RecQuest: Towards Estimating User Domain Knowledge in Conversational Recommender Systems

Ivica Kostric
University of Stavanger
Stavanger, Norway
ivica.kostric@uis.no

Ujwal Gadiraju
Delft University of Technology
Delft, Netherland
u.k.gadiraju@tudelft.nl

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Abstract

The ideal conversational recommender system (CRS) acts like a savvy salesperson, adapting its language and suggestions to a user's expertise level. However, most current systems treat all users as experts, leading to frustrating and inefficient interactions when users are unfamiliar with a domain. Systems that can adapt their conversational strategies to a user's knowledge level stand to offer a much more natural and effective experience. To enable such adaptation, a CRS must first be able to estimate a user's domain knowledge from interaction signals. Yet, accurately estimating knowledge typically requires tailored interactions to elicit those signals in the first place, creating a fundamental chicken-and-egg problem. In this work, we take a first step toward breaking this dependency by introducing a new task: estimating user domain knowledge directly from conversational transcripts. A key obstacle to such estimation is the lack of suitable data; to our knowledge, no existing dataset captures the conversational behaviors of users with varying levels of domain knowledge. Furthermore, in most dialogue collection protocols, users are free to express their own preferences, which tends to concentrate on popular items and well-known features, offering little insight into how novices explore or learn about unfamiliar features. To address this, we design RecQuest, a game-with-a-purpose data collection protocol that elicits varied expressions of knowledge while using a target-aware CRS to guide interactions, release the resulting dataset, and provide baseline methods and analyses to support future work on user-knowledge-aware CRS.

CCS Concepts

• **Information systems** → **Recommender systems**; **Personalization**.

Keywords

Conversational Recommender Systems; User Knowledge Estimation; User Modeling; Data Collection

ACM Reference Format:

Ivica Kostric, Ujwal Gadiraju, and Krisztian Balog. 2026. RecQuest: Towards Estimating User Domain Knowledge in Conversational Recommender Systems. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26), July 25, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3805713.3820438

ACM ISBN 979-8-4007-2600-2/2026/07
https://doi.org/10.1145/3805713.3820438

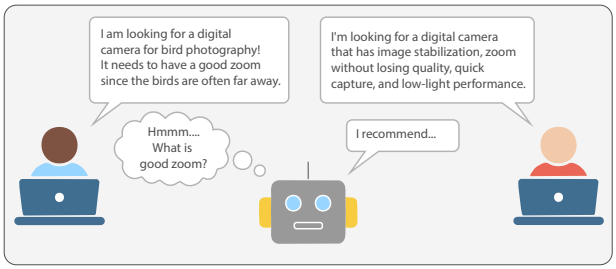

Figure 1: An example from our collected dataset illustrating the different levels of product specification between a novice and an expert while describing their needs for a digital camera purchase. Novices tend to describe goals and contexts, whereas experts tend to specify technical constraints and components more clearly.

1 Introduction

Conversational recommender systems (CRS) address the challenge of helping users find relevant items in vast catalogs by engaging them in interactive dialogues, typically involving preference elicitation based on item attributes [16]. Their effectiveness relies heavily on accurately capturing user preferences and interpreting their expressions about desired attributes [31]. However, most current CRS approaches typically assume that users understand item attributes clearly and can directly map them to their preferences [27, 40, 43]. In contrast, Kostric et al. [22] propose a usage-oriented preference elicitation strategy, specifically targeting users with low-domain knowledge. In isolation, both approaches can lead to limited and inefficient interactions, as users differ significantly in domain knowledge, vocabulary, and preferred interaction style. As illustrated in Figure 1, a novice might express their needs differently than an expert when buying a digital camera. As a result, systems that rely on a single interaction strategy force users to adapt to the system's language. However, recent work shows that task performance and user satisfaction improve significantly when a CRS instead adapts its interaction style to the user's domain knowledge [21].

To bridge this gap and enable the creation of more adaptive systems, we introduce the novel task of estimating user domain knowledge from conversations, allowing a CRS to tailor its interaction strategy—from preference elicitation to the presentation of recommendations and explanations. While empirical studies in information retrieval and recommender systems have shown that domain expertise shapes user interactions [28, 38], estimating this knowledge from multi-turn dialogues remains a novel problem in the CRS setting.

A major obstacle in advancing the task of user domain knowledge estimation is the lack of datasets that capture such domain knowledge within a conversational context. Existing CRS datasets rely on open-ended protocols in which participants discuss only the attributes they are familiar with [17, 26, 32]. This approach leads to conversations focused primarily on well-known aspects of the domain, providing limited evidence about how users express needs, ask questions, or refine preferences when encountering unfamiliar aspects. Given these challenges, a critical question is: ***(RQ1)*** *How can we design a data collection protocol that elicits signals of domain knowledge through engaging and natural conversations?* We address this by developing a game-with-a-purpose (GWAP)[1] [2, 37] protocol that combines open-ended interaction with a target-oriented objective. Specifically, participants are assigned a hidden target item, presented only as a concise background narrative. During the interaction, they must interact with the CRS to identify the item that best matches the narrative by providing preference statements, responding to elicitation prompts, and discussing the system's recommendations. This design is distinct from both known-item search [24] and open-ended preference elicitation, while combining elements of each: participants are looking for a particular item without knowing its exact identity, while the open-ended protocol encourages the articulation, refinement, and clarification of preferences that expose signals of domain knowledge.

An inherent challenge in estimating user domain knowledge is determining the ground truth. This leads us to ask: ***(RQ2)*** *How do objective knowledge scores compare to user self-assessments, and which provides a more reliable ground truth?* We address this by contrasting participants' self-assessed confidence ratings with their performance on a domain-specific technical questionnaire. We then determine the more robust ground truth by analyzing which measure better predicts actual task success, thereby identifying the calibration biases that undermine self-reported expertise.

With the ground truth established, we next examine which observable signals distinguish users with varying domain knowledge. Specifically, we ask: ***(RQ3)*** *What identifiable traces differentiate expert behavior from novice behavior, and to what extent do these patterns make users with different knowledge levels distinguishable in practice?* To answer this, we analyze the collected dialogues for measurable linguistic and interactional markers, including utterance length, vocabulary, and the frequency and specificity of the phrases used. Based on prior work [28, 38], we expect that knowledgeable users articulate their needs through more precise, attribute-oriented language, whereas users with limited domain knowledge rely on high-level or usage-based descriptions and exhibit more frequent clarification behavior.

