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 ♣Microsoft Research ◇Allen Institute for AI
{zeqiuwu1,rparish,sewon,ostendor,hannaneh}@uw.edu
chehao@microsoft.com raja@allenai.org

Abstract

In an information-seeking conversation, a user converses with an agent to ask a series of questions that can often be under- or over-specified. An ideal agent would first identify that they were in such a situation by searching through their underlying knowledge source and then appropriately *interacting* with a user to resolve it. However, most existing studies either fail to or artificially incorporate such agent-side initiatives. In this work, we present INSCIT (pronounced Insight), a dataset for **In**formation-**S**eeking **C**onversations with mixed-initiative **Int**eractions. It contains a total of 4.7K user-agent turns from 805 human-human conversations where the agent searches over Wikipedia and either asks for clarification or provides relevant information to address user queries. We define two subtasks, namely evidence passage identification and response generation, as well as a new human evaluation protocol to assess the model performance. We report results of two strong baselines based on state-of-the-art models of conversational knowledge identification and open-domain question answering. Both models significantly underperform humans and fail to generate coherent and informative responses, suggesting ample room for improvement in future studies.[1]

1 Introduction

Recently, there are increasing interests in developing conversational information-seeking systems (Choi et al., 2018; Adlakha et al., 2022; Saeidi et al., 2018; Feng et al., 2020), where the system assists users in finding information from knowledge sources (e.g., text corpus) via multi-turn conversational interactions. One important advantage of conversational information-seeking systems is that users do not need to come up with a very descriptive query by themselves (Webb and Webber,

---

[1]We release all data and code for experiments and human evaluation at https://github.com/ellenmellon/INSCIT.
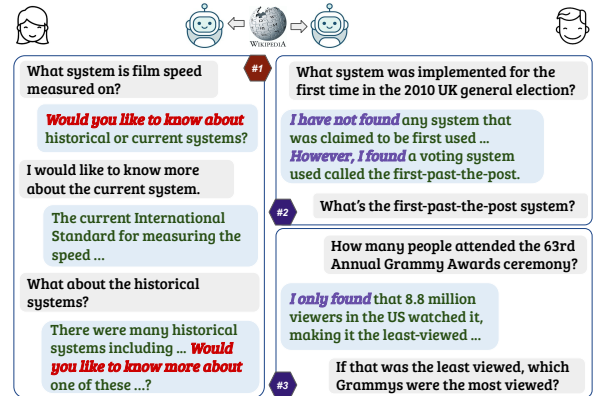

Figure 1: The user often asks under-specified questions that require clarification (left), or asks questions with no direct answer but where providing relevant information can possibly fulfill their information needs (right).

2009; Rieser and Lemon, 2009; Konstantinova and Orasan, 2013). Instead, as shown in Figure 1, they 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 (#1) or relax their queries and even proactively suggest relevant information (#2 & 3) that may partially satisfy the user's information needs. However, existing information-seeking conversation datasets rarely contain such mixed-initiative interactions. For example, most existing 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 the envisioned information-seeking interactions (Saeidi et al., 2018; Feng et al., 2020; Aliannejadi et al., 2021; Guo et al., 2021).

In this paper, we introduce INSCIT, a dataset for **In**formation-**S**eeking **C**onversations with mixed-initiative **Int**eractions, where agents take various

strategies, such as providing direct answers (72%), raising clarification questions (13%), and presenting relevant information as indirect answers (13%), to address users' information needs. It contains 805 natural human-human information-seeking conversations with 4.7K user-agent turns over diverse topics. To simulate realistic information-seeking scenarios, users write queries with minimal restriction, and human agents decide on different strategies to respond, after searching over Wikipedia and with evidence paragraphs annotated. Through a design of scalable annotation pipeline and careful quality control, we were able to collect high-quality data (over 96% annotations pass our validation).

We formulate two tasks for the conversational agent system, where it should (1) identify a set of evidence passages from Wikipedia, and (2) generate a response grounded in the evidence. Comparing with previous 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 information from *multiple* evidence passages to construct responses that reflect various *strategies*. Since handling queries with multiple evidence passages or no direct answer is more open-ended, we emphasize the need for human evaluation, and propose a more systematic human evaluation protocol that considers diverse aspects including coherence, factual consistency and information comprehensiveness, with both the predicted evidence passages and response provided as the evaluation input.

We present two strong baselines based on the state-of-the-art in open-domain question answering (Karpukhin et al., 2020; Izacard and Grave, 2020) and conversational knowledge identification (Wu et al., 2021). Results indicate that, while these systems achieve substantial improvements over trivial baselines, 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 information from multiple passages by leveraging different response 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' initial and follow-up information needs

Dataset	IR	Response Strategy		H-H Dialogue	Multi-Evidence
		CLAR	REL		
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.

with grounding in knowledge sources. Table 1 shows the comparison between INSCIT and existing 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 in a conversation 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 simply outputs `no answer` if there is no direct answer.

A few other studies create artificial conversations to address ambiguous user questions. For instance, Qulac (Aliannejadi et al., 2019) and the data collected in their follow-up work (Aliannejadi et al., 2021) are based on user queries containing a multi-faceted entity and intentionally annotated agent clarification questions; 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. User queries that are open-ended and ambiguous have also been observed in the single-turn question answering task (Min et al., 2020; Zhang and Choi, 2021; Sun et al., 2022), which is usually addressed by training a model to predict multiple conditional answers without further interactions. In contrast, INSCIT consists of human-human conversations

with natural information-seeking user requests and mixed agent initiative to address them.

Gustavo Penha and Hauff (2019) crawl conversations from Stack Exchange.[2] These conversations are mixed with information-seeking utterances and casual talk, and one grounding document is heuristically obtained for each conversation. In contrast, INSCIT consists of clean information-seeking dialogues with annotated and validated grounding knowledge for each agent turn.

Knowledge-Grounded Social Chat Different from information-seeking dialogues, 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) collect a dataset for training a conversational agent to select knowledge to present based on the user's background, with the aim to maintain the user's interest in the conversation.

3   Task Formulations

Before introducing our data, we first formulate two tasks for INSCIT, namely *passage identification* and *response generation*. These two tasks mimic how an agent responds to each information-seeking user request, where the agent first searches for relevant information over the knowledge source and then constructs 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 information from *multiple* evidence passages to construct responses using various strategies, rather than a single passage and a short answer.

At the nth agent turn, both tasks have the same input - the dialogue context $X = [u_1, a_1, u_2, a_2, ..., u_n]$, the corpus of all passage candidates $\mathcal{C}$ as well as the previously used passages $\{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_{n-1}\}$ where each $\mathcal{P}_i = \{p_i^1, p_i^2, ..., p_i^{|P_i|}\}$ is the set of passages used in the ith agent turn a_i. $\mathcal{C}$ is defined as all textual paragraphs[3] in the full Wikipedia dump.[4]

Instead of treating the *passage identification* as a ranking problem (i.e., passage retrieval), we require

the model to predict a *set* of passages $\bar{\mathcal{P}}_n$ from $\mathcal{C}$, that are relevant to the current user request u_n in the dialogue context X and provide evidence for the *response generation* task—generating the next agent response $\bar{a}_n$. Specifically, identifying 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 $\bar{a}_n$.

