Information Processing and Management 59 (2022) 102838

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

How humans obtain information from AI: Categorizing user messages in human-AI collaborative conversations

Yuhan Wei [a,b], Wei Lu [a,b], Qikai Cheng [a,b], Tingting Jiang [a], Shewei Liu [c,*]

[a] School of Information Management, Wuhan University, Wuhan, Hubei, China
[b] Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, Hubei, China
[c] Tk.cn Insurance CO., LTD, China

ARTICLE INFO

Keywords:
Human-AI collaboration
Conversational agent
Informational conversation
Customer service

ABSTRACT

Although there is an increasingly number of research about the design and use of conversational agents, it is still difficult for conversational agents to completely replace human service. Therefore, more and more companies have adopted human-AI collaborative systems to deliver customer service. It is important to understand how people obtain information from human-AI collaborative conversations. While the existing work relies on self-reported methods to elicit qualitative feedback from users, we have concluded a categorization system for user messages in human-AI collaborative conversations after a thorough examination of a real-world customer service log, which could objectively reflect the user's information needs. We categorize user messages into five categories and 15 specific types related to three high-level intentions. Two annotators independently classified the same set of 1,478 user messages from 300 conversations and reached a moderate consistency. We summarize and report the characteristics of different message types and compare their usage in sessions with only human, AI, or both representatives. Our results show that different message types vary significantly in usage frequency, length, and text similarities with other messages in a session. Also, the frequency of using different message types in our dataset seems consistent over sessions with different types of representatives. But we also observed some significant differences in a few specific message types across the sessions with different representatives. Our results are used to suggest some areas for improvement and future work in human-AI collaborative conversational systems.

1. Introduction

Human-AI collaborative conversation is a new dialogue system which employ a hybrid model involving both human representatives and AI outputs. For an incoming customer message, the system will determine whether it is replied by AI or human representatives through a series of calculations. Figure 1 shows an exemplar of human-AI collaborative conversation from an online customer service. In this dialogue, the customer is asking about some information about a product. The first two questions are answered by the chatbot, and the last one is answered by a human representative. Although the traditional way to provide customer service is that a user asks a question and a staff member answers it, there are some problems with human services, such as high labor cost, lack of

* Corresponding author at: School of Information Management, Wuhan University, Wuhan, Hubei, China.
E-mail addresses: yuhan_wei@whu.edu.cn (Y. Wei), weilu@whu.edu.cn (W. Lu), qikaicheng@whu.edu.cn (Q. Cheng), tij@whu.edu.cn (T. Jiang), liusw23@taikanglife.com (S. Liu).

https://doi.org/10.1016/j.ipm.2021.102838
Received 12 August 2021; Received in revised form 6 November 2021; Accepted 21 November 2021
Available online 20 December 2021

professional knowledge, slow response, and so on. An increasingly popular alternative service today is conversational agents such as Siri, Google, Alexa, and Cortana. A conversational agent is a kind of Artificial Intelligence which can respond to users' simple queries or commands to accomplish a single-turn QA or goal-oriented task, such as asking for time and scheduling appointments (Yang, Xu, & Chen, 2021). They provide information to users in a conversation model, allowing people to communicate in a natural way. Fully automatic conversational agents can answer questions well when the question is simple and common, or the topic is clear and specific. But when they are asked to provide unusual information or respond to follow-up questions, they are not as good at responding, which may frustrate the users (Radziwill & Benton, 2017). To improve customer satisfaction while reducing the labor cost, many commercial companies adopt a human-AI collaborative dialogue system, where workers can use intelligent assistants in daily work practices to provide services for customers (Chung, Ko, Joung, & Kim, 2020).

To build functional and natural human-AI collaborative dialogue systems, it is necessary to understand how users interact in human-AI collaborative environments (Osterlund et al., 2021). Previous studies have concluded that many common patterns of dialogue act in human-human conversations, such as statement of opinion, greetings, information request, and others (Pareti & Lando, 2018). These patterns indicate user intent and provide a fundamental basis for understanding user behavior. The development of conversational AI agents has focused the interest of this study on human-AI conversations, which can help improve the performance of conversational agents. In order to analyze and characterize dialogue acts in human-AI collaborative environments, we focus on studying user messages in an online text-based informational dialogue system, where users chat with a conversational agent or human service via text to address their information needs. Text-based informational conversation is important for people to acquire information, as it provides more efficient information access in a more natural and interactive way, particularly when it is difficult for the users to retrieve the required information by themselves. Taking the customer service on an E-commerce platform named Taobao.com as an example, over 77% of buyers on Taobao asked for information about products via text before placing an order (Jie & Zhang, 2011). Therefore, such text-based informational dialogue data contain very important clues to study the user intent in human-AI collaborative conversations.

Particularly, we are interested in what types of messages users send in human-AI collaborative conversations, and the underlying user intent behind these messages. We follow previous studies of dialogue acts and conclude a classification scheme of user messages based on a thorough examination of a real-world customer service log. Our classification scheme included five categories of messages (including 15 specific types) linked to three higher-level intentions: describing information, understanding information, and maintaining conversation. Two annotators independently classified the same set of 1,478 user messages from 300 conversations and came to a moderate consistency with Cohen's Kappa = 0.59. Also, the customer service system employed a hybrid model involving both human representatives and AI outputs. Thus, our dataset also included three types of dialogue: dialogue between customer and human representative, dialogue between customer and AI representative, and dialogue between mixed representatives—this makes it possible to also examine the differences in messages and message types in sessions with different representatives.

We are particularly interested in the following research questions:

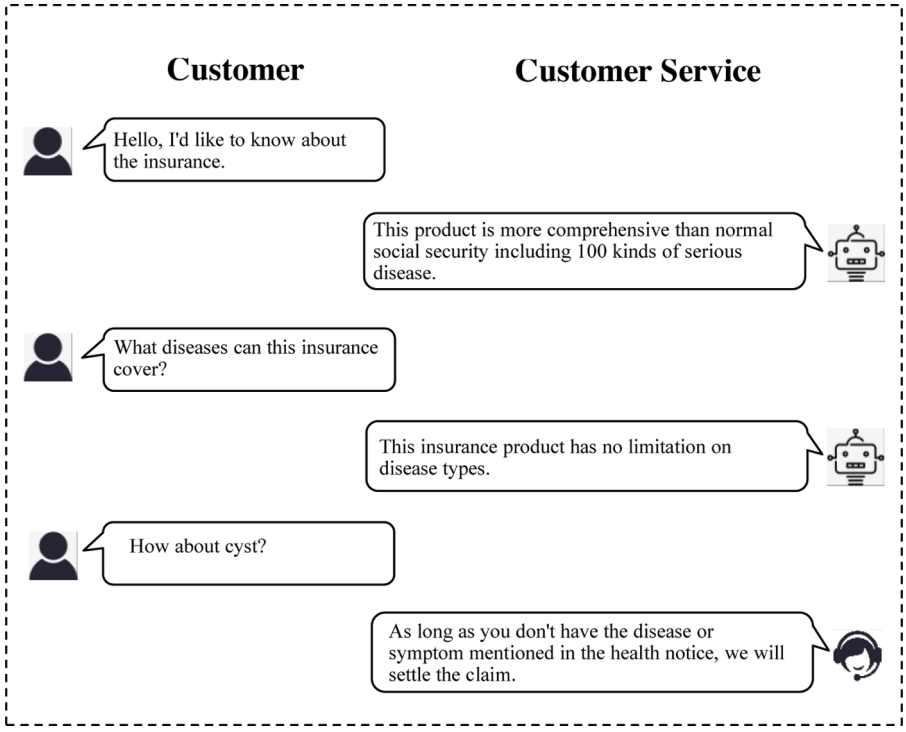

Fig. 1. An exemplar of dialogue from online customer service.