Finally, we investigate the feasibility of automatic knowledge estimation from dialogue. Specifically, ***(RQ4)*** *How effectively can large language models (LLMs) estimate users' domain knowledge from conversational data?* To answer this, we examine whether these models can infer knowledge levels directly from raw dialogue transcripts. This involves assessing how well-separated the knowledge categories are, and whether users at different expertise levels exhibit sufficiently distinct linguistic, behavioral, or interactional markers to enable robust classification.

In summary, our contributions are as follows:

- We introduce the task of estimating user domain knowledge in conversational recommenders directly from dialogues.
- We design RecQuest, a game-with-a-purpose data collection protocol that uses background narratives and attribute clues to elicit natural conversations rich in domain knowledge signals.
- We collect and release a novel CRS dataset capturing these domain-specific interactions. We note that this dataset is intended primarily to validate the proposed collection protocol and support initial feasibility analyses, rather than to serve as a large-scale supervised training benchmark.
- We conduct a detailed analysis of the collected dialogues to uncover the behavioral and linguistic patterns that distinguish users with different knowledge levels.
- We establish LLM-based baselines for estimating user domain knowledge from conversational data, and analyze the separability of knowledge levels on this challenging task.

The developed CRS, study protocols, and collected dialogues are made available at: https://github.com/iai-group/crs-knowledge.

2 Related Work

User domain knowledge has been previously investigated in the context of web search [28, 35, 38]. Prior work differs mainly in how expertise is determined and modeled. The two most common approaches for determining user knowledge are subjective self-assessments and objective knowledge tests. Self-assessment studies ask users to report their familiarity or perceived expertise, most often distinguishing between two and five categories [12, 29, 30]. Objective approaches instead—also popularly employed in the '*search as learning*' realm—rely on domain quizzes or knowledge tests whose scores define knowledge levels [10, 13, 19, 28, 39, 41, 42]. Research based on both approaches consistently finds that experts issue more structured and technical queries, use more precise and attribute-oriented vocabulary, reformulate less, while novices rely more on general, case-oriented descriptions and ask more clarification questions [30, 35]. Prior work on adaptive preference elicitation further shows that user characteristics and domain familiarity shape interaction behavior and satisfaction, with more knowledgeable users responding better to structured, attribute-focused elicitation, and less knowledgeable users benefiting more from guided and example-based approaches [20].

Our work is situated at the intersection of two key research areas. We first review the main paradigms of CRS, as they provide the architectural context for the system we developed for our data collection. We then critically examine existing dialogue collection methods, highlighting the gap our protocol directly addresses.

Conversational Recommender Systems. Research on CRS is broadly categorized into two main groups. *Attribute-based CRS* cast the dialogue as a slot-filling exercise [4, 8, 25], eliciting preferences from users about items or item attributes with fixed templates. *Generation-based CRS*, in contrast, generates free-form responses while recommending items, generally utilizing an end-to-end architecture that integrates the conversation and the recommendation

[1] Here, "game" means an incentivized task with goals and rules, designed to elicit useful data from participants, rather than a game-theoretic model of strategic interaction.

components [44]. Our work is relevant to both paradigms, as estimating user knowledge can, among others, help tailor the elicitation strategy and the explanatory language. Our work employs a hybrid CRS that combines attribute-based preference elicitation with natural language generation for recommendations and explanations. The system maintains an internal representation of user preferences and supports interactions entirely in natural language.

Dialogue Collection Methods. Significant effort has been dedicated to collecting datasets that capture how users interact with conversational recommenders [3, 6, 14, 17, 18, 26, 32–34]. Influential examples like ReDial [26] and INSPIRED [14] are created as a conversation between two participants, where one user (the "seeker") states open-ended preferences and another (the "recommender") suggests items that fit those preferences. However, these open-ended protocols often yield shallow conversations, with seekers frequently accepting recommendations uncritically. More importantly for our work, these protocols do not explicitly control for or capture user domain knowledge, making it difficult to study its effects. While other datasets use coached human-human protocols to emphasize naturalness (e.g., CCPE-M [32], MG-ShopDial [3]) or focus on human-machine settings (e.g., Schema-Guided Dialogue [33], MultiWOZ [6]), none explicitly incorporates user domain knowledge or controls for expertise.

3 Problem Statement

The realization of a truly adaptive conversational recommender system presents a fundamental "chicken and egg" problem. Ideally, a CRS should act like a savvy salesperson, adapting its language and suggestions to each user's level of expertise. However, to effectively tailor these interactions, the system must first estimate the user's domain knowledge; yet, accurate estimation often requires tailored interactions—such as specific elicitation strategies—to surface those knowledge signals in the first place. We aim to break this co-dependency by establishing the necessary groundwork for knowledge estimation independent of complex adaptation strategies. In this work, we focus on the first half of this challenge: designing a protocol to capture natural variations in domain knowledge and introducing baseline methods to estimate it. By isolating the estimation task, we provide the resources required to build systems that can eventually close the loop and personalize the experience dynamically.

Following prior work on modeling knowledge, we adopt a discretized representation, where users are grouped into a small number of categories [5]. Specifically, we categorize users into three relative levels of knowledge: `low`, `medium`, and `high`. Low domain knowledge is characterized by infrequent and superficial interactions with the domain, where users have limited exposure and may lack a precise technical vocabulary or jargon. In contrast, high domain knowledge corresponds to frequent, hobby-like engagement, where users are more familiar with detailed attributes and exhibit a refined understanding. Medium knowledge falls between these extremes, indicating occasional but shallow interactions.

4 The RecQuest Data Collection Protocol

To study knowledge estimation, we require data that captures the nuanced conversational behaviors of users with varying levels of expertise—a resource absent in existing open-ended datasets. Current data collection protocols typically elicit unconstrained preferences, which tend to focus on popular, widely known items and allow users to accept recommendations with minimal scrutiny. In such settings, successful interaction does not require users to articulate domain-specific constraints, use lesser-known domain terminology, or explore trade-offs, making domain expertise largely irrelevant and consequently difficult to observe. To address this, we introduce RecQuest, a game-with-a-purpose protocol that incentivizes users to explore domain boundaries by assigning them a specific target item hidden within a background narrative. This section details the game mechanics, the item collection process across five distinct domains (bicycles, laptops, digital cameras, running shoes, and smartwatches), and the CRS architecture designed to support these goal-directed interactions.

