InSCIt: Information-Seeking Conversations with Mixed-Initiative Interactions

Zeqiu Wu♠ Ryu Parish♠ Hao Cheng♣ Sewon Min♠
Prithviraj Ammanabrolu♢ Mari Ostendorf♠ Hannaneh Hajishirzi♠♢
♠University of Washington, USA ♣Microsoft Research, USA ♢Allen Institute for AI, USA
{zeqiuwu1,rparish,sewon,ostendor,hannaneh}@uw.edu
chehao@microsoft.com raja@allenai.org

Abstract

In an information-seeking conversation, a user may ask questions that are under-specified or unanswerable. An ideal agent would interact by initiating different response types according to the available knowledge sources. However, most current studies either fail to or artificially incorporate such agent-side initiative. This work presents InSCIt, a dataset for **In**formation-**S**eeking **C**onversations with mixed-initiative **Int**eractions. It contains 4.7K user-agent turns from 805 human-human conversations where the agent searches over Wikipedia and either directly answers, asks for clarification, or provides relevant information to address user queries. The data supports two subtasks, evidence passage identification and response generation, as well as a human evaluation protocol to assess model performance. We report results of two systems based on state-of-the-art models of conversational knowledge identification and open-domain question answering. Both systems significantly underperform humans, suggesting ample room for improvement in future studies.[1]

1 Introduction

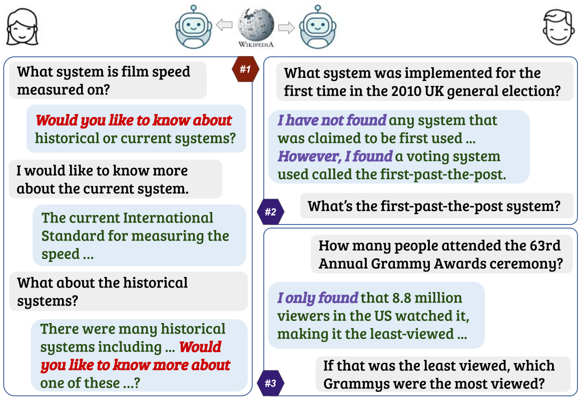
Recently, there has been increasing interest in developing conversational information-seeking systems (Choi et al., 2018; Adlakha et al., 2022; Saeidi et al., 2018; Feng et al., 2020) that assist users in finding information from knowledge sources (e.g., text corpus) via multi-turn conversational interactions. One important advantage of such conversational information-seeking systems is that users do not need to come up with a very descriptive query by themselves (Webb and Webber, 2009; Rieser and Lemon, 2009; Konstantinova and Orasan, 2013). In realistic settings, as shown in Figure 1, users can start with a request that is under-specified or has no direct answer, and through conversational interactions, the agent can collaboratively guide users to refine (left) or relax their queries and proactively suggest relevant information that may partially satisfy the user's information needs (right). This collaboration requires a mixed-initiative dialogue, where both the user and agent can direct the flow of the conversation.

Handling such realistic user requests poses challenges to a conversational agent system. A comprehensive search can result in *multiple passages* from the knowledge source, which may provide different components of an answer (possibly incomplete) or multiple answers that surface ambiguities in the user query. Depending on the available information, the agent needs to use *different strategies*, which might involve summarizing the results, providing partial information, or trying to clarify an ambiguity. However, existing information-seeking conversation datasets rarely contain conversations where agents initiate different interaction strategies. As a result, most conversational question answering (CQA) work focuses on user-initiative interactions, where the agent simply responds to user questions with direct answers or uses `no answer` for out-of-scope queries (Choi et al., 2018; Reddy et al., 2019; Adlakha et al., 2022). Other work studies clarification questions using artificially created data, failing to capture natural information-seeking interactions (Saeidi et al., 2018; Feng et al., 2020; Aliannejadi et al., 2021; Guo et al., 2021). In addition, most of them only rely on a single evidence string or passage for agent response construction.

To support research in mixed-initiative conversations, we introduce InSCIt (pronounced Insight), a dataset for **In**formation-**S**eeking **C**on-

[1] We open-source all data and code at `https://github.com/ellenmellon/INSCIT`.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 453–468, 2023. https://doi.org/10.1162/tacl_a_00559
Action Editor: Dan Goldwasser. Submission batch: 8/2022; Revision batch: 12/2022; Published 5/2023.



What system is film speed measured on?

Would you like to know about historical or current systems?

I would like to know more about the current system.

The current International Standard for measuring the speed ...

What about the historical systems?

There were many historical systems including ... *Would you like to know more about* one of these ...?

What system was implemented for the first time in the 2010 UK general election?

I have not found any system that was claimed to be first used ... *However, I found* a voting system used called the first-past-the-post.

What's the first-past-the-post system?

How many people attended the 63rd Annual Grammy Awards ceremony?

I only found that 8.8 million viewers in the US watched it, making it the least-viewed ...

If that was the least viewed, which Grammys were the most viewed?

Figure 1: InSCIt examples show that user queries can often be under-specified and require clarification (left), or have no direct answer but where providing relevant information may fulfill users' information needs (right).

Dataset	IR	Response Strategy		H-H	Multi-
		CLAR	REL	Dialogue	Evidence
InSCIt (ours)	✓	✓	✓	✓	✓
QuAC	✗	✗	✗	✓	✗
CoQA	✗	✗	✗	✓	✗
DoQA	✓	✗	✗	✓	✗
QReCC	✓	✗	✗	◐	✗
TopioCQA	✓	✗	✗	✓	✗
Qulac	✗	✓	✗	✗	✗
ShARC	✗	✓	✗	✗	✗
MultiDoc2Dial	✓	✓	✗	✗	✗
Abg-CoQA	✗	✓	✗	✗	✓

Table 1: Comparison of InSCIt with existing datasets of information-seeking conversations. *IR*, *CLAR*, *REL*, *H-H* stand for *Retrieval Needed*, *Clarification*, *No Direct but Relevant Answer*, and *Human-Human*. ◐ indicates the property only applies to part of the dataset.

versations with mixed-initiative **In**teractions, where agents take various strategies, such as providing direct answers (72%), raising clarifications (13%), and presenting relevant partial information (13%), to address users' information needs. It contains 805 natural human-human conversations with 4.7K user-agent turns over diverse topics, collected through a scalable annotation pipeline and careful quality control. To simulate realistic information-seeking scenarios, users write queries with minimal restriction, and human agents decide on different strategies to respond, after searching over the knowledge source (i.e., Wikipedia) for evidence passages.

