Towards Proactive Interactions for In-Vehicle Conversational Assistants Utilizing Large Language Models

Huifang Du[1], **Xuejing Feng**[1], **Jun Ma**[1,*], **Meng Wang**[1], **Shiyu Tao**[2], **Yijie Zhong**[1], **Yuan-Fang Li**[3] and **Haofen Wang**[1,*]

[1]Tongji University
[2]Beijing Technology and Business University
[3]Monash University
{duhuifang, fengxuejing, jun_ma, mengwangtj}@tongji.edu.cn,
{dun.haski, carter.whfcarter}@gmail.com,
{ms_taoshiyu}@163.com,
{yuanfang.li}@monash.edu

Abstract

Research demonstrates that the proactivity of in-vehicle conversational assistants (IVCAs) can help to reduce distractions and enhance driving safety, better meeting users' cognitive needs. However, existing IVCAs struggle with user intent recognition and context awareness, which leads to suboptimal proactive interactions. Large language models (LLMs) have shown potential for generalizing to various tasks with prompts, but their application in IVCAs and exploration of proactive interaction remain under-explored. These raise questions about how LLMs improve proactive interactions for IVCAs and influence user perception. To investigate these questions systematically, we establish a framework with five proactivity levels across two dimensions—assumption and autonomy—for IVCAs. According to the framework, we propose a "Rewrite + ReAct + Reflect" strategy, aiming to empower LLMs to fulfill the specific demands of each proactivity level when interacting with users. Both feasibility and subjective experiments are conducted. The LLM outperforms the state-of-the-art model in success rate and achieves satisfactory results for each proactivity level. Subjective experiments with 40 participants validate the effectiveness of our framework and show the proactive level with strong assumptions and user confirmation is most appropriate.

1 Introduction

In-vehicle conversational assistants (IVCAs) are an integral component in smart cockpits and play a vital role in facilitating human-agent interaction [Lee and Jeon, 2022]. They can deliver features including navigation, entertainment control, and hands-free phone operation [Braun *et al.*, 2019].

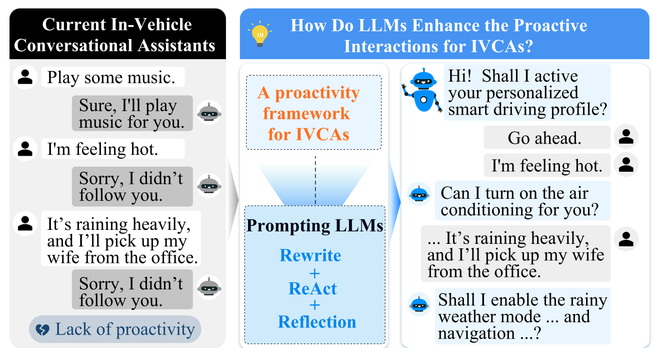

Figure 1: The motivation and main contributions of our work to exploring effective proactive interactions for IVCAs based on LLMs.

Despite the promising prospects and strides made in IVCAs' development, the proactive interactions from a human-centered perspective in the vehicle context are relatively limited. For example, existing IVCAs mostly passively receive and execute simple commands [Meck *et al.*, 2023; Meck and Precht, 2021; Lin *et al.*, 2018], though the proactive interaction concepts are proposed in some car manufacturers[1]. The issues above can be approached from two angles. **Firstly, current research lacks a clear and helpful definition of proactivity for IVCAs. Secondly, there are technical limitations to achieving satisfactory proactive interactions, such as poor intent recognition [Mi *et al.*, 2022] and context awareness [Shen *et al.*, 2022]**. As the demand for better interactions rises, IVCAs are expected to manage complex tasks and offer proactive support [Völkel *et al.*, 2021]. Particularly, through proactively providing information and addressing the anticipated issues[Kim *et al.*, 2020], IVCAs can offer personalized services and reduce driver cognitive load, thus improving driving safety and experience.

From the perspective of human-computer interaction (HCI) research, proactive behaviors of conversational assistants can

*Corresponding authors.

[1]https://www.mercedes-benz.de/passengercars/technology/mbux-zero-layer.html

be summarized into two essential elements [Peng *et al.*, 2019; Grant and Ashford, 2008; Parker and Collins, 2010]: *autonomy* (the capability to execute intended tasks), and *assumption* (anticipation of users needs). Many efforts center around these two dimensions. For instance, the Interface-Proactivity (IP) continuum was proposed to define five different proactivity levels of autonomy, ranging from zero to full autonomy [Isbell and Pierce, 2005]. Building upon the IP continuum, proactive dialogues are categorized into four levels: None, Notification, Suggestion, Intervention [Kraus *et al.*, 2021; Kraus *et al.*, 2020]. Besides, a three-level proactivity policy framework for decision-making support assistants was defined across the assumption and autonomy dimensions [Peng *et al.*, 2019]. Yet, how IVCAs in driving contexts proactively interact with users is still an open issue. Some studies try to understand the impact of proactivity on human-vehicle interaction from the viewpoint of interruptions [Kim *et al.*, 2020; Cha *et al.*, 2020]. They provide valuable case studies, but mainly focus on the timing and linguistic impacts, without offering comprehensive interaction strategies.

