Context-Aware and User Intent-Aware Follow-Up Question Generation (CA-UIA-QG): Mimicking User Behavior in Multi-Turn Setting

1st Shujing Dong
Amazon
Seattle, WA, USA
shujdong@amazon.com

2nd Yuan Ling
Amazon
Seattle, WA, USA
yualing@amazon.com

3rd Shunyan Luo
Amazon
Seattle, WA, USA
shunyl@amazon.com

4th Shuyi Wang
Amazon
Seattle, WA, USA
wanshuyi@amazon.com

5th Yarong Feng
Amazon
Seattle, WA, USA
yarongf@amazon.com

6th Joe Liu
Amazon
Seattle, WA, USA
joeliu@amazon.com

7th Hongfei Li
Amazon
Seattle, WA, USA
lihongfe@amazon.com

8th Ayush Goyal
Amazon
Seattle, WA, USA
ayushg@amazon.com

9th Bruce Ferry
Amazon
Seattle, WA, USA
bferry@amazon.com

***Abstract*—This paper introduces a Context-Aware and User Intent-Aware follow-up Question Generation (CA-UIA-QG) method in multi-turn conversational settings. Our CA-UIA-QG model is designed to simultaneously consider the evolving context of a conversation and identify user intent. By integrating these aspects, it generates relevant follow-up questions, which can better mimic user behavior and align well with users' conversational goals. When assessed using public Shopping datasets on Fashion domain, our approach demonstrates significant enhancements over CA-QG baseline models. Specifically, it achieves an improvement of up to 3% in BLEU, 7% in METEOR, and 8% in ROUGE-Lsum. Additionally, our findings show the efficacy of fine-tuning in enhancing the model's capacity to better mimic user behavior, CoT prompting with fine-tuned model yields superior performance compared to the ensemble method. Furthermore, we investigate the impact of model size, model type, and intent granularity, highlighting their impact to overall model performance. The importance of our work lies in its effectiveness to improve follow-up question generation from the user's perspective and application in developing user-centric conversational AI systems.**

***Index Terms*—multi-turn conversation, follow-up question generation, user-centric conversational AI**

I. Introduction

The recent advancements in foundation models, especially Language Models (LMs), such as BERT [8], T5 [18], GPT-3 [4], GPT-4 [15] etc., have demonstrated remarkable capabilities in various natural language processing tasks and empowered to many assistant tools like Amazon Alexa, Apple Siri, Bard, ChatGPT. The increasing intricacy and sophistication of those AI systems are indeed driving revolutionary changes to the way we interact with them. They have evolved beyond single-turn interactions and are now capable of engaging in multi-turn conversations while incorporating diverse forms of data including text, images, audio, etc.

Given the multi-turn and multi-modal capabilities of current conversational AI systems, various methods have been developed to benchmark their performance in different scenarios. Methods such as HELM [11], BIG-bench [19], MT-Bench [22] primarily focus on evaluating LMs in a closed multi-turn setting. In a multi-turn open-ended generation setting, Agent-Bench [13] evaluated LLM-as-Agent's reasoning and decision-making abilities in eight different scenarios. However, there is a noticeable gap in research specifically addressing in the integration and performance of LMs in complex, real-world scenarios where user interactions are more dynamic and unpredictable comparing with the curated evaluation data. Closing this gap and developing evaluation methods that better reflect real-world complexities like user intents is a valuable direction for advancing the field of conversational AI. Developing metrics that align with the intricacies of human-machine conversations is crucial. This ensures that advanced conversational AI systems not only generate accurate and contextually relevant responses but also deliver an overall satisfactory and engaging user experience. Previous approaches for automated dialogue evaluation at turn-level or dialogue-level cannot adequately capture the dynamic nature of conversational AI [9]. The evolving context, the cumulative impact of multiple turns, and the intricacies of user intent in extended conversations pose challenges that demand more nuanced evaluation strategies. PERSONA-CHAT [21], which endows dialogue data with a persona by profile, offers a method to generate personalized dialogues for benchmarking. Such ability to simulate users with diverse intentions and generate questions automatically from users' personalized perspectives is a pivotal component in the multi-turn conversations benchmarking process, as showed in Figure 1.