4   Our Data: INSCIT

Now we introduce INSCIT, a new dataset of information-seeking conversations where the agent interprets the user intent and provides *comprehensive* information grounded in Wikipedia via natural human-human *interactions*. In the following, we first present our data collection pipeline (§ 4.1),[5] and then explain how we define and control the data quality (§ 4.2). Lastly, we highlight characteristics of INSCIT that illustrate the diversity of user and agent turns, as well as interesting observations in agent initiatives and dialogue structures (§ 4.3).

4.1   Annotation Pipeline

We recruit user, agent and validation workers to annotate user / agent turns and validate agent annotations, respectively. Due to the asymmetric time spent by the 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). Thus, the annotation of each dialogue progresses in a pipeline, and no worker needs to wait for the other party. This pipeline (Wen et al., 2017) was introduced to collect human-human task-oriented dialogues and was inspired by the Wizard-of-Oz paradigm (Kelley, 1984). It has proved to be efficient while not hurting the conversation coherence by requiring each worker to read all previous utterances before annotation.

Figure 2 illustrates the overview of the 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, user and agent turns are annotated sequentially until the end of the conversation is reached.

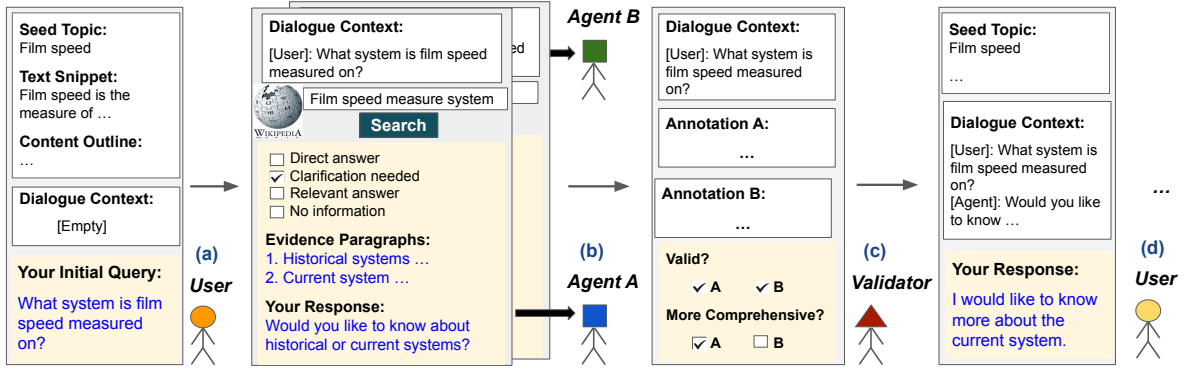
Figure 2: Each conversation is annotated in a chained user → agent → validator tasks. A single agent is dedicated to each user / validator task but two agents work in parallel to complete the agent turn annotation (see discussion in §4.1). White and yellow boxes are inputs and outputs for each annotation task, respectively.

Validation follows each user-agent turn. We collect up to 7 user-agent turns for each conversation.

Seed Document Selection To diversify conversation topics, we select 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 in Wikipedia. To further diversify the pool of seed documents, we leverage the top-down tree structure of Wikipedia categories[6] and sample Wikipedia pages at various tree depths under each of the first 4 categories. Weekly top-25 pages are from Wikipedia weekly top25 reports of the year 2021.[7] See Appendix A.1 for details. 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 Wikipedia article, and is instructed to ask a question they are interested to know 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 utterances in the conversation (i.e., dialogue context).

Agent Turn Different from the user worker, in addition to the dialog context, each agent worker

(Figure 2 (b)) is given all evidence paragraphs used by each previous agent turn as additional context. Then, the worker is told to use the provided search engine[8] to find answer(s) from Wikipedia for the current user request. They are asked to select all (up to 4) evidence paragraphs from Wikipedia that they use to construct their final response. Based on what they find, they can choose one of four response strategies: {*direct answer*, *clarification*, *relevant answer*, *no information*}. In contrast to a direct and precise 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 agent turn, we send two annotations to a validator for checking whether each annotation is valid (details in § 4.2). If both are valid, the validator is asked to rate which one is more comprehensive. An agent response is considered as more comprehensive if it contains more information relevant to the user request. The more comprehensive (or the only one valid) annotation[9] is then used to continue the conversation. The annotation is terminated if both annotations are invalid.

4.2 Quality Control

Worker Qualification As the agent annotation is challenging, to recruit agent workers, we manually review more than 150 submissions of a qualification task and select 24 highly qualified agent

---

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

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

[8]based on Google Search API from https://developers.google.com/custom-search and restricted to the https://en.wikipedia.org/ domain.
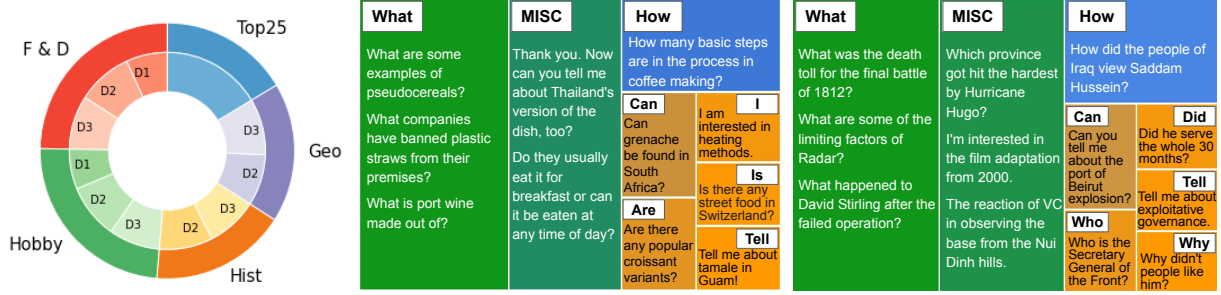
[9]We randomly select one if there is a tie.

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.*

workers who consistently produce valid annotations during the qualification. We use two different qualification tasks to select 35 qualified users and 10 validators before the annotation (details in Appendix A.1).

Annotation Control To discouraging 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 information request. Similarly, for each agent turn annotation, a validator (Figure 2 (c)) determines whether i) each selected evidence paragraph is properly used in the response; ii) the response 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. Only valid user and agent annotations are included in our final dataset.

To encourage workers to extensively search for relevant information to each user request, we assign a bonus to an agent worker if the validator labels their annotation to be 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% agent annotations are marked as less comprehensive. See more annotation details in Appendix A.1.