As for conversational technologies, extensive research has been conducted in the general domain [Young *et al.*, 2022], inspiring the studies of IVCAs. Still, conversation support for IVCAs, such as Adasa [Lin *et al.*, 2018] and CarExpert [Rony *et al.*, 2023], have mainly focused on providing driving-related knowledge with data derived from user manuals. In contrast, we try to address users' task-related requests in various driving scenarios. Furthermore, current dialogue models typically require large amounts of labeled data, incur high costs, and may not generalize well to other tasks. Recently, large language models (LLMs), such as GPT-4[2] and Vicuna[3], have shown impressive understanding and generation capabilities to many tasks with prompts (i.e. without updating model parameters). The ability to learn from limited data is highly advantageous, but few studies have been conducted to understand their viability for IVCAs.

In this paper, we investigate enhancing proactive interactions between users and IVCAs by mitigating the aforementioned two issues: the lack of clear proactivity definition for IVCAs and technical limitations. To thoroughly explore proactive interactions of IVCAs, it is imperative to establish a formal formulation and ensure a consistent implementation between IVCAs and users. Drawing on previous research [Peng *et al.*, 2019; Isbell and Pierce, 2005; Kraus *et al.*, 2021; Kraus *et al.*, 2020], we build a proactivity framework with five levels across assumption and autonomy dimensions while incorporating user control as a design constraint. Based on the framework, we investigate LLMs' feasibility in achieving different levels of proactivity for IVCAs. We prompt LLMs by proposing a "Rewrite + ReAct + Reflect" approach to get a response. Specifically, we first rewrite casual questions of users to be more normal in driving contexts. Then we not only prompt LLMs to reason by the proactive interaction instructions and search for external supportive knowledge but also make them reflect on whether the generated answers fulfill the designated level of proactivity. Our work also provides insights into how LLMs can integrate various in-vehicle information for understanding and decision-making tasks.

We extensively experiment with the LLM model gpt-3.5-turbo to investigate its capability to achieve various levels of proactivity for IVCAs. Results show that the LLM not only achieves a superior success rate (93.72%) than the state-of-art models for task-oriented dialogue but also satisfactory proactivity attainment rates for each proactivity level (more than 78%). Furthermore, we explore the effects of different levels of proactive interactions on human perception with 40 participants. For a more realistic setting, we develop an IVCA simulator based on the LLM to implement an actual conversation environment. Experimental results indicate that the proactivity level with strong assumptions and user confirmation is most preferred. As it offers natural and helpful assistance and user confirmations, it's considered the most appropriate. Notably, our work is the first to explore proactive interactions for IVCAs using LLMs, verifying the potential of LLMs for IVCAs. In summary, the main contributions are as follows:

- We establish a proactivity framework for IVCAs with five levels along the dimensions of assumption and autonomy while integrating user control as a design principle. The framework lays a theoretical foundation for systematically studying proactive interactions for IVCAs.
- To our knowledge, we are the first to investigate the potential of LLMs in improving proactive interaction experiences for IVCAs. We utilize a "ReAct + Reflect" strategy to prompt LLMs to achieve various levels of proactivity with satisfactory performance.
- Comprehensive experimental results show that our approach is feasible to enhance the interaction experience for users. We observe that proactivity significantly influences user perceptions and users prefer proactive interactions with strong assumptions and user control.

2 Related Work

The evolution of IVCAs has been a significant research subject within the context of human-computer interaction (HCI) and artificial intelligence (AI). This section explores related work in the areas of proactive interaction, prompting LLMs.

2.1 Proactivity of the Intelligent Assistants

Proactivity is determined by the following two factors: assumption, and autonomy in the domain of occupational and organizational psychology [Grant and Ashford, 2008; Parker and Collins, 2010]. Based on the two elements, proactive behaviors of assistants are often discussed in HCI [Peng *et al.*, 2019; Kraus *et al.*, 2021]. Among these studies, the challenges of *if*, *how*, and *when* to take proactive action for dialogue assistants are proposed [Nothdurft *et al.*, 2014] and become the guidelines for designing proactive assistants then. The *if* question stresses the necessity. Many studies demonstrate that proactive behaviors of an assistant system affect the user's perception [Peng *et al.*, 2019; Kraus *et al.*, 2020] and proactivity is considered one of the users' desired features for perfect assistants [Zargham *et al.*, 2022]. Regarding the *how* research, some works give examples to answer the *how* question [Zargham *et al.*, 2022;

[2]https://arxiv.org/abs/2303.08774
[3]https://lmsys.org/blog/2023-03-30-vicuna/

Meck *et al.*, 2023], but they mainly focus on specific features like linguistic styles, tone of voice, gestures, etc. There are also some works developing guidelines for general dialogues between humans and assistants [Isbell and Pierce, 2005; Peng *et al.*, 2019; Kraus *et al.*, 2020; Kraus *et al.*, 2021]. As for the research question of *when*, some studies try to find the balance between being helpful and being intrusive decided by proactivity from the viewpoint of interruptions [Kim *et al.*, 2020] and linguistic impacts [Cha *et al.*, 2020]. In this paper, we focus on building proactive interaction strategies tailored for IVCAs to respond to the *how* research question, which also lays the groundwork for opportune interactions.

