中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-23

论文引用格式:Liu Jialing, Zhang Xinjie, Jin Qin. Proactive Dialogue Systems in the Era of Large Language Models: Methods, Evaluation, and Challenges[J/OL]. Journal of Image and Graphics, XXXX:1-23. DOI: 10.11834/jig.250601. (刘嘉玲, 张鑫洁, 金琴. 大语言模型时代的主动对话系统:方法、评估与挑战综述[J/OL]. 中国图象图形学报,XXXX:1-23. DOI: 10.11834/jig.250601.)[DOI:10.11834/jig.250601]

大语言模型时代的主动对话系统:方法、评估与挑战综述

刘嘉玲,张鑫洁,金琴*

中国人民大学信息学院,北京　100872

摘　要:主动对话系统旨在赋予机器根据上下文自主发起交互、预测用户需求并提供适时服务的能力,是人机交互与AI智能体(Agent)领域的一个重要研究方向。早期研究主要关注如何引导用户实现系统预设目标,通过生成引导性提问、建议或信息来主动驱动对话进程,相关方法涵盖基于规则的策略设计、强化学习的动作优化及预训练模型的内容生成。然而,主动性不应仅局限于"说什么",时间感知与时机把握同样是关键维度:过早干预可能造成打扰,过晚响应则失去帮助价值,何时主动介入直接决定了用户体验的成败。近两年来,随着大语言模型的推理能力提升和多模态感知技术的成熟,关注"何时说"的对话研究快速涌现,但现有综述尚未对这一新兴方向进行系统分析,对主动引导方法的最新进展也缺乏及时跟进。因此,本文首先对现有对话系统研究进行了梳理,依据是否引入时序维度这一核心标准,将现有方法划分为意图主动(intentional proactive)与时序主动(temporal proactive)两大类。针对意图主动,本文从以目标为中心和以用户为中心两个视角解析大语言模型时代的建模范式和内容生成机制演进;针对时序主动,本文将其解构为时机预测、策略规划与响应生成三个关键环节,对各类研究工作的核心思想与技术特点进行了阐述与综合分析。在此基础上,本文进一步从LLM支撑角色、模型版本披露与评估可比性三个维度补充比较性讨论。此外,本文列举了当前用于主动对话的基准数据集,从数据规模、对话轮次及应用场景等方面进行了综合比较。最后,本文指出当前主动对话领域面临的数据稀缺、个性化与自适应能力不足、多模态融合局限及主动干预伦理边界模糊等核心挑战,并对未来的研究方向进行了展望。

关键词:主动对话;大语言模型;意图主动;时序主动;人机交互;多智能体

Proactive Dialogue Systems in the Era of Large Language Models: Methods, Evaluation, and Challenges

Liu Jialing, Zhang Xinjie, Jin Qin*

School of Information, Renmin University of China, Beijing 100872, China

Abstract: With the exponential growth of artificial intelligence technologies, particularly the advent of Large Language Models (LLMs) and advanced multimodal perception capabilities, the field of dialogue systems is undergoing a profound paradigm shift from passive, reactive interaction models to intelligent, proactive agents capable of autonomous engagement. Traditional reactive dialogue systems, which have long dominated the landscape of human-computer interaction,

收稿日期:2025-11-28;**修回日期**:2026-08-18

* **通信作者**:金琴 qjin@ruc.edu.cn

基金项目:国家自然科学基金项目(批准号:62576347)

Supported by: National Natural Science Foundation of China (Grant No. 62576347)

operate on a "request-response" basis where the system relies entirely on explicit user queries or commands to initiate a turn, a limitation that becomes increasingly apparent in complex, dynamic real-world scenarios where users may have latent needs, undefined goals, or require timely assistance without knowing how or when to ask. Addressing these limitations, Proactive Dialogue Systems represent a significant leap forward, endowing machines with the cognitive ability to anticipate user intentions, monitor environmental contexts, and initiate interactions to provide support, guidance, or emotional companionship. However, proactivity should not be limited solely to "what to say"; temporal awareness and timing are equally critical dimensions: intervening too early may cause disruption, while responding too late diminishes the value of assistance. The timing of proactive intervention directly determines the success or failure of user experience. In the past two years, with the enhanced reasoning capabilities of large language models and the maturation of multimodal perception technologies, research on dialogue systems focusing on "when to speak" has rapidly emerged. Yet, existing surveys have not systematically analyzed this emerging direction nor provided timely coverage of recent advances in proactive guidance methods. Therefore, this paper first reviews current dialogue system research and categorizes existing approaches into two main types—Intentional Proactive and Temporal Proactive—based on the core criterion of whether they incorporate a temporal dimension. In the realm of Intentional Proactive Dialogue, the system's proactivity is driven by the content and goals of the conversation rather than strict real-time environmental constraints. This category is further dissected into two subdimensions: Target-Oriented and User-Centric systems. Target-Oriented systems are engineered to guide the dialogue efficiently toward a predefined goal or conclusion, such as completing a specific task, making a recommendation, or converging on a topic. Research in this area focuses on overcoming the limitations of greedy, turn-by-turn generation by employing global planning mechanisms, such as Target-Constrained Bidirectional Planning (TRIP) and hierarchical goal decomposition, which allow the system to navigate complex dialogue trajectories while maintaining logical consistency and minimizing the "distance" to the desired outcome. Differently, User-Centric systems prioritize the user's experience, emotional state, and implicit needs, leveraging psychological frameworks like the Belief-Desire-Intention (BDI) model and empathetic computing strategies. These systems aim to build rapport and provide emotional support by proactively asking clarifying questions, offering comfort, or stimulating interest based on the user's historical preferences and personality traits, thereby transforming the agent from a mere tool into a companion that fosters long-term engagement. The survey places particular emphasis on the second major paradigm, Temporal Proactive Dialogue, which introduces the critical challenge of determining when to interact. Unlike intentional proactivity, which focuses on what to say, temporal proactivity requires the system to continuously sense the environment and user state to identify the "Goldilocks Time Window"—the optimal moment for intervention where the utility of assistance outweighs the cost of interruption. The survey deconstructs the architecture of temporal proactive systems into a closed-loop process comprising three tightly coupled components: Timing Prediction, Policy Planning, and Response Generation. Timing Prediction serves as the foundation, employing multimodal sensors (vision, audio) and social graph reasoning to detect cues such as hesitation, gaze direction, and environmental changes that signal a need for help. Theoretical underpinnings like Cognitive Load Theory are utilized to predict the user′s mental state, ensuring interventions occur at moments of low cognitive burden or high necessity. Policy Planning connects timing with action, transforming the "when" and "how" into a unified optimization problem where the system dynamically adjusts its strategy—choosing between explicit verbal interruptions or subtle visual cues—based on real-time feedback and reinforcement learning models trained to maximize long-term interaction value. Response Generation in this context moves beyond static text generation to dynamic, time-sensitive output. The survey highlights innovations in asynchronous architectures and "speak-before-asked" mechanisms, where the system generates responses that are contextually coupled with the evolving timeline of events, ensuring that advice is delivered with high temporal precision and relevance. Furthermore, the survey critically analyzes the infrastructure supporting this field, reviewing a wide array of benchmark datasets that have evolved from text-only corpora like MultiWOZ to complex, multimodal datasets such as ProactiveVideoQA and EgoTaskQA, which capture the nuances of physical interactions and rich environmental contexts necessary for training temporally aware agents. The paper also summarizes different evaluation criteria, while traditional metrics like BLEU, BERTScore and Perplexity assess linguistic quality, they fail to capture the essence of proactivity. Consequently, the survey investigates specific metrics, such as the "Proactive Area Under Curve" (PAUC) for assessing timing accuracy, along-

side multi-dimensional human evaluation protocols that measure Strategic Effectiveness, Reasoning Consistency, and Ethical Safety. The survey further examines how LLMs support proactive dialogue systems in different roles and discusses the comparability issues caused by inconsistent disclosure of model versions and deployment settings. Despite notable advances in proactive dialogue systems, their broad deployment in open-world settings remains hindered by several interrelated challenges. High-quality annotated data—particularly real-world datasets that integrate multimodal signals and capture fine-grained user feedback on proactive interventions—is still scarce. Current systems exhibit limited capacity to understand individual user differences and evolving cognitive states, resulting in weak adaptability and insufficient continual learning. Multimodal fusion remains nascent, with existing approaches struggling to effectively align heterogeneous cues from speech, text, and vision to support robust environmental awareness. Moreover, ethical and privacy concerns lack comprehensive frameworks, and there is a pressing need to ensure user controllability and transparency in how and when proactive behaviors are triggered. In conclusion, the paper not only charts current state of the art in proactive dialogue systems but also outlines a roadmap for future research, emphasizing the need for real-world datasets, dynamic and self-adaptive personalization, deeper multimodal spatiotemporal fusion, and rigorous ethical frameworks to realize the vision of AI agents that are true intelligent partners—capable of understanding not just what we say, but what we need, and intervening at precisely the right moment to enhance human capability and well-being.

Key words: proactive dialogue; large language models (LLMs); intentional proactive; temporal proactive; human-computer interaction; multi-agent

论文引用格式：Liu Jialing, Zhang Xinjie, Jin Qin. Proactive Dialogue Systems in the Era of Large Language Models: Methods, Evaluation, and Challenges[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.250601.（刘嘉玲，张鑫洁，金琴. 大语言模型时代的主动对话系统：方法、评估与挑战综述[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.250601.）

0 引　言

随着人工智能技术的迅猛发展，对话系统已从简单的问答交互向更加智能、自然、人性化的方向演进，并已被探索用于智能医疗（Xu等，2024a；文昕颢等，2026）、情感支持（Liu等，2021a）、社会规范感知（Zhan等，2023）等领域。然而，传统的对话系统大多属于被动式对话（reactive dialogue；Huang等，2020），即系统在对话过程中主要对用户输入进行理解与回应，依赖用户提供明确的查询或指令（如图1（a））。这种被动响应式的交互模式在面对复杂的用户需求和交互场景时，逐渐显露出局限性：系统缺乏对上下文的主动感知能力、无法提前预测用户潜在需求，也难以主动引导对话向目标导向或情感友好的方向发展，导致交互体验单一、效率低下，甚至可能引发用户的不满。

与传统被动式对话系统相比，主动对话（proactive dialogue）作为一种新兴的交互范式，正逐渐成为智能对话领域的研究热点。Liu等人（2025）指出，主动的人工智能应当像人类一样，不仅能对轮流的话语作出反应，还能在对话中形成自己的内心想法，并寻找合适的时机做出贡献。基于这一观点，我们将主动对话定义为系统不仅能响应用户输入，还能结合上下文、用户画像及外部知识等多维信息，预测用户意图并主动发起交流、引导话题或提供建议的交互模式。虽然这一模式极具潜力，但其在技术实现、用户体验评估、伦理规范等方面仍面临诸多挑战，如主动开启对话的时机、怎样确保主动干预的准确性和恰当性、如何建立有效的评估标准等。