4.1 Game Mechanics

Game Start. At the start, participants see a chat interface and a brief background story setting up a use scenario (Fig. 2). The story is generated from a specific item in the curated domain collection to anchor the conversation context without revealing the item itself. Each target item has two background-story variants: a concise (5 feature constraints, 80–100 words) and a more detailed version (10 feature constraints, 150–200 words); see Section 5.3 for details. This enables us to investigate whether the richness of contextual cues affects how participants formulate preferences or describe product features. Participants are instructed to find the item that best fits the scenario by engaging in a turn-based dialogue with the CRS.

Participant Turns. Participant turns are open-ended, allowing for various conversational strategies. Participants may state preferences, answer/ask questions, or critique recommendations. The ability to ask about an item's properties allows them to probe which constraints or features are satisfied—much like in real-world interactions between customers and salespeople. Feedback on recommendations is optional. Each recommendation includes an image and a brief system-generated explanation of its fit to the participant's stated needs. Participants may ask follow-up questions, refine their needs, or ignore the suggestion, using their own vocabulary.

A valid session requires the CRS to have produced at least two recommendations, which remain visible and can be selected at any point. Selecting an item triggers a confirmation prompt, asking whether they are certain of their choice, followed by a request for a brief feature-based justification. Participants are incentivized to identify the correct item by a monetary reward, fostering realistic behavior and increasing the external validity of the collected data.

CRS Turns. On each system turn, the CRS performs one of three possible actions: asking an elicitation question, answering a participant's question, or providing a recommendation. All actions are generated by a target-aware conversational agent guided by a mix of rule-based heuristics and LLM output.

- *Elicitation.* When a participant expresses several preferences at once, the CRS restricts them to at most three constraints. If more are listed, the system asks which ones are most important before continuing. This keeps early turns focused and prevents participants from giving the entire need description in a single turn.

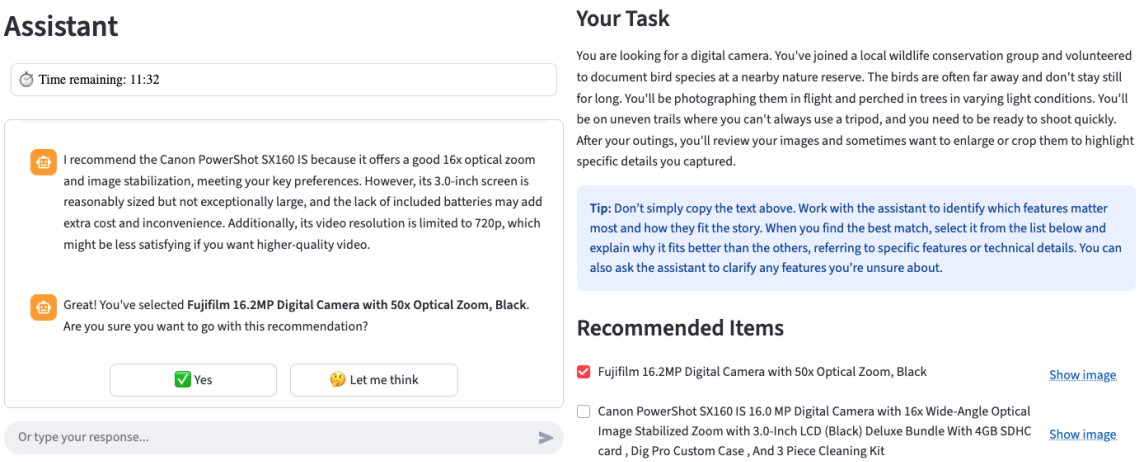

Figure 2: Screenshot of the game in action, showing a user who has obtained several recommendations during the conversation and is ready to make a final selection.

Elicitation questions are open-ended but specific to attributes present in the domain's schema (e.g., *"Is weight or range more important for your use?"*). Thus, they maintain engagement through incremental exchanges.

- *Answering questions.* When the participant asks about item attributes or domain facts, the CRS responds directly. All factual answers are LLM-generated and framed relative to the target item when applicable, so that information indirectly guides the user toward it without revealing the item's identity.
- *Recommendation.* A recommendation turn presents a single item from the domain's curated collection accompanied by an image and a short natural-language explanation. The explanation is generated through a two-step LLM procedure: (i) the system identifies key differences between the candidate item and the hidden target, and (ii) produces an explanation that highlights how those differences relate to the stated preferences (example in Fig. 2). The system may suggest a previously recommended item if new preferences align with it.

Game End. A session ends when the participant confirms a selection. The session length is a configurable parameter (limited to 15 minutes in our work). If the time expires, a soft grace period begins, allowing participants to finalize a selection from the recommendations already shown. On confirmation, participants provide a brief, feature-based justification, after which the dialogue concludes.

4.2 Item Collection

We curate item collections for each of the five domains used in RecQuest —bicycles, laptops, digital cameras, running shoes, and smartwatches—using data from the Amazon Review Dataset [15]. Each domain represents a semantically coherent item category with its own attribute structure, user intent patterns, and relevance criteria. This multi-domain setup allows us to evaluate knowledge estimation performance both within and across heterogeneous product types, providing a controlled yet diverse testbed for RecQuest. For every domain, we extract metadata, structured attributes, and item descriptions to create representations of available items. To ensure meaningful interaction, duplicates and highly similar items are removed using similarity-based filtering over item descriptions, resulting in diverse collections of distinguishable items.

4.3 CRS Architecture

Figure 3 shows the CRS architecture and data flow between its core components. At each participant's turn, the system processes the input, updates its internal state, and determines the next intent before producing a response.

Preference Summarizer. The CRS maintains a structured JSON-like representation of atomic preference statements. After each turn, an LLM-based preference summarizer updates this record by merging newly mentioned constraints with existing ones, ensuring all active needs are captured concisely.

Item Retrieval. A retrieval step is triggered when the preference record contains at least two statements or when new ones are added. Dense vector representations of items are pre-computed and held in memory. Current preferences are encoded to retrieve relevant candidates. If ≥ 3 preferences appear in a single turn, the system defers retrieval and asks the participant to prioritize which constraints matter most before continuing. This mechanism is designed to promote multi-turn interaction, encouraging the gradual refinement of needs through dialogue rather than attempting to complete the task in a single turn.