- **RQ1—**What types of messages do users send in human-AI collaborative conversations? What are the possible intentions behind sending these message types?
- **RQ2—**How do different types of messages vary in their characteristics and usage frequency?
- **RQ3—**Does the use of different message types vary in sessions with different representatives?

The rest of this article introduces our categorization scheme, data annotation process, and results.

2. Literature review

2.1. Human-AI Collaborative Conversation

With the development of machine learning capabilities, natural language processing produces natural and straightforward dialogue experiences for industries. As a novel and entertaining way to satisfy clients, conversational agents have shown their advantages in improving service quality and efficiency, and have aroused an increasing interest in the field of business (Bavaresco et al., 2020; Wazurkar, Bhadoria, & Bajpai, 2017). However, the grammatical complexity or semantics of the conversation limit the interaction between customers and conversational agents, such as inappropriate responses generation (Hori et al., 2019). Although humanizing conversational agents or increasing users' perception of human presence has been a major goal of conversational agent design (Go & Sundar, 2019; Van den Broeck, Zarouali, & Poels, 2019), it is more effective to build a human-AI collaborative conversational system where human representatives and conversational agents work together to provide services for customers in different scenarios (Liu et al., 2020) .

The concept of symbiotic computing was originated by Licklider (1960) in "Man-Machine Symbios". With the development of machine learning and natural language processing, AI began to be designed to understand the human perception and cognition (Brynjolfsson & Mitchell, 2017; Neff & Nagy, 2018). This makes it possible for AI to replace part of the human work or augment existing human skills. But in most cases, a workflow still needs human participation in the core work, especially when it involves complex tasks (Sterlund et al., 2020). The notion of a computer working as a partner with humans became an increasingly common metaphor for interaction design (Gray & Suri, 2019).

In the past decade, AI technology has been applied in the service industry and is expected to substantially change both marketing strategies and customer behaviors in the future. Some practices have proved that AI will be more effective if it augments human managers rather than replaces them (Davenport, Guha, Grewal, & Bressgott, 2020). Intelligent assistants are already helping both customers and workers to more easily interact with information at different points of service through more natural conversations (Maedche et al., 2019). For example, Isbister et al. (2000) designed an AI assistant that could support human-human communication in virtual environments. They found that the AI assistant made positive contributions to participants' experience of the conversation and even seemed to affect their style of behavior.

It is proved that there are notable differences in the content and quality between human-chatbot conversations and human-human conversations (Hill et al., 2015). Despite a growing body of research focusing on the design and use of intelligent assistant (Renjith, Sreekumar, & Jathavedan, 2020; Sun, Cheng, Wang, Qi, & Liao, 2021; Ye et al., 2018), it is necessary to find out how an intelligent assistant is perceived and used during human-human conversation, which is reflected in the user's behavior and intention. Some researchers tried to understand how users interact with a human-AI collaborative system that already existed. Hohenstein and Jung (2018) compared conversations between dyads using AI-assisted and standard messaging apps and elicited qualitative feedback from users of the AI-assisted messaging app through interviews. Instead of using self-reported methods, Kušen and Strembeck (2020) analyzed a dataset including more than 4.5 million tweets to characterized human-human and human-bot communication on Twitter by specific, but they simply considered the impact of robot on human-human conversations. In this paper, we would like to characterize user informational intentions by categorizing user messages in a real-world customer service log, and to compare the differences of user intentions in three types of conversation (human-human conversation, human-AI conversation, human-human conversation with AI assisted).

2.2. Dialogue Acts Classification

In the framework of dialogue systems, dialogue acts (DA) can be helpful to identify and model user intention (Renjith et al., 2020). Furthermore, DA information may be also used to increase the performance of the dialogue system. DA are fine-grained classification systems for user-system communications in conversational dialog systems (Oraby, Bhuiyan, Gundecha, Mahmud, & Akkiraju, 2019; Oraby, Gundecha, Mahmud, Bhuiyan, & Akkiraju, 2017) . Previous studies developed different dialogue acts categories with different granularity, and used the acts to manage conversations, generate responses, model users, and evaluate systems.

Defining taxonomies of dialogue acts has been studied extensively for decades, for human-human conversations. Many early studies use the Dialogue Act Markup in Several Layers (DAMSL) scheme (Core & Allen, 1997). The annotation scheme of DAMSL is based on spoken, task-oriented dialogues, and is fine-grained, with 220 tags divided into four categories depending on their roles in the conversation and characteristics. Stolcke et al. (2000) introduced an approach for modeling dialogue acts from human-human conversational speech which could detect and predict dialogue acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialogue act sequence. They focus on recognizing 42 major dialogue act types from the work presented by Jurafsky and Shriberg (1997), such as Statements and Opinions, Questions, Backchannels, Turn Exits, Answers and Agreements, etc.. Recently, this idea has been extended to human-machine conversations, which is more challenging for DA classification. Jiang et al.

(2015) defined several dialogue acts in intelligent voice assistants and used their transition patterns to predict conversation quality. Qu et al. (2018) introduced a labeled dialog dataset of question answering interactions between information seekers and providers from an online forum on Microsoft products and classified user intent in dialogs into 12 classes. Ahmadvand et al. (2019) fine-tuned their DA classification model trained on human-human conversations to human-machine conversations and demonstrated a promising result. In addition to users' dialogue acts, systems' dialogue acts can also be classified. For example, Wood, Eberhart & McMillan (Wood, Eberhart, & McMillan, 2020) targeted the problem of dialogue act classification for a virtual assistant for software engineers repairing bugs. But from the perspective of user intention, system acts usually do not need to be recognized because they can be defined while designing the system.

Approaches defining and analyzing dialog acts have also been applied to examine a particularly type of human discourse. For example, Sandor et al. (Sandor, Lagos, Vo, & Brun, 2016) proposed the detection of user issues and request types in technical forum question posts and presented a categorization system for detecting the proposed question post types based on discourse analysis. Tavakoli (Tavakoli, 2020) analyzed human-generated clarifying questions in a Community Question Answering website as a sample of conversation and discovered the patterns and types of such clarifying questions. Our work also follows previous studies of dialogue acts but applies similar methods to examine online text-based human-AI collaborative conversations.

3. User Message Category

3.1. Conversation Sessions

We examine online text-based conversations between users and customer service which is composed of a conversational agent and several human representatives. Particularly, we focus on the conversations where the primary goal is to address the user's information need, and the customer service provides information. We call such conversations *informational conversations*. Informational conversation is an important method to address people's information needs, especially in commercial services such as online customer support. Other types of conversations also exist—for example, a conversation can also be transactional (e.g., chatting with a colleague to make an appointment), discussional (e.g., debating with a friend about presidential candidates), or entertaining (e.g., conversations that are just for fun)—but we do not discuss them here.

We further define the two parties of an informational conversation:

- ***User***—The party who hopes to address an information need from a conversation.
- ***Customer service***—The party who provides information to the user during the conversation.

We note that the key difference between the users and customer service is not whether they ask questions or not, but their roles for addressing information needs. As we will discuss in the following sections, the customer service may also ask questions during a conversation to help users describe the information need.

