CoSRec: A Joint Conversational Search and Recommendation Dataset

Marco Alessio*
ISTI-CNR
Pisa, Italy
University of Pisa
Pisa, Italy
marco.alessio@isti.cnr.it

Simone Merlo*
University of Padua
Padua, Italy
simone.merlo@phd.unipd.it

Tommaso Di Noia
Politecnico di Bari
Bari, Italy
tommaso.dinoia@poliba.it

Guglielmo Faggioli
University of Padua
Padua, Italy
guglielmo.faggioli@unipd.it

Marco Ferrante
University of Padua
Padua, Italy
ferrante@math.unipd.it

Nicola Ferro
University of Padua
Padua, Italy
ferro@dei.unipd.it

Cristina Ioana Muntean
ISTI-CNR
Pisa, Italy
cristina.muntean@isti.cnr.it

Franco Maria Nardini
ISTI-CNR
Pisa, Italy
francomaria.nardini@isti.cnr.it

Fedelucio Narducci
Politecnico di Bari
Bari, Italy
fedelucio.narducci@poliba.it

Raffaele Perego
ISTI-CNR
Pisa, Italy
raffaele.perego@isti.cnr.it

Giuseppe Santucci
Sapienza University of Rome
Rome, Italy
santucci@dis.uniroma1.it

Nicola Viterbo
Politecnico di Bari
Bari, Italy
n.viterbo@studenti.poliba.it

Abstract

Conversational Information Access systems have experienced widespread diffusion thanks to the natural and effortless interactions they enable with the user. In particular, they represent an effective interaction interface for conversational search (CS) and conversational recommendation (CR) scenarios. Despite their commonalities, CR and CS systems are often devised, developed, and evaluated as isolated components. Integrating these two elements would allow for handling complex information access scenarios, such as exploring unfamiliar recommended product aspects, enabling richer dialogues, and improving user satisfaction. As of today, the scarce availability of integrated datasets — focused exclusively on either of the tasks — limits the possibilities for evaluating by-design integrated CS and CR systems. To address this gap, we propose CoSRec[1], the first dataset for joint Conversational Search and Recommendation (CSR) evaluation. The CoSRec test set includes 20 high-quality conversations, with human-made annotations for the quality of conversations, and manually crafted relevance judgments for products and documents. Additionally, we provide supplementary training data comprising partially annotated dialogues and raw conversations to support diverse learning paradigms. CoSRec is the first resource to model CR and CS tasks in a unified framework, enabling the training and evaluation of systems that must shift between answering queries and making suggestions dynamically.

*The authors contributed equally to this work.
[1]The dataset and code are available at: https://github.com/CAMEO-22/CoSRec

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730319

CCS Concepts

• **Information systems → Test collections**.

Keywords

Conversational Search, Conversational Recommendation, Joint Information Retrieval and Recommendation

ACM Reference Format:
Marco Alessio, Simone Merlo, Tommaso Di Noia, Guglielmo Faggioli, Marco Ferrante, Nicola Ferro, Cristina Ioana Muntean, Franco Maria Nardini, Fedelucio Narducci, Raffaele Perego, Giuseppe Santucci, and Nicola Viterbo. 2025. CoSRec: A Joint Conversational Search and Recommendation Dataset. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3726302.3730319

1 Introduction

Conversational Agents have revolutionized information access by enabling natural interactions, making it easier for diverse user groups, including children, the elderly, and individuals with visual impairments, to retrieve information. The widespread adoption of virtual assistants like Siri and Alexa, as well as chatbots such as ChatGPT, underscores the growing public interest in these systems. However, conversational information access introduces unique challenges compared to traditional media. These systems must hold the conversational state, interpret complex natural language constructs (*e.g.,* co-references, anaphoras, and ellipses), and adapt dynamically to user inputs. Information Retrieval (IR) and Recommender Systems (RS) are among the information access systems that benefit most from conversational interfaces. Conversational Search (CS) systems assist users in refining their information needs through multi-turn dialogues, while Conversational Recommendation (CR) systems guide users in exploring a catalogue of items to identify optimal recommendations. Despite their distinct objectives, CS and CR systems share significant commonalities, as both rely on iterative, multi-turn interactions to progressively refine user needs [19].

The development of Conversational Search and Recommendation (CSR) systems, which integrate both search and recommendation functionalities, could offer notable advantages. From a user perspective, such integration aligns naturally with real-world behaviour. For instance, a user may begin with a recommendation intent but later realize the need for additional information to refine their query. Conversely, a user starting with an informational need might eventually seek personalized recommendations. This seamless interplay between search and recommendation is already evident in commercial search engines like Google and Bing, which routinely incorporate recommendations into their result pages. Extending this integration to conversational systems would provide a more cohesive user experience. From a technical standpoint, recent research in joint IR and RS [46, 49, 57], though not yet conversational, has demonstrated the benefits of modeling these tasks together. For example, shared representation spaces have been shown to enhance system effectiveness by capturing the interplay between search and recommendation tasks. These findings suggest that integrating CS and CR into a unified conversational framework could yield similar improvements, paving the way for more effective and user-centric systems [55, 56, 59].

Traditionally, CS and CR have been treated as independent modules within information access tools, with little to no interaction or shared knowledge between them. This approach has hindered the development of joint search and recommendation systems, particularly in the context of conversational user-system interactions. One major obstacle is the scarcity of publicly available resources for training and evaluating joint CSR systems. While rich datasets exist for individual tasks, *e.g.,* the TREC CAsT collections [16–18, 44] for search and REDIAL [32] for recommendation, there is a notable absence of datasets tailored for joint scenarios. Moreover, the field lacks a well-defined formalization of the ideal testbed for evaluating joint CSR systems. This includes the structure and characteristics that an ideal dataset should possess to effectively assess the performance of such systems. Without a standardized framework, it becomes challenging to compare different approaches or measure progress in this emerging area. Addressing these gaps in resources and evaluation methodologies is crucial for advancing the development of integrated CSR systems and unlocking their full potential.

To facilitate the development of CSR systems, we introduce and release CoSRec, the first large-scale dataset explicitly designed for joint CSR tasks. CoSRec comprises approximately 9,000 user-system conversations generated by a Large Language Model (LLM) in the product search and recommendation domain. These conversations encompass a variety of interactions, including pure search, pure recommendation, and mixed search-and-recommendation utterances. As a result, a CSR system tested on CoSRec must accurately interpret the user's intent in each utterance and respond appropriately, taking into account the context of previous interactions. To ensure the quality of the dataset, a sample of approximately 3% of the conversations has been manually annotated to identify user intents and assess overall quality. Additionally, for 20 high-quality conversations, we provide utterance-level human-generated relevance judgments for items or documents, depending on the intent of the utterance. These annotations enable precise and effective evaluation of joint CSR systems. A key feature of CoSRec is its agnosticism toward underlying systems and evaluation paradigms. Traditional evaluation methods for CS systems focus on ranking documents based on information needs, while CR systems are evaluated based on the appropriateness of recommended items. However, there is no established standard for evaluating conversational systems, particularly joint CSR systems. To address this, CoSRec includes separate ground truths for search and recommendation tasks, allowing researchers to apply diverse evaluation paradigms and methodologies.

