa, 2013). Additionally, understanding the mechanisms behind human decision making has been shown to yield performance improvements and positive gains in collaborative tasks (Nikolaidis and Shah, 2013; Nikolaidis et al., 2015). Be it chatbots supporting collaborative learning (Tegos et al., 2011) or complex intelligent agents employed in industrial use, it is important to make design decisions for the chatting behavior of the agents. Various chatbot configurations are popular in practice. Some agents may change dynamically with the environment or help only on invocation. It has been seen that little details of the conversational behavior of AI chatbots may have a remarkable effect on its users (Wang et al., 2014; Al-Natour and Benbasat, 2009).

2.3 Search Engine Query Suggestion and Auto Completion

The design of our collaborative systems is illuminated by query suggestions and auto-completion in search engines. Similar to predictive search and related search queries shown in popular search engines today, we show provide candidate responses based on a conversation context. Such suggestions can make the human more informative with possibilities of improving the breadth of the topics covered in the conversation.

Search engine users can quickly find the required content when related queries are shown to

them. An example of such an application is related search keywords at the end of the result page in many popular search engines today. When the users search using a particular query, such related keywords provide a comprehensive recommendation to them and helps them quickly find the desired information. When users search the web for the desired information, they gradually refine their queries to reach the content. This logged information can be useful to provide recommendations to users who are not very familiar with the domain. Query logs containing the clicked URLs have been used in query recommendation ([Thirumalai et al., 2017](#)). [Fonseca et al. \(2003\)](#) proposed a novel method based on association rules to discover the related queries. [Feuer et al. \(2007\)](#) performed a controlled study using two search engines without phrase suggestions and one with it. They report that when the performance using the same search queries was compared, the engine with phrase suggestions had better precision and recall.

To get to the required information quickly even without less knowledge and effort, auto-completion is a common mechanism. Many popular web and desktop search engines like Google and Apple Spotlight have an integrated auto-completion system. Early approaches employed probability-based ranking based on the occurrence of queries in a background corpus ([Bhatia et al., 2011](#)). [Bast and Weber \(2006\)](#) proposed an auto-completion feature providing feedback on prefixes on their ability to fetch relevant documents. [Lev Finkelstein and Ruppin \(2002\)](#) used contextual information to select promising extensions for a query. [Di Santo et al. \(2015\)](#) emphasized that personalization to users outperform approaches that only consider past queries across multiple users in aggregate.

Chapter 3

Collaborative Conversation Systems

3.1 Collaborative Systems C1 and C2

We have designed two human-chatbot collaborative conversation systems (C1 and C2) as shown in Table 3.1. The systems provide users with a list of top-ranked chatbot responses as suggestions when they write their responses in a conversation. This collaboration design is illuminated by query suggestions and query auto-completion in search engines. Although both systems allow users to use the ideas provided by the chatbot suggestions to craft the conversation response, they differ significantly as they might encourage the user to write a response from scratch or directly edit the AI response.

C2 encourages workers to edit chatbot suggestions to formulate their responses because it presets the text field using the top-ranked chatbot result. Thus, this design “forces” users to examine chatbot suggestions, costing users more effort to write a response that is largely different from the preset content compared with an empty text box (as they might need to remove most of the preset text). On the other hand, C1 encourages workers to write responses by themselves because it provides the users with AI suggestions, but do not automatically copy the suggestion to the text field.

Previous studies suggest that both techniques can help search engine users by reducing the effort to input a query and providing ideas to formulate a query. We expect chatbot response

suggestions to help users similarly. For example, a user may simply make a few changes to a chatbot’s response as the final reply message, and chatbot results may provide users with the ideas and relevant information to write their responses too.

Table 3.1: Human-chatbot collaborative conversation systems.

	C1	C2
Cost to write own response	Low	High
Cost to reply by editing chatbot suggestions	High	Low

3.2 Retrieval and Generation-Based Chatbots Used

We have conducted multiple user studies followed by response quality judgment experiments on Amazon MTurk introduced in Appendix A.1. We used the Wizard of Wikipedia dataset (Dinan et al., 2019) where Wikipedia articles were treated as a knowledge source to assist in the task of question answering (QA). The dataset has 1,365 natural dialogue topics, each linked to their corresponding Wikipedia article with a total of 22,311 conversations and 201,999 dialogues. We start with a pilot user study described in Section 3.3 followed by the final large scale user study described in Section 3.4.

In the pilot user study, we used the Document Retriever proposed by Chen et al. (2017) for the Open-SQuAD dataset to find the articles relevant to the current conversation. With every lookup, we retrieved the first paragraph of the top 3 articles and then chose the top 2 sentences from each article that could be used as suggestions according to the setting in the Amazon Mechanical Turk task. A collaborative system employing such knowledge retrieval from Wikipedia can be used to assist humans in conversations (such as live customer support or emails) that aims to provide information related to domains in which the human participant might not sound.

For the final user study, we used the dialogue models provided by ParlAI (Dinan et al., 2019) introduced in Appendix A.2. These models have access to Wikipedia as a knowledge source.

We used the pre-trained models in various flavours in our experiments:

- **M1:** *retrieval model to produce ranked candidate responses from Wikipedia.*

This model can help in a use-case similar to model the pilot user study (assisting humans in conversations that aims to provide information) but instead of ordering the sentences by topic (in the pilot study), it ranks the responses using the knowledge selection process described below. Although the suggestions provided by M1 can be informative, they can lead to responses that are not fluent and human-like.

- **M2:** *retrieval model that selects candidate responses from human dialogues in the training set.*

By including the human responses for the conversation from the dataset, it corresponds to the use case where the same information is requested repeatedly (consider a customer service setting, where some topics are talked about frequently). M2 leads to response suggestions that are fluent and human-like, also containing the relevant information.

- **M3:** *retrieval model that selects candidate responses from human dialogues in the training set (like M2) but excludes the top suggestion in M2.*

By removing the actual conversation response, M3 corresponds to a scenario where there are very few occurrences of the same conversation in the past. M3 leads to response suggestions that are fluent and human-like but might not contain the relevant information.

- **M4:** *generation-based model that generates the next utterance word-by-word.*

M4 mimics a chatbot that generates the responses given the current dialog context. It does not pick a past human response but learns from the historical conversations to create a new response suggestion. Such responses can be less fluent and human-like than the retrieval models discussed before.

At each dialogue turn, the models were given the current dialogue context and the initial starting topic of the conversation (like Gardening). The first step is knowledge retrieval. The retriever (Dinan et al., 2019) uses an “inverted index lookup followed by term vector model scoring”. The current conversation context and articles from Wikipedia are compared as n-gram vectors and TF-IDF weighted bag-of-words. We followed the hashing proposed by Weinberger et al. (2009) to map the bigrams to 2^{24} bins with an unsigned murmur3 hash. The retriever is configured to operate on the last two conversation dialogue turns and the top 7 articles (first paragraph only) from Wikipedia are retrieved for each lookup and finally, separate sentences are formed from them.

To get the model output M1 and select the best knowledge sentences, we used an attention mechanism. For M2 and M3, the Transformer encoder proposed by Vaswani et al. (2017) (explained in Section 3.2.1) was used by encoding sentences in the memory and the dialogue context. Finally, standard dot-product attention between the dialogue context and the memory candidates was performed. For M4, we used the generative transformer memory network described in Section 3.2.2.

3.2.1 Retrieval Transformer Memory Network

Each knowledge sentence retrieved from Wikipedia $know_1, \dots, know_K$ in response to an input conversation context c is encoded with a Transformer explained above. To calculate the final input encoding, dot-product attention over: $enc(know_1), \dots, enc(know_K)$ is

performed and the resulting weighted sum of these vectors is added to $\text{enc}(c)$ to get the representation $r_{LHS}(know_1, \dots, know_K, c)$. The candidate responses $resp_1, \dots, resp_L$ are encoded with a separate Transformer to get $r_{RHS}(resp_i)$ for each i . We followed [Dinan et al. \(2019\)](#) to select the response $resp_l$ such that:

$$l = \arg \max_{i \in \{1, \dots, L\}} \frac{r_{LHS}(know_1, \dots, know_K, c)}{\|r_{LHS}(know_1, \dots, know_K, c)\|_2} \bullet \frac{r_{RHS}(resp_i)}{\|r_{RHS}(resp_i)\|_2}$$

The model is trained to minimize the cross-entropy loss.

3.2.2 Generative Transformer Memory Network

This model finds relevant knowledge sentence $know_{best}$, followed by an encoding step by concatenating it with the dialogue context. The decoder is allowed to attend over both the knowledge and dialogue to formulate the response. A beam search of 5 is employed to select the best response. Also, a shared Transformer encoder is used to encode all candidates $know_i$ and the dialogue history. Flattening of the encoded candidates into vectors is done using the normalization to produce an attention prediction over the memory. Finally, the full sequence encoding of the highest selected knowledge $know_{best}$ is concatenated with the encoding as proposed by [Dinan et al. \(2019\)](#).

The model is trained to minimize the negative log-likelihood of the response utterance. Additionally, supervision is added by forcing the knowledge selection to correctly choose knowledge candidate as the humans in the training set by adding cross-entropy loss over the knowledge attention, modulated by a weight λ :

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{NLL} + \lambda\mathcal{L}_{knowledge}$$

3.3 Experimental Design for the Pilot Study

To verify the design of the collaborative systems, we designed a crowdsourcing user study to evaluate the C1 and C2 systems introduced in Section 3.1 and compared them with human-only and chatbot-only baselines. Our study used a between-subjects design. We assigned each participant to use one of the following seven experimental systems:

- **H:** a baseline human-only system where users are only provided with a text box to input their replies without the help of a chatbot.
- **C1:** three variants of C1 providing one, three, or five chatbot top-ranked responses as suggestions.
- **C2:** three variants of C2 providing one, three, or five chatbot top-ranked responses as suggestions.

We created a conversation task pool based on the Wizard of Wikipedia dataset (Dinan et al., 2019). The pool included 90 conversations with one, two, or three rounds of existing utterances (30 for each case). For each conversation task, the participant needed to respond according to the previous utterances. The system did NOT further reply to the participant’s message (which would require either real human or human-like chatbots at the backend). We randomly sampled conversations from the dataset and manually removed those that were chit-chat or required extensive domain knowledge.

3.3.1 User Study Design

Table 3.2 shows sample conversations and suggestions for different context lengths. We required each participant to finish an experiment session of five minutes (we did not count

the time spent on instructions and a training task at the beginning of the experiment). The participants finished the conversation tasks randomly sampled from the pool (without replacement) one after another until five minutes. We particularly instructed them to provide informative responses instead of simple and short replies such as “Yes/No” and “I don’t know”. We paid each HIT (a five-minute experiment session) \$0.25 and provided an additional \$0.25 bonus for the top 10% performing HITs.

After accepting the HIT and completing the sample task, the system started a 5 minutes timer to complete the actual task. The Turker was asked to complete as many conversations as possible in this timeframe while maintaining the quality of the responses. C1 setting for this user study provided a “Show Suggestions” button below the input box for entering a reply message. It did not show chatbot suggestions until users click on the “Show Suggestions” button. After showing the suggestions, clicking on a suggestion appended its content to the current text in the input box. Users could make further edits to the auto-copied content and select other suggestions. Figures 3.1 and 3.2 respectively show screenshots of C1 before and after clicking the “Show Suggestions” button.

The C2 system had the same functionality as C1, but C2 presets the input box to the top-ranked chatbot response. Figures 3.3 and 3.4 respectively show screenshots of C2 before and after clicking the “Show Suggestions” button.

3.3.2 Response Quality Judgments

We evaluated response quality by informativeness, fluency, and humanness and collect human judgments on Amazon Mechanical Turk. Each judgment HIT page included the conversation and response to be judged (we had highlighted the response text) and several judgment questions. We paid each judgment HIT \$0.04. To improve the accuracy of our results we

Table 3.2: Sample conversation contexts and suggestions from Wikipedia retriever.

Context with one round of conversation:**Person1:** When walking through Walmart I always see so much fishing tackle and equipment!**Suggestions:**

Fishing tackle is the equipment used by fishermen when fishing.

Almost any equipment or gear used for fishing can be called fishing tackle.

Recreational fishing, also called sport fishing, is fishing for pleasure or competition.

It can be contrasted with commercial fishing, which is fishing for profit, or subsistence...

Fly fishing is an angling method in which an artificial fly is used to catch fish.

Original Response:

You mean the equipment people use when they go fishing? That fishing tackle I happen to know a lot about! Do yo u ever fish?

Context with two rounds of conversation:**Person 1:** I have blond hair, what do you think of people with that hair color?**Person 2:** Sometimes I wish I had blonde hair, or fair hair. They have low levels of eumelanin.**Person 1:** What is eumelanin? I like it, but sometimes I need a change, plus it gets thick.**Suggestions:**

Hair color is the pigmentation of hair follicles due to two types of melanin: ...

Generally, if more eumelanin is present, the color of the hair is darker; if less ...

Melanin (; from melas, black, dark) is a broad term for a group of natural pigments...

Melanin is produced by the oxidation of the amino acid tyrosine, followed by...

Red hair (or ginger hair) occurs naturally in 1-2% of the human population.

Original Response:

Eumelanin is just a dark pigment that humans have to give hue to hair.

Context with three rounds of conversation:**Person 1:** What allergies do you have?**Person 2:** I do not have any allergies or allergic diseases.**Person 1:** Do you know anything about allergies?**Person 2:** Yes I know food allergies are due to abnormal immsume response to food.**Person 1:** What are some common allergies?**Suggestions:**

A food allergy is an abnormal immune response to food.

The signs and symptoms may range from mild to severe.

Peanut allergy is a type of food allergy to peanuts.

It is different from nut allergies.

Anaphylaxis is a serious allergic reaction that is rapid in onset and may cause death.

Original Response:

They include itchiness and swelling of the tongue.

Client: I would really like to enjoy more time out in the sun, but I'm worried I might get skin cancer. Any advice?

You: It definitely should be a concern. There's actually three types of it, with the basal-cell, squamous-cell, and melanoma types of skin cancer.

Client: Well I've heard of melanoma before, but what's going on with those other two types? Are those more serious?

Please enter here...

Show Suggestions

Figure 3.1: C1 before clicking the “Show Suggestions” button.

asked three Workers to provide answers for each response and gave them 15 minutes to complete each item. Also, we only allowed workers with the HIT approval rate above 95% to participate in our assignments. We decided to automatically approve work in three days since we did not manually want to approve or reject the work. The questions and options are adapted from previous studies ([See et al., 2019](#); [Zhang et al., 2018](#)):

- Does the highlighted response contain sufficient information relevant to the conversation?
 1. Very Insufficient
 2. Slightly Insufficient
 3. Slightly Sufficient
 4. Very Sufficient

- How interesting or boring did you find the highlighted response?
 1. Very Boring
 2. Slightly Boring
 3. Slightly Interesting
 4. Very Interesting

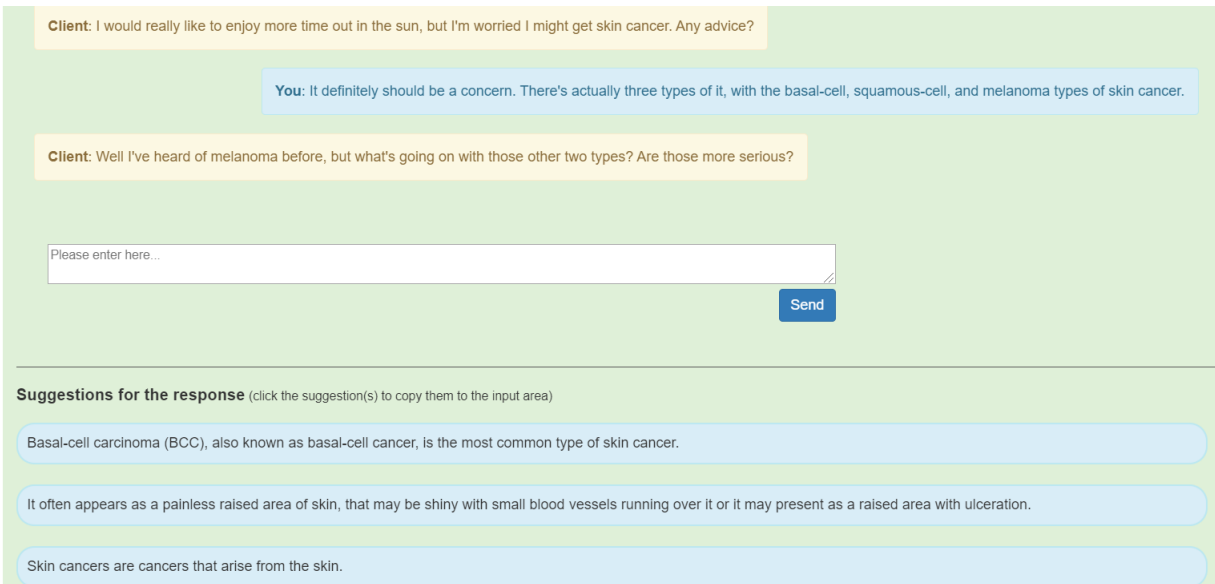


Figure 3.2: C1 after clicking the “Show Suggestions” button.