We formulate two tasks for the conversational agent system: (1) identify a set of evidence passages from Wikipedia, and (2) generate a response grounded in the evidence. Since handling queries with multiple evidence passages or no direct answer can be open-ended, we emphasize the need for human evaluation, and propose a systematic human evaluation protocol that considers diverse aspects including coherence, factual consistency, and information comprehensiveness.

We present two strong baselines based on the state-of-the-art in open-domain question answering (Karpukhin et al., 2020; Izacard and Grave, 2021) and conversational knowledge identification (Wu et al., 2021). While the systems achieve substantial improvements over a trivial baseline, there is still significant room for improvements, especially for scenarios requiring agent strategies other than providing a direct answer. Our analysis suggests that the key remaining challenges are improving passage identification and fusing comprehensive information from *multiple passages* by leveraging *different strategies*. We present detailed discussion and avenues for future work.

2 Related Work

Information-Seeking Conversations The aim of information-seeking conversations is to address the user's initial and follow-up information needs with grounding in knowledge sources. Table 1 compares InSCIt with previous information-seeking conversation datasets. Early CQA work, including QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019), requires the agent to answer each user question by reading a short passage. DoQA (Campos et al., 2020), QReCC (Anantha et al., 2021), and TopioCQA (Adlakha et al., 2022) extend the task to an open-domain setting where the knowledge source is a large document corpus. These studies only consider limited scenarios where the agent provides a direct answer based on a short text span in a single passage, or outputs `no answer` if there is no direct answer.

Ambiguous user queries have been observed in single-turn question answering tasks (Min et al., 2020; Zhang and Choi, 2021; Sun et al., 2022), but these are usually addressed by training a model to predict multiple conditional answers without further interaction. A few other studies create artificial conversations to address ambiguous user questions. For instance, Qulac (Aliannejadi et al., 2019) and the data collected in follow-up work

(Aliannejadi et al., 2021) are based on user queries containing a set of pre-specified multi-faceted entities, where agents choose from a fixed set of clarification questions that cover these ambiguities. ShARC (Saeidi et al., 2018), Doc2Dial (Feng et al., 2020), and MultiDoc2Dial (Feng et al., 2021) are rule-based information-seeking conversations in the social welfare domain that incorporate agent-side clarifications. Guo et al. (2021) create Abg-CoQA by rewriting conversations in the CoQA dataset to intentionally include ambiguous questions. In contrast, InSCIt consists of human-human conversations with natural information-seeking user requests and mixed agent initiative to address them.

Penha et al. (2019) crawl conversations from Stack Exchange[2] that are mixed with information-seeking utterances and casual talk. One grounding document is heuristically obtained for each conversation. In contrast, InSCIt contains validated grounding passages and only goal-oriented agent interactions.

Knowledge-Grounded Social Chat Instead of seeking for information, the user intent in social chat is mostly to conduct casual talk. Knowledge-grounded social chat systems (Ghazvininejad et al., 2018; Dinan et al., 2019; Zhou et al., 2018; Moghe et al., 2018) incorporate external knowledge with the purpose of making the conversations more engaging and informative. Rodriguez et al. (2020) trains a conversational agent to select knowledge to present based on the user's background, in order to maintain the user's interest in the conversation.

3 Task Formulations

We define two task formulations for InSCIt, namely *passage identification* and *response generation*. These two tasks mimic how an agent responds to each information-seeking user request, by first searching for relevant information over the knowledge source and then constructing the response based on the gathered information. Comparing with prior studies on open-domain information-seeking conversations (Anantha et al., 2021; Adlakha et al., 2022), the key challenges in our tasks come from identifying and fusing comprehensive information from *multiple passages* to construct responses using *different strategies*, rather than a single passage and a short answer.

At the n^{th} agent turn, both tasks have the same input: all previous utterances (i.e., dialogue context) $X = [u_1, a_1, u_2, a_2, \ldots, u_n]$, the corpus of all passage candidates $\mathcal{C}$, and the previously used passages $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{n-1}\}$ where each $\mathcal{P}_i = \{p_i^1, p_i^2, \ldots, p_i^{|P_i|}\}$ is the set of passages used in the i^{th} agent turn a_i. $\mathcal{C}$ is defined as all textual paragraphs (i.e., passages) in a full Wikipedia dump.[3]

For *passage identification*, we require the model to predict a *set* of passages $\bar{\mathcal{P}}_n$ from $\mathcal{C}$, containing comprehensive and relevant information to the current user request u_n in the dialogue context X, which serves as evidence for the *response generation* task—generating the next agent response $\bar{a}_n$. This is different from the passage retrieval task where only a ranked list of relevant passages is predicted. Identifying specific knowledge to be used in the response can be important for model interpretability purposes as well as for evaluating how well a model grounds the response generation in the knowledge source. Ideally, all factual information contained in $\bar{a}_n$ should be consistent with $\bar{\mathcal{P}}_n$, and every passage in $\bar{\mathcal{P}}_n$ should provide at least one unique information piece as evidence for $\bar{a}_n$.

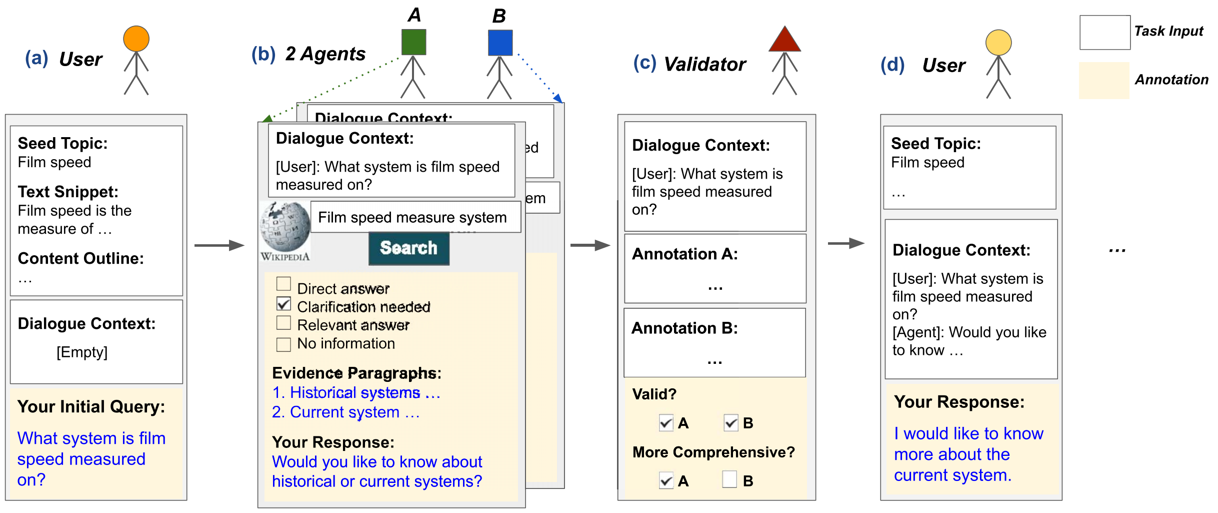
In interactive dialogues, each predicted evidence $\bar{\mathcal{P}}_i$ and response $\bar{a}_i$ are used in the dialogue context for later conversations. However, to use pre-collected dialogues with automatic evaluation metrics, the input context must be the same as that leading to the human reference response. This is also consistent with setups in previous information-seeking dialogue studies that are discussed in § 2. Therefore, the gold $\{\mathcal{P}_i\}$ and $\{a_i\}$ are used here as inputs in testing.