Our contributions can be summarized as follows:

- *Formalization of an Ideal CSR Collection Structure*: we define the ideal structure for a joint CSR collection, providing a framework to evaluate the performance of integrated CSR systems effectively.
- *Release of* `CoSRec-Raw`: we introduce `CoSRec-Raw`, a dataset comprising approximately 9,000 automatically generated conversations for joint search and recommendation tasks. Alongside the dataset, we provide a toolkit to generate additional conversations, enabling further research and scalability.
- *Release of* `CoSRec-Crowd`: we present `CoSRec-Crowd`, a subset of over 290 conversations manually annotated for quality. Each utterance in these conversations is labeled with its intent (search, recommendation, or joint search and recommendation), offering valuable insights for intent recognition and system evaluation.
- *Release of* `CoSRec-Curated`: we provide `CoSRec-Curated`, a high-quality subset of 20 deeply annotated conversations. For each utterance, we include manual (personalized) annotations identifying relevant passages or items, enabling precise and granular evaluation of CSR systems.

The remainder of the paper is organized as follows. Section 2 surveys the current state-of-the-art for the integration of search and recommendation tasks, and the main CS and CR datasets available. Section 3 describes the characteristics that the ideal CSR collection should have for testing properly a CSR system. Section 4 introduces the proposed dataset, along with an overview of its main elements. Section 5 outlines the source data from which CoSRec is built, the methodology employed to create the dataset, and the methodology used to generate all manual annotations. Section 6 analyzes the current limitations of our dataset. Finally, Section 7 draws some conclusions and outlines future work that will be built upon CoSRec.

2 Related Work

The strong relationship between search and recommendation has been a topic of discussion since the 1990s [8]. Recently, systems integrating both IR and RS functionalities have been developed [46, 49, 55–57, 59], demonstrating that jointly addressing search and recommendation tasks can lead to significant improvements. Similarly, in the conversational domain, the inherent similarities between CS and CR can be leveraged to build joint CSR systems, offering users seamless access to both search and recommendation capabilities. The potential benefits and configurations of such conversational systems have been explored by Di Noia et al. [19]. However, a major obstacle to advancing joint CSR systems is the lack of publicly available datasets specifically designed for this purpose. This gap in resources hinders the development, evaluation, and benchmarking of integrated CSR systems, limiting progress in this promising area of research.

The recent gain in popularity of CS and CR systems led to the creation of many conversational datasets. For CS, the TREC [50] conference represents a major dataset resource since it held many CS-related tracks: TREC CAsT [16–18, 44] and TREC iKAT [2]. In particular, the first three editions of TREC CAsT focused on context understanding and on the ability of systems to identify user needs in a conversational setting. Instead, the fourth edition of TREC CAsT also considered more realistic environments, including a mixed initiatives sub-task. Moreover, TREC iKAT started to take into account the personalization. Other CS-related datasets include QuAC [14], CANARD [21] and QReCC [5] which are mainly concerned with contextual Question Answering (QA). However, none of the mentioned CS datasets takes into account the (personalized) recommendation task. Indeed, even if it may seem that some search queries represent recommendation requests (*e.g.,* "what can I do in Rome?"), CS collections are usually based on open-world collections (crawled from the web) and not on closed-world catalogues as typical of CR and recommendation datasets [25, 32, 60].

For CR, datasets are usually created on top of existing recommendation datasets or online resources. These include the MovieLens dataset[2], the Amazon Reviews dataset [26], IMDB[3] and many others[4]. Dodge et al. [20] first proposed a dataset, known as FacebookRec, that combines Question Answering QA and recommendation. Many other CR datasets were then created, inspired by this early approach. These include: REDIAL [32], which is a CR dataset in the movie domain, TG-REDIAL [60] which considers Topic-Guided CR, INSPIRED [25] which focuses on sociable CR and many others [30, 31, 33, 37, 38, 48, 54]. However, none of the mentioned CR datasets consider the search task. Indeed, even if it may seem that some recommendation requests represent search queries (*e.g.,* "can you tell me the price of these shoes?"), CR datasets are based on close-world catalogues usually limited to a single or few domains and, therefore, are not suitable to provide satisfactory answers to open-world search queries. Moreover, in CR, differently from CS, personalization plays a key role. Indeed, most of the existing datasets include personalized conversations or some personalization features [25, 32, 33, 60].

A point of connection between CS and CR can be found in the product domain. Indeed, product search is a branch of IR that is mainly concerned with the retrieval of products instead of documents. Many existing approaches draw products from closed-world catalogues instead of open-world collections [27, 39]. This brings search really close to recommendation since the datasets used in product search are frequently shared with the ones used in recommendation. Amazon Reviews, for example, is a dataset built for recommendation that is employed also in product search [1, 10, 11]. This dataset does not contain search style queries, thus they are normally generated from the product categories or other elements. Moreover, several product search approaches represent adaptations of recommendation strategies [22]. Therefore, product search and recommendation is a natural domain of application for joint CSR systems. Indeed, CR systems usually focus on recommending items (*e.g.,* movies, music or products) and in CS there exists a sub-domain that focuses on conversational product search — CS systems which operate in the product domain — [29, 40]. This was noticed also by Bernard and Balog [9] who proposed MG-ShopDial, a dataset considering both search, recommendation and QA. However, MG-ShopDial offers a limited set of annotated conversations, without including evaluation or personalization data.

Recent advancements in LLMs have had a transformative impact across numerous fields. LLMs are increasingly being utilized to either replace or support human efforts in a wide range of tasks. In the field of Conversational Recommendation (CR), for instance, early datasets [25, 30, 32, 48] were primarily created through human participation. This often involved individuals recruited via crowdsourcing platforms engaging in simulated conversations to generate the necessary dialogue data. However, due to the high costs associated with human-generated data, recent CR datasets have increasingly turned to LLMs for conversations generation [31, 33]. This shift has demonstrated promising results, enabling the creation of high-quality datasets at scale. Either in CS and CR, it is common to assess the quality of conversations in a dataset by sampling a subset of the conversations and employing human assessors to evaluate the sampled dialogues. This step is fundamental, especially when LLMs are used in the generation process [33].

3 The Ideal Joint CSR Collection

In this section, we outline the key characteristics essential for a dataset designed to effectively evaluate joint Conversational Recommender Systems (CRS) systems. We define a conversation as a multi-turn interaction between a user and a system, where the user begins with an information need and iteratively refines it through dialogue with the system. This process continues until the search space is sufficiently narrowed and the user's need is satisfied.