2.2 Prompting Large Language Models

Recently, large language models (LLMs) have shown emergent abilities [Schaeffer *et al.*, 2023] and have led to a new paradigm in creating natural language processing systems. Unlike traditional methods that rely on a well-selected, labeled training dataset, LLMs have introduced a new technique, prompt engineering. In-context learning (ICL), prompting LLMs with a few examples [Dong *et al.*, 2022], can generalize to various tasks like summarization, question answering, and code generation without updating parameters. ICL is adopted in our study to transform users' diverse and casual expressions into formal questions. More Helpful prompting techniques are proposed to interface with LLMs [Wei *et al.*, 2022; Yao *et al.*, 2023; Yao *et al.*, 2022]. For example, chain of thought (CoT) [Wei *et al.*, 2022] shows intermediate reasoning steps of the examples to boost the prompting performance. Using the technique of Tree of Thoughts (ToT), LLMs can make thoughtful decisions by considering many different reasoning paths and self-evaluating options [Yao *et al.*, 2023]. The ReAct prompting framework leverages LLMs to produce reasoning traces and task-specific actions while enabling the collection of external information [Yao *et al.*, 2022]. It also enhances the trustworthiness and interoperability of LLMs by using the problem-solving process. We adopt ReAct prompting in our work to incorporate external knowledge and implement proactive interaction strategies. Additionally, we include a reflective function at the end to ensure that LLMs align with the desired level of proactivity.

3 Design of Proactive Interaction Strategies

In this section, we formulate proactive interaction behaviors reflecting the unique characteristics of interacting with IVCAs in driving contexts. Specifically, we apply the concept of proactivity, originally from the field of occupational and organizational psychology [Grant and Ashford, 2008; Parker and Collins, 2010], to the domain of HCI [Peng *et al.*, 2019], considering two essential factors: autonomy and assumption. The first factor, system autonomy, which refers to the ability to perform tasks without direct user commands, has been the subject of study in various earlier works. These include the autonomy scale definition in [Rau *et al.*, 2013], the IP continuum in [Isbell and Pierce, 2005], the three-level proactivity framework in [Peng *et al.*, 2019], and four-level proactivity in [Kraus *et al.*, 2020]. We follow the principles of autonomy as outlined in these works when designing the proactive behaviors of IVCAs. Regarding the system assumption, it is closely associated with the ability to anticipate users' potential intentions [Kraus *et al.*, 2020]. Many methods utilize human actions or poses, such as gaze and body positioning, to make predictive inferences. We attempt to make assumptions by comprehending the driver's utterances in driving contexts. Building upon prior work, we design the proactivity scales for IVCAs based on assumptions and autonomy as well. However, considering the direct implications of IVCAs on driving safety and user experience, we particularly account for the importance of user control [Kraus *et al.*, 2021; Kraus *et al.*, 2020]. We incorporate user control as a design principle or constraint, dividing the levels of proactivity into five levels based on assumptions and autonomy. Within each proactive level, we discuss the degree and manner of user control. The proactive interaction guidelines at the five levels are derived as follows:

Level 1. At this level, IVCAs make no assumptions and passively receive and execute the user's instructions. The user has full control over the behavior of IVCAs, and IVCAs will not take any action without instructions. For instance, "Driver: Please turn on the air conditioner. IVCAs: Sure."

Level 2. IVCAs at this level demonstrate some assumptions, which means IVCAs make preliminary judgments based on limited utterance information. They may identify potential issues or suggest possible solutions based on the assumptions. However, they rely on the user's confirmation before taking any proactive steps. For example, "Driver: I'm feeling hot. IVCAs: Shall I activate the air conditioning for you? Driver: Go ahead."

Level 3. IVCAs at this level show the same level of assumption ability as level 2. However, they can automatically take actions with minimal user inputs during the interaction, and they will execute actions based on these inputs. For instance, "Driver: I'm feeling hot. IVCAs: I will activate the air conditioning for you. How about 25 degrees Celsius okay? Driver: Sounds good. Thanks."

Level 4. At this level, IVCAs become highly adaptive, making assumptions based on extensive historical data and deep learning of user behavior. They may initiate conversations and offer suggestions, like providing personalized entertainment options and adjusting responses according to user preferences. However, users retain the right to confirm or adjust proposals before execution. For example, "IVCAs: Would you like me to set the air conditioning to your preferred temperature of 25 degrees Celsius? Driver: Yes, that would be helpful. IVCAs: The temperature has been set."

Level 5. IVCAs are adaptive at this level with strong assumptions like the level at 4. Additionally, they have high autonomy to execute their assumptions automatically with some explanations. However, users can still intervene to stop execution. For example, "IVCAs: You're in the car. I'll adjust the air conditioning to your preferred temperature of 25 degrees Celsius. Driver: No, thanks."

Our proactive interaction framework, clearly delineating five levels of proactivity, provides more specific guidance for the design of IVCAs. Additionally, the framework can serve as a benchmark for evaluating the level of proactivity in existing IVCAs, aiding in identifying shortcomings in current systems and guiding future improvement directions. Leveraging this framework, we conduct user studies to discover which level of proactivity in IVCAs is most appropriate.

4 Rewrite + ReAct + Reflect Prompting

In this section, we answer the question of "How to prompt LLMs to achieve accurate dialogue task completion and align with different levels of proactivity for IVCAs?". We give an introduction to the task and our prompting strategy. The overview of our "Rewrite + ReAct + Reflect" architecture is shown in Figure 2.