4 Our Data: InSCIt

We introduce InSCIt, a new information-seeking conversation dataset where the agent interprets the user intent and provides comprehensive information grounded in Wikipedia via natural human-human interactions. In this section, we present our data collection pipeline, quality control mechanisms, and analyses that show the characteristics and diversity of the user and agent turns.

[2] `https://stackexchange.com/`.

[3] We use the dump of 04/20/2022.



Seed Topic:
Film speed

Text Snippet:
Film speed is the measure of …

Content Outline:
…

Dialogue Context:
[Empty]

Your Initial Query:
What system is film speed measured on?

Dialogue Context:
[User]: What system is film speed measured on?

Film speed measure system

□ Direct answer
☑ Clarification needed
□ Relevant answer
□ No information

Evidence Paragraphs:
1. Historical systems …
2. Current system …

Your Response:
Would you like to know about historical or current systems?

Dialogue Context:
[User]: What system is film speed measured on?

Annotation A:
…

Annotation B:
…

Valid?
☑ A ☑ B

More Comprehensive?
☑ A □ B

Seed Topic:
Film speed
…

Dialogue Context:
[User]: What system is film speed measured on?
[Agent]: Would you like to know …

Your Response:
I would like to know more about the current system.

Figure 2: Each conversation is annotated in a series of user → agent → validator tasks. One worker is dedicated to each user/validator task but two workers work in parallel on the agent turn annotation (see discussion in § 4.1).

4.1 Data Collection Pipeline

We recruit user, agent, and validation workers[4] to create and annotate user/agent turns and validate agent annotations, respectively. Due to the asymmetric time spent by user and the agent workers in a conversation, we design a separate annotation task for each user or agent turn, following Wen et al. (2017) to annotate each dialogue in a pipelined fashion. This framework has proved to be efficient while maintaining the conversation coherence by requiring each worker to read all previous utterances. Our data collection has IRB approval and is deemed exempt.

Figure 2 illustrates the data collection and annotation pipeline. Each conversation starts with an initial user turn, where the worker asks a question after reading a text snippet from a seed document. Then, two agents independently search for relevant passages in Wikipedia, provide a response, and categorize their response. Validation follows after each user-agent turn. We refer to the retrieved passages, contributed text, and validations collectively as ‘‘annotations.’’ The user/agent/validation process is repeated for 7 turns or until responses are found to be invalid. Details for each step follow.

Seed Document Selection To diversify conversation topics, we sample seed Wikipedia articles, used for triggering initial user requests, from 5 different topic categories—food and drink, hobby, historical events, geography, and weekly top-25 pages. Additionally, we leverage the top-down tree structure of Wikipedia categories[5] and sample articles at various tree depths under each of the first 4 categories. Weekly top-25 pages are from Wikipedia weekly reports of 2021.[6] Figure 3 (left) shows the distribution of sampled seed documents under each category and their corresponding depths.

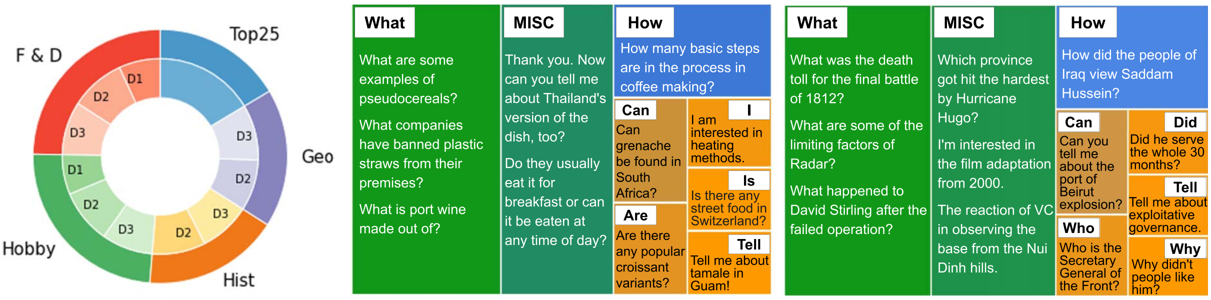
User Turn Here, a user worker is asked to write an *initial query* or *follow-up response* to continue the existing conversation. To trigger each conversation (Figure 2(a)), the user worker is presented with the leading paragraph of a seed article, and is instructed to ask a question they are interested in but cannot find the answer from the paragraph. The article content outline containing all section titles is also provided to help with the question construction. The annotation for each following user turn (d) starts after the completion of the previous agent annotation (b) and the validation step (c), based on all previous conversation utterances.

Agent Turn Different from the user worker, in addition to the dialogue context, each agent worker (Figure 2(b)) is given all evidence paragraphs used by each previous agent turn as additional

[4] We use Amazon Mechanical Turk (https://www.mturk.com/) for data collection.

[5] https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories.

[6] https://en.wikipedia.org/wiki/Category:Wikipedia_Top_25_Report.



What

What are some examples of pseudocereals?

What companies have banned plastic straws from their premises?

What is port wine made out of?

MISC

Thank you. Now can you tell me about Thailand's version of the dish, too?

Do they usually eat it for breakfast or can it be eaten at any time of day?

How

How many basic steps are in the process in coffee making?

Can

Can grenache be found in South Africa?

Are

Are there any popular croissant variants?

I

I am interested in heating methods.

Is

Is there any street food in Switzerland?

Tell

Tell me about tamale in Guam!

What

What was the death toll for the final battle of 1812?

What are some of the limiting factors of Radar?

What happened to David Stirling after the failed operation?

MISC

Which province got hit the hardest by Hurricane Hugo?

I'm interested in the film adaptation from 2000.

The reaction of VC in observing the base from the Nui Dinh hills.