Intent multiplicity. Depending on the information access scenario we consider — either CS or CR —, two interaction modalities might occur. If we consider a CS scenario, we assume the user seeks information held by the system. Therefore, the most typical interaction involves the user asking questions and the system answering them. The role of "questioner" and "answerers" might switch if we consider the mixed initiative interaction [3, 52], where the system can ask clarifying questions in case of ambiguous utterances. Nevertheless, the paradigm remains the same: the system holds the knowledge and the user must extract information from the interaction. We refer to the user's objective of extracting information from the system as *search intent.* If we focus on the CR use case, the roles are reversed. Here, the user holds the knowledge, meaning they are the only ones who truly know what they desire. The system's role is to identify the items that best match the user's

[2]https://grouplens.org/datasets/movielens/
[3]https://www.imdb.com/
[4]https://github.com/ACMRecSys/recsys-datasets

preferences. Once ready, the system provides information in the form of recommendations. We define the user's intention to receive recommendations as *recommendation intent*.

A key characteristic of the joint search and recommendation scenario is that the user may switch between intents from one utterance to another and, in some cases, even express mixed intents. For example, consider a user looking for a movie to watch (recommendation intent). The conversational agent might begin by asking questions to determine the desired characteristics of the movie—acting as the information seeker. However, the user might then request the plot of a specific film or biographical details of its lead actor (search intent). In this case, the system must shift roles and provide information. Therefore, in the joint CSR scenario, the system must seamlessly transition between roles—acting as either the information seeker or the information holder. Consequently, an ideal joint CRS should **interleave search-oriented, recommendation-oriented, and mixed-intent utterances** to evaluate its ability to adapt to different tasks.

Personalization. The ideal CSR dataset must rigorously test the personalization capabilities of joint CSR systems. Frequently, personalization is treated as a secondary aspect in information retrieval, where the primary focus is on the query content rather than the user who generated it. In contrast, personalization is a cornerstone of recommender systems, which rely on user history and preferences to determine which items to recommend [4]. In the conversational domain, personalization takes again a critical role. Even with the same information need, different users may follow entirely distinct paths to satisfy their information need [2, 34]. This inherent complexity makes the construction of evaluation collections and frameworks significantly more challenging compared to traditional information retrieval and recommendation systems. Regardless, we strongly argue in favour of CSR evaluation collections allowing to **assess joint CSR systems for personalized user needs.**

Independence from the evaluation paradigm. Information retrieval systems (including CS systems) are traditionally evaluated based on the set of documents returned in response to the user's query [17, 23]. Given a corpus of documents, relevance judgments determine which documents are relevant to the query. A system is considered more effective if it ranks relevant documents higher in the results list. Similarly, recommender systems and CR are evaluated based on how well they suggest items that users consider relevant, *i.e.,* items they have consumed or rated highly. To assess this, a common strategy is to split users' profiles into a training set and a test set, where the test set includes items the users have already consumed and liked [12, 13].

At the same time, recent advances in LLMs and generative models have demonstrated their remarkable ability to simulate human-like language [6, 43, 51]. These developments have given rise to new interaction paradigms, such as Retrieval Augmented Generation (RAG). In this paradigm, the system no longer retrieves or recommends a list of documents or items but instead generates a concise summary containing the key information sought by the user. This interaction paradigm appears particularly suited for the conversational scenario, where the system typically responds with an utterance composed of a few sentences at most. Nevertheless, the research community is still struggling to find a standard evaluation methodology to evaluate the generated content. Some evaluation measures, such as BLEU [45] or METEOR [7] measure the lexical overlap between the generated content and a canonical answer. Others, such as BERTscore [58], employ semantic representation models to quantify the semantic similarity between the system's response and the canonical answer. Finally, other paradigms, such as the one recently employed for the TREC RAG Track [47], employ atomic information units, also called *nuggets*, of the ideal answer, to determine how many of them are found in the generated response.

Since, at the time of writing, an evaluation paradigm has yet to become a standard, especially when it comes to the conversational setting, we argue that **the ideal joint CSR collection should be agnostic to the underlying evaluation paradigm**, allowing the practitioner to instantiate the one more suited to the setting. Such an ideal CSR collection would allow us to evaluate different classes of systems, either based on generative models or that answer to user's utterances with the classical "ten blue links". Furthermore, this allows us to evaluate more "classical" approaches that might still employ two different modules for the CR and CS tasks.

4 The Structure of CoSRec

CoSRec is a novel, multi-domain conversational dataset that allows us to overcome the existing limitations of CR and CS datasets by enabling us to consider CS and CR jointly. Product search and recommendation are a natural domain of application for joint CSR systems. CoSRec is based on conversations concerning products to exploit this natural point of connection between CS and CR. Importantly, this does not influence the information needs of the search intents, which are open-ended and general. Following the classic Cranfield paradigm for offline evaluation, CoSRec includes three elements: a set of *information needs*, *i.e.,* conversations, a set of suitable responses to such information needs, *i.e.,* a document corpus and an items catalogue, and a set of human-made *annotations*. At the same time, these elements have been adapted to fit our CSR scenario.

In the remainder of this Section, we report an overview of the elements constituting CoSRec, which are listed below:

- **Information needs**:
  - **Conversations**: sequences of utterances.
  - **User profiles**: summaries of the user interests for personalization.
- **Corpora**:
  - **MS-MARCO v2.1**: an IR corpus composed of passages derived from web documents.
  - **Amazon Reviews**: a recommendation dataset that includes both a product catalogue and users' reviews.
- **Human annotations**:
  - **Conversation quality assessments**: rating of the conversation quality.
  - **Intent labels**: labels specifying the user intent for a given utterance.
  - **Relevance judgments**: relevant documents or products for each intent occurrence.

Information Needs. In the CoSRec dataset, information needs are represented by conversations. CoSRec includes 9,249 conversations

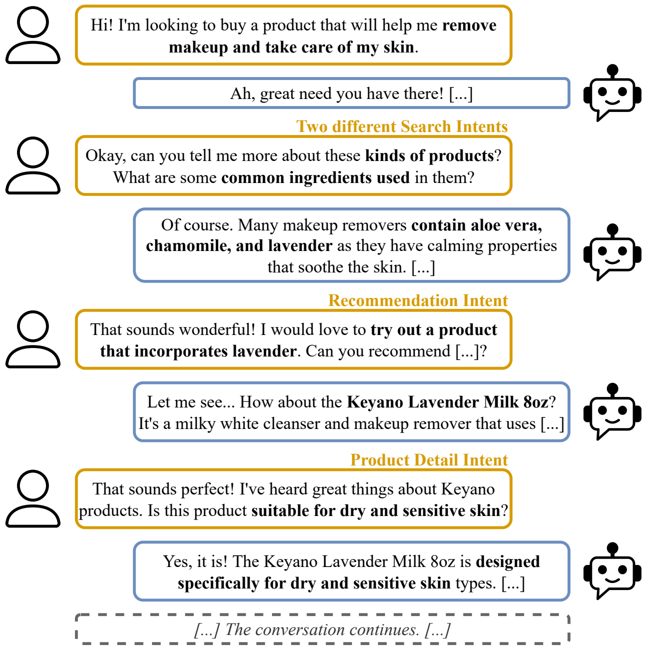


Figure 1: Example of conversation found in the CoSRec dataset.

split into 3 partitions: `CoSRec-Raw`: 8,938 non-annotated conversations containing 71,656 utterances; `CoSRec-Crowd`: 291 human-annotated conversations including 2,329 utterances; `CoSRec-Curated`: 20 deeply human-annotated conversations containing 150 utterances. Following the procedure described in Section 5.1.1, these conversations have been generated automatically. An example of conversation is shown in Figure 1. Each conversation is a multi-turn dialogue between a user and a system, where each turn corresponds to a user's utterance and a system's response. Hence, each user's utterance represents one or more information needs the system must satisfy. To account for the *intent multiplicity* requirement of the ideal joint CSR collection, these information needs require the system to answer with items drawn from a catalogue, *i.e.,* recommendation intent, or with information retrieved from a corpus of documents, *i.e.,* search intent. In `CoSRec-Crowd` and `CoSRec-Curated`, each intent is associated with a "canonical formulation" describing the information need in isolation and a series of human-made reformulations.