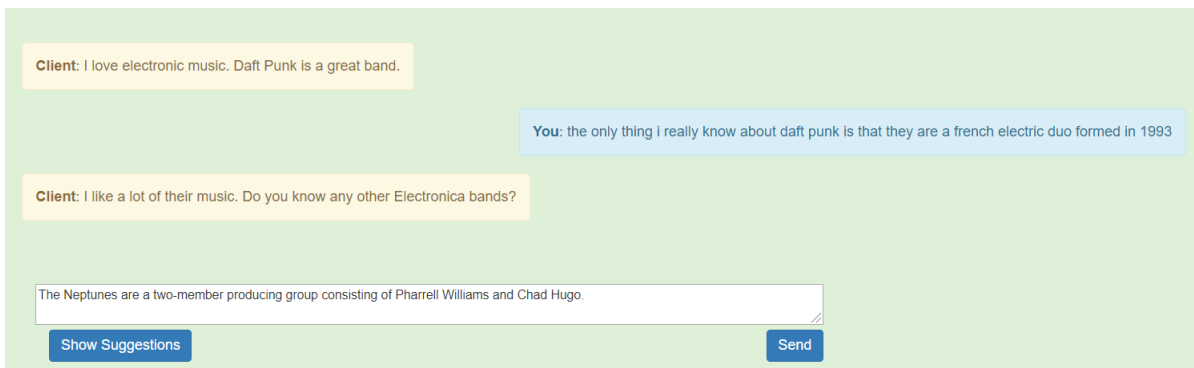


Figure 3.3: C2 before clicking the “Show Suggestions” button.

- How natural is the highlighted response in English?
 1. Very Unnatural
 2. Slightly Unnatural
 3. Slightly Natural
 4. Very Natural

- Do you think the highlighted response is provided by a bot or a human?
 1. Definitely a bot
 2. More likely a bot
 3. More likely a human
 4. Definitely a human

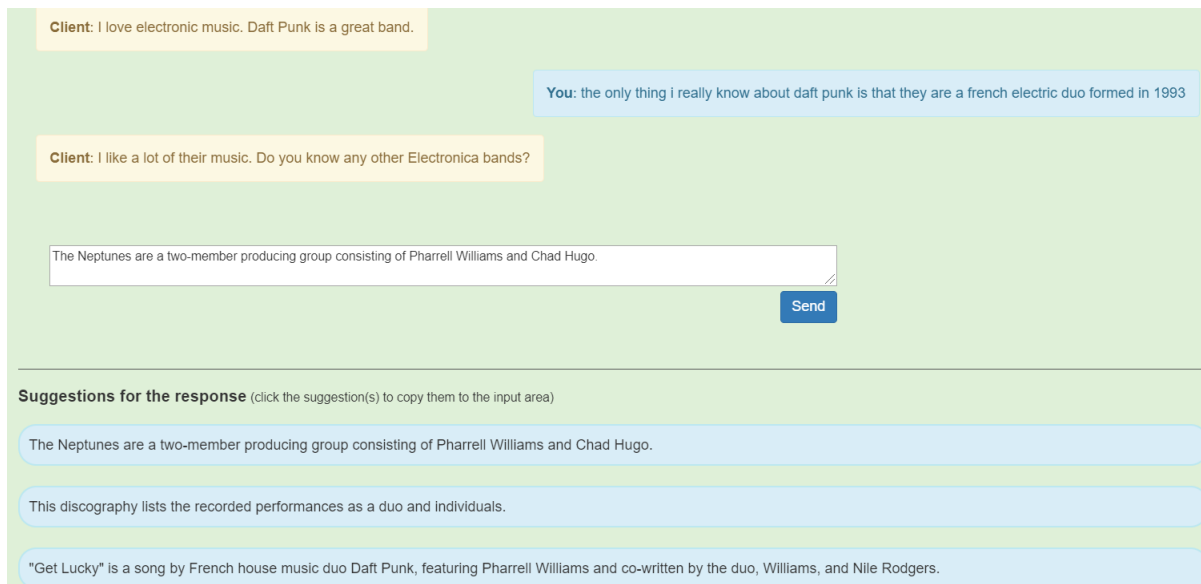


Figure 3.4: C2 after clicking the “Show Suggestions” button.

We judged the responses for all the complete conversations in H, C1, and C2 settings. We also used the chatbot for building C1 and C2 to generate responses for the 90 conversation tasks. We collected judgments for the chatbot responses in the same setting for comparison. For each conversation response, we used the mean ratings from three assessors as quality measures. Figure 3.5 shows the conversation in the task to rate user responses on MTurk. Chapter 4 presents the results and conclusions from the pilot study.

3.4 Experimental Design for the Final Study

After the pilot user study, we conducted a refined and large scale experiment built upon our experiences with the pilot study. We created a conversation task pool similar to the one described in Section 3.3 based on the Wizard of Wikipedia dataset (Dinan et al., 2019). Similar to the pilot user study, the pool included 90 conversations with one, two, or three rounds of existing utterances (30 for each case) and for each conversation task, the participant

Read the conversations between **Person 1** and **Person 2** carefully. Rate the next responses (**highlighted**) according to the criteria mentioned below and click **Submit** to complete the task.

Person 1: I grew up with 8 older siblings, pretty crazy right? I think it has something to do with a fertility factor, but I have no idea what that is really, can you explain it?

Person 2: Whoa 8!!! Fertility factor has to do with how likely a person is to have children.

Person 1: Oh I see, I guess my mother had quite a high fertility factor then.

Person 2: Yes she did. Fertility factors is based on a lot of things.

Person 1: Alright, can you elaborate on what some of those things are?

Person 2: some of the social factors that can influence fertility rates are: race, level of education, religion, use of contraceptive methods, abortion, impact of immigration

Figure 3.5: Judgement task to rate user responses.

needed to respond according to the previous utterances. We assigned each participant to use one of the following 21 experimental systems involving different combinations of models (M1, M2, M3, and M4) and experiment settings (C1 and C2) as shown in Table 3.3.

Table 3.3: Task settings for user study with retrieval and generative models.

Setting Name	Chatbot	Number of Suggestions	Configuration	Total Settings
Setting 1	Human-Only	-	-	1
Setting 2	M1	{1, 3, 5}	{C1, C2}	6
Setting 3	M2	{1, 3, 5}	{C1, C2}	6
Setting 4	M3	{1, 3, 5}	{C1, C2}	6
Setting 5	M4	1	{C1, C2}	2
				21

3.4.1 User Study Design

We required each participant to finish an experiment session of five minutes (do not count the time spent on instructions and a training task at the beginning of the experiment). Similar to the pilot user study, the participants finished the conversation tasks randomly sampled from the pool (without replacement) one after another until five minutes. We paid

each HIT (a five-minute experiment session) \$0.25 and provide an additional \$0.25 bonus for the top 10% performing HITs. After accepting the HIT and completing the sample task, the system started a 5 minutes timer to complete the actual task. The Turker was asked to complete as many conversations as possible in this timeframe while maintaining the quality of the responses. Unlike C1 setting for the pilot user study described in Section 3.3.1 we did not provide a “Show Suggestions” button but showed the suggestions by default as the user activity logs from the pilot study reflected that such a configuration to hide suggestions by default in C1 resulted to a lot of users not using the chatbot suggestions. Similar to the pilot study, the C2 system had the same functionality as C1, but C2 preset the input box with the top-ranked chatbot suggestion.

Also, during the HITs, we restricted the worker pool by setting the MTurk qualification *QualificationTypeId* with value *000000000000000000L0* which only allowed Turkers with prior acceptance rate above 95% to work on the task.

Appendix G shows sample conversations and suggestions for different context lengths in all the settings.

3.4.2 Response Quality Judgments

We evaluated response quality by 7 parameters: usefulness, correctness, reasonability, fluency, humanness, interestingness, and the overall quality. Each judgment HIT page included 5 conversations and responses to be judged (we have highlighted the response text) and judgment questions. Additionally, we randomly placed a verification conversation task in every HIT. We intentionally made the response to be judged in this verification conversation shown in Figure 3.7 absurd so that we could filter out the workers who gave high response quality ratings to this conversation.

The highlighted response provides **useful** information **relevant** to the conversation.

Strongly Disagree Disagree Neutral Agree Strongly Agree

The information provided by the highlighted response is **correct** and **accurate**.

Strongly Disagree Disagree Neutral Agree Strongly Agree

The highlighted response is **reasonable** and **logical** in its context.

Strongly Disagree Disagree Neutral Agree Strongly Agree

The highlighted response sounds like someone who speaks **fluent** and **natural** English.

Strongly Disagree Disagree Neutral Agree Strongly Agree

I believe the highlighted response is from **a real human** instead of a bot.

Strongly Disagree Disagree Neutral Agree Strongly Agree

The highlighted response reads **interesting** to me if I were person 1.

Strongly Disagree Disagree Neutral Agree Strongly Agree

How well would you rate the quality of the highlighted response if you were person 1?

1 (Very Poor) 2 (Poor) 3 (Okay) 4 (Good) 5 (Very Good)

Figure 3.6: Judgement task options to rate user responses.

We paid \$0.20 for each judgment HIT containing 6 conversations (5 user responses and 1 verification task). To improve the accuracy of our results we asked three Workers to provide answers for each response and give Workers 15 minutes to complete each item. Also, we only allowed workers with the HIT approval rate above 95% to participate in our assignments.

Similar to the quality judgement task for the pilot user study, the criteria and options were adapted from previous studies:

- The highlighted response provides useful information relevant to the conversation.

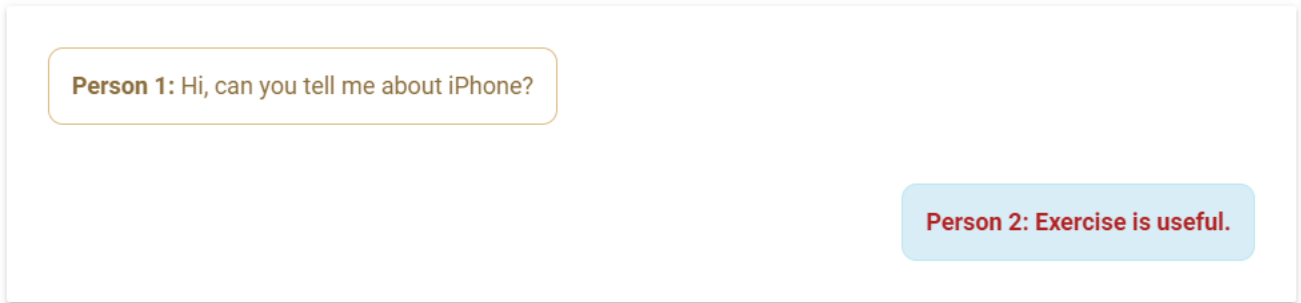


Figure 3.7: Verification task for judgement.

- | | | |
|----------------------|------------|-------------------|
| 1. Strongly Disagree | 3. Neutral | 5. Strongly Agree |
| 2. Disagree | 4. Agree | |

- The information provided by the highlighted response is correct and accurate.

- | | | |
|----------------------|------------|-------------------|
| 1. Strongly Disagree | 3. Neutral | 5. Strongly Agree |
| 2. Disagree | 4. Agree | |

- The highlighted response is reasonable and logical in its context.

- | | | |
|----------------------|------------|-------------------|
| 1. Strongly Disagree | 3. Neutral | 5. Strongly Agree |
| 2. Disagree | 4. Agree | |

- The highlighted response sounds like someone who speaks fluent and natural English.

- | | | |
|----------------------|------------|-------------------|
| 1. Strongly Disagree | 3. Neutral | 5. Strongly Agree |
| 2. Disagree | 4. Agree | |

- I believe the highlighted response is from a real human instead of a bot.

- | | | |
|----------------------|------------|-------------------|
| 1. Strongly Disagree | 3. Neutral | 5. Strongly Agree |
| 2. Disagree | 4. Agree | |

- The highlighted response reads interesting to me if I were person 1.

- | | | |
|----------------------|------------|-------------------|
| 1. Strongly Disagree | 3. Neutral | 5. Strongly Agree |
| 2. Disagree | 4. Agree | |

- How well would you rate the quality of the highlighted response if you were person 1?

- | | | |
|------------------|-------------|------------------|
| 1. 1 (Very Poor) | 3. 3 (Okay) | 5. 5 (Very Good) |
| 2. 2 (Poor) | 4. 4 (Good) | |

We only accepted the HITs meeting all of the following for the verification conversation (Figure 3.7):

1. The rating for criteria 1 is either ‘Strongly Disagree’ or ‘Disagree’
2. The rating for criteria 3 is either ‘Strongly Disagree’ or ‘Disagree’
3. The rating for criteria 7 is either ‘1 (Very Poor)’ or ‘2 (Poor)’

Overall, 14.7% of the HITs did not meet these criteria. Thus, we rejected them and did not pay these Turkers the \$0.20 MTurk fee.

We judged the responses for all the complete conversations in H, C1, and C2 settings. We also collected judgments for the chatbot responses in the same setting for comparison. For each conversation response, we collected judgments from three assessors and used their mean ratings as quality measures. Figures 3.8 and 3.6 shows the task and options respectively to rate user responses on MTurk.

Conversation 5

Person 1: Hi, I took a sewing class in middle school

Person 2: You mean the craft of fastening or attaching objects using stitches made with a needle and thread? That is cool.

Person 1: yes, I got to make my own pillow

Person 2: I like those Pillow Pets made from chenille.

Person 1: what is a chenille?

Person 2: A chenille is a velvet-like cord or yarn that people use to trimming furniture or make clothing or carpets.

Figure 3.8: Judgement task to rate user responses.

Chapter 4

Results from the Pilot Study

We collected 140 experiment sessions from the user study (20 for each of the seven systems) and judged the responses in the finished conversations. On average, an experiment session had 10.2 finished conversations.

In the judgment task, we judged each of the 1,159 user responses by 3 workers, ensuring that each worker works on unique tasks only. This resulted in a total of 3,477 assignments. Similarly, we created 1,272 assignments to judge the suggestions. Table 4.1 shows good and bad examples from the user study.

4.1 Response Time and Keystrokes

Figure 4.1 reports the time and keystrokes needed for completing a response in human-only (H) and collaborative (C1 and C2) settings. Compared with in a human-only system (H), participants in C1 and C2 used less time and fewer keystrokes to provide much longer responses when the systems only showed one suggestion (“1 sug.”). This suggests that the collaborative systems can improve the efficiency of users to formulate responses in a conversation. Our system has recorded an average of 60.2 keystrokes to finish a response in a human-only setting (H), compared with 37.0-39.8 in C1 ($p < 0.001$) and 15.9-32.4 in C2 ($p < 0.001$). Participants have spent an average of 25.6 seconds to finish a response, compared with 23.8 seconds in C1 (not significant at 0.05 level) and 14.8 in C2 ($p < 0.001$) when the

Table 4.1: Examples of collected responses from the user study

Conversation Context:

I like Anime, but i don't really know too much about it's history and techniques. I don't either, but I have read the manga series known as "Death Note". It has an interesting storyline. Have you heard of it?

No, I am new to Anime, nut Death Note sounds interesting. Who makes it?

Suggestions:

This is a list of the chapters of the Japanese manga series "Death Note", written by Tsugumi Ohba and illustrated by Takeshi Obata.

The individual chapters were originally serialized in Shueisha's "Weekly Shōnen Jump", from December 2003 to May 2006, with 108 chapters in all.

Death Note was first serialized in Shueisha's manga magazine Weekly Shōnen Jump from December 2003 to May 2006.

Good Response:

it is written by Tsugumi Ohba and illustrated by Takeshi Obata.

Bad Response:

I can picture the book but I can't remember the author's name! I will check and let you know.

systems showed only one suggestion (“1 sug.”).

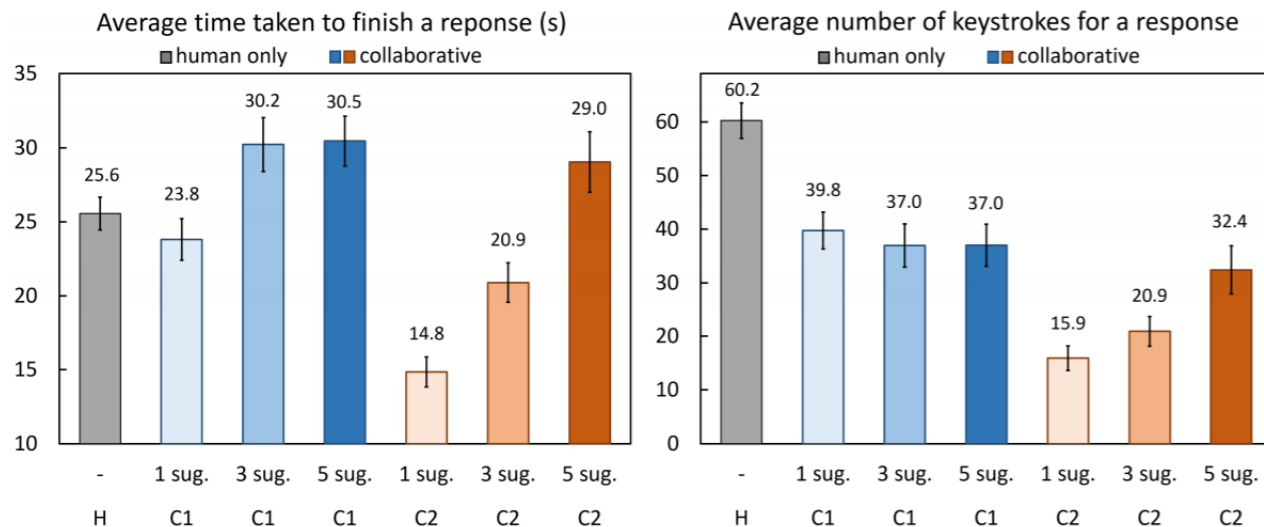


Figure 4.1: Time and keystrokes needed for completing a response in human-only (H) and collaborative (C1 and C2) settings.