How

How did the people of Iraq view Saddam Hussein?

Can

Can you tell me about the port of Beirut explosion?

Who

Who is the Secretary General of the Front?

Did

Did he serve the whole 30 months?

Tell

Tell me about exploitative governance.

Why

Why didn't people like him?

Figure 3: Left: seed document topic category breakdown ($D \rightarrow$ category *depth*). Middle and right: treemaps of top 7 (and other: MISC) first tokens in user turns from conversations under ‘‘food & drink’’ and ‘‘historical events’’ topic categories. *For each figure, the size of each colored area is proportional to its percentage in the data.*

context. Then, the worker is told to use the provided search engine[7] to find answer(s) from Wikipedia for the current user request. They are asked to select all (up to 4) evidence paragraphs from Wikipedia, which they then use to construct their response. They are also asked to categorize their response, choosing one of four response strategies: $\{$*direct answer* (DIRECT), *clarification* (CLAR), *relevant answer* (REL), and *no information* (NI)$\}$. In contrast to a direct and complete answer, we consider a response as a *relevant answer* when the agent finds information that only partially satisfies the user need (e.g., relax a constraint in the request). For each agent turn, we collect two different annotations to increase reference diversity.

Validation After each user turn, we send the two agent annotations to a validator (Figure 2(c)). For each agent turn annotation, the validator determines whether i) each selected evidence paragraph is properly used in the response; ii) the response is factually consistent with the evidence; iii) the response is coherent to the dialogue context; and vi) the labeled response strategy is faithfully reflected in the response. If both are valid, the validator is asked to rate which one is more comprehensive, where a tie is permitted. An agent response is considered more comprehensive if it contains more information relevant to the user request. The more comprehensive (or the only valid) annotation[8] is then used to continue the conversation. The annotation is terminated if both annotations are invalid, and we include the conversation up to the previous turn in our data.

4.2 Quality Control

Worker Qualification To recruit agent workers, we manually review >150 submissions of a qualification task and select 24 highly qualified workers who consistently produce valid annotations during the qualification. The qualification task consists of 12 agent annotation tasks, where each dialogue context is written by the first two authors of this paper. Similarly, we create different qualification tasks to select 35 qualified users and 10 validators who consistently produce reasonable user responses or validations based on our manual review.

Annotation Control To discourage users from chit-chatting or raising inappropriate requests (e.g., too subjective), each agent worker can decide to either continue the conversation or flag their previous user turn as incoherent or an invalid request. The validation process ensures that only valid agent annotations are included in our final dataset. To encourage extensive search for comprehensive information, we assign a bonus to an agent worker if their annotation is labeled as equally or more comprehensive than the other worker.

We constantly monitor the annotation process and send feedback to workers. Our user and agent workers have over 99% and 96% average passing validation rate, respectively. About 13% of agent annotations are marked as less comprehensive.

Worker Payment Structure We actively communicate with workers throughout the annotation process to clarify any questions they have and

[7] Based on Google Search API from `https://developers.google.com/custom-search` and restricted to the `https://en.wikipedia.org/` domain.

[8] We randomly select one if there is a tie.

	Train	Dev	Test	Total
# Convs	250	86	469	805
# Turns	1443	502	2767	4712
# Turns / Conv	5.8	5.8	5.9	5.9
# References / Turn	1.8	1.6	1.6	1.7
# Tokens / User	10.6	10.5	10.7	10.6
# Tokens / Agent	35.7	44.3	45.1	41.9
# Passages / Agent	1.5	1.7	1.6	1.6

Table 2: Overall statistics of InSCIt.

to give them feedback. We also check in with them early on to make sure they are satisfied with the pay and bonus structure. Most workers report that they are paid with an hourly rate of 15-20 USD, depending on their annotation speed. We pay 0.2/0.5/0.5 USD for each user/agent/validator annotation, plus a 0.1 USD bonus for each agent annotation if the worker passes validation over 80% of the time (all qualified). In addition, we assign a bonus of 0.3 USD to the agent annotation that is marked as equally comprehensive as its peer annotation by the validator, or 0.5 to those marked as more comprehensive or with multiple evidence passages found.[9] On average, we pay over 0.9 USD to each agent annotation.

4.3 Data Analysis

We collect 805 conversations, which includes 4712 user-agent turns after dropping agent annotations if their evidence passages cannot be found in the post-processed Wikipedia corpus.[10] Table 2 shows summary statistics of the train/ dev/test subsets of InSCIt. Word token counts are based on the spaCy (Honnibal et al., 2020) tokenizer. The *test set* contains conversations triggered with seed documents from all 5 topic categories, while the *training* and *dev sets* only contain those from ‘‘food and drink’’, ‘‘hobby’’, and ‘‘top-25’’. In the training set, we keep all valid agent annotations as well as their comprehensiveness comparison results.

In the dev and test sets, we did not include agent responses flagged as less comprehensive during validation. In addition, as discussed in § 4.2, we adjust the worker incentives to obtain more comprehensive responses when collecting dev/test sets, leading to the difference in the average agent turn length.

4.3.1 Diversity of User and Agent Turns

User Request We analyze the distribution of wh-words of user questions, as well as non-question user utterances (e.g., responses to clarification). The treemaps in Figure 3 (middle and right) show the 7 most frequent leading unigrams of user utterances in ‘‘food & drink’’ and ‘‘historical events’’ conversations, respectively. ‘‘MISC’’ refers to utterances with less frequent leading unigrams. Each box size is proportional to its percentage in the data. As shown, most user requests are ‘‘what’’ and ‘‘how’’ questions. There are also many user turns starting with words like ‘‘can’’ and ‘‘tell’’, most of which are responses to agent clarification questions. The user utterances are fairly long-tailed as ‘‘MISC’’ shares a large portion (about 30%) for both treemaps. Instead of being mostly factoid questions, open-ended user requests are well represented in InSCIt.

Agent Response Strategy Table 4 shows the diversity of agent response strategies in InSCIt. When no direct answer exists, agents in InSCIt can respond to the user with a *relevant answer* (see § 4.1). If no direct or relevant answer is found, the agent can then respond with *no information*. The average response length and number of evidence passages differ dramatically across various response strategies. Compared with direct or relevant answer cases, *clarification* responses tend to be shorter and are more likely to happen when more evidence passages are present. We also calculate that 30% *direct* or *relevant answer* agent turns have multiple evidence passages, which potentially require information summarization.

4.3.2 Analysis of Agent Initiatives

In this section, we present qualitative analysis to understand how different agent initiatives get triggered, with a focus on *clarification* and *relevant answer* agent responses.