尽管主动对话已引起广泛关注（Qi等，2026），但现有综述工作存在明显的局限性。已有综述（如Deng等，2025）全面总结了开放域、任务导向及信息寻求等场景下的关键技术，但其主要聚焦于“意图主动”（intentional proactive）型研究，即侧重于挖掘用户潜在需求以决定“主动说什么”，且多局限于文本模态。然而，真正符合人类习惯的交互不仅关乎内容，更依赖于对交互节奏的感知。为了实现更自然、更精准的干预，研究重心正从单纯的意图预测向融合多模态信息的“时序主动”（temporal proactive）转移。近期，Fang等人（2025b）提出的“黄金时间窗口”（goldilocks time window）概念强调了有效干预的

上下文自适应时间窗口的重要性。对于这一新兴的时序主动方向,已有综述尚未进行系统梳理,对意图主动领域的最新进展也缺乏及时跟进。

(a) 被动对话　　(b) 意图主动　　(c) 时序主动

图 1　(a)被动型对话系统;(b)意图主动型对话系统;(c)时序主动型对话系统

Fig. 1　(a)Reactive dialogue system;(b)intentional proactive system;(c)temporal proactive system

基于此,我们以是否引入"时序维度"为核心标准,将主动对话系统划分为两类:意图主动型(如图 1(b))与时序主动型(如图 1(c))。进一步地,本文采用"主导优化目标"而非"是否出现显式目标或意图建模"作为判别标准:若系统主要围绕语义目标、用户需求或会话结论来决定主动内容,则归为意图主动;若系统主要围绕干预时间窗口与时序收益来决定是否介入,则归为时序主动。两类方法在表示层面可以共享目标建模,但其方法论焦点与评估重点并不相同。如图 2 所示,我们将"意图主动"划分为"以目标为中心"(1.1 节)和"以用户为中心"(1.2 节)两个维度,将"时序主动"

图 2　根据是否引入时序维度,对主动对话中的研究问题进行分类及适用场景分析

Fig. 2　Taxonomy of proactive dialogue systems, in terms of the temporal dimension and their applicable scenarios analysis

细分为"时机预测"(2.1 节)、"策略规划"(2.2 节)和"响应生成"(2.3 节)三个层次。其中,"以目标为中心"聚焦于系统如何通过主动干预引导对话向预设目标推进,其核心在于对话的"目标驱动性";

而“以用户为中心”则关注系统如何感知并响应用户的即时状态、情绪与潜在需求，强调交互的“用户体验导向”。“时序主动”的三层次划分源于对话流程的时序维度与决策层次。“时机预测”解决“何时主动”的问题，即系统基于对话状态与上下文预测最佳干预时机；“策略规划”解决“如何主动”的问题，即设计具体的主动行为策略；“响应生成”则解决“具体说什么或做什么”的问题，即生成符合上下文的主动对话内容。该分类形成了完整的“预测-决策-执行”闭环，相较于传统分类中对时序因素的笼统描述，这种细分更精准地刻画了时序主动对话的动态机制。

为了系统地梳理主动对话领域的最新进展并填补现有综述的空白，本文以上述分类体系为纲领，对主动对话研究进行全面深入的总结。具体地，本综述在涵盖传统意图主动型研究的基础上，特别强调了“时序主动对话”这一新兴研究方向的重要性，系统梳理了近年来核心技术的最新进展；同时从技术架构、交互设计、评估机制等多个维度分析了主动对话系统面临的核心挑战与瓶颈。与此同时，本文新增从LLM支撑角色与模型披露粒度出发的横向比较，以增强不同方法之间的可比性分析。通过这样的全景式分析，本文期望可以为主动对话领域的后续研究提供有价值的参考与启示。

1 意图主动

本文将意图主动定义为一种以语义目标、用户意图为主导信号进行主动交互的对话模式，其核心特征是主动决策主要由“系统希望推进什么”与“用户想要什么”来驱动。基于此，我们进一步将意图主动划分为两个关键维度：以目标为中心（target-oriented）与以用户为中心（user-centric）。以目标为中心的对话系统旨在引导对话合理且高效地推进至预设目标或结论；以用户为中心的对话系统则通过类人交互机制，主动促进用户参与度、提供个性化建议及情感支持，从而构建更具陪伴性与情感共鸣等的对话体验。例如，CACAS系统（Su等，2025）虽然会动态调整策略时机，但其首要学习目标仍是“推进对话策略”，时序维度更多服务于语义目标，因此在本文的辨析中归入意图主动。

1.1 以目标为中心

以目标为中心的主动对话系统需要主动引导对话朝着指定的主题方向发展（Tang等，2019；Feng等，2025）。根据不同的应用，目标可以是主题关键词（Tang等，2019）、对话目标（Wang等，2024）、会话结论（Guo等，2024）等。

1.1.1 问题定义

以目标为中心的对话系统是通过引人入胜的对话，有效且高效地达到指定的目标（Deng等，2025）。具体定义为给定一个目标话题T，且该目标只有对话系统已知，用户未知。对话从任意初始话题x_0开始，对话系统将产生第t轮响应r_t，该响应呈现后的话题x_t需实现：

$$min_t x_t = T\#(1)$$

同时要求响应不会产生话题的突变，即应当密切衔接上下文产生适当的内容。

1.1.2 实现方法

在目标导向的主动对话系统中，真正困难的不是生成一句局部合理的回复，而是在用户并不知道系统目标的情况下，持续把对话推进到预设结论，同时避免话题跳变和错误累积。研究瓶颈主要有三类：逐轮顺序规划容易陷入局部最优，人工定义策略难以覆盖长程任务，而一旦生成内容与历史不一致，后续目标推进就会失去可信度。

针对第一类瓶颈，TRIP的核心洞察是：既然目标话题事先已知，规划就不必只从当前轮往前贪心展开，而可以让起点约束和终点约束同时向中间收缩。Wang等人（2024）提出目标约束双向规划（target-constrained bidirectional planning，TRIP），如图3（a），通过双Transformer解码器并行生成〈动作，话题〉序列，并用决策差距最小化、目标对比生成与双向协议约束解码来保证两端会合。这样做能够缓解顺序生成中的误差累积，但代价是方法默认目标可预先给定，而且双向协同解码也显著提高了训练与推理复杂度。

同类工作还发现，即使目标事先给定，系统也可能因为只盯当前轮而忽视场景状态与未来意图关键词的变化。为此，Li等人（2026）提出将用户画像与领域知识联合建模为会话场景（conversational scenario modeling，CSM），然后通过意图-关键词桥接（intent-keyword bridging，IKB）预测未来轮次的意图关键词，并将场景偏置注入生成流程，以动态引导系统朝预设目标前进。该方法在缩小预设目标与真实对话演进之间的落差方面表现出显著效果，但其性

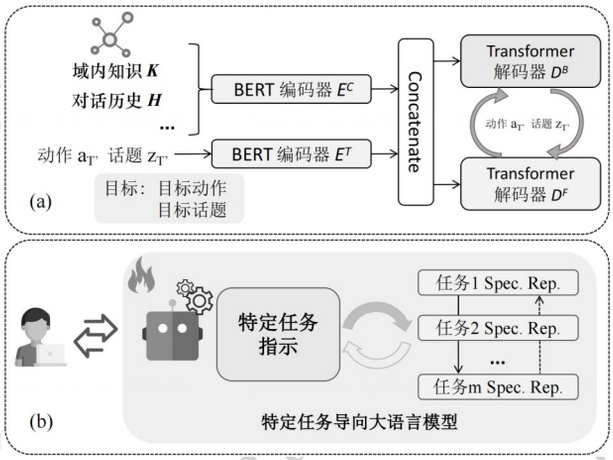

图3 (a)目标约束双向规划,(b)Spec-ToD

Fig. 3 (a)TRIP;(b)Spec-ToD

能可能更依赖用户画像和领域知识的质量,且跨域适配时需要对配置进行调整。

对于长程或复合目标无法靠局部搜索稳定完成,研究者便转向“先生成整条轨迹,再决定每一步”。Du等人(2025)的DiffTOD将对话规划形式化为条件引导的轨迹生成问题,利用扩散语言模型整体评估完整路径,试图从根源上减少逐轮决策偏航;HierTOD系统(Mo等,2025)则把复杂任务拆为高层目标分解与底层动作执行,用分层结构缓解单一策略同时处理“远目标”和“近动作”的负担。它们的共同代价是更依赖目标类型定义和子任务边界,一旦目标开放或分解失真,全局规划收益会明显下降。

另一条路线针对的是“人工定义策略过重”的问题。He等人(2024)提出的LDPP(learnable dialogue policy planner)认为,有效主动策略本来就隐含在真实对话日志中,与其手工枚举,不如直接抽取并学习这些策略模式;Wei等人(2025)的Learn-to-Ask则利用专家轨迹中的未来信息离线学习主动提问策略,试图绕开昂贵且偏差明显的用户模拟器。它们的共同优势是提升泛化与适应性,共同代价则是高度依赖日志质量和专家覆盖面,数据一旦偏窄,学到的主动性就容易迁移失效。

当规划能力增强后,新的瓶颈转为“怎样不说错、不断线”。Zhang等人(2025a)提出CRC框架,先识别响应与上下文的语义冲突,再基于反思结果重构输出;Guo等人(2024)的PCQPR则把主动提问与自我修正绑定起来,让问题既推进目标又不脱离上下文。这类方法的核心洞察是,目标导向主动性不能只靠更强规划,还要靠在线反思维持一致性;但相应代价是推理链路更长、时延更高,而且纠错质量仍受限于模型自身反思能力。

最后,上述思路在具体场景中继续分化。LUSTER(Feng等,2025)、DialogXpert(Rakib等,2025)和Wayfinding AI(Sayres等,2025)分别把目标推进与情感支持、健康咨询等需求绑定起来,Spec-TOD(Nguyen等,2025)则针对低资源场景用轻量级专用LLMs降低训练与部署成本。这说明目标导向主动系统最终仍需在任务成功率、情感自然度与部署成本之间取舍,并不存在对所有开放场景都通用的统一方案。

1.1.3 数据集

在以目标为中心的主动对话研究中,高质量、结构合理的数据集是推动模型设计的关键基础。如表1所示,近年来,多个面向目标导向对话的数据集相继开源,从信息寻求到多轮对话,为探索系统如何主动引导对话进程提供了丰富的实验资源。

针对开放域信息寻求中的主动性问题,多个新兴数据集引入了更复杂的会话动态。TopiOCQA(Adlakha等,2022)包含3,920个对话,平均每轮跨越13个问答回合并涉及4个主题转换,要求模型对同一会话的多个回合进行有效的检索,并使用会话历史构建有效的响应。在解决上下文依赖导致的意图模糊问题上,QReCC(Anantha等,2021)构建于大规模网络语料之上,配对了用户问题的重述形式与黄金答案,支持对会话中隐含意图的主动推断。而INSCIT(Wu等,2023)则聚焦于混合主动性的信息寻求对话,包含4.7k次用户-智能体交互,强调系统需根据用户反馈决定是否提问澄清、提供信息或调整检索策略。MultiDoc2Dial(Feng等,2021)进一步将目标导向对话置于多文档环境中,模拟用户在多文档中查找信息的场景,要求系统主动整合跨文档知识以达成用户目标。