4.3 Data Analysis

In this section, we first explain the data preparation and the overall statistics of INSCIT. Then, we dive into detailed discussions about the diversity of our user and agent turns (§ 4.3.1) and analysis of agent-side initiatives (§ 4.3.2). We include discussion about natural topic changes in INSCIT

conversations in Appendix A.2.

Data Preparation & Overall Statistics Table 2 shows the summary statistics of INSCIT. We collect 4712 user-agent turns from 805 conversations. To develop information-seeking systems, we split the data into train/dev/test sets. 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". The controlled topic distribution shift can be used to evaluate the developed model robustness. There are additional distribution changes from training to dev / test sets. First, we remove those agent responses flagged as less comprehensive from the reference set for dev and test sets. Instead, we keep all valid agent annotations as well as their comprehensiveness comparison results in the training set. This explains the variation among different sets in the number of references per turn in Table 2. In addition to this, we adjust the worker incentives when collecting dev / test sets, leading to the difference in the average agent turn length. Also, we drop agent annotations if their corresponding evidence passages cannot be found in the corpus during post-processing.[10] For word counting, we use the Spacy tokenizer.[11]

4.3.1 Diversity of User and Agent Turns

User Request The middle and right treemaps in Figure 3 show the 7 most frequent leading unigrams of user utterances from conversations under the "Food & Drink" and "Historical Events" topic categories respectively. "MISC" refers to utterances with less frequent leading unigrams. Each box size in each treemap is proportional to its percentage

---

[10]We use wikiextractor to process Wikipedia articles: https://github.com/attardi/wikiextractor.
[11]https://spacy.io/api/tokenizer

	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.

	DIRECT	CLAR	REL	NEI
% 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 3: Agent response strategy statistics. DIRECT, CLAR, REL, and NEI indicate direct answer, clarification, no direct but relevant answer, and not enough information, respectively.

in the data. As we can see, most user requests are "what" and "how" questions. There are also many user turns starting with words like "can", "tell", most of which are responses to agent clarification questions. The user utterances are fairly long-tailed as "MISC" shares a large portion for both treemaps. Instead of being mostly factoid questions, user requests are found to be fairly open-ended in our dataset.

Agent Response Strategy Different from previous work that mostly focuses on agent responses with either a direct answer or `no answer` to the current user request (Choi et al., 2018; Reddy et al., 2019; Adlakha et al., 2022), Table 3 shows the diversity of agent response strategies in INSCIT. When no direct or precise answer can be found, agents in INSCIT can respond to the user with a *relevant answer* (defined in § 4.1). If no direct or relevant answer is found, the agent can then respond with *no information*. As we can see from Table 3, 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 slightly shorter and are more likely to happen when more evidence passages are present.

Grounding Evidence Set In contrast to previous datasets that only allow a single string span or passage as the evidence for each agent turn, INSCIT contains cases where the grounding evidence passage sets are gathered from multiple paragraphs or even across documents. As shown in Table 2,

the average number of evidence passages per agent turn is 1.6. As expected, most *clarification* agent turns have multiple evidence passages (87% cases). Among *direct answer* or *relevant answer* cases, there are 30% of agent turns with multiple evidence passages, which potentially require some information summarization. We also observe that 90% of direct answer responses in INSCIT are longer than 15 words, while most previous datasets have agent turns with an average length shorter than that.

4.3.2 Analysis of Agent Initiatives

Fine-Grained Categorization To understand diverse agent initiatives at a more fine-grained level, we randomly sample and analyze 100 clarification and relevant answer responses respectively. As shown in the upper half of Table 4, we notice that in most cases, the agent raises a clarification when they either find a very long or too many answers (86%) or notice an ambiguous entity in the user request (13%). In 70% relevant answer cases (bottom half of Table 4), the agent would relax some constraint in the user request or provide evidence that no definite answer can be found. In the remaining cases (29%), they simply provide some relevant but side or partial information only. We also observe that in some rare cases (1%), the agent would point out some mistake (e.g., a false assumption) in the user request. We provide examples for such cases in Appendix A.2 to facilitate future investigations.

Clarification Occurrences We drill down to understand *when agents are more likely to raise clarification in a conversation*. We find that clarification questions are more frequently seen in the very beginning (example 2 in Table 4) than later in a conversation (18.8% vs. 11.5%). Furthermore, we notice that if a clarification is raised in the previous agent turn, the probability of the agent asking another clarification is 7.6% (Table 6), while the probability is 12.2% if the previous turn is non-clarification (example 1 in Table 4).

Response Strategy Selection Another interesting phenomenon we observe is that among examples where we obtain 2 agent references for which the validator mark as equally comprehensive, 23% of them take *different response strategies given the same dialogue context*. The potential reason is that the response strategy selection relies on what information (i.e., evidence) each agent finds. Due to the open-endedness of information-seeking queries in INSCIT and the large search space over Wikipedia,

Clarification	
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	
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 / partial information 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 4: Examples of clarification and no-answer-but-relevant-information agent responses. Prepending dots → omitted dialogue history. *Factual information from evidence passages is italicized in agent responses.*

it is common for the agent to land on different sets of evidence passages. The first example in Table 5 shows how different evidence triggers different agent response strategies. Additionally, even if two agents find the same evidence set, deciding whether it indicates an under-specified user request, a direct answer or only a relevant answer can be subjective, and we leave it to workers' own judgment. The second and third examples in Table 5 show cases where agents take different strategies to respond to the user when they find the same evidence. While the selection of response strategy can be subjective, Table 3 shows that more information pieces would more likely trigger the agent to refine the user request by raising clarification.

5 Experiment Setup

5.1 Systems

We build two systems for both 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, 2020; Adlakha et al., 2022). Here, the main function of the retriever is to gather a ranked set of 50 candidate evidence passages from the entire Wikipedia to facilitate passage selection and response generation for the later reader model. Different from prior work where the reader only predicts short answer strings, we need to adapt the reader to perform both evidence passage identification and response generation tasks based on the retrieved candidate set. We first describe the retrieval models used (§ 5.1.1),

and then introduce the two reader models that perform the two main tasks based on retrieval results (§ 5.1.2). We provide implementation and training details in Appendix C.

5.1.1 Retrieval Models

We experiment with two retrievers (BM25 and DPR) and choose the one with the best retrieval performance for readers to perform the two main tasks. **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 vector space. We finetune (FT) DPR on INSCIT. As the training set is relatively small in INSCIT, we also experiment with initializing it with a downloadable checkpoint[12] from pre-training (PT) DPR on a much larger open-domain conversational question answering dataset, TopioCQA (Adlakha et al., 2022).

5.1.2 Reader Models

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

---

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

Different Evidence; Different Response Strategies
Dialogue Context: . . .**<User>:** What are some of the health consequences of oats?
<Agent> (clarification): 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 answer):** 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 them themselves in any controversy but their film, V for Vendetta, 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. DIRECT, CLAR and REL indicate direct answer, clarification, and no direct but relevant information, respectively.