Decision Module. After preference updating and retrieval, the Decision Module applies a hybrid LLM and rule-based policy. It selects the next conversational intent based on the preference list, available recommendations, and the last message. It chooses one of four intents: redirect, answer, recommend, or elicit. *Redirect* is triggered when the dialogue drifts outside the designated domain. *Answer* is selected when the user asks factual or conversational questions. *Recommendation* is selected only if the LLM predicts recommendation and the following rule-based constraints are satisfied: retrieved

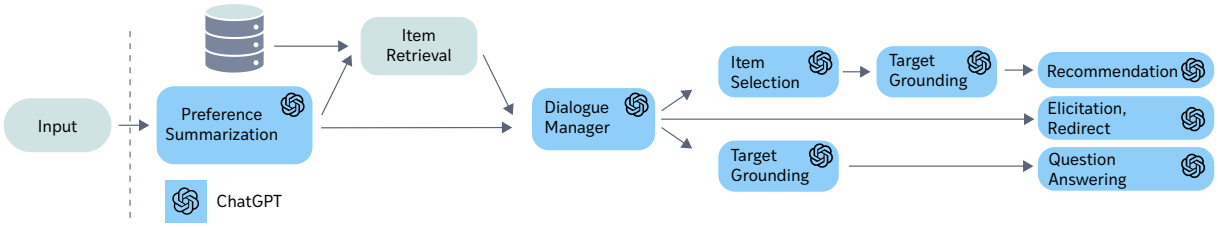

Figure 3: An overview of the CRS architecture, with the flow of data between different components.

candidates are available, the preference state contains new information, and at least three preferences have been stated; otherwise, the system falls back to *Elicitation* to gather additional constraints or clarify vague statements. Once the intent is selected, the system generates a response through specialized LLM modules.

Response Generation. *Recommendation* turns involve a multi-step process: an LLM selects the best candidate from the top-10 retrieved items given the current preferences and the latest exchange. A second model performs *target grounding* by comparing the selected candidate to the hidden target item, identifying which differences are relevant to the participant's stated needs. This ensures that explanations focus on meaningful contrasts. A third model generates the final natural language response that presents the recommended item and integrates the contrastive explanation. The *Answering* intent also uses target grounding, particularly for follow-up questions about recommended items. *Redirection* and *Elicitation* use a single generation step to produce a contextually appropriate response.

5 Data Collection

Using the RecQuest protocol, we conducted a comprehensive study to collect a dataset of dialogues rich in domain knowledge signals. Crucially, our goal was not to construct a large-scale training resource, but rather to validate the protocol and determine whether it effectively elicits observable signals of domain knowledge in conversation. The study followed a four-stage flow designed to balance participant expertise across domains. This section presents a descriptive analysis of the collected data as well as an assessment of the participants' subjective experiences to validate the protocol's engagement and naturalness.

5.1 Study Flow

Our study consists of four sequential stages: a screening questionnaire, a domain knowledge assessment, the main task, and an exit questionnaire. This flow is designed to balance participants across domains and expertise levels while collecting both self-reported and measured indicators of domain knowledge. While the self-reported knowledge categories used in our study offer structure, we acknowledge that the boundaries between them are soft, with many participants falling in the middle of the spectrum.

Screening. Participants first complete a brief self-assessment of their expertise corresponding to each of the five domains used in this work: bicycles, laptops, digital cameras, running shoes, and smartwatches. For each domain, they classify themselves into one of three categories: *novice*, *intermediate*, or *expert*. After the screening, they are assigned to the domain with the fewest participants for their expertise level at the time of assignment. Thus, the self-assessment provides a coarse estimate of knowledge, which is used to balance participation, ensuring that each expertise level is represented approximately equally across domains.

Domain Knowledge Questionnaire. Following the self-assessment, participants complete a short factual knowledge test in the domain to which they are assigned. This step provides us with the objective estimate of their true expertise. For each domain, we constructed a set of 30 factual statements using *ChatGPT 5.1*, guided by principles from Item Response Theory [1, 7]. In particular, the questions were designed to span a range of difficulties and to test both surface-level familiarity and deeper domain-specific knowledge. The questionnaire adopts the certainty-in-knowledge format proposed by Vidigal [36]. Each item presents a factual statement about the domain, and participants rate its correctness using one of five options: *Definitely True*, *Probably True*, *I don't know*, *Probably False*, or *Definitely False*. This captures both accuracy and confidence, and is more informative than binary correctness. To provide an additional validity check, human domain experts verified the suitability of the ChatGPT-generated questions.

Main Task. Participants are then presented with the instructions for RecQuest and perform the main task presented in Section 4.1. The assigned domain determines the item pool from which each participant is randomly assigned one of three target items and its corresponding background story. Each session lasts up to fifteen minutes, with the interaction process and system behavior governed by the game mechanics and dialogue flow previously outlined.

Exit questionnaire. After completing the main task, participants fill out an exit questionnaire in which they rate the quality of the interaction and provide open-ended feedback on their experience. These responses offer a subjective account of interaction satisfaction and help identify areas for improvement in both the system and the overall study design.

5.2 Participants

Participants were recruited from the Prolific platform,[2] residing in the US or Europe with English as first or primary language and had a $\geq$ 95% approval rate and $\geq$ 1000 previous submissions. Each participant could take part in only one task to avoid repeated exposure to the experimental setup. Participants completed the study fully online. After reading the task instructions, they proceeded through

[2] https://www.prolific.com/

Table 1: Descriptive statistics: number of dialogues (#Dial) and utterances (#Utt) are total counts, while values for the number of turns (#Turns), preferences (#Prefs), and recommendations (#Recs) are reported as mean ± standard deviation. Success rate (Success) is the proportion of dialogues where participants found the target item.

Domain	#Dial	#Utt	#Turns	#Prefs	#Recs	Success
Bicycle	79	1521	9.53 ± 4.71	10.00 ± 4.83	3.19 ± 1.35	67.9%
Digital Camera	79	1687	10.35 ± 4.42	10.76 ± 5.23	3.42 ± 1.72	29.1%
Laptop	98	2179	10.00 ± 4.86	10.47 ± 4.51	3.60 ± 1.27	29.6%
Running Shoes	179	3636	10.07 ± 5.37	9.90 ± 5.54	3.52 ± 1.43	37.1%
Smartwatch	80	1665	10.18 ± 3.63	9.61 ± 4.08	3.45 ± 1.19	37.5%
Total	**515**	**10688**	**10.21±4.81**	**10.10±5.03**	**3.46±1.41**	**39.3%**

the screening questionnaire, domain knowledge test, main task, and exit questionnaire. Sessions lasted for 18 minutes on average. Participants received an hourly wage of £8, in line with the platform's fair-wage guidelines, with a £1 bonus awarded on successful identification of the target items.