此外,若干通用型多轮对话数据集也为目标导向主动对话研究提供了支撑。MultiWOZ系列(2.0 - 2.2)(Budzianowski等,2018;Eric等,2020;Zang等,2020)以其大规模、多领域、多回合的特性,成为任务型对话系统的标准测试平台。Wang等人(2023)基于角色扮演方法的自动数据集管理框架构建了一个大规模的面向目标的个性化对话数据集——TOPDIAL,该数据集包含约18K个多轮对话。

Wang等人(2024)利用DuRecDial(Liu等,2020)及其扩展版本DuRecDial 2.0(Liu等,2021b)构建生成面向目标的对话。其中DuRecDial数据集包含约10k个多回合中文对话,而DuRecDial 2.0数据集包含16.5k个跨英语和中文的对话。在扩展后的数据集中,系统经常主动引导对话而不是被动地响应用户,具有丰富的交互行为,如聊天、问答、推荐等。

表1　以目标为中心的数据集

Table 1　Target-oriented datasets

名称	发表信息	内容	规模	对话级别
TopiOCQA	(Adlakha等,2022)	信息寻求	3,920	多轮
QReCC	(Anantha等,2021)	开放域对话	14k	多轮
INSCIT	(Wu等,2023)	混合主动交互信息寻求	4.7k	多轮
MultiDoc2Dial	(Feng等,2021)	目标导向对话	67k	多轮
MultiWOZ 2.0	(Budzianowski等,2018)	多领域任务导向	10k	多轮
MultiWOZ 2.1	(Eric等,2020)			
MultiWOZ 2.2	(Zang等,2020)			
TOPDIAL	(Wang等,2023)	面向个性化目标导向对话	18k	多轮
DuRecDial	(Liu等,2020)	多类型对话推荐	10k	多轮
DuRecDial 2.0	(Liu等,2021b)	中英文混合对话推荐	16.5k	多轮

注:表中所列数据集均不含显式时序标注,也不包含视觉、音频等多模态信号,整体上主要服务于意图主动中的以目标为中心研究;其中INSCIT、TOPDIAL与DuRecDial系列更强调混合主动、个性化引导或推荐式目标推进,但其主导优化对象仍是语义目标而非时间窗口。

1.1.4　评估协议

在主动对话系统的研究中,评估协议的设计对于衡量系统性能、理解其主动行为的有效性至关重要。当前的评估方法大致可分为自动评估、人类评估以及对话级评估三类,分别从不同维度刻画系统的语言生成质量、策略合理性及整体交互效能。

自动评估指标主要聚焦于生成话语的语言质量、知识一致性以及目标达成能力。早期工作常采用通用文本生成指标,如困惑度(Perplexity,PPL)和Distinct-n(DIST)(Li等,2016),前者反映生成语句的流畅性,后者衡量词汇多样性。此外,BLEU(Papineni,2002)通过计算生成话语与参考话语之间的n-gram重叠程度来评估表面相似性;而F1分数则在词级(或中文场景下的字符级)上综合衡量精度与召回率。进一步地,BERTScore(Zhang等,2020)基于预训练语言模型生成的上下文化词向量表示,通过计算候选回复与参考回复之间的语义匹配程度进行评估,能够在一定程度上弥补BLEU等表层重叠指标对同义改写和语义等价表达不敏感的不足。同时,P@K和R@K评估关键词候选目标集合中位置K的查准率和召回率。针对主动对话中特有的目标导向特性,研究者引入了更多任务相关的指标。例如,Wang等人(2024)使用goal success rate(Goal Succ.)来衡量系统是否成功引导对话达成预设目标(如完成特定任务或覆盖指定话题)。在知识密集型场景中,Knowledge F1(Know. F1)(Liu等,2020)进一步评估系统能否准确引用领域知识三元组中的关键信息(如主题或属性)。

鉴于自动指标难以全面捕捉对话的语用合理性与社会智能,人类评估仍是不可或缺的补充。受See等人(2019)启发,Hu等人(2025)提出了四个细粒度的人类评价维度:战略有效性(strategic effectiveness,SE)与推理一致性(reasoning consistency,RC),评分范围为{0, 1, 2},分别评估主动策略是否有效推动对话进程及其逻辑连贯性;专业性(professionalism, Prof)(评分{0, 1, 2, 3}),衡量回应是否符合领域规范(如心理咨询中的专业表达);道德安全性(ethical safety, ES)(二元评分{0, 1}),确保系统行为符合伦理准则,避免有害或不当建议。

在系统整体性能层面,由于真实用户实验成本

高昂且难以规模化，多数研究依赖用户模拟器进行端到端评估。常用指标包括：SR@t（success rate at turn t）：衡量系统在第t轮对话内达成目标的比例；以及#Turns：完成目标任务所需的平均对话轮次（Deng等，2025）。

综合上述评估协议可以看出，以目标为中心的方法在Goal Succ.、SR@t与#Turns等任务型指标上最容易体现优势，其有效性的根源在于通过全局规划、分层分解或离线策略学习显式压缩了对话搜索空间，从而减少逐轮贪心生成带来的偏航风险。从现有结果趋势看，TRIP、DiffTOD、HierTOD一类“先规划后生成”的方法更适合目标明确、任务链较长的场景；CRC、PCQPR等一致性与自我修正方法则更直接改善RC、Know. F1和专业性等质量维度，但其收益依赖于目标定义和上下文约束是否足够稳定；Spec-TOD等轻量化方案在低资源和低成本部署上更有吸引力。换言之，该类方法的本质优势在于“把主动性转化为可规划的任务推进”，其适用边界则是目标必须相对可表述、可验证，否则系统容易把主动引导退化为过度控制。

1.2 以用户为中心

如果说“以目标为中心”的主动对话强调任务导向的效率，那么“以用户为中心”的主动对话则聚焦于交互过程中对用户状态的深度理解与动态适应。Deng等人（2024）强调了建立以人为中心的主动对话智能体（proactive conversational agent，PCA）的重要性，在这一范式下，系统的主动性并非源于预设任务的推进需求，而是由对用户情绪、认知负荷、兴趣偏好或潜在意图（Lu等，2025）的实时感知所驱动。其核心目标是提升对话的自然性、共情能力与个性化水平，使系统不仅能“完成任务”，更能“理解人”。

1.2.1 问题定义

以用户为中心的主动对话可形式化定义为：给定当前对话历史$H=\{u_1, s_1, \ldots, u_t\}$（其中$u_i$表示用户话语，$s_i$表示系统回应）以及用户状态表征$U_t$（包括但不限于情绪、意图、注意力等隐变量），系统需在无需显式用户请求的情况下，主动判断是否应发起干预，并生成符合用户当前状态与长期需求的响应s_{t+1}，以优化用户体验指标（如满意度、流畅度等）。

1.2.2 实现方法

随着大型语言模型（LLMs）技术的快速发展，以用户为中心的主动对话系统已成为提升人机交互体验的关键方向。这类系统通过深入理解用户状态、情感和偏好，实现对话的主动引导与个性化服务，提高了用户满意度和交互质量。当前研究主要沿：用户状态感知、主动决策规划、个性化生成以及场景化交互机制四个层次展开。

1）用户状态感知与建模。问题在于用户往往不会直接表达情绪、意图和阻塞点，仅靠表层文本很难判断“该不该主动”。Satori（Li等，2025a）的解决方法是把主动性建立在心理状态而非字面需求上，因此，他们将信念-愿望-意图（belief-desire-and- intention，BDI）框架与多模态大模型结合，通过用户心理状态和环境背景共同决定主动引导方式。COCOON（Fu等，2025）进一步看到，情感支持场景的瓶颈不是信息缺失，而是直接追问会引发防御，于是采用积极倾听和温和探询去揭示隐性需求；PsyAdvisor（Hu等，2025）则用策略提示驱动揭示性提问，触及用户未言明的情绪与需求。它们的共同代价是对心理假设、场景标注和安全边界依赖较强，一旦状态判断失准，主动行为就可能从“体贴”变成“冒犯”。

2）主动决策与策略规划。即便已经识别出用户状态，系统仍要回答“先澄清还是先执行、主动到什么程度、什么时候收手”这些更难的问题。PRINCIPLES（Kim等，2025）的洞察是把策略覆盖不足看作记忆问题而非纯生成问题，因此构造合成策略记忆库，让系统能够根据用户反馈调用更平衡的对话策略。为应对复杂任务，如图4，CEP框架（Zhang等，2024）把模糊指令处理拆成澄清、执行和规划三个智能体，用分工来缓解单个LLM同时承担全部决策的负担；AppAgent-Pro（Zhao等，2025）则通过预测潜在需求提前展开多域信息挖掘。代价是系统链路更复杂，模块间误差会传递，并且对多智能体协同和外部工具稳定性要求更高。

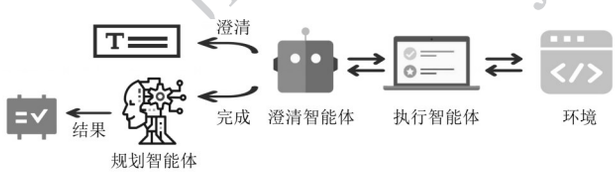

图4 CEP多智能体框架

Fig. 4 Clarification-Execution-Planning multi-agent framework

3）个性化响应生成与引导。真正的瓶颈不是能否生成个性化文本，而是系统是否掌握了足够可信的长期用户证据。面向用户主动聊天机器人UPC

(Wang等,2025c)直接把不足归因于现有系统缺乏用户导向主动性;Zhang等人(2025b)进一步指出,传统对齐多停留在通用价值或单轮偏好,难以解决冷启动与跨会话一致性,因此,他们提出PersonalAgent,将多轮对话划分为逐轮处理的记忆建模单元,在交互过程中持续推断和更新统一用户画像,从而支持长期一致的主动个性化;PaRT(Niu等,2025)则在此基础上把用户配置文件与当前上下文联合起来,实时决定检索什么、引导什么、最后说什么。这样做提升了主动内容的针对性与长期一致性,但代价是个性化收益高度依赖历史画像质量,并伴随隐私泄露和画像老化问题。

4)场景化交互机制与主动性调控。在真实应用中,更大的问题往往不是模型不会主动,而是主动过度或主动不足。Du等人(2024)在车载会话助手场景中把主动性细分为五个等级,并用"重写+反应+反映"策略控制不同干预强度;Li等人(2025b)在搜索辅助场景中使用目标自适应监督微调(G-SFT)和面向点击的强化学习(C-RL)同时优化目标跟踪与交互质量,把"是否继续追问"改写为可学习的交互收益问题;MapDia(Wu等,2025)则通过主题摘要、主题检索和主动主题切换,解决历史偏好被动遗忘的问题。它们的共同代价是场景依赖很强,一旦离开车载、搜索或长期陪伴等特定环境,主动性等级和奖励定义往往需要重新设计。