4.2 Response Length

Figure 4.2 reports the response length characteristic in human-only (H), collaborative (C1 and C2), and chatbot-only settings. Participants’ responses in a human-only setting (H) included 52.4 characters on average, compared with 101.1-125.1 characters in C1 ($p < 0.001$) and 140.0-177.5 characters in C2 ($p < 0.001$).

4.3 Response Quality

Crowdsourcing judgments suggest that human-only responses (H) are the least informative but the most fluent and human-like, while chatbot-only responses are informative but received the lowest fluency and humanness ratings. C1 and C2 collaborative systems reach a

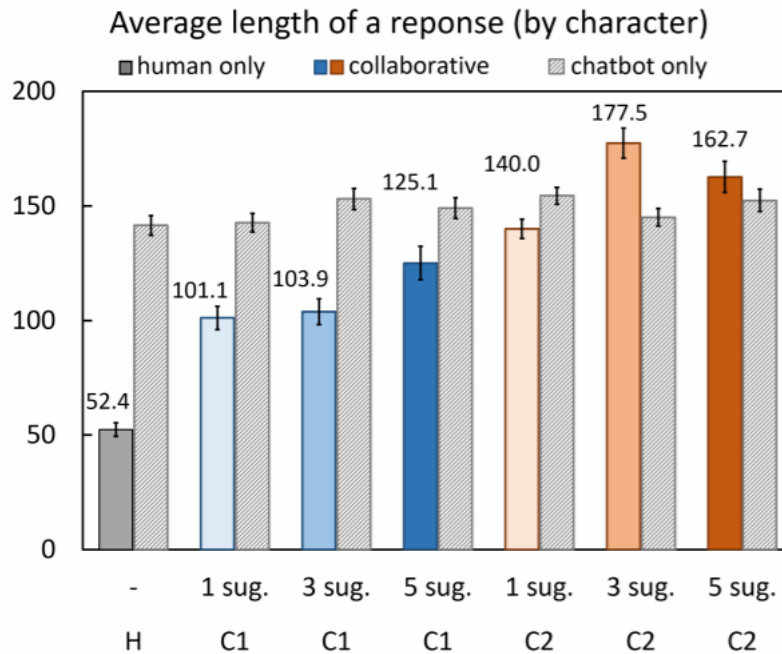


Figure 4.2: Response length characteristic in human-only (H), collaborative (C1 and C2), and chatbot-only settings.

balance between the two - their responses are as informative as chatbot-only responses but more fluent and human-like (though not as fluent and human-like as human-only responses).

4.3.1 Informativeness

Figure 4.3 reports informativeness of responses in human-only (H), collaborative (C1 and C2), and chatbot-only settings. According to the collected quality judgments, responses in both C1 and C2 settings were significantly more informative than human-only responses (H) and similarly informative to chatbot-only ones. Human-only responses received 2.72 informativeness ratings on average, which is statistically significantly lower than those in C1 (2.81-2.85 depending on the number of suggestions, $p < 0.05$) and C2 settings (2.81-2.89, $p < 0.05$). The human-only responses are also less informative than chatbot-only responses (2.72 vs. 2.78, $p = 0.076$), while C1 and C2 responses received informativeness ratings comparable

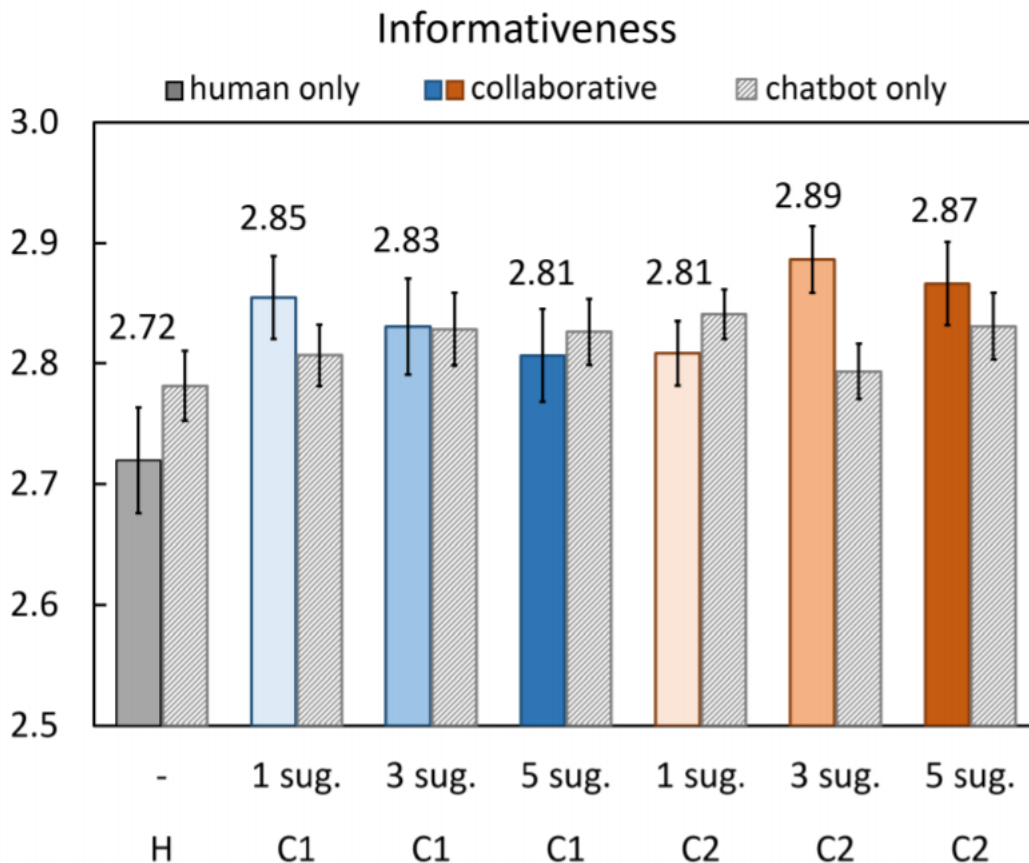


Figure 4.3: Informativeness of responses in human-only (H), collaborative (C1 and C2), and chatbot-only settings.

to chatbot-only responses.

4.3.2 Fluency and Humanness

Regarding the fluency and humanness of responses as reported in Figure 4.4, we observed that the human-only setting (H) received significantly higher ratings than both collaborative systems (C1 and C2) and the chatbot-only setting (the differences are all significant at 0.01 level). C1 responses are also significantly more fluent and human-like than chatbot-only responses (all differences are significant at 0.05 level), while C2 only outperformed chatbot in the humanness of responses.

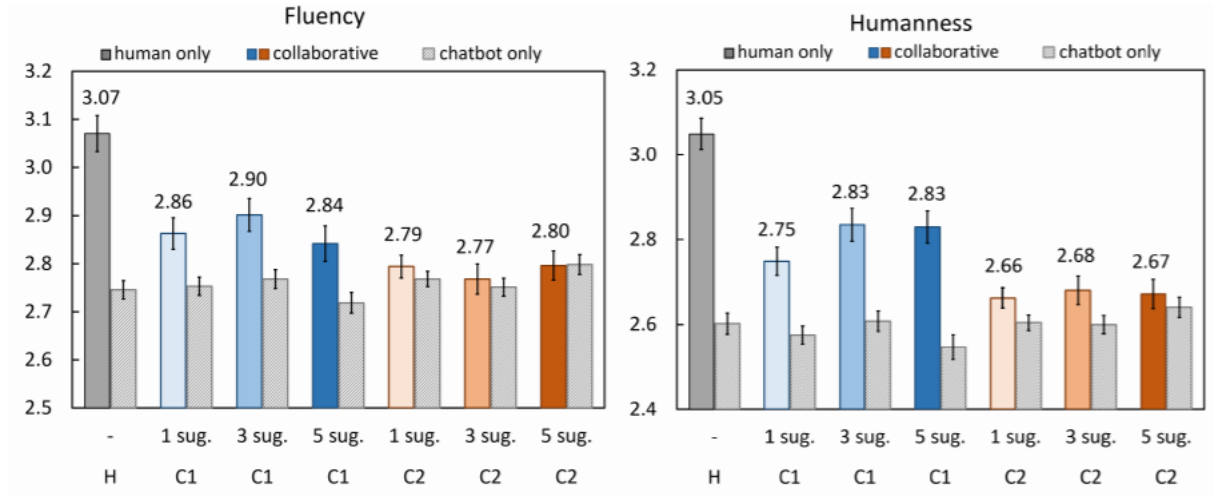


Figure 4.4: Fluency and humanness of responses in human-only (H), collaborative (C1 and C2), and chatbot-only settings.

4.4 Number of Chatbot Suggestions

We found that both C1 and C2 systems reach the ideal performance when showing only one chatbot suggestion. Providing more suggestions to users increased the time and keystrokes needed to finish a response but did not consistently improve or hurt response quality. For C1 setting, participants spent significantly longer time to finish a response when the system increased the number of suggestions from one (23.8 seconds) to three (30.2 seconds, $p < 0.01$ compared with one suggestion) and five (30.5 seconds, $p < 0.01$ compared with one suggestion). In contrast, the number of keystrokes needed to complete the response did not vary significantly regardless. For the C2 setting, increasing the number of chatbot suggestions from one to three and five has consistently increased both the time and keystrokes needed to finish a response. The time needed to finish a response increased from 14.8 seconds (one suggestion) to 20.9 (three) and 29.0 (five), where the differences are all statistically significant at 0.01 level. The number of keystrokes needed increased from 15.9 (one suggestion) to 20.9 (three) and 32.4 (five), where the differences are all statistically significant at 0.05 level.

4.5 C1 vs. C2

Comparing C1 and C2, the quality of C2's responses are more similar to chatbot-only responses. This is consistent with our expectation when designing the two systems, where we preset the text field to encourage users to synthesize chatbot results to a greater extent. This suggests our design has successfully met our purpose.

4.6 Conclusion and Discussion

Initial results state that chatbots might help human users improve the informativeness of their responses. Since our conversation tasks required certain knowledge to provide informative answers (though not requiring extensive domain expertise), we expect this finding to generalize to other conversation tasks with a similar goal. In contrast, both collaborative systems seem to reduce the fluency and humanness of responses compared with a human-only system. Also, our study shows that the use of simple design choices such as whether to preset the value of the input field or not might make the results more similar to those in a human-only or chatbot-only setting. This can provide an opportunity to balance response quality (informativeness vs. fluency and humanness) by the use of interface design. Further, the results indicate that it costs users a significant amount of effort (both time and keystrokes) to synthesize AI chatbot results with human wisdom.

After analyzing the results of the pilot user study we suspect that retrieval of Wikipedia sentences helps to create a collaborative system that can improve response quality and human efficiency. This gave us initial confidence in the feasibility of such systems to improve the conversations. But the results from the pilot user study can not be used to affirm the efficiency our collaborative settings C1 and C2 as:

- we did not filter the Turkers by their past HIT acceptance rate which might have made some our analysis inaccurate. Workers with high past performances generally follow the task instructions carefully and put in their best efforts.
- The use of the “Show Suggestions” button in C1 might have confused some workers preventing them from utilizing the chatbot suggestions.
- We only used a model that retrieves sentences from Wikipedia to act as response suggestions. We needed to diversify our model selection to claim some concrete findings.

Learning from our experiences from the pilot user study and to dive deeper into the results we conducted a final large scale user study after working on the limitations described above. Chapter 5 presents the results and conclusion for this final user study.

Chapter 5

Results from the Final User Study

We collected 1,050 experiment sessions from the user study (50 for each of the 21 systems) and judged the responses in the finished conversations. In the judgment task, we judge each of the total 5,139 unique user responses by 3 workers, ensuring that each worker works on unique tasks only. This resulted in a total of 15,417 assignments. Similarly, we created 276 assignments to judge the suggestions for chatbot-only baselines. Tables [F.1](#) - [F.4](#) shows good and bad examples from the user studies in different settings.

5.1 Response Time and Keystrokes

Figures [5.1](#) and [5.2](#) report the average time taken and keystrokes to finish a response in all the task settings introduced in Table [3.3](#). It took lesser time and fewer keystrokes to complete a conversation in collaborative settings 2-5 as compared to human-only setting 1. Since the experiments in the human-only setting took on an average 31.24 seconds as compared to 20.46-27.56 seconds ($p < 0.001$) in collaborative modes and 65.73 keystrokes as compared to 21.39-39.32 keystrokes ($p < 0.001$) in collaborative settings, it suggests that the collaborative systems can improve the productivity of users to formulate responses in a conversation.

Figures [5.3](#) and [5.4](#) shows the average time taken and keystrokes in C1 and C2 collaborative modes and compared them to the human-only baseline across settings 2-5. In general,

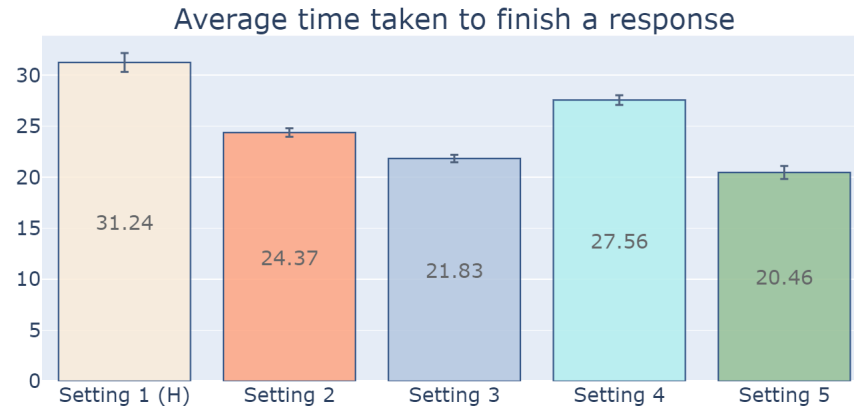


Figure 5.1: Average time taken to finish a response.

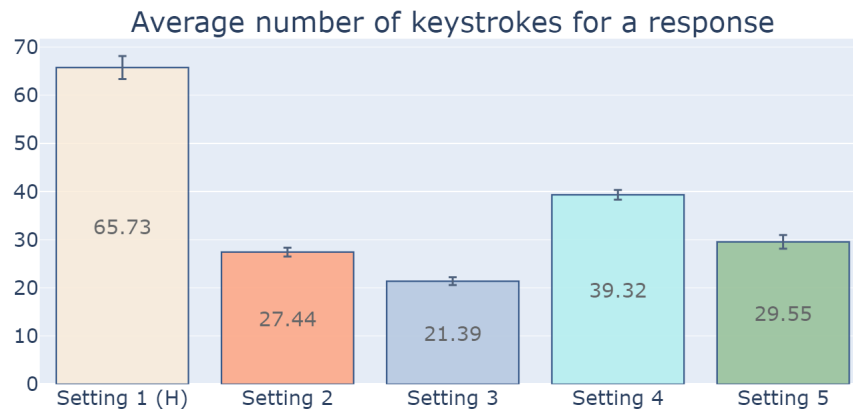


Figure 5.2: Average number of keystrokes for a response.

participants in collaborative settings were more efficient while finishing the conversations as compared to a human-only system (H), with responses in C2 taking lesser time and keystrokes than in C1. Thus, presetting the field with the best AI response improves the speed to formulate the response.

