Collecting High-quality Multi-modal Conversational Search Data for E-Commerce

Marcus D. Collins **Eugene Agichtein** **Oleg Rokhlenko** **Shervin Malmasi**
Amazon.com, Inc. Seattle, WA, USA
{collmr,eugeneag,olegro,malmasi}@amazon.com

Abstract

Continued improvement of conversational assistants in knowledge-rich domains like E-Commerce requires large volumes of realistic high-quality conversation data to power increasingly sophisticated LLM chatbots, dialogue managers, response rankers, and recommenders. The problem is worse for multi-modal interactions in realistic conversational product search and recommendation. Here, an artificial sales agent must interact intelligently with a customer using both textual and visual information, and incorporate results from external search systems, such as a product catalog. Yet, it remains an open question how to best crowd-source large-scale, naturalistic multi-modal dialogue and action data, required to train such an artificial agent. We describe our crowd-sourced task where one worker (the Buyer) plays the role of the customer, and other (the Seller) plays the role of the sales agent. We identify subtle interactions between one worker's environment and their partner's behavior mediated by workers' word choice. We find that *limiting* information presented to the Buyer, both in their backstory and by the Seller, improves conversation quality. We also show how conversations are improved through minimal automated Seller "coaching". While typed and spoken messages are slightly different, the differences are not as large as frequently assumed. We plan to release our platform code and the resulting dialogues to advance research on conversational search agents.

1 Introduction

In recent years, researchers have investigated new approaches to build automated agents capable of naturalistic conversations satisfying complex information needs. The need for such assistance is particularly acute in domains like E-Commerce, where customers may even know what questions to ask when shopping. Creating an automated agent to help such customers is challenging, as it must serve as a natural conversational interface to many specialized and general data sources, while maintaining context to return valuable responses or make proactive suggestions. In high-stakes domains like E-Commerce, experimentation on real customers is risky, presenting a significant barrier to training and validating such conversational agents. Most task-oriented conversation systems, especially for E-Commerce, have been evaluated on synthetic data even though such data may not reflect real world conversations (Gangadharaiah et al., 2018). Thus there is a clear need for high-quality and realistic data to be collected to allow training more natural automated agents. Moreover, it isn't clear *how to generate natural data that reflects real-world conversations in domains like E-Commerce*. Although Large Language Models such as ChatGPT can be used to generate conversation, these are not realistic as shopping conversations are dynamic and customer goals can change as new information is presented (Bernard and Balog, 2023).

We explore crowd task designs needed to collect the *right data* to train conversational search and recommender agents, roughly following a theoretical framework (Radlinski and Craswell, 2017). Such agents combine aspects of a mixed-initiative chat-like dialogues with goal oriented systems, creating a number of unique challenges. Consider Example 1. The Customer seeks help from the Agent with a specific goal, buying a TV set, but the conversation is open-ended. Unlike, say, airline reservation systems, there are few constraints on the utterances. In addition, Agent and Customer share initiative. The Agent must choose from several information sources to help construct a reply. Figure 1 shows a standard framework to provide such an experience. A language understanding module extracts information from customer utterances, used by the dialog manager to query sources, including product catalogs, QA databases, *etc.* A ranker selects among candidates to produce each system response.

Example Dialog

Customer: I'm looking for a good sized TV for my living room.
Agent: Nice! How much are you looking to spend?
Customer: 400 dollars
Agent: Do you mostly watch TV shows, or maybe you want something for gaming?
Customer: I like to watch movies and play video games
Agent: How big? What about resolution?
Customer: at least 42 inches. what's resolution?
Agent: number of pixels per inch. Higher is better.
Customer: I don't care
Agent: Ok, do you care about refresh rate? Or brand?
Customer: I like Sony. It should be at least 120 frames per second.
Agent: OK, Sony has 3 TVs that match. Do you have another criteria you care about?
Customer: Not really.
Agent: Here's some options for you, what do you think of these? (pastes in links or uses sharing widget)
Customer: Ooh, there's one I like. I'll take a look. [One of the shared options; system evaluates it, workers may continue if they like it.]

Example 1: An example dialog illustrating a desirable E-Commerce conversation with an automated agent (when we refer to an Agent, we mean an automated agent).

To train such an agent, we need a corpus of high-quality naturalistic, free-form, but goal-oriented conversations labeled with each action's intent and overall conversation success. Recently one such dataset was released (Moon et al., 2020) with a large number of conversations related to furniture and fashion. However, there is no exploration of the conversations' naturalness, or of how the multi-modal environment affects the results. Moreover, it isn't clear that such a corpus is ecologically valid (Vries et al., 2020), i.e., it isn't clear that the information and tools provided to the workers realistically emulate how customers would interact with automated or human sales agents. To ensure that, we aim here to understand how the conversation participants' multi-modal environment affects the conversation, and to add to the available conversational search and recommendation data.

We review prior work in the next section to put our contributions in context. Section 3 describes how we crowd-sourced conversations. Section 4 covers the development of our crowd-sourced annotations and automated measures of the conversation. We analyze the collected conversation data in Section 5, and demonstrate the effectiveness of the collected data to enable a robust conversational search and recommendation agent. Finally, we discuss future work and potential extensions in Section 6.

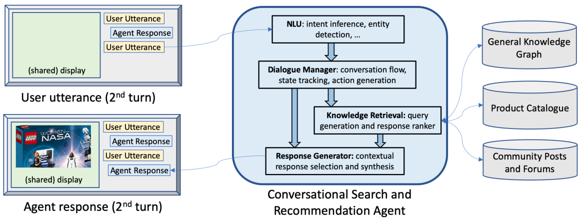

User Utterance
Agent Response
User Utterance
(shared) display
User utterance (2nd turn)
User Utterance
Agent Response
User Utterance
Agent Response
(shared) display
Agent response (2nd turn)
NLU: intent inference, entity detection, ...
Dialogue Manager: conversation flow, state tracking, action generation
Knowledge Retrieval: query generation and response ranker
Response Generator: contextual response selection and synthesis
Conversational Search and Recommendation Agent
General Knowledge Graph
Product Catalogue
Community Posts and Forums

Figure 1: Illustration of a conversational search system that can act as the Agent in Example 1.

