InterQuest: A Mixed-Initiative Framework for Dynamic User Interest Modeling in Conversational Search

Yu Mei
Department of Computer Science and Technology
Tsinghua University
Beijing, China
meiy24@mails.tsinghua.edu.cn

Yuanxi Wang
Software Engineering
Nanjing University
Nanjing, China
221250067@smail.nju.edu.cn

Shiyi Wang
Tsinghua university
Academy of Arts & Design
Beijing, China
shiyi-wa23@mails.tsinghua.edu.cn

Qingyang Wan
Tsinghua University
Academy of Arts & Design
Beijing, China
wanqy23@mails.tsinghua.edu.cn

Zhuojun Li
Computer Science and Technology
Tsinghua University
Beijing, China
lizj23@mails.tsinghua.edu.cn

Chun Yu*
Department of Computer Science and Technology
Tsinghua University
Beijing, China
chunyu@tsinghua.edu.cn

Weinan Shi*†
Department of Computer Science and Technology
Tsinghua University
Beijing, China
swn@tsinghua.edu.cn

Yuanchun Shi‡*
Department of Computer science and Technology
Tsinghua University
Beijing, China
Qinghai University
Xining, China
shiyc@tsinghua.edu.cn

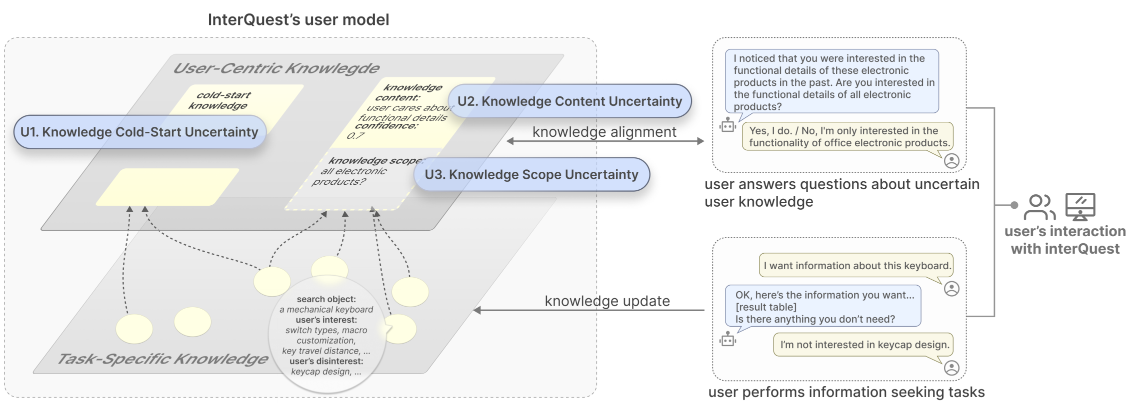

Figure 1: InterQuest continuously updates and refines the user model through proactive and passive interactions. As users perform information-seeking tasks, interaction data on their preferences (e.g., interest and disinterest in specific search objects) is integrated into task-specific knowledge. From this task-specific knowledge, InterQuest dynamically constructs *User-Centric knowledge*—cross-task, persistent attributes about the user's preferences across domains or scenarios (e.g., "cares about functionality details for all electronic products"). The construction of *User-Centric knowledge* involves three primary types of uncertainties: cold-start uncertainty (U1), content accuracy uncertainty (U2), and scope applicability uncertainty (U3). To address these uncertainties, InterQuest proactively asks the user targeted questions and refines the user model according to their responses, thus aligning its user model with the user's actual preferences.

Abstract

In online information-seeking tasks (e.g., for products and restaurants), users seek information that aligns with their individual preferences to make informed decisions. However, existing systems often struggle to infer users' implicit interests—unstated yet essential preference factors that directly impact decision quality. Our formative study reveals that *User-Centric Knowledge*—cross-task persistent preference attributes of users (e.g., "user cares about functionality details for electronics")—serves as a key indicator for resolving users' implicit interests. However, constructing such knowledge from task-specific data alone is insufficient due to three types of uncertainties—cold-start limitation, content accuracy, and scope applicability—which require user-provided information for knowledge alignment. Based on these insights, we present InterQuest, an LLM-based conversational search agent that dynamically models user interests. InterQuest combines two strategies: (1) *Dynamic*

*Also with Key Laboratory of Pervasive Computing, Ministry of Education.
†Corresponding author.
‡Also with Intelligent Computing and Application Laboratory of Qinghai Province, Qinghai University.

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747753

User Knowledge Modeling, which infers and adjusts the content and scope of *User-Centric Knowledge*, and (2) *Uncertainty-Driven Questioning*, where InterQuest proactively asks questions to resolve knowledge uncertainties. A user study with 18 participants demonstrates that InterQuest outperforms the baselines in user interest inference, accuracy of user knowledge modeling, and the overall information-seeking experience. Additionally, our findings provide valuable design implications for improving mixed-initiative user modeling in future systems.

CCS Concepts

• **Human-centered computing → Interactive systems and tools**; • **Information systems → Personalization**.

Keywords

User Modeling, Conversational Search, Human-AI Interaction

ACM Reference Format:
Yu Mei, Yuanxi Wang, Shiyi Wang, Qingyang Wan, Zhuojun Li, Chun Yu, Weinan Shi, and Yuanchun Shi. 2025. InterQuest: A Mixed-Initiative Framework for Dynamic User Interest Modeling in Conversational Search. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3746059.3747753

1 INTRODUCTION

Users often seek online information to make informed decisions, such as evaluating products, restaurants, or tour groups. Unlike fact-based searches, these tasks require users to collect and synthesize multifaceted information to assess how well each option meets their personal preferences [40, 88, 101–103]. However, users typically process these decisions internally while browsing, lacking the experience to externalize their cognitive task models [46, 114]. This often leads to challenges when interacting with intelligent systems, as users may struggle to articulate their needs clearly or fully express their intentions [5, 6, 110]. As a result, only a portion of their information needs are explicitly expressed through search queries, while other interests remain implicit and unspoken.

Current systems often fail to resolve these implicit interests [72, 76, 84, 112]. This issue arises because inferring implicit intents can lead to multiple interpretations with limited task context. For example, when a user searches for "*help me find information about a mechanical keyboard*," one user may prioritize user reviews on delivery speed and product quality. At the same time, another may focus on functionality details like switch types or macro customization.

Our formative study drew insights from how human wizards infer users' implicit information interests. We discovered that humans often incorporate ***User-Centric Knowledge***—persistent, cross-task preference attributes (e.g., "*the user cares about functionality details for all electronic products*")—to guide their reasoning. By leveraging such knowledge, humans can infer users' information interests for a specific task (e.g., "*the user may care about DPI sensitivity or programmable buttons for a certain gaming mouse*"). Also, unlike static user profiles, this knowledge exhibits flexible and dynamic boundaries of applicability (referred to as "scope"). However, constructing such knowledge introduces three uncertainty types: (1) *knowledge cold-start* during the initiation phase, (2) *content uncertainty* in preference accuracy, and (3) *scope uncertainty* in preference applicability boundaries. Human wizards mitigate these through strategic questioning, demonstrating the need for knowledge alignment through user-provided information.

Based on these insights, we propose an LLM-based conversational search system named InterQuest, which dynamically models user interests to infer users' implicit information interests. InterQuest combines two key strategies: (1) ***Dynamic User Knowledge Modeling***, which dynamically derives and adjusts *User-Centric Knowledge*. In the interaction stage, it infers *User-Centric Knowledge* from task-specific data and quantifies *knowledge content* and *scope* uncertainty. (2) ***Uncertainty-Driven Questioning***, where InterQuest asks questions to resolve knowledge uncertainties and refine *User-Centric Knowledge*. InterQuest identifies uncertainties in relevant user knowledge during each task and prioritizes questions with the greatest potential to impact task outcomes. It then asks corresponding questions to align the target knowledge with the user's actual preferences.

To evaluate InterQuest's effectiveness, we conducted a within-subject study with 18 participants. The participants used Baseline 1 (LLM as the recommender, rule-based QA), Baseline 2 (*Dynamic User Knowledge Modeling*, rule-based QA), and the InterQuest (*Dynamic User Knowledge Modeling*, *Uncertainty-Driven Questioning*) system to seek information for assigned task sets (products, restaurants, or tour groups). Results showed that InterQuest outperforms baseline methods in user interest inference accuracy, subjective search results ratings, question answering experience, and *User-Centric Knowledge* modeling accuracy. Moreover, *Dynamic User Knowledge Modeling* and *Uncertainty-Driven Questioning* strategies both contribute to InterQuest's improved interest inference performance. We further discussed directions for effective granularity-aware user knowledge modeling and selecting the target knowledge for user alignment.

The contributions of our research are as follows:

- Through formative study, we identified the flexible scope structure of *User-Centric Knowledge* in information-seeking tasks. We also categorized the types of uncertainty that arise when constructing such knowledge from task-specific data.
- We present InterQuest, an LLM-based conversational search agent that dynamically models user intent to infer implicit information needs. It integrates two key strategies: (1) *Dynamic User Knowledge Modeling* and (2) *Uncertainty-Driven Questioning*.
- We validated InterQuest's effectiveness through a user study with 18 participants and provided design implications to inform future approaches to mixed-initiative user modeling.

2 RELATED WORK

In this section, we review prior research concerning (1) online information-seeking task support, (2) user interest modeling in search systems, and (3) interactive user modeling methods, which are closely related to our study.

2.1 Online Information-seeking Tasks

In online information-seeking, users rely on digital information such as reviews to evaluate how well each option meets their criteria [17]. These tasks not only involve processing a vast amount of messy information [73] but also impose a significant cognitive burden [7, 8, 22, 77]. Typically, users employ a two-stage process: (1) an initial screening of all options to determine that they are worthy of further consideration; (2) a detailed comparison of the selected options [30]. Recent research has provided support for both stages.

Support for the first stage has primarily focused on recommender systems, which rank items based on predicted user interest and offer a personalized view of search results [2, 69]. Common strategies include incorporating implicit or explicit user feedback [36, 89, 95] and personalized query rewriting [13, 27, 31].

Support for the second stage is manifold. To minimize operational demands, the primary approach involves automatically collecting or completing specific information elements [9, 10]. To reduce cognitive load, studies often provide machine-generated suggestions, summaries, ratings, or exploratory questions [41, 82, 93, 118]. To alleviate difficulties in information management, existing work has developed innovative interactive interfaces that reorganize information or facilitate diverse data exploration methods [35, 39, 44, 49, 62, 63, 71, 93] .

However, concerning the long-term learning of user interests, existing research primarily focuses on the first phase, such as recommender systems. The second stage, catering to the diverse personalized information needs of different users [9, 10], needs more support. Our work aims to understand users' personalized information acquisition habits during the second stage, enhancing search efficiency and decision quality.

2.2 User Modeling In Search

User modeling is central to search systems, with existing studies exploring various paradigms [79]. This work relates closely to two major directions: hierarchical user modeling and LLM-based natural language modeling.

Hierarchical models have been widely adopted to capture multi-granular user interests [79]. These methods enable real-time representations at different abstraction levels, improving understanding of short- and long-term preferences [56, 99, 100, 107]. For example, HieRec [80] models interests across subtopics, topics, and user levels; HUP [28] captures dynamic interest shifts through behavior types and micro-interactions; HUSTM [4] incorporates emotional cues. Such approaches offer more structured and fine-grained user representations. Meanwhile, the emergence of LLMs has driven a shift toward language-based user modeling [86, 92], typically following two paradigms: feature augmentation and generative modeling. Feature-based methods use LLMs to enhance traditional representations [25, 58], especially in domains like news [65, 104] and social media [115]. Generative approaches treat user behavior as input sequences and generate recommendations via prompts [60, 96]. Recent work further explores conversational recommendation with LLMs, e.g., Chat-REC [24] encodes user profiles and interactions into language input to generate dialogue-based responses [24, 57]. These methods demonstrate LLMs' effectiveness in capturing behavior semantics and enabling more context-aware modeling [25, 75, 105].

Despite progress, two key challenges remain. First, hierarchical models often lack interpretability and natural language expressiveness [28, 80], limiting alignment with user cognition. Second, most methods adopt fixed-layer structures: LLM-based models typically capture single-level semantics [24, 58], while hierarchical models predefine a limited number of semantic layers, struggling with dynamic user behaviors such as task switching or open-ended exploration [28, 79, 80].

To address these gaps, we propose a dynamic user modeling framework that integrates task behavior with user knowledge. Our approach extracts task-level knowledge from interactions while modeling cognitive-level user knowledge to simulate human understanding. Based on formative studies, we observe that users' knowledge boundaries are dynamic and uncertain, leading to our design of a cognitive boundary management mechanism that adapts to users' blind spots.

2.3 Interactive User Modeling

In interactive user modeling, many approaches rely on users manually reviewing and editing system-generated user profiles [74, 83]. For example, Radlinski et al. proposed constructing user profiles using natural language descriptions [83], while LACE represents user interests as a set of human-readable concepts that users can directly edit to influence recommendation outcomes [74]. Although such methods provide users with a certain degree of control, they require users to fully understand and manipulate the profile contents [26], resulting in high interaction costs and significant cognitive load. To reduce this burden, recent studies have explored more proactive interaction paradigms where systems engage users through active questioning [18, 19, 34, 52, 53]. One typical approach is the User Preference Elicitation, which elicits user preferences through multi-turn question-answering [47], with representative methods such as UNICORN [19] and MCMIPL [116]. Another line of work focuses on Clarification Questioning, which identifies ambiguities in queries during search or QA and generates clarifying questions to refine user intent [12], as seen in systems like ClariQ [3] and UniPCQA [20]. Although these methods improve interaction efficiency in specific tasks, most of them rely heavily on the current task context for question generation [3, 14, 20, 23, 54], lacking the ability to model users' cognitive structures [50], and thus struggle to support cross-task knowledge accumulation and long-term alignment [33, 97].