Fine-Grained Categorization We randomly sample and analyze 100 clarification and relevant answer responses respectively. Table 3 (upper half) shows that in most cases, the agent raises

[9] At the beginning of our training set collection (before the collection of dev/test sets), we only assign a 0.3 USD bonus to agent annotations marked as more comprehensive. After communicating with our workers, we adjust our bonus structure, which leads to more comprehensive agent responses.

[10] We use wikiextractor to process Wikipedia articles: `https://github.com/attardi/wikiextractor`.

Clarification (CLAR)	
Too long / many answers (86%)	. . . **<Agent>:** In the Battle of New Orleans, . . . the Americans had 13 dead, totaling 304 dead. **<User>:** Were there any long-term consequences that came as a result of the War of 1812? **<Agent>:** There were many! Would you like to know what they were for *Bermuda, . . . , or Great Britain*?
Ambiguous entity (13%)	**<User>:** Washington University is classified as what for its high research activities? **<Agent>:** Do you want to know about *Washington University in St. Louis or in Baltimore, Maryland*?
No Direct but Relevant Answer (REI)	
Constraint relaxation / No definite answer (70%)	**<User>:** Was the Matrix franchise adapted from any other work? **<Agent>:** While not explicitly adapted from another work, *the Matrix did draw heavily on Jean* . . .
	. . . **<User>:** Who authored the Torah? **<Agent>:** *It was originally thought that a majority of the Torah was written by . . . However, it's now thought that . . . though the number of authors is up for debate.*
Relevant but side info only (29%)	. . . **<User>:** What countries have an ecological footprint under 5 hectares per person? **<Agent>:** *The world-average ecological footprint in 2013 was 2.8 global hectares per person* . . . But I don't have a list of countries with an ecological footprint under 5 global hectares per person.

Table 3: Examples of clarification and no-direct-but-relevant-answer agent responses. *Factual information from evidence passages is italicized in agent responses.*

	DIRECT	CLAR	REL	NI
% Turns	71.5	12.7	13.1	2.7
# Tokens	43.7	33.5	46.6	10.6
# Passages	1.5	2.8	1.4	0.0

Table 4: Agent response strategy statistics. DIRECT, CLAR, REL, and NI indicate direct answer, clarification, no direct but relevant answer, and no information.

a clarification when they find a long answer or too many answers (86%) or notice an ambiguous entity in the user request (13%). In 70% of relevant answer cases (bottom half of Table 3), the agent relaxes some constraint in the user request or provides evidence that no definite answer can be found. In 29% of these cases, they simply provide some relevant but side information only. We also observe that in rare cases (1%), the agent points out some mistake (e.g., a false assumption) in the user request.

Clarification Occurrences We next look at contexts where agents are more likely to ask for clarification in a conversation. Clarification questions occur more frequently at the very beginning (ex. 2, Table 3), rather than later in a conversation (18.8% vs. 11.5%). If a clarification is raised in the previous agent turn, the chance of a subsequent clarification (see Table 6) is 7.6%, compared to 12.2% if the previous turn is not a clarification (ex. 1, Table 3).

Response Strategy Selection In 23% of examples with 2 agent annotations marked as equally comprehensive by validators, workers take *different response strategies given the same dialogue context*. Of this set, 82% have different evidence passages labeled by the two workers, potentially due to the open-endedness of user queries in INSCIT and the large knowledge source. In addition, as suggested by our analyses in § 6.2, it is more likely that agents will choose different evidence passages when there is no direct answer to the question. As illustrated in the first example in Table 5, the different evidence passages often trigger different agent response strategies. The second and third examples show that even if two agents find the same evidence set, deciding whether it indicates an under-specified user request, a direct or only a relevant answer can be subjective.

5 Experiment Setup

5.1 Systems

We build two systems for each of the tasks formulated in § 3. Both systems build on retriever-reader models, inspired by recent advances in open-domain single-turn or conversational question answering (Karpukhin et al., 2020; Izacard and Grave, 2021; Adlakha et al., 2022). Here, the main function of the retriever is to gather a ranked set of top-k candidate evidence passages from the entire Wikipedia to facilitate passage identification and response generation for the later reader model. We first describe the retriever models, and

Different Evidence; Different Response Strategies

Dialogue Context: . . . **<User>:** What are some of the health consequences of oats?

<Agent> (CLAR): Oat related health concerns include oat allergies, and gluten-related disorders. Which type of health issue would you like to learn more of? *(evidence from Wikipedia page: Oat Sensitivity)*
<Agent> (DIRECT): Oat toxicity in people with gluten-related disorders depends on the oat cultivar consumed because the immunoreactivities of toxic prolamins are different among oat varieties. Furthermore, oats are frequently cross-contaminated with the other gluten-containing cereals . . . *(evidence from Wikipedia page: Oat)*

Same Evidence; Different Response Strategies

Dialogue Context: . . . **<User>:** Who is the major competitor of PepsiCo?

<Agent> (CLAR): Would you like to know PepsiCo's major competitor in the beverage market or the snack food market?
<Agent> (DIRECT): PepsiCo has two main competitors in different market sectors. The Coca-Cola Company is its largest competitor in the beverage market, and Kraft Foods is its largest competitor in the snack foods market.

Dialogue Context: . . . **<User>:** Were the directors involved in any controversy?

<Agent> (REL): I couldn't find anything about themselves in any controversy but their film . . . had a controversial storyline and themes.
<Agent> (DIRECT): The Wachowskis' V for Vendetta was considered to have controversial storyline and themes, but have been both criticized and praised by sociopolitical groups.

Table 5: Examples of two agent reference responses with different response strategies.

. . . **<User>:** What kinds of regional varieties are there?
<Agent>: Would you like to know about East Asia, Southeast Asia, South Asia, or Europe?
<User>: Tell me about East Asia.
<Agent>: Sorry, but each country is detailed as well, do you want to know more about congee in China, Japan, Korea or Taiwan?

Table 6: An example of consecutive clarifications.

then introduce the two reader models that perform the two main tasks based on retrieval results.

5.1.1 Retriever Models

We experiment with two retrievers: BM25 and DPR. **BM25** (Robertson and Zaragoza, 2009) uses sparse bag-of-word representations for ranking passages with regard to each query. We use Pyserini (Yang et al., 2017) in our experiments. **DPR** (Karpukhin et al., 2020) is a BERT-based (Devlin et al., 2019) dual encoder model that produces learned dense representations for queries and passages, and measures the relevance using the dot product similarity in the vector space. We finetune DPR on InSCIt. As the training set is small in InSCIt, we initialize it with a downloadable checkpoint[11] that is pre-trained on a much larger ($> 30\times$) open-domain conversational question answering dataset, TopioCQA (Adlakha et al., 2022).