In this paper, we introduce a Context-Aware and User Intent-Aware follow-up Question Generation (CA-UIA-QG) method to mimic user behavior in a multi-turn conversational setting. Our approach is uniquely tailored to generate follow-up questions that align closely with the user's intent, which

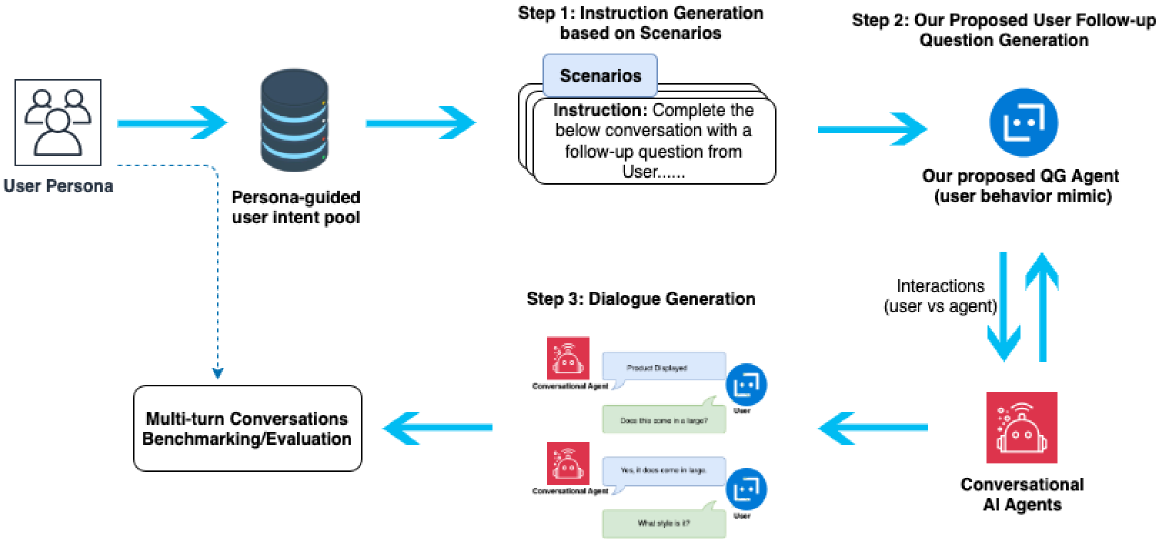

Step 1: Instruction Generation based on Scenarios
Step 2: Our Proposed User Follow-up Question Generation
User Persona
Persona-guided user intent pool
Scenarios
Instruction: Complete the below conversation with a follow-up question from User......
Our proposed QG Agent (user behavior mimic)
Interactions (user vs agent)
Step 3: Dialogue Generation
Conversational Agent
Product Displayed
User
Does this come in a large?
Conversational Agent
Yes, it does come in large.
User
What style is it?
Multi-turn Conversations Benchmarking/Evaluation
Conversational AI Agents

Fig. 1: An use case illustration of our proposed CA-UIA-QG method.

can be instructed by various user types and persona. In summary, the contribution of this work are two-fold: 1) To the best of our knowledge, this is the first study demonstrating a methodological approach to generate follow-up questions from the user's perspective. Our work lays the foundation for developing dialogues that represent user-AI agent interactions, uniquely reflecting various user types and personas. This is achieved by accurately capturing and responding to user intent at each turn of the conversation. Such an approach marks a significant step forward in creating more nuanced and user-centric Benchmarking process/evaluation metrics for conversational AI systems. 2) By combining context-awareness with user intent analysis, our approach generates follow-up questions that are closely aligned with user interests and conversational goals, thereby mimicking natural user behavior more effectively. Our experimental results demonstrated the effectiveness of our proposed CA-UIA-QG method.

II. Problem Statement

Our aim is to construct methods capable of generating personalized follow-up user questions that can interact with conversational AI agents in a dynamic, multi-turn conversational environment.

Our approach hinges on leveraging both the dialogue context and the user's intent at each turn. This strategy ensures that the questions generated for users are not only contextually informed but also highly personalized, thus accurately mirroring real user behavior.

A. Tasks Description

The problem can be defined as follows: Given a sequence of dialogue turns

$$D_{n-1} = \{U_1, A_1, U_2, A_2, ..., U_{n-1}, A_{n-1}\}, \qquad (1)$$

where U_i and A_i denotes user question and agent response (in this work, we assume AI-agents provide text-only response, and our method can be generalized to other types of responses), respectively, our objective is to learn a mapping function F that generate a user follow-up question U_n based on the preceding dialogue context D_{n-1}. Besides the dialogue turns D_{n-1}, users have different follow-up intention to the same dialogue, we denote such intent as I_n and discuss the mapping F in three different scenarios:

Scenario 1: CA-QG. In this scenario, the task is to predict the user question directly based on dialogue history without user intent:

$$U_n = F(D_{n-1}) \qquad (2)$$

Scenario 2: CA-UIA(latent)-QG. In this scenario, the object is to predict user question in two steps: first predict the user intent as an intermediate goal and then utilize this predicted user intent along with the previous dialogue turns to generate the subsequent user question.

$$U_n = F(D_{n-1}, \hat{I}_n) \qquad (3)$$

where $\hat{I}$ denotes the predicted user intent.