To satisfy the *personalization* requirement for a CSR collection, each conversation of the `CoSRec-Curated` partition is associated with at least 3 user profiles. Such user profiles are composed of two elements: a brief textual summary of the user's interests and a set of keywords. The users' profiles are constructed using the text of the users' past reviews and are used to steer the recommendation-oriented relevance judgments. Hence, they can be used to personalize the CSR system's responses.

Corpora. As mentioned earlier, in CSR, search and recommendation intents are satisfied by information drawn from a corpus and a catalogue of items, respectively. Mixed intents, on the other hand, require information from both data sources. Therefore, we need a corpus and a catalogue to serve as the foundation for the system's answers during evaluation. To this end, we rely on two publicly available resources: MS-MARCO v2.1 [41] for search intents and Amazon Reviews [26] for recommendation intents:

MS-MARCO v2.1 is a widely used IR corpus which contains over 113.5M passages obtained from web crawls. In CoSRec the set of passages derived from MS-MARCO v2.1 is used to support search intents.

Amazon Reviews is a traditional recommendation dataset that includes a product catalogue along with product metadata and user reviews. These two components enable both recommendation and personalization, as user interests can be inferred from the reviews. To ensure relevance, we use a refined version of Amazon Reviews, which we refer to as `AR-filtered`, focusing only on sensible items for recommendation and personalization.

Table 1: Distribution of labeled intents for `CoSRec-Curated` and `CoSRec-Crowd`.

Partition	# Labeled Intents			
	Search	Rec.	Pr. Det.	Total
Curated	43	62	38	143
Crowd	1,159	1,121	1,988	4,268

Table 2: Distribution of relevance judgments per scores and intent type in `CoSRec-Curated`.

Relevance	# Judgments		
	Search	Rec.	Total
Highly Relevant	571	5,950	6,521
Partially Relevant	696	4,097	4,793
Not Relevant	1,047	5,103	6,150

Annotations. Among the 9,249 conversations included in CoSRec, 311, corresponding to those in the `CoSRec-Crowd` and `CoSRec-Curated` partitions, are manually annotated to assess their quality. The quality evaluation considers four aspects, inspired by those employed in LLM-REDIAL [33]: fluency, informativeness, logicality, and coherence. Additionally, the same 311 conversations are labeled to identify user intents in each utterance. The considered intent types are: *search*: a request for general information about a topic related to a product; *recommendation*: a request for some product suggestion, according to user's requirements; *product detail*: a request for some details about a specific product. There is no limit to the number of intents that can be identified in a single user utterance. For each intent occurrence, the corresponding utterance (or a portion of it) is used to generate a canonical formulation along with reformulations. Since an utterance may contain multiple intents, it has as many canonical formulations (and corresponding reformulations) as the number of intents. Table 1 reports the number of labeled intents for each intent type in the `CoSRec-Crowd` and `CoSRec-Curated` partitions.

Based on the quality assessment results, 20 high-quality conversations are then selected and refined to form the CoSRec-Curated dataset. For each intent, the CoSRec-Curated reports manually-crafted relevance judgments. The relevance judgments for the search intent are built following a TREC-style paradigm and are not personalized. For recommendation intents, on the other hand, the relevance judgments are personalized on the users' profiles associated with the conversation. For completeness, we also report a non-personalized set of relevance judgments in the case of recommendation intents. Finally, for product detail, we report the product for which the user asked for more information. We opt for this fine-grained level of relevance judgments to satisfy the *independence from the evaluation paradigm* criterion. Moreover, building the relevance judgments independently for search and recommendation intents allows the evaluation of a wider spectrum of CSR systems: systems that employ sub-systems for the two tasks, systems that carry out the two tasks jointly with a unified module, systems which rely on personalization and systems which do not consider user preferences. Table 2 shows the number of human-made judgments organized by relevance level and intent type.

Comparison. We compare CoSRec with other existing CR and CS datasets. The outcomes of this detailed comparison are reported in Table 3. It shows that CoSRec has a comparable number of "test" conversations, utterances, and relevance judgments with respect to other CS datasets. Notably, it is the only dataset supporting both search and recommendation in the conversational context while also providing relevance assessments to enable evaluation. Furthermore, personalization plays a key role in recommendation, as evidenced by all the considered CR datasets supporting it. TREC iKAT, instead, is the only CS resource to support personalization. Given that iKAT is the most recent among the considered datasets, it highlights that the interest towards personalization is growing also within CS research. Consequently, the CoSRec dataset besides being among the few allowing to consider CS and CR jointly, it is also the only one including the necessary features to enable personalization.

5 Methodology

In this section, we describe how the different components of the datasets are generated and the annotation process.

5.1 Information Needs

Here, we describe the process of generating the information needs, including the conversations and the profiles of three users for each conversation.

5.1.1 Conversations Generation. CoSRec includes a total of 9,249 conversations generated by a LLM by iterating the following steps:

Step 1 We sample a random item from AR-filtered. This item is considered the target item.

Step 2 We concatenate the title and description of the target product and use it as a query to retrieve ten items from AR-filtered using BM25 as a retrieval model.

Step 3 We concatenate the description of the target product and the three reviews with the highest count of "useful" rates in the AR-filtered. We then use Llama 3.1 8B to extract a textual query from this string. Such query retrieves ten passages from MS-MARCO v2.1 using BM25.

Step 4 We fill in a prompt with i) the title and description of the target item and the three reviews rated "useful" the most; ii) the ten passages from Step 3; iii) the title, description, and the three most rated useful reviews of the ten products retrieved in Step 2. This prompt is fed to Llama 3.1 8B. The output is a simulated conversation between a user and a conversational agent about aspects related to the target item.

Step 5 We drop the conversation if: i) contains parts of the prompt or placeholder text; ii) is malformed or truncated; iv) is too long or too short.

It is important to emphasize that the conversations were not generated by making the LLM converse with itself. Instead, each conversation is the output of a single prompt, ensuring greater internal validity. The code used to generate the conversations, including prompts and filtering scripts, is available in the CoSRec repository for favoring reproducibility and extensions.