1.2.3 数据集

随着以用户为中心的主动对话研究不断深入,研究者们开始构建针对性更强的领域数据集,为系统开发和评估提供坚实基础。如表2所示,这些数据集不仅丰富了研究资源,也推动了主动对话从行为层面到认知机制的深入探索。

在情感支持场景中,ExTES(Zheng等,2023b)作为ESConv(Liu等,2021a)的扩展,收录了11,117个完整对话,特别强调系统需主动识别用户情绪并提供共情式回应。该数据集为情感支持型主动对话系统提供了高质量的训练和评估资源。更进一步,ProPsyC(Hu等,2025)数据集专门面向心理支持类主动对话,包含2,001个高质量多轮对话,并标注了策略决策逻辑与反应归因等解释性标签。该数据集不仅记录了对话行为,更揭示了"为何在此刻采取某种主动行为"的认知机制,标志着主动对话数据集从单纯的行为记录向认知机制建模演进。

在说服性对话这一非协作性任务中,P4G(Wang等,2019)提供了1,017个说服性对话,详细记录了"说服者"如何通过主动策略引导"被说服者"向慈善机构捐款的过程。该数据集为研究非合作性场景下的主动对话策略提供了宝贵资源,揭示了主动引导在说服性任务中的关键作用。

为支持更精准的用户状态理解,Wang等人(2025c)引入了ISCO-800数据集,这是一个用于构建对话智能体的多样化用户背景数据集,包含了丰富的用户属性和背景信息。在更广泛的应用场景中,GTEA(Li等,2015)记录了参与者执行日常生活任务的以自我为中心的视频,为理解用户在真实环境中的行为提供了基础;EgoTaskQA(Jia等,2022)则包含关于人类对世界的信念以及模型对人类信念的理解的问题,为构建具有心理理论的对话系统提供了数据支持。

对于主动行为评估,Lu等人(2025)提出的ProactiveBench是一个包含6,790个事件的综合数据集,旨在提炼基于LLMs的智能体的主动行为。该数据集通过系统化的事件分类和标注,为评估主动对话系统的性能提供了客观标准。针对更具挑战性的特定垂直场景,Yamasaki等人(2025)扩展了DO-I-DEMAND数据集,专门定义了在用户发表模棱两可言论的家庭场景下预期的主动援助行动,为系统在模糊情境下的主动干预提供了评估依据。在移动端主动助手场景中,ProactiveMobile(Kong等,2026)将用户画像、设备状态、世界信息与行为轨迹统一为设备端上下文,围绕多场景、多意图构建评测任务,并要求模型生成对应的可执行功能序列。这表明以用户为中心的主动系统数据资源正从静态对话样本扩展到真实设备生态中的连续辅助场景。

1.2.4 评估协议

当前针对以用户为中心的主动对话系统的评估,主要从主观体验、语义质量、特定应用场景和行为有效性四个层面展开。

首先,在主观体验方面,研究者常借助成熟的用户体验量表来量化用户对系统主动性的感知。例如,Li等人(2025a)在Satori系统中采用NASA任务负载指数(NASA TLX)评估用户的认知负荷,并结合可用性量表衡量整体交互满意度,从而间接反映系统主动行为是否自然、适度且不造成干扰。

其次,在语义与内容质量层面,多项工作引入了

表2　以用户为中心的数据集

Table 2　User-centric datasets

名称	发表信息	内容	规模	对话级别
ExTES	(Zheng等,2023b)	情感支持	11k	多轮
ESConv	(Liu等,2021a)	情感支持	1,053	多轮
ProPsyC	(Hu等,2025)	含解释性标注的主动式心理对话	2,001	多轮
P4G	(Wang等,2019)	说服	1,017	多轮
ISCO-800	(Wang等,2025c)	用户画像	800个用户背景	—
GTEA	(Li等,2015)	自我中心视角视频	4名受试者,71个动作类别,525个动作实例	—
EgoTaskQA	(Jia等,2022)	自我中心视角的视频问答数据集	40k	单轮
ProactiveBench	(Lu等,2025)	日常生活场景	6,790个事件	多轮
DO-I-DEMAND	(Yamasaki等,2025)	家庭环境模糊话语	400个场景	—
ProactiveMobile	(Kong等,2026)	移动设备上的主动交互智能	3660实例,14类场景	—

注:"—"表示不涉及该项内容

基于大语言模型的自动评估框架。Wang等人(2025c)提出利用LLM-as-judge策略构建"评论家"模块,从三个关键维度量化面向用户的主动性:1)用户兴趣契合度——即用户是否对系统主动发起的内容表现出兴趣;2)背景相关性——系统是否能基于用户历史或当前上下文选择恰当话题;3)响应价值——回应是否兼具信息性与吸引力。同一时期的Niu等人(2025)也采用LLMs作为评估器,聚焦于个性化(personalization)、信息有用性(informativeness)和会话连贯性(communication skills)三项指标。在此之前,以LLMs为基础的判断已显示出与人的判断相一致(Chiang等,2023;Zheng等,2023a)。

进一步地,针对特定应用场景(如情感支持),研究者整合了心理学领域的标准化量表以增强评估效率。Fu等人(2025)不仅采用GPT-4o与ESC-Rank框架(Zhao等,2024)进行自动评分,还引入六个基础对话质量指标(流畅度、多样性、同理心、信息度、类人度、技巧性),并结合两个经典心理测量工具:安慰反应量表(Clark等,1998)与变更能力评定量表(Jones等,2004),从而全面评估系统在情绪支持场景中主动干预的有效性。

在行为有效性层面,评估重点转向系统主动行为是否"恰到好处"且具有功能性。由于主动对话中往往存在多个合理动作(如提问、建议、共情等),传统准确率难以适用。Yamasaki等人(2025)因此提出三项替代指标:部分匹配(Is-in)、宏观召回(macro-Recall)与微观召回(micro-Recall),以更灵活地衡量预测动作与人类标注之间的覆盖关系。此外,Lu等人(2025)通过训练奖励模型来模拟用户对"任务适配性"的判断,实现对主动行为合理性的端到端评估。

值得注意的是,一些最新工作开始定义面向主动对话特性的专用指标。例如,Hu等人(2025)在PsyAdvisor中提出策略匹配率(strategy fit ratio,SFR)与主动提问率(proactive questioning ratio,PQR):SFR仅在系统策略引发用户积极反馈时才计为成功,强调主动行为的实际效果;PQR则统计主动提问在所有策略中的占比,用以衡量系统发起探索性交互的倾向性。这类指标将评估焦点从"是否主动"转向"主动是否有效",体现了评估理念从形式向功能的演进。

与以目标为中心的方法相比,以用户为中心的方法其有效性来自对"用户是否愿意接受主动性"的直接建模:系统不再只优化任务完成,而是同时优化认知负荷、情绪支持和兴趣契合度。这也意味着意图主动内部存在清晰分工:前者更适合任务边界清楚、成功标准明确的推荐、咨询与工作流场景,后者更适合陪伴、情感支持、搜索辅助等需要长期关系维护的场景。若以统一评估协议审视,意图主动研究的关键不在于生成是否更主动,而在于任务收益与

交互负担之间的平衡是否被显式量化。

1.3 本节小结

综合1.1节与1.2节可以看出，意图主动的内部差异首先体现为优化目标不同：以目标为中心的方法优先压缩到达任务终点的路径长度，因此在Goal Succ.、SR@t与#Turns等指标上更具优势；以用户为中心的方法优先提升可接受性与长期关系质量，因此在NASA TLX、SFR、PQR以及个性化相关指标上表现更具解释力。二者都依赖对用户意图或语义目标的建模，但本质差别在于“主动性究竟服务于任务推进还是关系维护”。从适用边界看，当目标清晰、可验证且交互成本可量化时，目标导向方法更具确定性；当用户需求隐性、情绪负荷显著或需要长期记忆支撑时，用户导向方法更有优势。这一比较也表明，意图主动并非简单的内容生成问题，而是围绕语义目标组织主动行为的一类方法论集合。

2 时序主动

时序主动（temporal proactivity）指对话系统在无显式用户触发的前提下，基于对时间上下文、用户状态与环境动态的理解，自主决定何时介入对话的能力。它并非孤立的时间点选择，而是一个贯穿“感知—决策—生成”全链路的协同过程，包含三个紧密耦合的核心环节：时机预测、策略规划与响应生成。

需要强调的是，时序主动并不排斥显式目标、用户意图或任务约束的建模；其与意图主动的本质区别在于，系统优化的第一目标是“何时介入才能使收益大于打扰成本”，而非仅仅决定“朝哪个语义目标推进”。因此，如果将时间窗口、干预必要性与中断代价作为主导决策变量的方法，即使同时包含明确目标建模，仍应归入时序主动范畴。例如，ProVox（Grannen等，2025）虽然包含用户意图与偏好建模，但其主动性主要体现在交互过程中提前提出下一步协作建议，而非仅进行语义目标选择，因此归入时序主动范畴。

2.1 时机预测

时机预测旨在建模并识别用户处于“可被有效辅助”状态的时刻。一方面，干预越接近用户执行错误或陷入困境的临界点，其纠正价值越高；另一方面，若系统过早响应，可能因误判用户意图而引入不必要的中断，增加认知负荷甚至引发反感。因此，精准的时机预测需在干预价值与中断成本之间取得动态平衡，并高度依赖对用户当前情境、认知状态及社交上下文的综合理解。

2.1.1 理论基础

在主动对话系统中，时机预测是实现有效干预的核心环节，其理论基础源于对人类社交互动和认知过程的深刻理解。如图5：“黄金时间窗口”（goldilocks time window；Fang等，2025b）概念揭示了干预时机的矛盾特性：干预过早会增加假阳性率（false positive rate），导致不必要的中断；干预过晚则会降低干预效率，使用户无法及时获得帮助。认知负荷理论（cognitive load theory；Sweller，2011）进一步为其提供了理论支撑，指出干预时机直接影响用户的认知负荷和接受度。

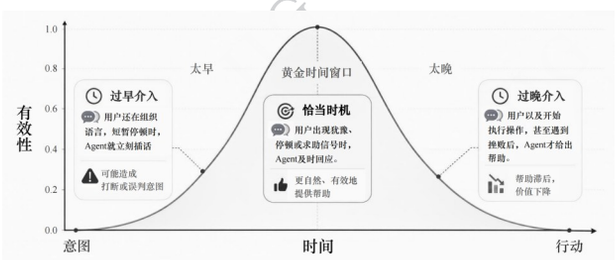

图5 黄金时间窗口

Fig. 5 Goldilocks time window

2.1.2 现有技术

当前时机预测的关键问题不是系统能否生成帮助内容，而是它如何判断“现在插话到底值不值”。之前的问题在于：过早介入会打断用户，过晚介入又会错失黄金窗口，而仅凭停顿长度或显式请求做规则触发，很难覆盖真实环境中的社交线索、认知负荷和时间约束。因此，现有方法逐渐把时机预测从表层检测转向对“干预收益-中断成本”的权衡。

