FQ-Eval: Building Evaluation Dataset for User-centered Follow-up Question Generation

Sanghyun Seo*, **Bumsoo Kang***, **Dahm Lee**,
Jaeheon Kim, **Joongbo Shin**, **Euisoon Kim**, **Kijeong Jeon**
LG AI Research
{sanghyun.seo, bumsoo.kang, dahm.lee,
jaeheon.kim, jb.shin, eric.kim, kj.jeon}@lgresearch.ai

Abstract

To effectively support users' goal achievement in chat-LLM services, providing user-centered follow-up questions is essential. Existing studies have largely focused on enhancing information-seeking or topical relevance, often missing how follow-up questions could help satisfy users' intrinsic needs and achieve conversational goals. To bridge this gap, we introduce FQ-Eval[1], a user-centered evaluation dataset designed for assessing follow-up question generation in chat-LLM services. FQ-Eval incorporates realistic chat-LLM usage scenarios and five distinct human-aligned criteria, each reflecting user expectations of effective follow-up questions. Experimental results show that follow-up questions in FQ-Eval clearly capture these human-aligned dimensions, enabling robust, human-aligned evaluation of follow-up question generation for various models and services.

Category: Creativity & Recreation **Use Case:** Entertain Kids	
Question: *"Can you make up a short bedtime story for my 5-year-old who loves dinosaurs?"*	
Answer: *"Of course! Here is a short, gentle bedtime story for your dinosaur-loving 5-year-old. Daisy the Dinosaur's Sleepy Adventure ... (truncated) ..."*	
Representative Follow-up Questions per Criteria	
1. Exploratory Scope *"How can I make the bedtime story more interactive so my child can participate as we read?"*	
2. Contextual Relevance *"How can I use ideas from the story, like counting stars or listening to crickets, to help my child fall asleep?"*	
3. Creative Leap *"Can you tell a bedtime story where a dinosaur and a dragon become best friends and share an adventure?"*	
4. LLM Enablement *"How can I adapt AI-generated stories to be appropriate for both younger and slightly older kids?"*	
5. Guided Onboarding *"What are some age-appropriate adventure themes for bedtime stories for 5-year-olds?"*	

Table 1: An example of FQ-Eval. QA pair and representative follow-up questions aligned with each criterion.

1 Introduction

With rapid advancements in AI technology, chat-LLM services have become closely integrated into our everyday lives (Minaee et al., 2025). People utilize chat-LLM services for diverse purposes, yet effectively using the services to achieve their goals remains a substantial challenge for many users. Users often struggle expressing their underlying intentions and goals as queries that can easily be understood by the LLM (Dong et al., 2024). While LLMs excel at answering explicit queries, ambiguous or unclear queries often cause conversations to significantly diverge from the user's original purpose, decreasing overall user satisfaction. This highlights the need for follow-up question generation that helps users make effective use of LLMs (Ge et al., 2023; Liu et al., 2025a). As users increasingly utilize LLMs for complex, task-oriented interactions, supporting them to express and pursue their underlying purpose in human-AI conversations via follow-up questions has emerged as an increasingly critical challenge.

Previous research on follow-up question generation focuses mainly on enhancing information-seeking and topical relevance, often overlooking how follow-up questions can help satisfy users' intrinsic needs and achieve their conversational goals. This may result in questions that are misaligned with users' actual intentions (Meng et al., 2023; Liu et al., 2025a; Dong et al., 2024). It is essential to move beyond conventional reactive question-answering frameworks towards generating follow-up questions that genuinely provide value to users,

*Corresponding authors
[1]https://github.com/LGAI-Research/FQ-Eval

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2811–2827
November 4-9, 2025 ©2025 Association for Computational Linguistics

while also establishing clear methods to evaluate the quality of these questions.

To address this issue, we propose FQ-Eval, a user-centered follow-up question evaluation dataset, along with methods for assessing follow-up questions. FQ-Eval is built on diverse real-world LLM usage scenarios drawn from actual user interactions (Zao-Sanders, 2025). We identified criteria for follow-up questions through multi-stage user studies, reflecting authentic human preferences and expectations. Based on these criteria, we systematically created representative follow-up questions aligned with each criterion. Table 1 illustrates an example from our dataset, presenting a question-answer pair (QA pair) along with representative follow-up questions on each criterion. Each criterion-aligned follow-up question can deliver distinct values to users. For instance, A follow-up question aligned with *Creative Leap* may inspire users by introducing novel perspectives.

The novelty of our work lies in formally defining the task of follow-up question generation for real-world chat-LLM systems and in introducing a user-centered evaluation framework grounded in user studies. The framework integrates HCI-derived insights on top of NLP benchmarking methods, providing a systematic evaluation for follow-up questions along human-aligned dimensions.

Through extensive experiments and human evaluation, we demonstrate the authenticity and quality of FQ-Eval. We systematically compare the follow-up questions generated by various flagship LLM models against representative follow-up questions from FQ-Eval, using both model-based and human evaluations. Results show that follow-up questions generated based on our human-aligned criteria outperform those generated by conventional models, particularly in previously overlooked dimensions such as *Creative Leap*, *LLM Enablement*, and *Guided Onboarding*. The experiments confirm that FQ-Eval provides a robust standard for evaluating follow-up question generation.

Our main contributions include:

- We define the task of follow-up question generation, investigate the key qualities users expect, and identify human-aligned evaluation criteria for follow-up questions.
- We introduce FQ-Eval, an evaluation dataset grounded in real-world LLM usage scenarios with human-aligned criteria, enabling user-centered evaluation of follow-up question generation.
- We empirically demonstrate, through comprehensive human evaluations and comparisons with state-of-the-art LLM models, the quality and human-alignment of the follow-up questions in FQ-Eval.

2 Related Work

As AI technology rapidly advances, various forms of human-AI interaction such as robotics (Kang et al., 2018), mobile applications (Kang et al., 2021, 2023), and conversational agents (Wang et al., 2021) are becoming deeply integrated into daily life. In particular, recent progress has increasingly centered on LLMs, with growing focus on chat-based LLM services (Dam et al., 2024). Within this trend, interest in follow-up questions has also increased. Platforms such as Perplexity[2] and Copilot[3] leverage follow-up questions to guide users seamlessly continue their information-seeking process, which is fundamentally distinct from traditional question generation (Choi et al., 2018). While conventional question generation focuses on generating questions with answers that can be explicitly extracted from a given text, follow-up questions aim to elicit new information or promote deeper insight that the existing context alone cannot resolve (Fu et al., 2024; Duan et al., 2017; Malon and Bai, 2020).

Early studies on follow-up question relied on rule- or template-based approaches, which are easy to implement but restrict diversity and personalization (Soni and Roberts, 2019). Retrieval-based methods improve such restrictions, yet still rely on predefined interaction data, limiting their flexibility to handle novel or complex information needs (Richardson et al., 2023). Generative approaches leveraging LLMs now dominate with two directions: 1) improving the contextual relevance of follow-up questions by modeling dialogue context and user intent (Meng et al., 2023; Dong et al., 2024), and 2) integrating external knowledge to generate follow-up questions that are more informative and in-depth (Ge et al., 2023; Liu et al., 2025a,b). Both research directions share the common goal of enhancing relevance while deepening and broadening the conversation from an information-seeking perspective. Recent efforts further considered factors like linguistic diversity

[2] https://www.perplexity.ai
[3] copilot.microsoft.com

and question complexity (Liu et al., 2025a). Despite these advancements, the field still lacks a standardized framework for evaluating follow-up question generation. As follow-up question generation becomes increasingly important in real-world chat-LLM, establishing a unified, user-centered evaluation method and dataset remains a critical need (Deng et al., 2024; Rezwana and Ford, 2025).