4.1 Task Formulation

We focus on achieving task-oriented conversations at the formulated proactivity levels in vehicles by prompting LLMs. Given the dialogue history $H_t = (q_1, r_1, ..., q_{t-1}, r_{t-1})$ and the user's current utterance q_t, we aim to get the correct response $P_{llm}(y|H_t, q_t)$ using question rewrite, ReAct prompt [Yao *et al.*, 2022] and the final reflect stage to identify and correct any biases or uncertainties in understanding the proactive interaction strategy in the response.

4.2 Question Rewriting

During conversations with IVCAs, users express themselves in diverse and casual ways, such as saying, "The smell in the car is a bit pungent." To facilitate user-centered interactions, IVCAs should be able to understanding and responding to various natural language expressions of users in driving contexts. To enhance the accuracy of completing dialogue tasks, aligning the user's natural inputs with the semantic space of the in-vehicle knowledge bases or contexts is necessary. LLMs have powerful language understanding ability and we use the ICL prompt technique to convert users' expressions into in-vehicle task-oriented questions. We utilize a few examples to help the LLMs better understand the rewriting question tasks. For instance, "The smell in the car is a bit pungent" is transformed into "Activate the car's fresh air circulation mode."

4.3 ReAct + Reflect

After rewriting the user's question, we improve proactive interactions for IVCAs using the ReAct + Reflect prompting strategy. ReAct [Yao *et al.*, 2022] prompts LLMs to trace reasoning and then execute task-specific actions, allowing the model to integrate external knowledge. The term "actions" refers to the functions that LLMs can employ. We include **search**[*question*] and **get_proactivity_strategy**[*number*] as the actions. The search action retrieves the most relevant knowledge with the rewritten question from the knowledge vector store to better support the conversation. The embedding model [4] is used to vectorize the rewritten question and the knowledge in databases. The get_proactivity_strategy action prompts the LLM to achieve the desired level of proactivity according to the strategies mentioned above. The search action takes the rewritten question as input, while get_proactivity_strategy takes the proactivity level number as input to get a specific proactive interaction strategy as illustrated in Section 3.

Additionally, some studies suggest that due to limitations in the model's memory capacity, LLMs could forget preceding information as the length of the prompt increases [Lu *et al.*, 2020]. In our work, the multiple reasoning and retrieved knowledge may lead to an excessively lengthy prompt, hindering LLMs from achieving the precise proactive level. As a solution, we implement a "reflect" stage before generating the final response. This stage encourages LLMs to assess whether their response aligns with our chosen proactive strategy and, if not, to regenerate the answer.

[4]https://github.com/FlagOpen/FlagEmbedding

5 Capability Experiments

We conduct experiments to verify the feasibility of using LLMs to improve IVCAs in proactive interactions.

5.1 Data Collection

We follow the data format of the In-Car dataset [Eric *et al.*, 2017] to construct multiple knowledge bases covering various scenarios. The In-Car dataset includes weather inquiries, calendar planning, and navigation data. We extend the dataset with in-car functions, environmental conditions, and user profiles. Fields for the knowledge base of each scenario are designed as comprehensively as possible to ensure they address the potential questions users may have within these scenarios. Based on the knowledge bases, we finally obtain a total of 1,302 queries.

5.2 Experimental Setups

We evaluate whether LLMs, prompted with our designed prompts, can reach each proactivity level with high quality using two metrics: *success rate* (the percentage of successfully achieved user requests or tasks within conversations) and *proactivity attainment rate* (the proportion of LLMs reaching the required level of proactivity). We employ the LLM gpt-3.5-turbo in our experiments. For *Success rate*, we compare the results of gpt-3.5-turbo with the state-of-the-art model TSCP [Lei *et al.*, 2018], LABES [Zhang *et al.*, 2020], Galaxy[He *et al.*, 2022] for task-oriented dialogue. TSCP is a sequence-to-sequence model with belief spans to track dialogue context. LABES is a dialog model that uses unlabeled data to improve belief state tracking. GALAXY leverages semi-supervised learning to improve dialogue performance. TSCP, LABES, and Galaxy were all fine-tuned on the training set of the In-Car dataset. As for the proactivity attainment rate, we follow the evaluation method [Sun *et al.*, 2023] to conduct scoring statistics:

$$Rate = \frac{\sum_{q \in Q} I(C = n)}{N_Q} \times 100\%, \tag{1}$$

where $Rate$ denotes the percentage of the scores labeled as n. n is from 1 to 5 in line with the proactivity levels. C means the conversation generated by the LLM. Besides, Q is the collected questions, and N_Q is the number of the queries.