Contributions and Research Questions Multi-modal, conversational search opens up many questions. We focus on how the workers' environment — the information and tools provided to each worker — affects their interactions with each other. To that end, we aim to answer four research questions:

R1 How does one worker's environment affect their behavior and language use?

R2 Can one worker's *environment* alter the other's behavior including language and feature use?

R3 How much do spoken and typed utterances differ in a conversational search environment?

R4 What Buyer and Seller environments yield the most realistic conversations?

Our work for the first time explores and analyzes the most effective conditions for priming workers in large-scale, crowd-sourced conversational data collection, with multiple interaction modalities, and introduces quantitative evaluation metrics for dialogue quality, which, as we demonstrate, could be used to train a conversational E-Commerce search and recommendation agent.

2 Related Work

Automated conversational agents have been an active area of study, and their use has exploded following the success of voice-based conversational agents such as Siri, Alexa, and Google Home Assistant. A long-term goal for dialogue systems is to coherently and engagingly converse with humans on a variety of topics (Guo et al., 2018; Venkatesh et al., 2018; Khatri et al., 2018). However, such systems require extensive engineering or extensive training data collection and annotation, or both. Below, we review prior approaches to collecting and annotating data for training conversational agents. We focus mostly on task-oriented agents in complex domains such as search and recommendation, i.e., information-oriented and transactional tasks (Radlinski and Craswell, 2017; Zhang et al., 2018).

2.1 Task Completion Agents

For well-structured tasks like travel reservations (Bobrow et al., 1977) or movie ticket purchases, rule-based dialogue systems can be effective, but require significant engineering to design possible responses and appropriate dialogue flow. Beginning with early systems such as Eliza, rule-based dialogue management (DM) systems (Bobrow et al., 1977) have been steadily improving in sophistication and flexibility (Chen et al., 2017).

Recently, end-to-end learning for automated conversational agents approaches has grown in popularity, due in part to improvements in neural architectures and the availability of general-purpose training data, e.g. (Serban et al., 2018). The idea of conversation has also been introduced as a way to elicit user interests for item recommendation (Christakopoulou et al., 2016). For example, Sun and Zhang (2018) introduced an end-to-end reinforcement learning framework for a personalized conversational sales bot, and Li et al. (2018) use a combination of deep-learning based models for conversational movie recommendation.

2.2 Knowledge-grounded Agents

Corpus-based chatbots mine human-human conversations, often collected via crowd-sourcing or by scraping online resources. (Serban et al., 2018) summarizes available corpora up to 2017, including online human-human chats, Twitter, and in-movie dialog. However, these resources are not helpful to train knowledge-, task- or information-oriented conversational agents to provide or recommend useful information for a specific topic like a purchasing decision. To improve the knowledge retrieval process, several teams have recently introduced frameworks to incorporate external knowledge in response generation as well as actively learn concepts through conversations (Dinan et al., 2018; Luo et al., 2019; Jia et al., 2017; Ghazvininejad et al., 2018). Despite these advances, the underlying knowledge is essentially encoded, e.g. in a neural network. This is not feasible for extensible, large, or frequently updated domains, such as product information, or sources lacking a rich search mechanism. The closest effort to this is a system outlined in (Gur et al., 2018), which learned to query a reservations system from extensive logs of human interactions with the system. Such logs are naturally not available for privacy reasons. Thus, to train such agents, a conversation collection tool must be specifically designed to incorporate dynamically retrieved external knowledge from a search engine, with the associated queries and actions.

2.3 Previous Conversation Collections

Recently, a number of shared tasks and challenges have pushed researchers to develop more intelligent chat bots capable of in-depth conversations on numerous topics, not just *small talk*. Resulting conversations have been evaluated both by crowd workers and live users as part of the Alexa Prize Conversational AI challenge (Venkatesh et al., 2018). Some public datasets have been made available as a by-product of the challenge. (Dinan et al., 2018) introduced a valuable resource for crowdsourcing conversations in a “Wizard of Oz” interface, used to collect restaurant reservation dialogs (CamRest676 dataset) (Wen et al., 2017); the Frame corpus in a more complex travel booking domain (El Asri et al., 2017), and a corpus of in-car navigation conversations (KVRET corpus) (Eric et al., 2017). Later, this approach was extended to conversations on multiple topics (Budzianowski et al., 2018). (Gopalakrishnan et al., 2019) complemented that work with a large corpus of topical conversations between crowd workers asked to discuss an assigned topic, but without specific suggestions or aspects to discuss.

A data set of *coached* movie discussions between a “Wizard” and an “Apprentice” was introduced by (Radlinski et al., 2019). Conversations did not have a specific goal, but unlike previous efforts the “Wizards” were instructed to follow a general script and ask prescribed questions about movie preferences. The resulting data set may be helpful for recommender systems. In other work, searchers asked “intermediaries”–other workers–to find information for them on complex tasks via voice input and output, with only the “intermediary” having access to the search engine (Trippas et al., 2017, 2018). The resulting data set, while valuable, was limited to a small number of participants in the laboratory study. The study’s open-ended nature makes it hard to scale to sufficiently large and robust data collection required to train effective automated search agents.

While there has been substantial recent work in leveraging Large Language Models (LLMs) such as ChatGPT and GPT-4 to generate conversations (Brown et al., 2020; Han et al., 2021; Li et al., 2023), these approaches have not had much success in the E-Commerce domain due to the dynamic range of customer behavior, which is quite different from info-seeking scenarios where LLMs excel. This trend is reflected in Bernard and Balog (2023), where the authors release a collection of 64 high-quality shopping conversations encompassing various goals. The size of this data also reflects the challenges in scaling data collection in this domain, a key factor that we try to address. Most recently, Joko et al. (2024) used LLMs to provide workers with guidance on what to say, ostensibly to simplify collection of these complex conversations, but they collected less than half the conversations we have.

These previous efforts provide large corpora of human-human conversations grounded on specific topics, but are neither sufficient to learn to *search and retrieve*, nor to *incorporate* external, dynamically retrieved information, nor to lead the dialogue towards a task completion, which is the focus of our work. Furthermore, the ability to share rich information items, such as product descriptions or picture, is critical for an effective search-oriented conversational system. Prior work left as open questions *how* to collect such conversation and action data for complex search and recommendation tasks; how interaction modality variations affect the richness, and ultimately overall quality, of the resulting dialogue; and even how to measure dialogue quality in such crowd-sourced efforts.