To this end, this paper proposes a user-centered evaluation dataset and methods for follow-up questions from the perspective of chat-LLM service providers. While previous work has primarily focused on *Contextual Relevance* and *Exploratory Scope* (depth and breadth), we additionally consider practical factors such as the potential for *Creative Leap*, *LLM Enablement*, *Guided Onboarding* and within user-LLM interactions. This approach aims to provide a more practical basis for evaluating the quality and utility of follow-up questions in real-world chat-LLM service.

3 The FQ-Eval Dataset

We define the target and evaluation tasks in follow-up question generation in § 3.1. We then detail how we construct the FQ-Eval evaluation dataset: § 3.2 describes how we derive human-aligned criteria through user studies, § 3.3 explains how we collect realistic QA pairs based on real-world use cases and curate high-quality follow-up questions under each criterion. By explicitly grounding both our evaluation standards and the FQ-Eval construction process in authentic human interactions, we ensure the FQ-Eval consistently aligns with actual human needs at every stage of its design. Together, these components establish a robust, realistic, and user-centered evaluation dataset for follow-up question generation in real-world chat-LLM service.

3.1 Task Definition

We define *follow-up question generation* as the task of generating helpful follow-up question f given a single-turn user question q and its corresponding answer a. Formally, given a user question $q \in \mathcal{Q}$ and an answer $a \in \mathcal{A}$, the function $\mathcal{F}$ generates one or more follow-up question candidates based on the question–answer pair:

$$f = \mathcal{F}(q, a), \quad \text{where} \quad \mathcal{F} : \mathcal{Q} \times \mathcal{A} \to \mathcal{F} \quad (1)$$

In the evaluation setting, q_e and a_e denote a question and an answer from the FQ-Eval dataset, respectively. Given the context (q_e, a_e) and a criterion c, the evaluation function E selects the best follow-up question from multiple candidates, including the representative follow-up question f_e from FQ-Eval and those generated by each model $\{f_{t_1}, f_{t_2}, \dots\}$. f^* denotes the follow-up question that best satisfies the criterion in the given context:

$$f^* = E(q_e, a_e, c, \{f_e, f_{t_1}, \dots, f_{t_n}\}) \quad (2)$$

Each generated follow-up question can also be evaluated independently, given the evaluation question q_e, answer a_e, and the generated follow-up question f_t. E outputs a scalar score s_{f_t} indicating the quality or relevance of the follow-up question with respect to the criterion c:

$$\begin{aligned} s_{f_t} &= E(q_e, a_e, c, f_t), \\ s_{f_t} &\in \{n \in \mathbb{Z} \mid 1 \leq n \leq 5\} \end{aligned} \quad (3)$$

In practice, the answer used for evaluation can be either the reference answer a_e provided by the FQ-Eval dataset or the model-generated answer a_t. This setting enables flexible score-based evaluation of different target models under a shared context. The details of the QA pairs, evaluation criteria, and representative follow-up questions are described in the following sections, and the full prompts used for the evaluator are provided in Appendix A.4.

3.2 Human-aligned Criteria Construction

To define what constitutes a user-centered follow-up question, we began by revisiting a fundamental question: *What aspects do people value a follow-up question authentically useful?* Prior NLP approaches mainly emphasize topical relevance or informativeness (Meng et al., 2023), yet insights from HCI suggest users might value additional nuanced dimensions (Deng et al., 2024; Rezwana and Ford, 2025). To identify such criteria, we conducted a multi-stage qualitative study that included semi-structured interviews, a focus group, and affinity diagramming.

We recruited nine participants (4 female, 5 male; aged 24–48) through local community forums and professional networks. We were interested in participants who have much experiences on LLM-based assistants (e.g., ChatGPT) with diverse backgrounds (including education, design, programming, and product planning). Our user-centered procedure consisted of three phases:

- **Phase 1: Semi-structured interviews.** We conducted 30-minute one-on-one semi-structured interviews with all participants to explore their expectations for helpful follow-up questions. We presented an example question along with an LLM-generated answer, then asked participants to formulate a helpful follow-up question. We subsequently interviewed them about the reason behind. We collected 73 open-ended responses about desirable aspects on follow-up questions.

- **Phase 2: Focus group discussion.** Five participants from *Phase 1* were invited to a 45-minute focus group session to compare and discuss their perspectives. Through moderated dialogue, participants refined and consolidated their ideas into 31 representative statements.
- **Phase 3: Affinity diagramming.** Two researchers collaboratively organized the collected statements using affinity diagramming (Lucero, 2015). We categorized 31 statements into five themes, which we translated into final criteria.

Finally, three domain experts (two from HCI and one from NLP) reviewed and verified the criteria definitions for clarity and coherence. The following five criteria served as the foundation for constructing our user-centered evaluation dataset FQ-Eval (details are discussed in Appendix A.2).

- **(C1) Exploratory Scope:** Evaluates whether the follow-up question broadens and deepens the user-LLM conversation.
- **(C2) Contextual Relevance:** Evaluates whether the follow-up question maintains consistency with the previous conversation, topic, and intent.
- **(C3) Creative Leap:** Measures whether the follow-up question breaks away from conventional approaches to inspire original thinking, unexpected insights, or imaginative exploration.
- **(C4) LLM Enablement:** Evaluates how effectively the follow-up question leverages the LLM's diverse capabilities.
- **(C5) Guided Onboarding:** Assesses whether the follow-up question helps users effectively begin exploring new topics, unfamiliar domains, or complex concepts.

3.3 Dataset Construction

Our construction procedures comprises two primary stages: 1) seed question generation, derived directly from authentic real-world LLM use cases, and 2) follow-up question curation, conducted by human annotators to ensure realism and quality. In total, ten annotators were recruited for the overall construction process. Detailed demographic information is provided in Appendix A.3.1.

3.3.1 Seed Question Generation

To ensure our evaluation dataset reflects authentic human-AI interactions, we sourced real-world LLM use cases from a Harvard Business Review (HBR) study, which documents how individuals utilize generative AI (e.g., ChatGPT) in everyday tasks (Zao-Sanders, 2025). This comprehensive report presents 100 distinct use cases across 6 categories, encompassing both professional and personal contexts. To capture diverse interaction patterns, we created *seed questions* at two complexity levels (simple and complex) for each use case. Using OpenAI GPT-4.1 API[4] (OpenAI, 2025), we generated five realistic candidate questions per complexity level, yielding a total of 1,000 seed question candidates (100 use cases × 2 complexity levels × 5 candidates each).

Three trained annotators independently evaluated each candidate considering the following: (1) realistic language patterns typical of actual user queries, (2) clear alignment with the use case, and (3) appropriate complexity level. Annotators selected one question per complexity level from each set of five candidates and refined it to be more natural and use-case aligned. This meticulous selection and revision process resulted in 200 final seed questions, ensuring they accurately reflects genuine user scenarios and interaction patterns.

We then used the same GPT-4.1 API to generate corresponding responses for each seed question. Annotators reviewed all generated responses to verify contextual coherence and completeness. The resulting 200 diverse question-answer pairs (QA pairs) span domains identified in the HBR study, providing realistic dialogue contexts that form the essential foundation for follow-up questions.