5.1.2 User Profile Generation. For the CoSRec-Curated partition, we generate a set of user profiles for every conversation, which are used to steer the personalized relevance judgments. These are the steps to generate such profiles:

Step 1 We sample three users, each having at least one review in the Amazon Reviews dataset for one of the ten items selected during conversation generation. In six conversations, only two users met these criteria and were therefore included.

Step 2 For each user, we extract the 10 reviews posted before reviewing the items considered in the previous step. These reviews are included in two prompts fed to Llama 3.1 8B to generate a textual and keyword-based description of the features important to the user.

We manually reviewed the profiles for consistency and meaningfulness, discarding unsatisfactory profiles.

5.2 Corpora

CoSRec builds upon two publicly available resources. Specifically, we employ Amazon Reviews [26], a large-scale dataset comprising product metadata (~48M products) and user reviews (~571M) from Amazon, collected from May 1996 to September 2023. We also exploit MS-MARCO v2.1 [41] segmented corpus[5], comprising over 113.5M passages extracted from web pages.

5.2.1 Product Catalogue Preprocessing. The Amazon Reviews dataset reports for each review: the rating, the title and the text of the review, the number of users that rated the review as helpful, a reference to the user that wrote the review and to the reviewed product, a boolean flag indicating if the user purchase of the product has been verified and other details. For each product instead, it provides: the title, the description, the features, the details (*e.g.,* materials, brand, sizes), the price, the category (also the hierarchy), and other information (*e.g.,* images). To build the product catalogue of CoSRec, we process Amazon Reviews dataset, generating AR-filtered. First, we merge each product's features and description (we refer to this string as "description"). Then we discard items with: i) empty

[5]This corpus is the same used in TREC RAG 2024 track [47]. More information can be found at https://trec-rag.github.io/annoucements/2024-corpus-finalization/.

Table 3: Characteristics of the test set of CoSRec and different CS and CR datasets.

	Dataset	Domain	# Convs.	# Utterances	# Assessed Utterances	Avg. Rel. Judgments	Supports Search	Supports Rec.	Supports Personalization
CSR	CoSRec-Curated	General	20	150	150	166.3	✓	✓	✓
	MG-ShopDial [9]	General	64	2196	–	–	✓	✓	✗
CS	QuAC [14]	General	1k	7.3k	7.3k	1	✓	✗	✗
	QReCC [5]	General	2.8k	16.5k	16.5k	1			✗
	TREC CAsT 2019 [17]	General	20	173	173	169.7			✗
	TREC CAsT 2020 [16]	General	25	216	208	194.5			✗
	TREC CAsT 2021 [18]	General	26	239	158	122.4			✗
	TREC CAsT 2022 [44]	General	18	205	163	256.1			✗
	TREC iKAT 2023 [2]	General	25	326	176	194.2			✓
CR	INSPIRED [25]	Movie	1k	35k	–	–	✗	✓	✓
	REDIAL [32]	Movie	10k	182k	–	–			
	TG-REDIAL [60]	Movie	10k	129k	–	–			
	LLM-REDIAL [33]	General	47.6k	482.6k	–	–			

title or description, written not in English, or with non-ASCII characters; ii) unavailable or without store, price, or category. In total, our filtered catalogue includes about 12.3M products. Furthermore, we drop empty reviews or reviews associated with a discarded item, ending with approx. 303.9M reviews.

5.3 Human Annotations: Conversation Quality Assessments

The CoSRec dataset comes with human-made quality assessments for a subset of 311 conversations (~3%) corresponding to `CoSRec-Crowd` and `CoSRec-Curated`. In particular, our annotation process involved 99 semi-expert human annotators.[6] Each conversation was assigned to five annotators to ensure that at least three quality assessments were available for each conversation. Such quality assessments are given on a 1 to 5 scale and concern 4 aspects:

Fluency A conversation is fluent when it is well organized, in regular English grammar, easy to understand, and has a continuous flow.

Informativeness A conversation is informative when the utterances include substantial content, communicate the user's needs, or deliver valuable information.

Logicality (*a.k.a.,* Inverse Perplexity) A conversation has a high logicality when its utterances are organized according to a logical flow and align with common reasoning.

Coherence A conversation is coherent when the user and the system follow each other without unexpected or inappropriate utterances. Furthermore, given the specific product search setting, the user's final utterance must be consistent with the needs expressed during the dialogue.

Figure 2a reports the distribution of the quality assessments averaged across the four aspects. 260 (approx. 83%) conversations have an average rating above 3.5. Figure 2b depicts the ratings distribution for each quality aspect considered. Only a few conversations received ratings below 4 in fluency and informativeness. In contrast, lower ratings were more frequent for logicality and coherence. This trend may be attributed to the characteristics of LLMs, which are effective in generating fluent and human-like text but often demonstrate limitations in logical reasoning and maintaining coherence [6, 53].

5.4 Human Annotations: Intents Labeling

The human annotators also associated intent labels to each utterance of the `CoSRec-Crowd` and `CoSRec-Curated` conversations. Each utterance is annotated with zero, one, or more among "search", "recommendation", and "product detail" intents.

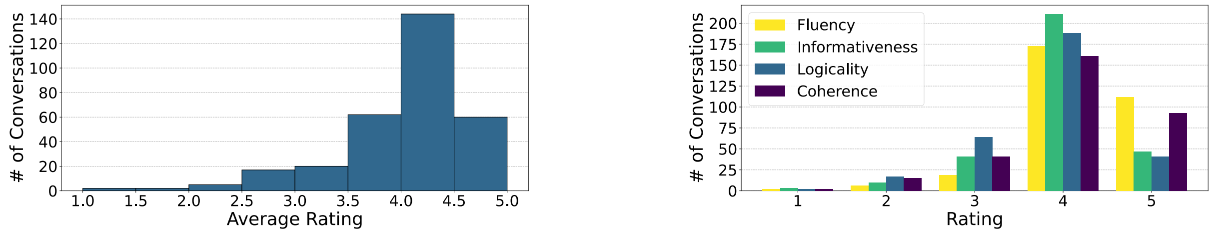
Search The user asked for general information about a topic related to the product they are discussing.

Recommendation The user asks for some products to be suggested, according to her requirements.

Product Detail The user inquires about details of the product being discussed (*e.g.,* price, brand, size).

The type of answer is the main difference between "search" and "product detail". A product detail question can be answered by inspecting the product's description using information extraction approaches [36, 42]. On the other hand, search intents denote open-ended questions whose answers are likely to be found on an external corpus. These labels are released as they are for the `CoSRec-Crowd` portion of the datset. In contrast, for the 20 `CoSRec-Curated` conversations, the authors of this paper further refined the labels by reviewing cases where annotators did not reach unanimity. Through discussion, they assigned the most appropriate label. After intent labeling, approximately seven utterances per conversation in the `CoSRec-Crowd` partition were annotated. Specifically, 1,159 utterances were labeled with the *search* intent, 1,121 with the *recommendation* intent, and 1,988 with the *product detail* intent. For each conversation in the `CoSRec-Curated` partition, following the label refinement process, an average of approximately 7.15 intents were identified: ~2.15 *search* intents, ~3.10 *recommendation* intents, and ~1.90 *product detail* intents.