. . .**<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.

Fusion-in-Decoder (FiD) FiD is a generative reader model that can be easily adapted to generate different formats of task output. 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 it to our tasks, we prepend a passage identifier (ID) to each of the top-k retrieved passages[13] and concatenate the passage with the dialogue context to be encoded by FiD. To perform the two tasks, the decoder generates a sequence of evidence passage IDs (*passage identification*), followed by the final response (*response generation*). In order to incorporate previously used evidence passages, we simply add them into the top-k retrieved passages as the augmented input to FiD.[14]

DIALKI + FiD In contrast to generating outputs for both tasks in an end-to-end fashion as the first system, the second system 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 to select an answer string span from a set of passages. DIALKI predicts a passage score for each input passage (e.g., each top-k retrieved passage). To adapt it for our passage identification task, we simply keep evidence passages (up to 4)[15] with ranking scores higher than γ for *multiple passage prediction*. The hyperparameter γ is tuned on the dev set. We apply the same method to incorporate previously used evidence passages into DIALKI as in FiD.

Trivial Baselines We report performance of three simple baselines. 1) **Most Frequent**: We use the most frequent evidence passage and agent response seen in the training set as the prediction. 2) **Random Previous Turn**: We randomly select a previous turn in the dialogue context and use its evidence and response as the prediction. For first-turn instances, we use the most frequent passage and response as in "Most Frequent". 3) **Last Turn**: We use the most recent agent turn in the dialogue context as the prediction. Similarly, we use "Most Frequent" for first-turn examples.

Human We collect one additional annotation for each agent turn in the test set and evaluate it as the

---

[13]We follow (Adlakha et al., 2022) to set k=50.

[14]The number of passages encoded by FiD is still 50, and thus those retrieved at a lower rank would potentially be dropped after adding previous evidence passages.

[15]same upper limit during our data annotation.

human performance. 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 both automatic metrics and a human evaluation protocol for the passage Identification (PI) and response generation (RG) tasks in § 3, as well as a metric to assess the impact of the first stage passage retrieval on the PI task.

Passage Retrieval We follow previous work (Karpukhin et al., 2020; Adlakha et al., 2022) to report HIT@K scores for retrieval performance:

$$HIT@K = \mathbb{1}\Big[\big|\mathcal{R}_K \cap \mathcal{P}\big| > 0\Big]$$

where $\mathcal{R}_K$ denotes the top K retrieved passages and $\mathcal{P} = \bigcup_{i=1}^{m} \mathcal{P}^i$ denotes the union of all passages in the reference.

Passage Identification As discussed in § 3, we allow multiple evidence passages in INSCIT. We measure the model performance by comparing the set of predicted evidence passages $\bar{\mathcal{P}}$ to the set of reference passage sets $\{\mathcal{P}^1, \dots, \mathcal{P}^m\}$ where $\mathcal{P}^i$ denotes the ith reference passage set. In INSCIT, m equals 1 or 2. We use the maximum F1 score (PI-F1) between $\bar{\mathcal{P}}$ and each $\mathcal{P}^i$ as the final score.

Response Generation For a generated agent response $\bar{a}$, we calculate the maximum BLEU (Papineni et al., 2002)[16] and token-level F1 (RG-F1) scores against the reference responses $\{a^1, \dots, a^m\}$ for evaluation, following previous studies (Feng et al., 2020; Adlakha et al., 2022).

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

Therefore, we design a human evaluation protocol to evaluate the model performance on both tasks. (See the interface and details in Appendix B). 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

---

[16]We use SACREBLEU specifically (Post, 2018).

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 evaluation scores (Choi et al., 2018; Anantha et al., 2021; Adlakha et al., 2022), a few studies collect human ratings on dimensions like "coherence" and "informativeness" of each response (Gao et al., 2022; Feng et al., 2022). However, they do not require models to predict evidence and thus ignore the factual consistency between the response and the knowledge source.

We provide outputs for both tasks of two (or more) systems to a human judge and ask them to rate on a 4- or 5-point Likert scale[17] on the first 3 dimensions for each system output. Each evaluator ranks the system responses in the aspect of *response comprehensiveness* (ties are permitted). We have 3 raters for each agent turn and take the average rating scores or rank places on each dimension for each individual system. Since human evaluation can be time-consuming and costly, we run it on a sampled test subset with 50 conversations (290 examples in total) and encourage future studies to report on the same subset.

6 Experiment Results

6.1 Quantitative Results

Passage Retrieval We first report the performance for passage retrieval in Table 7. BM25 underperforms DPR models significantly. We observe that pretraining (PT) alone can already work better than simply finetuning (FT). This can be explained by the fact that the pretraining data (TopioCQA) is more than 30 times larger than INSCIT. Karpukhin et al. (2020) have a similar observation on retrieval for single-turn question answering tasks. Finally, we achieve the best retrieval results with both pretraining and finetuning DPR (PT+FT), and we use the retrieval output from this setting for reader models to do passage identification and response generation throughout the remaining experiments.

Automatic Evaluation (PI & RG) Table 8 shows the overall automatic evaluation results for our *main tasks*. The three simple baselines perform very poorly. DIALKI + FiD achieves much

---

[17]The 4-point scale is used only for coherence to discourage "neutral" ratings. We report all scores after normalizing to the 1-5 scale.

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

Table 7: *Passage retrieval* performance. PT and FT refer to pretraining on TopioCQA and finetuning on INSCIT.

Method	Dev			Test		
	PI-F1	BLEU	RG-F1	PI-F1	BLEU	RG-F1
Most Frequent	0.0	0.5	6.7	0.0	0.4	5.1
Rand Prev Turn	7.1	3.4	13.1	6.6	3.1	13.2
Last Turn	10.5	4.2	14.1	9.8	3.9	14.2
DPR (PT+FT) +						
FiD	17.1	8.8	21.6	17.5	9.6	22.2
- Prev Passages	16.2	9.8	22.5	17.8	9.3	22.3
DIALKI + FiD	**21.5**	**16.6**	**26.6**	**23.7**	**16.0**	**27.8**
- Prev Passages	17.8	15.3	25.3	20.2	15.2	26.0
Human	-	-	-	52.5	33.8	43.5

Table 8: Automatic evaluation for **main tasks**: *passage identification (PI)* and *response generation (RG)*.

Method	Evidence Utility	Factual Consistency	Coherence
FiD	2.35	2.52	3.76
DIALKI + FiD	**4.33**	**4.74**	**3.77**
Human	4.76	4.77	4.85

Table 9: Human evaluation scores on dimensions rated with Likert scales.