3 Crowdsourced Conversation Task

Our crowd tasks pair "Buyers" with "Sellers" in a E-Commerce search simulation. Our goal is to simulate the in-store experience of asking a salesperson for assistance. Therefore, we wanted to learn how Seller behavior changes when Buyers come with varied shopping-related knowledge and needs. It has been long-held that voice and text interactions are very different, but this has largely been tested for simple text search, not conversation. To answer our research questions, we tested several product search and display features, and additionally the impact of a voice interface for the Buyer. We collected 1,500 conversations, on which we based on our analysis. We then collected 1503 more conversations under what we found were the best conditions, described in Table 1, which we publicly release.[1]

Below we describe key features of our conversation task. Further details, including the modified ParlAI/Mechanical Turk (Miller et al., 2017) framework, audio transcription, and the catalog used, may be found in the Appendices.

3.1 Layout and Conversation Flow

Each worker's interface (Figure 2) displays products at left, and a chat pane right for interacting with their partner. Both are given instructions before beginning the chat, but can view them again at any time. In particular, both workers are made aware they will be chatting in real time with another person, that the text and other interactions will be stored for future research use. The Seller first picks product categories they are familiar with. To increase variety, we kept track of each worker's choices, which they could not repeat within one week, unless they work through all categories in that time. The Buyer then chooses the category from the Seller's options, and opens the conversation with a request.

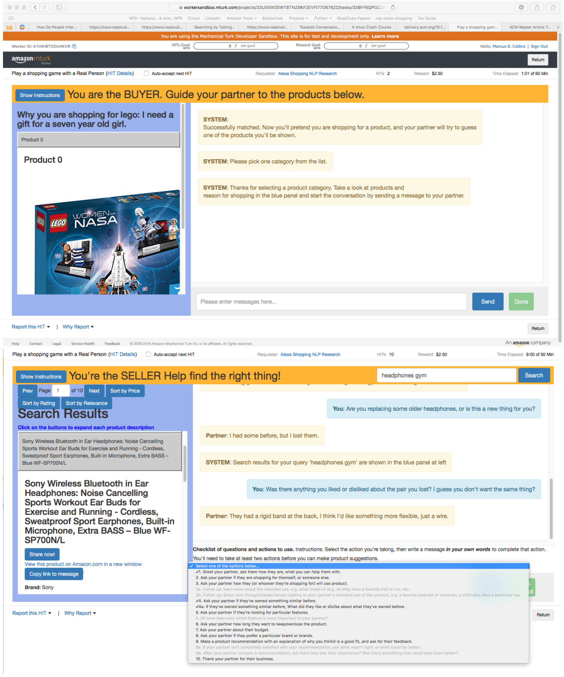
Buyer view In the left pane, the Buyer sees context about why and for what they are shopping (their "persona"), and three target products which match their context. The Buyer's goal is to guide the Seller to one of those products by asking and answering questions. The info shown varies depending on the experimental conditions (Section 3.3), but can include product title, details (price, age range, etc.), description, and images. The personas describe a shopping mission, e.g., "My four year old daughter loves Star Wars", or "I want really good headphones for home listening to classical music, but I don't have an unlimited budget". Each persona is specific to a particular product category.

Seller view Seller have a search box at upper right, and a limited interface for sorting and scanning search results. They have several options for

Topic	Count	Mean Turns
books	135	6.5
headphones	303	6.9
Lego	252	6.3
pet food	209	6.6
running shoes	187	6.8
smartwatches	294	6.6
vitamins	123	6.7

Table 1: Statistics of the data we will release. One turn is an exchange between the two users. These were collected under condition B.IIc, see Section 3.3 for details.

[1] https://github.com/marcuscollins/woa-ecommerce-conversations


workersandbox.mturk.com/projects/33U5EW3SWT8TNZBKF2EVFI77D678ZD/tasks/33BFF6QPI3Z2
You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. Learn more
Worker ID: A1IHH8TODU4KVR
HITs Goal 0 / set goal
Reward Goal 0 / set goal
Hello, Marcus D. Collins | Sign Out
amazon mturk Worker
Return
Show Instructions
Product 0
Send
Done
Report this HIT
Why Report
Help Contact Legal Service Health Feedback
© 2005-2019 Amazon Mechanical Turk Inc, or its affiliates. All rights reserved.
An amazon company
Prev Page 1 of 10 Next Sort by Price
Sort by Rating Sort by Relevance
Share now!
View this product on Amazon.com in a new window
Copy link to message
Search

Play a shopping game with a Real Person (HIT Details) □ Auto-accept next HIT Requester Alexa Shopping NLP Research HITs 2 Reward $2.50 Time Elapsed 1:01 of 60 Min

You are the BUYER. Guide your partner to the products below.

Why you are shopping for lego: I need a gift for a seven year old girl.

SYSTEM: Successfully matched. Now you'll pretend you are shopping for a product, and your partner will try to guess one of the products you'll be shown.

SYSTEM: Please pick one category from the list.

SYSTEM: Thanks for selecting a product category. Take a look at products and reason for shopping in the blue panel and start the conversation by sending a message to your partner.

Please enter messages here...

Play a shopping game with a Real Person (HIT Details) □ Auto-accept next HIT Requester Alexa Shopping NLP Research HITs 10 Reward $2.50 Time Elapsed 9:00 of 50 Min

You're the SELLER Help find the right thing! headphones gym

Search Results

Click on the buttons to expand each product description

Sony Wireless Bluetooth in Ear Headphones: Noise Cancelling Sports Workout Ear Buds for Exercise and Running - Cordless, Sweatproof Sport Earphones, Built-in Microphone, Extra BASS – Blue WF-SP700N/L

Sony Wireless Bluetooth in Ear Headphones: Noise Cancelling Sports Workout Ear Buds for Exercise and Running - Cordless, Sweatproof Sport Earphones, Built-in Microphone, Extra BASS – Blue WF-SP700N/L

Brand: Sony

You: Are you replacing some older headphones, or is this a new thing for you?

Partner: I had some before, but I lost them.

SYSTEM: Search results for your query 'headphones gym' are shown in the blue panel at left

You: Was there anything you liked or disliked about the pair you lost? I guess you don't want the same thing?