EgoSoD的出发点是很多不合时宜的干预并非因为内容错误而是系统没有读懂社交场。Wang等人（2025b）基于EgoSocial提出主动社交检测方法EgoSoD（egoSocial detect），将交互场景、视觉提示与音频提示整合到社交图谱推理模块中，只在用户分心或对话流受阻等必要时刻介入。其核心洞察是：时机不是单一时间点分类，而是多方社会关系和互动状态共同决定的。代价则是方法依赖高密度、多模态社交标注，跨场景泛化与部署成本都较高。

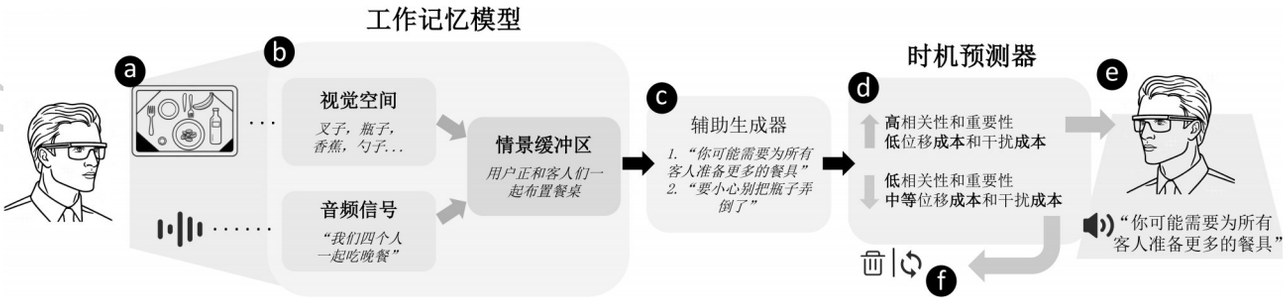

图6　ProMemAssist工作流程图（a）多模态输入（视觉和听觉）；（b）工作记忆模型；（c）辅助生成器；（d）时间预测器；（e）具有高预测价值的消息将以主动语音辅助的方式进行推送；而（f）价值较低的消息则可能会被推迟处理，以供日后评估，或直接丢弃。

Fig. 6　ProMemAssist workflow. (a) multimodal sensor input (visual and auditory); (b) working memory model; (c) assistance generator; (d) the timing predictor; (e) messages with high predicted utility are delivered as proactive voice assistance; while (f) lower-utility messages may be deferred for later evaluation or discarded.

还有问题来自用户正在处理任务、系统却不知道打断代价有多高，ProMemAssist（Pu等，2025）便借用工作记忆（working memory，WM）模型来估计“现在说会不会增加认知负担”。如图6所示，它通过多模态传感器构建用户WM的实时状态，将环境视觉空间与语音记忆项编码为情节组块，再同时评估援助价值与中断成本，只有当净收益足够高时才发出提示。这样做提升了介入选择性，但代价是工作记忆建模对感知质量和认知假设都十分敏感，一旦估计偏差过大，系统就会过度保守或过度打断。

另一类工作把瓶颈归因于系统缺少“内部准备”过程。Liu等人（2025）提出Inner Thoughts框架，让主动对话AI在公开交流之外并行运行一条“内部思维”序列，通过触发、提取、思维形成、评估与参与五个阶段持续生成并评估内在动机，只在表达动机足够强时选择发言。其洞察是，恰当时机来自内部动机成熟，而非外部输入一出现就立即响应；代价则是决策过程更长，且内部动机强度难以外部校验。

当问题转向长时视频或事件流场景时，研究者开始把时机预测视为可训练、可评测的独立能力。TIMELYCHAT（Jang等，2025）要求模型显式预测适当时间间隔，强调“何时回应”比“说什么”更基础；Dispider（Qian等，2025）则用感知、决策和反应分离的异步架构缓解“既要实时盯住画面，又要生成细致回复”的冲突；ProactiveVideoQA（Wang等，2025d）进一步通过PAUC（proactive area under curve）量化响应时机与内容质量的联合效果。它们共同说明，时机预测的收益不仅来自更准的分类器，还依赖训练目标、系统架构和评估协议的共同设计，代价则是系统更复杂，基准与部署条件也更难统一。

与这类能力直接对应的评测资源也开始出现。ProAgentBench（Tang等，2026）将主动辅助显式拆解为“时机预测+帮助内容生成”两个层次，并基于500余小时真实用户会话构建含28，000余事件的benchmark，说明时机预测已经不再只是响应生成的附属模块，而被视为可独立比较的核心能力。相应代价是。这类真实工作流基准的场景噪声更高、历史依赖更强，模型对长期记忆和上下文质量也更敏感。

2.1.3　关键挑战

尽管时序主动性的研究取得了初步进展，但时机预测作为一个核心环节，仍面临着若干严峻的、相互关联的挑战。这些挑战贯穿于数据、评估与模型泛化等多个层面，制约着技术的进一步发展和实际应用。

首要挑战源于数据标注的极高成本与复杂性。时机预测需要的是对“最佳介入时刻”这一极其微妙且主观的概念进行标注，这远比对静态内容分类要困难。标注者需要深刻理解对话的上下文、社交动态以及用户的潜在认知状态，这使得大规模获取高质量、高一致性的标注数据变得异常艰难（Zhang等，2025c）。尽管有研究尝试通过合成数据来缓解这一问题，但其生成的数据往往难以完全复现真实人类互动中充满不确定性和细微差别的复杂场景。

其次，该领域尚未建立起统一且可靠的评估标准。如何科学地量化“时机恰当”，是一个尚未解决的开放性问题。传统的自然语言生成指标，如困惑度或BLEU、BERTScore分数，完全无法捕捉时机选择的优劣。而像PAUC（Wang等，2025d）这样考虑了时间动态的新指标虽具启发性，但仍在推广和共识

建立的早期阶段。缺乏公认的基准和评估协议，导致不同研究之间的性能对比困难，也使得模型优化的方向不够明确。更重要的是，最终的评判标准——用户体验（如帮助性、自然度、非侵入性）通常依赖于成本高昂且难以规模化的人类评估，这进一步给研究进程带来了巨大的挑战。

最后，现有时机预测模型的泛化能力普遍不足。一个在特定任务或封闭数据集上表现良好的模型，往往难以迁移到新的场景中。例如，一个在任务指导场景中，如ProMemAssist（Pu等，2025），基于工作记忆模型做出精准预测的系统，可能无法适应开放域社交对话，如EgoSocial（Wang等，2025b），中依赖于社交线索的时机判断。模型的这种难泛化性，根源在于其对多样化、动态变化的上下文缺乏深层且鲁棒的理解能力。当前的多模态大模型虽然能感知到社交互动的发生，但仍难以精准把握那稍纵即逝的最佳介入点，这揭示了当前技术在实现通用社会智能方面存在的差距。

2.2 策略规划

时序主动中的策略规划是一种将策略执行时机与策略内容紧密耦合的决策机制，旨在实现“恰到好处”的主动交互。与传统反应式交互不同，该规划方法通过实时感知环境上下文、用户状态和任务需求，动态确定最佳策略执行时机，使系统能够在用户需要之前主动提供支持，同时避免不必要的干扰。其核心在于将“何时”与“如何”的决策过程整合为一个统一的优化问题，确保策略在时序维度上与用户认知状态和任务需求之间的匹配。

2.2.1 问题定义

设$C(t)$表示时刻t的环境上下文信息，U表示用户状态，T表示任务目标，S表示策略。时序主动中的策略规划可定义为：

$$\left(t^{*},S^{*}\right)=\arg\max_{t,S}\ V\left(C(t),U,T,S\right)\#(2)$$

式中：t^{*}是最优干预时机；S^{*}是与t^{*}匹配的最佳策略；$V(C(t),U,T,S)$是价值函数，衡量在t时刻执行策略S的预期回报。

该函数定义体现了时序主动中策略规划的核心思想：在给定当前环境、用户和任务条件下，同时优化时机选择和策略生成，以最大化交互的长期价值。通过这一联合优化，系统能够动态适应环境变化，在适当的时机执行最合适的策略，从而提供自然、有效且不干扰的主动交互体验。

2.2.2 现有技术

当前策略规划的难点已经不是单独回答“何时响应”，而是如何把“何时介入”和“介入后采取什么行动”联合起来优化。之前的瓶颈在于，仅有时机判断并不能保证策略合适；在同样的时机下，澄清、建议、等待或求助会带来完全不同的用户体验。因此，现有方法普遍把策略规划看成一个受用户反馈、环境状态和任务语义共同约束的序贯决策问题。

一条路线试图解决“策略只看历史文本，却看不见环境变化”的瓶颈。Real-World Assistive Agents（Chang等，2025）和ContextAgent（Yang等，2025）通过视频、音频和可穿戴上下文实时判断是否需要主动调用工具，把环境变化、用户状态和工具使用价值统一纳入决策。其洞察是，策略规划不能脱离现场情境独立完成；代价则是多模态感知链路更长，一旦环境判断失误，就会同时放大“何时介入”和“如何介入”两类错误。

更难的瓶颈在于系统需要在指令尚未明确前就预判是否主动，研究重点便转向意图预测与反馈修正的耦合。ProVox（Grannen等，2025）利用元提示协议和当前交互状态预测用户意图，在明确指令前主动给出建议；AssistantX（Sun等，2024）则用多智能体框架在突发事件下动态调整策略并向人类求助。它们共同说明，时序策略规划并不是先决定时机、再决定动作的串行流程，而是面向长期交互收益的联合决策过程；代价则是系统更依赖在线反馈质量与外部模块稳定性。

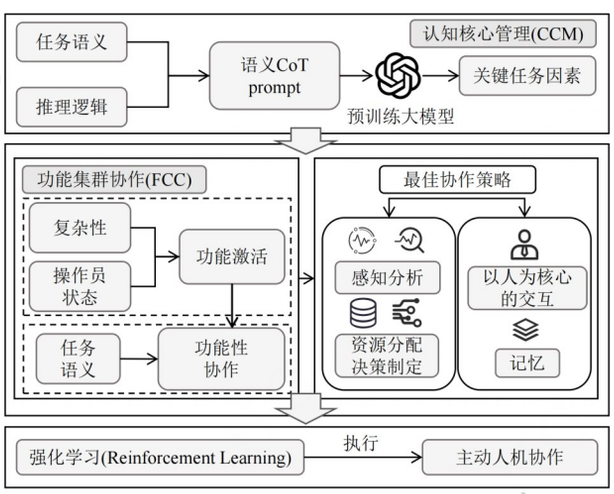

图7 CCM-FCC框架

Fig. 7 CCM-FCC framework

当任务复杂到单一代理难以同时完成理解、推理和执行时,方法开始转向结构化决策。CCM-FCC框架(Ding等,2025)把瓶颈定义为关键任务因素难以及时提取,因此借助语义链提示驱动的认知核心、功能激活模块和深度强化学习,根据任务复杂性与操作员状态生成协作策略。如图7所示,这类方法的洞察是把主动规划嵌入认知约束和模块协同之中,而不是把它当作一句建议的表面生成问题。代价是系统工程复杂度显著上升,调试、解释和实时部署都更困难。