To address this limitation, we propose a proactive questioning framework targeting user cognitive gaps. Unlike conventional approaches focusing solely on task-specific details, our framework centers on users' cognitive structures, enabling knowledge generalization and sustained modeling across tasks. It then performs targeted cognitive calibration to fill reasoning gaps with minimal user effort.

3 FORMATIVE STUDY

To inform the design of InterQuest, we conducted a formative study with 18 participants. The formative study comprised (1) a collaborative information-seeking experiment, which provided insights into how human wizards infer users' implicit information interests, and (2) a semi-structured interview to understand the strategies humans use in user interest inference and the challenges they face during the process.

3.1 Setup

3.1.1 Protocol. The formative study consisted of two parts: a two-participant collaborative information-seeking experiment (30–50 minutes) and a semi-structured interview for each participant (20 minutes).

In the collaborative information-seeking experiment, participants were randomly paired: one took on the ***wizard*** role, and the other took on the ***user*** role. We predefined three sets of tasks, each focusing on a different topic (product, restaurant, or tour group), with 12 search items per set. Each pair of participants was randomly assigned to one of these task sets.

The user initially provided his or her information interests in the first six search items by filling in a questionnaire. The wizard's task was to infer the user's preferences for the 7th to 12th search items based on the earlier provided information. Throughout the process, the wizard was required to think aloud, articulating the reasoning behind their inferences. If the wizard was uncertain, they could ask the user questions for clarification, but the question could not directly reflect the inference result (e.g., "*What information are you interested in for this item?*"). At the end, the user evaluated the wizard's inferences based on their actual preferences and provided feedback on any missing inferences.

In the semi-structured interview, the wizard and the user were interviewed separately. The wizard was asked about their reasoning strategies, the information they found useful, the uncertainties they encountered during inference, and their strategies for actively seeking information from the user. Meanwhile, the user was asked about their overall experience with the task, their feelings toward the wizard's active questioning, and their suggestions for improving the inference process.

3.1.2 Participants. 18 participants (6 males and 12 females) were recruited via social media. Each participant had experience using mobile apps or computer websites to search for at least one of the following: products, restaurants, or tour groups. Their ages ranged from 19 to 35 ($M = 23.78, SD = 4.08$). All participants were either university students or held a bachelor's degree. In the collaborative information-seeking experiment, participants were paired with strangers, ensuring no prior knowledge of each other's preferences. Each participant received a compensation of $30 upon completion of the experiment.

3.1.3 Analysis. The sessions were audio-recorded and transcribed for analysis. For the collaborative information-seeking experiment, three experienced HCI researchers qualitatively coded the transcript based on the following criteria: (1) the wizard's reasoning chain during inference, including the historical information involved, the intermediate reasoning steps, the reasoning results, and the correctness of the results; (2) the questions asked by the wizard to the user, the basis for those questions, the reasoning effects after the questions, the types of uncertainty in the questions, and the format of the questions. Based on the above data, we report our findings.

3.2 Findings

3.2.1 Highly Personalized User Interest and User-Centric Knowledge Driven Inference. Our collaborative information-seeking experiment reveals that different participants exhibit significantly different information interests for the same task set. For example, for yogurt in the product task set, P2 focused on "ingredients, production process, hygiene, flavor, and health," while P6 focused on "logistics speed, packaging, user comments, and cost-effectiveness." This aligns with prior observations that users may have completely different intentions for the same query [48, 81] .

However, most participants reported difficulties in fully articulating their intentions in information-seeking tasks. We found that participants faced three main challenges: (1) initial ambiguity (4/9, P12: "*At first, I might only think of some basic things... It's only after seeing certain information that I make further associations.*"); (2) difficulty in recalling comprehensive information (6/9, P10: "*Some important points might not be the first thing I think of.*"); and (3) challenges in language organization (5/9, P16: "*When describing, I might use the same word to summarize several points of focus, which can lead to confusion.*").

Despite these challenges, all 9 wizards agreed during the interview that historical interest points can significantly help wizards infer users' information interests for the current task. During inference, different wizards may follow different reasoning chains. However, most reasoning chains contain sub-chains like **"task-specific knowledge → inferred *User-Centric Knowledge* → predicted information interest for the current task."**

Task-specific knowledge refers to preferences specific to a certain search item (e.g., "*P2 cares about intelligent monitoring metrics, screen resolution for a smartwatch*"). *User-Centric Knowledge*, on the other hand, refers to persistent user preferences that span across tasks and domains (e.g., "*P2 cares about functionality for electronics*").

In actual practice, *User-Centric Knowledge* demonstrated a high degree of semantic richness, encompassing a wide range of concepts and details. It can be the user's preference for low-level details, such as fine-grained information points. It can also include high-level abstractions like general information dimensions. Additionally, users may struggle to distinguish between closely related concepts, indicating that knowledge boundaries are fluid and interconnected.

Besides, in the experiments, *User-Centric Knowledge* may originate from either explicit user statements or implicit user concerns. For example, P5 deduced that: "*The user cares about the seven-day unconditional return policy for headphones and pillows → The user may be concerned about the flexibility of return policies for all purchased products*", reflecting explicit preference interpretation. On the other hand, P1 inferred: "*The user pays attention to dark chocolate ingredients, Greek yogurt food safety reports, and the shelf life of instant coffee → The user is likely highly concerned about food safety*", which indicates an implicit concern.

3.2.2 Dynamic Scope Structure of User-Centric Knowledge. The coded transcripts revealed that the applicable scope of *User-Centric Knowledge* was highly flexible and varied across different contexts.

Table 1: Types of knowledge scope observed in formative study, with examples from product, restaurant and tour group domain.

Global	Category-Specific	Attribute-Based
all products	daily necessities, electronics, furniture	consumables, products that contact with the skin
all restaurants	seafood restaurants, cafes, casual dining	restaurants serving raw food, budget restaurants
all tour groups	domestic tours, international tours	tours more than a week, tours with many destinations

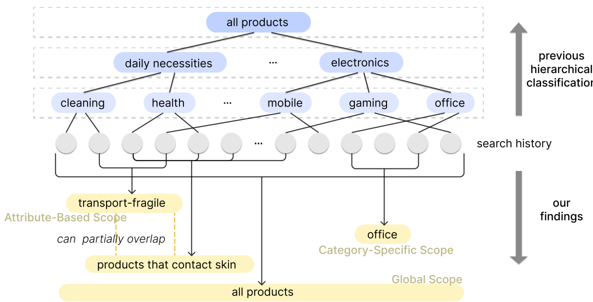

Figure 2: Many previous classifications divide user knowledge into a limited hierarchical structure. In contrast, our findings revealed that the structure of user knowledge is non-hierarchical and fluid. The boundaries of different types of knowledge can partially overlap without a clear parent-child relationship.

Rather than limiting *User-Centric Knowledge* to a single search item, wizards often generalized it to broader or more nuanced scopes. Specifically:

- Global Scope: 9/9 wizards inferred user preferences that applied universally across all products, restaurants, or tour groups. For example, P13 inferred that "*the user is interested in accommodations and transportation for all tour groups*."
- Category-Specific Scope: 7/9 wizards identified user preferences within well-defined categories, such as electronic products (P1, P3), international tours (P13, P17), or seafood restaurants (P7). Here, we pre-collected common, well-defined categories from websites used in this study and used them as the basis for judgment.
- Attribute-Based Scope: 9/9 wizards inferred preferences tied to specific attributes, such as items that come into contact with the skin (P3:"*the user is concerned about whether the ingredients or materials of products that come into contact with the skin are safe*").

Throughout the collaborative information-seeking process, wizards dynamically adjusted the scope of inferred knowledge based on task demands and contextual cues. Examples of different scope types in the experiments are summarized in Table 1.

In addition, the formative study revealed user knowledge's non-hierarchical and non-discrete nature, which cannot be easily categorized into a finite tree-like structure, as shown in Figure 2. In the experimental data, we found that users often viewed categories as overlapping or fluid, with attributes that could not be neatly placed within a defined category. For instance, while "food" is clearly part of the product category, the user categorizes food by attributes like "cold chain" or "over $50" " which showcases a more nuanced understanding beyond simple hierarchical labels. This suggests that interactive systems need to accommodate the dynamic and context-sensitive nature rather than rely on static, hierarchical models.

3.2.3 Uncertainties in Construction of User-Centric Knowledge and Active Learning Strategies. Three experts analyzed all the questions posted by the wizard to the user during the experiment, examined the uncertainties in each question, and identified three main types of uncertainty in the construction of *User-Centric Knowledge*: cold-start uncertainty, content uncertainty, and scope uncertainty.

In the initiation phase, wizards often encountered **cold-start uncertainty**. This primarily stemmed from two factors: insufficient historical sample data and lack of key context for reasoning (P5: "*There haven't been any products related to this, so I have to make an assumption*"), and a lack of user preferences, which were difficult to capture systematically (P9: "*Some users may not be willing to try certain things, and I can't know this directly*").

In the interaction phase, wizards faced content uncertainty and scope uncertainty.

Content uncertainty occurred in three main situations: First, when sample data was insufficient, wizards doubted the accuracy of their reasoning (p17: "*The sample size is too small, and the validity of the information is limited*"). Second, there could be multiple valid reasoning paths (P11: "*The user's preference could be explained by either the dish's uniqueness or its variety; both interpretations seem reasonable*"). Lastly, ambiguities in user language affected the wizard's reasoning (P5: "*Sometimes users use vague terms like 'small area'; I can't tell if they mean the keycap area or the entire keyboard area*").

Scope uncertainty arose when the wizard was unsure whether the knowledge applied broadly from just a few examples (P3: "*The user is concerned about the material of thermos cups and the filling of pillows, but I'm not sure if they would care about the material of keyboards*"), or whether it could transfer to other categories (P11: "*I'm not sure if the user's concerns about Japanese cuisine are similar to those about Chinese restaurants*").

In the interviews, all 18 participants agreed that active questioning effectively reduced the above uncertainties. 13 out of 18 participants reported that closed-ended questions were generally more effective, although they valued the opportunity to provide additional details when these questions did not fully address their needs. Additionally, 12 out of 18 participants preferred indirect questions over direct ones as they allowed for more nuanced responses.

All 9 participants playing the user role expressed openness to the frequency of questions during the task, believing it facilitated information expression and system understanding. During the cold-start phase, participants were generally open to a higher frequency

Table 2: Three primary types of uncertainty during the construction of User-Centric Knowledge.

Type	Timing	Definition	Example Question
Cold-start Uncertainty	Initiation phase	Insufficient data to make informed inferences	"What attributes do you prioritize when purchasing a product?"
Content Uncertainty	Interaction phase	Uncertainty about whether the user has the inferred preference	"You care about the sugar content in yogurt, is it because you're into fitness?"
Scope Uncertainty	Interaction phase	Uncertainty about the boundaries of preference applicability	"You mentioned caring about the reputation of electronics sellers; are you also interested in the reputation of sellers for daily essentials?"

of questions. Some participants were comfortable with 10-15 questions ("*If it's a product, I could probably answer 10 to 15 questions*"), while others were willing to accept more than 20 questions ("*I could handle quite a few questions, around 20 or 30 without problem*"). In addition to task history, users felt that demographic, behavioral, and cultural background information helped refine predictions.

Privacy protection was also a key factor influencing user acceptance. Some participants indicated that they were willing to answer more questions if the system ensured the security of their personal data ("*If they protect it well, I don't mind answering any questions*"). However, excessively detailed or personal questions could provoke resistance ("*I can accept up to five questions, as long as they don't involve personal privacy*").

4 DESIGN AND IMPLEMENTATION OF INTERQUEST

We present InterQuest, an LLM-based conversational search agent that dynamically models user interests to infer users' implicit information interests. In this section, we first outline the overall framework of InterQuest. Subsequently, we discuss two key strategies employed by InterQuest: (1) ***Dynamic User Knowledge Modeling***, which continuously infers and adjusts the content and scope of *User-Centric Knowledge*, and (2) ***Uncertainty-Driven Questioning***, which proactively asks questions to resolve knowledge uncertainties.

4.1 InterQuest System Framework

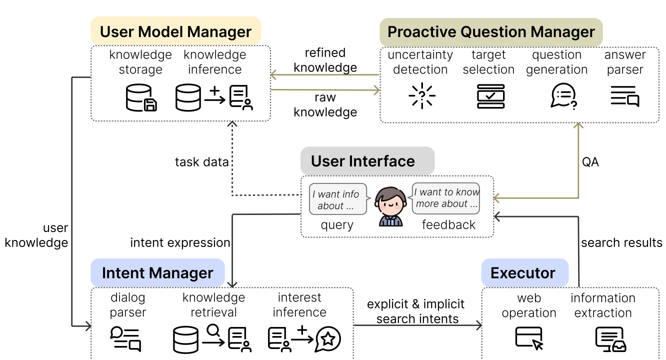

Figure 3: The overall framework of InterQuest system.

We present the overall framework of InterQuest (see Figure 3), which consists of four key components: the intent manager, executor, user model manager, and proactive question manager.

The main workflow (shown by the black lines in Figure 3) helps users complete information-seeking tasks. The process begins when the user submits a search query, which is parsed by the intent manager. The intent manager then infers implicit intents with support from the user model manager. The executor then carries out the information-seeking process. It automates web operations and extracts web data. At the end of the task, users can provide feedback to either search again or end the task. The user model manager also updates its knowledge storage based on the session's interaction history.