Scenario 3: CA-UIA-QG. This task is to predict the user question directly using the provided user intent and the contextual information from the previous dialogue turns. The user intent at the current turn I_n is explicitly provided:

$$U_n = F(D_{n-1}, I_n) \qquad (4)$$

B. User Intent Types Definition

According to the shopping datasets we use in Section III-E, three user intent types are defined based on the depth and breadth of information they encompass in granularity:

Coarse-Grained Intent Types I: At the highest level of abstraction, Coarse-Grained Intent Types I provide a broad understanding of user intentions. These intents serve as a foundational framework for classifying user interactions into fundamental categories such as seeking information (e.g., ”ASK”), confirming details (e.g., ”CONFIRM”), providing information (e.g., ”INFORM”), prompting system actions (e.g., ”PROMPT”), making direct requests (e.g., ”REQUEST”), and indicating errors or uncertainties (e.g., ”Misc”). Coarse-Grained Intent Types I offer a high-level overview of user engagement patterns.

Coarse-Grained Intent Types II (with additional actions): Building upon the Coarse-Grained Intent Types I, Coarse-Grained Intent Types II introduce a finer level of granularity by including additional specific actions and behaviors within each intent category. These actions provide a more detailed characterization of user intentions. For instance, within the “ASK” category, users may inquire about various aspects, including checking prices, adding items to a cart, comparing products, and more. Coarse-Grained Intent Types II serve to capture a broader range of user behaviors, offering a deeper understanding of user interactions.

Fine-Grained Intent Types: At the most granular level, Fine-Grained Intent Types provide a detailed breakdown of user intent by incorporating specific attributes, parameters, and details related to the interaction. Fine-Grained Intent Types delve into the specifics, such as inquiring about the price of clothing items in different sizes, checking the availability of particular furniture configurations, or seeking detailed product information. These intent types aim to capture the intricate nuances of user requests, enabling a highly detailed analysis of user behavior and preferences.

This hierarchical categorization of intent types allows for a comprehensive examination of user interactions, from a high-level understanding of user goals to a fine-grained explanation of user queries and preferences. See details in Table VI and Table VII in Appendix A.

III. User Follow-Up Question Generation Methodology

In this section, we describe our proposed methods for different scenarios discussed in Section II-A.

A. CA-QG

We use text generation with LMs for user question generation. Table I provides details of input format for user question generation. First we concatenate the dialogue history into a text sequence, a model M is to generate user’s follow-up question, given the task instruction and the dialogue context.

TABLE I: Model input for question generation in CA-QG.

Task	Input
Instruction:	Complete the below conversation with a follow-up question from User.
Context:	System : This one is the highest rated, at 4.84. User : Who designs it? System : This is made by Fancy Nails. User :

B. CA-UIA(latent)-QG

For Scenario 2, we have two goals: user intent prediction and user question generation. Given the assumption that understanding user’s intent can be beneficial to the question generation, we use two methods to generate latent intents.

1) Ensemble method: This method utilizes two specialized models, each fine-tuned for a specific part of the task. As showed in Table II, The first model focuses on predicting the user’s intent based on the dialogue history. The output of this model – the predicted intent – then serves as an input for the second model, which is responsible for predicting the user’s next question.

TABLE II: Input format for question generation in Scenario 2 task with ensemble method.

Stage	Input
Model 1	*Model Input:* System : This one is the highest rated, at 4.84. User : Who designs it? System : This is made by Fancy Nails. User : *Expected Output:* ASK-CHECK: CLOTHING.availableSize
Model 2	*Instruction:* Complete the below conversation with a follow-up question from User. Considering user intent when generating the question. *Context:* System : This one is the highest rated, at 4.84. User : Who designs it? System : This is made by Fancy Nails. User : *User Intent:* the intent of user question is to ask and check the available size of the clothing

C. Chain-of-Thought (CoT) Method

CoT prompting [20] mimics human reasoning processes, which perfectly aligned with our assumption that before we generate user question, we first get user’s intent, then the intent can guide us to better user question generation. In this method, the model is prompted to generate a “chain of thought” – a sequence of intermediate steps or reasonings – leading to the final prediction.

As showed in Table III, we design few-shot CoT and zero-shot CoT for infering user intent and generating user question.

D. CA-UIA-QG

For scenario 3 task, we have given user intent as model input, then the inputs for next question generation is similar as the input defined for Model 2 in Table II. Instead of having a predicted user intent, here we have the intent given explicitly in our CA-UIA-QG model here.