3.2. User Message Categories and Intentions

We designed a classification scheme for user messages in human-AI collaborative conversations based on user's information needs. We followed SWBD-DAMSL and MSDialog (Jurafsky & Shriberg, 1997; Qu et al., 2018) to ensure that the classification scheme is reasonable, easy to annotate and train automatic classifiers. This classification scheme focues on assisting human-AI collaborative systems to understand user's information needs, which is of significance for indentifying when to change from AI to human staff, while previous taxonomies were mainly designed to improve the response accuracy of human-machine conversations. In addition to inheriting categories from previous taxonomies, we refined the classification scheme by taking into account the context of user messages, and create a pair of new categories that adapt to the human-AI collaborative conversations. Therefore, we categorize user messages in human-AI collaborative conversations into three possible intentions, five higher-level categories and 15 specific types, as shown in Figure 2. Table 1 illustrates the descriptions and examples of user message categories:

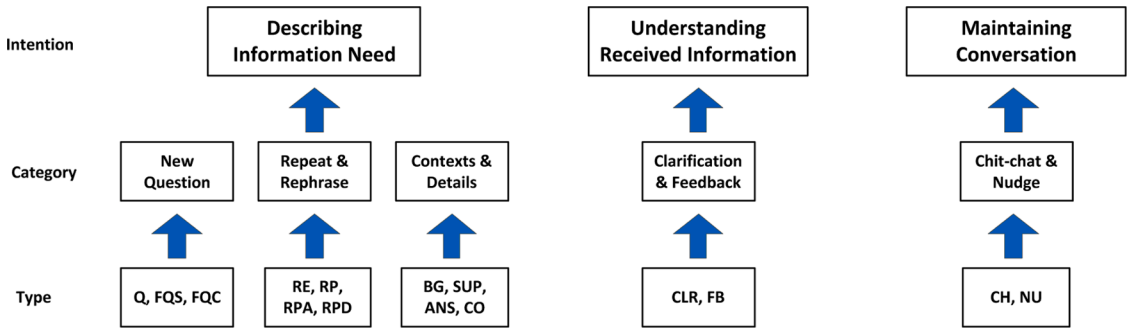

Fig. 2. User message categories and related user intentions during a conversation.

Table 1
Descriptions and examples of user message categories.

Intention	Category	Type	Description	Example
Describing Information Need	New Question	Q	A question that is not a follow up of any previous messages	So, it means that I won't get compensated if I don't spend over ¥ 10,000 a year, right?
		FQS	A question following up on a past message sent by the user.	I should ask social security or other health insurance for compensation if less than ¥ 10,000, right?
		FQC	A question following up on a past response from the customer service.	How long is afterwards?
	Repeat & Rephrase	RE	Restating a previous question without any change	I would like to know if I can get my insurance fee back.
		RP	Restating a previous question with only wording differences	I was asking if I can get my insurance fee back.
		RPA	Restating a previous question with more content	I do not have an insurance yet. I am hospitalized now. Will it be covered?
		RPD	Restating a previous question with less content	So, child has no social security, right?
	Contexts & Details	BG	User's self-initiated messages providing background information before a question.	I was hospitalized for fracture last year
		SUP	User's self-initiated messages providing supplement information after a question.	I am 49 years old.
		ANS	User's messages responding to questions from the customer service.	Yes, I have.
		CR	Correcting typos or incorrect details in a previous message	My family name is Lee, not Li.
Understanding Received Information	Clarification & Feedback	CLR	Clarifying the meaning of the customer service's responses.	So, I can only be covered after 30 days
		FB	Providing (negative) feedback and requests to the customer service	Sorry, I can't understand, can you make it simpler?
Maintaining Conversation	Chit-chat & Nudge	CH	Individual messages such as greeting and goodbye messages	Hello!
		NU	Trying to notify the customer service and urge them to respond	Is anyone there?

- ***Describing information need***—A substantial number of user messages aim to describe information needs, which is not necessarily a one-shot process. On one hand, a user may have several related but different needs. On the other hand, a single message may not convey a need well. We concluded three message categories for describing information needs: those for asking new questions, refining previous questions, and providing contexts and supplementary details.
- ***Understanding received information***—The intention of some of the user's messages is to understand the information received from the customer service. This includes messages about clarifying the received responses and providing feedback to the customer service's responses.
- ***Maintaining conversation***—Some of the user messages do not directly address information needs but help maintain an active conversation, e.g., chit-chat messages such as greetings, nudge messages that check if the customer service is available.

We summarize this classification system based on two annotators' examination of a real-world dataset and previous studies of conversational dialog system acts. The dataset includes informational conversations between customers (users) and online representatives (customer service) from a health insurance provider company in China. The representatives had both human workers and AI chatbots—we decided not to categorize customer service responses as we are just interested in how users interact with the human-AI collaborative customer service. All messages are in Chinese (Mandarin). We report examples translated into English in the paper and enclose the original Chinese messages in the appendix for reference.

3.2.1. New Questions

A user message may describe a new question whose information need has not been asked before in the conversation session. We further divide such messages into three types depending on whether and how these messages relate to past messages in the same session:

- ***Question (Q)***—A question that is not a follow up of any previous messages.
- ***Follow-up Question, Self (FQS)***—A question following up on a past message sent by the user.

Table 2
An example message classified as Follow-up Question, Self (FQS).

Q	So, it means that I won't get compensated if I don't spend over ¥ 10,000 a year, right? Correct.
FQS	*I should ask social security or other health insurance for compensation if less than ¥ 10,000, right?* Yes, correct.

Table 3
An example message classified as Follow-up Question, Customer Service (FQC).

Q	If I am hospitalized, do I need to pay for the expenses myself first?
	Hello, this insurance is a reimbursement insurance. You need to pay for the expenses first yourself and claim reimbursement afterwards from the insurance company with your supporting documents.
FQC	*How long is afterwards?*
	Sorry, I can't understand what you meant.

- ***Follow-up Question, Customer Service (FQC)***—A question following up on a past response from the customer service.

Tables 2 and 3 show example messages that were classified as follow-up questions. We use shaded cells for user messages in all the following examples. The message in Table 2 is classified as FQS because it is related to the previous user message regarding health insurance coverage with fewer than ¥ 10,000 annual expenses. The message in Table 3 is classified as FQC because it is following up the customer service's response mentioning claiming reimbursement afterwards.

3.2.2. Repeat & Rephrase

A user message may also restate a question where the same information need has been expressed before in the session. We further divide such messages into four types depending on the difference between the message and previous ones stating the same information need:

- ***Repeat (RE)***—Restating a previous question without any change.
- ***Rephrase (RP)***—Restating a previous question with only wording differences.
- ***Rephrase, Add (RPA)***—Restating a previous question with more content.
- ***Rephrase, Delete (RPD)***—Restating a previous question with less content.

Tables 4, 5, and 6 show some example messages that were classified as rephrased questions in our dataset. The key difference between rephrased questions and FQS (a question following up a previous user question) is whether the same information need has been expressed before (despite any content difference). As the example messages show, a rephrased question states the same information with some previous user messages, even though some content may have been added or removed. In contrast, the example FQS message in Table 2 is related to a previous Q message, but it conveys a different question. The reason for repeating and rephrasing questions is mostly because the customer service was not able to provide effective responses.

3.2.3. Contexts and Details

A user message may not ask a question but provide contexts and details to enrich information needs. Such messages can be either self-initiated or elicited by the customer service. We summarized four types of such messages:

- ***Background (BG)***—User's self-initiated messages providing background information *before* a question.
- ***Supplement (SUP)***—User's self-initiated messages providing supplement information *after* a question.
- ***Answer (ANS)***—User's messages responding to questions from the customer service.
- ***Correction (CO)***—Correcting typos or incorrect details in a previous message.

Tables 7 and 8 show examples that were classified as Background and Supplement messages in our dataset. We define Background and Supplement messages as those that are not questions and only describe contexts or details. In contrast, users may also include more contexts and details while rephrasing a question, but such rephrased messages are stating questions. We suspect that an important reason for sending Background and Supplement messages is that users may not want to draft long messages in online text chatting, especially on mobile devices. In such a case, it is natural to split a long question into separate messages, where some of the messages may only provide contexts and details. Table 8 also shows a message classified as Answer. This message also provides context to the problem but is elicited by the customer service.