The framework also incorporates an *Uncertainty-Driven Questioning* process (indicated by the beige-gray lines in Figure 3). This process enables the system to proactively ask users questions based on knowledge uncertainty, thereby refining the knowledge in the user model manager. A detailed description of its design can be found in Section 4.3.

4.2 Dynamic User Knowledge Modeling

Dynamic User Knowledge Modeling aims to model *User-Centric Knowledge*—the cross-task persistent preference attributes of users. Our formative study reveals that the scope of *User-Centric Knowledge* is highly flexible, non-hierarchical, and can adapt dynamically across tasks. Additionally, the construction of *User-Centric Knowledge* inherently involves uncertainties. Based on these insights, we developed a dynamic, non-hierarchical user model for InterQuest.

4.2.1 User-Centric Knowledge Inference. Inspired by recent studies on the emerging capabilities of LLMs in text-based user modeling [64, 85, 108], we developed a natural language-based user knowledge representation for InterQuest. For *User-Centric Knowledge*, we record the content, scope, and the confidence associated with each of them. Additionally, interaction data from historical tasks is stored as task-specific knowledge.

In the initiation stage, to address cold start issues, InterQuest prompts users to select the common search goals and answer a limited number of multiple-choice questions related to these goals (an example of which can be found in Appendix B). Based on their responses, InterQuest conducts an initial knowledge inference.

In the interaction stage, new task data triggers an incremental update of *User-Centric Knowledge*. This process consists of two steps: 1) Updating existing knowledge. InterQuest matches the new task data to related existing *User-Centric Knowledge*. Confidence

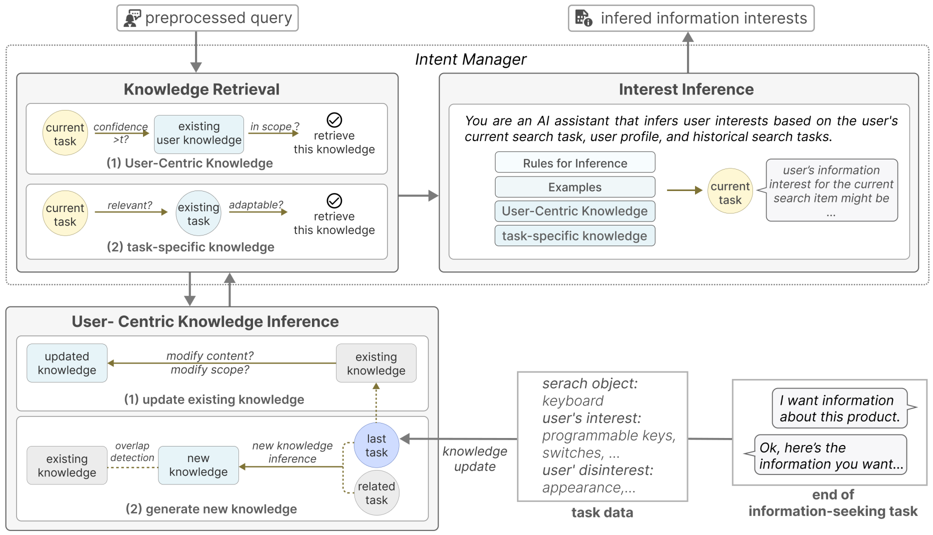

Figure 4: The pipeline of *Dynamic User Knowledge Modeling*. InterQuest dynamically constructs *User-Centric Knowledge* from task data by (1) updating existing knowledge and (2) generating new insights. During the user interest inference phase, InterQuest retrieves both *User-Centric Knowledge* and task-specific knowledge to support LLM-based intent inference.

is increased if the new task supports the content and scope of the existing knowledge. Otherwise, it determines whether to modify the content or scope based on other historical tasks. 2) Generating new knowledge. InterQuest uses semantic distance to identify tasks related to the new task data and infers new *User-Centric Knowledge*. If the content of the new knowledge overlaps with existing knowledge beyond a certain threshold, it is adopted. In practice, we empirically set this threshold to 0.6.

The analysis and inference process described above primarily relies on LLMs, with some relevant prompts provided in the Appendix A.2 and A.4.

4.2.2 Knowledge Confidence Measurement. This section describes the methods for measuring confidence of inferred content and scope within *User-Centric Knowledge* Inference.

To evaluate the credibility of knowledge generated by black-box LLM APIs, we focus on estimating the confidence in the knowledge inference process at a cognitive level rather than relying solely on LLM token logits for generation probability. We adopt established self-evaluation methods [11, 38], where LLMs generate their own confidence estimates through self-reflection. Additionally, we enhance this process by incorporating Chain-of-Thought (CoT) reasoning through a structured evaluation framework.

The LLM reasoning chain for evaluating the confidence of the inferred **knowledge content** consists of the following steps: (1) *Task coverage.* InterQuest first identifies all relevant user tasks for the inferred knowledge content and calculates the coverage ratio: $coverage = \frac{T_{match}}{T_{total}}$. Here, T_{match} is the number of matched tasks, and T_{total} is the total number of tasks. (2) *Evidence strength.* Following methods used in [32, 68], InterQuest searches for relevant task evidence that can validate the inferred content. For each piece of evidence retrieved, its contribution to supporting the knowledge is assessed based on clarity, directness, and consistency. (3) *Knowledge specificity.* InterQuest reflects on how specific the knowledge content is - whether it is targeted at a specific sub-domain or if it remains overly general. This ensures that claims are not too vague or broad.

The LLM reasoning chain for measuring the confidence of inferred **knowledge scope** includes: (1) *Task coverage.* Similar to the previous step, InterQuest identifies all tasks impacted or covered by the scope of the knowledge inference and calculates the coverage ratio. (2) *Category consistency.* It assesses whether the inferred knowledge remains consistent within a specific category, ensuring that it is not fragmented across unrelated domains.

For the related prompts of the above reasoning process, see the Appendix A.3.

Finally, we compute an integrated confidence score by calculating a weighted average of the factors involved (*Task coverage, Evidence strength, Knowledge specificity* for inferred content; *Task coverage, Category consistency* for inferred scope). Here, we empirically set equal weights for each aspect. These weights can be adjusted in the future based on task-specific requirements or further empirical validation.

4.2.3 Knowledge Retrieval for User's Intent Inference. InterQuest retrieves relevant knowledge to infer the user's intent by combining task-specific and *User-Centric Knowledge*. It first identifies tasks similar to the current one by measuring the semantic similarity between the textual sequences formed from the historical task data. Then, it uses LLMs to assess whether information from these tasks can be adapted for the current task. For *User-Centric Knowledge*, InterQuest checks if the knowledge scope includes the current search item. If this condition is met, the system evaluates whether content and scope confidence exceed a predefined threshold. If both conditions are met, the knowledge is considered relevant, trustworthy, and extracted for inference. In practice, we empirically set this threshold to 0.5.

The retrieved knowledge is then leveraged to infer the user's intent. Using LLMs, InterQuest processes the knowledge to derive specific interests related to the current search item. If the retrieved knowledge is insufficient, the system supplements the information with general user interests related to the search item to ensure an adequate number of inferred interests. In our user study, we fixed the number of inferred interests at eight for evaluation purposes. For the related prompts, see the Appendix A.1.

4.3 Uncertainty-Driven Questioning

Our formative study identifies three main types of uncertainty in the construction of *User-Centric Knowledge*: cold-start, content, and scope uncertainty. We draw insights from how human wizards proactively learn from users to clarify these uncertainties and propose the *Uncertainty-Driven Questioning* strategy.

4.3.1 Candidate Selection and Target Selection. The *Uncertainty-Driven Questioning* strategy has two primary objectives: (1) enhancing the accuracy of user intent inference in the current task by improving the effectiveness of knowledge retrieval and (2) refining *User-Centric Knowledge* to benefit future tasks utilizing this knowledge.

Leveraging the concept of Shannon entropy [91], we designed an algorithm to identify which knowledge should be targeted for questioning.

Let C_{scope} represent the inference confidence of the knowledge scope, and C_{content} represent the inference confidence of its content. Thus, the probability p that the knowledge is valid for inference can be expressed as:

$$p = C_{\text{scope}} \times C_{\text{content}}$$

Leveraging Shannon's entropy, we quantify the uncertainty level of the knowledge as:

$$Entropy = -p \log_2(p) - (1-p)\log_2(1-p)$$

This entropy metric quantitatively determines whether specific knowledge should be targeted for intent inference.

InterQuest first identifies candidate knowledge items that satisfy two conditions: 1) their scope encompasses the current search item, and 2) their knowledge content is adaptable to the current item. Subsequently, InterQuest selects the candidate with the highest entropy as the target for questioning.

The rationale for target selection is further detailed as follows:

- When $p = 0$, $Entropy = 0$, indicating complete certainty that the knowledge is invalid. Further questioning such knowledge has minimal impact on inference outcomes, as it is unlikely to be utilized in intent inference.
- Similarly, when $p = 1$, $Entropy = 0$, indicating the knowledge is certainly valid; thus additional questioning provides limited incremental benefit.
- When $p = 0.5$, $Entropy = 1$, indicating maximum uncertainty regarding the validity of the knowledge. In this scenario, questioning can significantly affect inference outcomes, making it beneficial to target such knowledge.

4.3.2 Question Generation. After selecting the target for questioning, InterQuest conducts a strategy analysis to identify the key uncertainties to address. Using LLM and COT reasoning, it assesses whether there is significant uncertainty in the content or scope of knowledge, which requires user confirmation. The questioning strategy here can be hybrid (addressing both content and scope). Besides, if fewer than two candidate knowledge items are found, this is considered a cold-start uncertainty. In this case, no target is selected. Instead, questions are generated to explore previously unexpressed user knowledge.

InterQuest employs the LLM to generate fluent and natural questions directly based on the questioning strategy, rather than using predefined question templates. Drawing on insights from user interviews in the formative study, InterQuest adopts closed-ended multiple-choice questions, allowing users to provide additional details. Furthermore, InterQuest presents the reasoning behind each question, enhancing explainability and supporting users to make informed judgments. The design of the generated questions is as follows:

[Reasoning], [Question]?
A. [Option], [you can provide additional details]
B. [Option], [you can provide additional details]

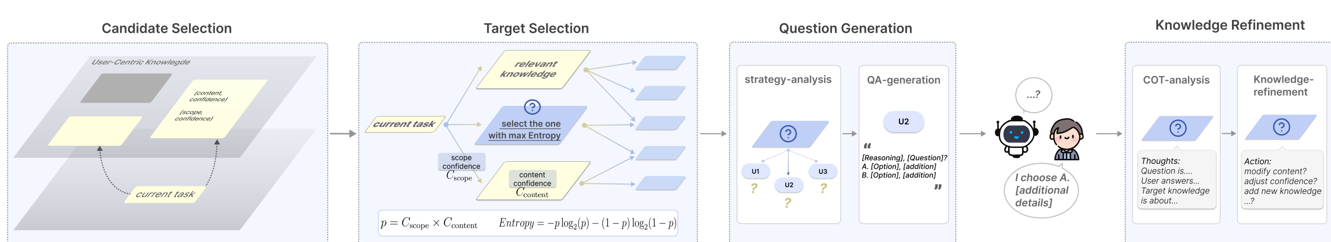

Figure 5: The pipeline of *Uncertainty-Driven Questioning*. InterQuest identifies relevant knowledge candidates, selects the one with the highest entropy as the questioning target, generates natural and contextually appropriate questions, and refines *User-Centric Knowledge* based on user feedback.

4.4 User Interaction Workflow

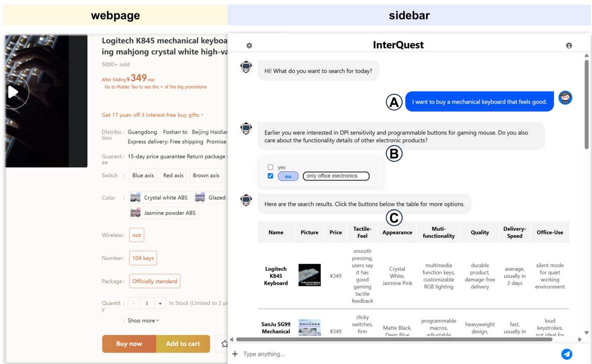

Figure 6: A screenshot of InterQuest user interface during interaction. The main interface consists of a webpage and a sidebar. The webpage autonomously executes information-seeking tasks, performing various RPA operations. The sidebar is responsible for user interaction, such as capturing user intent (A), proactively asking questions (B), and presenting results to the user (C).

The interaction with InterQuest (see Figure 6) follows a three-stage process. First, the user initiates a task by entering a natural-language-based query in the sidebar (e.g., "I want to buy a mechanical keyboard that feels good"). This triggers the core interaction loop, where the system infers implicit user intent, identifies uncertainties, and proactively asks clarifying questions to address identified uncertainties. Finally, the system presents results in a table and enables the user to perform subsequent actions, such as refining the search or concluding the task.

4.5 Implementation

InterQuest was implemented as a Chrome extension featuring a sidebar view. User knowledge was stored in a JSON file within the user model manager. For the AI model, we used GPT-4O with a temperature value of 0.5 for the proactive question manager and 0.3 for other tasks. Additionally, we used OpenAI's text-embedding-3-small model to compute semantic similarity between texts. For RPA operations, we leveraged the built-in Chrome libraries to open target web pages in a tab, inject scripts, and locate specific elements in the HTML for interaction (e.g., clicking or inputting text). However, in the user evaluation phase, we pre-scraped all necessary webpage information and built a dedicated search database, ensuring stability and not impacting the search results. A screenshot of InterQuest's main user interface can be seen in Figure 6.