		Win	Tie	Lose
DIALKI+FiD vs FiD	*Evidence Utility*	**212 (73%)**	51 (18%)	27 (9%)
	Factual Consistency	**219 (76%)**	52 (18%)	19 (7%)
	Coherence	**100 (34%)**	98 (34%)	92 (32%)
	Comprehensiveness	**191 (66%)**	28 (10%)	71 (24%)
DIALKI+FiD vs Human	*Evidence Utility*	35 (12%)	160 (55%)	**95 (33%)**
	Factual Consistency	39 (13%)	198 (68%)	**53 (18%)**
	Coherence	12 (4%)	98 (34%)	**180 (62%)**
	Comprehensiveness	46 (16%)	23 (8%)	**221 (76%)**

Table 10: Human evaluation on system comparison, where win/lose refers to DIALKI+FiD.

better performance than FiD in all evaluation metrics. The reason that the pipelined approach beats the end-to-end FiD model may be due to the small training set size, i.e. not enough data for training an effective model to process all inputs and predict various outputs simultaneously. Both systems still substantially underperform humans. We also observe that incorporating previously used evidence passages leads to better performance for DIALKI + FiD, while not necessarily for FiD, indicating that some form of information summary performed by DIALKI is necessary to avoid the model being overwhelmed by the large input information. The reason for *imperfect human performance* on passage identification is two-fold. As discussed in § 4.3, 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 used for reporting the human performance do not go through the validation process, which may lead to annotation mistakes or being less comprehensive than the reference responses.

Although Table 7 shows that TopioCQA can be leveraged to improve retrieval performance, it is not straightforward to leverage it for our two main tasks. It does not come with the passage identification task, and only has short answers or `no answer` as their agent responses. We observe poor zero-shot response generation performance by training FiD on TopioCQA and evaluating on INSCIT, reported in Appendix D. Therefore, we

leave the question of how to leverage other existing datasets to improve main task performance on INSCIT for future exploration.

Human Evaluation (PI & RG) Table 9 and 10 present our human evaluation results.[18] We observe that overall humans substantially outperform DIALKI+FiD. However, the difference is much smaller in factual consistency. When compared with FiD, DIALKI+FiD has better evidence utility, factual consistency and response comprehensiveness. However, response coherence is similar for both systems.

6.2 Analysis

Seen / Unseen Topic Categories As explained in § 4, there is a topic distribution shift from training to test set. We carry out a breakdown evaluation here. Specifically, we divide test conversations into seen (food & drink, hobby and top-25) and unseen (historical events, geography) topic categories during training. Interestingly, based on the automatic scores shown in Figure 4, both models generalize well on conversations from unseen topic categories. FiD has a closer trend with Human, which has slightly better performance on unseen conversations.

Performance by Reference Response Strategy Table 11 shows the model performance breakdown

---

[18]Our inter-rater agreements, reported as Krippendorf's alpha, are 0.66, 0.64, 0.42 and 0.37 for evidence utility, factual consistency, coherence and comprehensiveness respectively, which can be interpreted as good or moderate agreements.
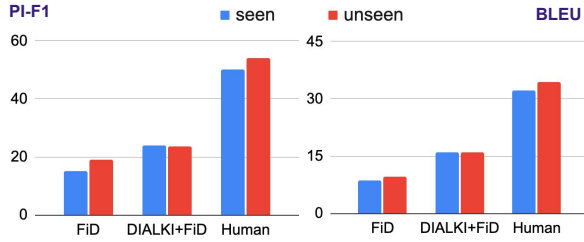
Figure 4: PI-F1 and BLEU scores on test conversations with seen and unseen topic categories. RG-F1 has the same trend as BLEU.

	PI-F1	BLEU	RG-F1
FiD			
direct answer	21.1 (-65%)	10.4 (-73%)	32.1 (-33.1%)
clarification	9.8 (-73%)	3.3 (-74%)	15.7 (-47.1%)
relevant answer	6.7 (-79%)	3.8 (-78%)	18 (-38.8%)
DIALKI+FiD			
direct answer	28.4 (-53%)	18.6.1 (-52%)	24.5 (-48.9%)
clarification	10.8 (-70%)	5.3 (-58%)	13.9 (-53.2%)
relevant answer	15.8 (-50%)	6.2 (-64%)	15.6 (-46.9%)

Table 11: PI (passage identification) F1 and BLEU scores by reference response strategy. Percentage values indicate the performance gap with humans.

by reference response strategy. Specifically, we assign test examples into a response strategy category (i.e., direct answer, clarification or relevant answer) if all of their reference responses are of that response strategy. We observe a bigger gap compared to humans for both FiD and DIALKI+FiD systems for "clarification" and "relevant answer" categories, suggesting that existing state-of-the-art models still struggles at distilling relevant information and communicate it effectively via generation.

Evidence Utility by # Evidence Passages Although DIALKI+FiD achieves relatively high evidence utility score in the human evaluation, we show that it is very sensitive to the number of evidence passages predicted. Note that the model (as well as Human) only predicts up to 4 passages, which is the same upper limit during our data annotation. Table 12 illustrates that its generated responses fail to include information from all predicted evidence passages when there are multiple of them, while the same issue does not hold for humans. Our further analysis shows that DIALKI+FiD rarely constructs clarification questions or summarizes information from difference passages as humans do, which indicates a remaining modeling challenge to fuse and present information from different evidence pieces.

# Predicted Evidence Passages	1	2	3	4
DIALKI + FiD	4.81	3.41	2.54	2.78
Human	4.81	4.52	4.71	4.91

Table 12: Human evaluation scores on *evidence utility* by the number of predicted evidence passages.

7 Conclusion & Future Work

To conclude, we introduce INSCIT, a new open-domain information-seeking conversational dataset grounded in Wikipedia, with mixed-initiative user-agent interactions. We formulate two tasks with INSCIT as well as a new human evaluation protocol to assess the model performance, and present the results of two strong baselines.

While both models outperform trivial baselines, they still significantly underperform humans in our two tasks based on both automatic and human evaluation. Such performance gap becomes even bigger for scenarios that require different agent strategies other than providing a direct answer. As the evidence passage identification task directly impacts response generation by providing grounding knowledge, it is especially important to build effective models for predicting relevant passages in the future. A key challenge that remains in response generation is how to fuse and present (comprehensive) information from multiple evidence passages by leveraging different response strategies. Future efforts are also required to improve the response coherence.

Given the small size of our training data, it would be an interesting future direction to address the above issues by applying transfer learning from existing information-seeking conversation or question answering resources.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. In *Transactions of*

the Association for Computational Linguistics, volume 10, pages 468–483.

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.

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.

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.

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.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau 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.

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.

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.

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.

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.

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.

M. Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *AKBC*.

Alexandru Balan Gustavo Penha and Claudia Hauff. 2019. Introducing mantis: a novel multi-domain information seeking dialogues dataset.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering.

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.

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.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.

Natalia Konstantinova and Constantin Orasan. 2013. Interactive question answering.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *EMNLP*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

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.

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.

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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.

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.

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.

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, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

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.