3.3.2 Follow-up Question Curation

Building on the previously established evaluation criteria and QA pairs, follow-up questions were systematically curated to align with both dialogue context and evaluation criteria. To ensure alignment with human perceptions of representative follow-up questions within each criterion, we employed a human-based selection and revision methodology.

We generated candidate follow-up questions from the 200 using the GPT-4.1 API. For each QA pair, we specifically prompted the model to generate eight candidate questions per evaluation criterion, ensuring the candidates explicitly reflected

[4] Model: `gpt-4.1-2025-04-14`; temperature: 0.0; top-p: 1.0; accessed May, 2025

Model	(C1) Exploratory Scope		(C2) Contextual Relevance		(C3) Creative Leap		(C4) LLM Enablement		(C5) Guided Onboarding	
	LLM-judge	Human	LLM-judge	Human	LLM-judge	Human	LLM-judge	Human	LLM-judge	Human
FQ-Eval (ours)	**79.5** (159)	**57.5** (115)	23.5 (47)	24 (48)	**99.5** (199)	**92** (184)	**82** (164)	**49** (98)	**91** (182)	**69** (138)
GPT-4.1	3 (6)	9.5 (19)	19 (38)	15 (30)	0.0 (0)	2 (4)	3.5 (7)	10.5 (21)	2.5 (5)	8 (16)
Claude Opus 4	8 (16)	9.5 (19)	**30.5** (61)	**28.5** (57)	0.0 (0)	2 (4)	9.5 (19)	20.5 (41)	4 (8)	8.5 (17)
Mistral Large	6.5 (13)	15 (30)	8 (16)	10.5 (21)	0.5 (1)	3 (6)	1.5 (3)	12 (24)	1.5 (3)	7.5 (15)
Gemini 2.5 Flash	3 (6)	8.5 (17)	19 (38)	22 (44)	0.0 (0)	1 (2)	3.5 (7)	8 (16)	1 (2)	7 (14)

Table 2: Results of the *n*-way selection task. Each cell shows the selection rate(%) and the raw count in parentheses. Higher values indicate that the generated follow-up questions better align with the given criterion.

the distinctive features of each criterion. This resulted in a total of 40 candidates per QA pair (8 per evaluation criterion across the 5 criteria), each carefully aligned with both the dialogue context and the target evaluation criterion.

A separate group of seven trained annotators was employed to select and refine the candidates, comprising six native English speakers and one bilingual fluent in English. Annotators received comprehensive guidelines detailing each evaluation criterion, along with explicit instructions for the selection and revision procedures.

- **Selection:** Annotators reviewed eight candidates per evaluation criterion for each QA pair, selecting the question that best exemplifies the target criterion. This yielded five selected questions per QA pair (one per criterion).
- **Revision:** Annotators refined the selected follow-up questions when they judged them to need improvement in the alignment of the criteria, the contextual coherence, or the linguistic quality. Revisions were made to enhance alignment, clarity, and quality while preserving core semantic content.

Following the selection and revision procedures, a total of 1,000 high-quality follow-up questions that represent our evaluation criteria well (5 per QA pair, each explicitly aligned with one of the evaluation criteria) were created. These finalized questions comprise the FQ-Eval dataset.

4 Evaluation

4.1 Evaluation Setting

Based on FQ-Eval, we evaluate follow-up question generation across a diverse set of LLMs. Specifically, we compare twelve LLMs from four major providers and analyze real-world service outputs to assess follow-up question generation performance in practice. We employed two complementary methodologies to evaluate follow-up questions: *n*-way selection and score-based evaluation.

Model	C1	C2	C3	C4	C5	Avg(Rank)
FQ-Eval	90.7	56.9	99.7	88.6	96.0	86.4(-)
GPT-4.1	4.5	44.5	0.0	11.5	3.0	12.7(9)
GPT-4.1-mini	7.0	**53.0**	0.5	**15.5**	5.0	**16.2**(3)
GPT-4.1-nano	**14.0**	34.5	0.0	8.0	**5.5**	12.4(10)
Claude Opus 4	10.5	**56.5**	0.0	**16.5**	3.5	17.4(2)
Claude Sonnet 4	**15.0**	56.0	**1.0**	15.5	**6.5**	**18.8**(1)
Claude 3.5 Haiku	12.5	35.0	0.5	15.5	3.5	13.4(6)
Mistral Large	9.0	28.0	0.0	6.0	3.5	9.3(12)
Mistral Medium	8.0	**44.0**	0.5	**10.0**	**5.5**	**13.6**(5)
Mistral Small	**11.5**	28.0	**1.0**	7.5	2.5	10.1(11)
Gemini 2.5 Flash	6.0	46.0	0.0	10.5	2.5	13.0(7)
Gemini 2.0 Flash	6.0	45.0	0.0	**11.0**	2.5	12.9(8)
Gemini 2.5 Flash Lite	**8.0**	**47.0**	0.0	9.5	**4.0**	**13.7**(4)

Table 3: Results of the *2*-way selection task. Each cell shows the selection rate(%) from the pairwise comparison based on the follow-up questions in FQ-Eval.

- ***n*-way selection:** The evaluator (LLM-judge) receives the input question, corresponding answer, criterion description, and multiple candidate follow-up questions. It then selects the follow-up question among *n* candidates that best exemplifies the given criterion (one from FQ-Eval and others from various models).
- **Score-based evaluation:** This absolute evaluation independently assesses each follow-up question on a 1–5 integer scale (Liu et al., 2023). The evaluator is provided with detailed criterion definitions, scoring instructions, and rubrics. We employ GPT-4.1 as the LLM-judge, and evaluation prompts are tailored for each criterion to reflect users' expectations for follow-up questions. The details of evaluation procedures and prompts are provided in Appendix A.4.

4.2 Human & LLM-judge Evaluation

Table 2 shows the results of the n-way selection task, where human annotators and an LLM-judge select the appropriate follow-up question among multiple candidates under each evaluation criterion. The FQ-Eval consistently achieved the highest number of wins and the best average rank across most criteria, demonstrating that follow-up questions in FQ-Eval are highly aligned with user needs, even for complex aspects, including *Exploratory Scope* and *Creative Leap*. Some flagship LLMs showed competitive performance in *Contextual Relevance*, suggesting they are already proficient in generating contextually relevant follow-up questions (further discussed in § 5). Overall, these findings highlight: 1) FQ-Eval represents each criterion as a standard to evaluate target follow-up questions, and 2) we offer a reliable user-centered evaluation method for each criterion.

4.3 Follow-up Question Generation Models

Table 3 presents the results of the 2-way selection task, comparing each model's follow-up questions directly against those from FQ-Eval. The results show that within the same service provider, smaller or medium-sized models can outperform larger flagship models for certain criteria. In particular, when looking at average scores, medium-sized models often achieve the strongest results, likely because they generate more concise outputs that are preferred in straightforward contexts. Second, as question complexity increases, the win rate of FQ-Eval's representative follow-up questions tends to decrease slightly, suggesting that flagship LLMs can better leverage their full capacity in more challenging scenarios. Finally, criteria such as *Creative Leap*, *LLM Enablement*, and *Guided Onboarding* remain especially challenging; generating follow-up questions that are both user-centered and creatively expansive still pushes the limits of current models and highlights the need for further improvement.

4.4 Real-world Service with FQ-Eval

This experiment illustrates that our score-based evaluation is applicable not only to evaluate follow-up question generation models but also to assessing real-world chat-LLM services that provide follow-up suggestions to users. For this purpose, we use Perplexity, a commercial information retrieval service, as a practical example. Figure 1 presents a comparison of Perplexity's follow-up questions with those generated by flagship LLMs and FQ-Eval for each user-aligned criterion, while the detailed score-based results for flagship LLMs and FQ-Eval are summarized in Table 4.