5.1.2 Reader Models

Our two readers are based on state-of-the-art models in open-domain question answering and conversational knowledge identification—Fusion-in-Decoder (Izacard and Grave, 2021) and DIALKI (Wu et al., 2021).

Fusion-in-Decoder (FiD) FiD is a generative reader model. It first encodes all retrieved passages with a given query, and then decodes the task output (e.g., an answer string) by attending over all encoded passages. To adapt FiD to our tasks, we prepend a passage identifier (ID) to each of the top-k retrieved passages (here, $k = 50$, following Adlakha et al. [2022]) and separately concatenate each passage with the dialogue context for encoding. Given the 50 encoded contextualized passage vectors, the decoder generates a sequence of evidence passage IDs (*passage identification*), followed by the final response (*response generation*). After the first turn, the encoded passage vectors associated with $\{\mathcal{P}_i, \ldots, \mathcal{P}_{n-1}\}$ are concatenated with the top-k retrieved passages, limiting k to give a total of 50. In training, we use the same hyperparameters as in Adlakha et al. (2022), with the batch size adjusted for the memory constraint and training steps adjusted to have the same epochs.

DIALKI + FiD The second reader adopts a pipelined approach to perform the two tasks. It first uses DIALKI (Wu et al., 2021) to select evidence passages and then feeds the identified passages into FiD to generate the agent response. DIALKI is a state-of-the-art conversational knowledge identification model that incorporates dialogue structure with a multi-task learning framework. DIALKI predicts a passage score for each input passage (i.e., each top-k retrieved passage). To adapt it

[11] https://github.com/McGill-NLP/topiocqa.

for passage identification, we simply keep evidence passages (up to 4, as in data collection) with ranking scores higher than γ for *multiple passage prediction*, where γ is tuned on the dev set. We apply the same method to incorporate previously used evidence passages into DIALKI as in the first reader model. We set the number of input passages of DIALKI to be 50 and keep other original hyperparameters. Parameters in FiD are the same as the first reader model, except that the number of input passages is 4 in the DIALKI+FiD system.

Trivial Baseline: Last Turn We report performance of a simple baseline: use the most recent agent turn in the dialogue context and associated evidence ($\bar{\mathcal{P}}_n = \mathcal{P}_{n-1}$; $\bar{a}_n = a_{n-1}$). For first-turn instances, we use the most frequent evidence passage and agent response seen in the training set as the prediction. We also tried using a random previous turn as the prediction, which gives lower scores than using the last turn.

Human We collect one additional annotation for each agent turn in the test set and evaluate it as the human performance. These additional annotations are annotated by the same agent workers we select in § 4.2. Note that these additional prediction data do not go through the same validation step as those that are used as references.

5.2 Evaluation

Below, we describe automatic metrics and a human evaluation protocol for the passage identification (PI) and response generation (RG) tasks in § 3.

Passage Identification INSCIT allows for multiple evidence passages, so we measure the model performance by computing the F1 score (PI-F1), comparing the set of predicted evidence passages $\bar{\mathcal{P}}_n$ to the set of reference passages $\mathcal{P}_n$. For turns where there are two valid reference annotations, we use the maximum F1 score between the two.

Response Generation For a generated agent response $\bar{a}$, we calculate the SACREBLEU score (Post, 2018) (BLEU in tables) and token-level F1 (RG-F1) scores against the reference response, following previous studies (Feng et al., 2020; Adlakha et al., 2022). Again, when there are two valid annotations, we use the maximum.

Human Evaluation As the two tasks are dependent on each other, decoupled automatic evaluations may not capture aspects like factual consistency between predicted passages and the response. Moreover, handling queries with multiple evidence passages or no direct answer can be open-ended.

Therefore, we design a human evaluation protocol to evaluate the model performance on both tasks.[12] Specifically, we focus on the evaluation of 4 dimensions: 1) *evidence passage utility*: how many predicted evidence passages are used in the generated response; 2) *factual consistency* between the predicted response and evidence; 3) response *coherence* with the dialogue context; and 4) response *comprehensiveness*: how much information, that is both relevant to the user request and factually consistent with the predicted evidence, is contained in the response. While most prior work on information-seeking dialogues only relies on automatic metric scores (Choi et al., 2018; Anantha et al., 2021; Adlakha et al., 2022), a few studies collect human ratings on dimensions like response ‘‘coherence’’ and ‘‘informativeness’’ (Gao et al., 2022; Feng et al., 2022). However, as they do not require models to predict evidence, the factual consistency between the response and the knowledge source cannot be evaluated (Nakano et al., 2021).

We provide outputs for both tasks of our two systems and ‘‘Human’’ to a human judge. We ask them to rate the first 3 dimensions for each system output on a 4- or 5-point Likert scale[13] and then rank the system responses in terms of *response comprehensiveness* (ties are permitted). We have 3 raters for each agent turn and take the average rating score or rank place on each dimension for each system. Since human evaluation can be time-consuming and costly, we run it on a sampled test subset with 50 conversations (290 turns) and encourage future studies to report on the same subset.

The inter-rater agreement measured as Krippendorf's alpha is 0.66, 0.64, 0.42, and 0.37 for EU, FC, CO, and COMP, respectively, which can be interpreted as good or moderate agreements. We observe two main types of coherence disagreements: 1) some workers are more strict and

[12] We release the code at https://github.com/ellenmellon/INSCIT/tree/main/eval/human_eval.

[13] The 4-point scale is used only for coherence to discourage *neutral* ratings. We report all scores normalized to a 1-5 scale.

Retriever	**Reader**	PI-F1	BLEU	RG-F1
Last Turn	Last Turn	10.5	4.2	14.1
BM25	FiD	14.1	9.4	22.5
	DIALKI + FiD	17.0	13.8	24.8
DPR	FiD	17.1	8.8	21.6
	DIALKI + FiD	**21.5**	**16.6**	**26.6**

Table 7: Automatic scores on the **dev** set.

indicate one response as more preferred due to minor differences (e.g., a connecting word), or 2) both responses are incoherent, but in very different ways (e.g., have very different content). Similarly, most comprehensiveness disagreements involve either: 1) two responses that are similar except that one includes additional side information, or 2) two responses that provide different answers but both are good.

