When to Ask: A Framework for Interruption-Aware Conversational Recommendation

Sanjana Arun
sanjanaarun@ebay.com

Abstract

Conversational recommender systems have made significant progress in preference elicitation through natural language dialog, and large language models have only accelerated that trajectory. Yet almost all of this work addresses the same sub-problem: *what* to ask. The prior question of *whether* to ask at all has received little formal treatment. Asking a clarifying question is not free—it imposes friction, consumes an interaction turn, and risks session abandonment—but existing frameworks offer little principled basis for deciding when those costs outweigh the value of additional preference signal. We call this the *interruption problem* and argue that it is logically prior to, and distinct from, question generation. We characterize the interruption decision as a cost-benefit tradeoff shaped by three factors—query ambiguity, result confidence, and user friction cost—and propose a taxonomy of four decision scenarios from their interaction. We ground the framework in e-commerce sponsored search, a setting where intent under-specification is pervasive and interruption costs carry direct revenue consequences, and we propose a research agenda for empirical validation.

CCS Concepts

• **Information systems** → **Recommender systems**; *Query intent*;
• **Human-centered computing** → *Interactive systems and tools.*

Keywords

conversational recommendation, clarification, interruption decision, query ambiguity, preference elicitation, sponsored search, large language models

1 Introduction

Consider a user who types "Sony WH-1000XM5" into a sponsored search interface. The query is fully specified: brand, model, and generation are all present. A conversational system that has learned to detect underspecified queries might still surface a clarifying question—"Do you prefer over-ear or on-ear headphones?"—because the form factor was not explicit. The user, who came to the session knowing exactly what they want, encounters an unexpected interrogation. They leave. No clicks, no purchases. The question was technically informative; asking it was nonetheless the wrong choice.

This failure mode does not appear prominently in the conversational recommendation literature. Systems built on large language models (LLMs) have grown markedly more capable at generating contextually rich clarifying questions and interpreting free-form responses [5]. That capability also makes the problem more pressing: when a clarifying question costs almost nothing to generate, over-asking becomes the default failure mode. Yet the field's dominant framing treats clarification as a question generation problem: given a query, produce a good question [1, 11]. The decision that precedes generation—should we ask anything at all?—is rarely modeled with the cost of interruption in view.

We call this the *interruption problem*. When a user issues a query, the system faces a choice among at least four distinct responses: return results as-is; silently re-rank using available context; surface a clarifying question before returning anything; or return a result set deliberately diversified across plausible intents. Existing frameworks typically conflate these choices, defaulting to one or two without articulating the decision criteria.

This paper has two goals. First, we characterize the interruption problem as a cost-benefit decision shaped by three factors and argue it deserves treatment as a first-class problem in its own right. Second, we ground the framework in e-commerce sponsored search—a domain where the interruption decision has measurable revenue consequences—and sketch what empirical validation would require.

2 Background

2.1 Conversational Recommender Systems

Conversational recommendation has been studied through slot-filling dialogue [4], multi-turn preference elicitation [7], bandit-based elicitation [10], and open-ended LLM-driven interaction [5]. Critiquing-based systems [3] allow users to refine recommendations through structured feedback. Some of this work does learn when to ask versus when to recommend [7, 9], as a policy over attribute slots that often penalizes additional turns in the reward. We return to these systems in Section 2.3. Most of the literature nonetheless treats asking as the default mode of progress: the agent's task is selecting which aspect of preference to probe, rather than weighing the cost of probing in context.

2.2 Clarification in Information Retrieval

The IR community has developed dedicated methods for generating clarifying questions for ambiguous queries [1, 8, 11]. Zamani et al. [11] frame the task as ranking candidate questions by expected utility given the query; Aliannejadi et al. [1] extend this to open-domain conversational settings. Both treat the ask decision as a given: they trigger on queries identified as ambiguous, without modeling the cost of the interruption itself or asking whether ambiguity is sufficient justification for asking.