5.2 Response Length and Number of Conversations

Figures 5.5 and 5.6 report the average length of the responses and number of conversations in all the task settings. Participant's responses in human-only setting 1 on an average con-

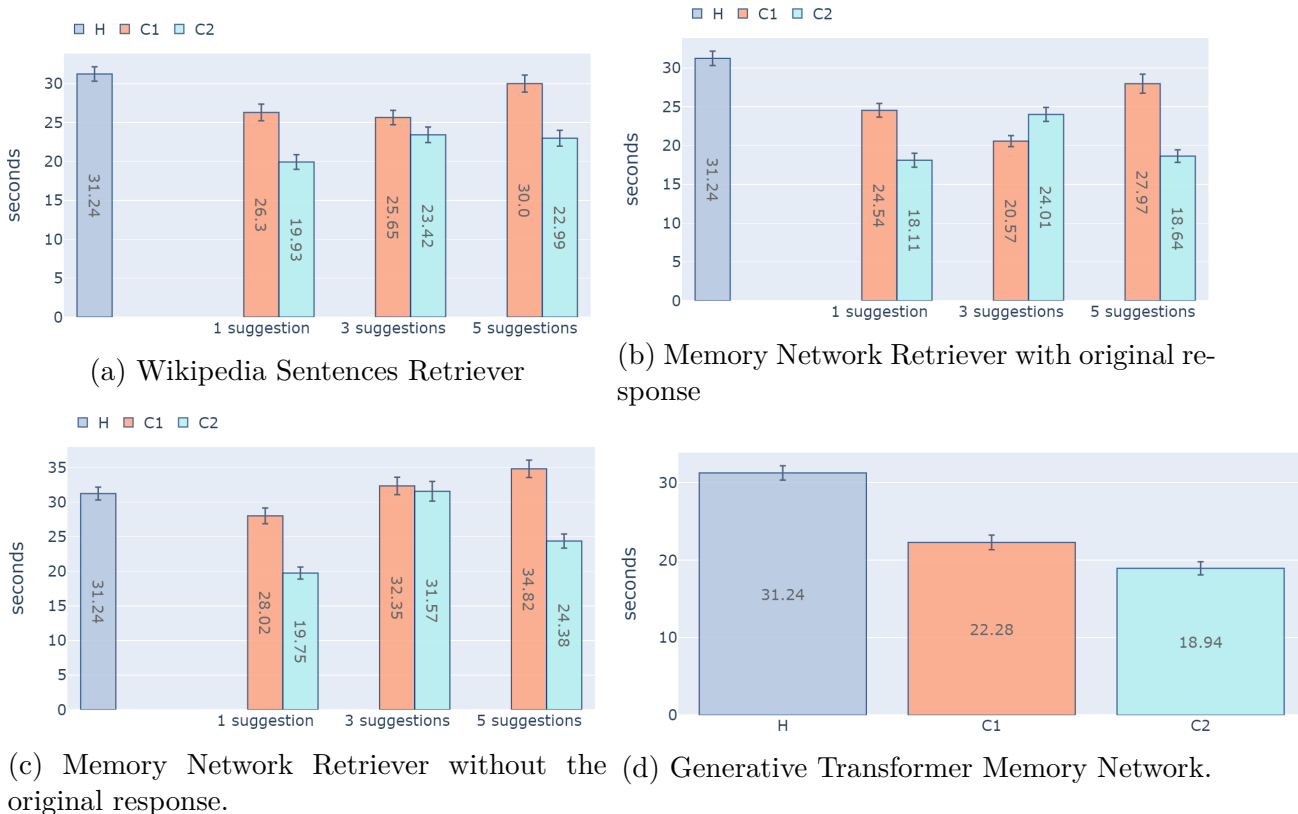
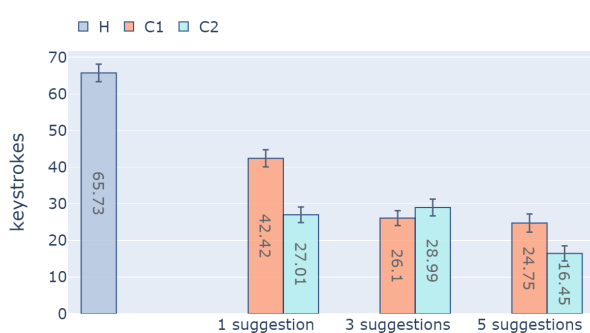
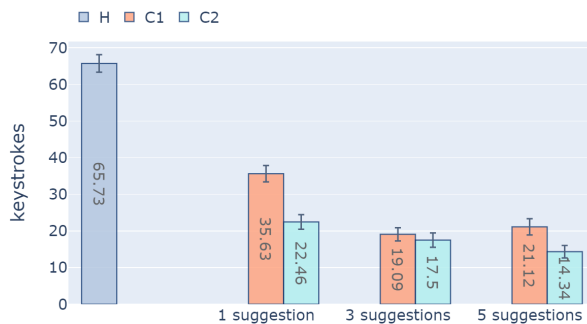


Figure 5.3: Average time taken to finish a response in collaborative settings.

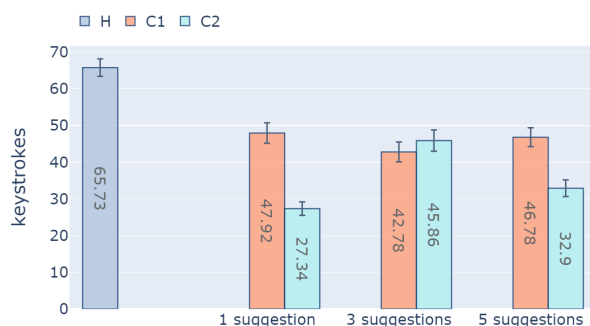
sists of 60.9 characters as compared to 75.49-145.24 ($p < 0.001$) characters in collaborative settings. It suggests that compared to a human-only system (H), participants in collaborative systems used less time and fewer keystrokes to provide much longer responses to more number conversations. Figures 5.7 - 5.10 show the average response lengths in C1 and C2 collaborative modes and compares them to the human-only and chatbot-only baselines across settings 2-5. Collaborative and chatbot-only responses on an average are longer than human-only baselines, with the responses in C2 being longer than those in C1. When 3 and 5 suggestions are shown, responses in C2 are longer than the corresponding chatbot-only baselines. In setting 2 (Memory Network Retriever with original response), although the number of keystrokes decreased as the number of suggestions increased, the time taken to finish a response does not change much. Figure 5.11 reports that in setting C2 users were



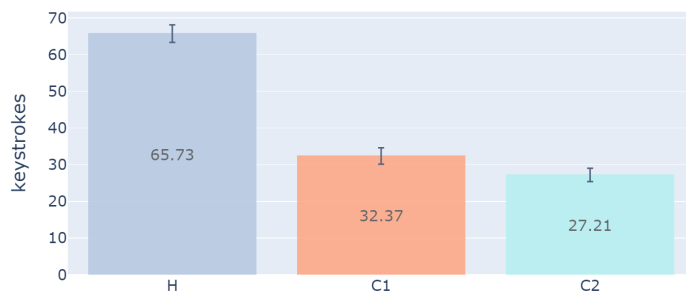
(a) Wikipedia Sentences Retriever



(b) Memory Network Retriever with original response



(c) Memory Network Retriever without the original response



(d) Generative Transformer Memory Network

Figure 5.4: Average number of keystrokes to finish a response in collaborative settings.

able to complete a higher number of conversations than in C1.

5.3 Response Quality

Figures 5.12 and 5.13 reports that with the average response correctness and usefulness ratings of 3.78 and 3.87 respectively for human-only setting 1, responses in setting 2 and 3 are more useful (with ratings between 3.9 - 3.92) and correct (with ratings between 3.96 - 3.99). Participants in the response judgment study reported that the responses in settings 4 and 5 are less useful (3.39 - 3.53) and correct (3.46 - 3.64) as compared to the human-only baseline.

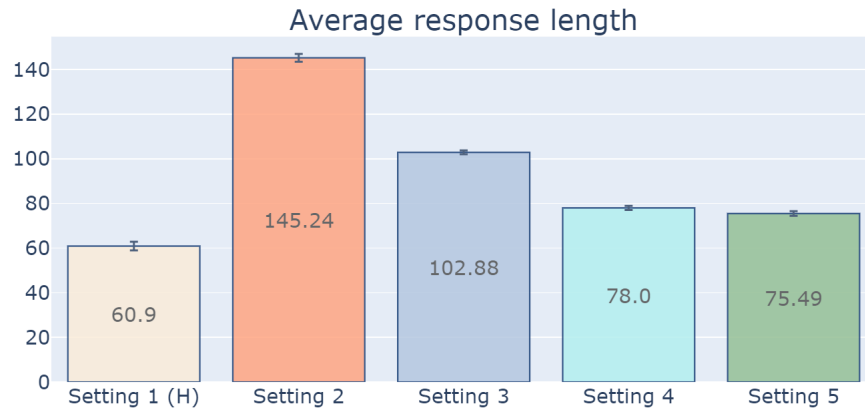


Figure 5.5: Average response length.

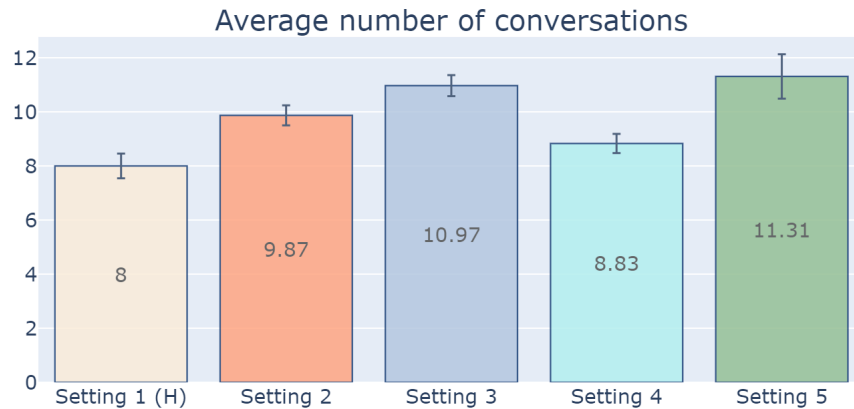


Figure 5.6: Average number of conversations.

Figures E.1 - E.4 reports the average response usefulness and correctness in C1 and C2 collaborative modes and compares them to the human-only and chatbot-only baselines across settings 2-5. While responses in setting 2 and 3 with collaborative mode C2 are reported to be more useful and correct than those in C1 and chatbot-only baselines, in settings 4 and 5 responses in C2 are less useful and correct than C1. Further, in settings 4 and 5, human-only responses are more useful and correct than C1 and C2.

Figure 5.14 reports that responses in collaborative settings 2 and 3 (with rating 4.01 and 4.07 respectively) are more fluent than those in settings 4 and 5 (with rating 3.84 and 3.85 respectively) and similar to human-only responses. Similarly, Figure 5.15 reports that the

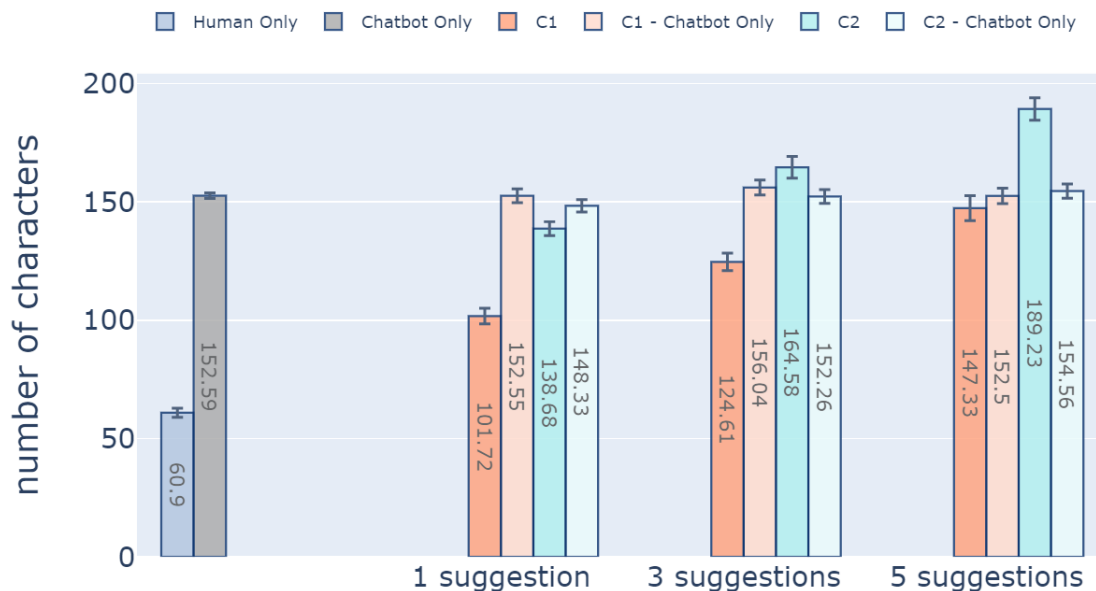


Figure 5.7: Average response length with Wikipedia Sentences Retriever.

responses in the human-only setting 1 (with rating 3.77) are more human-like as compared to those in collaborative settings 2-5. Amongst the collaborative modes, setting 3 responses (with rating 3.69) were more human-like as compared to other settings (with ratings between 3.31 - 3.42).

Figures E.5 - E.8 reports that in general, the response humanness is greater in C1 as compared to C2. While the human-only baseline has more humanness than collaborative modes C1, C2, and the chatbot-only baselines, it has better fluency ratings than settings 4 and 5 and comparable to settings 2 and 3. Chatbot-only responses, in general, are less fluent and human-like than collaborative and human-only settings. Amongst the collaborative systems, system 3 provides the best response humanness and fluency.

Figures 5.16 and 5.17 report that responses in collaborative settings 2 and 3 are more interesting (with ratings 3.84 and 3.91 respectively) than those in the human-only setting (rating 3.71) and other collaborative settings. On the other hand, responses in settings 2 and 3 are slightly less logical (with ratings 3.86 and 3.88 respectively) than those in the human-only



Figure 5.8: Average response length with Memory Network Retriever with original response.

setting (rating 3.91) but more logical than responses in settings 4 and 5. Figures E.9 - E.12 reports that while in setting 2 and 3, the human-only responses are less interesting than C1, C2, and chatbot-only responses, in settings 4 and 5 they are more interesting and logical than C1, C2, and chatbot-only responses. Compared to chatbot-only baselines, the responses in C1 and C2 are more interesting in settings 4 and 5 and more logical in all the settings.

Figure 5.18 reports the overall response quality as judged by the crowdsource workers. Responses generated in collaborative settings 2 and 3 (with ratings 3.77 and 3.89 respectively) have higher overall quality than those in settings 4 and 5 (with rating 3.35 and 3.43 respectively) and human-only setting (rating 3.74).

As reported in Figures E.13 - E.14 in settings 2 and 3, C2 produces responses with higher overall quality than C1 and human-only baseline. In settings 4 and 5, the overall quality in C1 and C2 is better than chatbot-only responses but worse than human-only baselines.

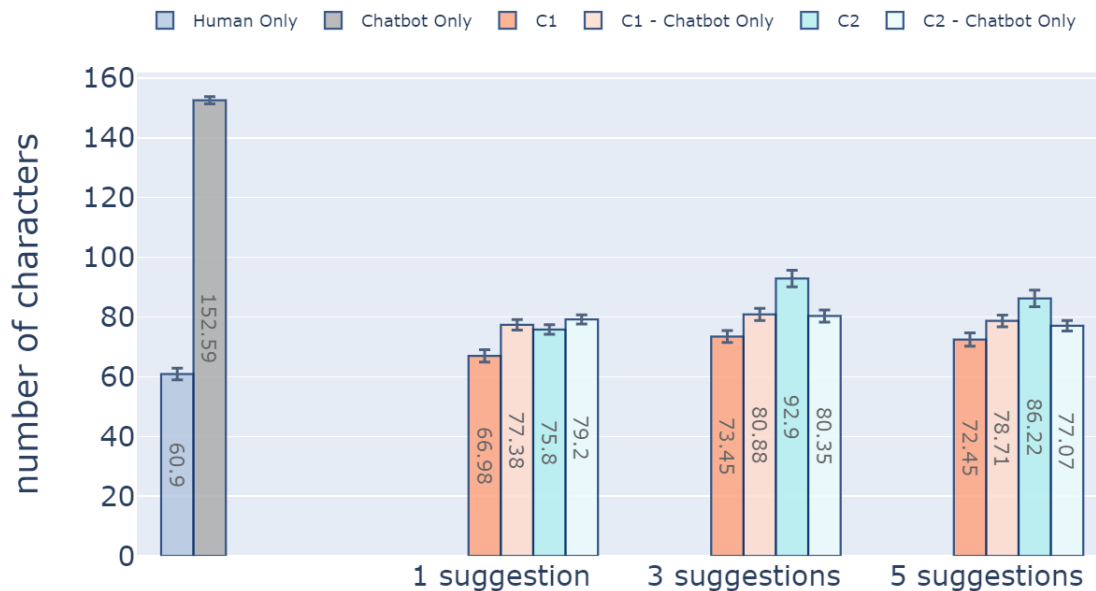


Figure 5.9: Average response length with Memory Network Retriever without the original response.

5.4 Retrieval vs Generative Models

Although the average time taken by the generative model setting 5 is lesser than that of retrieval models, responses in settings 2 and 3 are longer with fewer keystrokes than setting 5. Except for the humanness quality rating where the generative model setting 5 was rated better than the retrieval setting 2, settings 2 and 3 produced responses with better quality judgments in all the criteria mentioned in Section 3.4.2. Thus, overall retrieval models are better than generative models in human-chatbot conversational collaboration tasks in our research.