Partner: They had a rigid band at the back, I think I'd like something more flexible, just a wire.

Checklist of questions and actions to use. Instructions: Select the action you're taking, then write a message *in your own words* to complete that action. You'll need to take at least two actions before you can make product suggestions.

✓ Select one of the options below...
✓1. Greet your partner, ask them how they are, what you can help them with.
2. Ask your partner if they are shopping for themself, or someone else.
3. Ask your partner how they (or whoever they're shopping for) will use product.
3a. Follow up: learn more about the intended use, e.g. what breed of dog, do they have a favorite trail to run, etc..
3b. Follow up: Share your thoughts/experiences relating to your partner's intended use of the product, e.g. a favorite podcast or musician, a child who likes a particular toy.
✓4. Ask your partner if they've owned something similar before.
✓4a. If they've owned something similar before, What did they like or dislike about what they've owned before.
5. Ask your partner if they're looking for particular features.
5. (If more than one) which feature is most important to your partner?
6. Ask your partner how long they want to keep/own/use the product.
7. Ask your partner about their budget.
8. Ask your partner if they prefer a particular brand or brands.
9. Make a product recommendation with an explanation of why you thinkit is a good fit, and ask for their feedback.
9a. If your partner isn't completely satisfied with your recommendation, ask what wasn't right, or what could be better.
9b. After your partner accepts a recommendation, ask them how was their experience? Was there something that could have been better?
10. Thank your partner for their business.

Figure 2: A screenshot of the worker chat windows: Buyer view (top), in priming condition B (Section 3.3) with no product details available, and Seller view (bottom), in a "coached" condition. Search results, search box, message history, and the Seller checklist are shown.

sharing products, set by the experimental conditions. In all conditions, sellers can describe the product or provide a product URL to the Buyer. In some conditions, Sellers are allowed to copy/paste the URL directly from the product description. Another option is to click "Share Now" on the product and enter a message; product details and message are then immediately displayed to the Buyer. We experimented with a "recommendation list" that allowed the Seller to display several products at once to the Buyer for comparison. In one setting Sellers were "coached" with a variety of actions before, visible in Figure 2, and described below.

Ending the conversation The conversation continues at least until Buyer and Seller meet a minimum number of turns (usually 5). The Seller can then make a formal guess (by pressing the "Guess" button), which the Buyer can confirm or reject. The Buyer may also end the conversation at any point after the minimum number of turns. Or, the two can continue the conversation as long as they wish.

3.2 Product Search and Catalog

We selected seven categories: Lego, smartwatches, books, vitamins, running shoes, headphones, and pet food, each with 3-4 personas and 2,207 total products. We chose these categories to cover diverse but still common interests (gifts, technology, recurring purchases), but we did not do any statistical analysis showing them to be the most common amongst real shoppers. The exact products were chosen by searching Amazon.com using queries based on each of the personas, taking the top 100 results, and removing duplicates. More details are in Appendix A.3.

3.3 Experimental Settings

We tested four variables: Buyer *priming*, Seller sharing tools, coaching Sellers, and Buyer voice versus text message entry, each detailed below. To save cost and maximize benefit, we only tested voice transcription and seller coaching with some base conditions. These settings represent different ideas of what information a customer would have when shopping, and the tools a salesperson could use to make effective suggestions. For instance, salespeople are usually trained, and even scripts of how to interact with customers, reflected in our "coaching" condition. Buyers come to the store with different prior knowledge. Indeed, as we'll show, the aspects we condition them to focus on heavily influence their behavior.

Buyer Priming Priming describes what the Buyer is shown about their persona and target products. We used three settings:

A. one product image per product, with no other details or persona.
B. one product image per product and the persona.
C. full product details and persona.
D. a single image, with a pop-up displaying additional images, and the persona.

Seller sharing settings Sellers are assigned one of three levels for how they can share products. At each level Seller has all tools in the lower levels. So, Sellers in condition *II* can still share URLs in message text in addition to sharing a product in a modal dialog.

I. Only URLs can be shared, and only by including them in the message text.
II. All attributes of a single product can be shared in a modal dialog shown to the Buyer, with an optional message from the Seller.
III. The Seller can create a list of multiple products to share in a similar modal dialog.

Seller coaching "Coached" Sellers first choose an action from a list (Table 3, Figure 2) before sending a message to the Buyer. This yields dialog act labels and suggests important actions Sellers should take. To develop the list we consulted a separate set of Mechanical Turk workers about retail sales. We used coaching only with priming *B* and sharing *II* and refer to it as condition B.IIc.

Voice transcription We finally tested voice transcription for condition B.IIc. only. For privacy reasons no audio was stored, only the text transcript. Workers are asked to test their audio transcription first, shown how to correct the transcription if it has errors, and reminded that only text, no audio, will be stored. See Appendix A.2 for further details.

4 Evaluation

Noting that evaluating conversations remains difficult and subjective, we use a variety of measures to understand workers' behavior in our crowd-sourced conversations. We quantify conversations with automated linguistic measures, language modeling, worker surveys, and manual annotation.

4.1 Automated Measures

We computed a number of common language features derived from tri-gram language models and parsing. We use SpaCy[2] for dependency parsing and POS tagging. From the parse tree, we compute utterance-level mean and maximum token depths. We computed perplexity using the NLTK `lm` module.[3] We build the language model from all available conversations, and compute mean perplexity per word at the conversation level, sometimes limited to just Buyer or Seller utterances.

4.2 Language Modeling

To show that Buyer priming influences *Seller* behavior we developed a Poisson regression of word usage in different priming conditions; model details are in Appendix B.

4.3 Manual Annotation

Separate crowd workers evaluated several aspects of conversation quality at both chat and message level. They labeled many chats at once to improve task understanding and consistency of results. Herein, we focus on annotations of overall chat quality. Often these are highly subjective, so we focused on questions that were either quantitative, e.g. "Did the Buyer ask questions about the products?" or had unconventional framing, e.g. "Would you hire this Seller, if you owned the store?". These proved to be the questions that most clearly distinguish between different experimental conditions. We built a Plackett-Luce ranking model (Plackett, 1975; Luce, 1959) to learn a quantitative score for each condition. More details are in Appendix B.