1) Model Training: To capture the dependence of user question generation on text input of dialogue context, user intent and instruction, we use ground truth user question $\hat{U}_n$, a loss function L tries to maximise similarity between generated user question U_n and ground truth question $\hat{U}_n$. Due to the impracticality of retraining large LMs in terms of data and computational resources, fine-tuning is a widely

TABLE III: CoT Prompting for question generation in CA-UIA (latent)-QG.

Few-shot-CoT	**Q:** User : I'm looking for a nice neutral coloured jacket, but none of these are doing it for me. Is there anything else you could recommend? System : How about this sweater by Home Store? It's $254. What size are you looking for? User : ***A: Let's think step by step. First, predict next-turn user intent, then based on intent, predict user question.*** Here is an example: **Q:** System : This one is the highest rated, at 4.84. User : Who designs it? System : This is made by Fancy Nails. User : **A:** given the context, the predicted follow-up user intent is to ask and check the available size of the clothing. Thus the predicted user question: Is it available in XL?
	(Output) given the context, the predicted follow-up user intent is to inform user's preference about clothing size. Thus the predicted user question: Oh I like that one. I'm an xs, but I can make a small work too
Zero-shot-CoT	**Q:** User : I'm looking for a nice neutral coloured jacket, but none of these are doing it for me. Is there anything else you could recommend? System : How about this sweater by Home Store? It's $254. What size are you looking for? User : ***A: Let's think step by step. First, predict next-turn user intent, then based on intent, predict user question.***
	(Output) given the context, the predicted follow-up user intent is to inform user's preference about clothing size. Thus the predicted user question: Oh I like that one. I'm an xs, but I can make a small work too

adopted approach to enhance the performance of pretrained LMs in specific scenarios. In cases where the model sizes are relatively smaller, we employ full-model fine-tuning. For larger models, we adopt Low-Rank Adaptation (LoRA) fine-tuning method [7], [10] to achieve efficient fine-tuning. We also demonstrated the comparisons between full fine-tuning and LoRA fine-tuning in Appendix C.

E. Tasks and DataSets

We use a public available shopping dataset for our experiments: SIMMC datasets (Situated Interactive MultiModal Conversations) [14]. SIMMC datasets is in the domain of interactive shopping and is collected using the SIMMC platform [6]. It covers two domains: 1) Fashion and 2) Furniture. The SIMMC datasets have originally been utilized for three tasks: 1) Structural API call prediction; 2) Response generation; and 3) Dialog state tracking. All three tasks are targeting on predicting for AI agents. We further extend the datasets to serve our follow-up question generation task, which concentrates on understanding and predicting the behavior and intentions of the user in the conversation.

As showed in Table IV, we processed the SIMMC datasets with Fashion domain for user question generation tasks. The datasets include training and test. For given context, coarse_intent_I, coarse_intent_II, and fine_intent_I, our task aims at predicting the user_question, which represents the follow-up user question in the given context.

TABLE IV: SIMMC Fashion Datasets for User Question Generation Task.

Dataset	Total Conversations	Total Turns
Training	3,929	17,267
Test	982	4,415

F. Evaluation Metrics

We employed a set of automated metrics for assessing the performance of our question generation outputs. These metrics include BLEU [16], METEOR [3], ROUGE-1, ROUGE-L, and ROUGE-Lsum [12] as automated metrics for our evaluation.

G. Selection of LMs

We use three types of pre-trained Language Models (LMs).

FLAN-T5 Series Flan-T5 models [5] are enhanced versions of T5 [18] that has been finetuned in a mixture of tasks. We experiment with a set of Flan-T5 models[1] including flan-t5-base (220M), flan-t5-large (0.8B), flan-t5-xl (3B), and flan-t5-xxl (11B).

Falcon Series Falcon-7B and Falcon-40B [1] are causal decoder-only model built by TII [17].

Claude Claude2 [2] are general purpose large language models. They use a transformer architecture and are trained via unsupervised learning, RLHF, and Constitutional AI (including both a supervised and Reinforcement Learning (RL) phase).

For the Flan-t5 and Falcon series models that exceed 3 billion parameters, constraints in computational resources, coupled with the observation that the performance gap between LoRA fine-tuning and full fine-tuning narrows as model size increases, led us to choose LoRA fine-tuning for these larger models, which is further elaborated in Section Appendix C. For Claude 2 we use it directly via Amazon Bedrock [2].

IV. Experimental Results

In this section, we demonstrate our overall experimental results in Table V and discuss the observations made based on results.

Significant performance edge of CA-UIA-QG model over CA-QG model.