5.3 Generating Game States

Item Collection. To operationalize the RecQuest protocol, we instantiated item collections and target scenarios. From the Amazon dataset, we extracted all items belonging to categories related to the five domains. The raw collections ranged from 1,000 to 200,000 items per domain and included many near-duplicates differing only in brand or minor variations. To clean this data and maintain sufficient item differentiation, we narrowed each domain to a curated collection of 50 items. We achieved this by embedding item descriptions and applying k-means clustering to the embeddings, selecting the items nearest to the cluster centroids. The same embedding representations were later used for retrieval during gameplay.

Target Items. To create the interactive scenarios, we prepared a total of 30 background stories. For each domain, we selected three unique items from the curated collection and generated two versions of a story for each item: a *long* version and a *short* version. Each story was designed to contain a fixed number of implicit feature constraints—ten for the long version and five for the short—sufficient to uniquely identify the target item. Using an LLM, we converted these structured constraints into natural narratives without explicitly mentioning the corresponding item features.

LLM Usage. The CRS pipeline relies on multiple calls to an LLM, with different prompts and inputs for each component of the system. For all LLM calls in this study, we employed *gpt-4.1-mini-2025-04-14*, which early tests indicated offered a good balance between speed and output quality. All prompts used in the LLM-based components of the system can be found in the GitHub repository.

5.4 Analysis

We now address ***RQ1***, assessing whether the RecQuest protocol achieves its primary goal: eliciting rich signals of domain knowledge through natural, engaging conversation. To answer this, we evaluate the RecQuest protocol through two complementary lenses: a quantitative analysis of dialogue statistics to verify that the protocol successfully induces the iterative constraint refinement necessary to surface knowledge signals, and a qualitative analysis of participant feedback to validate that the resulting interactions were perceived as natural and engaging.

Quantitative Analysis: Interaction Depth. The aggregate dialogue statistics provide strong evidence that the protocol elicits rich interaction signals. Overall, we collected a total of 515 high-quality conversations totaling over 10,000 utterances across five domains. A statistical overview of the collected data is presented in Table 1. Crucially, dialogues average 10.21 turns and contain 10.10 preference statements. This density of preference expression, combined with an average of 3.46 recommendations per session, indicates that participants did not merely search for items, but engaged in iterative refinement over multiple exchanges. This confirms that the protocol successfully drives the articulation of detailed constraints, which serve as the proxies for domain knowledge in our analysis.

Qualitative Analysis: User Experience. To validate the "engaging and natural" aspect of RQ1, we examine the feedback from the post-task questionnaire. Responses were overall positive (mean rating of 3.89/5.0), with comments highlighting the educational and conversational nature of the system: *P13: "I appreciated how the chatbot patiently asked clarifying questions to understand my priorities, which made the recommendations feel personalized and relevant. The detailed comparisons between different e-bikes, including specific features like battery capacity, motor type, braking system, and handling, helped me make an informed decision*," and *P66: "The chatbot is very friendly and human like. I enjoyed every bit.*"

Negative comments primarily focused on the study constraints, such as the limited collection (e.g., *P396: "It could do with offering more [laptop] models.*") or the interaction constraint that prohibits information dumps (e.g., *P502: "was quite frustrating because it didn't take into account all of my wants all at once.*"). However, this latter frustration indirectly validates the protocol design: it confirms that the system successfully prevented keyword-search behaviors, forcing users to explore the domain through dialogue. Overall, the positive feedback on personalization and the fluid interaction confirms our protocol's effectiveness in creating an engaging experience at scale.

6 Estimating User Domain Knowledge

In this section, we begin by comparing participants' self-reported expertise with their actual performance on the domain-knowledge

questionnaires (**RQ2**). To move past the well-documented unreliability of self-assessments [9, 11], we investigate multiple strategies for grouping participants by objective scores on the knowledge questionnaires. We then examine behavioral patterns that distinguish experts from novices in real interaction settings (**RQ3**). Building on these validated ground-truth labels, we finally evaluate whether LLMs can infer user knowledge directly from dialogue transcripts (**RQ4**). Our approach uses both zero-shot and few-shot prompting to estimate overall knowledge levels, circumventing dependence on large training datasets while leveraging the natural language understanding capabilities of modern pre-trained models.

6.1 Methods

We estimate user domain knowledge directly from conversation transcripts using LLMs in zero-shot and few-shot settings. All methods share a unified prompt structure, shown in Fig. 4, comprising a fixed task description, a three-level expertise definition, placeholders for the domain, conversation history, optional few-shot examples, and a constrained JSON output format, to ensure consistency across settings. We evaluate three prompt variants. The *holistic* variant conditions on the full conversation history and produces a single overall knowledge estimate based on all user turns. The *evidence-based* variant also uses the full conversation but first instructs the model to extract domain-specific phrases introduced by the user and to ground its prediction explicitly in this extracted evidence. The *incremental* variant operates turn-by-turn: given the conversation history, a previous estimate, and the latest user utterance, the model updates the predicted knowledge level only when new evidence appears, with the restriction that the estimate may increase or remain stable but not decrease—a pragmatic design choice that prioritizes stability over perfect Bayesian updating. Each variant is evaluated in both zero-shot and few-shot settings, with few-shot prompts providing three domain-matched examples (one for each label). We additionally train a supervised fine-tuned model using the same prompt structure and annotated conversations.

6.2 Experimental Details

We evaluate two off-the-shelf LLMs for zero-shot and few-shot prompting: *gpt-4.1-mini-2025-04-14*—also used during data collection—and *Meta-Llama-3.3-70B*, an open-weight model. Supervised fine-tuning is performed on a 4bit quantized *Meta-Llama-3-8B-Instruct* model with QLoRA adapters and an encoder-only *ModernBERT-large* model suitable for a classification task, using 310 labeled conversations, with 205 held out for evaluation. Both models were trained for 10 epochs and $2e-5$ learning rate. All models are evaluated using identical prompts, and expertise definitions.

```
Role: You are an expert system for evaluating a user's domain knowledge in
**domain** domain based on their conversation with a recommender system.

Task:
Estimate the user's expertise level based on their utterances.

Knowledge Levels:
Level 1 (Novice): No independent technical knowledge.
Level 2 (Intermediate): Some correct use of domain terminology.
Level 3 (Expert): Demonstrates knowledge and independent use of domain
concepts.

Input Handling:
⟨if mode = holistic⟩
    You are given the entire conversation history.
    Assess all user turns jointly and produce a single overall estimate.
⟨elif mode = phrase-identification⟩
    You are given the entire conversation history.
    First, extract domain-specific phrases that were introduced by the user,
    then base the knowledge estimate on those phrases.
⟨elif mode = incremental⟩
    You are given the conversation history, a previous knowledge estimate,
    and the latest user utterance.
    Update the estimate only if the latest turn provides new evidence.
    The level may increase or remain the same, but never decrease.
⟨end if⟩

Input Data:
[conversation history]

⟨if mode = incremental⟩
    Previous estimate: [previous estimate]
    Latest user utterance: [latest user utterance]
⟨end if⟩

⟨if fewshot⟩
    [Examples]
⟨end if⟩

Format:
{
⟨if mode = phrase-identification⟩
    "user_initiated_phrases": [...],
⟨end if⟩
    "knowledge_level": <1, 2, or 3>,
    "reasoning": "<brief explanation>"
}

All responses must be valid JSON.
Output:
```

Figure 4: LLM prompt used for knowledge estimation. Placeholders for variables are in [square brackets], while conditional logic is marked by ⟨if⟩...⟨end if⟩.