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.

A Additional Details of INSCIT

A.1 Annotation

Figure 6 and 7 show examples of user and agent annotation task interfaces. We provide more annotation details below.

Seed Document Selection We use PetScan[19] to sample Wikipedia pages at various tree depths under each of the 4 categories—"food and drink", "hobby", "historical events" and "geography". We use different Wikipedia category keywords for each of them to find corresponding articles via PetScan, as shown in Table 13. We filter out overlapped pages between these categories, and pages with fewer than 150 outgoing links or 3000 content words.

We sample articles from depth 1-3 for "food and drink" and "hobby". Since there are very few articles under depth 1 for "historical events" and "geography", we only sample articles from depth 2-3. For example, under the food and drink category, general pages like "Wine" and "Rice" usually appear at depth level 1. More specific topics like "Hong Kong Cuisine" and "Dairy Farming" are at level 2. Less well-known or related topics like "Callinectes sapidus" and "Chipotle" can be found at level 3.

Topic Category	Keywords
Food & Drink	Foods, Cuisine, Drinks
Hobby	Hobbies
Historical Events	Disasters by continent, Historical events by country, Wars
Geography	Nature reserves, Bodies of water, Landforms

Table 13: Keywords used for collecting documents under each topic category from PetScan.

Additional Annotation Instructions After the initial user turns (Figure 2 (d)), each user worker is instructed to read the dialogue context and respond to the last agent turn by providing a clarification, asking a new or follow-up question, or raising concerns (e.g., if there is a misunderstanding of the previous request), given the same seed article. Each agent worker is asked to limit their responses to be at most 2 sentences.

We define "response comprehensiveness" for the validation task in the instruction, that an agent response is considered as more comprehensive if it

contains more information relevant to the user request. Additionally, we also specifically ask validators to focus on the "information scope". For example, if there exist multiple answers (answer A and B) to the user query, a direct answer containing both A and B or a clarification like "Do you want to know about answer A or B?" should be considered as equally comprehensive while both of them are more comprehensive than a direct answer containing A only.

Qualification Before Annotation We restrict our annotation tasks to workers in English-speaking countries and with more than 5000 HITs and at least a 98% acceptance rate. Depending on each worker's annotation speed, most of our workers are paid with an hourly rate of 15-20 USD. To recruit agent annotators, we manually review more than 150 submissions of a qualification task and select 24 highly qualified agent workers. Similarly, we select 35 qualified users and 10 validators before the annotation. We launch a different user qualification task, and since the user annotation is a much easier task, we only filter out 7 spamming workers during the qualification. Most of our select agent workers also participate in the user qualification task and all of them pass the qualification. We instruct agent workers to try not to annotate for their own user request turns, and we find only fewer than 5% user-agent turns are annotated by a same worker. The 10 validators are selected from our qualified agents with a different qualification task. We instruct our validators to avoid validating their own annotations unless they forget, and we find only fewer than 5% agent annotations are self-validated.

Quality Control During Annotation Each user worker can only get bonus and retain their qualification in our annotation if their following agent decides to continue their conversation, as discussed in Section 4. Similarly, our agents need to pass validation in order to get bonus and remain in our annotation. Besides verdicting whether each agent annotation is valid or not, our validators are also instructed to fix typos in each agent response. At the same time, we allow agent workers to check their mistakes flagged by validators, so that we can improve in later annotation tasks. In our pilot study with more than 500 examples, our agent validators reach over 90% agreement in deciding whether an annotation is valid or not.

We also assign additional bonus if agent workers

...<User>: Why did Wilson *want to get elected* as the Prime Minister of the United Kingdom the second time?	
<Agent>: I wasn't able to find that, but I did find that when Wilson entered office for the second time, he had privately admitted that he had *lost his enthusiasm for the role* ...	

Table 14: An example showing that the agent finds evidence that is contradicting to an assumption made in the user query.

generate responses that are marked as equally or more comprehensive than the other workers. We collect training data first before collecting the development and test sets. This bonus structure was added and adjusted during the training data collection based on constant monitoring and feedback from workers during the annotation process. As a result, we observe variations in the average agent response length among different data subsets.

Each validator rate the response comprehensiveness comparison between 2 annotations at a 1-5 scale. After analysis, we notice that in most cases where the validator gives a score of 2-4, the 2 annotations actually find different sets of evidence passages, or have very similar or slightly different comprehensiveness. Therefore, we only consider scores of 1 and 5 as a sufficient indicator that one annotation is more comprehensive than the other.

A.2 Data Analysis

User Turns   In Table 2, we observe longer user turns than reported in previous datasets (Reddy et al., 2019; Choi et al., 2018; Anantha et al., 2021; Adlakha et al., 2022). This can be partly because that in our instruction, we encourage our user workers to ask interesting questions although we leave the definition of "interestingness" to users. Another reason is that we observe in some cases, user turns contain "reaction" sentences to the previous agent turns before raising their request. For example, the user may say "Oh, sorry I should've been more specific." after the agent asks a clarification.

Other Relevant Answer Cases   Besides the two categories of relevant answer agent responses discussed in Table 4, we find about 1% cases where the agent would identify some issue (e.g. a false assumption) in the user request by providing relevant information as evidence. Table 14 shows such an example. Although we do not find such situations to be common in INSCIT, it would be an interesting phenomenon for future studies.
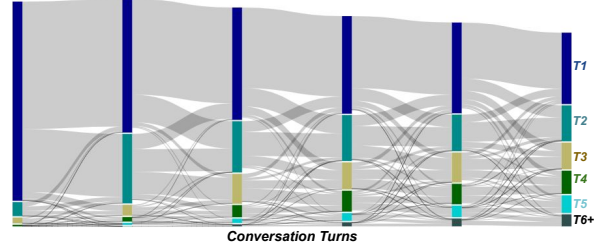

Figure 5: Flow diagram of topic changes in up to 6 turns in INSCIT conversations. "T" stands for "Topic" and the topic indices indicate the order of their first occurrences in the conversation.

A.2.1 Natural Conversation Topic Changes

Following Adlakha et al. (2022), we consider the topic(s) of each conversational turn as the Wikipedia article(s) in which the agent finds evidence passages.[20] Figure 5 shows the flow of topic switches for up to 6 turns in each conversation. For each conversational turn, every pair of (previous topic, current topic) leads to a flux in the diagram. We can see that INSCIT conversations contain frequent topic changes, and Table 15 shows an example illustrating the naturalness of such topic changes.