[6] Human annotators were voluntary master students recruited from the Information Retrieval and Recommendation course and Machine Learning course at Polytechnic University of Bari. The maximum workload was estimated in 3 hours, to ensure a fair effort.



(a) Distribution of the conversations based on the rating averaged on the quality aspects.

(b) Number of conversations for each rating value for each quality aspect.

Figure 2: Quality of the LLM-generated conversations

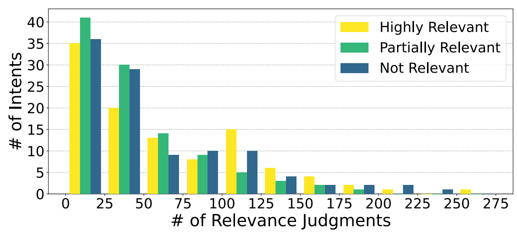


Figure 3: Distribution of highly relevant, partially relevant, and not relevant judgments across intents.

Along with intent labels, human annotators also provided a stand-alone formulation. This formulation is a self-explanatory textual description of the information need, independent of the conversation's context, as it fully encapsulates it. In most cases, the stand-alone formulation for a search intent resembles a query, while for a recommendation intent, it is akin to an ideal product description. Since each conversation, and therefore each utterance, was annotated by multiple annotators, we define the longest stand-alone formulation as the *canonical formulation*. In contrast, the others are considered *reformulations*. As with intent labeling, the stand-alone formulations in `CoSRec-Curated` were carefully reviewed to correct typographic errors and ensure consistency. After this revision, each intent-labeled utterance in `CoSRec-Curated` has a minimum of 1, a maximum of 5, and an average of 2.79 stand-alone formulations.

5.5 Human Annotations: Relevance Judgements

The `CoSRec-Curated` portion of the dataset contains a total of 17,464 relevance judgments for user intents related to search and recommendation. Each intent has between 26 and 452 judgments, with an average of 166.3. Figure 3 shows the distribution of relevance judgments across intents for each relevance level. Of these judgments, 11,314 indicate either high or partial relevance, with per-intent values ranging from 3 to 311 (average: 107.75). In the following, we describe in more detail the relevant judgment collection procedure for each type of intent.

5.5.1 Search Intents Relevance Judgements. To construct the relevance judgments, we follow the standard TREC-style annotation procedure, which involves two steps: building a pool and judging the documents. To form the pool, for each search intent, we consider every stand-alone formulation (the canonical formulation and all reformulations) of the utterance that generated it. Such formulations are used to retrieve the documents from MS-MARCO v2.1. In detail, we use BM25 to retrieve 1000 documents. This set of documents is then reranked using SPLADE [24], TCT ColBERT [35] and Contriever [28]. Thus, for each stand-alone formulation, we obtain 4 ranked lists. All the ranked lists are pooled together, with a pooling depth of 10.

During the relevance judgment phase, each *(search intent, document)* pair was evaluated to ensure at least three human relevance judgments. Assessors had access to (i) the canonical formulation of the intent, (ii) the conversation up to the utterance from which the intent was derived, and (iii) the document text. Based on this information, they assigned a relevance judgment on a 0-2 rating scale, defined as follows:

0 – Not Relevant The document is completely unrelated to the query. It does not contain any, even partial, information useful to provide the correct response to the user.

1 – Partially Relevant The document contains some information related to the query but does not provide a complete response. It may include partial information or details about some particular facets of the topic.

2 - Highly Relevant The document is sufficient to provide a complete and meaningful response.

The gap between partial and high relevance forces the annotators to take a clear stance, easing their work and avoiding ambiguity. To account for possible systematic biases introduced by the assessors, we combine the relevance judgments through majority voting. To break ties, we weigh the relevance judgments of the assessors by considering their average Cohen's κ [15] with every other assessor as a measure of their reliability.

5.5.2 Recommendation Intent Relevance Judgements. When gathering recommendation intent relevance judgments, we introduce a personalization aspect to satisfy the *personalization* requirement of a CSR collection. In detail, this calls for a few changes compared to the search intents relevance judgements collection. In detail, we generate a set of queries to retrieve the products from `AR-filtered` for each generated user profile associated with the conversation. This set of queries contains the canonical formulation of the intent, all reformulations, and their concatenation with the generated user's keyword-based profile. Additionally, we generate the queries for the so-called "cold-start" profile, *i.e.,* an empty profile. In this case, we use only the canonical formulation and the reformulations. Using these queries, as before, we employ BM25, TCT ColBERT, Contriever, and SPLADE to obtain the (re-)ranked product lists from the `AR-filtered` dataset for each profile, including the "cold-start" profile. We then extract the pool of products to be judged — one for each profile.

In this case, the human assessor assigned relevant judgments to triples *(recommendation intent, product, profile)*. Notice that each triple is judged by at least two assessors and the possible ties are broken in the same way as for the search intents. Thus, the assessors had access to i) the canonical formulation of the intent, ii) the conversation up to the utterance from which the intent was derived, iii) the textual description of the profile, iv) the product title and its description extracted from `AR-filtered` dataset. In the case of the "cold-start" profile, we do not provide any profile information. The human assessors were instructed to imagine impersonating the user corresponding to the profile and consider their preferences when assigning the relevance judgments. On the other hand, in the case of the "cold-start" profile, they were asked to express relevance judgments irrespective of the characteristics of any user. As for documents, relevance judgments are on a scale of 0-2 and correspond to:

0 – Not Relevant The product is completely unrelated to the request for the considered user.

1 – Partially Relevant The product is partially related to the request for the considered user. Despite this, the user might prefer similar products or alternatives.

2 - Highly Relevant The product is completely related to the request for the considered user.

Product Detail Intents. With a "product detail" intent, the user wishes to satisfy an information need about a specific product and its description. Therefore, the ground truth, in this case, corresponds to the identifier of the product the question is about. If the human assessor annotated the utterance with "product detail", we searched the corresponding product on the `AR-filtered` dataset. After manually ensuring that the LLM that generated the conversation did not hallucinate the product, we link the product identifier to the utterance. We plan to extend future versions of CoSRec, by also annotating the specific span of text where the answer to the question is found.

6 Limitations

While a deployed system can be evaluated via A-B testing and other online experiments, the development of a system—especially during the first phases—requires an experimental collection that can be used offline. Traditionally, IR experimental collections are built employing IR systems to retrieve the documents later annotated by the experts. Similarly, RS collections rely on historical data logs derived from a real-life service. This cannot be done in the CSR domain as there exists no system from which we can take the logs as data. Indeed, traditional CS and CR systems operate separately. At the same time, currently, LLMs are not meant to retrieve documents from a corpus and items from a catalog and behave as interfaces rather than information access engines. This raises a "chicken-and-egg" situation: the community still lacks CSR systems to extract the data from and the data to develop CSR systems. Consequently, we were forced to build CoSRec treating and annotating search and recommendation intents separately. Since the conversations did not occur in a real-life scenario and were generated by an LLM, some utterances might feel unnatural to a human reader. Nevertheless, this collection may represent the cornerstone for the CSR research domain and to start collecting joint CSR data: using CoSRec, the research community can develop CSR systems whose logs can be used as future collections.