2.3 Mixed-Initiative Interaction

The closest prior work on the timing of system initiative comes from the HCI and dialogue systems literature. Horvitz's principles of mixed-initiative interaction [6] identify task difficulty, user engagement signals, and time cost as factors governing when a system should take initiative versus defer—recognizable precursors of the framework we develop below. That work was aimed at decision

support systems, not retrieval, and predates LLM-based recommenders by more than two decades. The core intuition transfers; the operationalization does not.

Deciding whether interaction is worth its cost is therefore not new: Horvitz frames mixed initiative as acting when the expected value of engaging the user exceeds its cost, and the ask-or-recommend policies above apply a similar logic within conversational recommendation. We retain this expected-value logic. Our contributions are three. First, we treat friction as an explicit, context-dependent input rather than a fixed per-turn penalty. Second, we enlarge the action space with two *implicit* alternatives to asking—silent re-ranking and diversification across plausible intents—in which the user's subsequent behavior, rather than an explicit answer, is the information channel. Third, we separate the interruption policy from question generation, so that each can be developed and evaluated independently.

3 Characterizing the Interruption Problem

Let q denote a user query and $\mathcal{H}$ the available context (session history, user profile, platform signals). Upon receiving q, a conversational recommender must choose one of four responses before surfacing results to the user:

- **Proceed (P):** Return the current top-k results unchanged.
- **Re-rank (RR):** Silently adjust result ordering using $\mathcal{H}$, without any interaction turn.
- **Interrupt (I):** Surface a clarifying question before returning results; use the user's response to revise retrieval or ranking.
- **Diversify (D):** Return a result set that deliberately spans multiple plausible intents implied by q, letting the user's subsequent click behavior resolve ambiguity implicitly.

The choice among $\{P, RR, I, D\}$ is the *interruption decision*. We argue it is governed by the interaction of three factors.

3.1 Query Ambiguity (A)

A measures the degree to which q admits multiple distinct intents that would produce meaningfully different optimal result sets. When the entropy of the intent distribution implied by q (conditioned on $\mathcal{H}$) is large, A is high. "Sneakers" is highly ambiguous; "Nike Air Max 90 size 10 white" is not. A can be estimated from the empirical diversity of click distributions over historical queries lexically similar to q, or by prompting an LLM to enumerate plausible alternative interpretations and measuring their divergence.

3.2 Result Confidence (C)

C measures the system's belief that the current top-k results satisfy the user's actual intent given q and $\mathcal{H}$. This is not reducible to retrieval score alone: a system can retrieve high-scoring items that are nonetheless a poor match for the user's unstated preference. In practice, C is low when the top-k is dispersed across product categories (suggesting the system has not resolved the query) and higher when it converges on a recognizable interpretation also supported by prior session context.

Table 1: Canonical interruption scenarios. F distinguishes Interrupt from Diversify when both A and C favor intervention. A is conditioned on $\mathcal{H}$.

Scenario	A	C	F	Action
Interrupt	High	Low	Low	I: Clarify
Diversify	High	Low	High	D: Diverse results
Re-rank	Low	Low	Any	RR: Use context
Proceed	Low	High	Any	P: Return directly

3.3 User Friction Cost (F)

F captures the expected cost of choosing I in the current context. We restrict F to the cost side of the decision. It is shaped by session depth (an additional turn late in a session costs more relative to the interaction budget than early on) and by platform affordance (a chat-native interface sets lower friction expectations than a search box, where unsolicited questions feel like an obstruction). User familiarity affects the *benefit* side instead: a returning user with a rich interaction history provides less marginal value from clarification than a cold-start user, because the system already holds substantial preference signal. Such history reduces effective A rather than changing the price of a question, and treating the two as part of F would conflate distinct mechanisms. Unlike A and C, F is not a property of the query but of the context in which it arrives.

4 A Taxonomy of Interruption Scenarios

The three factors (A, C, F) jointly determine which action dominates. Table 1 summarizes the four canonical scenarios.

Interrupt (high A, low C, low F). The clearest case for interruption. The query is genuinely ambiguous, the system cannot confidently rank results, and the context makes asking cheap. A user querying "running shoes" on a conversational shopping interface early in a session has left surface, distance preference, and price range all unresolved; one well-chosen question substantially collapses the space. The key enabling condition is low F: the platform affords dialogue, and the user has not yet spent their interaction budget.