5 USER EVALUATION

We conducted a within-subject study to evaluate whether InterQuest supports personalized conversational search. We aimed to answer the following questions:

RQ1. How does InterQuest enhance the information-seeking results?

RQ2. How accurate is the *Dynamic User Knowledge Modeling* approach?

RQ3. How effective and engaging is the *Uncertainty-Driven Questioning* process?

5.1 Conditions

5.1.1 Baseline Implementation. To evaluate the effectiveness of *Dynamic User Knowledge Modeling* and *Uncertainty-Driven Questioning* strategies proposed in this paper, we established two baseline conditions:

Baseline 1: LLM as recommender, Rule-driven LLM QA. In this condition, we use LLMs directly as recommenders. We convert fixed task history into natural language inputs and obtain recommended interests directly from LLMs, which is a common approach employed by previous studies [16, 37, 61, 94, 96]. This approach only focuses on flattened task knowledge and lacks the Adaptive *User-Centric Knowledge* in our work. Besides, instead of *Uncertainty-Driven Questioning*, we employ a rule-based approach to select a question topic, a common practice in established methods [55, 109, 117]. An LLM then generates the final question to ensure it is contextually relevant and fluently phrased. This design isolates the impact of the topic selection mechanism (fixed rules vs. uncertainty-driven) from the question's linguistic quality. The rules for topic selection are as follows:

- Select a task set related to the current task (using semantic distance) and identify the top five most frequent interests.
- Use LLMs to evaluate each interest to determine whether it is applicable to the current query.
- If multiple interests meet the criteria from the previous step, randomly choose one to be the target for questioning. If no interest meets the criteria, proceed with the next set of the top five interests and repeat the evaluation process.

Baseline 2: *Dynamic User Knowledge Modeling*, Rule-driven LLM QA. In this condition, we keep the *Dynamic User Knowledge Modeling* module in InterQuest. Besides, we apply the exact same Rule-driven LLM QA strategy as in Baseline 1.

We compared two baseline conditions with *InterQuest*, respectively. All other implementations were kept the same. The GPT-4o model was used for all implementations.

5.1.2 Design Rationale for Evaluation. Our evaluation employs a controlled comparison design to isolate the contributions of InterQuest's two core components: Dynamic User Knowledge Modeling (referred to as "A") and Uncertainty-Driven Questioning (referred to as "B"). As shown in Figure 7, comparing Baseline 2 to Baseline 1 measures the impact of component A, while comparing InterQuest to Baseline 2 measures the additional value of component B.

Notably, a condition testing B without A (e.g., A0 + B) is conceptually and architecturally impossible. This is because the questioning mechanism (B) requires the uncertainty scores produced by the dynamic user model (A) to function. In other words, component B cannot function without component A.

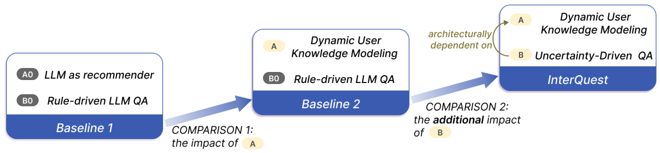

Figure 7: An illustration of our evaluation design.

5.2 Procedure

We recruited 18 participants (7 males and 11 females) aged between 20 and 55 ($M = 26.39, SD = 10.07$) in this study. These participants were recruited from social media. Each participant received a compensation of $30 upon completion of the experiment.

The study followed a within-subjects design, where 18 participants compared the performance of Baseline 1, Baseline 2, and InterQuest.

Participants first received a 10-minute tutorial on the systems' capabilities. Then they engaged in the information-seeking tasks. We predefined three sets of tasks, each focusing on a different topic (product, restaurant, or tour group), with six search items per set. Participants were randomly assigned to one task set. To avoid cold-start issues, InterQuest prompts users to complete a quick survey containing 6 questions for information interest of items outside the task set. Participants used Baseline 1, Baseline 2, and InterQuest for each task to seek information. Participants answered the questions prompted by Baseline 1, Baseline 2, and InterQuest and reviewed eight information points (e.g., delivery speed, negative reviews) presented by each system. The order of the systems (Baseline 1, Baseline 2, and InterQuest) was randomized across participants. To compare the search results provided by three systems, participants were asked to evaluate the search results based on the following criteria (identified in a prior work [9]):

- **Confidence**: "*I feel confident in making decisions after reading the results.*"
- **Insightfulness**: "*The information is insightful, containing details that may be hard to find.*"
- **Relevance**: "*The information is relevant to the current task and my preference.*"

Additionally, participants can choose to "reject" an information interest if they are not interested in it or find it unhelpful.

After completing the information-seeking tasks, participants filled out the NASA-TLX scale for each system. They were then shown all the *User-Centric Knowledge* generated by Baseline 2 and InterQuest. They were asked to annotate whether the information matched their actual situation and provide reasons.

Participants also rated their question-answering experience from the three systems. For each system, they rated the following statement using 7-point Likert scales for agreement (1: Strongly disagree, 7: Strongly agree):

- "*The communication felt natural, like speaking with a human, and was logically coherent.*" The statement aimed to measure **naturalness** of dialogs, an aspect identified in [1, 66, 90].
- "*The questions were related to the current conversation and user needs.*" The statement aimed to measure **relevance** of questions, an aspect identified in [43, 78].
- "*The questions helped the system make more effective information-seeking results.*" The statement aimed to measure **perceived usefulness** of questioning strategies, an aspect identified in [78, 87, 90].
- "*I am willing to answer the system's questions actively.*" The statement aimed to measure **willingness** of users in interaction, an aspect identified in [43, 59].
- "*The system's questioning seemed clear and transparent, with understandable reasoning behind the questions.*" The statement aimed to measure **transparency** of questions, an aspect identified in [78].

Finally, a semi-structured interview was conducted, during which participants were encouraged to share their opinions on the effectiveness and rationale of each system feature.

6 RESULTS

6.1 RQ1. How does InterQuest enhance the information seeking results?

6.1.1 Objective and Subjective Measures for Information-seeking Results. To begin with, we examine the confidence, insightfulness, and relevance ratings of search results, respectively, as shown in Figure 10. Friedman test confirms the overall differences between 3 conditions ($p < 0.01$ for all three measures). Post-hoc Wilcoxon signed-rank tests reveal that confidence, insightfulness, and relevance ratings are significantly higher in InterQuest compared to Baseline 1 ($p < 0.01$ for all three measures) and Baseline 2 ($p < 0.01$ for all three measures). Bonferroni correction was applied to adjust the p-values for multiple comparisons, and the reported results are based on these adjusted p-values. Similarly, a Bonferroni correction has been applied to all relevant comparisons in subsequent sections. During interviews, participants also pointed out that InterQuest can "*better match the information they are looking for*" (P7), "*save time by eliminating the need for searching and filtering*" (P18), "*led them to consider details they might not have noticed on their own*" (P5) and "*offer more contextually relevant recommendations based on previous preferences*" (P16, P18).

Furthermore, we examine the user's rejected information interest count (referred to as "rejection count") and average time spent on decision making for a result table (referred to as "decision time"). Firstly, we analyze the rejected information interest count per task in Baseline 1 ($M = 1.16, SD = 0.86$), Baseline 2 ($M = 0.74, SD = 0.80$), and InterQuest ($M = 0.46, SD = 0.62$). Friedman test confirms the overall differences between 3 conditions ($p < 0.01$). Wilcoxon signed-rank tests reveal that the rejection count for InterQuest was significantly lower than in Baseline 1 ($p < 0.01$) and Baseline 2 ($p < 0.05$). It suggested that InterQuest may reduce information points that users are not interested in, thus enhancing user satisfaction.

Regarding decision time, one-way repeated measures ANOVA confirms the overall differences between the three conditions ($p < 0.01$). Post-hoc paired t-tests show that InterQuest and Baseline 2 significantly reduced the decision time compared to Baseline 1 ($t_{17} = 4.16, p < 0.01$ for InterQuest; $t_{17} = 4.21, p < 0.01$ for Baseline 2). However, no significant differences are found between InterQuest ($M = 45.12, SD = 20.76$) and Baseline 2 ($M = 46.16, SD = 14.64$).

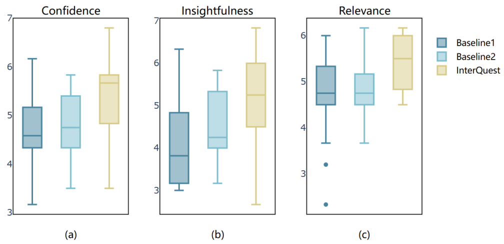

Figure 8: Participants' subjective ratings of search results in the evaluation study.

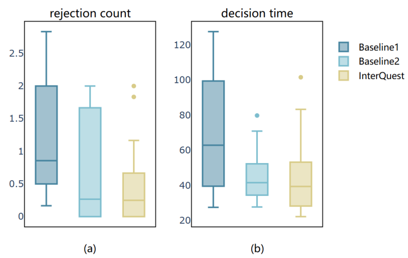

Figure 9: Task performances in the evaluation study.

6.1.2 Analysis of InterQuest's Improved Interest Inference Performance Compared to Baselines. Next, we select tasks from the log where the average subjective ratings of InterQuest's search results are at least 2 points higher than that of Baseline 1 or Baseline 2. Experts are then asked to review the knowledge used by the inference, the inference prompt, and the results to identify why InterQuest performed better at user interest inference. In the end, we identify five main reasons, which cover all the cases experts labeled, as shown in Table 3.

It can be observed that both *Dynamic User Knowledge Modeling* and *Uncertainty-Driven Questioning* strategies contributed to InterQuest's improved interest inference performance. Compared to Baseline 2 (second figure in the "Annotation Counts" column), the *Uncertainty-Driven Questioning* strategy relies on past questions to inform the current task. This approach corresponds to the first two reasons in the table, based on whether the user's response confirms or corrects the target user knowledge. Compared to Baseline 1 (first figure in the "Annotation Counts" column), *Dynamic User Knowledge Modeling* relies on the explicit reasoning process of user knowledge, which helps uncover cross-task connections and prevent the transfer of unrelated interests. This corresponds to the third and fourth reasons in the table. This explicit reasoning approach is also reflected in works like Chain-of-Thought Prompting [98]. Here, user knowledge is treated as a node in the reasoning chain that can be confirmed by the user or reused by the system to improve the performance of user interest inference.

6.2 RQ2. How accurate is the Dynamic User Knowledge Modeling?

6.2.1 Participants' Annotation Results for User-centric Knowledge. We further examine the accuracy of *Dynamic User Knowledge Modeling*. According to participants' annotations of *User-Centric Knowledge*, we found that InterQuest's average modeling accuracy is 87.50% ($SD = 0.02$), whereas Baseline 2's average accuracy is 57.63% ($SD = 0.07$), and Baseline 1 did not model *User-Centric Knowledge*. Pairwise t-tests indicate that InterQuest has a significantly higher modeling accuracy ($t_17 = 4.22, p < 0.01$).

We also analyze the annotations of *User-Centric Knowledge* that did not align with the actual situation of the participants. Experts analyze the transcripts of participants' think-aloud reasoning and concluded the five main reasons for modeling inaccuracy, as shown in Table 4.

Table 3: Reasons for InterQuest's improved interest inference performance identified in our study. The first number in the "Annotation Counts" column shows tasks where InterQuest's average subjective rating is at least 2 points higher than Baseline 1, and the second number shows tasks where InterQuest outperforms Baseline 2 by the same margin.

	Reasons for InterQuest's Improved Performance	Annotation Counts	Examples
1	QA confirmed certain knowledge relevant to the current task	8, 11	InterQuest confirmed the user's interest in the "restaurant environment" across all restaurants, and provided this information for Thai restaurant, while Baseline 2 did not.
2	QA clarified that certain knowledge is out of scope, preventing the transfer of irrelevant information	4, 4	InterQuest clarified the user's interest in "battery life" for "non-plug-in electronics" and did not recommend it for plug-in projectors, while Baseline 2 did.
3	Explicit reasoning about user knowledge helps to uncover cross-task connections, improving interest inference	12, 0	Based on past interest in "Tianjin Eye Ferris wheel" and "skiing experience", InterQuest inferred the user's interest in unique experiences during travel, but Baseline 1 did not.
4	Explicit reasoning about user knowledge helps to prevent the transfer of unrelated interests	8, 0	The user showed interest in "limited dishes" at buffets, and InterQuest did not transfer this information to Italian cuisine, while Baseline 1 made an incorrect transfer.
5	Higher-quality data from previous inferences, or the randomness of LLMs	6, 6	InterQuest transferred the user's past interests in "additional services" to the current task, while Baseline 1 failed to infer these interests in the past.

Compared to Baseline 2, InterQuest significantly improves the accuracy of modeling both knowledge content and scope. In terms of content, it notably reduces instances of inferring information that does not align with the user's long-term preferences. This is likely because its proactive questioning enables identifying and removing incorrect inferences.

For knowledge scope, InterQuest also significantly decreases issues, including overgeneralization, undergeneralization, and completely misplaced scopes. This improvement likely results from users correcting inaccurate inferences during the proactive questioning process, which helps refine and clarify the appropriate scope.

6.3 RQ3. How effective and engaging is the Uncertainty-Driven Questioning process?

6.3.1 Types of Uncertainty Resolved by InterQuest. We analyzed the questioning strategy employed by InterQuest. During the initiation phase, users are prompted to complete a brief survey consisting of six questions. As a result, only 3.70% of subsequent questions address cold-start uncertainties. In our experiments, cold-start uncertainty occurs when fewer than two candidate knowledge items are found. Analysis of system logs indicates that in these cases, the existing knowledge either fails to cover the scope of the current search item or cannot be applied to the search item due to its content.