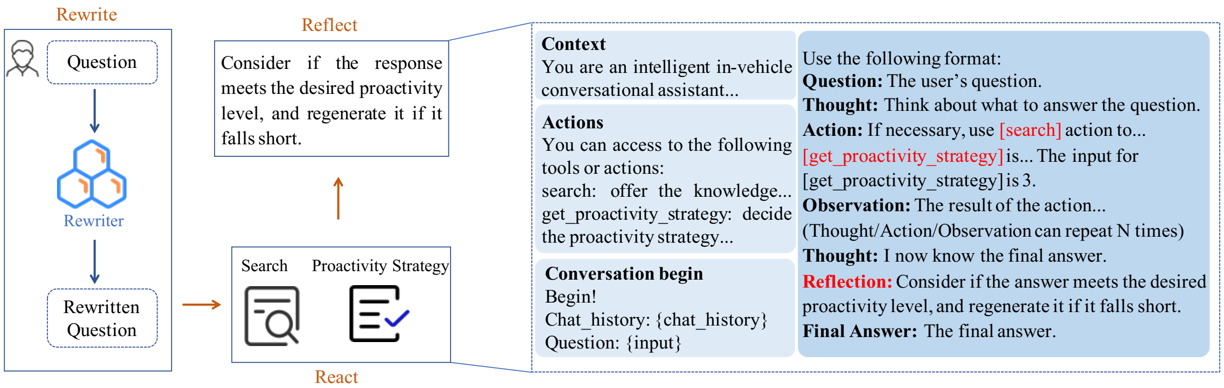

Figure 2: The overview of "Rewrite + ReAct + Reflect" prompting. The right is an example prompt based on the ReAct framework incorporating the rewritten question and reflection step.

Conversations	Level 1	Level 2	Level 3	Level 4	Level 5
number	210	301	179	177	408

Table 1: The conversation statistics for each proactivity level.

5.3 Result Annotation and Analysis

LLMs may generate different expressions with correct answers. Therefore, we adhere to the common practice of evaluating language generation quality using human ratings. We assign a 0 when the task is not completed. As for proactivity, we utilize scoring scales from 1 to 5, representing different proactivity levels according to the principles outlined in our framework. We involve three experts, none of whom are the authors of this paper, to annotate the results. Two specialize in computer science, and the third is from the HCI field. We ask them to rate every dialogue independently. When the annotators assign different scores for the same dialogue, the majority principle is used to resolve the inconsistencies. When all three are different, we discard the dialogue directly, resulting in 1,275 dialogues. The conversation counts of each proactivity level are shown in Table 1.

The experimental results are shown in Table 2, gpt-3.5-turbo achieves the success rate (93.72%), greatly outperforming the other models. TSCP, LABES, and Galaxy struggle to respond to users' naturally expressed demands, such as "I'm feeling hot" while this is easily manageable for LLMs (as shown in Figure 1). This also highlights the capability of LLMs to anticipate user intent within the assumption dimension. The score distribution in Table 3 illustrates that using our "Rewrite + ReAct + Reflect" prompts for various proactivity levels, over 78% of the conversations receive scores within the expected ranges. Besides, it can be observed that the results of "Rewrite + React + Reflect" are superior to those obtained with the ReAct strategy alone (in parentheses), confirming the effectiveness of the Reflect stage.

The reasons why some conversations are beyond the expected levels may be because of (1) Context influence. Some studies [Mishra *et al.*, 2022] indicate instructions within a context significantly impact LLMs, while there is no definitive conclusion about the best way to formulate the instruction. (2) Organization of demonstration. Prompt engineering depends on how demonstrations are organized [Mishra *et al.*, 2022]. It's important to carefully manage the demonstration format, the number, and the order of demonstration examples.

Model	Success rate
TSCP	64.32
LABES	61.60
GALAXY	69.00
gpt-3.5-turbo	**93.72**

Table 2: Performance of different models on success rate.

6 Subjective Experiments

Drawing from the capability experiments, it is clear that LLMs exhibit competence in dialogue comprehension and proactive interaction. In this section, we focus on validating our proactivity framework for IVCAs and assessing the effects of the five proactivity levels on user perceptions.

6.1 Simulator Design

We leverage gpt-3.5-turbo as the conversation engine and Alibaba ChatUI, a popular Web UI design language and React library, to develop an IVCA simulator. Furthermore, we add the functions of automatic speech recognition (ASR) and text-to-speech (TTS). So participants can interact with the simulator using natural language like the real interaction between drivers and the IVCA simulator. Our simulator includes five levels of proactive interaction, and users are required to select a specific level before engaging in dialogue. We select 10 questions and their corresponding knowledge bases for each proactivity level, totaling 50 questions.

6.2 Setup and Procedure

The IVCA simulator is integrated into the vehicle's Human-Machine Interfaces, appearing on an iPad before participants enter the vehicle (as shown in Figure 3). The experiment takes place in a stationary vehicle with a 240° curved screen displaying a dynamic environment. Before starting, participants receive a safety briefing, sign consent forms, and com-

Strategies	Level 1	Level 2	Level 3	Level 4	Level 5
1	**87.88 (84.90)**	0 (3.01)	12.12 (12.09)	0	0
2	6.32 (5.81)	**78.25 (77.92)**	15.44 (16.27)	0	0
3	1.18 (1.87)	16.57 (18.01)	**82.25 (80.12)**	0	0
4	0	0	0 (0.08)	**90.12 (87.69)**	9.88 (12.23)
5	0.54 (0.14)	0.54 (0.62)	0 (0.32)	17.79 (18.09)	**81.13 (80.83)**

Table 3: The proactivity attainment rates at each level. The numbers on the left indicate the specific proactivity strategy used in the prompt, while the percentage on the right represents the proportion reaching each proactivity level. Values in parentheses show outcomes from the ReAct strategy without Reflect stage.

plete a pre-trial questionnaire covering demographics, personality traits, and potential confounding variables. During the experiment, participants engage in five levels of proactive interactions. After each session, they complete a post-condition questionnaire and take part in a brief interview with a researcher. Each test session lasts approximately one hour.