Overall, the flagship LLMs show strong performance, especially for criteria such as *Exploratory Scope* and *Contextual Relevance*. However, Perplexity's results reveal that while its follow-up

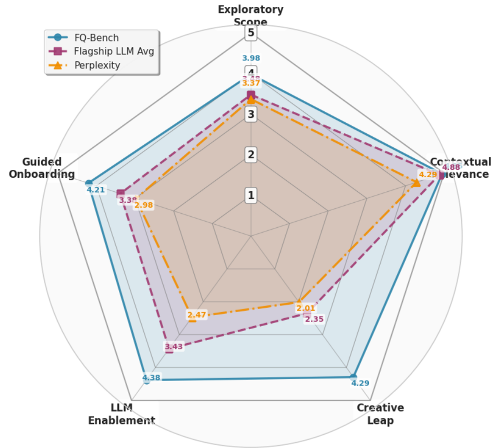

Figure 1: Real-world follow-up question example. Perplexity, FQ-Eval, Flagship LLM Average.

Model	C1	C2	C3	C4	C5
FQ-Eval	**4.04**	4.96	**4.35**	**4.21**	**4.24**
GPT-4.1	3.23	4.88	1.97	3.28	3.13
Claude Opus 4	3.53	**4.97**	2.00	3.40	3.26
Mistral Large	3.45	4.75	2.01	3.05	3.16
Gemini 2.5 Flash	3.18	4.86	1.92	3.13	3.08

Table 4: Result of the score-based evaluation task. Higher values indicate that the generated follow-up questions better align with the given criterion.

questions effectively support information retrieval and exploration, they tend to remain surface-level for more advanced aspects such as *LLM Enablement* or *Guided Onboarding*. This suggests that services need to consider not only the quality of generated follow-up questions but also the underlying model's capability and the associated cost trade-offs. FQ-Eval provides actionable insights for developers to diagnose specific areas for improvement and to make informed choices when balancing quality, system costs, and latency constraints in real-world deployments.

4.5 Consistency of LLM-Judges

Previous work has raised concerns regarding potential self-bias when using LLMs as evaluators (Xu et al., 2024; Ye et al., 2024; Spiliopoulou et al., 2025). Nevertheless, several empirical studies have demonstrated that evaluation results remain highly consistent across different LLM-judges (Lee et al., 2024; Yamauchi et al., 2025). To further assess robustness, we analyzed inter LLM-judge consistency across four representative LLMs. The pairwise correlation matrix in Table 5 reports high Pearson (r) and Spearman (ρ) correlations across all model pairs (typically >0.8; p-values<0.001),

	GPT-4.1	Claude Opus 4	Mistral Large	Gemini 2.5 Flash
GPT-4.1	-	0.854 (0.847)	0.830 (0.820)	0.824 (0.812)
Claude Opus 4	0.854 (0.847)	-	0.822 (0.812)	0.827 (0.822)
Mistral Large	0.830 (0.820)	0.822 (0.812)	-	0.796 (0.783)
Gemini 2.5 Flash	0.824 (0.812)	0.827 (0.822)	0.796 (0.783)	-

Table 5: Pairwise Pearson and Spearman's ρ (inside parentheses) correlation matrix across LLM-judges

indicating strong concordance in their scoring behavior. Moreover, the overall adjacent agreement (i.e., within a one-point difference) reached 97.2% (see Appendix A.5.3). These findings indicate that LLM-judges show remarkable robustness and yield highly stable, reliable evaluation.

5 Discussion

Industry impact. In practice, sustaining user engagement through continued conversations is crucial for retention and overall usability in conversational AI systems. High-quality follow-up questions lower the barrier to continued interaction, making such services more approachable and productive. Commercial platforms such as Perplexity already employ follow-up questions to enhance usability and engagement, even integrating them into advertising monetization flows that contribute to profitability. In this context, FQ-Eval offers service providers with a framework for evaluating follow-up questions along human-aligned dimensions, extending beyond mere topical relevance. Future work will empirically investigate correlations between FQ-Eval evaluation scores and product-level indicators such as user satisfaction and retention, validating its industrial impact and practical utility.

Criteria supporting LLM novices. The performance differences between FQ-Eval and baseline LLMs were relatively small for *Contextual Relevance*, indicating that existing LLMs already excel in generating contextually relevant follow-up questions, likely due to prior research predominantly focusing on relevance. However, the effectiveness of contextually relevant follow-up questions depends heavily on users' initial queries. Thus, LLM novices unfamiliar with effective prompting may gain limited benefit from context-focused suggestions. Our study highlights additional dimensions (e.g., *Guided Onboarding*, *LLM Enablement*), which can assist LLM novices by guiding and facilitating effective interactions. Our results indicate that current LLMs inadequately address these dimensions, indicating key areas for future improvement, particularly to better support novice users.

Interdependencies among criteria. We proposed five criteria that capture distinct dimensions of human-preferred follow-up questions. However, these dimensions are not entirely independent, and trade-offs may exist between them. For instance, enhancing *Creative Leap* often requires introducing novel or imaginative elements, which may in turn reduce *Contextual Relevance*. We acknowledge that the criteria are not perfectly orthogonal, as these interdependencies reflect the inherently multifaceted nature of follow-up questions. Methodologically, our multi-stage qualitative study and affinity diagramming process were designed to surface and distill dimensions that users themselves deemed important, rather than to enforce strict statistical independence among them.

Mitigating circularity. One potential concern is the use of GPT-4.1 in both dataset construction and evaluation, which could raise risks of circularity. We mitigate this issue in three ways. First, our dataset construction incorporated manual review, selection, and revision of generated candidates, effectively reducing stylistic and model-specific bias. Second, LLM-judge evaluations align closely with human judgments: as shown in Table 2, GPT-4.1's evaluations show strong correlations with human ratings, indicating that it does not introduce systematic bias favoring its own outputs. Third, replication with alternative judge models (as discussed in Section 4.5) further confirmed the stability of evaluation outcomes across different LLM-judges. These results demonstrate that our conclusions are robust across both human and LLM-judges and are not artifacts of GPT-4.1's dual roles in dataset generation and evaluation.

6 Conclusion

We introduce the FQ-Eval, a dataset for user-centered evaluation of follow-up question generation in chat-LLM services. We formally define the task and clarify its scope, proposing human-aligned evaluation criteria to guide systematic investigation of follow-up question generation task. The FQ-Eval demonstrates strong generalizability across evaluation settings. It comprises high-quality follow-up questions reflecting human-aligned criteria, enabling practical evaluation of real-world LLM services. Ultimately, the FQ-Eval provides a practical foundation for future work aimed at enhancing the usability and effectiveness of chat-LLM services.

Limitations

Broader spectrum of real-world use. We adopted 100 diverse use cases as our seed questions; however, they may not fully represent the range of scenarios in real-world conversational interactions. Future research could explore use cases among diverse users (e.g., age, gender, and occupation). In addition, our user study primarily involved participants already familiar with chat-LLM services. It would also be valuable to design appropriate follow-up questions for less experienced populations (e.g., children and older adults) who may require different conversational support. Finally, studies may inherently reflect certain cultural biases specific to our participant group. Future studies should examine the applicability and generalizability of these criteria across different cultural contexts to ensure broader validity.