Topic	Dialogue Turns
Intellectual property	**<U>:** How is intellectual property protected such as photographs? **<A>:** Are you interested in intellectual property law in the U.S., U.K., Germany, or ... ?
Copyright law of the U.S.	**<U>:** I'm interested in the U.S. **<A>:** In the US there are copyright laws to protect original works ...
Copyright law of Canada	**<U>:** How about in Canada? **<A>:** The copyright law of Canada governs the legally enforceable rights ...
Internet censorship in the United Kingdom	**<U>:** How are the intellectual property rights on photographs in the U.K. deal with terrorism? **<A>:** Copyright law of the United Kingdom, contains regulations against incitement to terrorism.
Terrorism Act 2000	**<U>:** What is the anti-terrorism law? **<A>:** The Terrorism Act 2000 (c.11) is the first of a number of general Terrorism Acts ...

Table 15: Natural topic changes in a conversation.

B   Additional Details of Human Evaluation

Figure 8 and 9 show interfaces of our human evaluation. As mentioned in Section 5, *response comprehensiveness* is defined as: how much information,

---

[20]Note that we allow multiple evidence passages which can potentially be found in multiple articles.

that is both relevant to and factually consistent with the predicted evidence, is contained in the response. Additionally, we also specifically ask evaluators to focus on the "information scope". For example, if there exist multiple answers (answer A and B) to the user query, a direct answer containing both A and B or a clarification like "Do you want to know about answer A or B?" should be considered as equally comprehensive while both of them are more comprehensive than a direct answer containing A only.

C Training and Implementation Details

C.1 Retrieval Models

BM25 We use Pyserini (Yang et al., 2017) with the parameters $k_1 = 0.9$ and $b = 0.4$, following Karpukhin et al. (2020).

DPR We use the released DPR training scripts from Adlakha et al. (2022) for finetuning. We set the batch size to be 48. The learning rate is set to be $1e-6$ and $1e-5$ respectively with or without pretraining on TopioCQA respectively. All other parameters are kept the same. We follow Feng et al. (2021) to create a hard negative for each example by sampling a retrieved passage from BM25, with one of the gold evidence passage as the query. We keep all other parameters unchanged in the original codebase. Each training process is run on 4 A100 or A40 GPUs.

C.2 Reader Models

FiD We also follow the code released by Adlakha et al. (2022) for training FiD. We take the top 50 retrieved passages and encode each of them concatenated with the dialogue context by a separator token `[SEP]`. In order to adapt FiD to our two tasks, in front of each passage, we prepend a passage id (pid), formatted as `article title:passage position`, where the passage position refers to the order of the passage appearing in the article. The model is trained to decode a sequence of evidence passage ids and the final response in the format of

```
pid1 | pid2 ... answer:  response
```

During inference, we parse the decoded sequence to get all predicted passage ids and the final response. We remove passage ids that are duplicates or do not exist in the corpus.

For each training example, we randomly select one reference agent turn. We also experiment with always selecting the reference agent turn marked as more comprehensive than the other reference[21], but it gives similar performance.

The maximum encoder input length is 384 and the decoder sequence length is 100. We train for 7 epochs and the number of warm-up steps is 50. We use 12 as the batch size. All the other parameters in the original code are unchanged. Each training process takes about 2-3 hours with 4 A40 GPUS. We select the best model based on the PI-F1 score on the dev set.

DIALKI+FiD We use the public code released by Wu et al. (2021) for training DIALKI. We also take 50 passages and encode each of them concatenated with the dialogue context by a separator token `[SEP]`. As DIALKI is designed to select only one positive passage, to create each training example, we only include one randomly sampled gold reference passage in the input and 49 negative passages from the top retrieved passages. Then, the best model is selected based on the single passage selection accuracy. During inference, we simply feed top 50 retrieved passages into the model. In order to perform multiple passage identification, we simply keep evidence passages (up to 4)[22] with ranking scores higher than γ. The hyperparameter γ is tuned on the dev set.

To train FiD in this two-step system, we create training examples by taking a dialogue context and each set of its reference evidence passages as the input. As a result, the maximum number of input passages is 4. FiD is trained to decode the agent response only. The best model is selected by the RG-F1 score on dev set. During inference, we feed FiD with the predicted passages from DIALKI.

The maximum encoder sequence length is 384 for both DIALKI and FiD. We keep original parameters in DIALKI unchanged. Parameters in FiD is the same as the first reader model, except that the number of input passages is 4 in this setting. Each training process is run on 4 A40 and 2 RTX6000 GPUs for DIALKI and FiD respectively.

For each experiment, we observe similar performance or training curves for 2-3 runs and report numbers on a single run.

D Additional Details of Experiments

Metric Correlations We observe similar trends between the automatic and human evaluation re-

---

[21]Randomly select one if there is a tie.
[22]same upper limit during our data annotation.

Human Eval Dimension	Automatic Metric	Pearson	Spearman
Evidence Passage Utility	PI-F1	0.08	0.07
	BLEU	0.06	0.07
	RG-F1	0.07	0.07
Factual Consistency	PI-F1	0.05	0.04
	BLEU	0.00	0.00
	RG-F1	-0.01	-0.01
Response Coherence	PI-F1	0.20	0.19
	BLEU	0.22	0.28
	RG-F1	**0.30**	**0.33**
Response Comprehensiveness	PI-F1	0.26	0.27
	BLEU	0.26	0.26
	RG-F1	**0.27**	**0.28**

Table 16: Pearson and Spearman correlation coefficients between each human evaluation dimension and each automatic metric. Both "evidence passage utility" and "factual consistency" human ratings have poor correlation with all three automatic metrics. Response F1 scores have the best correlation with response "coherence" and "comprehensiveness". Except for coefficients calculated for "factual consistency", all numbers have a p-value less than 0.05.

sults in terms of system comparison. However, as shown in Table 16, we calculate the Pearson and Spearman correlation coefficients between human-rated response coherence and comprehensiveness and automatic scores to be moderate and weak. Among all three automatic metrics, response F1 scores have the highest correlation with the two human evaluation dimensions. Evidence utility and factual consistency has poor correlation with all three automatic scores. This indicates that existing grounded generation metrics are lacking in truly reflecting the human judgements.

Zero-Shot Performance As DPR retriever followed by FiD is the state-of-the-art model on TopioCQA, we use the publicly released checkpoint[23] trained on TopioCQA to do inference on INSCIT. Since TopioCQA only considers the task of response generation, we skip the evaluation for passage identification. Table 17 shows that the model generalizes poorly from TopioCQA to INSCIT, which indicates that agent responses in INSCIT have a very different distribution with TopioCQA.

Fixed Passage Setting We experiment with a setting where the pool of evidence passage candidates is much smaller than the open-domain setting and fixed throughout each conversation. In specific, we gather passages from all Wikipedia article sections that include at least one evidence passage

Model	TopioCQA		INSCIT	
	R-EM	R-F1	BLEU	R-F1
DPR Retriever + FiD (Dev)	33.0	55.3	4.1	16.9
DPR Retriever + FiD (Test)	33.4	55.8	4.0	17.1

Table 17: Results of DPR Retriever + FiD, trained with TopioCQA, on TopioCQA and INSCIT (zero-shot). Response exact match (EM) is a metric used in TopioCQA.