6.3 Hypotheses

Previous works suggest that highly proactive behaviors of assistants will negatively influence users' perceptions, diminishing appropriateness and helpfulness [Peng *et al.*, 2019; Sun *et al.*, 2017]. Conversely, moderate proactive behaviors are associated with promoting a positive human-computer interaction relationship [Kraus *et al.*, 2021; Kraus *et al.*, 2020]. We hypothesize that:

H1. All five levels of user-perceived proactivity are effective, which implies that as the level of proactivity increases, IVCAs will be perceived as significantly more autonomous.

H2. Compared with proactivity levels of L5 and L1-L2, IVCAs at level L4 will be perceived as significantly more helpful, natural, acceptable, and appropriate, and exhibit the highest level of usability.

We measure the IVCA's autonomy, helpfulness, and appropriateness (adapted from [Lee *et al.*, 2010; Pu *et al.*, 2011; Sun *et al.*, 2017; Torrey *et al.*, 2013; Peng *et al.*, 2019]). Naturalness is investigated through "the naturalness of the interactive experience" of the IVCA [CAO *et al.*, 2023]. We utilize a reliable questionnaire for assessing the acceptance [Van Der Laan *et al.*, 1997]. Furthermore, usability is measured using a voice usability scale [Zwakman *et al.*, 2020]. A 7-point Likert scale measures all items in these questionnaires.

6.4 Participants

In this within-subjects design with repeated measures, 40 participants are recruited, with each evaluating five proactivity levels in a randomized order. 40 participants (P1-P40, 21 females and 19 males) from the local university and some technology companies participate in our experiments. Participants major in a diverse range of fields, and their ages range from 18 to 35 ($M = 28.75, SD = 2.47$). Thirty-two of them report that they have experience interacting with physical or virtual conversational assistants. All participants are not native English speakers but they all have fluent spoken and written English skills with a TOEFL score higher than 88 or an IELTS score above 6.5 assessed in the past two years.

6.5 Results

We use repeated measures ANOVA (Analysis of Variance) to compare the differences among groups with different proactivity levels. The data are checked for sphericity using Mauchly's test, and where violated, Greenhouse-Geisser and Hyunh-Feldt corrections are applied [Field, 2013]. We summarize the statistical analysis and user evaluation results in perceived autonomy, helpfulness, naturalness, acceptance, appropriateness, and usability during the interaction. Quantitative results are visualized in Figure 4.

Autonomy. The results show that the perceived autonomy of five groups is effective ($F(2.02, 78.60) = 29.88, p < .001, \eta^2 = 0.43$). The L5 group, operating without user control, is perceived as the most autonomous ($M = 6.47, SD = 0.85$), followed by the L4 ($M = 6.32, SD = 0.89$), L3 ($M = 5.83, SD = 1.13$), L2 ($M = 4.95, SD = 1.77; p < .001$), and finally L1 ($M = 3.68, SD = 2.35; p < .001$). Nevertheless, the Bonferroni post-hoc test reveals that the differences between L5 and L4, as well as L5 and L3, are not statistically significant. H1 verified.

Helpfulness. The results demonstrate a significant relationship between proactivity level and perceived helpfulness ($F(1.86, 72.42) = 19.37, p < .001, \eta^2 = 0.33$). Noteworthy findings from the Bonferroni post-hoc test indicate that participants in the L4 group, which retains a certain degree of user control, show significantly higher helpfulness ($M = 5.98, SD = 1.10$) compared to both the L5 ($M = 5.47, SD = 1.02; p < .05$), L3 ($M = 5.45, SD = 1.04, p < .05$), L2 ($M = 4.93, SD = 1.27, p < .05$), and L1 ($M = 3.76, SD = 2.02, p < .001$) groups.

Naturalness. Participants also perceive that they depend significantly more on the naturalness of the L4 group ($M = 6.19, SD = 0.72$) compared to the L5 ($M = 4.28, SD = 1.69; p < .001$), L3 ($M = 5.74, SD = 0.90, p < .001$), L2 ($M = 5.07, SD = 1.20, p < .05$), and L1 ($M = 4.17, SD = 1.73, p < .001$) groups in the Bonferroni post-hoc test ($F(1.86, 72.29) = 30.60, p < .001, \eta^2 = 0.44$).

Acceptance. Similarly, the L4 ($M = 6.42, SD = 0.96$) demonstrates significantly higher acceptance compared to the L5 ($M = 6.20, SD = 0.96; p < .001$), L3 ($M = 5.80, SD = 1.04, p < .05$), L2 ($M = 5.35, SD = 1.20, p < .001$), and L1 ($M = 4.67, SD = 1.64, p < .001$) groups, as revealed by the Bonferroni post-hoc test; ($F(2.17, 84.54) = 19.68, p < .001, \eta^2 = 0.34$).

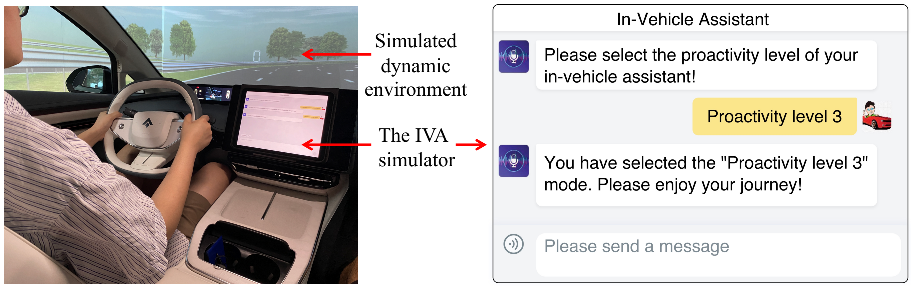

Figure 3: The experimental environment. The IVCA simulator is set in front of the driver.