CoSRec supports personalization, but only to a limited extent. Specifically, user profiles are not factored into the generation of conversations. As a result, according to the *personalization* criterion of the ideal CSR collection, different users with the same information needs follow the same sequence of utterances to reach an answer. Furthermore, CoSRec only incorporates personalization for the recommendation intents. This limitation arises because of the scarcity of publicly available data suitable for personalization within the CS context. Finally, personalization does not concern users' purchase history but only their concept of "relevance", preventing the full application of collaborative filtering strategies. Therefore, as before, the main goal is to employ CSR as a starting point to develop CSR system and collect appropriate, user-centric data to enhance personalization.

7 Conclusions and Future Work

In this work, we have outlined the key features of an ideal joint CSR collection and introduced CoSRec, a novel dataset designed for the CSR context. CoSRec comprises 9.2k conversations encompassing pure search, pure recommendation, and mixed search-and-recommendation utterances, all generated using LLMs. A subset of 311 conversations has been human-annotated to evaluate their quality and to label user intents. Additionally, for 20 high-quality conversations, CoSRec provides relevance judgments for each labelled intent, personalized for recommendation scenarios. While CoSRec is still limited in terms of personalization, it meets all the requirements of an ideal CSR collection. We believe that CoSRec will foster research in the area by providing a robust foundation for developing and evaluating CSR systems. To ensure reproducibility and encourage extensions, we make all code, scripts, prompts, and the dataset publicly available. Future work will focus on the generation and labelling of new conversations and the improvement of personalization, by including it in the generation process and extending it to the search intents. Furthermore, the current version of CoSRec will allow the development of actual integrated CSR systems that can be used to collect additional data, ground truth labels, and conversations.

Acknowledgments

This work has received support from CAMEO, n. 2022ZLL7MW PRIN 2022. Moreover, we acknowledge the support of "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" (PE00000013), "Extreme Food Risk Analytics" (EFRA), GA n. 101093026, funded by the European Commission under the NextGeneration EU programme.

References

[1] Qingyao Ai, Yongfeng Zhang, Keping Bi, and W. Bruce Croft. 2020. Explainable Product Search with a Dynamic Relation Embedding Model. *ACM Trans. Inf. Syst.* 38, 1 (2020), 4:1–4:29. doi:10.1145/3361738

[2] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2023. TREC iKAT 2023: The Interactive Knowledge Assistance Track Overview. In *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14-17, 2023 (NIST Special Publication, Vol. 500-xxx)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec32/papers/Overview_ikat.pdf

[3] Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing Mixed Initiatives and Search Strategies during Conversational Search. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 16–26. doi:10.1145/3459637.3482231

[4] Mohammed Fadhel Aljunid, Manjaiah Doddaghatta Huchaiah, Mohammad Kazim Hooshmand, Wasim A. Ali, Amrithkala M. Shetty, and Sadiq Qaid Alzoubah. 2025. A collaborative filtering recommender systems: Survey. *Neurocomputing* 617 (2025), 128718. doi:10.1016/J.NEUCOM.2024.128718

[5] 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, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 520–534. doi:10.18653/V1/2021.NAACL-MAIN.44

[6] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR* abs/2312.11805 (2023). doi:10.48550/ARXIV.2312.11805 arXiv:2312.11805

[7] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. https://aclanthology.org/W05-0909/

[8] Nicholas J. Belkin and W. Bruce Croft. 1992. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Commun. ACM* 35, 12 (1992), 29–38. doi:10.1145/138859.138861

[9] Nolwenn Bernard and Krisztian Balog. 2023. MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2775–2785. doi:10.1145/3539618.3591883

[10] Keping Bi, Qingyao Ai, and W. Bruce Croft. 2021. Learning a Fine-Grained Review-based Transformer Model for Personalized Product Search. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 123–132. doi:10.1145/3404835.3462911

[11] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W. Bruce Croft. 2019. Conversational Product Search Based on Negative Feedback. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 359–368. doi:10.1145/3357384.3357939

[12] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Inf. Retr. J.* 23, 4 (2020), 387–410. doi:10.1007/S10791-020-09371-3

[13] Pablo Castells and Alistair Moffat. 2022. Offline Recommender System Evaluation: Challenges and New Directions. *AI Mag.* 43, 2 (2022), 225–238. doi:10.1002/AAAI.12051

[14] 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, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2174–2184. doi:10.18653/V1/D18-1241

[15] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104 arXiv:https://doi.org/10.1177/001316446002000104

[16] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf

[17] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *CoRR* abs/2003.13624 (2020). arXiv:2003.13624 https://arxiv.org/abs/2003.13624

[18] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. TREC CAsT 2021: The Conversational Assistance Track Overview. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021 (NIST Special Publication, Vol. 500-335)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/Overview-CAsT.pdf

[19] Tommaso Di Noia, Guglielmo Faggioli, Marco Ferrante, Nicola Ferro, Fedelucio Narducci, Raffaele Perego, and Giuseppe Santucci. 2024. CAMEO: Fostering Joint Conversational Search and Recommendation. In *Proceedings of the 32nd Symposium of Advanced Database Systems, Villasimius, Italy, June 23rd to 26th, 2024 (CEUR Workshop Proceedings, Vol. 3741)*, Maurizio Atzori, Paolo Ciaccia, Michelangelo Ceci, Federica Mandreoli, Donato Malerba, Manuela Sanguinetti, Antonio Pellicani, and Federico Motta (Eds.). CEUR-WS.org, 290–301. https://ceur-ws.org/Vol-3741/paper33.pdf

[20] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.06931

[21] Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 5917–5923. doi:10.18653/V1/D19-1605

[22] Jon Eskreis-Winkler, Yubin Kim, and Andrew Stanton. 2023. XWalk: Random Walk Based Candidate Retrieval for Product Search. In *Proceedings of the 2023 SIGIR Workshop on eCommerce co-located with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), Taipei, Taiwan, July 27, 2023 (CEUR Workshop Proceedings, Vol. 3589)*, Surya Kallumadi, Yubin Kim, Tracy Holloway King, Shervin Malmasi, Maarten de Rijke, and Jacopo Tagliabue (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3589/paper_22.pdf

[23] Nicola Ferro and Maria Maistro. 2024. Evaluation of IR Systems. In *Information Retrieval: Advanced Topics and Techniques*, Omar Alonso and Ricardo Baeza-Yates (Eds.). ACM Books, Vol. 60. ACM, 111–191. doi:10.1145/3674127.3674132

[24] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2288–2292. doi:10.1145/3404835.3463098

[25] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8142–8152. doi:10.18653/V1/2020.EMNLP-MAIN.654

[26] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian J. McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. *CoRR* abs/2403.03952 (2024). doi:10.48550/ARXIV.2403.03952 arXiv:2403.03952