5.5 C1 vs C2

In general, as compared to C1, responses in C2 took lesser time and keystrokes but were longer. In settings 2 and 3, except for humanness rating, responses in C2 were better

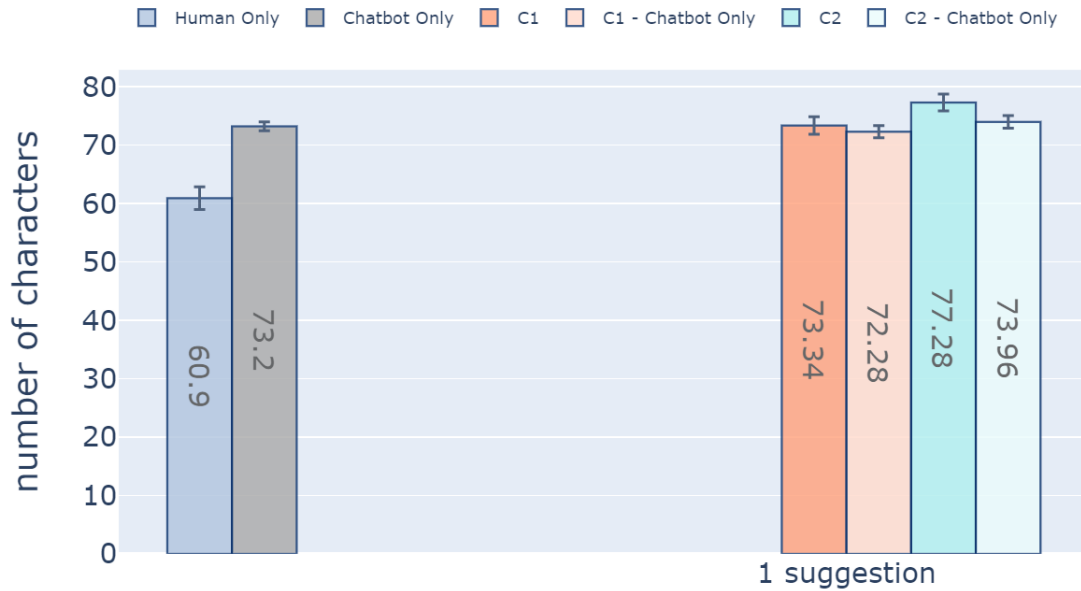
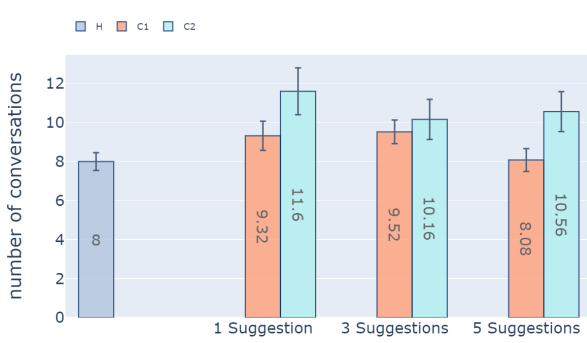


Figure 5.10: Average response length with Generative Transformer Memory Network.

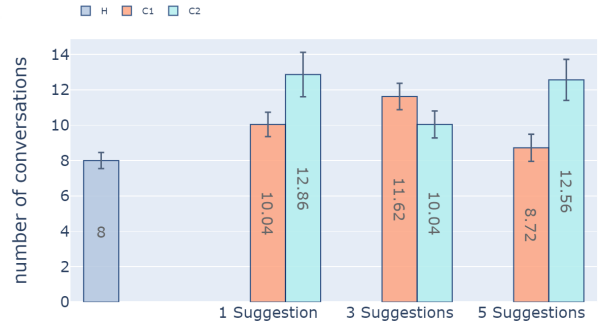
than C1 in all qualitative criteria. In settings 4 and 5, responses in C1 had better quality ratings than those from setting C2. These observations are consistent with our expectations when designing the two systems, where we preset the text field in C2 to encourage users to synthesize chatbot results to a greater extent. Since the overall performance of models in settings 4 and 5 does not match the settings 2 and 3, presetting irrelevant suggestions in the input field makes C2 perform inferior to C1.

5.5.1 Number of suggestions

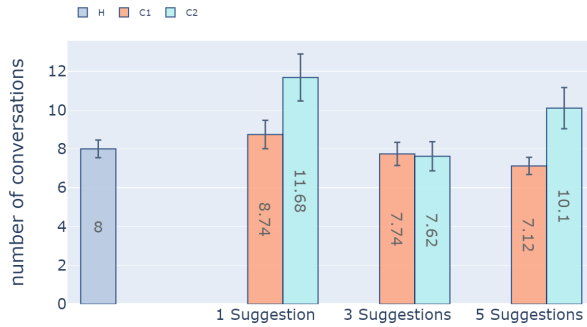
We noticed that compared to 5 suggestions, showing only 1 suggestion to participants reduced the time taken but increased the keystrokes needed to finish the responses. The length of the responses when 1 suggestion was shown was shorter as compared to those when 5 suggestions were shown. Further, increasing the number of responses did not consistently improve or hurt the response quality attributes except the humanness of responses which was reduced in some cases. For the C1 setting, with the increase in the number of suggestions the humanness



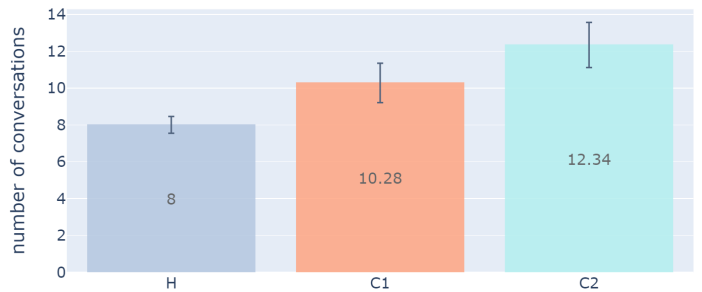
(a) Wikipedia Sentences Retriever



(b) Memory Network Retriever with original response



(c) Memory Network Retriever without the original response



(d) Generative Transformer Memory Network

Figure 5.11: Average number of conversations in collaborative settings.

and logicalness of the response decreased in setting 2. For both C1 and C2, the length of the responses increased with more suggestions.

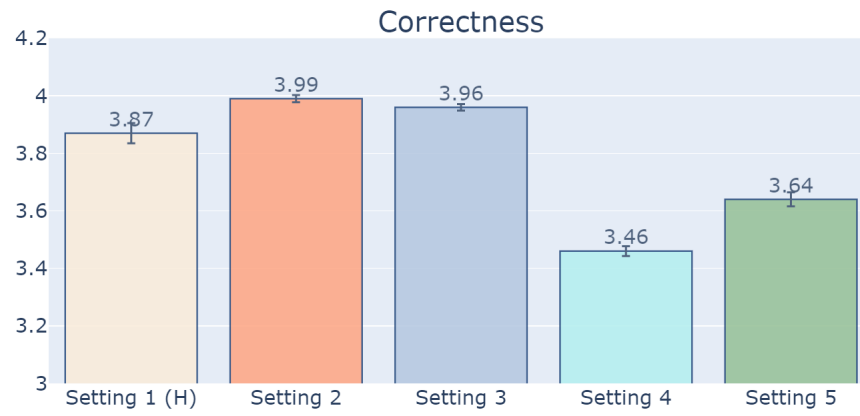


Figure 5.12: Response correctness.

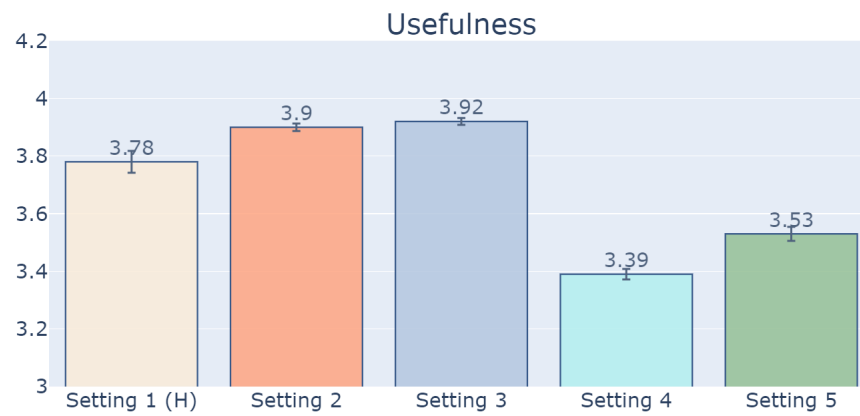


Figure 5.13: Response usefulness.

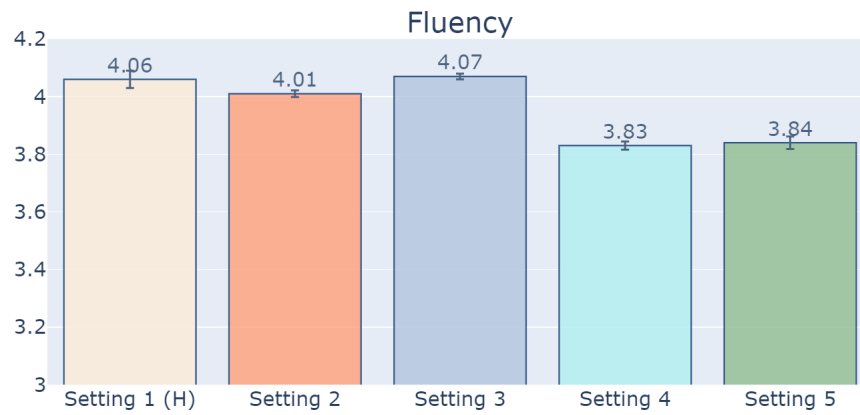


Figure 5.14: Response fluency.

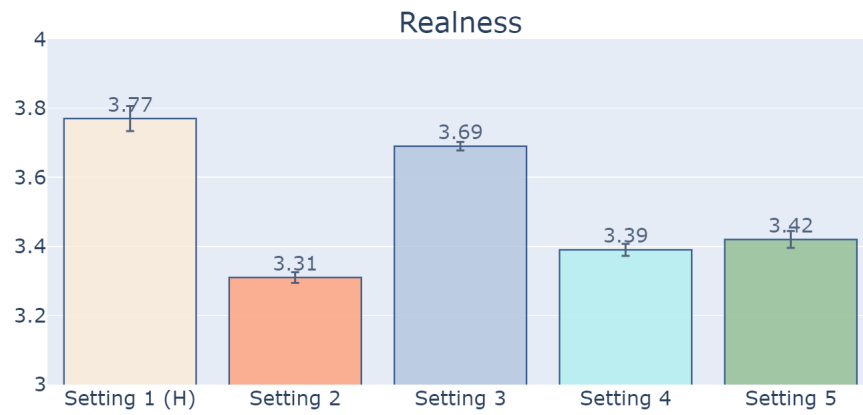


Figure 5.15: Response realness.

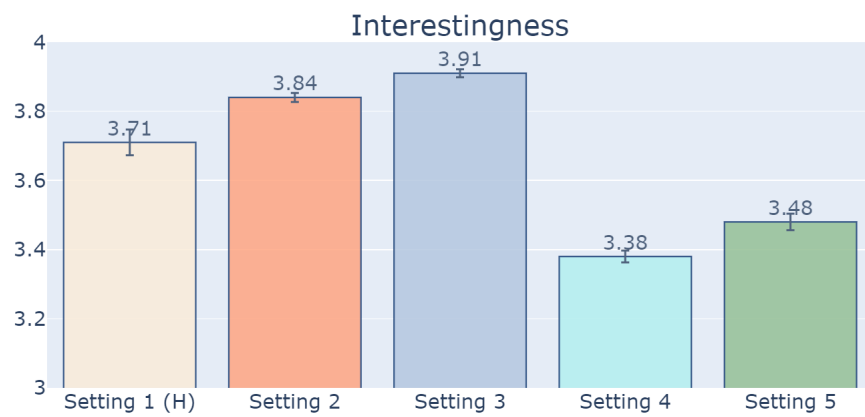


Figure 5.16: Response interestingness.

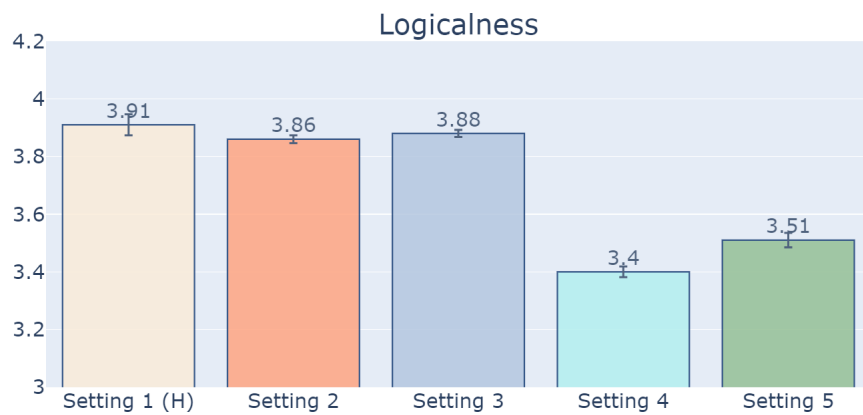


Figure 5.17: Response logicalness.

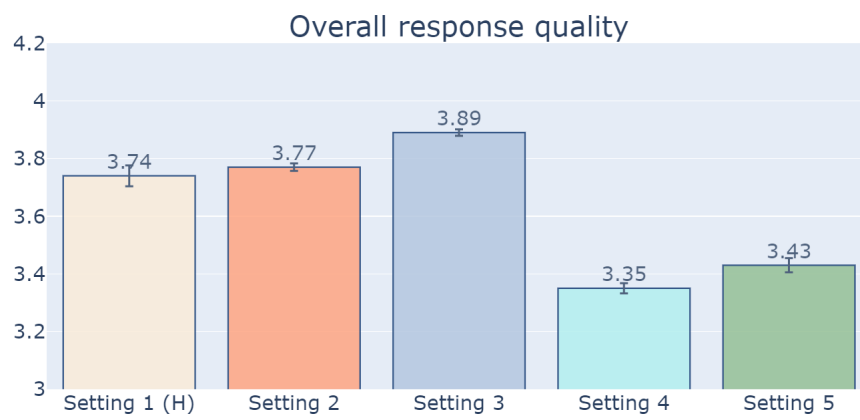


Figure 5.18: Overall response quality.

Chapter 6

Discussion and Conclusion

We present, to the best of our knowledge, the first study towards a novel human-chatbot collaborative conversation system. Our study has some important implications for the research and design of human-chatbot collaborative systems. Our conversation tasks require certain knowledge to provide informative answers (though not requiring extensive domain expertise). We expect these findings to generalize to other conversation tasks if the goal of the conversation is to provide information.

- It costs users a significant amount of effort (both time and keystrokes) to synthesize AI chatbot results with human wisdom. We found that collaborative systems reduced the effort and time to provide longer responses to conversations. This finding can be beneficial as such systems can help humans to handle more customer queries and respond efficiently to emails and messages.
- For conversations tasks that require to provide answers to common and repeated questions, the collaborative system with retrieval models producing utterances from previous conversations provides responses that have better overall quality than human-only responses and are more correct, useful, and informative. Further, these responses are fluent but not as real as human-only responses. Compared to a chatbot-only response, such responses are more useful, correct, logical and have a higher overall quality. Thus, collaborative systems can provide relevant and fluent responses in various real-world

conversation scenarios. Consider the customer sales department of a large organization that regularly gets similar queries related to its services. With collaboration with chatbots, human agents can effortlessly respond to the customer within seconds.

- For conversations tasks that require to provide answers to new topics/ questions, the collaborative system with retrieval models fetching sentences from external sources like Wikipedia provides responses that have better overall quality than human-only responses and are more useful, correct, and interesting but less fluent, human-like and logical than human-only responses. Compared to a chatbot, such responses are equally human-like but better in every other qualitative aspect discussed in the thesis. Such a system can be used as a personal assistant acting as a source of information to craft responses that have all the required details to make the response informative.
- our study shows that we can use simple design choices to encourage users to synthesize chatbot results by different extents, making the results more similar to those in a human-only or chatbot-only setting. Collaborative modes in which the input field was preset with AI suggestions produced responses that were longer but needed lesser time and keystrokes. We also show that such a collaborative setting, when used with retrieval models, makes the responses more useful, correct, interesting and useful but less human-like with a similar fluency. It provides an opportunity to balance response quality by using the interface design.
- providing more chatbot suggestions increased the response length, reduced the time taken but increased the effort to craft a response without any significant quality improvement or degradation except for humanness of responses which reduced in some cases with more suggestions provided to the user. This indicates that future system design can benefit from reducing the cost of human-AI collaboration and identifying the ideal occasions and strategies for collaboration.

Tables 6.1 and 6.2 summarises the results across various attributes for human and chatbot-only baselines and collaborative systems C1 and C2 for best performing Settings 2 and 3.

Table 6.1: Results for Setting 2.

	Human	C1	C2	Chatbot
Speed	Slow	Medium	Fast	-
Effort (keystrokes)	High	Medium	Low	-
Length	Short	Medium	Long	Long
Usefulness	Low	Medium	High	Medium
Correctness	Medium	Medium	High	Medium
Humanness	High	Low	Low	Low
Fluency	High	Medium	High	Low
Interestingness	Low	Medium	High	Medium
Logicalness	High	Medium	High	Low
Overall Quality	Medium	Medium	High	Low

Table 6.2: Results for Setting 3.

	Human	C1	C2	Chatbot
Speed	Slow	Medium	Fast	-
Effort (keystrokes)	High	Medium	Low	-
Length	Short	Medium	Medium	Long
Usefulness	Low	Medium	High	Medium
Correctness	Low	Medium	High	Medium
Humanness	High	Medium	Low	Low
Fluency	High	High	High	High
Interestingness	Low	Medium	Medium	Medium
Logicalness	High	High	High	Low
Overall Quality	Low	High	High	Medium

6.1 Limitations

Although we carefully inspected the conversation pool used to exclude conversations that were mainly chit-chat, we feel that some of the selected conversation topics were commonplace which might have led the user study participants to use their wisdom without feeling

a need to collaborate with the chatbot. Our results could have been more accurate if we had segregated our analysis by the complexity of the conversation topics.