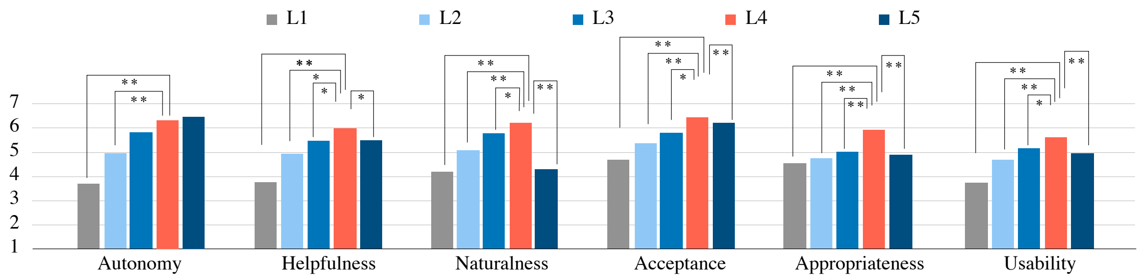

Figure 4: Mean scores of all subjective ratings between the five levels on a 7-point Likert scale ($*: p < .05, **: p < .001$).

Appropriateness. The results demonstrate a significant relationship between proactivity level and perceived appropriateness $(F(2.63, 102.61) = 22.108, p < .001, \eta^2 = 0.36)$. The Bonferroni post-hoc test further verifies that all pairwise comparisons are significantly different ($p < .001$). Specifically, participants in the L4 group suggest that it is notably more appropriate ($M = 5.92, SD = 0.81$) than the L5 ($M = 4.88, SD = 0.87; p < .05$), L3 ($M = 5.00, SD = 0.84$), L2 ($M = 4.75, SD = 0.81$), and L1 ($M = 4.55, SD = 0.83$).

Usability. The effect on the usability rating reaches statistical significance $(F(2.26, 88.04) = 28.96, p < .001, \eta^2 = 0.43)$. To be specific, the L4 group demonstrates the highest usability values ($M = 5.62, SD = 0.99$) compared with the L5 ($M = 4.94, SD = 0.87; p < .001$), L3 ($M = 5.14, SD = 0.87; p < .05$), L2 ($M = 4.67, SD = 0.93; p < .001$), and L1 ($M = 3.74, SD = 1.35; p < .001$) groups. Therefore, H2 verified.

Based on the user evaluations, hypotheses H1 and H2 are accepted. We find that different levels of proactivity do have a significant impact on autonomy, helpfulness, naturalness, acceptance, appropriateness, and usability (all $p < .001$). However, the highest level of autonomy (L5) shows varying degrees of decrease in acceptance, naturalness, appropriateness, helpfulness, and appropriateness compared to level 4, with the most pronounced decrease observed in naturalness. This indicates that a high degree of autonomy to some extent exceeds user cognitive demands.

7 Discussions and Future Work

We show the potential of LLMs in enhancing proactive interaction for IVCAs. By offering the "Rewrite + ReAct + Reflect" prompts for different proactivity levels, our approach shows advantageous results in the capability experiments. LLMs sometimes generate "hallucinations", providing reasonable but inaccurate information, so we should improve response reliability in future work. Additionally, LLMs lack transparency in decision-making. To address this, we would explore the model's ability to explain its decisions and enable users to understand how they generate the responses. Our proactivity framework significantly impacts user perceptions across autonomy, helpfulness, naturalness, acceptance appropriateness, and usability. Users express that the IVCA at the fourth level, which demonstrates strong anticipatory capabilities while maintaining user control, is most helpful, appropriate, and natural. Also, proactive interaction should consider task difficulty and timing to provide comprehensive strategies. The limited number of test questions and short testing duration for each level may also introduce bias. We aim to optimize these issues in our future work.

8 Conclusion

We explore how LLMs enhance user-centered interactions for IVCAs by introducing a framework with five proactivity levels. In addition, we recognize the potential of LLMs for IVCAs and devise a "Rewrite + ReAct + Reflect" approach to customize prompts for different proactivity levels. Our experiments show the feasibility of LLMs. User studies reveal that different proactivity levels significantly impact user perception of autonomy, helpfulness, naturalness, acceptance, appropriateness, and usability, which validates the effectiveness of our proactivity framework. Our study offers valuable insights and interaction strategies for IVCAs using LLMs, benefiting future research and practical uses in this field.

Acknowledgements

This work was supported by the National Nature Science Foundation of China (Grant No. 62176185), the Central Universities' Basic Research Fund (Grant No. 22120240256), and the National Key Laboratory of Information Systems Engineering (Grant No. PU52221147).

Contribution Statement

Huifang Du and Xuejing Feng contributed equally.