In our comparative analysis of the CA-UIA-QG and CA-QG models, it's evident that the CA-UIA-QG model outperforms CA-QG model significantly. Particularly the LoRA-FT-Flan-t5-XXL variant, showcases a remarkable performance

[1] https://huggingface.co/docs/transformers/model_doc/flan-t5

[2] https://aws.amazon.com/bedrock/claude/

advantage. For instance, with LoRA-FT-Flan-t5-XXL, the CA-UIA-QG model achieved performance gains compared to CA-QG model with 3.2% of BLEU, 7.4% of METEOR, 8.7% of ROUGE-1, 8.6% of ROUGE-L, and 8.7% of ROUGE-Lsum. This trend of superior performance by the CA-UIA-QG model is consistently observed across different pre-trained LMs, including all flan-t5 series baseline models, falcon-7b baseline models, Claude 2 - Zero-Shot, and LoRA fine-tuning models. The significant performance advantage of CA-UIA-QG model over CA-QG model underscores the efficacy of our user intent-aware approach in question generation. This highlights how effectively the CA-UIA-QG model, with its dual focus on evolving conversational context and user intent recognition, contributes to generating more pertinent follow-up questions. Such an approach not only closely mimics user behavior but also aligns more accurately with the user's conversational objectives, demonstrating the benefits of integrating both contextual understanding and intent awareness in our model.

Fine-tuning improves performance. Our experimental findings demonstrate a noteworthy enhancement in model performance through the application of fine-tuning. For CA-QG, the performance of the fine-tuned model demonstrates a significant elevation. Specifically, LoRA finetuned Flan-t5-xxl with integrated context surpasses the baseline by 4.1% in BLEU, 9.8% in METEOR, 8.3% in ROUGE-1, 7.5% in ROUGE-L, and 7.5% in ROUGE-Lsum. A similar positive trend is evident in CA-UIA-QG, where the model improvement of fine-tuned Flan-t5-xxl with preceding turns and intent records a 7.0% gain in BLEU, 16.1% in METEOR, 15.8% in ROUGE-1, 15.1% in ROUGE-L, and 15.1% in ROUGE-Lsum. CA-UIA(latent)-QG further highlights the benefits derived from fine-tuning. In this scenario, the objective is to generate the next turn without explicit intent information for that turn. Significant improvements are observed in the LoRA-fine-tuned Flan-t5-xxl model compared to the non-fine-tuned ones. this fine-tuned model outperforms Flan-t5-xxl, yielding a 12.8% improvement in ROUGE-1 and a 12.1% improvement in ROUGE-Lsum, while maintaining a consistent BLEU and METEOR score. Additionally, LoRA-fine-tuned Flan-t5-base with only 220 million parameters, surpasses Flan-t5-xxl with 11 billion parameters in all five metrics. The superior performance of CA-UIA-QG model fine-tuned with context and intent in all experiments has proven the effectiveness of fine-tuning in infusing knowledge and pattern from training data to mimic user behavior. Even in scenarios where the input lacks explicit intent information, the fine-tuned CA-UIA-QG model consistently outperforms pre-trained models in generating the next turn.

CoT outperforms ensemble model. In practical applications, the reliable availability of next-turn intent remains a significant challenge. In instances where explicit next-turn intent is not provided, selecting an optimal model for generating subsequent turns becomes crucial. The outcomes of CA-UIA(latent)-QG underscore the efficacy of applying CoT to a model fine-tuned for both context and intent. This approach surpasses the performance of an ensemble model incorporating a BERT classifier trained to predict next-turn intent, and a finetuned model to generate next turn with predicted intent and preceding turns. The superior performance of CoT is posited on the potential inadequacy of intent prediction accuracy (discussed in Section V-C) in effectively guiding the generation of subsequent turns. In comparison with ensemble method, CoT instructs the model to first predict the intent and then generate the next part of the conversation based on that decision. This approach takes advantage of the finetuned model's learning of intent and context in the training data, making it better at mimicking how real users naturally talk.

V. Ablation Studies

A. *Importance of Model Size*

In our analysis of the T5 Series models, we noted a consistent trend: an increase in model size leads to a corresponding improvement in performance. This pattern holds true in the baseline results for both the CA-QG Model and the CA-UIA-QG model. Furthermore, this pattern is also apparent in the results of full fine-tuning and LoRA fine-tuning methods, as illustrated in Figure 3. Similarly, within the Falcon series models, Falcon-7b and Falcon-40b, we observed the same phenomenon. An increase in the model size is paralleled by enhanced performance.