6 Experiment Results

6.1 Quantitative Results

Table 7 shows the overall automatic evaluation results of all systems for our main tasks (PI and RG) on the *dev* set. The simple baseline performs very poorly. Using retrieval results from DPR (vs. BM25) leads to the best overall performance for both tasks. For both BM25 and DPR retrievers, DIALKI + FiD achieves better performance than FiD in all metrics. A possible reason could be that the smaller number of context vectors used with DIALKI+FiD is better suited to learning from limited data than the end-to-end FiD approach. DIALKI leverages previous evidence passages in passage identification, so its following FiD response generation has only 4 context vectors (vs. 50 for FiD). This hypothesis is supported by the observation that incorporating previously used evidence hurts the RG performance slightly for FiD but for DIALKI+FID it helps (roughly 1 point decrease vs. increase in scores, respectively, with DPR).

Table 8 shows both automatic and human evaluation results on the *test* set for FiD and DIALKI+FiD with the DPR retriever, confirming the dev set findings. Experiments with BM25 also confirm dev set trends. Figure 4 presents comparative human evaluation results. DIALKI+FiD greatly outperforms FiD except in coherence where scores are similar. DIALKI+FiD substantially underperforms humans in both automatic and human scores, except for factual consistency where the difference is small. This could indicate that, although DIALKI+FiD generates responses consistent with the predicted evidence, it identifies less relevant passages which lead to less coherent and less informative responses.

Retriever	**Reader**	**Automatic**			**Human**		
		PI-F1	BLEU	RG-F1	EU	FC	CO
DPR	FiD	17.5	9.6	22.2	2.35	2.52	3.76
	DIALKI + FiD	**23.7**	**16.0**	**27.8**	**4.33**	**4.74**	**3.77**
–	Human	52.5	33.8	43.5	4.76	4.77	4.85

Table 8: Automatic scores on the **test** set, and human scores on 50 *sampled* test conversations (290 turns) for dimensions rated with Likert scales: evidence utility (EU), factual consistency (FC), and coherence (CO).

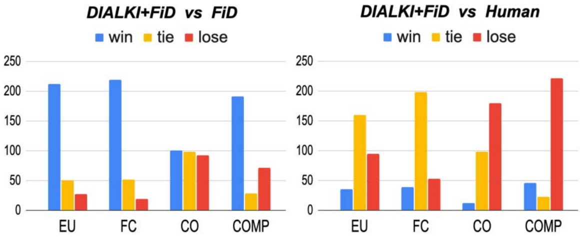


Figure 4: Human evaluation on system comparison for all dimensions: evidence utility (EU), factual consistency (FC), coherence (CO), and response comprehensiveness (COMP). Win/lose refers to DIALKI+FiD.

The reason for *imperfect human performance* on passage identification, shown in Table 8, is two-fold. Due to the open-endedness of information-seeking queries in InSCIt and the large search space over Wikipedia, annotators may find different (but both valid) sets of evidence passages. In addition, annotations corresponding to the Human ''system'' do not go through the validation process, so they could have errors or be less comprehensive.

6.2 Analysis

Passage Retrieval Table 9 reports the performance for passage retrieval in HIT@k scores, following Karpukhin et al. (2020) and Adlakha et al. (2022). HIT@k is calculated as $\mathbb{1}\left[|\mathcal{R}_K \cap \mathcal{P}| > 0\right]$, where $\mathcal{R}_K$ denotes the top K retrieved passages and $\mathcal{P}$ denotes the union of the two reference passage sets (or a single reference set if only one is valid). We evaluate both BM25 and DPR models used in our main experiments, as well as two DPR ablations: with pretraining (PT) on TopioCQA or finetuning (FT) on InSCIt only.

Retriever	Dev		Test	
	HIT@20	HIT@50	HIT@20	HIT@50
BM25	35.3	48.0	35.6	48.1
DPR (FT only)	62.5	70.1	51.3	60.8
DPR (PT only)	66.4	76.3	68.4	**77.5**
DPR	**71.1**	**79.8**	**69.9**	**77.5**

Table 9: *Passage retrieval* results. PT and FT refer to pretraining on TopioCQA and finetuning on InSCIt.

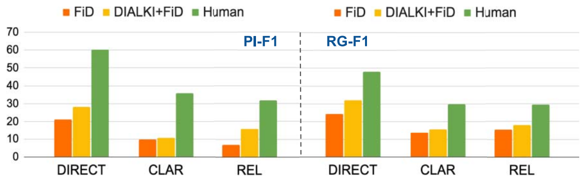


Figure 5: PI-F1 and RG-F1 scores by reference response strategy (direct answer, clarification, relevant answer) on the one-strategy test subset, excluding instances where two references differed in strategy.

BM25 underperforms DPR models significantly, which explains the main task performance differences between BM25 and DPR in Table 7. DPR with PT alone is more effective than FT only, which can be explained by the much larger training data in TopioCQA. The best retrieval results are achieved with PT and FT combined. We do not leverage TopioCQA for pretraining on our two main tasks, because 1) it does not come with the passage identification task and only has short answers or `no answer` as their agent responses; 2) we observe poor zero-shot response generation performance on InSCIt for FiD trained on TopioCQA.

Passage Identification & Response Generation Performance Breakdown Figure 5 shows the system and task performance breakdown by *reference response strategy* (direct answer, clarification, and relevant answer) for the test set, excluding examples where two annotations differed in the response strategy category (16%). DPR is used for retrieval. Only RG-F1 is shown for response generation; trends for BLEU are similar. For all response types, DIALKI+FiD is similar or outperforms FiD, but significantly underperforms humans. For both systems and humans, the non-direct-answer responses have lower automatic scores. The lower PI-F1 scores for humans suggest that the retrieval task is more difficult (with more variety in evidence) when a simple direct answer is not available. Lower automatic response generation scores may be explained by lower retrieval scores (less reliable evidence), larger number of passages, and/or challenges in learning non-direct-answer response strategies. Note that for both systems, the largest percentage gap with respect to human scores is for clarifications.

Response Generation Results with Human Evidence Passages To explore the above hypotheses, we generated responses using the DIALKI+FiD response generator with passages selected in the ‘‘Human’’ annotation of the test data. The resulting responses had 26.5 and 37.4 for BLEU and RG-F1 scores, respectively, compared to 16.0 and 27.8 when using DIALKI passages. We sample and analyze 20 examples each of single and multiple ‘‘Human’’ evidence passages. Given multiple evidence passages, most DIALKI+FiD responses either do not use all passages or introduce incorrect facts. With one passage, responses are consistent with the evidence but not always as comprehensive as for humans. In the 20 examples with multiple passages, DIALKI+FiD asks one clarification, whereas humans ask nine.