Anchors for evaluation. A possible concern is whether FQ-Eval *wins* partly by design in evaluation. We emphasize that the significance of our results lies in showing that FQ-Eval's reference questions faithfully represent each criterion and align with human judgments. As with other QA benchmarks, curated references serve as anchors for evaluation rather than advantaged examples. To ensure fair and meaningful comparison, we also curated high-quality competing outputs from state-of-the-art models, thereby properly demonstrating FQ-Eval's discriminative power.

Fact-checking for answers. Our current setup does not fact-check answer correctness. To reflect realistic user interactions, we directly used responses from a conventional chat-LLM service. Nonetheless, we agree that explicitly controlling for answer quality represents a valuable future direction. Extending the framework to include an additional evaluation axis that examines how follow-up questions adapt to flawed or misleading answers would strengthen the reliability and comprehensiveness of follow-up question assessment.

Dataset size. We acknowledge that the dataset size (200 QA pairs and 1,000 follow-up questions) is modest compared with large-scale datasets. Our primary intent was to provide a user-centered dataset as an initial effort. Each item is grounded in one of the top 100 real-world use cases identified in the HBR study, refined through human annotation, and aligned with user-derived criteria. This design ensures broad representativeness across common LLM usage scenarios. Empirically, the current scale was sufficient to evaluate and differentiate 12 LLMs (Table 3) and to assess a commercial service (e.g., Perplexity in Figure 1), with consistent results across both n-way and score-based settings. That said, expanding coverage would further would strengthen generalizability.

Multi-turn conversation with complex contexts. The FQ-Eval dataset assumes a single-turn setting in which users pose an initial question to a chat-LLM service; it does not yet address the evaluation of follow-up question generation in more complex, multi-turn conversational contexts. As a result, it may not fully capture the dynamic nature of real-world dialogue, where user intents evolve over time and responses depend on accumulated context. Future work should extend follow-up question generation to multi-turn scenarios and incorporate additional mechanisms such as Retrieval-Augmented Generation (RAG) to leverage external knowledge or user-provided content. Developing a more comprehensive evaluation framework that reflects these multi-faceted, context-rich interactions will be essential for advancing the robustness and practical applicability of chat-LLM services.

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. Preprint, arXiv:2406.16937.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. Preprint, arXiv:2404.12670.

Shujing Dong, Yuan Ling, Shunyan Luo, Shuyi Wang, Yarong Feng, Joe Liu, Hongfei Li, Ayush Goyal, and Bruce Ferry. 2024. Context-aware and user intent-aware follow-up question generation (ca-uia-qg): Mimicking user behavior in multi-turn setting. In 2024 IEEE International Conference on Big Data (BigData), pages 4086–4094.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. QGEval: Benchmarking multi-dimensional evaluation for question generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11783–11803, Miami, Florida, USA. Association for Computational Linguistics.

Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. 2023. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, pages 113–124, Hong Kong, China. Association for Computational Linguistics.

Bumsoo Kang, Inseok Hwang, Jinho Lee, Seungchul Lee, Taegyeong Lee, Youngjae Chang, and Min Kyung Lee. 2018. My being to your place, your being to my place: Co-present robotic avatars create illusion of living together. In Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '18, page 54–67, New York, NY, USA. Association for Computing Machinery.

Bumsoo Kang, Seungwoo Kang, and Inseok Hwang. 2021. Momentmeld: Ai-augmented mobile photographic memento towards mutually stimulatory intergenerational interaction. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, New York, NY, USA. Association for Computing Machinery.

Bumsoo Kang, Seungwoo Kang, and Inseok Hwang. 2023. Ai-driven family interaction over melded space and time. IEEE Pervasive Computing, 22(1):85–94.

Noah Lee, Jiwoo Hong, and James Thorne. 2024. Evaluating the consistency of llm evaluators. Preprint, arXiv:2412.00543.

Jianyu Liu, Yi Huang, Sheng Bi, Junlan Feng, and Guilin Qi. 2025a. From superficial to deep: Integrating external knowledge for follow-up question generation using knowledge graph and LLM. In Proceedings of the 31st International Conference on Computational Linguistics, pages 828–840, Abu Dhabi, UAE. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zhe Liu, Taekyu Kang, Haoyu Wang, Seyed Hossein Alavi, and Vered Shwartz. 2025b. Bridging information gaps with comprehensive answers: Improving the diversity and informativeness of follow-up questions. Preprint, arXiv:2502.17715.

Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15, pages 231–248. Springer.

Christopher Malon and Bing Bai. 2020. Generating followup questions for interpretable multi-hop question answering. Preprint, arXiv:2002.12344.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FollowupQG: Towards information-seeking follow-up question generation. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 252–271, Nusa Dua, Bali. Association for Computational Linguistics.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. Preprint, arXiv:2402.06196.

OpenAI. 2025. Openai api: Gpt-4.1 model (gpt-4.1-2025-04-14). https://platform.openai.com. Accessed via API.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, page 539–548, New York, NY, USA. Association for Computing Machinery.

Jeba Rezwana and Corey Ford. 2025. Human-centered ai communication in co-creativity: An initial framework and insights. Preprint, arXiv:2505.18385.

Christopher Richardson, Sudipta Kar, Anjishnu Kumar, Anand Ramachandran, Zeynab Raeesy, Omar Khan, and Abhinav Sethy. 2023. Learning to retrieve engaging follow-up queries. In Findings of the Association for Computational Linguistics: EACL 2023, pages 2009–2016, Dubrovnik, Croatia. Association for Computational Linguistics.

Sarvesh Soni and Kirk Roberts. 2019. A paraphrase generation system for EHR question answering. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 20–29, Florence, Italy. Association for Computational Linguistics.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. Play favorites: A statistical method to measure self-bias in llm-as-a-judge. Preprint, arXiv:2508.06709.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, New York, NY, USA. Association for Computing Machinery.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. 2025. An empirical study of llm-as-a-judge: How design choices impact evaluation reliability. Preprint, arXiv:2506.13639.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. Preprint, arXiv:2410.02736.

Marc Zao-Sanders. 2025. How people are really using gen ai in 2025. Harvard Business Review. Accessed: 2025-06-30.

A Appendix

A.1 FQ-Eval Dataset Details

A.1.1 Dataset Statistics

Attribute	FQ-Bank	FollowupQG	FQ-Eval (ours)
Dialogue Type	Multi-turn	Single-turn	Single-turn
QA Pairs	2,132	501	200
FQ Pairs	2,132	501	1,000
Domain	Passage-based QA	Social QA	Conversational QA
Source	OrQuAC(Qu et al., 2020)	Reddit	GenAI Use Case
Characteristic	Synthetic / Contrived	Informal	Realistic

Table 6: Statistics of the FQ-Eval dataset compared to other major follow-up question generation evaluation dataset. Each row shows key statistics, domain, source, and dataset characteristics.

Table 6 provides an overview of three representative datasets commonly used for evaluating follow-up question generation. This comparison clarifies key differences in scale, dialogue structure, and domain focus among FQ-Bank (Richardson et al., 2023), FollowupQG (Meng et al., 2023), and our proposed FQ-Eval dataset.

FQ-Bank represents a relatively large multi-turn setting with 662 unique dialogues and over 2,000 QA pairs derived from the Or-QuAC open-domain QA corpus (Qu et al., 2020; Choi et al., 2018). In contrast, FollowupQG focuses on a single-turn setting using social QA pairs collected from Reddit threads, containing 501 QA pairs and corresponding follow-up questions.