Also, we did not deploy a live chatbot at the backend to continuously produce candidate responses for the next round of conversation. Lack of such a system prevented us to gauge the effect of collaborative systems as a conversation develops beyond a single response. Further, while analyzing the user logs, we noticed that some participants tended to ignore/ delete the chatbot suggestions from the input field and wrote their responses from scratch. We are not sure whether such participants did not like the suggestion or could not understand the task.

Finally, we did not include any metric to capture the sentiment related to the ease of use of our user study design. Such a metric could have provided us with areas where the system can improve to further improve the productivity and efficiency to craft the responses.

Bibliography

Sameh Al-Natour and Izak Benbasat. The Adoption and Use of IT Artifacts: A New Interaction-Centric Model for the Study of User-Artifact Relationships. *J. AIS*, 10, 09 2009. doi: 10.17705/1jais.00208.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, Dec. 2014. doi: 10.1609/aimag.v35i4.2513. URL <https://aaai.org/ojs/index.php/aimagazine/article/view/2513>.

Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 337–346, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702509. URL <https://doi.org/10.1145/2702123.2702509>.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction. In *CHI 2019*. ACM, May 2019. URL <https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/>.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *AAAI*, 2019.

Holger Bast and Ingmar Weber. Type Less, Find More: Fast Autocompletion Search with a Succinct Index. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 364–371, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933697. doi: 10.1145/1148170.1148234. URL <https://doi.org/10.1145/1148170.1148234>.

Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query Suggestions in the Absence of Query Logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 795–804, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307574. doi: 10.1145/2009916.2010023. URL <https://doi.org/10.1145/2009916.2010023>.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL <https://www.aclweb.org/anthology/P17-1171>.

Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In *Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii, USA, 2019. URL http://pubs.wi-kassel.de/wp-content/uploads/2018/09/JML_706.pdf.

Tilman Deuschel. On the Influence of Human Factors in Adaptive User Interface Design. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, page 187–190, New York, NY, USA, 2018. Association

for Computing Machinery. ISBN 9781450357845. doi: 10.1145/3213586.3213587. URL <https://doi.org/10.1145/3213586.3213587>.

Giovanni Di Santo, Richard McCreddie, Craig Macdonald, and Iadh Ounis. Comparing Approaches for Query Autocompletion. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 775–778, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336215. doi: 10.1145/2766462.2767829. URL <https://doi.org/10.1145/2766462.2767829>.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017.

Anca D Dragan and Siddhartha S Srinivasa. A Policy-Blending Formalism for Shared Control. *Int. J. Rob. Res.*, 32(7):790–805, June 2013. ISSN 0278-3649. doi: 10.1177/0278364913490324. URL <https://doi.org/10.1177/0278364913490324>.

Alan Feuer, Stefan Savev, and Javed A. Aslam. Evaluation of Phrasal Query Suggestions. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*, page 841–848, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938039. doi: 10.1145/1321440.1321556. URL <https://doi.org/10.1145/1321440.1321556>.

Leah Findlater and Joanna Mcgreneire. A comparison of static, adaptive, and adaptable menus. pages 89–96, 04 2004. doi: 10.1145/985692.985704.

Bruno M. Fonseca, Paulo B. Golgher, Edleno S. de Moura, and Nivio Ziviani. Using Association Rules to Discover Search Engines Related Queries. In *Proceedings of the First Conference on Latin American Web Congress, LA-WEB '03*, page 66, USA, 2003. IEEE Computer Society. ISBN 0769520588.

Jeffrey Heer. The Partnership on AI. *AI Matters*, 4(3):25–26, October 2018. doi: 10.1145/3284751.3284760. URL <https://doi.org/10.1145/3284751.3284760>.

Kristina Höök. Steps to take before intelligent user interfaces become real. *Interacting with Computers*, 12:409–426, 02 2000. doi: 10.1016/S0953-5438(99)00006-5.

Anthony Jameson. *Adaptive Interfaces and Agents*, page 305–330. L. Erlbaum Associates Inc., USA, 2002. ISBN 0805838384.

Ece Kamar. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, March 2016. URL <https://www.microsoft.com/en-us/research/publication/directions-hybrid-intelligence-complementing-ai-systems-human-intelligence/>.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, page 5686–5697, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858529. URL <https://doi.org/10.1145/2858036.2858529>.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15*, page 126–137,

- New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333061. doi: 10.1145/2678025.2701399. URL <https://doi.org/10.1145/2678025.2701399>.
- Walter S. Lasecki, Jeffrey P. Bigham, James F. Allen, and George Ferguson. Real-Time Collaborative Planning with the Crowd, year = 2012. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 2435–2436. AAAI Press.
- Yossi Matias Ehud Rivlin Zach Solan Gadi Wolfman Lev Finkelstein, Evgeniy Gabrilovich and Eytan Ruppín. Placing Search in Context: The Concept Revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, January 2002. ISSN 1046-8188. doi: 10.1145/503104.503110. URL <https://doi.org/10.1145/503104.503110>.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1127. URL <https://www.aclweb.org/anthology/D16-1127>.
- Claire Liang, Julia Proft, Erik Andersen, and Ross A. Knepper. Implicit Communication of Actionable Information in Human-AI Teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300325. URL <https://doi.org/10.1145/3290605.3300325>.
- Brian Y. Lim and Anind K. Dey. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, page 195–204, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584317. doi: 10.1145/1620545.1620576. URL <https://doi.org/10.1145/1620545.1620576>.

- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4640. URL <https://www.aclweb.org/anthology/W15-4640>.
- Ewa Luger and Abigail Sellen. “Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, page 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858288. URL <https://doi.org/10.1145/2858036.2858288>.
- Christopher D. Manning and Mihail Eric. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue. *ArXiv*, abs/1701.04024, 2017.
- Scott McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher Kelly, Dominic King, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577:89–94, 01 2020. doi: 10.1038/s41586-019-1799-6.
- A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. ParLAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476*, 2017.
- Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018*

CHI Conference on Human Factors in Computing Systems, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173580. URL <https://doi.org/10.1145/3173574.3173580>.

Stefanos Nikolaidis and Julie Shah. Human-Robot Cross-Training: Computational Formulation, Modeling and Evaluation of a Human Team Training Strategy. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, page 33–40. IEEE Press, 2013. ISBN 9781467330558.

Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. Efficient Model Learning from Joint-Action Demonstrations for Human-Robot Collaborative Tasks. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 189–196, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450328838. doi: 10.1145/2696454.2696455. URL <https://doi.org/10.1145/2696454.2696455>.

Donald A. Norman. How Might People Interact with Agents. *Commun. ACM*, 37(7):68–71, July 1994. ISSN 0001-0782. doi: 10.1145/176789.176796. URL <https://doi.org/10.1145/176789.176796>.

Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173677. URL <https://doi.org/10.1145/3173574.3173677>.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. *CoRR*, abs/1511.06732, 2015.

- Alan Ritter, Colin Cherry, and William B. Dolan. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D11-1054>.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? How controllable attributes affect human judgments. 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)*, pages 1702–1723, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1170. URL <https://www.aclweb.org/anthology/N19-1170>.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>.
- S. Tegos, S. Demetriadis, and A. Karakostas. MentorChat: Introducing a Configurable Conversational Agent as a Tool for Adaptive Online Collaboration Support. In *2011 15th Panhellenic Conference on Informatics*, pages 13–17, 2011.
- C. Thirumalai, K. S. Sree, and H. Gannu. Analysis of cost estimation function for facebook web click data. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 2, pages 172–175, 2017.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V.

- Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. A Dataset for Research on Short-Text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D13-1096>.
- Weiquan Wang, Yi Zhao, Lingyun Qiu, and Yan Zhu. Effects of emoticons on the acceptance of negative feedback in computer-mediated communication. *Journal of the Association of Information Systems*, 15(8):454–483, 2014. ISSN 1536-9323.
- Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature Hashing for Large Scale Multitask Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, page 1113–1120, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553516. URL <https://doi.org/10.1145/1553374.1553516>.
- Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics*, 45(1):163–197, March 2019. doi: 10.1162/coli_a_00345. URL <https://www.aclweb.org/anthology/J19-1005>.
- Rui Yan, Yiping Song, and Hua Wu. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of*

the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, page 55–64, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340694. doi: 10.1145/2911451.2911542. URL <https://doi.org/10.1145/2911451.2911542>.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? pages 2204–2213, 01 2018. doi: 10.18653/v1/P18-1205.

Tiancheng Zhao and Maxine Eskenazi. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Los Angeles, September 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3601. URL <https://www.aclweb.org/anthology/W16-3601>.

Appendices

Appendix A

Tools Used

A.1 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) is a popular crowdsourcing website where a variety of on-demand tasks can be created by people called *Requesters*. These task assignments are completed by remote workers colloquially known as *Turkers*. Requesters post assignments known as Human Intelligence Tasks (HITs) which can be used for a variety of user studies and data collection. Turkers can see a list of existing assignments with details like nature, length and payment amount corresponding to them.

The Turkers using the MTurk platform have access to a worker dashboard that provides an overview of their total earnings, HIT status and other account summaries. Requesters can require the Turkers to fulfil some qualifications before they can work on an assignment. These qualifications are useful to filter out Turkers based on their geographical location, performance on past HITs, etc. to create the desired candidate pool. Figure [A.1](#) shows a typical list view of available assignments.

To create an assignment HIT, the requester can either use the MTurk open application programming interface (API) or templates on the MTurk requester portal. In recent years, with the launch of AWS Software Development Kits (SDK) for MTurk, Amazon has significantly increased the features provided by the MTurk API.

The screenshot shows the Amazon Mechanical Turk worker task portal. At the top, there is a navigation bar with the Amazon MTurk logo, 'HITS', 'Dashboard', and 'Qualifications' links. A search bar labeled 'Search All HITS' is on the right. Below the navigation bar, there are tabs for 'All HITS' and 'Your HITS Queue'. The main content area displays 'HIT Groups (1-20 of 5076)' with options to 'Show Details' and 'Hide Details', and a dropdown for 'Items Per Page' set to 20. A table lists various HIT groups with columns for Requester, Title, HITs, Reward, Created, and Actions.

Requester	Title	HITs	Reward	Created	Actions
Flow MTurk	Parking Sign Transcription	15,253	\$0.08	9/12/2019	Preview, Quality
Sam Rawal	CSE 572: Bio Annotation Task	9,060	\$0.00	2/20/2020	Preview, Accept & Work
Data Science Group, The New Yo	Article Emotion Tagging	6,967	\$0.04	43m ago	Preview, Accept & Work
Tim Scullin	Collect data from a Website	4,614	\$0.05	11d ago	Preview, Accept & Work
Thorsten Beier	Los Wochos2	3,987	\$0.01	12/3/2019	Preview, Accept & Work
Aleksander Madry	Label new images for machine learning dataset	3,391	\$0.21	1/3/2020	Preview, Accept & Work
Drishti Labs	Label the actions / objects and bounding boxes ecm:newcoral-zone2b-actions-N...	3,313	\$0.00	10d ago	Preview, Quality
Rece Capture	Receipt Transcription	3,105	\$0.03	1m ago	Preview, Quality

Figure A.1: Amazon Mechanical Turk worker task portal.

Some advantages of using MTurk for crowdsourcing experiments:

- **Efficiency:** MTurk helps to efficiently manage repetitive tasks by dividing the assignments into microtasks ensuring that the tasks get completed parallelly and quickly.
- **Flexibility:** MTurk provides global on-demand access to crowd workers and removes the burden associated with dynamically scaling the resources.
- **Reduced Cost:** MTurk prevents the Requesters from the overhead costs associated with hiring and managing a temporary workforce. It offers a pay-per-task model which leads to significantly lower costs to achieve the result.

A.2 ParlAI

ParlAI (pronounced “par-lay”) [Miller et al. \(2017\)](#), is an open-source platform developed by Facebook Open Source and implemented in Python, for training and evaluating AI models using open-sourced learning agents. It provides:

- a framework that can be used for training, evaluating, and sharing dialogue models
- Parallel training over many datasets
- integration with Amazon Mechanical Turk

There are various kinds of tasks in ParlAI:

1. Question answering: It has only one turn per speaker. Its evaluation is simpler if the dataset is labelled.
2. Sentence completion: This scenario requires an agent to complete a sentence by filling a missing word in the dialogue.
3. Goal-oriented dialogue: This represents a class of a much more realistic class of dialogue to achieve a predefined real-world goal like customer service.
4. Chit-chat dialogue: These tasks are abstract discussions like two people discussing about sports.
5. Visual dialogue: These tasks are composed of images and text.

Agents, Teachers, and Worlds

- World: It is the environment for the task and can be very simple like just two people in a conversation.
- Agent: An agent acts/ speaks in the world. It can be a machine learning model, a chatbot , or a human (Turker).
- Teacher: It is an agent that talks to the learner to teach it.

Figure [A.2](#) shows the observation structure in ParlAI used by agents while in a task.

Observation/action dict	
Passed back and forth between agents & environment.	
<i>Contains:</i>	
<code>.text</code>	<i>text of speaker(s)</i>
<code>.id</code>	<i>id of speaker(s)</i>
<code>.reward</code>	<i>for reinforcement learning</i>
<code>.episode_done</code>	<i>signals end of episode</i>
<i>For supervised dialog datasets:</i>	
<code>.label</code>	
<code>.label_candidates</code>	<i>multiple choice options</i>
<code>.text_candidates</code>	<i>ranked candidate responses</i>
<code>.metrics</code>	<i>evaluation metrics</i>
<i>Other media:</i>	
<code>.image</code>	<i>for VQA or Visual Dialog</i>

Figure A.2: The observation/action dict in ParlAI.

Appendix B

User Study Setup

We inherited the *MTurkManager*, *MTurkWizardOfWikipediaWorld* and *WIZARD* classes from ParLAI’s Wizard of Wikipedia task [Dinan et al. \(2019\)](#) to begin the experiment setup. Then we created the following experiment configuration:

- *max_resp_time*: time limit in seconds for entering a dialog message (default: 600 seconds). If the user did not interact with the system for more than 10 minutes, he/she was treated as inactive and the HIT was expired for efficient resource consumption. A new HIT was pushed in the pool in such a scenario.
- *max_onboard_time*: time limit in seconds to begin the task. (default: 300 seconds). The users on the Amazon Mechanical Turk can preview the task before accepting it. When a task is being previewed, Amazon Mechanical Turk reserves the HIT for the user. For efficient resource consumption, the HIT was automatically expired if the user did not accept the task within 5 minutes. A new HIT was pushed in the pool in such a scenario.
- *auto_approve_delay*: duration to wait before auto-approving the work. (default: 3 days). Since we have collected a huge amount of user conversations, it was not practical to manually scan the sessions and approve the task. Thus, if a session consisted of at least 5 rounds of conversations with 5 minutes of duration, we auto-approved it after 3 days and the user was paid the fees.

- *suggestions_config*: The flag accepted one of the value from “no”- do not show any suggestion and “show”- show number of suggestions as configured using the *suggestions_number* flag. (default: “show”).
- *textbox_filled*: The flag accepted one of the integer values from 0 - do not prefill the textbox (C1) and 1 - prefill the textbox with the best chatbot suggestion (C2) after task initialization. (default: 0).
- *suggestions_number*: Number of AI-generated suggestions (from 1, 3, and 5) shown in the experiment. (default: 5).
- *shuffle_configs*: If the flag’s value was 1 the system overrides and randomly sets the value of the flags: *suggestions_config*, *textbox_filled*, and *suggestions_number*. (default: 1).
- *context_count*: The number of rounds of conversation in the given dialogue context. It accepts one of the integer values from 1 - one round of conversation in the context, 2 - two rounds of conversation in the context, 3 - three rounds of conversation in the context, and 0 - shuffle all contexts and randomly select the contexts during the conversations. (default: 0).
- *timer*: The duration of the experiment in seconds. (default: 300).

Once the experiment was initiated, pre-existing conversations and their suggestions were loaded according to the experiment configuration settings. A typical conversation object contained the: *conversation context*, *suggestions*, *context length information*, *file index of the conversation*, and *actual response from the dataset*. This information was very useful while analyzing the results after the experiment.

To create our user experiment, we used the AWS Python SDK to access the MTurk API. At the beginning of each HIT, we selected one of the settings from all the possible combinations of C1, C2, the number of suggestions (1, 3, and 5) and the AI chatbot model. This setting remained constant for the task duration of 5 minutes. To optimize resource consumption and automatically end the sessions where the user accepted the HIT but did not participate in the experiment, we set a time limit of 10 minutes for entering dialogue messages. If the task remained inactive for 10 minutes, we expired that session and recreated the HIT. We provided a reward of \$0.25 per HIT which was auto-approved after 3 days. After a worker completed the assignment, we assigned a qualification to the worker that prevented them from accepting another task.

The UI of the user study experiment has been created using Flask¹. The experiment began with a sample task with detailed instructions about the task. The user could click on the suggestion(s) to copy them to the text area and edit. We captured all the keys pressed and sentences selected for later analysis. Users might also copy the text from our suggestions or external sources like Wikipedia and Google search results and we logged such activities by special keywords «copy» and «paste» to easily identify them in the logs. Additionally, if the top response was pre-populated in the textbox, the identifier «autotop» was used.