Correction messages make up a very small fraction of our dataset (0.9%). They include both messages that only rectify the incorrect part of a previous message and those restating a corrected question. Theoretically, we can further conclude an individual type, "Rephrase, Correction", for the latter case. But here we simply count all these messages into one type since they are very rare in our dataset.

Table 4
An example message classified as Rephrase.

Q	Will my insurance fee be returned?
	Our insurance has a high compensation rate. You can pay as low as ¥ 100 to be compensated up to ¥ 6 million. You can check the insurance fee for different ages by clicking on micro app—health insurance—estimate my first-year insurance fee.
RP	*I was asking if I can get my insurance fee back.*

Table 5
An example message classified as Rephrase, Add.

Q	Will it be covered if I am hospitalized? Our insurance is a compensatory health insurance. The covered expenses mainly include hospitalization expenses, specialized clinics expenses, surgery expenses, ER expenses before and after hospitalization. Our coverage is not limited to hospitalization expenses.
RPA	*I do not have an insurance yet. I am hospitalized now. Will it be covered?*

Table 6
An example message classified as Rephrase, Delete.

Q	My child is 3. Should I check the no social security option? Please wait a second while I am answering your question. Hello! Both people with and without social security are eligible for our insurance, but the fees are different.
RPD	*So, child has no social security, right?*

Table 7
An example message classified as Background (BG).

BG	*I was hospitalized for fracture last year* You are eligible for the insurance if you have been hospitalized in the past two years for the following reasons: A) labour; B) acute respiratory diseases; C) acute gastroenteritis or appendicitis; D) gallstones that did not relapse in two years; E) benign gallbladder polyps; F) accidental hospitalization recovered in 5 days without sequelae or loss of any organ.
Q	Am I not eligible?

Table 8
Example messages classified as Supplement (SUP) and Answer (ANS).

Q	Will the insurance rate increase every year? As one gets older, the risk of having an accident or disease increases too, and the insurance rate also increases. But our insurance aims to be inclusive. Even for people over 60 years old, they only need to pay a little more than ¥ 1000 a year (i.e., roughly ¥ 100 monthly rate) to have an insurance that can compensate up to ¥ 6 million. It is a highly cost-effective product that everyone can afford.
SUP	*I am 49 years old.* Please wait a second while I am answering your question. May I ask if you have social security? [rose emoji]
ANS	*Yes, I have.* Hello, the rate for 49 years old with social security is ¥ 858 a year.

3.2.4. Clarification & Feedback

A user's message may not describe an information need but be to communicate with the customer service to better understand the responses. We have concluded two types of messages for this purpose:

- ***Clarify (CLR)***—Clarifying the meaning of the customer service's responses.
- ***Feedback (FB)***—Providing (negative) feedback and requests to the customer service.

Table 9 shows an example message classified as Clarify (CLR) in our dataset. A Clarify message may also be a question, which makes it look similar to a Follow-up question related to the customer service (FQC). However, the key difference between them is that a Clarify message does not convey new information need, but a follow-up question does express one by our definition.

Table 10 shows an example message classified as Feedback in our dataset. The message provides feedback to the previous responses and requests for a simpler answer. Feedback messages take only a small fraction of all the messages in our dataset (0.9%). Most of them expressed negative feelings towards customer service responses. The purpose of the Feedback messages is to notify customer service and urge them to improve response quality.

3.2.5. Chit-chat & Nudge

We have also identified many Chit-chat (CH) messages and a very small number of Nudge (NU) messages in the dataset. Chit-chat includes individual messages which not provides any information needs, but only used to maintain the conversation, such as greeting and goodbye messages. In this paper, those user messages that are difficult to be classified into other categories but has the attribute of maintaining conversation were classified as Chit-chat. We define Nudge messages as those trying to notify the customer service and urge them to respond (e.g., is anyone there, are you still there). These messages do not describe information need, but they are useful for keeping the conversation natural and active.

Table 9
An example user message classified as Clarify (CLR).

Q	How soon I will be covered after buying this insurance?
	There is a 30-day waiting period starting the midnight of the next day you have purchased our insurance.
	The waiting period is also called the observation period, which refers to the 30-day period after purchasing the insurance. During this period, the customer's medical expenses caused by diseases will not be covered. The purpose of having a waiting period is to guarantee the benefits of healthy customers and to avoid people buying this insurance while they have already been diagnosed with certain diseases such that we can ensure the benefits of the majority.
CLR	*So, I can only be covered after 30 days*
	Correct.

Table 10
An example user message classified as **Feedback**.

Q	Okay, effective starting the midnight of the next day of purchasing, may I explain this item as that I can get covered for serious diseases starting the next day after I purchased this insurance?
	Hello, customers are not covered for diseases within 30 days after purchasing this insurance.
	But the waiting period does not include accidents [happy emoji].
FB	*Sorry, I can't understand, can you make it simpler?*
	Please wait a second while I am answering your question.
	To make it simpler: after successfully purchasing our insurance, you have no waiting period for accident compensation, but there is a 30-day waiting period if you are hospitalized for diseases, which is the same as the observation period you mentioned [happy emoji].
	Does that make more sense to you?

4. Data and Annotation

4.1. System

Our dataset comes from a company's online text-based customer support log. The company is a primary health insurance provider in China. Thus, all the conversations are related to health insurance. The company provides online customer support through a popular mobile messaging app in China where customers can communicate with the customer service by sending messages.

The company's online customer service employs a human-AI collaborative model, including a group of human representatives and an AI chatbot. For an incoming customer message, the system first computes a chatbot message response. If the chatbot's confidence level (an internal measure indicating the quality of the generated message) is below a certain threshold, the system will try to switch to a human representative. If all human representatives are busy, the system will let the customer wait. If the waiting time exceeds a threshold, the system will send an automatic response, "Please wait for a second while I am answering your question," to notify the customer. We also include this automatic message in the conversation if the customer received it. If all human workers are offline (such as during nighttime), the system will respond with AI chatbot messages even if they have low confidence scores.

4.2. Dataset

We created a dataset of informational conversations based on the company's customer service conversation log. The log includes conversations between October 2017 and January 2018. Although we couldn't get the latest service log due to the limitation of privacy policy, we believe that the current dataset still has the significance of analyzing the user intention. Therefore, we further divide a conversation session into three types based on the types of representatives involved:

- **AI-only**—All the responses are from an AI chatbot.
- **Human-only**—All the responses are from a human representative.
- **Hybrid**—The session includes responses from both AI chatbot and human representative.

We randomly sampled 100 sessions for each type. The dataset includes 300 conversations, involving 1,478 customer messages and 1,936 representative responses. We examined all the conversations manually to make sure they are informational conversations. We have excluded sessions where customers had sent multiple consecutive messages without receiving any responses (0.6% of the sessions in the log belong to this type). We further define a *round* of a conversation as the period from one user message (inclusive) to the next user message (exclusive). Thus, each round in our dataset includes one and only one user message but may have one or multiple customer service responses.

Note that our selection of AI-only, Human-only, and Hybrid sessions is quasi-experimental. Particularly, a conversation may end up being AI-only just because the chatbot has high confidence scores for all customer messages. Thus, we suspect the complexity and difficulty of customer questions in the three types of sessions may vary. One should be cautious when reading our results comparing the three session types, because the differences may not entirely come from the different types of representatives involved in the conversations.

4.3. Annotation Procedure and Consistency

Two of the authors independently annotated the dataset to categorize user messages into different types. All the messages are using Chinese Mandarin, and both annotators are also native Chinese Mandarin speakers, to make sure they can correctly understand the messages.