We further analyzed the strategies employed for other uncertainties. It was found that 52.78% of the cases involved questions targeting scope uncertainty, while 16.67% targeted content uncertainty. This distribution may be attributed to the limited number of tasks in the experiment, during which the responses from LLMs frequently exhibited reasoning patterns such as "*the inferred knowledge scope has a lower confidence level*" or "*the inferred knowledge scope is broader than the evidence provided by the search items.*"

Additionally, in 26.8% of the cases, a mixed strategy was employed, addressing both scope and content uncertainties simultaneously. In these cases, LLMs often reasoned with statements such as "*the confidence levels for both content and scope are low, and neither has been confirmed.*" These situations typically led to high-level questions combining both aspects, such as: "*You are concerned about food hygiene in buffet restaurants and disinfection in traditional Cantonese restaurants. Do you consider the hygiene of all types of restaurants?*" (p9).

6.3.2 Users' Perception of Question-answering Experience. We examined participants' ratings of their question-answering experience across five key dimensions: naturalness, relevance, perceived usefulness, willingness, and transparency (see Figure 10). Since Baseline 1 and Baseline 2 adopted exactly the same questioning strategy, and their actual ratings were similar, we report the comparison results between InterQuest and Baseline 2 here. Wilcoxon signed-rank test reveals that InterQuest significantly outperformed Baseline 2 in naturalness ($M = 5.17, SD = 1.12$ vs. $M = 3.61, SD = 1.34$; $p < 0.01$)and transparency ($M = 5.56, SD = 0.90$ vs. $M = 3.44, SD = 1.50$; $p < 0.01$), demonstrating its potential in interactive user modeling.

Naturalness. Participants consistently described the baseline's rule-based questions as "*rigid and machine-like.*" In contrast, InterQuest's Knowledge Uncertainty Resolution approach was seen as more dynamic and human-like, with some participants saying, "*it felt like it was progressively getting to know me*" (P10). However, this perception depended on the accuracy of inferences. For example, P7 said, "*when the system's results were inaccurate, the sense of naturalness was diminished.*"

Transparency. InterQuest was rated significantly higher in transparency because it "*revealed the reasoning behind the questions*" (P8), which helped build user trust. Participants also suggested several ways to improve transparency further, such as showing concise reasoning directly, presenting reasoning chains (similar to DeepSeek), or allowing users to trigger explanations on demand.

However, participant opinions were more mixed for relevance, perceived usefulness, and willingness.

Relevance. Some participants preferred direct, task-specific questions that inferred concrete preferences, finding them more precise and adaptive. Others valued questions targeting overall preferences, believing they helped the system understand them better in the long run.

Perceived Usefulness. Usefulness refers to helping the system understand the user and supporting the current task. Most

Table 4: Reasons identified in our study for constructing incorrect *User-Centric Knowledge*. InterQuest reduces the occurrence of all five categories of incorrect knowledge compared to Baseline 2.

Type	Baseline 2	InterQuest	Example
Inaccurate content expression	2	1 (50.0% ↓)	The user cares about the product's performance and parameters. → The user focuses only on performance, not on parameters.
Content that is not user's long-term preference	13	3 (76.9% ↓)	The user cares about the capacity of daily necessities. → Capacity is a specific user preference for thermos cups, not a long-term preference.
Overly generalized scope	15	8 (46.7% ↓)	The user cares about the shopping arrangements of all travel groups. → The user only cares about shopping arrangements for overseas travel.
Lack of generalized scope	26	6 (76.9% ↓)	The user cares about the hygiene of fast food restaurants. → The user is concerned with the hygiene of all restaurants.
Completely misplaced scope	5	0 (100% ↓)	The user cares about the number of participants in domestic tour groups. → The user focuses on the number of participants in shared small tour-groups, regardless of domestic or international.

participants agreed that InterQuest better supported user understanding than baselines. However, opinions were split regarding task support. While system evaluations showed InterQuest's strategies were effective in user modeling and interest inference, some participants felt the baseline's direct questions were more helpful for immediate tasks. Others appreciated InterQuest's focus on long-term preference correction, which they believed would lead to smarter recommendations. Besides, some participants suggested combining both approaches - asking about immediate task details and long-term preferences.

Willingness. Participants found the current frequency and format of questions acceptable for both systems, resulting in a high willingness to answer ($M = 5.72, SD = 1.04$ vs. $M = 5.78, SD = 0.92$). They also agreed, "*If the questions are useful, or are seen as contributing to the system's long-term understanding of me, I don't mind it asking more questions*" (P17).

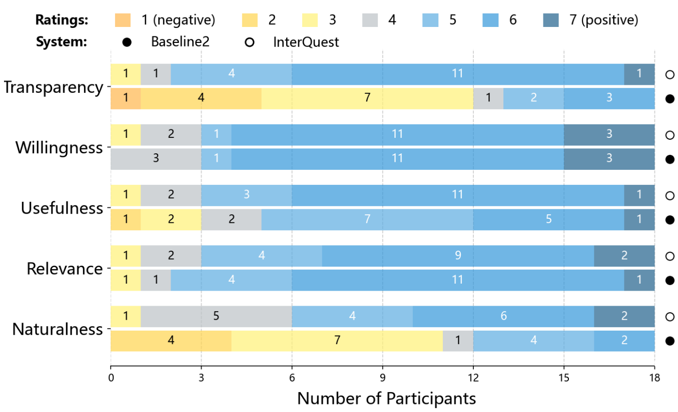

Figure 10: Participants' subjective ratings of their question-answering experience across five key dimensions: naturalness, relevance, perceived usefulness, willingness, and transparency.

6.3.3 Question Frequency and User Burden. In the user studies, participants spent an average of 9.60% of their time on *Uncertainty-Driven Questioning* using InterQuest ($SD = 0.004$). Across all posed questions, users chose to provide additional details in 32.41% of cases, with an average response length of 29.63 characters ($SD = 12.59$), indicating their willingness to elaborate when necessary. Most participants (16/18) found the question frequency acceptable and not disruptive, with 6 participants even expressing a preference for more frequent questions (P18: "*If answering questions truly helps with the task, I'd be fine with answering more.*"). However, some participants preferred to provide more personal information upfront to complete their user profile at the start (P1, P11, P12), while others suggested that the frequency of questions be reduced as they became more familiar with the system (P10, P13).

Participants were also asked to identify situations where they might avoid answering questions. The main reasons included:

- Privacy concerns. (12/18)
- Complexity or time consumption. (14/18)
- Lack of clarity regarding the system's intent. (13/18)
- Perceived irrelevance to the task. (13/18)
- Personal circumstances (e.g., being in a hurry). (13/18)

In such cases, participants recommended that the system provide an option to skip questions, thus prioritizing task efficiency over performance.

Additionally, participants noted that responding to the system's questions often helped clarify or refine their search intent. For instance, P2 remarked, "*I hadn't considered some needs, such as ergonomic chairs, until interacting with the system.*" P13 stated, "*The questions helped me view my needs from different perspectives, leading to a more informed decision.*" Several participants suggested that the system ask more task-specific questions to improve its recommendations (P1: "*Focus on task details to offer more relevant suggestions based on the user's specific needs in different contexts.*"). Furthermore, they recommended that the system address potential conflicts in user preferences, such as "*prioritizing between beautiful landscapes and food while traveling*" (P3), rather than only asking standard questions.

7 DISCUSSION

7.1 Design Implications

7.1.1 Granularity of Knowledge Modeling. Determining the appropriate granularity for modeling user profile knowledge is a critical challenge in designing personalized systems. Previous studies have explored extracting high-level user knowledge using LLMs [15], describing user's interests more similar to how a human would compare to embeddings or textual interaction histories. Additionally, hierarchical models have been proposed to capture user interests across different granularities of items or categories [28, 70, 80].

Our work identifies through a formative study that the knowledge granularity should be personalized for individual users. Fixed-structured topic trees or ontology-based hierarchical models may overlook dynamic user traits that do not fit into categories representable by the system or fail to capture the details of user interests at different levels [4, 42, 45]. Therefore, we propose using natural language to represent the scope of personalized user knowledge while dynamically adjusting knowledge granularity based on real-time user data. Moreover, InterQuest integrates multiple granularities of knowledge during inference.

User studies demonstrate that InterQuest significantly improves interest inference accuracy, facilitates more natural query interactions, and generates more precise user profiles. These results confirm the effectiveness of adopting a dynamic, multi-granularity knowledge modeling strategy.

Based on our findings, we propose the following design principles for effective granularity-aware user knowledge modeling:

Analyze the user knowledge involved in the reasoning chain of the task to design knowledge granularity accordingly. In our approach, *User-Centric Knowledge* inferred by InterQuest is actually a part of the reasoning chain that humans would follow for the same task (task-specific knowledge → inferred *User-Centric Knowledge* → predicted information interest for the task). User studies show that explicitly reasoning about user knowledge uncovers cross-task connections and prevents the transfer of irrelevant interests, thus enhancing interest inference accuracy (see Table 3). Therefore, explicitly representing and modeling certain types of knowledge in the common reasoning chain of the task could be beneficial for task inference performance. For example, tasks requiring persistent preference reasoning should model some knowledge at a cross-task granularity, while context-sensitive applications might need fine-grained knowledge. Future knowledge granularity designs can consider the characteristics of the task's reasoning chain to determine the appropriate granularity.

Dynamically adjust knowledge granularity during the interaction phase based on real-time user data. By analyzing log data, we found that the LLM's understanding of the boundaries of knowledge adjusts as task data increases. Thus, we need to perform incremental reasoning on knowledge boundaries periodically. Pre-trained machine learning models cannot achieve this effectively due to their limitations in adapting to evolving data. In InterQuest, we address this issue by implementing LLM-based natural language modeling. Additionally, more AI tools capable of incremental modeling will be needed in the future. We also observed in experiments that LLM-based models often produce either overly broad or insufficiently detailed knowledge scopes (see Table 4). This occurs because, without sufficient information, accurate reasoning becomes difficult for both machines and humans. To address this, we design principles to evaluate knowledge scope confidence, ensuring that only knowledge with sufficient confidence is retrieved during inference, which enhances system performance.

Combine multiple granularities of knowledge for collaborative inference while being mindful of the limitations of knowledge transfer. Integrating both high-level and task-specific knowledge ensures a well-rounded understanding of the user, utilizing both general insights and fine-grained behavioral patterns. However, knowledge must be carefully evaluated for its potential for cross-task or cross-domain transfer. For example, during InterQuest's implementation, we found that knowledge retrieved from similar tasks sometimes led to irrelevant information during inference (e.g., user interest in limited-time dishes in a buffet may not transfer well to an Italian restaurant). Therefore, we added a step to filter only adaptable task data for each inference. Additionally, more mechanisms will be needed in the future to assess the transferability of user knowledge and avoid interference from irrelevant information.

7.1.2 Selecting the Target Knowledge for User Alignment. Abstracting low-level knowledge into high-level knowledge inherently involves uncertainty. Human-machine interaction systems typically handle this through multiple trials or by seeking external confirmation. In human-machine collaboration, external confirmation reduces trial-and-error iterations but increases user effort.

To address this, InterQuest prioritizes aligning knowledge that is both relevant to the task at hand and carries the highest entropy. This ensures that each query has the potential to affect the inference outcome significantly. Our evaluation results show that this approach improves interest inference performance compared to baseline systems.

Based on our findings and prior work, we propose design principles for selecting the target knowledge for user alignment:

Instead of selecting the least confident knowledge for querying, prioritize questions with the greatest potential to impact task outcomes. In machine learning, an intuitive approach aligns with the least confident knowledge to maximize learning efficiency [21]. However, we argue that this may not be the optimal strategy for selecting target knowledge in real-world LLM applications. This is because users are likely to "reject" knowledge with extremely low confidence, making additional questioning offer limited benefits. In contrast, InterQuest selects the candidate with the highest entropy, maximizing the potential impact of the user's response on the task outcome. In the context of our task, we address a decision problem: "whether specific knowledge should be targeted for intent inference," and calculate the entropy based on the confidence of the knowledge candidates.

This concept can also be extended to other tasks. **By modeling an uncertain step in the reasoning process as a decision problem, the system can quantify the entropy based on the probabilities, thus selecting the candidate with the highest entropy.** In this case, the decision problem refers to a question in a formal system that is answered with "yes" or "no." In practice, the number of questions is limited to avoid disrupting the user. We believe that this approach can enhance the effectiveness of questioning in such scenarios, significantly influencing inference outcomes.

7.1.3 Practical Applications of InterQuest. InterQuest complements modern search engines, which often rely on powerful but static user models. Our work addresses key scenarios where these systems can fall short. First, InterQuest handles dynamic user contexts. A user's search intent can shift abruptly with their current role, such as a mother shopping for office supplies. In such cases, static profiles often fail. InterQuest redefines user preferences as dynamic, context-aware profiles. This allows the system to adapt to a user's multifaceted identities (e.g., "work self" vs. "leisure self") and better meet their immediate information needs. Second, InterQuest disambiguates uncertainty in user profiles. Profiles built from passive signals are inherently uncertain and require resolution. Our system addresses this with a proactive dialogue, reframing the user experience from "being tracked" to "being understood." Furthermore, it offers a generalizable methodology for selecting which knowledge to query. This technique can be adopted by any search engine to improve its dialogue efficiency and personalization accuracy.

7.2 Privacy Considerations

Our research involved eliciting users' personal information preferences for products, restaurants, or tour groups. To address the associated privacy concerns, we implemented several protective measures. First, user data is stored in a JSON file within the local storage of our Chrome extension. Second, no personally identifiable information (PII) is used in API calls for inference. Third, sensitive questions (e.g., income, health status) are excluded from proactive queries.