In comparing coached/un-coached Sellers, annotators chose the better Seller from two randomly-paired conversations in the same product category. We modeled the ordered pair data with a Logistic Regression accounting for the presentation order.

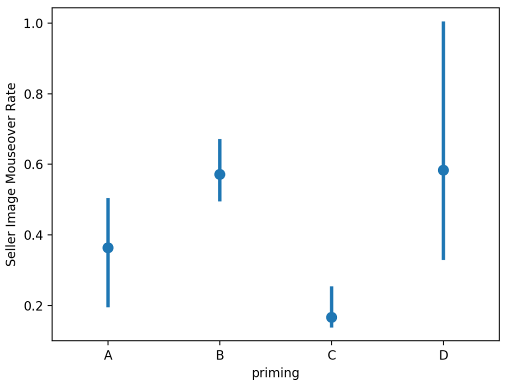


Figure 3: Median and 95% confidence bounds for the product image mouse-over rate per conversation turn, for each of the four priming conditions.

5 Results

Below, we answer our research questions. We show in particular that Buyers' priming (i.e., their *environment*) significantly alters their behavior and language use (R1), that Buyers' environment significantly alters *Sellers'* behavior (R2). Finally, we find that while there are some differences between spoken and typed conversational messages, these differences are not as large as would be expected from studies of keyword-style spoken and typed search queries (R3). We address what are the "best" conditions (R4) throughout this section.

5.1 Buyer priming influences Seller behavior

When Buyers were given the most details (priming *C*), *Sellers* viewed fewer products and scrolled less over product images (Figure 3). We guessed that Buyers focused on the first attributes they saw, and mention only these to Sellers, who naturally

[2] `www.spacy.io`

[3] `www.nltk.org/api/nltk.lm.html`

Budget words		Brand and rating words	
word	*coefficient*	*word*	*coefficient*
around $	-3.0	stars	-2.7
dollars	-2.4	adidas	-2.3
pay	-1.8	apple watch	-2.5
pay $	-2.8	audio-technica	-3.5
something $	-3.5	pegasus	-3.4
us	-3.3	plantronics	-3.8
usd	-4.4	saucony	-2.6

Table 2: Interaction coefficients between priming with details and specific words. The model is log-linked (Eqn. 1), so a coefficient of -3.0 indicates the word is used e^3 20 times less often when showing fewer details.

focus on what the Buyer says. To understand this, we investigated Buyers' word choice in different priming conditions.

The language model shows that Buyers *not* shown product details used budget- and brand-related words much less than Buyers who are shown product details. Table 2 lists words significantly ($p < 0.01$) influenced by product detail priming.

Buyers' word choice is clearly influenced by priming. To show that Buyer word choice led to the *Sellers'* behavior, we tested whether Buyer priming and Seller image mouse-overs are *conditionally independent* of each other, given Buyers' use of words identified by our language model. χ^2-tests of hover-rate and priming reveal that Buyers' word use is what influences Sellers. Specifically, we constructed the tables $P(h, p)$ and $P(h, p|\{w\}_{bb})$ with (binned) hover rate h, priming p, and words $\{w\}_{bb}$ from Table 2. The table conditioned on $\{w\}_{bb}$ yields a χ^2 p-value of 0.34, while without conditioning p is essentially zero, indicating that Buyer use of brand and budget words drives Sellers to ignore other details and images.

We conclude that providing certain product details results in more formulaic, less diverse language from Buyers, i.e., environment clearly influences their language use (R1). This leads Sellers to focus on fewer product aspects and examine fewer products to find a good fit for the Buyer, so one worker's environment clearly alters the other's behavior (R2). To generate the most natural and interesting conversations (R4), the Buyer should not see details like brand and price.

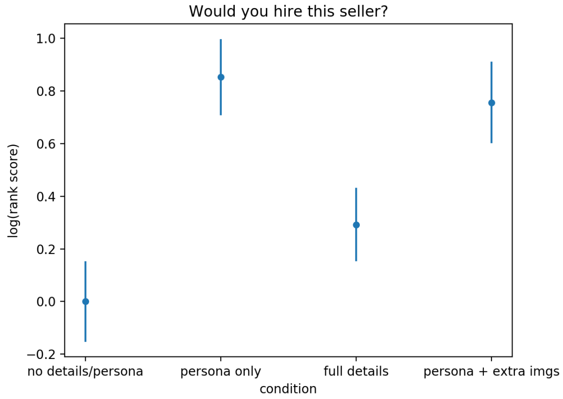


Figure 4: Plackett-Luce scores for the question "Would you hire this Seller, if you owned the store?" with 95 % confidence limits for the four priming conditions.

5.2 Sellers are rated higher if Buyers see fewer details

Annotators ranked priming condition B (personas and images but no product details) highest on the question "Would you hire this Seller, if you owned the store?" (Figure 4). Buyer priming affects not just click and hover actions, but overall Seller quality as well (R2). Our findings demonstrate that the best conversations come when we show Buyers only minimal product details.

5.3 Multi-modal sharing may impact the conversation

So far, we have focused on how the Buyer's priming affects both Buyer's and Seller's behavior. Do the Seller's options they have for sharing results have similar impact? We made a Plackett-Luce model of sharing conditions' impact on annotators responses to "Would you hire this Seller...?" We determined scores for conditions I, II, and III to be 0 ± 0.31, -0.19 ± 0.32, and -0.61 ± 0.37 respectively; that is, annotators felt Sellers did a better job when using only text to share products. However, we find that Buyers rated conversations as more natural in condition II, where Sellers were able to share a single product at a time with complete details. The ratings, on a scale of 1-4, were 2.61 ± 0.030, 2.67 ± 0.027, and 2.48 ± 0.066 for conditions I, II, and III respectively. For the best conversations (R4) we should limit Sellers to sharing simple, single results with perhaps one image.

5.4 Coached Sellers are Preferred

Specifically focusing on how to generate the highest quality conversations (R4), we hypothesize that some kind of coaching should improve Seller quality and indeed this proves true.

Coached Sellers are rated higher than uncoached Sellers We randomly sampled 30 conversations each from B.II with and without coaching and generated pairs of conversations in the same category. We then asked annotators to choose which Seller they preferred of two conversations from the same category. Annotators preferred coached Sellers in 60% of cases, $p \approx 0.013$.