[27] Zhizhang Hu, Shasha Li, Ming Du, Arnab Dhua, and Douglas Gray. 2024. Multimodal Learning with Online Text Cleaning for E-commerce Product Search. In *Proceedings of the ACM SIGIR Workshop on eCommerce 2024 co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024), Washington D.C., USA, July 18, 2024 (CEUR Workshop Proceedings, Vol. 3843)*, Surya Kallumadi, Yubin Kim, Tracy Holloway King, Maarten de Rijke, and Vamsi Salaka (Eds.). CEUR-WS.org. https://ceur-

ws.org/Vol-3843/paper_20.pdf

[28] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=jKN1pXi7b0

[29] Vahid Sadiri Javadi, Martin Potthast, and Lucie Flek. 2023. OpinionConv: Conversational Product Search with Grounded Opinions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2023, Prague, Czechia, September 11 - 15, 2023*, David Schlangen, Svetlana Stoyanchev, Shafiq Joty, Ondrej Dusek, Casey Kennington, and Malihe Alikhani (Eds.). Association for Computational Linguistics, 66–76. doi:10.18653/V1/2023.SIGDIAL-1.6

[30] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A. Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 1951–1961. doi:10.18653/V1/D19-1203

[31] Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A Review-driven Persona-Knowledge Grounded Conversational Recommendation Dataset. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 1105–1120. doi:10.18653/V1/2024.FINDINGS-ACL.65

[32] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9748–9758. https://proceedings.neurips.cc/paper/2018/hash/800de15c79c8d840f4e78d3af937d4d4-Abstract.html

[33] Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 8926–8939. doi:10.18653/V1/2024.FINDINGS-ACL.529

[34] Allen Lin, Ziwei Zhu, Jianling Wang, and James Caverlee. 2023. Enhancing User Personalization in Conversational Recommenders. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben (Eds.). ACM, 770–778. doi:10.1145/3543507.3583192

[35] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling Dense Representations for Ranking using Tightly-Coupled Teachers. *CoRR* abs/2010.11386 (2020). arXiv:2010.11386 https://arxiv.org/abs/2010.11386

[36] Pai Liu, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Zongsheng Li, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. 2024. A Survey on Open Information Extraction from Rule-based Model to Large Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 9586–9608. https://aclanthology.org/2024.findings-emnlp.560

[37] Yuanxing Liu, Weinan Zhang, Baohua Dong, Yan Fan, Hang Wang, Fan Feng, Yifan Chen, Ziyu Zhuang, Hengbin Cui, Yongbin Li, and Wanxiang Che. 2023. U-NEED: A Fine-grained Dataset for User Needs-Centric E-commerce Conversational Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2723–2732. doi:10.1145/3539618.3591878

[38] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1036–1049. doi:10.18653/V1/2020.ACL-MAIN.98

[39] Chen Luo, Rahul Goutam, Haiyang Zhang, Chao Zhang, Yangqiu Song, and Bing Yin. 2023. Implicit Query Parsing at Amazon Product Search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 3380–3384. doi:10.1145/3539618.3591858

[40] Heli Ma, Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Yi Bin, and Yang Yang. 2024. Ask or Recommend: An Empirical Study on Conversational Product Search. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, Edoardo Serra and Francesca Spezzano (Eds.). ACM, 3927–3931. doi:10.1145/3627673.3679875

[41] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[42] Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A Survey on Open Information Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, 3866–3878. https://aclanthology.org/C18-1326/

[43] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). doi:10.48550/ARXIV.2303.08774 arXiv:2303.08774

[44] Paul Owoicho, Jeff Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas, and Svitlana Vakulenko. 2022. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/Overview_cast.pdf

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. doi:10.3115/1073083.1073135

[46] Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco De Nadai, and Hugues Bouchard. 2024. Bridging Search and Recommendation in Generative Retrieval: Does One Task Help the Other?. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, Tommaso Di Noia, Pasquale Lops, Thorsten Joachims, Katrien Verbert, Pablo Castells, Zhenhua Dong, and Ben London (Eds.). ACM, 340–349. doi:10.1145/3640457.3688123

[47] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. *CoRR* abs/2406.16828 (2024). doi:10.48550/ARXIV.2406.16828 arXiv:2406.16828

[48] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, Satoshi Nakamura, Milica Gasic, Ingrid Zuckerman, Gabriel Skantze, Mikio Nakano, Alexandros Papangelis, Stefan Ultes, and Koichiro Yoshino (Eds.). Association for Computational Linguistics, 353–360. doi:10.18653/V1/W19-5941

[49] Zihua Si, Zhongxiang Sun, Xiao Zhang, Jun Xu, Xiaoxue Zang, Yang Song, Kun Gai, and Ji-Rong Wen. 2023. When Search Meets Recommendation: Learning Disentangled Search Representation for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 1313–1323. doi:10.1145/3539618.3591786

[50] Nicola Stokes. 2006. *TREC: Experiment and Evaluation in Information Retrieval* Ellen M. Voorhees and Donna K. Harman (editors) (National Institute of Standards and Technology), Cambridge, MA: The MIT Press (Digital libraries and electronic publishing series, edited by William Y. Arms), 2005, x+462 pp; hardbound, ISBN 0-262-22073-3. *Comput. Linguistics* 32, 4 (2006), 563–567. doi:10.1162/COLI.2006.32.4.563

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). doi:10.48550/ARXIV.2302.13971 arXiv:2302.13971

[52] Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2020. An Analysis of Mixed Initiative and Collaboration in Information-Seeking Dialogues. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2085–2088. doi:10.1145/3397271.3401297

[53] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation From Deductive, Inductive and Abductive Views. *CoRR* abs/2306.09841 (2023). doi:10.48550/ARXIV.2306.09841 arXiv:2306.09841

[54] Hu Xu, Seungwhan Moon, Honglei Liu, Bing Liu, Pararth Shah, and Philip S. Yu. 2020. User Memory Reasoning for Conversational Recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5288–5308. doi:10.18653/V1/2020.COLING-MAIN.463

[55] Hamed Zamani and W. Bruce Croft. 2018. Joint Modeling and Optimization of Search and Recommendation. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018 (CEUR Workshop Proceedings, Vol. 2167)*, Omar Alonso and Gianmaria Silvello (Eds.). CEUR-WS.org, 36–41. https://ceur-ws.org/Vol-2167/paper2.pdf

[56] Hamed Zamani and W. Bruce Croft. 2020. Learning a Joint Search and Recommendation Model from User-Item Interactions. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 717–725. doi:10.1145/3336191.3371818

[57] Hansi Zeng, Surya Kallumadi, Zaid Alibadi, Rodrigo Frassetto Nogueira, and Hamed Zamani. 2023. A Personalized Dense Retrieval Framework for Unified Information Access. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 121–130. doi:10.1145/3539618.3591626

[58] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[59] Kai Zhao, Yukun Zheng, Tao Zhuang, Xiang Li, and Xiaoyi Zeng. 2022. Joint Learning of E-commerce Search and Recommendation with a Unified Graph Neural Network. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 1461–1469. doi:10.1145/3488560.3498414

[60] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 4128–4139. doi:10.18653/V1/2020.COLING-MAIN.365