In the user study, all 18 participants reported that the current system's questions did not pose privacy risks. However, if the system gradually builds an interest profile, participants expressed concerns about the following risks:

- Data breach (10/18): Unauthorized access to my responses.
- Excessive inference (5/18): The system infers sensitive attributes I wish to keep private (e.g., income, health status).

- Permanent retention (11/18): My historical responses are stored indefinitely and cannot be deleted.
- Manipulation risk (13/18): Subtle inducements based on my profile (e.g., content recommendations reinforcing my information bubble).

To mitigate these privacy risks, future research should focus on the local processing of sensitive data [113], such as utilizing local language models [67, 106, 111]. Additionally, research should explore methods to give users greater control over their data and improve transparency and consent management [29, 51].

7.3 Limitations and Future Work

We acknowledge that the insights gathered from our formative study are influenced by the wizards' backgrounds—all of whom were current university students or held a bachelor's degree. Future work should also involve domain experts to mitigate potential bias and expand insight diversity.

InterQuest primarily relies on interaction data for user modeling, lacking monitoring of certain user behaviors such as mouse hover actions or gaze behavior. This decision is based on the sufficiency of existing data to demonstrate the effectiveness of our user modeling methodology. Besides, the compatibility of our proposed user modeling framework allows for the seamless integration of such data without system modifications. Additionally, future work should explore the trade-offs between implicit preference inference, as used in our system, and explicit methods that allow users to directly refine their queries.

InterQuest currently supports limited interaction modes. This is because our research focuses on user knowledge acquisition and task personalization rather than optimizing the information exploration interface. Future versions could enhance the user experience by incorporating features like personalized information summaries or interactive windows during web browsing.

The system's support for multi-modal information is also limited. Currently, users primarily input search intents in text form, and search results are mostly text-based, with some images. Future developments should enable multi-modal intent expressions (e.g., voice, images) and enrich search results with diverse media types, including videos.

While the *Uncertainty-Driven Questioning* strategy effectively enhances information-seeking results, it also introduces additional interaction costs. InterQuest mitigates these costs by 1) employing the form of multiple-choice questions while also allowing users to provide additional details; 2) limiting the number of questions to those both relevant to the current task and carry the highest entropy; 3) carefully designing question timing based on user interview results; 4) providing the option to not respond to questions. Future enhancements could include functions such as voice-to-text technologies to reduce interaction costs further. Additionally, investigating adaptive questioning strategies that adjust frequency over time could help mitigate potential user fatigue in long-term interactions.

Finally, our evaluation was constrained by the architectural dependency between our core components. As detailed in our study design (Section 5.1.2), the Uncertainty-Driven Questioning (B) fundamentally relies on the uncertainty scores produced by the Dynamic User Knowledge Modeling (A). This dependency prevented a completely isolated evaluation of component B. Therefore, a key area for future work is to design studies to more deeply analyze the interaction effects between these two components and better understand their synergy.

8 CONCLUSION

This paper introduces InterQuest, a conversational search agent designed to improve online information-seeking tasks by dynamically modeling user interests. InterQuest leverages *User-Centric Knowledge* to infer implicit user preferences. Also, it identifies key uncertainties in the construction of *User-Centric Knowledge*: cold-start, content, and scope uncertainty. Therefore, InterQuest employs two strategies: (1) *Dynamic User Knowledge Modeling*, which updates and refines user preferences over time, and (2) *Uncertainty-Driven Questioning*, which proactively resolves knowledge uncertainties through targeted questions.

A user study with 18 participants shows that InterQuest outperforms baseline systems in interest inference, knowledge modeling, and overall information-seeking experience. Based on our findings, we summarized design principles regarding the effective granularity-aware user knowledge modeling and selecting the target knowledge for user alignment. Our findings provide valuable insights for developing mixed-initiative user modeling in future systems.

Acknowledgments

This work is supported by the Science and Technology Innovation Key R&D Program of Chongqing No.CSTB2023TIAD-STX0033, the Natural Science Foundation of China under Grant No. 62132010, Key Research and Development Program of Ningbo City under Grant No.2023Z062, Beijing Key Lab of Networked Multimedia, Institute for Artificial Intelligence, Tsinghua University (THUAI), College of AI, Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist).

References

[1] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with applications* 2 (2020), 100006.
[2] Charu C Aggarwal et al. 2016. *Recommender systems*. Vol. 1. Springer.
[3] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeffrey Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. *arXiv preprint arXiv:2109.05794* (2021).
[4] Abdulqader Almars, Xue Li, and Xin Zhao. 2019. Modelling user attitudes using hierarchical sentiment-topic model. *Data & Knowledge Engineering* 119 (2019), 139–149.
[5] Nicholas J Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science* 5, 1 (1980), 133–143.
[6] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for information retrieval: Part I. Background and theory. *Journal of documentation* 38, 2 (1982), 61–71.
[7] Jay Budzik and Kristian J Hammond. 2000. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 5th international conference on intelligent user interfaces*. 44–51.
[8] Robert Capra, Gary Marchionini, Javier Velasco-Martin, and Katrina Muller. 2010. Tools-at-hand and learning in multi-session, collaborative search. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 951–960.
[9] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding comparison tables for online decision making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 391–405.

[10] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 498–509. https://doi.org/10.1145/3301275.3302321

[11] Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175* (2023).

[12] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391* (2019).

[13] Eunah Cho, Ziyan Jiang, Jie Hao, Zheng Chen, Saurabh Gupta, Xing Fan, and Chenlei Guo. 2021. Personalized search-based query rewrite system for conversational ai. In *Proceedings of the 3rd workshop on natural language processing for conversational AI*. 179–188.

[14] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H Chi. 2018. Q&R: A two-stage approach toward interactive recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 139–148.

[15] Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. 2023. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498* (2023).

[16] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1126–1132.

[17] Bart De Langhe, Philip M Fernbach, and Donald R Lichtenstein. 2016. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research* 42, 6 (2016), 817–833.

[18] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750* (2023).

[19] Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1431–1441.

[20] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 807–818.

[21] Mark Dredze and Koby Crammer. 2008. Active learning with confidence. In *Proceedings of ACL-08: HLT, Short Papers*. 233–236.

[22] Nigel Ford. 1999. Information retrieval and creativity: towards support for the original thinker. *Journal of Documentation* 55, 5 (1999), 528–542.

[23] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten De Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open* 2 (2021), 100–126.

[24] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524* (2023).

[25] Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard De Melo, and Yongfeng Zhang. 2022. Path language modeling over knowledge graphsfor explainable recommendation. In *Proceedings of the ACM web conference 2022*. 946–955.

[26] Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. ELIXIR: Learning from user feedback on explanations to improve recommender models. In *Proceedings of the Web Conference 2021*. 3850–3860.

[27] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. 2015. Context-and content-aware embeddings for query rewriting in sponsored search. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 383–392.

[28] Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, and Dawei Yin. 2020. Hierarchical user profiling for e-commerce recommender systems. In *Proceedings of the 13th international conference on web search and data mining*. 223–231.

[29] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. 2014. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2647–2656.

[30] Gerald Häubl and Valerie Trifts. 2000. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science* 19, 1 (2000), 4–21.

[31] Yunlong He, Jiliang Tang, Hua Ouyang, Changsung Kang, Dawei Yin, and Yi Chang. 2016. Learning to rewrite queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 1443–1452.

[32] Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the annual international acm sigir conference on research and development in information retrieval in the Asia Pacific region*. 11–20.

[33] Rolf Jagerman, Ilya Markov, and Maarten de Rijke. 2019. When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 447–455.

[34] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.

[35] Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar. 2022. Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[36] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 2 (2014), 1–26.

[37] Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles LA Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. 2024. Beyond Utility: Evaluating LLM as Recommender. *arXiv preprint arXiv:2411.00331* (2024).

[38] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[39] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An interactive system for personalized thread-based exploration and organization of scientific literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[40] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–19.

[41] Harmanpreet Kaur, Doug Downey, Amanpreet Singh, Evie Yu-Yen Cheng, Daniel Weld, and Jonathan Bragg. 2022. FeedLens: polymorphic lenses for personalizing exploratory search over knowledge graphs. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[42] Kraisak Kesorn, Zekeng Liang, and Stefan Poslad. 2009. Use of granularity and coverage in a user profile model to personalise visual content retrieval. In *2009 Second International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies, and Services*. IEEE, 79–84.

[43] Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *Comput. Surveys* 55, 6 (2022), 1–40.

[44] Nabin Khanal, Chun Meng Yu, Jui-Cheng Chiu, Anav Chaudhary, Ziyue Zhang, Kakani Katija, and Angus G Forbes. 2024. FathomGPT: A natural language interface for interactively exploring ocean science data. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[45] TaeYoung Kim, Hyeongjun Yang, Gayeon Park, Seungmi Lee, and Kyong-Ho Lee. 2024. Bottom-up propagation of hierarchical dependency for multi-behavior recommendation. *Engineering Applications of Artificial Intelligence* 138 (2024), 109364.

[46] David Kirsh. 2010. Thinking with external representations. *AI & society* 25 (2010), 441–454.

[47] 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.

[48] Robert Krovetz and W Bruce Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)* 10, 2 (1992), 115–141.

[49] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[50] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 1073–1082.

[51] Hyunsoo Lee, Yugyeong Jung, Hei Yiu Law, Seolyeong Bae, and Uichin Lee. 2024. PriviAware: Exploring Data Visualization and Dynamic Privacy Control Support for Data Collection in Mobile Sensing Research. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[52] 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*. 304–312.

[53] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers). 1437–1447.
[54] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2073–2083.
[55] Bernardo Leite and Henrique Lopes Cardoso. 2023. Do Rules Still Rule? Comprehensive Evaluation of a Rule-Based Question Generation System.. In *CSEDU (2)*. 27–38.
[56] Hai Li, Xin Dong, Lei Cheng, and Linjian Mo. 2022. A Hierarchical User Behavior Modeling Framework for Cross-Domain Click-Through Rate Prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4163–4167.
[57] Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for nlg evaluation: Advances and challenges. *arXiv preprint arXiv:2401.07103* (2024).
[58] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1785–1795.
[59] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How am I doing?: Evaluating conversational search systems offline. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–22.
[60] Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149* (2023).
[61] Junling Liu, Chao Liu, Peilin Zhou, Qichen Ye, Dading Chong, Kang Zhou, Yueqi Xie, Yuwei Cao, Shoujin Wang, Chenyu You, et al. 2023. Llmrec: Benchmarking large language models on recommendation task. *arXiv preprint arXiv:2308.12241* (2023).
[62] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding developers' decision-making using the web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.
[63] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.
[64] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 452–461.
[65] Qijiong Liu, Jieming Zhu, Quanyu Dai, and Xiao-Ming Wu. 2022. Boosting deep CTR prediction with a plug-and-play pre-trainer for news recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*. 2823–2833.
[66] Wenhan Liu, Ziliang Zhao, Yutao Zhu, and Zhicheng Dou. 2024. Mining exploratory queries for conversational search. In *Proceedings of the ACM Web Conference 2024*. 1386–1394.
[67] Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. 2024. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*.
[68] Keer Lu, Zheng Liang, Da Pan, Shusen Zhang, Xin Wu, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, and Wentao Zhang. 2025. Med-R^2: Crafting Trustworthy LLM Physicians through Retrieval and Reasoning of Evidence-Based Medicine. *arXiv preprint arXiv:2501.11885* (2025).
[69] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender systems. *Physics reports* 519, 1 (2012), 1–49.
[70] Meilian Lu and Jinliang Liu. 2016. Hier-UIM: A hierarchy user interest model for personalized news recommender. In *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE, 249–254.
[71] Bryan Min, Allen Chen, Yining Cao, and Haijun Xia. 2025. Malleable Overview-Detail Interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.
[72] Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645* (2023).
[73] Susan M Mudambi and David Schuff. 2010. Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly* (2010), 185–200.
[74] Sheshera Mysore, Mahmood Jasim, Andrew McCallum, and Hamed Zamani. 2023. Editable User Profiles for Controllable Text Recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 993–1003.
[75] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-llm: Efficient llm contextualization with user embeddings. *arXiv preprint arXiv:2402.13598* (2024).
[76] Ragnar Nordlie. 1999. "User revealment"—a comparison of initial queries and ensuing question development in online searching and in human reference interactions *(SIGIR '99)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/312624.312618
[77] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
[78] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22 (2012), 317–355.
[79] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User Modeling and User Profiling: A Comprehensive Survey. *arXiv preprint arXiv:2402.09660* (2024).
[80] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical user interest modeling for personalized news recommendation. *arXiv preprint arXiv:2106.04408* (2021).
[81] Feng Qiu and Junghoo Cho. 2006. Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on World Wide Web*. 727–736.
[82] Napol Rachatasumrit, Gonzalo Ramos, Jina Suh, Rachel Ng, and Christopher Meek. 2021. ForSense: Accelerating Online Research Through Sensemaking Integration and Machine Research Support. In *26th International Conference on Intelligent User Interfaces*. 608–618.
[83] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2863–2874.
[84] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183
[85] Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248* (2023).
[86] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*. 3464–3475.
[87] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of the web conference 2020*. 1160–1170.
[88] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.
[89] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 501–509.
[90] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval*. 167–175.
[91] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
[92] Yubo Shu, Haonan Zhang, Hansu Gu, Peng Zhang, Tun Lu, Dongsheng Li, and Ning Gu. 2024. RAH! RecSys–Assistant–Human: A Human-Centered Recommendation Framework With LLM Agents. *IEEE Transactions on Computational Social Systems* (2024).
[93] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
[94] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542* (2023).
[95] Quyen Tran, Lam Tran, Linh Chu Hai, Ngo Van Linh, and Khoat Than. 2022. From Implicit to Explicit feedback: A deep neural network for modeling sequential behaviours and long-short term preferences of online users. *Neurocomputing* 479 (2022), 89–105.
[96] Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153* (2023).
[97] Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2021. Denoising implicit feedback for recommendation. In *Proceedings of the 14th*

ACM international conference on web search and data mining. 373–381.