Diversify (high A, low C, high F). The same ambiguity and confidence profile as Interrupt, but the context makes interruption costly. A power user late in a long session on a keyword search interface is unlikely to tolerate an unexpected question. Here the system should return results that span the major plausible interpretations of q, drawing on result diversification methods [2]: the user's next click reveals their intent without requiring an explicit response. Diversify trades the precision of a direct answer for lower abandonment risk.

Re-rank (low A, low C). Given $\mathcal{H}$, the query is specific enough that the intent is largely resolved, but the system is not confident its current top-k is the right match—perhaps because the user's history encodes a preference that the query does not express. The right response is not to ask, but to let $\mathcal{H}$ do the work. A user who has historically purchased mid-range electronics and queries "headphones" has low effective A once that history is taken into account, and probably does not want to be asked about their budget;

the system should surface mid-range options without making them re-specify what their click history already reveals.

Proceed (low A, high C). The query is specific and the system is confident the top-k is a good match. The opening example belongs here: the user's query fully specifies their intent, and any interruption reduces satisfaction without improving result quality.

Toward a utility formulation. The factors are not independent, and Table 1 is not exhaustive. A formal reading clarifies their relationship. Let $\pi(z \mid q, \mathcal{H})$ denote the distribution over intents z. Then A is the spread of π over intents that induce different optimal result sets, and C is the expected utility of the current top-k under π. Clarification is worthwhile only if the expected value of the information it yields exceeds the friction cost, in the spirit of expected-value-of-information formulations for clarification [8]:

$$\text{Interrupt only if } \ \mathrm{VOI}(q, \mathcal{H}) > F, \tag{1}$$

where $\mathrm{VOI}(q, \mathcal{H})$ is the expected gain in utility from the best available clarifying question, and F is expressed in the same units. High A together with high C is therefore coherent: when a single robust result set serves all plausible intents, little value remains to be gained by asking, and Proceed dominates. Diversify is the case in which no single set is robust, but a mixed set captures part of the value without spending a turn. A full utility formulation covering the $A \times C \times F$ space is the natural next step.

5 Grounding in E-Commerce Sponsored Search

E-commerce sponsored search is a useful testbed for the interruption framework because it combines high rates of structurally ambiguous queries with a revenue model in which interruption costs are directly measurable. In sponsored search, platforms earn revenue on clicks, not sessions; an abandoned session following an unwanted clarifying question is a lost conversion whose cost can be estimated. This makes the interruption decision unusually tractable to evaluate.

Several features of this domain also stress-test the framework in ways that organic search or recommendation benchmarks do not.

Supply-side externalities. In sponsored search, the result set reflects both organic relevance and advertiser bids. A broad-match advertiser may have listings eligible for many interpretations of a category-level query. An interruption that narrows user intent—"Are you looking for running shoes or casual sneakers?"—removes some listings from contention without those advertisers having any opportunity to respond. The interruption decision thus has externalities for the supply side that buyer-facing frameworks do not model; it is not just a UX choice but a marketplace governance decision.

Session length as a hard constraint. Sponsored search sessions are short. A single interruption in a two-query session consumes half the available interaction budget. This makes the session-depth component of F particularly load-bearing in this domain and suggests that the interruption threshold should decay as a session progresses.

Ad engagement as an ambiguity signal. Sponsored search generates a behavioral signal not present in organic search: which ads a user clicked, skipped, or dwelled on across prior sessions. A user with a consistent history of clicking Apple product ads does not need to be asked which phone brand they prefer. Ad engagement history effectively reduces A for returning users, shifting the decision away from Interrupt or Diversify toward Re-rank or Proceed. This is a potentially high-value source of ambiguity signal that, to our knowledge, the clarification literature has not examined.

6 Research Agenda

The framework raises four concrete empirical questions.

Can A be measured reliably? One approach: compute the entropy of the purchase outcome distribution over queries lexically similar to q in historical logs. High outcome entropy is a candidate signal of ambiguity, although it also reflects noise and exploration, which would need to be separated out. An LLM-based alternative—enumerate plausible interpretations and measure their cosine divergence—provides a signal without historical data but requires calibration. It is not obvious these two signals agree, and the conditions under which they diverge are worth studying in their own right.