Another important attribute we logged during the experiment session was the time taken to complete a conversation task. As soon as a conversation was loaded we began a timer that ended and reset after the user submitted the response. We recorded such value for every conversation task in the HIT. After the user pressed the Send button, we passed on the logs to the server that ultimately saved the session log in a file. A typical data structure sent after each conversation is given below:

{

¹<https://palletsprojects.com/p/flask>

```

text: response by the user,
id: MTurk worker ID,
message_id: message ID,
episode_done: if the experiment has been completed,
loggedKeys: keys/ activities logged,
selectedSentences: AI responses selected,
suggestionsShown: if the suggestions were shown,
timeTaken: time taken to finish a response
}

```

Turkers could preview our assignment as shown in Figure B.1 before accepting it. The welcome page gave all the necessary details to provide an overview of the task.

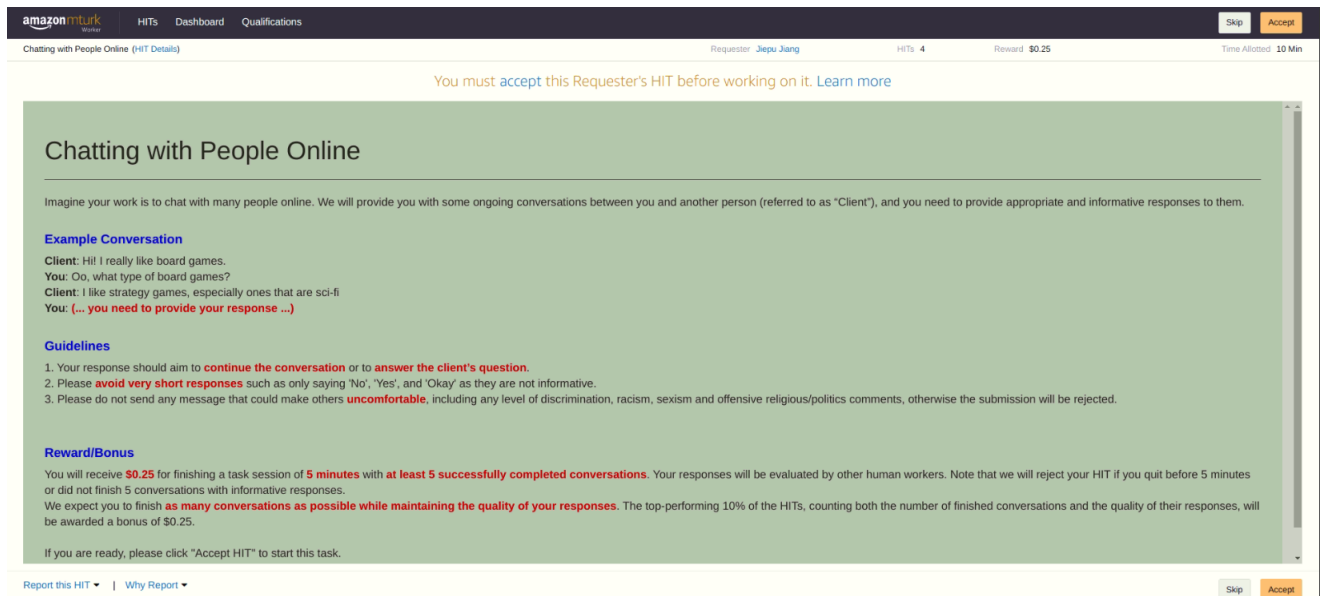


Figure B.1: User study preview page.

After accepting the task, we provided a sample task to the Turkers to establish the expectation and help them get accustomed to the user interface. Users could take as long as they

wanted to work on the sample task. The instructions for the sample task were dynamically generated based on the experiment settings. Figure B.2 shows the instructions for C1 setting.

Please refer the Appendix D for instructions in different configurations.

Sample Task

This is a sample task to help you get familiar with the system.

We will show you an ongoing conversation between you and a "client". Please read the conversation carefully. Your work is to provide **appropriate** and **informative** responses to the "client". Please **avoid very short responses** such as only saying 'No', 'Yes', and 'Okay' as they are not informative.

After completing this sample task, you will see a **timer** on the top. You need to complete as many conversations as possible while maintaining the quality of your responses. The top-performing 10% of the HITs will be awarded a **bonus**.

To help you finish the conversations, our AI system has generated some response suggestions for you below the conversation. Clicking on a suggestion will copy the text to the textbox. Please feel free to make any edits in the copied text to craft your response.

Please note that:

1. You are free to **NOT** use any suggestions; you can craft your responses from scratch if you want.
2. We do **NOT** guarantee the quality of the suggestions. Your HIT may be rejected if you simply copy and submit suggestions as they may not be appropriate and informative.

After finishing the response text, press the **Send** button to complete the current conversation.

Figure B.2: Instructions for C1 setting.

Appendix C

Judgement Task Setup

Amazon Mechanical Turk provides a large set of templates and infrastructure that make it easy to get started creating HITs for workers to complete. For our judgment experiment, we created a template from scratch to fit our needs. We provided our new project with a name as well as a title and description that was shown to workers. Figure C.1 shows the details shown in the worker tasks pool.

Describe your task to Workers

Title

Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description

Give more detail about this task. This gives Workers a bit more information before they decide to view your task.

Keywords

Provide keywords that will help Workers search for your tasks.

Figure C.1: Basic details about the judgement task shown to the workers.

To design the user interface we used Crowd HTML Elements¹ by extending the script:

```
<script src="https://assets.crowd.aws/crowd-html-elements.js"></script>
```

Additionally, we used Javascript and CSS to prepare the judgement task UI. We employed

¹<https://docs.aws.amazon.com/sagemaker/latest/dg/sms-ui-template-reference.html>

the crowd-card to display the conversation context:

```
<crowd-card>{conversation}</crowd-card>
```

Here the *conversation* is a variable that contains the HTML formatted conversation context.

To provide workers with ample time to understand and complete the assignment, we set a time limit of 15 minutes to work on a single task. Only workers with a HIT approval rate of over 95% were allowed to participate in the assignments. For every assignment, we auto paid the reward after 3 days of completion.

Appendix D

User Study Instructions

Sample Task

This is a sample task to help you get familiar with the system.

We will show you an ongoing conversation between you and a "client". Please read the conversation carefully. Your work is to provide **appropriate** and **informative** responses to the "client". Please **avoid very short responses** such as only saying 'No', 'Yes', and 'Okay' as they are not informative.

After completing this sample task, you will see a **timer** on the top. You need to complete as many conversations as possible while maintaining the quality of your responses. The top-performing 10% of the HITs will be awarded a **bonus**.

You will be shown an input field and **Send Button** below. **Fill in an appropriate response** in the input field and **Click on the Send Button** to complete the task.

Figure D.1: Instructions for setting with no suggestions.

Sample Task

This is a sample task to help you get familiar with the system.

We will show you an ongoing conversation between you and a "client". Please read the conversation carefully. Your work is to provide **appropriate** and **informative** responses to the "client". Please **avoid very short responses** such as only saying 'No', 'Yes', and 'Okay' as they are not informative.

After completing this sample task, you will see a **timer** on the top. You need to complete as many conversations as possible while maintaining the quality of your responses. The top-performing 10% of the HITs will be awarded a **bonus**.

To help you finish the conversations, our AI system has generated some response suggestions for you below the conversation. Clicking on a suggestion will copy the text to the textbox. Please feel free to make any edits in the copied text to craft your response.

Please note that:

1. You are free to **NOT** use any suggestions; you can craft your responses from scratch if you want.
2. We do **NOT** guarantee the quality of the suggestions. Your HIT may be rejected if you simply copy and submit suggestions as they may not be appropriate and informative.

After finishing the response text, press the **Send** button to complete the current conversation.

Figure D.2: Instructions for C1 setting.

Sample Task

This is a sample task to help you get familiar with the system.

We will show you an ongoing conversation between you and a "client". Please read the conversation carefully. Your work is to provide **appropriate** and **informative** responses to the "client". Please **avoid very short responses** such as only saying 'No', 'Yes', and 'Okay' as they are not informative.

After completing this sample task, you will see a **timer** on the top. You need to complete as many conversations as possible while maintaining the quality of your responses. The top-performing 10% of the HITs will be awarded a **bonus**.

To help you finish the conversations, our AI system has generated some response suggestions for you below the conversation.

The AI system has **already copied** the text of the best suggestion to the textbox for you. Please feel free to make any edits in the copied text to craft your response.

Please note that:

1. You are free to **NOT** use any suggestions; you can craft your responses from scratch if you want.
2. We do **NOT** guarantee the quality of the suggestions. Your HIT may be rejected if you simply copy and submit suggestions as they may not be appropriate and informative.

After finishing the response text, press the **Send** button to complete the current conversation.

Figure D.3: Instructions for C2 setting.

Appendix E

Results

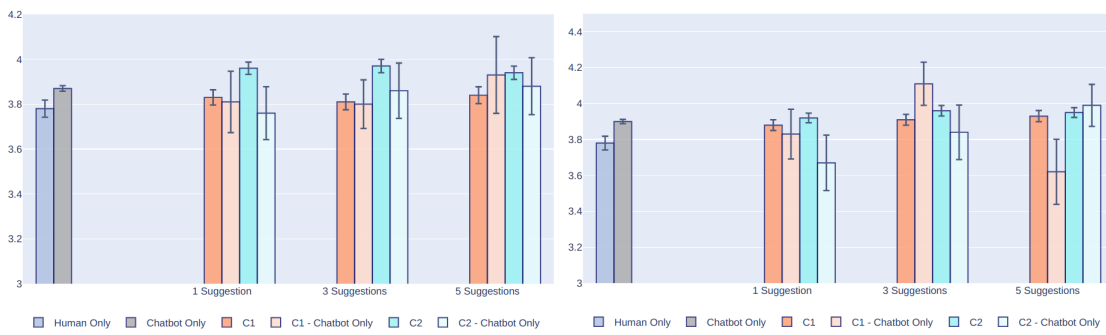


Figure E.1: Average response usefulness in Settings 2 and 3.

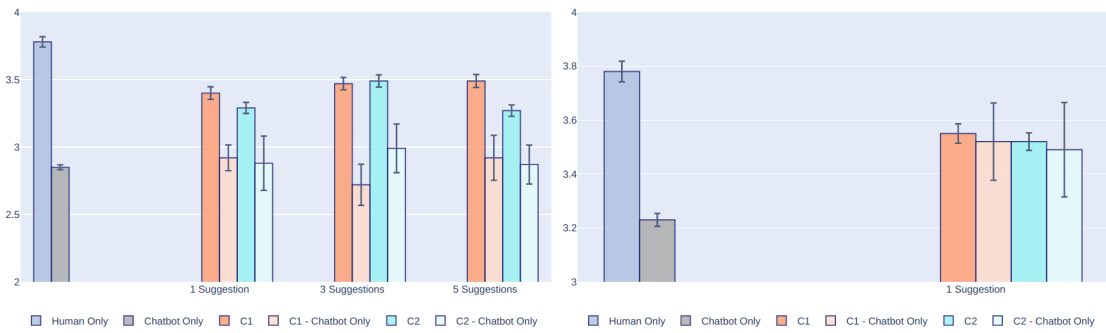


Figure E.2: Average response usefulness in Settings 4 and 5.

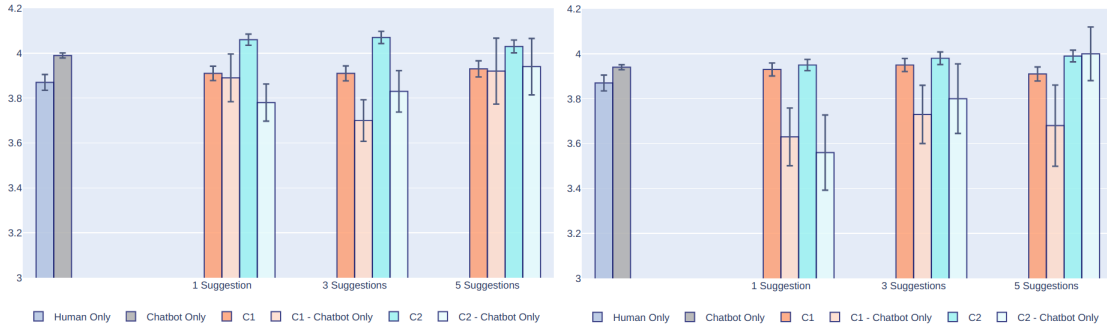


Figure E.3: Average response correctness in Settings 2 and 3.

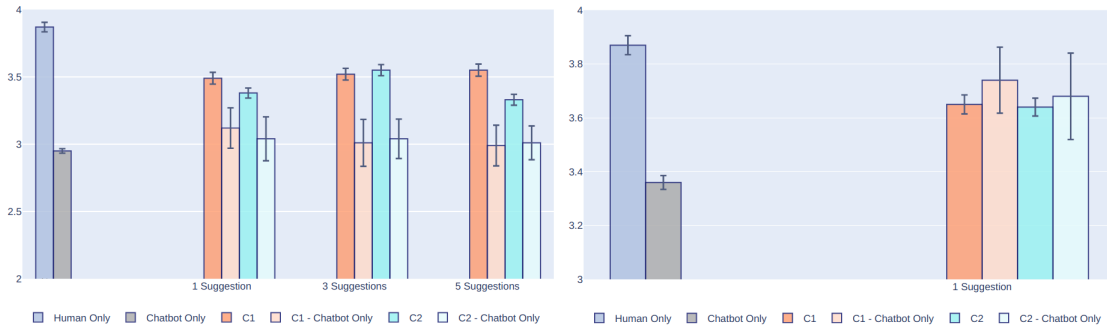


Figure E.4: Average response correctness in Settings 4 and 5.

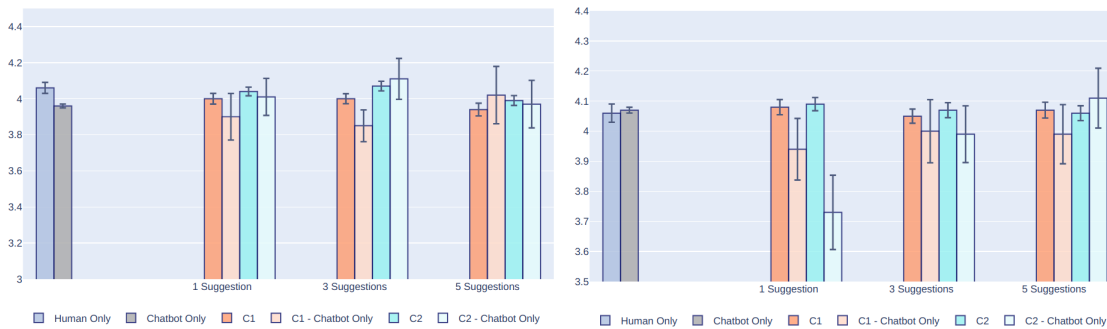


Figure E.5: Average response fluency in Settings 2 and 3.

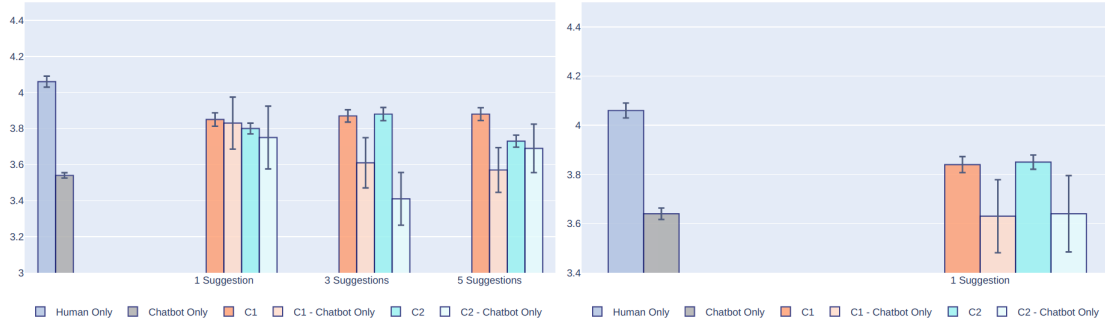


Figure E.6: Average response fluency in Settings 4 and 5.

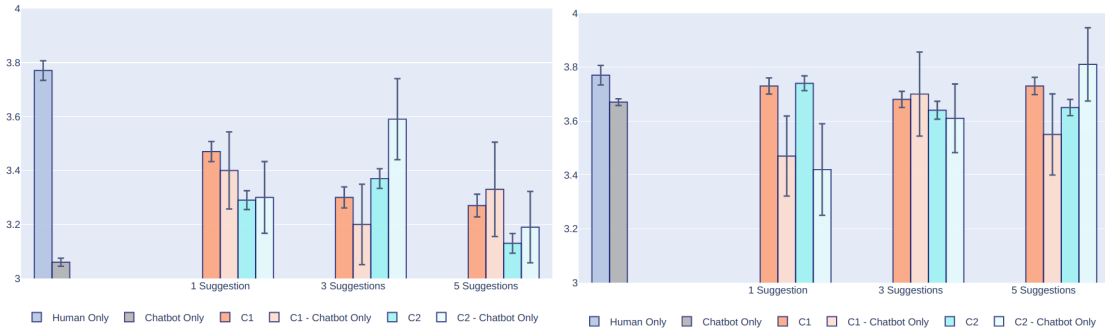


Figure E.7: Average response humanness in Settings 2 and 3.