Coaching Sellers results in better dialog from both Buyers and Sellers We analyzed both Buyer and Seller linguistic features to see what might make for more convincing Sellers. An example dialog from this experiment is shown in Appendix C. Coached Sellers used more long utterances (Mann-Whitney test $p \approx 0$), which we suspect indicate to Buyers that Seller is engaged. Coached Sellers' utterances have 12.6% higher perplexity in a 2-gram language model built over all conversations ($p \approx 7.7 \times 10^{-5}$). And, coached Sellers use slightly more complex language, measured by dependency tree depth (1.82 ± 0.049 average token depth versus 1.71 ± 0.035, $p = 0.012$.) We were surprised to find that the conversation partners weakly prefer simpler language. For instance both Buyers and Sellers rate their partner's message clarity slightly lower as the 2-gram perplexity per word increases. (Spearman's $\rho = -0.24$, $p = 0.00003$. Moreover, there is no correlation between annotators' "hire this Seller" rating and any of these linguistic features.

We examined detailed aspects of the dependency parsing and find that coached Sellers use fewer compound words and clausal compounds (e.g., "Let's see what **we can find**") but more compound descriptions, indicating language that is more descriptive but less complex. We observe that Buyers language use also appears to be different when Sellers are coached or not. At this point, we have no firm conclusions what, if any, linguistic features influence annotators ratings of conversation quality.

While we still don't have clear evidence explaining why coached Sellers do a better job, we do conclude that Seller coaching improves conversations overall, helping to answer R4.

5.5 Spoken and Typed Queries are Different

Experiments by Guy (2016), based on web searches in the Yahoo mobile app which had an option to speak the query, are frequently cited. In that study, voice queries were longer and there are noticeable differences in the queries used. In particular, the most distinctive tri-grams in voice reflect fully formed natural language questions, while text queries more strongly resemble keyword queries. Voice queries are much more likely to start with "wh-" question words.

Do those findings hold true in conversational systems (R3)? We tested a variant of priming condition B with voice transcription for Buyers only; Sellers still typed queries. Our findings are quite different than previous work. Perplexity is not significantly different between voice and typed utterances. Buyer voice utterances are on average a word shorter than typed utterances (10.0 vs. 9.0, $p \approx 0$). Surprisingly, Seller utterances are also shorter (10.3 vs. 8.7 words, $p \approx 0$), even though Sellers only type their responses; again Sellers seem to adapt to Buyer language. Finally, unlike Guy (2016) we find at most small differences in parts-of-speech usage between Buyers with and without voice transcription. The largest difference is that 11.7% of Buyer tokens are pronouns with voice, versus 11.0% without voice, just a 0.7% absolute difference. We see a much lower fraction of nouns than in search queries. Search queries were largely dominated by nouns, while we see roughly equal fractions of nouns, pronouns, and verbs.

Taken together, we find that while there are significant differences in voice and typed utterances in a task-oriented conversation, they are not as marked as for individual web search queries. Conversation in general likely leads to more complete and structured sentences, and the use of back-references like anaphora, while people typing web queries will focus on keywords and not use anaphora.

5.6 Results Summary

We have shown in a number of ways that worker environment (i.e., Buyer priming, multi-modal sharing, and coaching sellers) impacts worker behavior (R1). For instance, Buyers view fewer images when presented with other product details, and their language is altered by what they view; when we similarly prime Sellers by coaching them on what questions to ask, they use longer sentences and more diverse language. And we showed that at least

Buyer priming does impact Seller behavior (R2), although it still isn't clear whether the opposite is true, that Seller coaching or different forms of result sharing impact Buyer behavior or language. We found minimal differences in language use for Buyers who spoke or typed their messages (R3) in contrast to general search queries, which are quite different when spoken or typed. Most importantly, we found several ways in which to improve overall conversational quality (R4): limiting Buyer priming to personas and images, coaching Sellers, and limiting multi-modal sharing to single results, or simply sharing links or product titles.

6 Conclusions

We have highlighted several important results that should advance future efforts to crowd-source conversations for effective conversational multi-modal search. First, spoken and typed messages are not as different as previously thought. We attribute this to the conversational nature of the task. This suggests that transfer learning approaches that take advantage of the more plentiful text-based conversations are a promising avenue for voice systems as well.

Importantly, conversational partner's behavior–both "private" behaviors like mouse-overs as well as the language used to communicate–affects the other partner's behavior as well, and we found we can influence both behaviors through priming. Our results show that task design must both direct the desired behaviors as much as possible (e.g., Seller coaching) but must avoid providing too much information. We should look for opportunities to influence Seller behavior in the Buyer's environment as well. We emphasize that the more structure the Seller has, the better the resulting conversations.

Our findings have implications for voice assistants as well: workers will do what we've taught them to do, and ask questions only about the information we present to them. Therefore, to enable good conversational systems for search and exploration, strategies to prime customers with a knowledge of the actions they can take, and the information they can obtain are critical. For instance, if a system presents "price" as a key attribute to customers, our results show that customers are more likely to focus on price in their product exploration.

References

Nolwenn Bernard and Krisztian Balog. 2023. Mgshopdial: A multi-goal conversational dataset for e-commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan*, SIGIR '23.

Daniel G Bobrow, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd. 1977. Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824, New York, NY, USA. ACM, Association for Computing Machinery.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Rashmi Gangadharaiah, Balakrishnan Narayanaswamy, and Charles Elkan. 2018. What we need to learn if we want to do and not just talk. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 3 (Industry Papers), pages 25–32, New Orleans - Louisiana. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. Topic-based evaluation for conversational bots. In *NIPS*.

Izzeddin Gur, Ulrich Rueckert, Aleksandra Faust, and Dilek Hakkani-Tur. 2018. Learning to navigate the web. *arXiv preprint arXiv:1812.09195*.

Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 35–44, New York, NY, USA. Association for Computing Machinery.

Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10*, pages 206–218. Springer.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA. ACM.

Robin Jia, Larry Heck, Dilek Hakkani-Tür, and Georgi Nikolov. 2017. Learning concepts through conversations in spoken dialogue systems. In *Proc. of ICASSP*, pages 5725–5729. IEEE.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. *arXiv*.