7 Experimental Results

Here, we present our findings corresponding **RQ2**, **RQ3**, and **RQ4**.

7.1 RQ2: Establishing Ground Truth

To characterize user behavior accurately, we first require a reliable method for labeling expertise. We compare the reliability of subjective self-assessments against objective knowledge scores.

Figure 5 presents the distribution of item difficulties per domain, measured by the proportion of participants answering each question correctly. The spread is well-balanced: no item was prohibitively difficult (each was solved by at least 15% of participants), very few were trivial (solved by more than 90%), and most fell between 40% and 80%. This distribution indicates that the questionnaire is reasonably well-suited for distinguishing between different knowledge levels in our study setting.

We then examined how participants' self-reported expertise relates to their objective knowledge scores (Figure 6). The correlation is weak, with notable discrepancies—particularly among participants who identified as Intermediate or Expert yet performed poorly on the objective test. To empirically determine which metric serves as a more reliable ground truth, we compared task success rates under both labeling schemas. Intuitively, higher domain knowledge

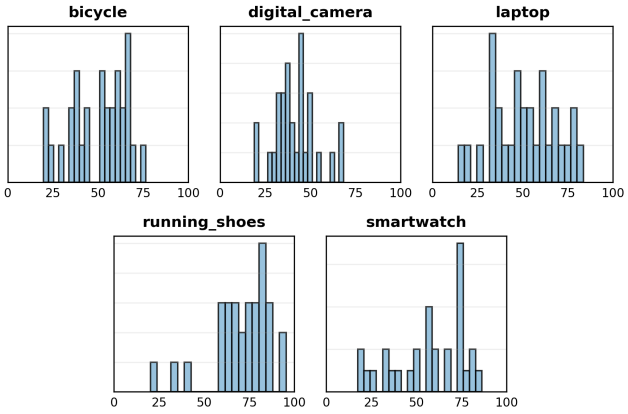

Figure 5: Distribution of questions by fraction of participants who correctly answered them by domain.

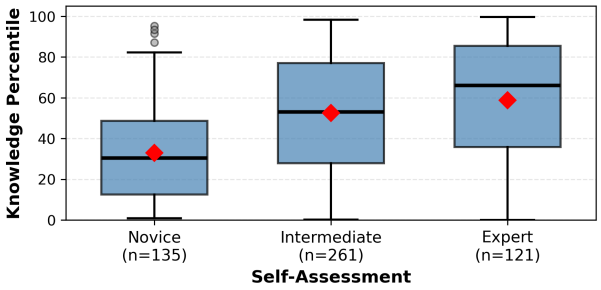

Figure 6: Relationship between objective and self-assessed knowledge.

should facilitate the search process and lead to better outcomes. Indeed, when using objective scoring, Experts outperformed Novices (40.4% vs. 35.9%). However, when grouping users by self-assessment, this trend paradoxically reversed: self-identified Novices achieved a higher success rate (44.4%) than self-identified Experts (36.4%). This inversion—where perceived expertise negatively correlates with actual task performance—provides compelling empirical evidence for the Dunning–Kruger phenomenon [23], where overconfidence masks a lack of competence.

This confirms that self-assessment is a noisy and misleading signal. In contrast, objective scores offer a more granular and verifiable basis for labeling. Consequently, for the remainder of our analysis (RQ3 and RQ4), we use objective score percentiles as an operationalized ground-truth proxy, defining **Novices** ($\leq$ 20th percentile) and **Experts** ($\geq$ 80th percentile).

7.2 RQ3: Behavioral and Linguistic Traces

Having established a robust ground truth, we examine how domain knowledge manifests within the conversation. Our goal is to determine whether experts and novices exhibit distinct behavioral patterns that are consistent enough to separate the two groups in practice. We analyze these interaction traces across three dimensions: dialogue flow, task outcomes, and linguistic specificity.

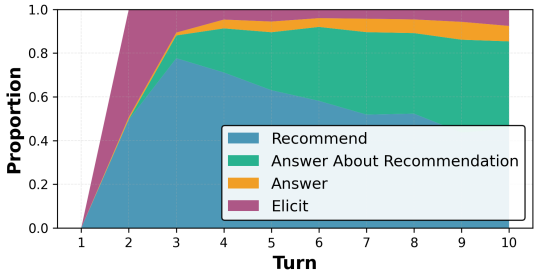

Figure 7: Dialogue intent progression.

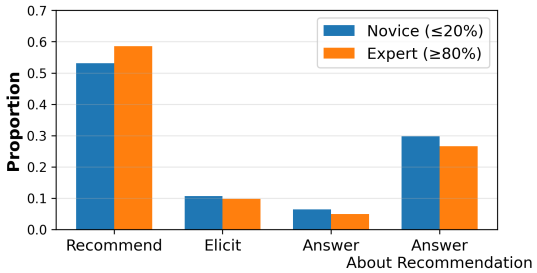

Figure 8: Distribution of dialogue acts.

Dialogue Patterns and Task Outcomes. The dialogue intent progression, (Fig. 7 shows that *Recommendation* dominates early interactions, peaking around the third turn before gradually declining, as *Answer About Recommendation* intents increase over time. Although all groups follow this general trajectory, we observe subtle differences. Experts receive slightly more recommendations, consistent with their more frequent preference updates and refinements. Novices, in contrast, ask more clarifying questions, indicating a greater need for guidance. Beyond the flow of conversation, we also analyze *success rate*: the percentage of participants who correctly identified the target item. Success varied substantially across domains, from 29.1% (running shoes) to 67.8% (bicycles), with an overall rate of 39.3%. Experts (40.4%) outperformed novices (35.9%), suggesting that domain knowledge helps users steer the system more effectively.