B. *Impact of Pre-trained LMs*

In our experiments, Flan-t5-xxl model with fine-tuning has the best performance, compared to Falcon and Claude 2. For the same type of models, we observe the pattern that the larger the model is, the better performance is. But that's not true when we have different types of models. As showed in Table V, fine-tuning Falcon-40b has no better results than fine-tuning flan-t5-xxl (11B), the first one has larger parameters size. Similarly, Claude 2 yields the lowest performance results across all scenarios despite being the largest model. Across all experiments, it is evident that the Flan-t5 series consistently produce superior results compared to Claude 2. This discrepancy may arise from the nature of next turns in shopping conversations, which tend to be prototypically short. Claude 2, the largest model size in this experiments, generates longer turns and infuses knowledge from its knowledge base that may not be present in the test dataset. For instance, with a given intent "chitchat", Claude 2 generates user next turn like "Thank you for sharing the details about the chairs. Since we're just chatting, may I ask what your plans are for the holidays this year?", which is not situable for the shopping scenario and much longer than the ground truth turn "Thank you for your help". Claude 2 also sometimes doesn't take on the "User" role, instead of generating user's question, it outputs a turn as the AI assistant. For example, Claude 2 generates user's next turn as "What kind of material and fit do you prefer for the jacket? I can look for some options with neutral colors in those styles." which takes on the assistant role, not the user role defined in the instruction.

TABLE V: Overall Model Performance.

Model	Model Size	BLEU	METEOR	ROUGE-1	ROUGE-L	ROUGE-Lsum
CA-QG:						
Without Model Fine-Tuning						
Flan-t5-base	220M	0.00790	0.10342	0.13708	0.12794	0.12779
Flan-t5-large	0.8B	0.01391	0.13330	0.17031	0.16082	0.16078
Flan-t5-xl	3B	0.01443	0.13595	0.17795	0.16964	0.16938
Flan-t5-xxl	11B	**0.01616**	**0.14085**	**0.18282**	**0.17352**	**0.17349**
Falcon-7b	7B	0.01169	0.12254	0.12715	0.11262	0.11268
Falcon-40b	40B	**0.01625**	**0.13504**	**0.14488**	**0.12998**	**0.13001**
Claude 2 - Zero-Shot	130B	0.00612	0.12497	0.11714	0.09795	0.09860
Claude 2 - One-Shot		**0.01413**	**0.13294**	**0.14579**	**0.12806**	**0.12831**
With Model Fine-Tuning						
LoRA-FT-Flan-t5-xxl	11B	**0.05680**	**0.23847**	**0.26533**	**0.24838**	**0.24810**
LoRA-FT-Falcon-7b	7B	0.04972	0.22942	0.24816	0.23116	0.23112
CA-UIA(latent)-QG:						
Ensemble:						
Bert + Flan-t5-xxl	11B	0.01771	0.14453	0.18385	0.17323	0.17310
Bert + LoRA-Flan-t5-xxl	11B	0.05008	0.22626	0.25356	0.23652	0.23664
CoT:						
zero-shot-Flan-t5-xxl	11B	0.01395	0.12909	0.16883	0.15839	0.15839
zero-shot-LoRA-Flan-t5-base	220M	0.04060	0.25821	0.28490	0.24806	0.24790
zero-shot-LoRA-Flan-t5-xxl	11B	0.03845	0.25770	**0.29702**	**0.28030**	**0.27942**
zero-shot-Claude 2	130B	0.01100	0.11311	0.12545	0.10812	0.10805
dynamic-few-shot-Claude	130B	0.01561	0.14842	0.14431	0.12332	0.12339
CA-UIA-QG:						
Without Model Fine-Tuning						
Flan-t5-base	220M	0.00779	0.10376	0.13871	0.13020	0.13018
Flan-t5-large	0.8B	0.01606	0.13983	0.18076	0.16897	0.16899
Flan-t5-xl	3B	0.01693	0.14381	0.18765	0.17800	0.17834
Flan-t5-xxl	11B	**0.01913**	**0.15199**	**0.19441**	**0.18346**	**0.18366**
Falcon-7b	7B	0.01571	0.15535	0.14769	0.13148	0.13148
Claude 2 - Zero-Shot	130B	0.01376	0.17766	0.16547	0.13774	0.13776
Claude 2 - One-Shot		**0.01492**	**0.20657**	**0.17949**	**0.14879**	**0.14927**
With Model Fine-Tuning						
LoRA-FT-Flan-t5-xxl	11B	**0.08899**	**0.31249**	**0.35239**	**0.33466**	**0.33473**
LoRA-FT-Falcon-7b	7B	0.06941	0.27778	0.29721	0.27823	0.27792

C. *Different Intent Types Granularity*

In the ensemble method, our model performance on coarse-grained intent types I, coarse-grained intent types II, and fine-grained intent types achieved accuracy scores of 59%, 52%, and 54% accordingly (see details in Appendix B). We see most error cases of coarse-grained intent types I are in distinguishing between "ASK" and "INFORM" intents, as well as between "ASK" and "REQUEST" intents. In experiments, We observed that combining fine-grained intent types with coarse-grained intent types in the instructions to models can lead better results.