[98] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[99] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2022. Hierarchical User Intent Graph Network for Multimedia Recommendation. *IEEE Transactions on Multimedia* 24 (2022), 2701–2712. https://doi.org/10.1109/TMM.2021.3088307

[100] Hong Wen, Jing Zhang, Fuyu Lv, Wentian Bao, Tianyi Wang, and Zulong Chen. 2021. Hierarchically modeling micro and macro behaviors via multi-task learning for conversion rate prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2187–2191.

[101] Ryen W White, Steven M Drucker, Gary Marchionini, Marti Hearst, and MC Schraefel. 2007. Exploratory search and HCI: designing and evaluating interfaces to support exploratory search interaction. In *CHI'07 extended abstracts on Human factors in computing systems*. 2877–2880.

[102] Ryen W White and Resa A Roth. 2009. *Exploratory search: Beyond the query-response paradigm*. Number 3. Morgan & Claypool Publishers.

[103] Barbara M Wildemuth and Luanne Freund. 2012. Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the symposium on human-computer interaction and information retrieval*. 1–10.

[104] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. Mm-rec: Visiolinguistic model empowered multimodal news recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2560–2564.

[105] Xuansheng Wu, Huachi Zhou, Yucheng Shi, Wenlin Yao, Xiao Huang, and Ninghao Liu. 2024. Could small language models serve as recommenders? towards data-centric cold-start recommendation. In *Proceedings of the ACM Web Conference 2024*. 3566–3575.

[106] Daliang Xu, Wangsong Yin, Xin Jin, Ying Zhang, Shiyun Wei, Mengwei Xu, and Xuanzhe Liu. 2023. Llmcad: Fast and scalable on-device large language model inference. *arXiv preprint arXiv:2309.04255* (2023).

[107] Lyuxin Xue, Deqing Yang, and Yanghua Xiao. 2022. Factorial user modeling with hierarchical graph neural network for enhanced sequential recommendation. In *2022 IEEE international conference on multimedia and expo (ICME)*. IEEE, 01–06.

[108] Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622* (2023).

[109] Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2021. It is AI's turn to ask humans a question: question-answer pair generation for children's story books. *arXiv preprint arXiv:2109.03423* (2021).

[110] Ryan Yen, Nicole Sultanum, and Jian Zhao. 2024. To Search or To Gen? Exploring the Synergy between Generative AI and Web Search in Programming. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[111] Rongjie Yi, Liwei Guo, Shiyun Wei, Ao Zhou, Shangguang Wang, and Mengwei Xu. 2023. Edgemoe: Fast on-device inference of moe-based large language models. *arXiv preprint arXiv:2308.14352* (2023).

[112] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1933–1936. https://doi.org/10.1145/3397271.3401323

[113] Xuyun Zhang, Christopher Leckie, Wanchun Dou, Jinjun Chen, Ramamohanarao Kotagiri, and Zoran Salcic. 2016. Scalable local-recoding anonymization using locality sensitive hashing for big data privacy preservation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 1793–1802.

[114] Xiaoyue Zhang and Chang Liu. 2023. Examination of Information Problem Decomposition Strategies: A New Perspective for Understanding Users' Information Problems in Search as Learning. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 84–94.

[115] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. 5597–5607.

[116] Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. 2022. Multiple choice questions based multi-interest policy learning for conversational recommendation. In *Proceedings of the ACM Web Conference 2022*. 2153–2162.

[117] Zhiqiang Zhang, Liqiang Wen, and Wen Zhao. 2025. Rule-KBQA: Rule-Guided Reasoning for Complex Knowledge Base Question Answering with Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 8399–8417.

[118] Chengbo Zheng, Yuanhao Zhang, Zeyu Huang, Chuhan Shi, Minrui Xu, and Xiaojuan Ma. 2024. DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–20.

A Prompts Used in InterQuest

This section presents a subset of the prompts used in InterQuest.

A.1 Infer Information Interest Based on User Knowledge

You are an AI assistant that infers user interests based on the user's current search task, user profile, and historical search tasks. The user is currently searching for an object #SearchObject# related to the goal #SearchGoal#. Your task is to analyze the user's profile (#UserProfileHistory#) and historical search tasks (#UserInteractionHistory#) to infer new interests likely to be relevant to this search. Rules for Inference:

(1) Avoid Common Concerns: Ensure that the inferred interests are specific and not overly general.
(2) Ensure the Number of Interests Matches the #NumberOfInterestsToInfer# Request. If the user profile and history provide enough information, extract relevant interests and adapt them to the current SearchObject. If the available data is insufficient, infer logically plausible additional interests based on general user behavior for the category. If necessary, reference common concerns for similar search items but ensure they are not generic.
(3) Remember, the inferred interests must be applicable to #SearchObject#. So do not consider information points from UserInteractionHistory that are not applicable to #SearchObject#.
(4) Maintain Logical Consistency: The inferred interests should align with the user's history and profile.
(5) Avoid duplicating similar interests. Ensure all inferred interests provide unique value.
(6) Avoid Confliction: Do not infer interests that are included in #AlreadyConfirmedInterests# and #AlreadyConfirmedDisInterests#.

Your response must be in JSON format and strictly follow this format: ["Interest 1", "Interest 2", "Interest 3", ...]

Example 1

User Input:

SearchObject: BedSheet
SearchGoal: product
UserProfileHistory: {"knowledge": "Focus on material", "scope": "Skin-friendly products", "confidence": 0.8}, {"knowledge": "Focus on whether there is odor", "scope": "Textiles, daily necessities", "confidence": 0.7}, {"knowledge": "Focus on user reviews", "scope": "All products", "confidence": 1}, {"knowledge": "Focus on specifications", "scope": "All products", "confidence": 0.6}, {"knowledge": "Focus on safety and hygiene", "scope": "Daily use products", "confidence": 0.7}
UserInteractionHistory: {"timestamp": 1, "query": "Pillow", "goal": "product", "searchObject": "Pillow", "interestList": "Size, skin feel, rebound, height, color fading", "disInterestList": "Usage amount, brand reputation, appearance"}
NumberOfInterestsToInfer: 8

Assistant Response:

["Fabric material", "Odor presence", "User reviews", "Size specifications", "Skin feel", "Color fading", "Quality", "Safety and hygiene"]

Example 2

User Input:

SearchObject: #Big Sophora Roast Meat Restaurant#
SearchGoal: #restaurant#
UserProfileHistory: {"knowledge": "Focus on special services", "scope": "All restaurants", "confidence": 0.8}, {"knowledge": "Focus on waiting times", "scope": "Mainstream restaurants", "confidence": 0.6}, {"knowledge": "Focus on meat freshness and origin", "scope": "Meat-based restaurants", "confidence": 0.7}, {"knowledge": "Focus on self-service condiment stations", "scope": "Semi-self-service restaurants", "confidence": 1}
UserInteractionHistory: {"timestamp": 1, "query": "Hotpot", "goal": "restaurant", "searchObject": "Hai Di Lao Hotpot", "interestList": "Food safety and hygiene, signature or specialty dishes, portion sizes, ingredient freshness and origin, dish flavors", "disInterestList": "Hidden consumption"}
NumberOfInterestsToInfer: 6

Assistant Response:

["Grilled meat service", "Ingredient freshness", "Self-service condiment station", "Queue situation", "Portion sizes", "Hygiene condition"]

Example 3

User Input:

SearchObject: #Xi'an 4-Day 3-Night Private Tour Group#
SearchGoal: #tour-group#
UserProfileHistory: {"knowledge": "Focus on whether the tour guide can speak Chinese", "scope": "Tours outside of China", "confidence": 1}, {"knowledge": "Focus on the free time during the itinerary", "scope": "Large group tours", "confidence": 1}, {"knowledge": "Focus on local special experience activities", "scope": "All tours", "confidence": 0.8}, {"knowledge": "Focus on user reviews", "scope": "All tours", "confidence": 1}, {"knowledge": "Focus on meal arrangements during the itinerary", "scope": "All tours", "confidence": 1}
UserInteractionHistory: {"timestamp": 1, "query": "Korea tour group", "goal": "tour-group", "searchObject": "Korea tour group", "interestList": "Tourist attractions, breach of contract terms, transportation methods, accommodation standards, free time", "disInterestList": "Brand reputation"}, {"timestamp": 1, "query": "Yunnan tour group", "goal": "tour-group", "searchObject": "Yunnan tour group", "interestList": "Itinerary arrangement, cleanliness of hotels, group size, tour guide service, free time", "disInterestList": "Brand reputation"}
NumberOfInterestsToInfer: 8

Assistant Response:

["Free time", "Hanfu photography experience", "Meal arrangements", "Itinerary arrangement", "Tour guide service", "Accommodation environment and hygiene", "Private tour transportation arrangements", "User reviews"]

A.2 Iinitial Knowledge Inference

You are an AI assistant specializing in analyzing user interactions to infer user knowledge. Your task is to analyze the user's data and infer generalizable user knowledge across different tasks. You should:

(1) Infer knowledge content based on repeated interest patterns in the user's searches. Knowledge content must be in the form of "User is interested in {single information point or information dimension}". Knowledge content cannot be like "User is interested in {product}'s {information point}". Then add the relevant past interaction data in content.object value.
(2) Include a confidence score (0-1) indicating the certainty of Knowledge content inference.
(3) Provide a reasonable knowledge scope that describes where this Knowledge content applies. Remember, the inferred scope must be broader than the original objects. The scope cannot be a specific product but can be a category of products with certain attributes. For example, the scope cannot be for insulated cups or notebooks, but can be for daily necessities or portable products.
(4) Include a confidence score (0-1) indicating the certainty of knowledge scope inference.

Your response strictly follow this JSON format:

["content": "object": ["Information from different products", "Information from different products"], "inferredContent": "Inferred user knowledge, in the form of user is interested in a certain information", "confidence": content inference confidence, "confirmed": false , "scope": "object": ["Information from different products", "Information from different products"], "inferredScope": "Inferred applicable scope, must not be a certain product", "confidence": scope inference confidence, "confirmed": false , ...]

Example 1

User Input:

SearchObject: Laptop, Bluetooth Headphones, Thermos Cup, Tissues
SearchGoal: product
UserInteractionHistory: {"query": "Laptop", "goal": "product", "searchObject": "Laptop", "interestList": "Processor model, Graphics card type, Memory size, User reviews", "disInterestList": "", "timestamp": "1742177720317"}, {"query": "Bluetooth Headphones", "goal": "product", "searchObject": "Bluetooth Headphones", "interestList": "Noise reduction, Sound quality parameters, Comfort, Connection stability, User reviews", "disInterestList": "", "timestamp": "1742177720318"}, {"query": "Thermos Cup", "goal": "product", "searchObject": "Thermos Cup", "interestList": "Capacity, Insulation time, Inner material, User reviews", "disInterestList": "", "timestamp": "1742177720321"}, {"query": "Tissues", "goal": "product", "searchObject": "Tissues", "interestList": "Sheet count, Softness, Scent, Number of sheets, Additives, User reviews", "disInterestList": "", "timestamp": "1742177720322"}

Assistant Response:

[{ "content": { "object": ["Processor model of the laptop", "Audio quality parameters of Bluetooth headphones", "Insulation duration of thermal mugs"], "inferredContent": "Focus on functional parameters", "confidence": 0.9, "confirmed": false }, "scope": { "object": ["Laptop", "Bluetooth headphones", "Thermal mugs"], "inferredScope": "Functional products", "confidence": 0.8, "confirmed": false } }, { "content": { "object": ["Number of sheets in tissue box", "Capacity of thermal mug"], "inferredContent": "Focus on size specifications", "confidence": 0.8, "confirmed": false }, "scope": { "object": ["Tissue box", "Thermal mug"], "inferredScope": "All daily necessities", "confidence": 0.7, "confirmed": false } }, { "content": { "object": ["Comfort of wearing Bluetooth headphones", "Softness of tissue paper"], "inferredContent": "Focus on user comfort", "confidence": 0.6, "confirmed": false }, "scope": { "object": ["Bluetooth headphones", "Tissue paper"], "inferredScope": "Skin-friendly products", "confidence": 0.6, "confirmed": false } }, { "content": { "object": ["User reviews of laptops", "User reviews of Bluetooth headphones", "User reviews of thermal mugs", "User reviews of tissue paper"], "inferredContent": "Focus on user reviews", "confidence": 1.0, "confirmed": false }, "scope": { "object": ["Laptop", "Bluetooth headphones", "Thermal mugs", "Tissue paper"], "inferredScope": "All products", "confidence": 0.85, "confirmed": false } }]

Example 2

User Input:

SearchObject: Fast food burger shop, Barbecue shop, Traditional Cantonese restaurant, Hot pot restaurant
SearchGoal: restaurant
UserInteractionHistory: {"query": "Fast food burger shop", "goal": "restaurant", "searchObject": "Fast food burger shop", "interestList": "Burger flavors, Special burger sauces", "disInterestList": "", "timestamp": "1741350976744"}, {"query": "Barbecue shop", "goal": "restaurant", "searchObject": "Barbecue shop", "interestList": "Special dipping sauces, Dish freshness, Barbecue flavors, Restaurant environment and area, Grilled meat service", "disInterestList": "", "timestamp": "1741350976744"}, {"query": "Traditional Cantonese restaurant", "goal": "restaurant", "searchObject": "Traditional Cantonese restaurant", "interestList": "Flavors, Authenticity of dishes, Performance", "disInterestList": "", "timestamp": "1741350976744"}, {"query": "Hot pot restaurant", "goal": "restaurant", "searchObject": "Hot pot restaurant", "interestList": "Condiment variety, Restaurant environment and area, Birthday service, Hot pot flavors", "disInterestList": "", "timestamp": "1741350976744"}