We (including both annotators and another author) first discussed and then produced an initial classification scheme based on a small sample of the data (including 100 messages). The initial scheme also borrowed ideas from previous studies of dialog system acts. The initial scheme had included 11 types and did not involve Background, Supplement, Correction, and Feedback. The two annotators discussed cases that they could not categorize into the initial scheme during the annotation process and gradually enriched the scheme to the form we introduced in Section 3.

The two annotators' results have a moderate consistency—the overall Cohen's Kappa on the whole dataset (including the three types of sessions) is 0.59. The agreements are higher in human-only sessions (0.67) but lower in hybrid ones (0.55). Further, the two annotators discussed the messages on which they disagreed and came to an agreement on them all. Table 11 reports the consistency between each annotator's results and the final agreed types after discussion.

For 11 out of the 15 message types (excluding Q, CH, Nudge, and Feedback), the two annotators had also identified their related messages/responses in the session. For example, the message related to an FQS is the user's previous message that the FQS followed up, and the message related to a CLR is the response that the user was trying to clarify. The two annotators also had high agreement on the identified related messages. Among those messages where both annotators agreed on the message type, they also agreed on 83.6% of the related messages.

5. Message types and characteristics

5.1. Overall Statistics

Table 12 reports overall statistics about sessions, rounds, and messages in our dataset. The results suggest that the three types of sessions in our dataset are very different in many aspects.

First, users had used many more rounds to finish conversations in the Hybrid sessions (7.26 on average, in contrast to 4.42 for AI-only and 3.10 for Human-only sessions). These differences consequently made Hybrid sessions differ greatly to the other two types in the number of messages and responses at a session level, although users had received significantly more responses during a round in Human-only sessions (1.73) than in Hybrid ones (1.32). Note that every AI session round included consistently one response because the chatbot is designed to always respond with only one message for each request. We suspect the high number of rounds in a session may indicate that customers had low conversation quality in Hybrid sessions, as previous studies had also identified long search sessions and dialog sessions as negative signals for search/conversation quality.

Second, the messages and responses from AI and Human sessions also differ significantly in length (by the number of Chinese characters), though we found neither of them had any significant difference to messages in Hybrid sessions. The length of responses also differs greatly between AI-only and human-only sessions. This suggests that AI and human representatives are providing very different responses, and user messages may also be different in these two session types (13.08 vs. 15.61 characters).

To conclude, many statistics in Table 12 show that the conversations and messages in the three session types have lots of differences in our dataset. Such differences may come from the influence of the different types of representatives in these sessions, our selection criterion when building this dataset, or both. Yet further investigation is needed to better understand such differences.

5.2. User Message Type Distribution

Although the three types of sessions in our dataset vary significantly in many statistics, we found that the frequency of using message types in the sessions is mostly consistent, with noticeable differences only in a few message types. Table 13 reports the percentage of each message type in the three types of sessions. We group some message types because they appeared a minimal number of times in our dataset—we group all four types of Repeat and Rephrase message types together as *REP*, and we group Correction, Nudge, and Feedback as *OTHER*.

First, our results show that the frequency of using different types of messages is highly consistent across sessions with various representatives. We compared the overall distribution of message types in the three types of sessions using a Kruskal-Wallis H test. The test results suggest no significant differences between any of the session types regardless of using the original 15 message types ($p = 0.191$) or the grouped 10 types ($p = 0.537$). This indicates that 1) our message classification scheme is highly generalizable and can be applied to different types of sessions, and 2) the use of different message types seems relatively stable when communicating with AI and human representatives (though it is unclear whether the results would remain the same if the customers were told which type of representatives they were talking with).

Second, results in Table 13 also disclosed the popularity of the message types in conversations. In all three types of sessions, Q, FQO, and CH remain the top three most popular types. About half of the messages in the sessions are directly asking new questions (Q, FQS, and FQO make up 55% of all messages), with over 1/3 being follow-up questions (19.8% out of 55%). Additionally, the purpose of 18.3% of the messages is to provide contexts and details (BG, SUP, ANS, and Correction). Users also used 16.5% of the messages (Chit-chat and Nudge) to maintain conversations. In contrast, restating questions (REP) and understanding customer service's responses (CLR and Feedback) take up only 5.1% and 5% of the total, respectively. It is worth noting that we compared the distribution of user

Table 11
Cohen's Kappa amongst the two annotators' results and the final results after discussion.

	ALL	AI-only	Human-only	Hybrid
Cohen's Kappa: Annotator 1 vs. 2	0.59	0.58	0.67	0.55
Cohen's Kappa: Final vs. Annotator 1	0.74	0.71	0.82	0.73
Cohen's Kappa: Final vs. Annotator 2	0.80	0.83	0.84	0.77
% agreed related messages	83.6%	81.2%	85.7%	84.2%

Table 12
Mean and standard error of various statistics for sessions, rounds, and messages in different sessions. We test significant differences using a one-way ANOVA with the Tukey HSD post hoc test.

User Message Category	ALL	AI-only	Human-only	Hybrid	$P < 0.05$ Differences
# messages & responses / session	11.38 (0.54)	8.84 (0.76)	8.47 (0.42)	16.83 (1.19)	Hybrid > AI, Human
# user messages / session	4.93 (0.24)	4.42 (0.38)	3.10 (0.15)	7.26 (0.51)	Hybrid > AI > Human
# AI responses / session	2.37 (0.19)	4.42 (0.38)	–	2.68 (0.26)	AI > Hybrid
# human responses / session	4.09 (0.28)	–	5.37 (0.28)	6.89 (0.59)	Hybrid > Human
# rounds / session	4.93 (0.24)	4.42 (0.38)	3.10 (0.15)	7.26 (0.51)	Hybrid > AI > Human
# messages & responses / round	2.31 (0.02)	2.00 (0.00)	2.73 (0.05)	2.32 (0.02)	Human > Hybrid > AI
# user messages / round	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	–
# AI & human responses / round	1.31 (0.02)	1.00 (0.00)	1.73 (0.05)	1.32 (0.02)	Human > Hybrid > AI
# AI responses / round	0.48 (0.01)	1.00 (0.00)	–	0.37 (0.02)	–
# human responses / round	0.83 (0.03)	–	1.73 (0.05)	0.95 (0.04)	Human > Hybrid
user message length (# chars)	14.14 (0.28)	13.08 (0.51)	15.61 (0.56)	14.16 (0.41)	Human > AI
Customer Service response length (# chars)	58.91 (1.18)	86.83 (2.60)	46.53 (2.18)	52.96 (1.49)	AI > Human, Hybrid

Table 13
Distribution of user message categories in sessions with AI, human, and hybrid representatives. For each message type, we test significant differences of three session types using the Chi-square test with Bonferroni correction.