2.2.3 关键挑战

时序主动中的策略规划在实际应用中面临多重关键挑战,这些挑战制约了系统在真实环境中的有效部署和用户体验。

1)多模态感知与上下文理解的局限性。现有系统在多模态感知方面存在不足。Su等人(2025)的研究表明,“当前的语境识别模型无法理解躺、坐、站等对理解语境至关重要的活动”,同时,Yang等人(2025)也指出,“现有的主动智能体要么完全依赖于来自封闭环境的观察,进行直接的LLMs推理,要么使用基于规则的主动通知,从而导致用户意图理解不佳”。这种感知能力的局限性使得系统难以准确捕捉环境上下文和用户状态,从而影响时机判断的准确性。

2)个性化与动态适应的困难。系统难以实现真正的个性化和动态适应。Xu等人(2024b)的研究揭示了阻碍聊天机器人有效回忆的两个主要挑战:信息分散和缺乏主动指导。同时,偶尔会因为建议中的新信息而出现思维中断,表明系统难以适应不同用户的认知能力和偏好变化。

3)多用户场景的处理能力有限。当前系统大多只能处理单用户场景。Su等人(2025)明确指出,“我们目前的设置只能处理单用户场景。我们计划探索区分家庭中多个人的声音的方法”。但在真实环境中,多用户交互是常态,现有系统缺乏有效处理多用户环境的能力,导致在家庭、办公室等多用户场景中性能的下降。

4)隐私与伦理问题。随着系统收集更多用户数据,隐私和伦理问题日益突出。Xu等人(2024b)的研究提到,“我们的研究涉及使用个人照片来唤起个人记忆”,并采取了“告知参与者将照片提交给第三方服务的隐私风险,并征得他们的同意”等措施。这表明,时序主动中策略规划在收集和使用用户数据时面临重大隐私挑战。近期,MineContext项目(ByteDance,2025)受到了广泛关注,这是一个主动的情境感知型人工智能合作伙伴。它通过利用截图和内容理解(包括文档、图片、视频、代码和外部应用数据),可以看到并理解用户的数字世界上下文,然而其对用户的隐私性是一个巨大的隐患。

5)策略泛化能力不足。系统在跨场景泛化方面存在明显缺陷。Ding等人(2025)的研究指出:“虽然具有记忆和交互功能的AI智能体具有更强的适应性,但其针对任务的设计导致缺乏整体认知,从而限制了其泛化能力”。这导致系统在新场景中需要大量重新训练,难以在不同环境中提供一致的高质量服务。

2.3 响应生成

在主动对话系统中,响应生成不再仅是语言内容的构造,更关键的是对何时响应与如何与时序上下文协同的精准把握。相较于传统被动式对话系统遵循“请求—响应”的静态模式,主动系统需在动态演进的交互流中识别干预时机、预测用户潜在需求,并在恰当时刻嵌入辅助信息,从而实现“先于提问而答”或“伴随行为而助”的智能交互范式。这种能力的核心在于将时序维度深度融入响应生成机制,使系统行为与人类活动节奏、认知负荷及任务进程形成同步。

2.3.1 时序响应的特征

时序主动对话中的响应生成具有三个关键特征:1)主动性(proactivity):响应并非由用户显式触发,而是由系统基于环境感知、用户行为或任务进展主动发起,体现“先于需求”的干预逻辑;2)时机敏感性(timing sensitivity):响应的有效性高度依赖于其插入对话流的精确时刻。过早可能造成干扰,过晚则失去意义,因此系统须具备对响应时机的判断能力;3)上下文-时间耦合性(context-time coupling):响应内容不仅依赖静态对话历史,还需融合动态的时间演化信息(如任务阶段、情绪变化、环境事件),实现内容与时机的双重适配。

2.3.2 现有技术

时序响应生成的核心问题是让内容本身服从时机约束:同一句话,早说可能打断,晚说又会失去价值。传统生成模型默认收到输入就立即回答,很少显式建模触发时点、干预强度和跨会话适应。因此,

当前方法把响应生成重写为“何时说、说多少、说成什么形式”的联合优化问题。

LLAMAPIE(Chen等,2025)针对的瓶颈是,大模型擅长生成但不适合全天候、低成本地盯住对话流。它的洞察是将“看时机”和“写内容”拆开:轻量级模型持续监测停顿和语义模糊点,只有判定值得介入时才唤起LLM生成简短、非侵入式提示。Compeer(Liu等,2024)进一步看到,真正重要的不是任何停顿,而是情绪转折或关键决策等重大事件,因此把事件性质直接纳入响应规划。如图8所示,Alpha-Service(Wen等,2025)则把这一思路扩展到持续视频场景,由中央处理单元统筹服务机会检测和讲解触发。代价是两阶段或中央统筹架构虽然减少了无意义发言,但也引入更多的调度延迟与系统耦合。

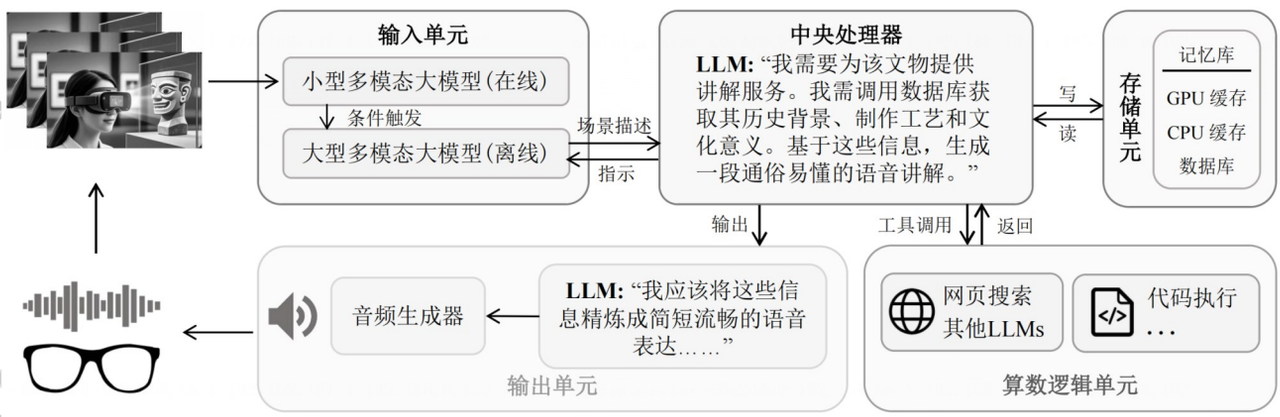

图8　Alpha-Service框架

Fig. 8　The architecture of Alpha-Service

如果单一文本信号不足以判断该如何说,多模态方法就是让视觉、听觉和社会线索共同决定响应形式。Islam等人(2023)把面部表情、生理信号和语言模型结合起来服务情绪干预;SocialMind(Yang等,2024)利用AR眼镜捕捉眼神、语调和姿态来生成社会建议;RoboOmni(Wang等,2025a)则用全模态LLMs把感知、思考、谈话和执行串成端到端链路;Pask(Xie等,2025)通过持续音频监控在用户困惑点主动解释。收益是响应更贴近现场语境,代价则是感知噪声、算力消耗和时延都会更直接地反映到生成质量上。

另一类工作试图解决“当下合适,长期未必合适”的问题。Mirai(Fang等,2025a)通过建模个体习惯周期,把响应从即时提示扩展为对未来行为节点的预干预;Alpha-Service中的存储单元同样利用跨会话偏好历史生成定制化讲解。其核心洞察是,时序响应的价值常常取决于长期模式而非单次语境;代价则是需要持续收集个人数据,并面临画像老化与隐私风险。

最后,Memoro(Zulfikar等,2024)把瓶颈转向“如何几乎不被感觉到地帮助用户”,通过“无查询模式”在自然间隙插入低干扰提示;PROASSIST(Zhang等,2025c)则提供训练数据和自动评测,让“响应时机恰当性”与“内容相关性”可以被系统比较。这说明时序响应生成的改进不仅依赖更强的生成器,还依赖对干预强度和评测标准的同步约束。

2.3.3　关键挑战

时序主动对话系统在响应生成层面面临着多维度的系统性挑战,这些挑战深刻影响着系统在真实场景中的应用效果与用户体验。

计算与能源限制构成了技术落地的核心瓶颈,Alpha-Service(Wen等,2025)等系统部署在资源受限的边缘设备上,如AI眼镜,需要同时处理连续视频流分析和大模型推理,这与设备有限的算力和能效形成尖锐矛盾,导致系统在真实场景中难以达到理想的响应速度与服务质量。通用化与个性化的平衡难题贯穿系统设计始终,过度依赖通用策略导致服务缺乏针对性,而过度个性化则可能陷入“信息茧房”,影响服务的多样性和公平性。同时,冷启动问题在缺乏历史交互数据的情况下尤为突出,使得系统难以在用户首次使用时提供高质量的个性化服务。

真实环境中的稳健性与可扩展性挑战同样不容忽视,系统在极端照明、背景噪声或多人交互等复杂

场景中表现不稳定,多模块协作的稳定性与错误恢复机制也有待完善。隐私与伦理问题随着系统持续感知用户环境信息而日益凸显,即使采用本地化存储,长期行为记录与个性化偏好学习仍可能引发隐私泄露担忧,而用户对主动干预的接受度与信任建立更是系统成功的关键,部分用户可能对AI的积极干预感到不适或产生依赖,需要可解释的决策机制与用户反馈接口来逐步建立信任,这使得系统设计在技术实现与人文关怀之间面临双重考验。

数据与评估体系的不足进一步制约了系统的发展,高质量、大规模的真实世界时序主动对话数据稀缺,现有数据集多基于合成内容,对话与视频内容的一致性有待提高,同时评估指标多聚焦静态内容质量,缺乏对"时机恰当性"、"干预必要性"等时序维度的统一量化标准,使得系统优化缺乏明确导向。这种数据与评估的局限性导致系统开发陷入"数据-性能-数据"的循环困境,难以有效推进技术进步,也阻碍了研究者对时序主动对话系统真实价值的准确评估。

2.4 本节小结

若在PAUC、干预必要性、用户满意度、内容相关性与非侵入性等指标下综合比较,时序主动三层方法呈现出明显的功能分工:时机预测决定系统

能否"少打扰而不错失",因而最直接影响PAUC与误触发率;策略规划决定系统在正确时刻采取何种行动,更关联任务成功率、长期收益与满意度;响应生成则最终决定干预是否被感知为有用、自然且值得继续互动,对应内容相关性、帮助性与人类评分。从代表机制看,时机预测更依赖社交线索推理、工作记忆建模、内部思维与时间上下文建模,优势是能够直接降低误触发并提升干预必要性判断,局限则是数据标注昂贵、最佳时机主观性强且跨场景迁移仍有限;策略规划更依赖反馈驱动决策、强化学习、多智能体协同与认知模型约束,优势是更适合平衡长期收益、用户满意度与任务完成效率,局限则是系统复杂度高,对在线反馈与环境感知质量更敏感;响应生成则更常依赖异步触发、全模态响应生成、低干扰表达与长期时序适应,优势是能够把内容相关性与时机恰当性联合起来,局限则是受硬件算力、时延、鲁棒性与部署约束影响更明显。