Linguistic Markers of Expertise. To further answer RQ3, we first analyze basic linguistic features, such as the number of turns and utterance length. Experts engaged in slightly longer dialogues (10.68 messages on average) and used longer utterances (17.3 words) than novices (9.87 messages, 14.0 words).

Next, we compute domain-specific TF-IDF scores to quantify differences in keyword usage between the novice and expert groups. For each domain, we construct joint TF-IDF models over all documents to ensure a shared vocabulary and comparable scaling. Then, we average TF-IDF values within each group. The resulting normalized differences indicate how strongly each keyword is associated with Experts versus Novices (ranging from -1 to +1).

Table 2 lists representative keywords highlighting the contrasts in keyword usage. Across domains, novices tend to use more general or evaluative terms (e.g., good, recommend, comfort), reflecting

Table 2: Discriminative keywords by domain and expertise.

Domain	Novice ($\leq 20th$ percentile)	Expert ($\geq 80th$ percentile)
Bicycle	recommend, local, new, need, stay	battery, range, assist, brake, gravel, trail, motor
Digital Camera	camera, bird, look, zoom, stability	lens, FujiFilm, autofocus, fast, travel
Laptop	best, game, work, time, speed, smooth	power, performance, light, windows, lenovo
Running Shoes	comfort, look, breathable	trail, Salomon, grip, protect
Smartwatch	smartwatch, good, brightness	battery life, charging, appearance

Table 3: Results of automatic knowledge estimation.

	Model	Accuracy	Macro-F1
Holistic	GPT-4.1 (zero-shot)	0.305	0.296
	GPT-4.1 (few-shot)	0.378	**0.363**
	LLaMA 3.3 (zero-shot)	0.354	0.354
	LLaMA 3.3 (few-shot)	**0.490**	0.357
Incremental	GPT-4.1 (zero-shot)	0.359	0.332
	GPT-4.1 (few-shot)	0.349	0.347
	LLaMA 3.3 (zero-shot)	0.364	0.347
	LLaMA 3.3 (few-shot)	**0.490**	**0.356**
Evidence-based	GPT-4.1 (zero-shot)	0.339	0.321
	GPT-4.1 (few-shot)	0.417	0.379
	LLaMA 3.3 (zero-shot)	**0.417**	**0.382**
	LLaMA 3.3 (few-shot)	0.213	0.154
Supervised	Llama 3-8B	0.519	0.286
	ModernBERT	**0.524**	**0.353**

subjective impressions or purchase intentions. Experts, in contrast, rely more on technical or domain-specific vocabulary (e.g., brake, autofocus, performance), emphasizing specifications and functional attributes. These patterns suggest that Experts approach the CRS with a more analytical, comparison-oriented communication style, while novices rely on broader, experience-based descriptions.

This TF-IDF analysis provides preliminary evidence for a clear hypothesis: *Experts tend to communicate through specific product attributes, while novices tend to communicate through intended usage or scenarios.* To test this hypothesis systematically, we designed an LLM-based annotation pipeline. We few-shot-prompted *gpt-4.1-mini-2025-04-14* to analyze each user utterance and extract phrases corresponding to two distinct categories: (i) *Specific Attributes*: Mentions of technical features, components, or specifications (e.g., "disc brakes"). (ii) *Intended Usage*: Descriptions of goals, contexts, or scenarios (e.g., "for daily rides to work"). This annotation moves beyond simple keyword counts and enables per-utterance quantification of communication style. The results strongly support our hypothesis, particularly at the conversation level. We found that Experts mentioned significantly more specific attributes on average (9.01 per conversation) than Novices (6.80). While Experts also provided more usage-based context (5.38 vs. 4.59), their overall linguistic preference still leaned more heavily toward attributes (62.6% of expressions) compared to Novices (59.8%).

7.3 RQ4: Automatic Knowledge Estimation

In this section, we address RQ4 by examining how effectively large language models can estimate user domain knowledge from conversational interactions.

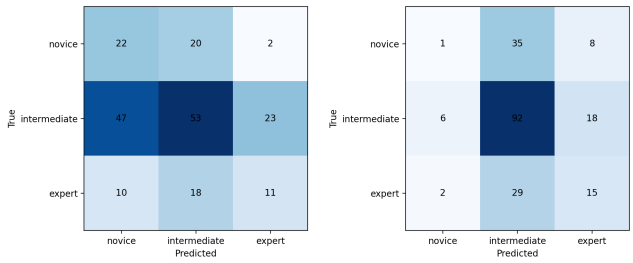

Figure 9: Confusion matrix comparison of best performing prompting model (Evidence-based, zero-shot LLama) left, and ModernBERT on the right.

Table 3 reports classification performance using both accuracy and macro-averaged metrics to capture global correctness and class-level balance. Zero-shot LLaMA consistently yields strong macro-F1 scores, indicating robust alignment across all expertise levels. We observe contrasting few-shot effects: adding in-context examples consistently improves GPT-4.1 but degrades LLaMA 3.3, particularly in the evidence-based setting where performance drops substantially. Among strategies, incremental estimation proves most effective. While the fine-tuned ModernBERT achieves the highest overall accuracy (0.524), it struggles with underrepresented categories, showing higher variability than the prompted LLMs. Ultimately, these moderate results highlight the inherent difficulty of the task: expertise cues are rarely explicit in single utterances, but rather distributed indirectly across multiple conversational turns.

To further understand the challenges in automatic knowledge estimation, Figure 9 contrasts confusion matrices for two of the best performing methods: zero-shot LLaMA and supervised ModernBERT. LLaMA shows distributed predictions with errors mostly confined to adjacent classes. Conversely, ModernBERT strongly favors the Intermediate class, maximizing overall accuracy at the expense of recall for Novices and Experts. This distinction explains the observed trade-off: ModernBERT over-centralizes predictions to optimize accuracy, whereas LLaMA maintains a better balance across the spectrum, resulting in higher macro-F1.