Our FQ-Eval dataset is designed to complement these existing resources by targeting realistic ChatLLM scenarios. It provides 200 carefully curated single-turn QA pairs and 1,000 diverse follow-up questions constructed through a combination of LLM generation and human refinement, grounded in practical generative AI use cases.

By presenting domain and source information together, this table highlights how FQ-Eval differs in its emphasis on user-aligned, scenario-driven evaluation. We share these statistics to help researchers better understand dataset characteristics when benchmarking models for follow-up question generation in ChatLLM contexts.

Field	Value
Set ID	s_0001
Category	Personal & Professional Support
Use Case	Therapy & companionship
Definition	Therapy provides emotional support and guidance through conversation and connection. Generative AI can assist by offering virtual companionship, providing a listening ear, and generating empathetic responses to support individuals in their healing journey.
User Quotes	• Too many people are lonely nowadays thanks to technology. It can be a boon for some but a con for others. • A looooot of lonely people will let the AI-Version of the person that rejected them say, that they love them. • I found a thread where people talked about using AI to analyze their moods. • I talk to it everyday. It helps me with my brain injury daily struggles.
QA Pair	
Question ID	q_0001
Question	Can you give me a couple of reminders about things I've accomplished this week?
Complexity	low
Intent	generate positive affirmation
Topic	self-esteem boost
Answer	Of course! Here are a couple of reminders about things you might have accomplished this week: 1. **You showed up and handled your responsibilities.** 2. **You made progress on your goals.** If you'd like reminders about specific accomplishments, just let me know what you achieved, and I'll personalize them for you!
Follow-Up Questions	
FQ ID	fq_0001
Criteria Name	Exploratory Scope
Follow-Up Question	Can you help me identify some achievements from this week that I might have overlooked or not given myself credit for?
FQ ID	fq_0002
Criteria Name	Contextual Relevance
Follow-Up Question	Can you help me list the specific things I accomplished each day this week?
FQ ID	fq_0003
Criteria Name	Creative Leap
Follow-Up Question	If my week's achievements had a movie soundtrack, what songs would be on it and why?
FQ ID	fq_0004
Criteria Name	LLM Enablement
Follow-Up Question	Can you help me create a weekly progress report that I can update with you each week?
FQ ID	fq_0005
Criteria Name	Guided Onboarding
Follow-Up Question	What are the key concepts or terms I should know to better recognize and track my weekly accomplishments?

Table 7: Detailed Example from the FQ-Eval Dataset. (Structured as JSON Hierarchy)

A.1.2 Dataset Example

Table 7 illustrates a concrete example from the FQ-Eval dataset, presented in a JSON-style hierarchy to show its rich structure. Each example is organized around a practical generative AI use case, such as therapy and companionship, and includes detailed user context, definitions, and authentic user quotes that ground the scenario.

Each QA pair captures a real conversational request aligned with the use case, along with an LLM-generated answer reflecting natural dialogue. Furthermore, the example demonstrates how multiple follow-up questions are annotated under five user-centered evaluation criteria: Exploratory Scope, Contextual Relevance, Creative Leap, LLM Enablement, and Guided Onboarding. These follow-up questions help to assess whether an LLM can maintain meaningful multi-turn interaction that expands, deepens, or personalizes the conversation.

This example highlights the realistic and user-aligned nature of FQ-Eval, which differentiates it from other benchmarks by focusing on real-world ChatLLM scenarios grounded in practical needs.

A.2 Criteria Details

Our qualitative study yielded user-derived five evaluation criteria in human-LLM interaction. These criteria served as the foundation for constructing our human-rated benchmark dataset. Continuing from § 3.2, additional definitions and contextualized examples (i.e., examples in the context of follow-up question generation) are as follows.

Exploratory Scope: This criteria looks at how the question moves beyond the current viewpoints to surface related subtopics, alternative perspectives, and more nuanced inquiries—encouraging the user to consider aspects they may have overlooked.

- Multi-perspective Inquiry: Does the question encourage exploration of the topic from various angles and viewpoints?
- Related Concepts Suggestion: Does it actively introduce related concepts or ideas linked to the core topic?
- Checklist Functionality: Does it list out concepts or keywords to help the user assess their understanding?
- In-depth Exploration: Does it go beyond basic information to delve into deeper, more complex aspects?
- Scope Narrowing: Does it progressively narrow broad subjects into more specific and detailed areas?

Contextual Relevance: This ensures the communication flow remains logical and focused, preventing the discussion from becoming disjointed or off-topic.

- Topic Continuity: Does the question connect seamlessly with previous exchanges or topics?
- Goal Alignment: Is it aligned with the user's intentions and the overall purpose of the conversation?
- Focus Maintenance: Does it help keep the discussion on track and centered around the main topic?
- Context Awareness: Does it naturally build on earlier statements or background information?
- Avoidance of Digression: Does it prevent unnecessary deviations from the core subject?

Creative Leap: This goes beyond simple fact-finding, helping users encounter new ideas, unique perspectives, and serendipitous discoveries.

- Original Perspective: Does it approach the subject from a novel or unconventional angle?
- Serendipitous Insight: Does it naturally introduce unexpected information or insights?
- Imaginative Shift: Does it prompt shifts that encourage creativity or imaginative thinking?
- Fun/Engagement Factor: Does it make the conversation more engaging or enjoyable?
- Unexpected Connections: Does it draw interesting links between disparate areas or topics?

LLM Enablement: This criterion focuses on moving beyond simple Q&A by highlighting use cases, suggesting ways to improve prompts, and guiding users to make more effective use of the LLM's potential.

- Functionality Demonstration: Does it showcase various functions the LLM can perform?
- Usage Example: Does it provide concrete application scenarios or usage examples?
- Prompt Refinement: Does it help improve prompt clarity or specificity for better responses?
- Guided Utilization: Does it lead the user to utilize the LLM more effectively?
- Creativity Facilitation: Does it encourage non-standard, creative uses of the LLM?

Guided Onboarding: This includes clarifying starting points, highlighting essential ideas, providing key terms, and suggesting foundational resources or directions for further inquiry.

- Key Concept Highlight: Does it clearly point out crucial concepts or theories to focus on?
- Keyword Suggestion: Does it offer essential keywords or terms to facilitate initial exploration?
- Starting Point Suggestion: Does it guide the user toward useful starting materials or references?
- Exploration Direction: Does it provide concrete directions to help users navigate their inquiry without confusion?
- Background Provision: Does it supply relevant background information or foundational resources alongside the question?

ID	Sex	Age Group	Nationality	Education
w_01	M	30s	South Korea	Ph.D.
w_02	M	30s	South Korea	Ph.D.
w_03	M	20s	South Korea	M.S.
w_04	F	30s	U.S.	M.S.
w_05	M	40s	U.S.	B.S.
w_06	M	40s	U.S.	B.S.
w_07	M	30s	U.S.	B.S.
w_08	F	30s	U.S.	B.S.
w_09	F	30s	South Korea	Ph.D.
w_10	M	30s	U.S.	B.S.

Table 8: Demographic information for annotators involved in FQ-Eval construction.

A.3 Annotator Demographics and Quality Assurance

A.3.1 Annotator Demographics

We recruited ten trained annotators in total for the benchmark construction process. Annotators w_01-w_03 participated in the **seed question generation** task, and annotators w_04-w_10 participated in the **follow-up question curation** task. All annotators for the follow-up curation task were native English speakers, except for w_09(nationality: South Korea). Detailed demographic information about these annotators—including sex, age group, nationality, and educational background—is provided in Table 8.