User Message Type	ALL	AI-only	Human-only	Hybrid	$P < 0.05$ Differences
Q (new query)	35.5%	38.9%	48.1%	28.0%	Hybrid < AI < Human
FQS (follow-up query, self)	5.3%	5.4%	4.8%	5.5%	–
FQO (follow-up query, the other person)	14.5%	13.3%	14.2%	15.4%	–
REP (repeat & rephrase)	5.1%	6.8%	2.9%	5.0%	–
RE (repeat without any change)	0.9%	0.7%	1.0%	1.0%	–
RP (rephrase; wording difference)	1.8%	3.2%	0.0%	1.8%	Human < AI
RPA(rephrase; added some content)	1.8%	2.0%	1.3%	1.9%	–
RPD(rephrase; removed some content)	0.5%	0.9%	0.6%	0.3%	–
CLR (clarify)	4.1%	3.4%	4.2%	4.4%	–
ANS (answer)	3.2%	0.2%	4.5%	4.4%	AI < Human, Hybrid
CH (chit-chat)	16.2%	15.8%	13.2%	17.6%	–
BG (background information)	6.6%	7.9%	1.6%	7.9%	Human < AI, Hybrid
SUP (supplementary information)	7.6%	6.3%	4.8%	9.5%	Human < Hybrid
OTHER (other types)	2.0%	1.8%	1.6%	2.3%	–
CO (correction)	0.9%	0.7%	0.3%	1.2%	–
NU (nudge)	0.3%	0.2%	0.6%	0.1%	–
FE (feedback)	0.9%	0.9%	0.6%	1.0%	–

Kruskal-Wallis H test (H_0: message category distributions in AI-only, Human-only, and Hybrid sessions are not significantly different): $P = 0.537$.

message categories in previous studies and found that the results was very different. Although different schemes often do not cover all aspects necessary for open-domain human-machine or human-human interaction, which may be an important reason for the difference of distribution results. For example, New Query accounts for 35.5% in our study, 13% in Qu's study (Qu et al., 2018) and 7.5% in

Ahmadvand's study (Ahmadvand et al., 2019). Studies have confirmed that these differences mainly come from the dataset used. Conversations in different situations result in the characteristics of user message category distribution (Mezza et al., 2018). However, the distribution of user message categories provides important guidelines for designing chatbots that can better respond to different types of user messages.

Third, the message type distribution also suggests that a substantial number of messages in a conversation session are closely linked with some other messages/responses in the same session. By our definition, FQS, FQO, REP, CLR, ANS, BG, SUP, and Correction messages all have related messages or responses. They take up 47.1% of all users' messages. This shows that the messages and responses in a conversation are highly related to each other, suggesting the importance of modeling context information in designing chatbots.

Fourth, we also observed significant differences for a few specific message types across different session types. Some of the differences are related to the settings of the customer service system, e.g., we only observed one case of Answer in AI-only sessions because the AI chatbot does not provide questions as responses. For the other differences, we suspect one possible reason lies in the question-routing strategy in the customer service system—for example, if the system has higher confidence scores for a certain message type, those messages are less likely to be routed to human representatives, and thus will have a lower percentage in the human-only sessions. However, further study is required to understand the differences regarding using messages in sessions with different representatives.

5.3. Characteristics of User Message Types

After examining the content of user messages, we found that different types of messages vary greatly in length (Figure 3) and in their similarity to the identified related messages/responses in the session (Figure 4). Here, we measure message length by the total number of Chinese characters. We measure the similarity of a user message to its related message/response by the percentage of the common content (by Chinese character unigrams or bigrams) in the user message itself.

Note that a Chinese word typically includes one, two, or three characters, where we can roughly equivalate a Chinese character to a word root or stem in English. The whole Chinese character set includes over 50,000 different characters, with about 3,500 frequently used ones. We did not examine messages by words because we found that out-of-the-box word segmentation tools[1] did not work well on our dataset (probably because the text messages are noisy and used many verbal expressions).

Figure 3 shows that some types of messages are much longer than others. Particularly, FQO, REP, ANS, and CLR messages are significantly longer than BG, SUP, CH, and OTHER messages in our dataset using a Tukey HSD post hoc test (the difference of each pair is at least significant at 0.05 level).

Figure 4 reports the percentage of overlap character unigrams (solid color bars) and bigrams (pattern-filled bars) between a user message and its identified related message/response in the dataset. We did not report results for Q and CH because they do not have related messages/responses. Figure 4 also shows the overlap values with a most recent user message (baseline 1) and a most recent customer service response (baseline 2) across the whole dataset for comparison. The results for overlap unigrams and bigrams are mostly consistent.

Results show that, except BG, SUP, and ANS, the other types of user messages share much more common content with their identified messages or responses than two random adjacent messages/responses (baseline 1 and 2). This also demonstrates that the

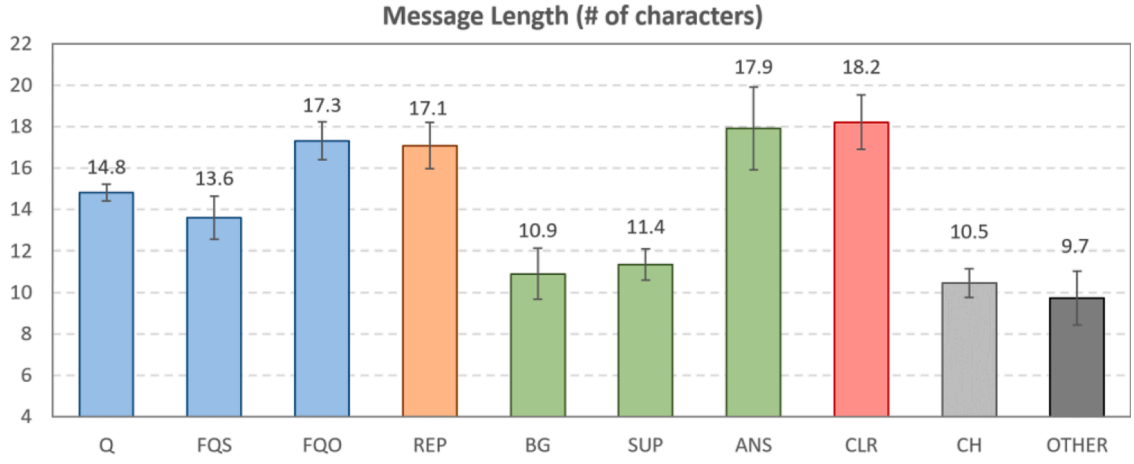

Fig. 3. Length of messages classified into different types (by the number of Chinese characters).

[1] The Chinese writing system does not put a white space between words; thus we need to use NLP tools to segment words from text.

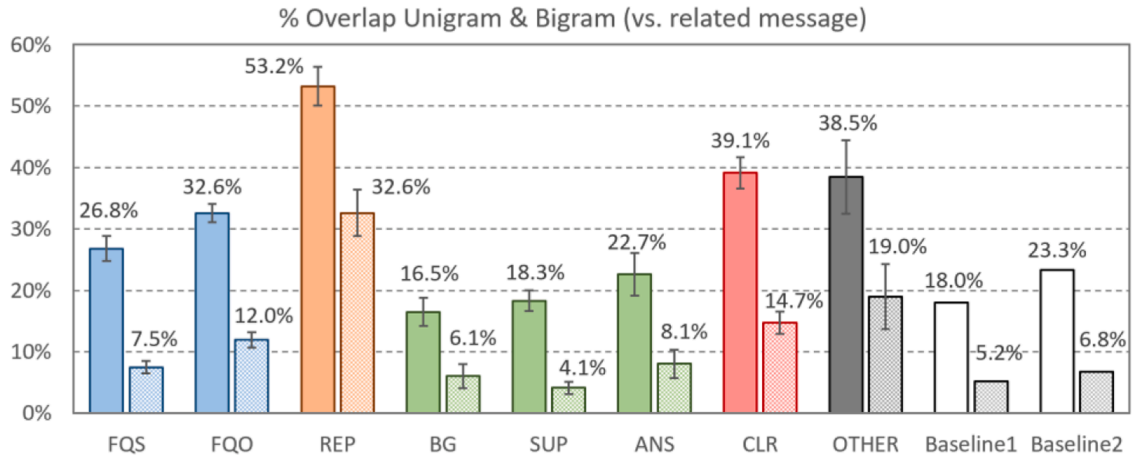

Fig. 4. The percentage of overlap unigrams and bigrams (pattern-filled bars) between a message and its identified related message/response. Baseline 1 is the average overlap with a most recent user message. Baseline 2 is the average overlap with a most recent answer response.

manually labeled related messages/responses are probably accurate.