需要说明的是,由于主动对话尤其是时序主动仍是近两年快速发展的新方向,大量工作建立在自定义场景、自建数据和闭源模型之上,因此目前尚难给出同一基准、同一骨干模型和同一部署约束下的完全公平实验对比。基于这一现实,本文将比较重点放在方法原理、评估协议、适用边界与披露完整度上,以避免产生表面数值可比但实际条件并不一致的误导性结论。

3 未来方向与讨论

3.1 数据集构建与优化

主动对话系统的发展离不开高质量的数据集支持。现有的数据集多为通用型,涵盖了多种对话场景,但大多缺乏针对个性化、情境化对话的专门设计。未来的研究应当注重构建更加丰富和精细化的个性化对话数据集,包括用户行为模式、历史交互、认知状态以及环境变化等多方面信息的集成。这些数据集不仅应包含用户对话的基本文本信息,还应结合多模态数据,如视觉、音频、手势和生理数据等,以全面反映用户在实际情境中的需求和反应。

此外,数据集的收集方式也需要更加灵活和多样,结合用户长期的行为历史与实时反馈,形成具有长期价值的交互数据。这些数据的积累不仅有助于系统的持续学习,还能为用户提供更加精准和个性化的服务。

3.2 个性化与自适应能力提升

当前主动对话系统多为通用设计,缺乏对用户个体差异和动态状态的深度理解。未来应发展系统的持续学习能力,能够基于用户历史交互、个人偏好(如对中断的容忍度)及实时行为模式,动态调整时机预测策略。具体而言,系统应集成长期记忆模型,将当前工作记忆内容与用户过往任务习惯和认知特征关联,实现对援助价值的个性化推断。同时,引入轻量级交互式反馈机制(如允许用户确认、推迟或否决系统干预),既能提升用户体验,又能为模型提供直接学习信号,形成系统与用户共同适应的良性循环。

针对许多对话系统在没有关于用户的任何信息时就开始对话的情况,可采用元学习的跨用户迁移机制或基于少量交互反馈的在线微调,实现无需大量历史数据的快速适配。此外,引入多样性约束防止个性化过度收敛,避免陷入"信息茧房"。通过这些方法,即使在缺乏用户历史数据的情况下,系统也

能提供初步但有效的个性化服务，并随着交互次数增加逐步优化其响应。

在动态个性化机制方面，系统应开发更精细的个性化策略，包括“使用互动历史记录个性化回复”、“根据用户环境确定信息优先级”、“定制聊天机器人角色”和“自定义描述样式”等。这些方法使系统能够根据用户历史行为、当前环境和认知状态动态调整策略，提供更符合用户需求的主动支持。

3.3 多模态感知与交互技术深化

多模态融合是突破当前时序主动对话系统感知局限的关键路径。未来系统应整合视觉、音频等多模态数据，实现更全面的环境感知。如CACAS系统（Su等，2025）计划“整合视觉和音频数据来实现能够识别活动的多通道活动识别方法”；ProVox框架（Grannen等，2025）也指出“可以集成视觉语言模型，甚至视觉语言动作模型，用于人-机器人协作”，使系统能够更准确地理解用户意图和环境，为交互提供更丰富的信息。

在交互通道扩展方面，未来研究也可整合超越当前音视频的感知信号，如眼动追踪、手势和生理数据等，以更精准地估算用户注意力焦点和认知负荷。干预方式也应从单一语音或文本扩展到触觉反馈、界面微调等更丰富、更轻柔的形式，其强度和模态可根据预测的用户心理可用性进行自适应调整，实现最小化的侵入感。同时，通过融合先进的多模态大模型直接从原始视频/音频生成上下文响应，发展环境自适应策略，实现噪声环境中以语音为主导、视觉清晰时启用细粒度动作指导的动态切换，提升系统在复杂环境中的表现。

3.4 评估框架与伦理设计完善

当前主动对话系统评估缺乏统一标准，需要建立多维度、细粒度的评估框架。未来工作应超越单一智能体指标，将自动化指标与深入的人类评估相结合，开发成本效益高、可扩展的自动化评估基准，例如探索利用大语言模型进行模拟用户评估的新范式，以加速实验迭代。评估体系应涵盖“时机准确性”、“干预必要性”等关键维度，构建全面的时序主动对话评估标准。

此外，从本文所覆盖工作看，LLM在主动对话系统中已不再只是响应生成器，而至少承担四类支撑角色：1）规划器/决策器，负责目标分解、策略搜索与主动行为选择；2）用户状态解释器，用于从历史对话、多模态线索和长期记忆中抽取隐式需求；3）评估器，以LLM-as-judge形式近似人类判断；4）多模态执行器，将感知、推理与响应生成端到端耦合起来。值得注意的是，当前文献对模型家族、版本与部署形态的披露并不一致：少数工作明确报告GPT-4o等具体模型，部分工作仅表述为“轻量级专用LLMs”或“全模态LLMs”，还有一些工作只说明使用通用LLM而未给出版本、参数规模或是否闭源。这种报告缺口会直接影响性能比较、成本评估与可复现性判断。因此，后续综述与实验应将“LLM骨干披露完整度”纳入标准报告项，至少同步报告模型版本、模态能力、是否经过SFT/RL适配、推理时延与部署成本。

如表3所示，当前主动对话研究最突出的问题是：LLM究竟以何种角色介入系统、该角色对应的模型配置是否被充分披露。对于本综述比较而言，

如果缺乏对底座模型版本、参数规模、模态能力与部署约束的同步报告，就很难判断性能提升究竟源于主动策略本身，还是源于更强的通用模型能力。

在知识层面，主动系统需与外部知识源（如知识图谱和搜索引擎）进行深度融合，确保干预行为建立在可验证事实基础上，减少“幻觉”问题。同时，隐私保护与伦理设计成为系统可持续发展的关键。未来研究可以专注于在移动设备上使用本地视觉语言模型来增强数据隐私，允许用户在开始会话前过滤内容，并设计透明化交互机制，赋予用户对主动干预的最终决定权。

4 结　语

本文围绕主动对话系统这一前沿研究方向，系统梳理并整合了近年来在该领域的关键进展。针对“主动对话”这一核心问题，本文从方法论层面出发，依据是否引入时序维度这一判别标准，将现有研究划分为意图主动与时序主动两大范式，并分别对其建模逻辑、技术路径与实现机制进行了结构化解析。

在意图主动方面，本文进一步区分了以目标为中心和以用户为中心两类建模范式，揭示了前者侧重任务达成效率、后者强调个性化需求理解的不同设计取向；在时序主动方面，本文将其解构为时机预测、策略规划与响应生成三个环节，阐明了当前研究在感知用户状态、预判干预窗口及生成上下文适配响应等方面的代表性方法及其局限性。

表3 LLM对主动对话系统的支撑方式与披露情况(代表性示例)

Table 3 Representative LLM support roles and disclosure status

代表性工作	主要支撑角色	模型/版本披露情况	评估性判断
LUSTER,Spec-TOD,AppAgent-Pro	规划器/决策器	多为LLM或“轻量级专用LLMs”,具体骨干与版本披露不统一	便于比较规划范式,但难以公平分离“方法设计收益”与“基座模型能力收益”,成本比较也不充分
Satori,PaRT,MapDia	用户状态解释器	多为多模态大模型或通用LLMs,通常未明确参数规模与部署形态	有助于提升个性化、长期记忆与上下文建模能力,但用户建模收益与骨干模型能力高度耦合
Wang等(2025c),Niu等(2025),COCOON	评估器	Fu等(2025)明确采用GPT-4o,其余工作多仅说明使用LLM-as-judge	自动评估效率较高,但judge偏差、版本敏感性与闭源模型可复验性仍需单独控制
LLAMAPIE,RoboOmni,Pask	多模态执行器	多为大语言模型或“全模态LLMs”,版本、时延和硬件约束披露有限	能够强化时序响应与环境适配,但部署成本、推理时延与鲁棒性之间的横向比较仍不充分

此外,本文对当前主流的主动对话数据集进行了横向比较,涵盖数据规模、对话轮次、场景多样性等维度,并阐述了面向任务成功率、用户满意度及介入时机准确性的多维评估指标体系。在此基础上,指出现有评估框架仍缺乏统一标准,尤其在时序主动维度上缺少细粒度、可复现的评测基准。

进一步地,本文强调意图主动与时序主动的分界应以主导优化目标而非是否存在显式目标建模来判断,并将LLM骨干与支撑方式的透明报告视为后续比较研究的必要条件。

最后,本文总结了若干制约主动对话系统发展的共性挑战:1)高质量标注数据稀缺,尤其缺乏包含多模态信息及用户对主动干预细粒度反馈的真实场景数据集;2)现有系统缺乏对用户个体差异、动态认知状态的深度理解,自适应与持续学习能力不足;3)多模态融合仍处于初级阶段,现有系统难以有效协同语音、文本、视觉等异构信号以提升环境适应能力;4)伦理与隐私问题尚未形成系统性解决方案,用户对主动行为的可控性与透明度需求亟待满足。

本综述不仅为理解主动对话系统的演进脉络提供了清晰的分类框架,也为后续研究在方法创新、系统实现与伦理治理等方面指明了可行路径,有助于推动人机交互向更自然、智能与可信的方向发展。

参考文献(References)

Adlakha V, Dhuliawala S, Suleman K, de Vries H and Reddy S. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. Transactions of the Association for Computational Linguistics, 10: 468 - 483 [DOI: 10.1162/tacl_a_00471]

Anantha R, Vakulenko S, Tu Z C, Longpre S, Pulman S and Chappidi S. 2021. Open-domain question answering goes conversational via question rewriting//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics: 520 - 534 [DOI: 10.18653/v1/2021.naacl-main.44]

Baddeley A. 2012. Working memory: theories, models, and controversies. Annual Review of Psychology, 63: 1 - 29 [DOI: 10.1146/annurev-psych-120710-100422]

Budzianowski P, Wen T H, Tseng B H, Casanueva I, Ultes S, Ramadan O, et al. 2018. MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 5016 - 5026 [DOI: 10.18653/v1/D18-1547]

Chang R C. 2025. Enabling real-world assistive agents: from live vision to proactive context-aware information delivery//Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. New York, NY, USA: Association for Computing Machinery: 2 [DOI: 10.1145/3746058.3758468]

Chen T C, Batchelder N S, Liu A, Smith N A and Gollakota S. 2025. Llamapie: proactive in-ear conversation assistants//Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics: 13801 - 13824 [DOI: 10.18653/v1/2025.findings-acl.710]

Chiang C H and Lee H Y. 2023. Can large language models be an alternative to human evaluations? //Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics: 15607 - 15631 [DOI: 10.18653/v1/2023.acl-

long.870]

Clark R A, Pierce A J, Finn K, Hsu K, Toosley A and Williams L. 1998. The impact of alternative approaches to comforting, closeness of relationship, and gender on multiple measures of effectiveness. Communication Studies, 49(3): 224 - 239 [DOI: 10.1080/10510979809368533]