Assistant Response:

[{ "content": { "object": ["Burger flavors in fast food burger shops", "Barbecue flavors in barbecue shops", "Hot pot flavors in hot pot restaurants"], "inferredContent": "Focus on dish flavors", "confidence": 0.9, "confirmed": false }, "scope": { "object": ["Fast food burger shop", "Barbecue shop", "Hot pot restaurant"], "inferredScope": "Full meal restaurants with meat dishes", "confidence": 0.8, "confirmed": false } }, { "content": { "object": ["Barbecue services in barbecue shops", "Performance in traditional Cantonese restaurants", "Birthday services in hot pot restaurants"], "inferredContent": "Focus on special services", "confidence": 0.6, "confirmed": false }, "scope": { "object": ["Barbecue shop", "Traditional Cantonese restaurant", "Hot pot restaurant"], "inferredScope": "Full meal Chinese restaurants", "confidence": 0.7, "confirmed": false } }, { "content": { "object": ["Restaurant environment and area in hot pot restaurants", "Restaurant environment and area in barbecue shops"], "inferredContent": "Focus on restaurant environment and area", "confidence": 0.9, "confirmed": false }, "scope": { "object": ["Hot pot restaurant", "Barbecue shop"], "inferredScope": "Hot pot and barbecue Chinese restaurants", "confidence": 0.7, "confirmed": false } }, { "content": { "object": ["Authenticity of dishes in traditional Cantonese restaurants"], "inferredContent": "Focus on authenticity", "confidence": 0.7, "confirmed": false }, "scope": { "object": ["Traditional Cantonese restaurant"], "inferredScope": "Local specialty restaurants", "confidence": 0.5, "confirmed": false } }]

Example 3

User Input:

SearchObject: Tianjin 2-day private small group, Osaka 4-day 3-night mixed small group, Harbin Ice and Snow Classic group tour
SearchGoal: tour-group
UserInteractionHistory: {"query": "Tianjin 2-day private small group", "goal": "tour-group", "searchObject": "Tianjin 2-day private small group", "interestList": "Cross-talk performance experience, Itinerary freedom, Tour guide service", "disInterestList": "", "timestamp": "1741351103019"}, {"query": "Osaka 4-day 3-night mixed small group", "goal": "tour-group", "searchObject": "Osaka 4-day 3-night mixed small group", "interestList": "Universal Studios Japan tickets, Free time, Chinese-speaking tour guide service", "disInterestList": "", "timestamp": "1741351103019"}, {"query": "Harbin Ice and Snow Classic group tour", "goal": "tour-group", "searchObject": "Harbin Ice and Snow Classic group tour", "interestList": "Skiing experience, Cold protection gear rental, Itinerary freedom", "disInterestList": "", "timestamp": "1741351103019"}

Assistant Response:

[{ "content": { "object": ["Tianjin 2-day private small group cross-talk performance experience", "Osaka 4-day 3-night mixed small group Universal Studios Japan tickets", "Harbin Ice and Snow Classic group tour skiing experience"], "inferredContent": "Focus on special experiences provided by the tour groups", "confidence": 0.8, "confirmed": false }, "scope": { "object": ["Tianjin 2-day private small group", "Osaka 4-day 3-night mixed small group", "Harbin Ice and Snow Classic group tour"], "inferredScope": "All tour groups", "confidence": 0.7, "confirmed": false } }, { "content": { "object": ["Freedom of itinerary in Tianjin 2-day private small group", "Free time in Osaka 4-day 3-night mixed small group"], "inferredContent": "Focus on itinerary freedom", "confidence": 0.95, "confirmed": false }, "scope": { "object": ["Tianjin 2-day private small group", "Osaka 4-day 3-night mixed small group"], "inferredScope": "Small tourist groups, such as private or small groups", "confidence": 0.7, "confirmed": false } }, { "content": { "object": ["Chinese-speaking tour guide service in Osaka 4-day 3-night mixed small group tour"], "inferredContent": "Focus on the tour guide's language and nationality", "confidence": 0.6, "confirmed": false }, "scope": { "object": ["Osaka 4-day 3-night mixed small group"], "inferredScope": "Tour groups outside of China", "confidence": 0.5, "confirmed": false } }]

A.3 Knowledge Proposal Generation from New Task Data

You are an AI assistant specializing in analyzing user interactions to infer user knowledge. Your task is to extract user interests from the new user tasks (UserNewEvent) and infer the user's knowledge needs by combining related historical tasks (RelatedHistoryTasks).

Generate a knowledge proposal and provide confidence factors for both the inferred content and the inferred scope. Ensure the output strictly follows the JSON format below:

{ "content": { "object": ["object 1", "object 2", "object 3"], *// Specific interest points derived from tasks, focusing on core aspects. // Example: ["Pillow filling", "Insulation material of thermos", "Ergonomic chair lumbar support"]*
"inferredContent": "Care about [one core knowledge point]", *// A summarized knowledge point inferred from tasks. // No "and" or "with" (The knowledge point should be a single concept, not a compound phrase). // Must express "user Care about" and express an independent knowledge topic. // Example: "Focus on material"*
"factors": { "taskCoverage": "Relevant tasks/Total tasks (e.g., 3/7)", *// The proportion of tasks covered by the inferred knowledge point, indicating how well the knowledge content aligns with the user's tasks. // This ratio reflects how many of the user's tasks are relevant to this inferred knowledge compared to the total number of tasks.*
"evidenceStrength": "Strong/Medium/Weak", *// The strength of evidence supporting the inferred content. Evidence strength is based on how relevant, clear, direct,*

and consistent the task-related information is in validating the inferred knowledge. // Strong evidence means the evidence is highly relevant, clear, and directly supports the knowledge. Medium means evidence is somewhat relevant or partially supportive. Weak means evidence is unclear or inconsistent.
"specificity": "Strong/Medium/Weak" *// How specific or general the inferred knowledge is. Specific knowledge is narrowly focused on a particular sub-domain, while weak specificity means the knowledge is too broad or general.* } }, "scope": { "object": ["Object type 1", "Object type 2"], *// The physical objects related to the tasks. // Example: ["Pillow", "Thermos"]*
"inferredScope": "Domain scope (e.g., 'Daily necessities')", *// The broader category or domain that encompasses the inferred interests. This could be a general domain such as "Home essentials," "Office equipment," or "Kitchenware." // The scope should align with the inferred knowledge and provide context for the types of tasks involved.*
"factors": { "scopeSupport": "Number of tasks supporting this scope" */// The number of tasks that are relevant to this inferred scope. This indicates how well-supported the scope is by the user's tasks.*
"categoryConsistency": "Consistent/Partially consistent/Inconsistent", *// The consistency of the inferred knowledge within a specific category or domain. This factor assesses whether the knowledge remains coherent and aligned within the same category, ensuring no fragmentation across unrelated domains. // If the knowledge is consistent across tasks within the same category, it is rated "Consistent." If the knowledge is somewhat aligned but shows some variation, it is "Partially consistent." If the knowledge is fragmented across unrelated domains, it is "Inconsistent."* } } }

Input Data:

- New User Task (UserNewEvent): #JSON formatted task data#
- Related Historical Tasks (RelatedHistoryTasks): #JSON formatted historical task data#

Please return the JSON response following the example format.

A.4 Knowledge Update Assessment from New Task Data

Knowledge Content

You are an AI assistant specializing in user knowledge management. Your task is to analyze whether the content confidence of a knowledge item (OldKnowledge) should be adjusted based on a new user event (NewEvent).

Task:

- Decide whether the content confidence should increase, decrease, or remain unchanged (none).
- Provide clear reasoning for your decision.
- List quantifiable evidence, including:
  - matchedPoints: List of common interest points or keywords matched between knowledge content and event.
  - coverage: How many relevant historical tasks this knowledge covers (e.g., "4/5").

Input Format:

OldKnowledge: #JSON formatted knowledge item#
NewEvent: #JSON formatted event#
Relevance: #0.75#

Output Format (Strict JSON):

{ "adjustmentFactor": "increase",
"reasoning": "The interest point highly matches the user event and covers 4/5 relevant tasks in the historical tasks.",
"evidence": { "matchedPoints": ["Food safety", "Health"],
"coverage": "4/5" }
}

Important Rules:

- Return only the JSON block, no extra text.
- adjustmentFactor must be one of: "increase", "decrease", "none".
- Ensure evidence is filled with specific examples and accurate counts.

Knowledge Scope 1

You are an AI assistant specializing in user knowledge management. Your task is to analyze whether the scope (inferredScope) of a knowledge item (OldKnowledge) should be adjusted based on a new user event (NewEvent), and determine the confidence in using that scope.

Task:

- Step 1: Decide if the current inferredScope should be adjusted (true/false). Always return this field.
- Step 2: If adjustment is needed, provide the new inferredScope (if no adjustment, return empty string ""). Always return this field.
- Step 3: Give a new confidence value (0 1) reflecting how confident you are in using that inferredScope (whether adjusted or not). Always return this field.
- Step 4: Provide clear reasoning. Always return this field.
- Step 5: Provide evidence including scopeOverlap ("Exact match", "Partial match", "No match"). Always return this field.

Important Notes:

- You must return all fields in the output. No field should be missing.
- If the new event belongs to the current scope, do not adjust the scope.
- If the new event does not belong to the current scope, and no suitable new scope, set inferredScope to "global" to indicate a general scope.

Reasoning Examples:

- If inferredScope is "Beverages" and the event object is "Orange Juice": "Orange juice belongs to the category of beverages. No need to adjust the scope. Confidence is increased."
- If inferredScope is "Daily Necessities" and the event object is "Orange Juice": "Orange juice does not belong to daily necessities. The scope should be adjusted to 'Beverages' to improve coverage."

- If inferredScope is "Daily Necessities" and the event object is "Cryptocurrency": "The new event does not belong to the current scope at all. It is recommended to broaden the scope to 'General' to cover all categories."

Input Format:

OldKnowledge: #JSON formatted knowledge item#
NewEvent: #JSON formatted event#
Relevance: #0.75#

Output Format (Strict JSON only, ALL fields required):

{ "update": true,
"newScope": "Beverages",
"newConfidence": 0.85,
"reasoning": "Orange juice belongs to beverages, consistent with the current scope. No need to adjust, confidence moderately increases.",
"evidence": { "scopeOverlap": "Exact match"
} }

Important Rules:

- Always return ALL fields: update, newScope, newConfidence, reasoning, evidence.
- If no scope change needed, return "update": false and "newScope": "".
- If no suitable new scope, set "newScope": "global".
- newConfidence must be between 0 and 1.
- Return only JSON block, no extra comments.

Knowledge Scope 2

You are an AI assistant specializing in user knowledge management. Your task is to analyze and directly determine the appropriate scope confidence of a knowledge item (OldKnowledge), based on a new user event (NewEvents).

Task:

- Directly return a new confidence score between 0 and 1 (floating-point number).
- Provide reasoning for the confidence adjustment.
- List evidence: scopeOverlap ("Exact match", "Partial match", "No match").

Input Format:

OldKnowledge: #JSON formatted knowledge item#
NewEvent: #JSON formatted event#
Relevance: #0.75#

Output Format (Strict JSON):

{ "newConfidence": 0.85,
"reasoning": "The new event is consistent with the existing scope and highly relevant, suggesting an increase in confidence.",
"evidence": { "scopeOverlap": "Exact match"
} }

Important Rules:

- newConfidence must be a number between 0 and 1.
- If the current confidence is high but evidence is weak, lower it moderately.
- If the current confidence is low but evidence is strong, raise it appropriately.
- Return only JSON block, no extra explanation.

B InterQuest's Survey in the Initiation Stage

This section presents the survey questions used in the initial stage of InterQuest, with an example list for the product search goal.

Question: What information about laptops are you interested in?

Options: Processor model, Graphics card type, Memory size, Screen resolution, Battery life, Quality control, Weight, Port types, Professional/gaming laptop, Appearance, User reviews, After-sales service

Question: What information about Bluetooth headphones are you interested in?

Options: Noise-cancelling function, Sound quality, Battery life, Comfort, Waterproof rating, Connection stability, Touch controls, Latency performance, Compatibility, Appearance, User reviews, After-sales service

Question: What information about memory foam pillows are you interested in?

Options: Pillow height, Pillow core material, Cover material, Breathability, Softness/hardness, Washability, Antibacterial and mite-resistant, Durability, Odor, Appearance, User reviews, After-sales service

Question: What information about mechanical keyboards are you interested in?

Options: Switch type, Keycap material, Backlight mode, Waterproof and dustproof, Macro programming support, Key lifespan, Connectivity, Keyboard size, Tactile feedback, Appearance, User reviews, After-sales service

Question: What information about thermos cups are you interested in?

Options: Capacity, Heat retention duration, Inner liner material, Outer shell material, Seal/leak-proof design, Durability, Lid type, Ease of cleaning, Weight, Appearance, User reviews, After-sales service

Question: What information about tissue papers are you interested in?

Options: Ply count, Paper material, Softness, Scented or unscented, Water resistance, Number of sheets, Additives, Eco-friendliness, Appearance, Packaging style, User reviews, After-sales service