8 Conclusion and Future Directions

In this work, we introduce the novel task of estimating user domain knowledge from dialogue interactions, a crucial first step that enables the creation of adaptive conversational recommender systems. Our primary contribution is the design and validation of RecQuest, a gamified data collection protocol that successfully elicits natural, knowledge-rich interactions. By systematically creating information gaps between users and the system, RecQuest provokes the linguistic and behavioral traces necessary for studying domain knowledge in a CRS context. A key finding of our study is the necessity of objective knowledge assessment. Our comparison of self-reports versus objective scores revealed significant calibration biases, indicating that self-assessment is a noisy signal of expertise in this setting. Future research in this area must therefore rely on verified competence rather than user confidence. Finally, we established baselines for the automatic estimation of user knowledge. Our initial LLM-based experiments indicate that

while broad distinctions between novices and experts are possible, fine-grained estimation remains an inherently difficult task. The subtle and evolving nature of knowledge signals in short conversational turns poses a significant challenge for current models, and the current results should thus be interpreted as initial evidence of task feasibility rather than robust practical performance.

Our protocol and dataset provide a foundation for studying this challenge and supporting future CRS that adapt elicitation and explanation strategies to users with different levels of understanding. More broadly, our results suggest that robust estimation will likely require additional data, and that the protocol may need refinement depending on the target application domain and conversational context. Finally, an inherent limitation of our current study is the assumption that differences in conversational behavior stem primarily from domain expertise. We recognize that real-world interactions are heavily influenced by confounding factors like individual verbosity, communication style, language proficiency, and personality traits. Isolating true knowledge signals from these individual characteristics remains an important challenge for future work.

Acknowledgments

An unrestricted gift from Google partially supported this research.

References

[1] Frank B Baker and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques.* CRC press.

[2] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. 2022. Ready player one! eliciting diverse knowledge using a configurable game. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. 1709–1719.

[3] 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 '23)*. 2775–2785.

[4] Nolwenn Bernard, Ivica Kostric, and Krisztian Balog. 2024. IAI MovieBot 2.0: An Enhanced Research Platform with Trainable Neural Components and Transparent User Modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*. 1042–1045.

[5] Peter Brusilovsky and Eva Millán. 2007. User Models for Adaptive Hypermedia and Adaptive Educational Systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer, 3–53.

[6] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. 5016–5026.

[7] Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. 2016. Item response theory. *Annual Review of Statistics and Its Application* 3, 1 (2016), 297–321.

[8] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16)*. 815–824.

[9] Michael J Cole, Xiangmin Zhang, Jinging Liu, Chang Liu, Nicholas J Belkin, Ralf Bierig, and Jacek Gwizdka. 2010. Are self-assessments reliable indicators of topic knowledge? *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[10] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval (CHIIR '16)*. 163–172.

[11] Junhua Dang, Kevin M King, and Michael Inzlicht. 2020. Why are self-report and behavioral measures weakly correlated? *Trends in cognitive sciences* 24, 4 (2020), 267–269.

[12] Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. 2021. Identifying users' domain expertise from dialogues. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. 29–34.

[13] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 conference on human information interaction & retrieval (CHIIR '18)*. 2–11.

[14] 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 '20)*. 8142–8152.

[15] Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders. arXiv:2403.03952 [cs.IR]

[16] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2022. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5 (2022), 1–36.

[17] Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. De Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing Personal LAPS: LLM-Augmented Dialogue Construction for Personalized Multi-Session Conversational Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 796–806.

[18] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul 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 '19)*. 1951–1961.

[19] Julia Kiseleva, Alejandro Montes García, Jaap Kamps, and Nikita Spirin. 2015. The Impact of Technical Domain Expertise on Search Behavior and Task Outcome. arXiv:1512.07051 [cs.IR]

[20] Bart P. Knijnenburg and Martijn C. Willemsen. 2009. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In *Proceedings of the third ACM conference on Recommender systems (RecSys '09)*. 381–384.

[21] Ivica Kostric, Krisztian Balog, and Ujwal Gadiraju. 2025. Should We Tailor the Talk? Understanding the Impact of Conversational Styles on Preference Elicitation in Conversational Recommender Systems. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*. 164–173.

[22] Ivica Kostric, Krisztian Balog, and Filip Radlinski. 2024. Generating Usage-related Questions for Preference Elicitation in Conversational Recommender Systems. *ACM Transactions on Recommender Systems* 2, 2 (2024), 1–24.

[23] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121–1134.

[24] Jin Ha Lee, Allen Renear, and Linda C. Smith. 2006. Known-Item Search: Variations on a Concept. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2006), 1–17.

[25] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. 304–312.

[26] 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 (NIPS '18, Vol. 31)*. 9748–9758.

[27] 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 '23)*. 770–778.

[28] Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How Does Domain Expertise Affect Users' Search Interaction and Outcome in Exploratory Search? *ACM Transactions on Information Systems* 36, 4 (2018), 1–30.

[29] Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web (WWW '13)*. 897–908.

[30] Taehyung Noh, Haein Yeo, Myungjin Kim, and Kyungsik Han. 2023. A Study on User Perception and Experience Differences in Recommendation Results by Domain Expertise: The Case of Fashion Domains. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. 1–7.

[31] Dhanya Pramod and Prafulla Bafna. 2022. Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review. *Expert Systems with Applications* 203 (2022), 117539.

[32] 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 '19)*. 353–360.

[33] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *Proceedings of the AAAI conference on artificial intelligence (AAAI '20)*. 8689–8696.

[34] Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a Neural Conversational Agent with Dialogue Self-Play, Crowdsourcing and On-Line Reinforcement Learning. In *Proceedings of the 2018 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers) (NAACL '18). 41–51.

[35] Diana Tabatabai and Bruce M. Shore. 2005. How experts and novices search the Web. *Library & Information Science Research* 27, 2 (2005), 222–248.

[36] Robert Vidigal. 2025. Measuring Belief Certainty in Political Knowledge. *Political Behavior* 47, 2 (2025), 529–551.

[37] Luis Von Ahn. 2006. Games with a purpose. *Computer* 39, 6 (2006), 92–94.

[38] Ryen W. White, Susan Dumais, and Jaime Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*. 132–141.

[39] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 75–84.

[40] Gangyi Zhang, Chongming Gao, Wenqiang Lei, Xiaojie Guo, Shijun Li, Hongshen Chen, Zhuozhi Ding, Sulong Xu, and Lingfei Wu. 2025. Vague Preference Policy Learning for Conversational Recommendation. *ACM Transactions on Information Systems* 43, 3 (2025), 1–27.

[41] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. 2011. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*. 1225–1226.

[42] Xiangmin Zhang, Jingjing Liu, Michael Cole, and Nicholas Belkin. 2015. Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology* 66, 5 (2015), 980–1000.

[43] Canzhe Zhao, Tong Yu, Zhihui Xie, and Shuai Li. 2022. Knowledge-aware Conversational Preference Elicitation with Bandit Feedback. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. 483–492.

[44] 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 '20)*. 4128–4139.