Impact of Response Type Prediction for Response Generation As explained in § 4.3, the agent response type depends on the selected evidence passages. To analyze how incorporating response types can help with response generation, we conduct a controlled experiment to generate agent responses with the dialogue context and oracle evidence passages as the input to FiD, and compare the performance when no/oracle/predicted response type is given. For examples that have two labels with different sets of evidence passages, we split them into two separate instances. To predict the response type, we use a sequence classification model based on BERT-base (Devlin et al., 2019), given the dialogue context and oracle evidence passages. To provide the oracle or predicted response type as the response generation model input, we simply append a formatted string—`response type: {response_type_name}`[14]—at the end of the dialogue context, when feeding it to FiD.

[14] Candidate response type names are ‘‘direct answer,’’ ‘‘clarification,’’ ‘‘no answer but relevant information,’’ and ‘‘no answer and no information.’’

Model Input	Dev		Test	
	BLEU	RG-F1	BLEU	RG-F1
DC+OEP+RT (Oracle)	32.6	48.7	31.6	47.4
DC+OEP+RT (Predicted)	32.6	46.3	31.7	45.4
DC+OEP	32.0	45.3	30.6	44.3

Table 10: Automatic RG scores for FiD with inputs: dialogue context (DC), oracle evidence passages (OEP), and different (oracle/predicted/no) response types (RT).

The response type classification model gives an overall accuracy of 0.75, compared to 0.73 when predicting everything as ‘‘direct answer.’’ Table 10 shows that adding either oracle or predicted response types improves BLEU and RG-F1 scores, compared with no response type being used, with greater gains in RG-F1 for oracle response type. We observe consistent performance gains on examples with either ‘‘direct answer’’, ‘‘clarification’’, or ‘‘no information’’ oracle response types, but not for the ‘‘relevant answer’’ response type.

7 Conclusion & Discussions

In summary, we introduce InSCIt, a new open-domain information-seeking conversational dataset grounded in Wikipedia, with mixed-initiative user-agent interactions. InSCIt supports two tasks (passage identification and response generation), for which we present results of two strong baselines, with best results obtained with the pipelined DIALKI+FiD system. We also introduce a human evaluation protocol.

Future Work Both models significantly underperform humans in both tasks in all metrics. The relative performance gap is greatest for scenarios that require the agent to provide a non-direct answer. We find that passage identification significantly impacts response generation (particularly coherence) by providing relevant grounding knowledge. Thus, improving methods for selecting relevant passages is critical for future work. Key challenges that remain in response generation are how to fuse and present comprehensive information from multiple passages and learning when and how to use non-direct response strategies. Given the small size of our training data, another future direction is to explore transfer learning using existing information-seeking conversation or question answering resources.

Our work focuses on different strategies that can be adopted by the agent to better address user requests in a conversational question answering setting, assuming the user will either ask an information-seeking question or provide a clarification to the agent. Exploring more user-side strategies would be interesting for handling system errors and other types of conversations (e.g., negotiations).

In contrast to Wikipedia passages, information sources used in practice (e.g., the whole Web) can often contain less trustworthy information. In such cases, retrieving evidence passages containing the same answer and predicting the trustworthiness of each answer based on all such retrieved passages can be a promising direction.

Another direction that is worth future exploration lies in the design of evaluation metrics. We follow previous studies to evaluate the model performance when given a fixed human-human dialogue context. However, as pointed out by Li et al. (2022), an interactive dialogue system often needs to handle dialogue contexts containing errors made by the model itself. Therefore, it is important for future work to develop new methods for automatic evaluation and scalable human evaluation in the interactive setting.

8 Ethical Considerations for Dataset Collection

Our work is primarily intended to encourage future work in information-seeking conversation *factually grounded* in given knowledge sources. Our knowledge sources come from Wikipedia articles, where the content follows principles emphasizing on a neutral point of view and reliable sources. Before and during the data collection, we carefully guide our user workers not to ask subjective or opinion-driven questions, and our agent workers not to include any content without evidence from the knowledge sources in their conversational responses. Therefore, all contents exposed to our workers during data collection should contain minimal risk to the workers. Our data collection has IRB approval from University of Washington and is deemed exempt. We also actively communicated with the workers to address any concern they had and we usually replied back within an hour during the whole data collection process. This communication also helped us to make sure that our workers were compensated fairly. As explained in § 4.2, most of our workers report that they are paid with an hourly rate of 15-20 USD.

Acknowledgments

This research was supported by NSF IIS-2044660, ONR N00014-18-1-2826, DARPA MCS program (N66001-19-2- 4031), a Sloan fellowship and gifts from AI2. We thank Tao Yu and all members in H2Lab and TIAL labs at University of Washington who participated in our data collection pilot studies. We thank Kevin Everson for proofreading our paper and providing valuable suggestions. We also thank members of University of Washington's NLP groups who provided feedback and insights to this work. We specifically want to thank our TACL reviewers and action editor for their detailed comments that lead to great improvement of our paper.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483. https://doi.org/10.1162/tacl_a_00471

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.367

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. https://doi.org/10.1145/3331184.3331265

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.44

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.652

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1241

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Song Feng, Siva Patel, and Hui Wan. 2022. DialDoc 2022 shared task: Open-book document-grounded dialogue modeling. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 155–160, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.dialdoc-1.18

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.498

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.652

Chang Gao, Wenxuan Zhang, and Wai Lam. 2022. UniGDD: A unified generative framework for goal-oriented document-grounded dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.66

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5110–5117. https://doi.org/10.1609/aaai.v32i1.11977

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coQA: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.74

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

Natalia Konstantinova and Constantin Orasan. 2013. Interactive question answering. *Emerging Applications of Natural Language Processing: Concepts and New Research*, pages 149–169. https://doi.org/10.4018/978-1-4666-2169-5.ch007

Huihan Li, Tianyu Gao, Manan Goenka, and Danqi Chen. 2022. Ditch the gold standard: Re-evaluating conversational question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8074–8085, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.466

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018.Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1255

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing mantis: A novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

V. Rieser and O. Lemon. 2009. Does this list contain what you were searching for? Learning adaptive dialogue strategies for interactive question answering. *Natural Language Engineering*, 15(1):55–72. https://doi.org/10.1017/S1351324908004907

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389. https://doi.org/10.1561/1500000019

Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. 2020. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.655

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1233

Haitian Sun, William Cohen, and Ruslan Salakhutdinov. 2022. ConditionalQA: A complex reading comprehension dataset with conditional answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3627–3637, Dublin, Ireland. Association for Computational Linguistics.

N. Webb and B. Webber. 2009. Special issue on interactive question answering: Introduction. *Natural Language Engineering*, 15(1):1–8. https://doi.org/10.1017/S1351324908004877

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.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1253–1256, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3077136.3080721

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.586

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1076