Results also show that overlap percentages vary a lot in different message types. The overlap unigrams take up over 50% of the content in REP messages, 30%–40% of the content in FQO, CLR, and OTHER messages, and lower proportions in other message types. The highest percentage of overlap content in REP messages is not surprising since the intention of the REP messages is to restate their related messages. The high overlap content percentages in FQO and CLR messages are probably because the users need to refer to the overlap content when asking follow-up questions or clarifying previous responses.

To conclude, results in this section show that different message types vary in characteristics related to their contents, which provides potential opportunities to recognize message types automatically based on contents.

5.4. Message Type Transition: After a Q Message

We further examine the use of different message types in a contextual manner—such as right after or before a message type. Figure 5 illustrates the transition probabilities to different message types after a Q message, i.e., the chances of having different message types if the previous user message is a Q. We also calculate the chance that the Q message is the last user message in the session (Q→END). We separately examine each individual session type and all sessions. We focus on Q messages because they are the most common message type.

Figure 5 shows that the use of different message patterns right after a Q message is mostly consistent across different session types, with some differences. The top three most frequent message types after Q are Q, FQO, and END (which means that the Q is the last user message of the session). Additionally, CH, SUP, FQS, and REP are also relatively frequent types. This suggests that the main pattern of an informational session is to keep on asking questions (including follow-up questions), occasionally with other messages to refine and complement the questions or clarify received responses.

We have also observed some differences in message type transition in different sessions. Particularly, we noticed that the chance of Q→REP is much lower in human-only sessions, indicating that human-only sessions probably have better response quality (such that users do not need to restate the same needs multiple times). To conclude, these results indicate the possibility that the usage of message types may be related to the type of representatives in a session, but further study is required to verify this due to our quasi-experimental design.

6. Discussion

First, we have concluded and introduced a classification scheme for categorizing user messages in human-AI collaborative conversations based on users' informational needs. Although many previous studies examined user request patterns in conversational dialog systems, and intelligent personal assistants, but our study and scheme are novel from two aspects: 1) informational conversation shares some similarities with but is very different from these applications, e.g., it provides more interactive communication and direct access to the information, it focuses on information seeking and acquisition tasks compared with dialog systems and intelligent assistants; 2) we design our classification scheme from a novel aspect and link the message categories with higher-level user intentions for information acquisition.

We have also demonstrated that the classification scheme is practical and actionable. As we described, we have successfully annotated a real-world dataset with highly specialized conversation topics (medical health). Our annotators do not have any prior knowledge related to this specialized topic, but they were still able to come to very reasonable agreements during the annotation. We acknowledge that the messages in our data are in Chinese, but our classification scheme does not include rules or details related to the

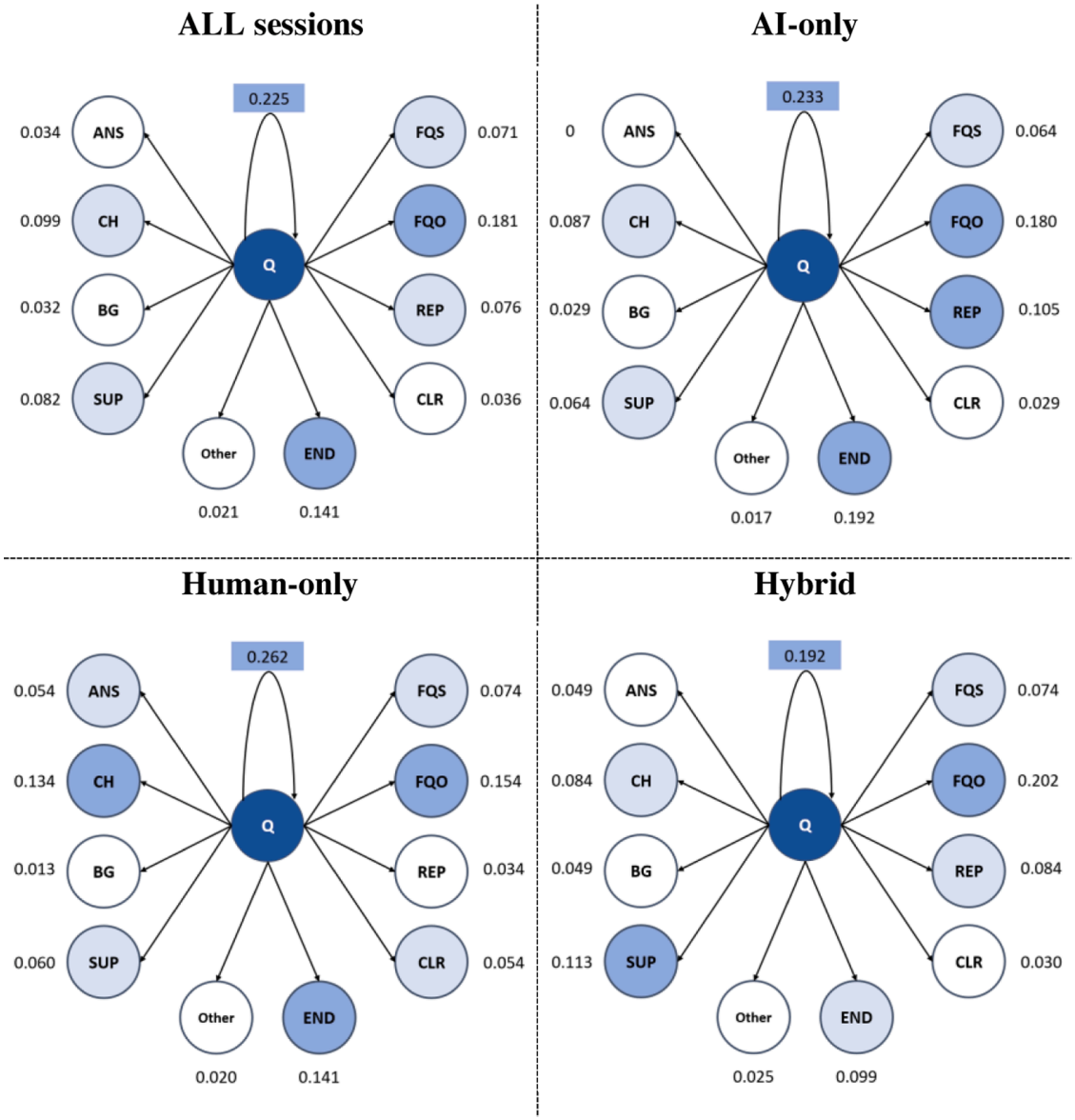

Fig. 5. User message category transition probabilities after a Q message (Q→?) in different types of sessions.

language used in the conversation. This suggests that the proposed scheme is likely to generalize to other scenarios and for lay people.

Second, we have also presented detailed comparisons among the different message types. The results provide insights to understand different types of user messages in a session. Particularly, we have observed that many types of messages vary significantly in content characteristics, such as length and their similarities with other messages in a session. On the one hand, this provides a second look into the validity of our classification scheme and data annotation consistency, because many of the observed differences can be explained well based on the definition of our message types (e.g., most message types have high content similarities with their related messages or responses). On the other hand, this offers clues to design techniques to recognize message types automatically—for example, message length and content similarity with previous messages and responses may be effective features for automatically classifying user message types.

Third, we have also presented an initial exploration of the possible relations between user message types and sessions with different types of representatives (AI-only, human-only, and hybrid). Our initial observation is that the use of message types is quite stable across sessions with different types of representatives, but we did also observe that the usage frequency for some specific message types (such as Q, REP, BG, and SUP) can be significantly different in different types of sessions. This provides a basis for further studies to examine the relationship between different types of representatives and users—we believe this is a fundamental research question in human-AI collaborative conversations as it is more and more common to offer hybrid chat services.