How large is F in practice? Session abandonment rates following clarifying questions—segmented by session depth, query specificity, and platform context—would give the field a first empirical estimate of friction cost. To our knowledge, no published work reports these numbers, despite their practical importance.

Can an interruption policy be learned? The decision among $\{P, RR, I, D\}$ can be framed as a contextual bandit problem over features derived from (A, C, F) estimates, or as a reinforcement learning problem when delayed credit assignment matters, with reward defined over long-horizon outcomes (session completion rate, return visit probability) rather than immediate click-through. This framing separates the policy from the question generator, allowing the two to be developed and evaluated independently.

What does offline evaluation look like? Running live experiments to test interruption policies is expensive and risks user harm. Historical logs record only the actions that were actually taken, so offline comparison requires either off-policy estimation or a simulation-based benchmark in which a user model responds to hypothetical clarifying questions based on their subsequent behavior. Constructing such a benchmark, and validating that it predicts live outcomes, is a non-trivial methodological contribution.

7 Discussion and Limitations

The framework presented here is conceptual: (A, C, F) are factors, not algorithms, and the taxonomy is illustrative rather than exhaustive. Several hard problems remain unaddressed. First, the three factors are not independent: users with richer $\mathcal{H}$ tend to have lower effective A, and depth and platform shape F independently of the query. The interactions among factors may matter as much as the factors themselves. Second, the taxonomy treats the four action types as discrete, but real systems operate on a continuum—a partially clarifying question that also surfaces results is not captured. Third, the framework does not specify which clarifying question

to ask within the Interrupt scenario; that remains the province of existing question generation work.

The framework also inherits the LLM-era problem of evaluation: what does it mean to ask a "good" clarifying question less often? Metrics that reward session engagement may penalize helpful interruptions; metrics that penalize abandonment may overcorrect in the other direction.

8 Conclusion

Conversational recommenders have learned to ask better questions. They have not yet learned to ask fewer of them. The interruption problem—when, not what—is a distinct design challenge with real consequences, and one the field has largely inherited rather than designed. The three-factor framework here is a starting point, not a solution: its value is in making the decision explicit so that it can be studied, measured, and optimized rather than defaulted on.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT GPT-5.6 Sol and Claude Sonnet 4.6 to check grammar and to suggest refinements for some sentences. After using these tools, the author(s) reviewed and edited the content as needed, and take full responsibility for the content in this work.

References

[1] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proc. 42nd ACM SIGIR*, pp. 475–484. ACM, 2019.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. 21st ACM SIGIR*, pp. 335–336. ACM, 1998.

[3] L. Chen and P. Pu. Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1):125–150, 2012.

[4] K. Christakopoulou, F. Radlinski, and K. Hofmann. Towards conversational recommender systems. In *Proc. 22nd ACM SIGKDD*, pp. 815–824. ACM, 2016.

[5] L. Friedman, S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, B. Chu, Z. Chen, and M. Tiwari. Leveraging large language models in conversational recommender systems. arXiv:2305.07961, 2023.

[6] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proc. CHI '99*, pp. 159–166. ACM, 1999.

[7] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proc. WSDM '20*, pp. 304–312. ACM, 2020.

[8] S. Rao and H. Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proc. ACL '18*, pp. 2737–2746. ACL, 2018.

[9] Y. Sun and Y. Zhang. Conversational recommender system. In *Proc. 41st ACM SIGIR*, pp. 235–244. ACM, 2018.

[10] S. Yang, H. Yuan, X. Zhang, M. Wang, H. Zhang, and H. Wang. Conversational dueling bandits in generalized linear models. In *Proc. 30th ACM SIGKDD*, pp. 3806–3817. ACM, 2024.

[11] H. Zamani, S. Dumais, N. Craswell, P. Bennett, and G. Lueck. Generating clarifying questions for information retrieval. In *Proc. WWW '20*, pp. 418–428. ACM, 2020.