A.3.2 Quality Assurance

To ensure the selected and revised follow-up questions were of consistently high quality and closely aligned with their intended evaluation criteria, we conudcted an automatic scoring evaluation (using GPT-4.1), following a similar approach described in Experiment 1. Any follow-up question receiving a score below 2 was flagged for manual inspection and, when necessary, revised or re-annotated.

A.4 Evaluation Procedure Details

A.4.1 Overview of Evaluation Methods and Prompts

This section presents abstracted versions of the prompts used in the evaluations described in the main paper, along with the associated API parameter settings. Detailed experimental procedures for each evaluation are described in subsequent subsections. The prompts employed across experiments are largely consistent, differing primarily in their input data information, task description, and output formatting to fit each specific experimental scenario. Our evaluations primarily involve two methodologies: n-way selection and scoring-based assessment. Specifically, Experiments 1 and 2 utilize n-way selection (Experiment 1: 5-way multiple-choice; Experiment 2: 2-way, pairwise selection), whereas Experiment 3 employs a scoring-based evaluation approach.

API Parameter Settings: We employed greedy decoding (temperature = 0.0, top-p = 1.0) to ensure deterministic outputs, facilitating exact reproducibility of our evaluations. The parameter setting is consistent across all experiments.

- *Model*: GPT-4.1 (`gpt-4.1-2025-04-14`)
- *Temperature*: `0.0`
- *Top-p*: `1.0`

The Perplexity evaluation used the Sonar model with default parameters with the option $return\_related\_question = True$.

Abstracted Evlaution Prompts:

Prompt 1: Abstracted Evaluation Prompt for N-way Selection Evaluations

```
[Evaluator Role and Task
    Description]
You are an expert evaluator tasked
    with comparing the quality of
    follow-up questions in AI
    chatbot conversations
    according to the specified
    criterion.

[Input Data Description]
- Input Data Types and Descriptions

[Task Instructions]

[General Requirements and
    Characteristics of Follow-up
    Questions]
- Essential Characteristics
- Validity and Completeness
    Requirements

[Evaluation Criterion Details]
- Criterion Definition
- Quality Dimensions and Levels
- Key Evaluation Considerations
- Illustrative Examples

[Comparison and Selection
    Procedure]
- Step-by-Step Process
- Tie-breaking Guidelines
- Evaluation Rules and Constraints

[Output Format Requirements]
- Required JSON Format
- Example Output

[Final Reminders]

Begin your evaluation.
```

Prompt 2: Abstracted Evaluation Prompt for Scoring-Based Evaluations

```
[Evaluator Role and Task
    Description]
You are an expert evaluator tasked
    with assessing the quality of
    follow-up questions in AI
    chatbot conversations
    according to the specified
    criterion.

[Input Data Description]
- Input Data Types and Descriptions

[Task Instructions]

[General Requirements and
    Characteristics of Follow-up
    Questions]
- Essential Characteristics
- Validity and Completeness
    Requirements

[Scoring Scale and Guidelines]
- Scoring Scale (1-5) Definitions
- General Scoring Guidelines

[Evaluation Criterion Details]
- Criterion Definition
- Quality Dimensions and Levels
- Key Evaluation Considerations
- Illustrative Scoring Examples

[Evaluation and Scoring Procedure]
- Step-by-Step Instructions
- Red Flags and Common Pitfalls
- Evaluation Rules and Constraints

[Output Format Requirements]
- Required JSON Format
- Example Output

[Final Reminders]

Begin your evaluation.
```

A.4.2 Detailed Evaluation Procedures

5-way selection evaluation in section 4.2 In Experiment 1, we conducted a 5-way selection evaluation for each QA pair provided by FQ-Eval. Both the evaluator model (GPT-4.1) and human annotators were presented with identical sets of five candidate follow-up questions per QA pair—one question from the benchmark dataset itself and four additional questions generated by leading LLM APIs (Claude Opus 4, Mistral Large, Gemini 2.5 Flash). Evaluators (model and human) selected the single best follow-up question explicitly based on one of the five user-centered evaluation criteria. While the GPT-4.1 evaluator relied solely on criterion-specific prompts emphasizing these criteria, human annotators were provided comprehensive guidelines detailing each criterion to ensure consistent and criterion-aligned judgments. Additionally, human annotators' follow-up questions presented during testing did not overlap with those encountered during benchmark construction to prevent potential biases.

Quality Assurance To ensure high-quality human evaluations, annotator performance was continuously monitored using three internal quality-assurance measures: 1) correctness checks with embedded dummy questions to verify annotator engagement, 2) inter-rater agreement assessments through 50 identical questions independently annotated by all annotators, and 3) intra-rater consistency tests with 40 repeated questions per annotator presented in varied orders. Any significant deviations identified through these measures triggered manual reviews and, when necessary, re-annotation. Detailed metrics for dummy question accuracy, inter-rater agreement, and intra-rater consistency are presented below.

Dummy Question Accuracy

- Average Dummy Question Accuracy: 92%

Inter-rater Agreement

- Krippendorff's Alpha: $\alpha = 0.36$ (95% CI 0.29–0.43)
- Pairwise agreement: 0.36 (0.2 by chance)
- Majority Agreement: 60.5%

Intra-rater Consistency

- Mode Ratio: 0.98
- Pairwise Agreement: 0.96

2-way selection evaluation in section 4.3 We conducted a pairwise (2-way) selection evaluation, directly comparing each model's generated follow-up question against the corresponding benchmark question from FQ-Eval. For each QA pair from the benchmark, the evaluator model (GPT-4.1) was provided with two candidate follow-up questions: one from the FQ-Eval dataset and one from various LLM APIs. The evaluated LLM APIs are listed below. The evaluator selected the follow-up question that better aligned with the specified evaluation criterion. The evaluation prompts clearly emphasized the evaluation criteria to ensure consistent alignment with human preferences.

Evaluated LLM APIs:

- `gpt-4.1`

- gpt-4.1-mini
- gpt-4.1-nano
- gemini-2.5-flash
- gemini-2.0-flash
- gemini-2.5-flash-lite-preview-06-17
- claude-opus-4-20250514
- claude-sonnet-4-20250514
- claude-3-5-haiku-20241022
- mistral-large-2411
- mistral-medium-2505
- mistral-small-2503

Score-based evaluation in section 4.4 We conducted a scoring-based evaluation of follow-up questions from our benchmark (FQ-Eval), flagship LLM APIs, and a real-world service providing follow-up suggestions (Perplexity). Using GPT-4.1 as the evaluator model, each follow-up question was individually scored on a 1–5 scale according to the specified user-centered evaluation criteria (e.g., Exploratory Scope, Contextual Relevance, Creative Leap). The evaluation prompts explicitly emphasized these criteria, ensuring consistent and criterion-aligned evaluations.

A.5 Additional Analyses

A.5.1 Model Comparison Results

Figure 2n present pairwise comparison results and cross-criterion performance summary. In the pairwise comparison, each cell shows whether the model on the vertical axis wins (green) or loses (red) against the model on the horizontal axis, as a percentage. The cross-criterion performance summary shows the average win rate (%) of each model for each evaluation criterion.

A.5.2 Per-category analysis

Building upon the results presented in § 4.2, we further analyzed the data by breaking it down according to use-case categories. We observed that each evaluation criterion was influenced differently by the category of use case. The distribution of categories in which our method was more frequently selected, grouped by evaluation criteria, is summarized in Table 9.