VI. Conclusion

We introduce methods capable of generating context-aware and intent-aware follow-up questions. Our experiments, conducted on a publicly available shopping dataset, have demonstrated the effectiveness of our approach. This advancement not only improves user behavior mimic in conversational settings but also opens new avenues for more personalized and dynamic interactions with AI agents, paving the way for more intuitive and user-centered benchmarking and evaluation process with conversational AI systems. Our future work includes using user types/persona to guide the turn-level user intent and follow-up question generation, and generating user follow-up actions and questions in multi-modalities settings.

References

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[2] Anthropic. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf, 2023. Accessed: 2023-12-05.

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[6] Paul A Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. Simmc: Situated interactive multimodal conversational data collection and evaluation platform. *arXiv preprint arXiv:1911.02690*, 2019.

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Edward Guo. Robertaiq: An efficient framework for automatic interaction quality estimation of dialogue systems. 2021.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[11] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[13] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

[14] Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, et al. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*, 2020.

[15] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[17] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[19] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[21] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Appendix

A. Appendix: Intent Types

The details of Coarse-Grained Intent Types are showed in Table VI and Fine-Grained intent types are showed in Table VII.

TABLE VI: Coarse-Grained Intent Types

Coarse-Grained Intent Types I	**Coarse-Grained Intent Types II (with additional actions)**
ASK	ASK-ADD_TO_CART
	ASK-CHECK
	ASK-COMPARE
	ASK-DISPREFER
	ASK-GET
	ASK-PREFER
	ASK-REFINE
CONFIRM	CONFIRM-ADD_TO_CART
	CONFIRM-CHECK
	CONFIRM-COMPARE
	CONFIRM-GET
	CONFIRM-PREFER
	CONFIRM-REFINE
INFORM	INFORM-ADD_TO_CART
	INFORM-CHECK
	INFORM-DISPREFER
	INFORM-GET
	INFORM-PREFER
	INFORM-REFINE
PROMPT	PROMPT-ADD_TO_CART
	PROMPT-DISPREFER
	PROMPT-GET
	PROMPT-PREFER
	PROMPT-REFINE
REQUEST	REQUEST-ADD_TO_CART
	REQUEST-CHECK
	REQUEST-COMPARE
	REQUEST-GET
	REQUEST-REFINE
Misc	Misc-CHITCHAT
	Misc-MISSING_LABEL
	Misc-MULTIPLE
	Misc-UNSUPPORTED
	Misc-UNSURE

B. Appendix: Confusion Matrix of Coarse-Grained Intent Classification

C. Appendix: Full Fine-Tuning vs LoRA Fine-Tuning

We conducted a comparative analysis of performance with CA-UIA-QG model using full fine-tuning and LoRA fine-tuning techniques across three models: Flan-t5-base, Flan-t5-large, and Flan-t5-xl. Figure 3 reveals that full fine-tuning uniformly outperforms LoRA fine-tuning across all models. However, it's noteworthy that as the model size increases, the performance gap between the two fine-tuning methods narrows. This trend suggests that the relative advantage of full fine-tuning diminishes with larger model sizes.

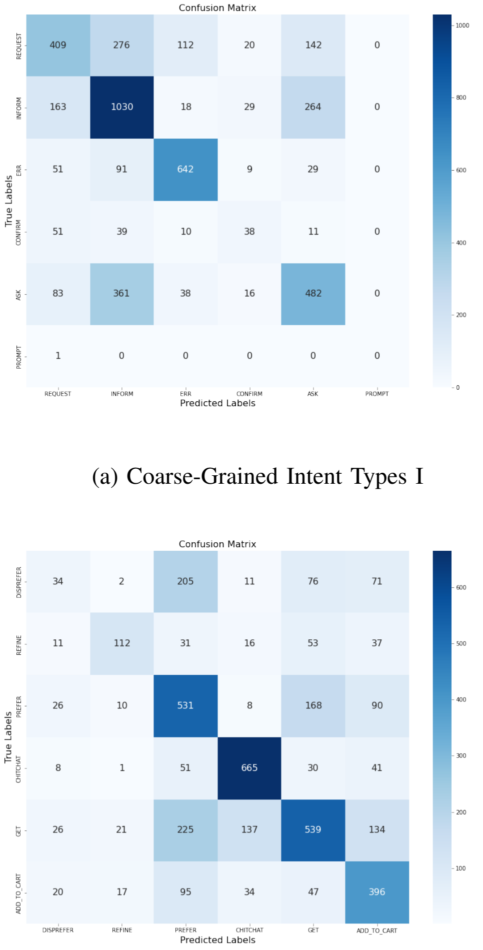
An example in the datasets mentioned in Section III-E is showed in Table VIII. For given context, coarse_intent_I, coarse_intent_II, and fine_intent_I, our task aims at predicting the user_question, which represents the follow-up user question in the given context.