References

[Braun *et al.*, 2019] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[CAO *et al.*, 2023] Jianqin CAO, Jingyu ZHANG, Liang ZHANG, and Xiaoyu WANG. The psychological structure and influence of interactive naturalness. *Acta Psychologica Sinica*, 55(1):55, 2023.

[Cha *et al.*, 2020] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. Hello there! is now a good time to talk? opportune moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–28, 2020.

[Dong *et al.*, 2022] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[Eric *et al.*, 2017] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[Field, 2013] Andy Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[Grant and Ashford, 2008] Adam M Grant and Susan J Ashford. The dynamics of proactivity at work. *Research in organizational behavior*, 28:3–34, 2008.

[He *et al.*, 2022] Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10749–10757, 2022.

[Isbell and Pierce, 2005] Charles L Isbell and Jeffrey S Pierce. An ip continuum for adaptive interface design. In *Proc. of HCI International*, volume 10, 2005.

[Kim *et al.*, 2020] Auk Kim, Jung-Mi Park, and Uichin Lee. Interruptibility for in-vehicle multitasking: influence of voice task demands and adaptive behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

[Kraus *et al.*, 2020] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 107–116, 2020.

[Kraus *et al.*, 2021] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. The role of trust in proactive conversational assistants. *IEEE Access*, 9:112821–112836, 2021.

[Lee and Jeon, 2022] Seul Chan Lee and Myounghoon Jeon. A systematic review of functions and design features of in-vehicle agents. *International Journal of Human-Computer Studies*, 165:102864, 2022.

[Lee *et al.*, 2010] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210. IEEE, 2010.

[Lei *et al.*, 2018] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, 2018.

[Lin *et al.*, 2018] Shih-Chieh Lin, Chang-Hong Hsu, Walter Talamonti, Yunqi Zhang, Steve Oney, Jason Mars, and Lingjia Tang. Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 531–542, 2018.

[Lu *et al.*, 2020] Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[Meck and Precht, 2021] Anna-Maria Meck and Lisa Precht. How to design the perfect prompt: A linguistic approach to prompt design in automotive voice assistants–an exploratory study. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 237–246, 2021.

[Meck *et al.*, 2023] Anna-Maria Meck, Christoph Draxler, and Thurid Vogt. How may i interrupt? linguistic-driven design guidelines for proactive in-car voice assistants. *International Journal of Human–Computer Interaction*, pages 1–15, 2023.

[Mi *et al.*, 2022] Fei Mi, Yasheng Wang, and Yitong Li. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11076–11084, 2022.

[Mishra *et al.*, 2022] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics, 2022.

[Nothdurft *et al.*, 2014] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. Finding appropriate interaction strategies for proactive dialogue systems—an open quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication*, volume 110, pages 73–80, 2014.

[Parker and Collins, 2010] Sharon K Parker and Catherine G Collins. Taking stock: Integrating and differentiating multiple proactive behaviors. *Journal of management*, 36(3):633–662, 2010.

[Peng *et al.*, 2019] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. Design and evaluation of service robot's proactivity in decision-making support process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[Pu *et al.*, 2011] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164, 2011.

[Rau *et al.*, 2013] Pei-Luen Patrick Rau, Ye Li, and Jun Liu. Effects of a social robot's autonomy and group orientation on human decision-making. *Advances in Human-Computer Interaction*, 2013:11–11, 2013.

[Rony *et al.*, 2023] Md Rashad Al Hasan Rony, Christian Süß, Sinchana Ramakanth Bhat, Viju Sudhi, Julia Schneider, Maximilian Vogel, Roman Teucher, Ken E Friedl, and Soumya Sahoo. Carexpert: Leveraging large language models for in-car conversational question answering. *arXiv preprint arXiv:2310.09536*, 2023.

[Schaeffer *et al.*, 2023] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

[Shen *et al.*, 2022] Junxiao Shen, Boyin Yang, John J Dudley, and Per Ola Kristensson. Kwickchat: A multi-turn dialogue system for aac using context-aware sentence generation by bag-of-keywords. In *27th International Conference on Intelligent User Interfaces*, pages 853–867, 2022.

[Sun *et al.*, 2017] Mingfei Sun, Zhenjie Zhao, and Xiaojuan Ma. Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 556–567, 2017.

[Sun *et al.*, 2023] Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*, 2023.

[Torrey *et al.*, 2013] Cristen Torrey, Susan R Fussell, and Sara Kiesler. How a robot should give advice. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 275–282. IEEE, 2013.

[Van Der Laan *et al.*, 1997] Jinke D Van Der Laan, Adriaan Heino, and Dick De Waard. A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, 5(1):1–10, 1997.

[Völkel *et al.*, 2021] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R Cowan, and Heinrich Hussmann. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[Wei *et al.*, 2022] 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.

[Yao *et al.*, 2022] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[Yao *et al.*, 2023] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[Young *et al.*, 2022] Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629, 2022.

[Zargham *et al.*, 2022] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Völkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–14, 2022.

[Zhang *et al.*, 2020] Yichi Zhang, Zhijian Ou, Huixin Wang, and Junlan Feng. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. *arXiv preprint arXiv:2009.08115*, 2020.

[Zwakman *et al.*, 2020] Dilawar Shah Zwakman, Debajyoti Pal, Tuul Triyason, and Chonlameth Arpnikanondt. Voice usability scale: measuring the user experience with voice assistants. In *2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS)*, pages 308–311. IEEE, 2020.