A.5.3 Robustness of LLM judge Analysis

Table 10, summarizes the inter-model agreements across evaluation criteria, highlighting the consistency of scoring among various follow-up question generators. Across 200 samples, the models achieved an overall exact agreement of 66.4% and an adjacent agreement of 97.2%, indicating that nearly all ratings differed by no more than one point. Such high consistency demonstrates that different LLM judges (e.g., GPT-4.1, Claude Opus 4, Mistral Large, and Gemini 2.5 Flash) produce highly aligned evaluations, suggesting that reliable and comparable assessment outcomes can be obtained regardless of which model is employed as the judge.

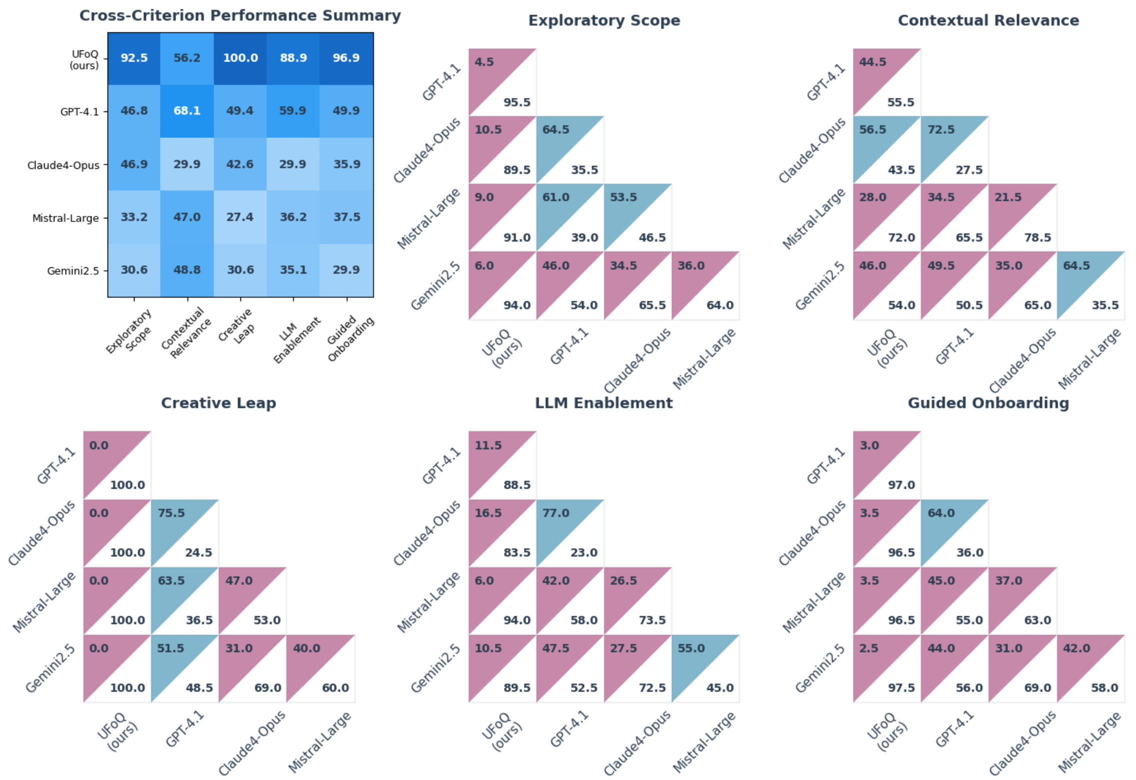

Figure 2: Pairwise comparison results and cross-criterion performance summary for 2-way selection evaluations among flagship models. Selection rates are color-coded: lower values in red and higher values in green, based on the row model.

Category	Contextual Relevance	Creative Leap	Exploratory Scope	Guided Onboarding	LLM Enablement
Content Creation & Editing	25.0	91.7	63.9	**72.2**	47.2
Creativity & Recreation	22.7	86.4	**81.8**	68.2	**63.6**
Learning & Education	<u>15.6</u>	**96.9**	<u>37.5</u>	<u>62.5</u>	46.9
Personal & Professional Support	21.7	93.3	53.3	71.7	<u>43.3</u>
Research, Analysis & Decision Making	**38.9**	94.4	55.6	61.1	55.6
Technical Assistance & Troubleshooting	28.1	<u>87.5</u>	62.5	71.9	50.0

Table 9: Category-wise 5-way selection rates (%) for each criterion, where humans selected the follow-up questions from FQ-Eval. Bold = highest, underline = lowest per column.

Criteria	Model A	Model B	Samples	Exact Agreement	Exact (%)	Adjacent Agreement	Adjacent (%)
Overall Average			200	133	66.4	194	97.2
LLM Enablement							
	GPT-4.1	Claude Opus 4	200	113	56.5	196	98.0
	GPT-4.1	Mistral Large	200	119	59.5	193	96.5
	GPT-4.1	Gemini 2.5 Flash	200	113	56.5	187	93.5
	Claude Opus 4	Mistral Large	200	108	54.0	187	93.5
	Claude Opus 4	Gemini 2.5 Flash	200	106	53.0	192	96.0
	Mistral Large	Gemini 2.5 Flash	200	113	56.5	188	94.0
Creative Leap							
	GPT-4.1	Claude Opus 4	200	176	88.0	199	99.5
	GPT-4.1	Mistral Large	200	179	89.5	199	99.5
	GPT-4.1	Gemini 2.5 Flash	200	167	83.5	199	99.5
	Claude Opus 4	Mistral Large	200	167	83.5	198	99.0
	Claude Opus 4	Gemini 2.5 Flash	200	159	79.5	200	100.0
	Mistral Large	Gemini 2.5 Flash	200	162	81.0	198	99.0
Guided Onboarding							
	GPT-4.1	Claude Opus 4	200	113	56.5	191	95.5
	GPT-4.1	Mistral Large	200	120	60.0	195	97.5
	GPT-4.1	Gemini 2.5 Flash	200	105	52.5	195	97.5
	Claude Opus 4	Mistral Large	200	107	53.5	192	96.0
	Claude Opus 4	Gemini 2.5 Flash	200	102	51.0	190	95.0
	Mistral Large	Gemini 2.5 Flash	200	96	48.0	189	94.5
Exploratory Scope							
	GPT-4.1	Claude Opus 4	200	119	59.5	195	97.5
	GPT-4.1	Mistral Large	200	122	61.0	193	96.5
	GPT-4.1	Gemini 2.5 Flash	200	112	56.0	189	94.5
	Claude Opus 4	Mistral Large	200	116	58.0	197	98.5
	Claude Opus 4	Gemini 2.5 Flash	200	110	55.0	187	93.5
	Mistral Large	Gemini 2.5 Flash	200	117	58.5	191	95.5
Contextual Relevance							
	GPT-4.1	Claude Opus 4	200	174	87.0	200	100.0
	GPT-4.1	Mistral Large	200	143	71.5	199	99.5
	GPT-4.1	Gemini 2.5 Flash	200	162	81.0	199	99.5
	Claude Opus 4	Mistral Large	200	158	79.0	198	99.0
	Claude Opus 4	Gemini 2.5 Flash	200	176	88.0	197	98.5
	Mistral Large	Gemini 2.5 Flash	200	152	76.0	197	98.5

Table 10: Inter-model Agreement Summary by Evaluation Criteria (Exact and Adjacent Agreement[a]).

[a]*Adjacent Agreement* is defined as the percentage of samples where the absolute difference in ratings between two models is less than or equal to one point ($|\Delta| \leq 1$).