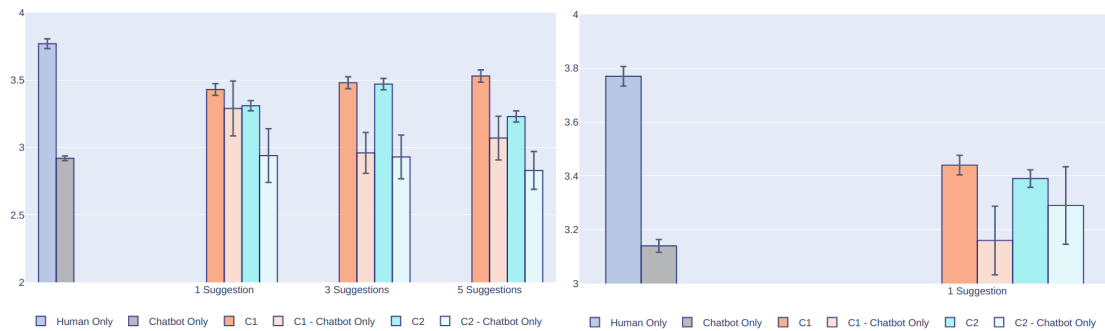


Figure E.8: Average response humanness in Settings 4 and 5.

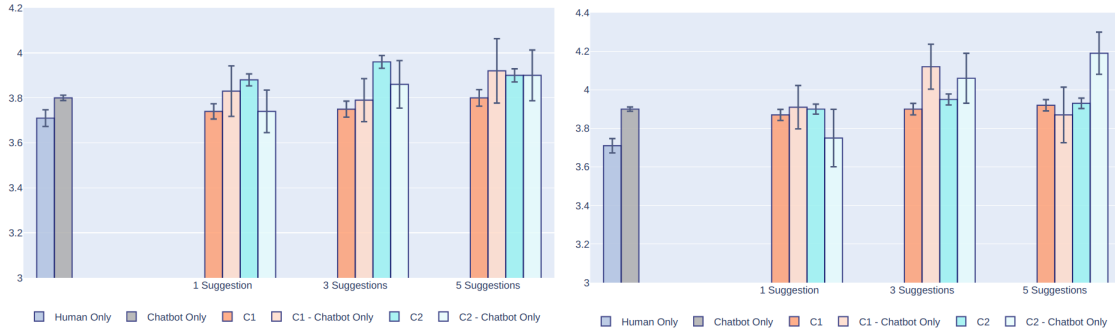


Figure E.9: Average response interestingness in Settings 2 and 3.

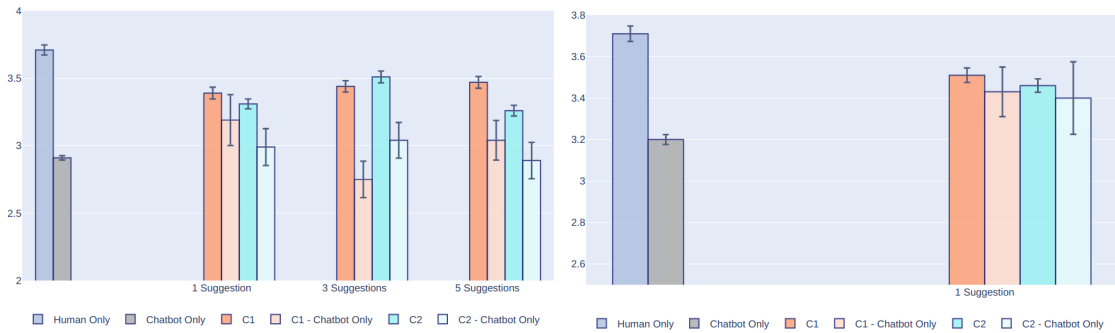


Figure E.10: Average response interestingness in Settings 4 and 5.

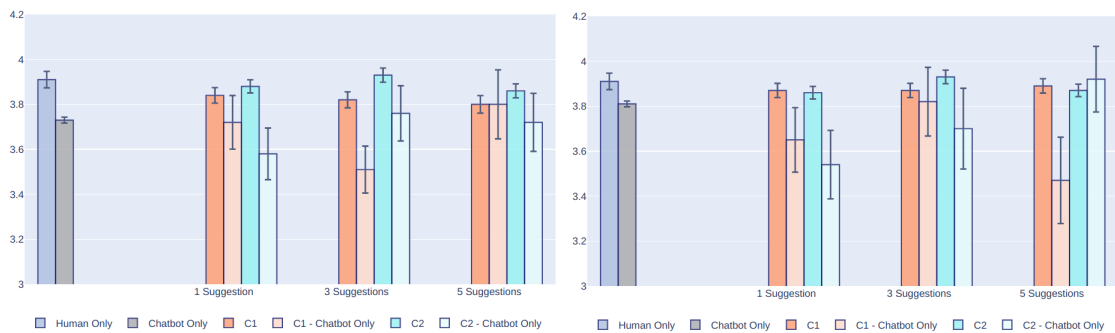


Figure E.11: Average response logicalness in Settings 2 and 3.

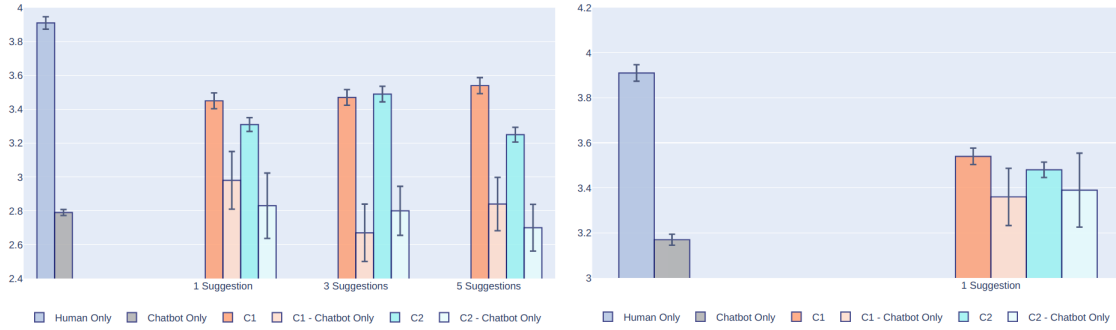


Figure E.12: Average response logicalness in Settings 4 and 5.

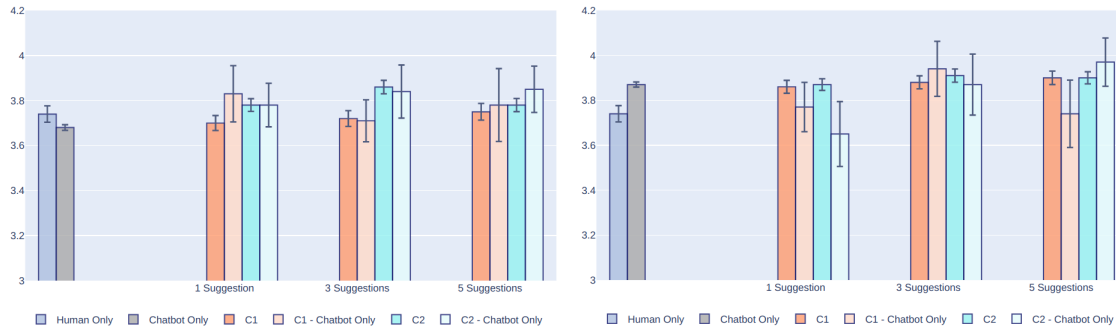


Figure E.13: Average overall response quality in Settings 2 and 3.

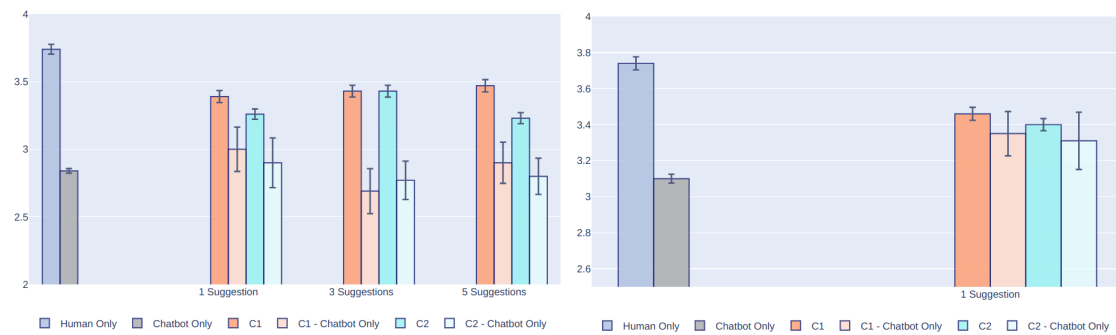


Figure E.14: Average overall response quality in Settings 4 and 5.

Appendix F

Examples from User Studies

Table F.1: Examples of collected responses from the user study in Setting 2.

Conversation Context:

Justin Beiber seems to really jump in and out of the spotlight constantly
 I think Justin Drew Bieber, the Canadian singer is a fantastic singer
 some of his music isnt too bad, but i dont care much for it
 For a musican that was discovered for his awesome Youtube videos and signed I think
 he must be good
 when was that? seems like its been a while.

Suggestions:

After a talent manager discovered him through his YouTube videos covering songs in 2008 and signed to RBMG, Beer released his debut EP, 'My World', in late 2009.

Good Response:

It was around 2008 he was discovered by a talent manager, he was soon signed to RBMG. His first song came out in 2009.

Bad Response:

Not quite long, they just opened a new restaurant at Woodland, California and let you know.

Table F.2: Examples of collected responses from the user study in Setting 3.

Conversation Context:

I spent the summer holiday in canada,though I dont know much, can you help out?
 Canada is a country located in north America
 Ok, when was it founded?
 It has ten provinces and three territories
 Any renowned sites?

Suggestions:

The country has a severely cold winter climate
 It is unlikely. The first jump in history was back in 1797.
 We used to get a lot more over in Oklahoma but now we don't get much honestly.

Good Response:

Quebec is well-known for its unique French-Canadian culture.

Bad Response:

It is unlikely. The first jump in history was back in 1797.

Table F.3: Examples of collected responses from the user study in Setting 4.

Conversation Context:
 Hey! I really like Yoga. What about you?
 I like yoga due to the mental practices.
 it does help clear the mind. Though the stretches are quite nice too.
 I prefer Hatha yoga.
 Oh? What does Hatha specialize in

Suggestions:
 I started yoga and I feel great!

Good Response:
Hatha practices involve the breath, body and mind.

Bad Response:
I have no clue what Hatha specializes in.

Table F.4: Examples of collected responses from the user study in Setting 5.

Conversation Context:
 Hi, I took a sewing class in middle school
 You mean the craft of fastening or attaching objects using stitches made with a needle and thread? That is cool.
 yes, I got to make my own pillow
 I like those Pillow Pets made from chenille.
 what is a chenille?

Suggestions:
 I like to sew, it is the process of fastening objects with stitches using a needle and thread

Good Response:
Chenille is a woven fabric that can be made from a variety of different fibers, including cotton, silk, wool, and rayon.

Bad Response:
Im not sure to be honest.

Appendix G

Sample Conversation and Suggestions

Table G.1: Sample conversation contexts and suggestions for Wikipedia Retrieval.

Context Length: 1**Conversation Context:**

When walking through Walmart I always see so much fishing tackle and equipment!
I really dont know how any of it works, do you?

Suggestions:

Fishing tackle is the equipment used by fishermen when fishing.

Almost any equipment or gear used for fishing can be called fishing tackle.

Recreational fishing, also called sport fishing, is fishing for pleasure or competition.

It can be contrasted with commercial fishing, which is fishing for profit, or subsistence...

Fly fishing is an angling method in which an artificial fly is used to catch fish.

Context Length: 2**Conversation Context:**

I have blond hair, what do you think of people with that hair color?

Sometimes I wish I had blonde hair, or fair hair. They have low levels of eumelanin.

What is eumelanin? I like it, but sometimes I need a change, plus it gets thick.

Suggestions:

Hair color is the pigmentation of hair follicles due to two types of melanin: eumelanin...

Generally, if more eumelanin is present, the color of the hair is darker; if less eumelanin...

Melanin (; from melas, black, dark) is a broad term for a group of natural pigments...

Melanin is produced by the oxidation of the amino acid tyrosine, followed by...

Red hair (or ginger hair) occurs naturally in 1–2% of the human population.

Context Length: 3**Conversation Context:**

What allergies do you have?

I do not have any allergies or allergic diseases.

Do you know anything about allergies?

Yes I know food allergies are due to abnormal immune response to food.

What are some common allergies?

Suggestions:

A food allergy is an abnormal immune response to food.

The signs and symptoms may range from mild to severe.

Peanut allergy is a type of food allergy to peanuts.

It is different from nut allergies.

Anaphylaxis is a serious allergic reaction that is rapid in onset and may cause death.

Table G.2: Sample conversation contexts and ranked candidate responses from Wikipedia (M1)

Context Length: 1

Conversation Context:

do you know anything about the Narcissus plant?

Suggestions:

Narcissus is a genus of predominantly spring perennial plants in the Amaryllidaceae...
 Various common names including daffodil, daffadowndilly, narcissus, and jonquil are...
 'Narcissus' has conspicuous flowers with six petal-like sepals surmounted by a cup- or...
 The flowers are generally white or yellow (also orange or pink in garden varieties),...
 'Narcissus' were well known in ancient civilization, both medicinally and botanically,...

Context Length: 2

Conversation Context:

The beach is my happy place. Do you like the beach?

Yes! I love the beach, especially wild beaches. They're my favorite to come across.

Wild beaches? I'm not familiar with that term.

Suggestions:

Wild beaches, also known as undeveloped or undiscovered beaches, are not developed...
 Wild beaches can be valued for their untouched beauty and preserved nature.
 Beaches typically occur in areas along the coast where wave or current action deposits...
 Although the seashore is most commonly associated with the word 'beach', beaches...
 'Beach' may refer to: The former are described in detail below; the larger geological...

Context Length: 3

Conversation Context:

Hi there. You play the piano?

No, i dont, The piano is an acoustic, stringed musical instrument

I took a few piano lessons when I was little. Do you play an instrument?

yes, but he word 'piano' is a shortened form of pianoforte

I didn't know that. When Bach was alive did he call it a pianoforte?

Suggestions:

The Italian musical terms 'piano' and 'forte' indicate 'soft' and 'loud' respectively...
 The first forte pianos in the 1700s had a quieter sound and smaller dynamic range...
 An acoustic piano usually has a protective wooden case surrounding the soundboard...
 The Bach House in Eisenach, Thuringia, Germany, is a museum dedicated to...
 On its 600 me it displays around 250 original exhibits, among them a Bach music autograph.

Table G.3: Sample conversation contexts and candidate responses from Retrieval Transformer Memory Network

Context Length: 1

Conversation Context:

I love watching shows on Netflix, do you know what it is?

Suggestions:

Me too. It streams media and video on demand. You can even rent DVD's through mail.

I know all those channels I wonder if reruns come on still.

That's what I love about it too. What reality shows do you watch?

I know! And they have some new episodes coming on it in 2019.

I love to surf YouTube, I always end up watching interesting videos on there.

Context Length: 2

Conversation Context:

I love Taco's! a Taco bell recently opened up near us and i cant get enough of it.

tacos are a traditional mexican dish composed of corn or flour tortillas, which taco bell sells a lot of

I didnt know that! Do you know when taco's started being made?

Suggestions:

It is not exactly known as the origins are generally theoretical that it derived from Mexico

Ok, I'll have to try to mix it up next time I make it

Yea and then they revised the formula in 1958

Not that I know of.

I don't know about that but I do know jaguar xj220 was the fastest car in 1992 and 1993

Context Length: 3

Conversation Context:

Hi there. You play the piano?

No, i dont, The piano is an acoustic, stringed musical instrument

I took a few piano lessons when I was little. Do you play an instrument?

yes, but he word 'piano' is a shortened form of pianoforte

I didn't know that. When Bach was alive did he call it a pianoforte?

Suggestions:

Yes, he did com The first forte pianos in the 1700s had a quieter sound

He actually attended the Academy of Contemporary Music as an undergrad.

Hi there. You play the piano?

I love Pope, would this be the era where there was mostly no sound?

Ballet came about from the Italian Renaissance around the 15th century, then caught on...

Table G.4: Sample conversation contexts and candidate responses from Generative Transformer Memory Network

Context Length: 1

Conversation Context:

Who founded Coors Brewing Company?

Suggestion:

I know that Coors was founded in 1873 by German immigrants to Jacob Schuyler.

Context Length: 2

Conversation Context:

I just discovered star trek and I really like watching star trek.

Gene Roddenberry created it based upon science fiction and it is American media. was the show created in the mid 60s?

Suggestion:

The first star trek was created by jams Roddenberry.

Context Length: 3

Conversation Context:

Hey! I really like Yoga. What about you?

I like yoga due to the mental practices.

it does help clear the mind. Though the stretches are quite nice too.

I prefer Hatha yoga.

Oh? What does Hatha specialize in

Suggestion:

Hatha yoga is one of the most commonly known types of yoga