K. Sparck Jones, S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779 – 808.

Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Angeliki Metanillou, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Contextual topic modeling for dialog systems. In *IEEE 2018 Spoken Language Technology (SLT)*. IEEE.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems*, pages 9748–9758.

Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. AutoConv: Automatically generating information-seeking conversations with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada. Association for Computational Linguistics.

R. Duncan Luce. 1959. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and Sons, New York, NY, USA.

Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6794–6801.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and Interactive Multimodal Conversations. *arXiv*.

Robert L. Plackett. 1975. The Analysis of Permutations. *Appl. Statist*, 24(2):193–202.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 117–126. ACM.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, pages 92 – 96.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems. *Dialogue & Discourse*, 9(1):1–49.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 235–244. ACM.

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 32–41. ACM.

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 325–328. ACM.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On evaluating and comparing conversational agents. In *NIPS*.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards Ecologically Valid Research on Language User Interfaces. *arXiv*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

A Implementation details

A.1 ParlAI and Mechanical Turk frameworks

We modified ParlAI's (Miller et al., 2017) Mechanical Turk platform for our experiments. We incorporate features for logging worker click and scroll behavior, product search and sorting, and multi-modal product search result sharing. We also implemented audio transcription capabilities with Amazon AWS Transcribe. Finally, the entire system is configurable to easily deploy experiments with different combinations of conditions. We plan to release our code in the near future.

We required participants to use modern web browsers on non-mobile devices and originate in predominantly English-speaking countries: US, UK, Canada, Australia, and New Zealand. We required participants to have completed 1000 or more accepted HITs with $> 98\%$ acceptance rate on Mechanical Turk.

A.2 Voice transcription

Buyers were asked to test their voice transcription beforehand to ensure it worked correctly, as it did not work with all browsers. To begin, workers clicked a green "Start Transcription" button, which then flashes red and reads "Stop Transcription". A red flashing bar indicates that transcription is in progress. Transcription lags a few seconds, but is generally real time. Buyers can edit the transcription to correct any errors, but we found this was rare. We stored both raw and edited transcripts for later analysis.

For privacy reasons, no audio is kept. Sellers never used voice transcription.

A.3 Catalog

Each category has 3-4 personas (23 total), and each of those is assigned three target products. The catalog is completed with roughly 100 related products for each persona, some of which overlap (2,207 total). The product search feature is a very simple keyword search over product title and description; this is sufficient to locate products in our small catalog. Search is implemented using the `whoosh` Python package.[4] Sellers can sort search results by price and rating as well, to help them adapt to specific Buyer personas focused on value or quality.

A.4 Seller dialog acts

Table 3 lists all "coaching" actions available to Sellers in condition B.IIc. As shown in Figure 2 some "follow-up" actions are unavailable until others are used first.

B Evaluation details

B.1 Poisson word-usage model

We used `statsmodels` (Seabold and Perktold, 2010) to implement several different models describing word usage. Based on deviance, we found

[4] https://whoosh.readthedocs.io

#	text
1	Greet your partner, ask them how they are, what you can help them with.
2	Ask your partner if they are shopping for themselves, or someone else.
3	Ask your partner how they (or whoever they're shopping for) will use product.
3a	Learn more about the intended use, e.g. what breed of dog, do they have a favorite trail to run, etc...
3b	Share your thoughts or experiences relating to your partner's intended use of the product, e.g. a favorite podcast or musician, a child who likes a particular toy.
4	Ask your partner if they've owned something similar before.
4a	If they've owned something similar before, what did they like or dislike about what they've owned before.
5	Ask your partner if they're looking for particular features.
5a	(If more than one) which feature is most important to your partner?
6	Ask your partner how long they want to keep/own/use the product.
7	Ask your partner about their budget.
8	Ask your partner if they prefer a particular brand or brands.
9	Make a product recommendation with an explanation of why you think it is a good fit, and ask for their feedback.
9a	If your partner isn't completely satisfied with your recommendation, ask what wasn't right, or what could be better.
9b	After your partner accepts a recommendation, ask them how was their experience? Was there something that could have been better?
10	Thank your partner for their business.

Table 3: **Seller message actions.** Questions with a letter following the index number can only be asked as follow-ups to the corresponding unlettered question.

the best model to be

$$\log y = \boldsymbol{\theta_w} \cdot \boldsymbol{w} + \boldsymbol{\theta_t} \cdot \boldsymbol{t} + \boldsymbol{\theta_{wt}} \cdot (\boldsymbol{w} \otimes \boldsymbol{t}) + \boldsymbol{\theta_{pt}} \cdot (\boldsymbol{p} \otimes \boldsymbol{t}) + \boldsymbol{\theta_{wp}} \cdot (\boldsymbol{w} \otimes \boldsymbol{p}), \quad (1)$$

where $\vec{w}$ are n-gram word features ($n \geq 3$), $\vec{p}$ the priming conditions, and $\vec{t}$ the chosen topic. $\otimes$ indicates the Cartesian product, and all $\vec{\theta}$ are (one-d) parameter vectors. Note that this model explicitly captures the interactions between the chosen topic and priming conditions. Statistical feature selection was performed to reduce the model dimension. We apply χ^2 tests to contingency tables of Buyer word occurrence and Seller behavior features to determine significant words.

B.2 Manual annotation details

As with the main crowd task, we took standard measures to ensure quality results, such as using "gold" test questions and excluding annotators with below 80% accuracy, and requiring a minimum time working on each task. Nonetheless, we found workers predictably over-rate the quality of conversations on any given aspect, resulting in skewed distributions and low variance. For example, annotators were probably overly accepting of conversation quality, as measured by the question "Would you hire this seller?" (Figure 5

Even filtering out low quality annotators (say, with the approach of (Ipeirotis et al., 2010)) simple statistical comparisons failed to reveal significant results, although they did provide some hints at differences between the conditions. To overcome the homogeneity of annotators' responses, we aggregated each worker's ratings to make a partial ranking over the experimental conditions. Workers who rate all chats the same are dropped, implicitly removing low-quality annotations. This partial ranking was then modeled using the Plackett-Luce framework mentioned in the main text.

B.3 Seller preference model

In evaluating Sellers in paired conversations, annotators showed a significant bias towards the second conversation of two presented; this *Context Effect* is probably most familiar in multiple-choice surveys, where it is addressed by randomly ordering the choices, as we have done here[5].