TABLE VII: Fine-Grained Intent Types

Fine-Grained Intent Types
CLOTHING.price
CLOTHING.availableSizes
CLOTHING.customerRating
CLOTHING.info
CLOTHING.size
CLOTHING.color
CLOTHING.brand
CLOTHING.hemLength
CLOTHING.embellishment
CLOTHING.pattern
CLOTHING.skirtStyle
...
FURNITURE.price
FURNITURE.color
FURNITURE.dimensions
FURNITURE.info
FURNITURE.material
FURNITURE.decorStyle
FURNITURE.intendedRoom
ATTRIBUTE
FURNITURE.hasStorage
FURNITURE.height
FURNITURE.width
...

TABLE VIII: An Example in SIMMC Datasets for User Question Generation Task.

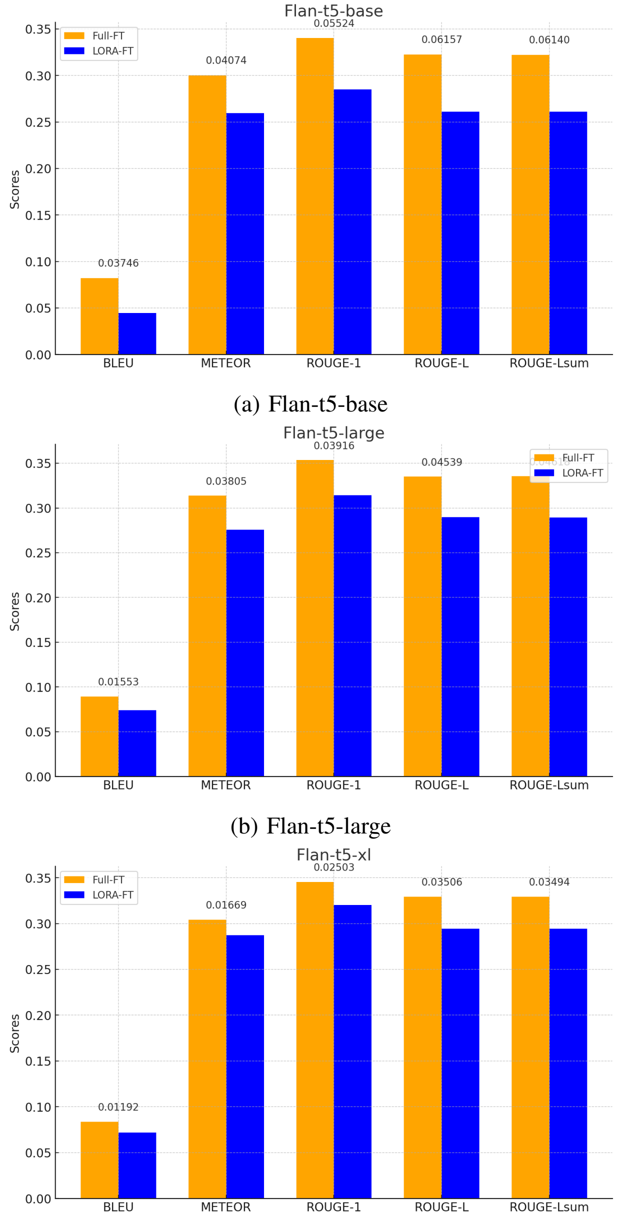
CONTEXT	System : This one is the highest rated, at 4.84. User : Who designs it? System : This is made by Fancy Nails. User :
USER_QUESTION	Is it available in XL?
COARSE_INTENT_I	ASK
COARSE_INTENT_II	ASK-CHECK
FINE_INTENT_I	CLOTHING.availableSizes



(a) Coarse-Grained Intent Types I

(b) Coarse-Grained Intent Types II

Fig. 2: Confusion Matrix of Coarse-Grained Intent Classification (Fashion domain).



(a) Flan-t5-base

(b) Flan-t5-large

(c) Flan-t5-xl

Fig. 3: CA-UIA-QG Model Performance of Full Fine-Tuning vs LoRA Fine-Tuning.