Method	Dev			Test		
	KI-F1	BLEU	R-F1	KI-F1	BLEU	R-F1
FiD	**43.1**	**25.6**	**35.5**	**45.5**	**24.7**	**35.6**
DIALKI + FiD	41.7	24.5	33.3	44.7	24.6	34.7

Table 18: Fixed passage setting results.

used in a conversation as the candidate passages for that conversation. Table 18 shows that both systems achieve much higher scores than in the open-domain setting. Interestingly, in contrast to the open-domain setting, FiD slightly outperforms DIALKI + FiD.

---

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

Read the following ongoing dialogue (some are empty) and write the next user response following our instructions.

Short Paragraph from a Wikipedia Article:

Orienteering is a group of sports that require navigational skills using a map and compass to navigate from point to point in diverse and usually unfamiliar terrain whilst moving at speed. Participants are given a topographical map, usually a specially prepared orienteering map, which they use to find control points. Originally a training exercise in land navigation for military officers, orienteering has developed many variations. Among these, the oldest and the most popular is foot orienteering. For the purposes of this article, foot orienteering serves as a point of departure for discussion of all other variations, but almost any sport that involves racing against a clock and requires navigation with a map is a type of orienteering.

Ongoing Dialogue:

User: What is the oldest form of orienteering?
Agent: The oldest and the most popular is foot orienteering.
User: What is foot orienteering?
Agent: Foot orienteering is the oldest formal orienteering sport and is usually a timed race in which participants start at staggered intervals, are individually timed, and are expected to perform all navigation on their own. The control points are shown on the orienteering map and must be visited in the specified order and standings are determined first by successful completion of the course, then by shortest time on course.

Type the next user response:

What is an example of foot orienteering that's still being played today?

Feedback (Optional)

Wikipedia Article Content Outline:

Contents

- 1 History
- 2 Variations
- 3 Governing bodies
    - 3.1 International
    - 3.2 National
    - 3.3 Regional
    - 3.4 Local
    - 3.5 Related sports
- 4 Competition and results
    - 4.1 Basics
    - 4.2 Map
    - 4.3 Courses
    - 4.4 Controls and control description sheet
    - 4.5 Control card and punching
    - 4.6 Results
    - 4.7 Safety
    - 4.8 Personal clothing
    - 4.9 Personal equipment
- 5 Competition types
    - 5.1 Long
    - 5.2 Middle
    - 5.3 Relay
    - 5.4 Score
    - 5.5 Sprint
    - 5.6 Ultrasprint
    - 5.7 Night
    - 5.8 String
    - 5.9 Precision
- 6 The Olympics
- 7 World Orienteering Championships

Figure 6: User annotation interface.

STEP 1: Read the following ongoing dialogue and try to understand and respond to the current user request.

Ongoing Dialogue:

User: Did Money Heist win a Golden Globe award?

Do you think the request is:

☐ chit-chat or an invalid / uninterpretable request
☑ valid and clear for you to start the answer search

STEP 2: Write a Query for Answer Search in Wikipedia

| Money Heist | | Search |

< Go back to search results

☆A   ↓ ☆ ✎

Money Heist (Spanish: *La casa de papel*, *The House of Paper*) is a Spanish heist crime drama television series created by Álex Pina. Netflix acquired global streaming rights in late 2017.[1]

In the first four programming seasons, the Spanish TV series garnered a total of 30 nominations by winning 26 times.

In 2017 the series received two Award nominations (one victory) at the Fotogramas de Plata, other two (no victories) at the FesTVal and got an Award win at the Premios Iris. In total, therefore, five nominations and two wins.

Money Heist
awards and nominations

LA CASA DE PAPEL

Awards and nominations

Award	Won	Nom
Unión de Actores	3	10

STEP 3: After your answer search, which of the following best describes your situtaion:

☐ No answer nor any relevant info found
☐ Long / multiple answers found and / or clarification needed
☑ Short answer found and can be directly provided
☐ No direct answer but relevant info / partial answer found

STEP 4: Copy and paste the FIRST SENTENCE of each evidence paragraph for your decision above. The Wikipedia page title of each paragraph will be entered automatically. Change it if you think it is incorrect.

Paragraph First Sentence:

In 2017 the series received two Award nominations (one victory) at the Fotogramas de Plata

Wiki Page Title:

List of awards and nominations received by

Delete Evidence Paragraph & Page Title

Add Evidence Paragraph & Page Title

FINAL STEP: Type the next agent response with the answer(s) (limit to 1-2 sentences only).

Money Heist was nominated for numerous awards in 2017, 2018, 2019 and 2020, but it was never nominated for, nor awarded, a Golden Globe.

Feedback (Optional)

Figure 7: Agent annotation interface.

Figure 8: Human evaluation interface 1: inputs from 3 systems for evaluation.

How many evidence paragraphs are *properly used* in each system response?

- **System 1:** ◯ All ◯ Most ◯ Half ◯ Very Few ◯ None ◯ No Evidence Predicted
- **System 2:** ◯ All ◯ Most ◯ Half ◯ Very Few ◯ None ◯ No Evidence Predicted
- **System 3:** ◯ All ◯ Most ◯ Half ◯ Very Few ◯ None ◯ No Evidence Predicted

How much factual information in the system response is *grounded* in AND *consistent* with its evidence paragraphs or the dialogue context, ignoring commonsense knowledge?

- **System 1:** ◯ All ◯ Most ◯ Some ◯ Very Little ◯ None ◯ No Factual Info In the Response
- **System 2:** ◯ All ◯ Most ◯ Some ◯ Very Little ◯ None ◯ No Factual Info In the Response
- **System 3:** ◯ All ◯ Most ◯ Some ◯ Very Little ◯ None ◯ No Factual Info In the Response

How *coherent* is the response with the given dialogue context?

- **System 1:** ◯ Very ◯ Somewhat ◯ Somewhat not ◯ Very not
- **System 2:** ◯ Very ◯ Somewhat ◯ Somewhat not ◯ Very not
- **System 3:** ◯ Very ◯ Somewhat ◯ Somewhat not ◯ Very not

How *fluent* is the response?

- **System 1:** ◯ Very ◯ Somewhat ◯ Somewhat not ◯ Very not
- **System 2:** ◯ Very ◯ Somewhat ◯ Somewhat not ◯ Very not
- **System 3:** ◯ Very ◯ Somewhat ◯ Somewhat not ◯ Very not

Which system response is the *most comprehensive* one?

☐ System 1   ☐ System 2   ☐ System 3

Which system response is the *least comprehensive* one?

☐ System 1   ☐ System 2   ☐ System 3

Figure 9: Human evaluation interface 2: rating dimensions. Note that we also have a question for evaluating "fluency", which we do not include in the human evaluation results. This question is mostly created to help human raters to decouple "coherence" to the dialogue context and the language "fluency". Both systems in our experiments turn out to have similar fluency as humans.