Our work also has some limitations. We leave them for future work. First, we acknowledge that our dataset included only a very

specialized topic (health insurance) and is in Chinese. We advise future work to further verify the generalizability of our scheme and findings (though we believe that our study can easily be replicated in other languages, e.g., one can examine word-based unigram and bigram overlap in English language datasets). Second, our selection of sessions with different types of representatives is quasi-experimental, and we are aware that the selection may be affected by some message characteristics (i.e., the chatbot's confidence for responding to these messages). And the selection may also be affected by user's preference when they talking to different representatives. Although users will not be informed of which agent they are engaging when using the dialogue system, users can clearly feel whether they are having conversation to human or AI. Users may adopt different converssation preferences when talking to different representatives, for example, users tend to add more background information to help AI better understand their problems when talking with AI. Thus, we also suggest further studies to utilize randomly assigned experiments to examine the relationship between the types of representatives and the use of messages.

7. Conclusion

In this paper, we introduced a fine-grained hierarchical classification scheme for user messages in human-AI collaborative conversations based on users' informational needs. Our scheme included five categories of messages (15 specific types) linked to three higher-level user intentions: describing information needs, understanding information, and maintaining conversation. The detailed message types share some similarities with previous studies of dialogue acts, but we put a special focus on the function of the message for assisting users during the conversation to acquire information. We have also examined the annotation results on a real-world dataset and reported statistics comparing different message types and in sessions with different representatives.

To the best of our knowledge, this is the first classification scheme for human-AI collaborative conversations which shed light on understanding a wide range of real-world applications, especially human-AI collaborative customer services. We believe our novel classification scheme provides significant guidance on future work related to online informational conversation, including work for both understanding human factors and designing systems and interactive techniques. For example, researchers may apply our scheme to annotate and examine informational conversations, and systems may design techniques for classifying user message types, recognizing related messages, and prepare specialized responses accordingly in the future.

8. Author contributions

Yuhan Wei: Conceptualization, Methodology, Formal analysis,Writing -Original Draft
Wei Lu: Supervision, Formal analysis, Writing - Review & Editing
Qikai Cheng: Conceptualization, Methodology, Data Curation
Tingting Jiang: Investigation, Writing - Review & Editing
Shewei Liu: Investigation, Data Curation

Declaration of Competing Interest

The authors declare no competing interests.

References

Ahmadvand, A., Choi, J. I., & Agichtein, E. (2019). Contextual Dialogue Act Classification for Open-Domain Conversational Agents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1273–1276).

Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., et al. (2020). Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review, 36*, Article 100239.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science (New York, N.Y.), 358*(6370), 1530–1534.

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research, 117*, 587–595.

Core, M. G., & Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. *AAAI fall symposium on communicative action in humans and machines, 56*, 28–35.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304–316.

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop silicon valley from building a new global underclass*. Boston, MA: Houghton Mifflin Harcourt.

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Computers in human behavior, 49*, 245–250.

Hohenstein, J., & Jung, M. (2018). AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems(Paper LBW089* (pp. 1–6).

Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y., Inaba, M., et al. (2019). Overview of the sixth dialog system technology challenge: DSTC6. *Computer Speech and Language, 55*, 1–25.

Isbister, K., Nakanishi, H., Ishida, T., & Nass, C. (2000). Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 57–64).

Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., et al. (2015). Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 506–516).

Jie, G., & Zhang, Z. (2011). User Satisfaction of Ali Wangwang, an Instant Messenger Tool. In A. Marcus (Ed.), *Design, user experience, and usability. theory, methods, tools and practice. duxu 2011. lecture notes in computer science*. Berlin, Heidelberg: Springer. vol 6770.

Jurafsky, D., & Shriberg, E. (1997). Switchboard SWBD-DAMSL labeling project coder's manual. *Technická Zpráva*, 97–02.

Kušen, E., & Strembeck, M. (2020). You talkin' to me? Exploring Human/Bot Communication Patterns during Riot Events. *Information Processing & Management, 57*(1), 102–126.

Licklider, J. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics, HFE-1*, (1), 4–11.
Liu, J., Gao, Z., Kang, Y., Jiang, Z., He, G., Sun, C., et al. (2020). Time to transfer: Predicting and evaluating machine-human chatting handoff. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(7), 5841–5849.
Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., et al. (2019). AI-based digital assistants: Opportunities, threats, and research perspectives. *Business and Information Systems Engineering, 61*(4), 535–544.
Mezza, S., Cervone, A., Tortoreto, G., Stepanov, E. A., & Riccardi, G. (2018). ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. *COLING 2018*.
Neff, G., & Nagy, P. (2018). Agency in the Digital Age: Using Symbiotic Agency to Explain Human–Technology Interaction:. *A networked self and human augmentics, sentience* (pp. 113–123). London: Routledge.
Oraby, S., Bhuiyan, M., Gundecha, P., Mahmud, J., & Akkiraju, R. (2019). Modeling and Computational Characterization of Twitter Customer Service Conversations. *ACM Transactions on Interactive Intelligent Systems, 9*(2–3), 1–28.
Oraby, S., Gundecha, P., Mahmud, J., Bhuiyan, M., & Akkiraju, R. (2017). How May I Help You?" Modeling Twitter Customer ServiceConversations Using Fine-Grained Dialogue Acts. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 343–355).
Osterlund, C., Jarrahi, M. H., Willis, M., Boyd, K., & Wolf, C. T. (2021). Artificial intelligence and the world of work, a co-constitutive relationship. *Journal of the Association for Information Science and Technology, 72*(1), 128–135.
Pareti, S., & Lando, T. (2018). Dialog Intent Structure: A Hierarchical Schema of Linked Dialog Acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.
Qu, C., Yang, L., Croft, W. B., Trippas, J. R., Zhang, Y., & Qiu, M. (2018). Analyzing and Characterizing User Intent in Information-seeking Conversations. Paper presented at the. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 989–992).
Radziwill, N., & Benton, M. (2017). Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional, 19*(3), 25–35.
Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing & Management, 57*(1), Article 102078.
Sandor, A., Lagos, N., Vo, N.-P.-A., & Brun, C. (2016). Identifying User Issues and Request Types in Forum Question Posts Based on Discourse Analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 685–691).
Sterlund, C., Jarrahi, M. H., Willis, M., Boyd, K., & Wolf, C. T. (2020). Artificial intelligence and the world of work, a co-constitutive relationship. *Journal of the Association for Information Science and Technology, 72*, 128–135.
Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., et al. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics, 26*(3), 339–373.
Sun, H., Cheng, D., Wang, J., Qi, Q., & Liao, J. (2021). Pattern and content controlled response generation. *Information Processing & Management, 58*(5), Article 102605.
Tavakoli, L. (2020). Generating Clarifying Questions in Conversational Search Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 3253–3256).
Van den Broeck, E., Zarouali, B., & Poels, K. (2019). Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior, 98*, 150–157.
Wazurkar, P., Bhadoria, R. S., & Bajpai, D. (2017). Predictive analytics in data science for business intelligence solutions. In *Proceeding of 7th International Conference on Communication Systems and Network Technologies* (pp. 367–370).
Wood, A., Eberhart, Z., & McMillan, C. (2020). Dialogue Act Classification for Virtual Agents for Software Engineers during Debugging. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops* (pp. 462–469).
Yang, Z., Xu, W., & Chen, R. (2021). A deep learning-based multi-turn conversation modeling for diagnostic Q&A document recommendation. *Information Processing & Management, 58*(3), Article 102485.
Ye, N., Fuxman, A., Ramavajjala, V., Nazarov, S., McGregor, J. P., & Ravi, S. (2018). PhotoReply: Automatically Suggesting Conversational Responses to Photos. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1893–1899).