[5] https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n439.xml

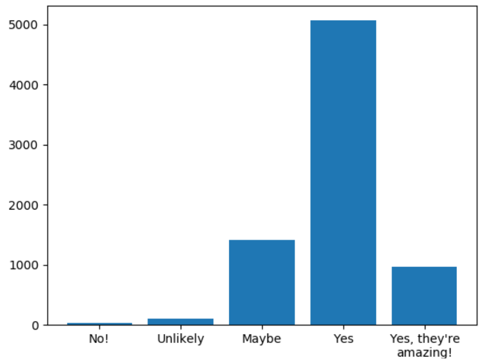


Figure 5: Distribution of responses to the question "Would you hire this Seller, if you owned the store?"

C Example dialog

Note. It has been suggested to us that the Buyer seems to "forget" the age of their child, and suggested this is an attempt to extend the conversation and meet the minimum number of turns. We suspect instead they failed at first to fully read their persona, which included the child being three years old, not six. In any event, we don't find this mistake particularly unnatural.

D Case Study: Retrieval-based Product Search and Recommendation Agent

We evaluated our data by creating a simple automated conversational search agent from it, as illustrated in Figure 1. Crowd workers tested the automated agent in a human-in-the-loop setting.

D.1 Conversational Agent Prototype

We collected 1,500 additional conversations in condition B.IIc. Our agent retrieves the most relevant archived response for the current context, using `whoosh` with BM25 ranking (Jones et al., 2000), for simplicity. It performs product search as needed. Queries are formed from Buyer utterances, with smaller weight for older utterances.

Heuristics improve the candidate responses. A simple bag-of-words logistic regression classifier identifies the product category; once known, our agent limits results to responses for just that category. We limit some dialog acts, like budget or brand questions, to be used at most once per conversation. (Recall Sellers labeled dialog acts as part of their task). Finally, we prevent similar responses from being given more than once.

After a minimum three turns, Product search is triggered if a candidate response is a product recommendation or mentions a product, or if no other

Conversation between a LEGO Buyer and a coached Seller

Buyer hi there, im looking for some legos for my kids. are these good for 6 year olds?
Seller They sure are! They are great for hand/eye coordination and problem solving skills! Do you know what kind of set you may be looking for? Do they have any specific likes or interests that may translate to a lego set we have for you?
Buyer awesome, they love trucks! anything out there with trucks?
Seller (*searches "lego truck set", "lego truck set 6"* They do! I'll send a suggestion in just a moment. First, do you have a particular budget?
Buyer ok perfect, no particular budget, just don't want to break the bank. my child is actually 3, i got them mixed up with my niece somehow. didnt have enough coffee!
Seller (*searches "lego truck set 3"*) How about this? This has bigger LEGO pieces so it is less complicated. (*shares details for LEGO Duplo Big Construction Site 10813*)
Buyer That is perfect! thanks so much.
Seller Yay! I am so glad. Your child will love this set! Before you go, was everything to your satisfaction? Is there anything more I can do for you?
Buyer Absolutely everything went great. I'm sure they'll love it too. Thank you again!

Example 2: Conversation in condition B.II with coaching. Note that we use Seller and Buyer here, as this is a real conversation between two crowd workers in our task, *not* between an automated Agent and a real Customer.

suitable response is available. When the agent cannot retrieve a response, it constructs one using the following strategies, in order: find another response in the archive which mentions the desired product and still scores well; construct a response from product description highlights based on the current context; or use a default generic response. Despite the simplicity of the agent, we will show that due to the high quality of our conversational corpus, the agent often performs on par with experienced crowd-workers.

D.2 Results of Human-in-the-Loop Agent Test

The agent's best 3-5 candidate responses were presented to human crowd workers acting as Sellers, who could select one of the candidates, or create their own response. We can then evaluate conversation quality in two ways. We asked annotators to rate the conversations individually, as above. We also quantify how often Sellers used the agent-recommended responses in each conversation, i.e., whether the agents' response was accepted by the human crowd worker.

Figure 6 summarizes our findings. We grouped conversations by the fraction f of Seller responses generated by the computer agent. So, if the Seller used the agent response without editing in three of five turns, then $f = 0.6$. A significant fraction of Sellers appeared to either be unaware of how to use the agent recommendations or didn't want to use them. We separate these conversations into the '0' group. Fig. 6 shows that there is no statistically significant difference between the different bins in f. Furthermore, there is no statistical difference from conditions B.II or B.IIc, whose quartiles are shown in red and blue, respectively.

The distribution of f is shown at the bottom of Figure 6. About 10 % of conversations do not use the agent recommendations at all. Overall, slightly more than 40 % of responses used in all conversations were from the agent. Given our simple retrieval-based agent, the results are promising and demonstrate the value of our corpus. In future work we aim to improve our agent by exploring more sophisticated dialogue management, response ranking models, and generalization.

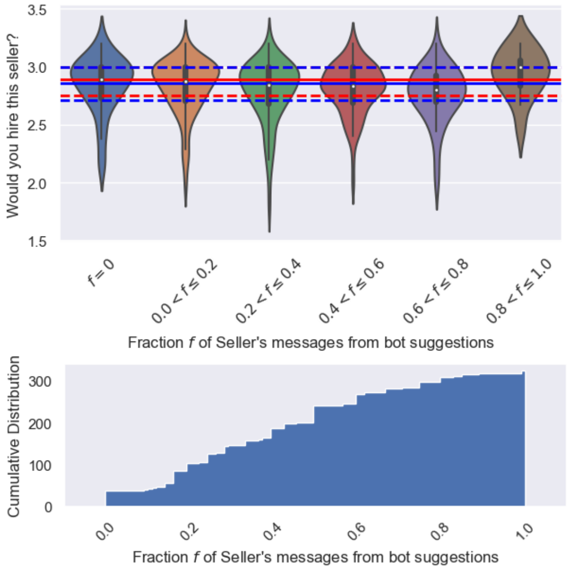


Figure 6: (Top) Violin plot of annotator ratings on the question "Would you hire this Seller?", grouped by the fraction of responses in each conversation generated by the retrieval agent. (Bottom) Cumulative Distribution of retrieval-agent response use fraction f.