Cowan N. 2010. The magical mystery four: how is working memory capacity limited, and why? Current Directions in Psychological Science, 19(1): 51 - 57 [DOI: 10.1177/0963721409359277]

Deng Y, Liao L Z, Lei W Q, Yang G H, Lam W and Chua T S. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. ACM Transactions on Information Systems, 43(3): 1-45 [DOI: 10.1145/3715097]

Deng Y, Liao L Z, Zheng Z H, Yang G H and Chua T S. 2024. Towards human-centered proactive conversational agents//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington DC, USA: Association for Computing Machinery: 807 - 818 [DOI: 10.1145/3626772.3657843]

Ding P F, Zhang J, Zhang P, Li H S and Wang D X. 2025. Ccm-fcc: llm-powered cognition-centered ai agent framework for proactive human-robot collaboration. Robotics and Computer-Integrated Manufacturing, 98: 103145 [DOI: 10.1016/j.rcim.2025.103145]

Du H F, Feng X J, Ma J, Wang M, Tao S Y, Zhong Y J, et al. 2024. Towards proactive interactions for in-vehicle conversational assistants utilizing large language models//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Jeju, Korea: International Joint Conferences on Artificial Intelligence Organization: 869 [DOI: 10.24963/ijcai.2024/869]

Du H W, Peng B and Ning X. 2025. Planning with diffusion models for target-oriented dialogue systems[EB/OL].[2025-04-23]. https://arxiv.org/pdf/2504.16858.pdf

Eric M, Goel R, Paul S, Sethi A, Agarwal S, Gao S, et al. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines//Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association: 422-428. [DOI: 10.18653/v1/2020.lrec-1.53]

Fang C M, Samaradivakara Y, Maes P and Nanayakkara S. 2025a. Mirai: a wearable proactive ai "inner-voice" for contextual nudging//Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery: 399 [DOI: 10.1145/3706599.3719881].

Fang C M, Zulfikar W, Samaradivakara Y, Nanayakkara, S, and Maes P. 2025b. The goldilocks time window for proactive interventions in wearable ai systems. https://arxiv.org/pdf/2504.09332.pdf

Feng S T, Lin H C, Lubis N, Niekerk C V, Heck M, Ruppik B, et al. 2025. Emotionally intelligent task-oriented dialogue systems: Architecture, representation, and optimisation [EB/OL].[2025-07-02]. https://arxiv.org/pdf/2507.01594.pdf

Feng S, Patel S S, Wan H and Joshi S. 2021. Multidoc2dial: modeling dialogues grounded in multiple documents//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics: 6162 - 6176 [DOI: 10.18653/v1/2021.emnlp-main.498]

Fu X, Li H Z, Wang B C, Yang H, Zhao Y Y and Qin B. 2025. Look beyond feeling: unveiling latent needs from implicit expressions for proactive emotional support//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics: 21582 - 21609 [DOI: 10.18653/v1/2025.emnlp-main.1094]

Grannen J, Karamcheti S, Wulfe B and Sadigh D. 2025. Provox: personalization and proactive planning for situated human-robot collaboration. IEEE Robotics and Automation Letters, 10(8): 8451 - 8458 [DOI: 10.1109/LRA.2025.3583460]

Guo S S, Liao L Z, Zhang J, Li C P and Chen H. 2024. PCQPR: proactive conversational question planning with reflection [EB/OL].[2024-10-02]. https://arxiv.org/pdf/2410.01363.pdf

He T, Liao L Z, Cao Y X, Liu Y X, Sun Y H, Chen Z R, et al. 2024. Simulation-free hierarchical latent policy planning for proactive dialogues//Proceedings of the AAAI Conference on Artificial Intelligence. 39(22): 24032 - 24040 [DOI: 10.1609/aaai.v39i22.34577]

Hu Y X, Liu D N, Liu B, Chen Y D, Cao J X and Liu Y. 2025. Psyadvisor: a plug-and-play strategy advice planner with proactive questioning in psychological conversations// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics: 12205 - 12229 [DOI: 10.18653/v1/2025.acl-long.596]

Huang M L, Zhu X Y and Gao J F. 2020. Challenges in building intelligent open-domain dialog systems. ACM Transactions on Information Systems, 38(3): 21 [DOI: 10.1145/3383123]

Islam R and Bae S W. 2023. Revolutionizing mental health support: an innovative affective mobile framework for dynamic, proactive, and context-adaptive conversational agents//Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp '23), GenAI4PC Symposium.

Jia B X, Lei T, Zhu S C and Huang S Y. 2022. Egotaskqa: understanding human tasks in egocentric videos//Advances in Neural Information Processing Systems. Vol. 35. Red Hook, NY: Curran Associates, Inc.: 3343 - 3360.

Jang S B, Jeon M J, Lee J H, Lee S H, Lee D H and Yu H J. 2025. From what to respond to when to respond: timely response genera-

tion for open-domain dialogue agents[EB/OL].[2025-06-17]. https://arxiv.org/pdf/2506.14285.pdf

Jones S. 2004. Putting the person into person-centered and immediate emotional support: emotional change and perceived helper competence as outcomes of comforting in helping situations. Communication Research, 31 (3): 338 - 360 [DOI: 10.1177/0093650204263436]

Kim N Y, Ong K T I, Hwang Y J, Kang M S, Jihn I I, Kim G Y, et al. 2025. PRINCIPLES: synthetic strategy memory for proactive dialogue agents//Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics: 21329 - 21368 [DOI: 10.18653/v1/2025.findings-emnlp.1164]

Kong D Z, Feng Z Z, Liang Q L, Wang H, Sun H F, Yang C P, et al. 2026. ProactiveMobile: A comprehensive benchmark for boosting proactive intelligence on mobile devices//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, Nevada, USA: IEEE/CVF: 27503-27513.

Li C Y, Wu G D, Chan G Y Y, Turakhia D G, Castelo Quispe S, Li D, et al. 2025a. Satori: towards proactive ar assistant with belief-desire-intention user modeling//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery: 1229 [DOI: 10.1145/3706598.3714188]

Li J W, Galley M, Brockett C, Gao J F and Dolan B. 2016. A diversity-promoting objective function for neural conversation models//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics: 110 - 119.

Li M D, Li Y C and Kong F. 2026. Enhancing target-guided proactive dialogue systems via conversational scenario modeling and intent-keyword bridging[EB/OL].[2026-05-12]. https://arxiv.org/pdf/2605.11964.pdf

Li X Y, Li X, Gao L, Liu Y D, Wang X Y, Wang S Q, et al. 2025b. Proactive guidance of multi-turn conversation in industrial search [EB/OL].[2025-05-30]. https://arxiv.org/pdf/2505.24251.pdf

Li Y, Ye Z F and Rehg J M. 2015. Delving into egocentric actions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE: 287 - 295.

Liu S Y, Zheng C J, Demasi O, Sabour S, Li Y, Yu Z, et al. 2021a. Towards emotional support dialog systems//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics: 3469 - 3483 [DOI: 10.18653/v1/2021.acl-long.269]

Liu T J, Zhao H Z, Liu Y H, Wang X B and Peng Z H. 2024. Compeer: a generative conversational agent for proactive peer support//Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. Pittsburgh, PA, USA: Association for Computing Machinery: 117 [DOI: 10.1145/3654777.3676430]

Liu X B, Fang S T, Shi W Y, Wu C S, Igarashi T and Chen X A. 2025. Proactive conversational agents with inner thoughts// Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery: 1-19 [DOI: 10.1145/3706598.3713760]

Liu Z M, Wang H F, Niu Z Y, Wu H, Che W X and Liu T. 2020. Towards conversational recommendation over multi-type dialogs// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics: 1036 - 1049 [DOI: 10.18653/v1/2020.acl-main.98]

Liu Z M, Wang H F, Niu Z Y, Wu H and Che W X. 2021b. Durecdial 2.0: a bilingual parallel corpus for conversational recommendation// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics: 4335 - 4347 [DOI: 10.18653/v1/2021.emnlp-main. 356]

Lu Y X, Yang S Z, Qian C, Chen G R, Luo Q Y, Wu Y S, et al. 2025. Proactive agent: shifting llm agents from reactive responses to active assistance//The Thirteenth International Conference on Learning Representations.

Miller G A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2): 81 - 97 [DOI: 10.1037/h0043158]

Mo L B, Jiang S, Maharaj A V, Hishamunda J B and Li Y Y. 2025. Hiertod: a task-oriented dialogue system driven by hierarchical goals//Proceedings of the 6th Workshop on Data Science with Human-in-the-loop (DaSH) at VLDB 2025.

Nguyen V Q, Chieu N Q, Pham H V and Bui K H N. 2025. Spec-tod: a specialized instruction-tuned llm framework for efficient task-oriented dialogue systems//Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Avignon, France: Association for Computational Linguistics: 133 - 145.

Niu Z H, Xie Z Y, Cao S S, Lu C G, Ye Z Y, Xu T, et al. 2025. Part: enhancing proactive social chatbots with personalized real-time retrieval//Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. Padua, Italy: Association for Computing Machinery: 4269 - 4274 [DOI: 10.1145/3726302.3731946]

ByteDance. 2025. https://github.com/volcengine/MineContext/tree/main

Papineni K, Roukos S, Ward T and Zhu W J. 2002. Bleu: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics: 311 - 318.

Pu K, Zhang T, Sendhilnathan N, Freitag S, Sodhi R and Jonker T R.

2025. Promemassist：exploring timely proactive assistance through working memory modeling in multi-modal wearable devices//Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. New York，NY，USA：Association for Computing Machinery：56［DOI：10.1145/3746059.3747770］

Qi H G，Tan M K，Gao L，Liu Y，Cong R M，Shi Z W，et al. 2026. Top 10 frontier challenges in computer image and graphics. Journal of Image and Graphics，31(1)：0001-0010（齐洪钢，谭明奎，高林，刘越，丛润萌，史振威，等. 2026. 图像图形领域十大前沿科技问题. 中国图象图形学报，31(1)：0001-0010）［DOI：10.11834/jig.2600001］

Qian R，Ding S R，Dong X Y，Zhang P，Zang Y H，Cao Y H，et al. 2025. Dispider：enabling video llms with active real-time interaction via disentangled perception，decision，and reaction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，NJ，USA：IEEE：24045 - 24055.

Rakib T B A，Mehrish A，Soon L K，Lim W H，and Poria S. 2025. Dialogxpert：driving intelligent and emotion-aware conversations through online value-based reinforcement learning with llm priors［EB/OL］.［2025-05-23］.
https://arxiv.org/pdf/2505.17795.pdf

Sayres R，Hao Y X，Ward A，Wang A，Freeman B，Zhan S，et al. 2025. Towards better health conversations：the benefits of context-seeking［EB/OL］.［2025-09-14］.
https://arxiv.org/pdf/2510.18880.pdf

See A，Roller S，Kiela D and Weston J. 2019. What makes a good conversation? How controllable attributes affect human judgments. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Volume 1（Long and Short Papers）. Minneapolis，Minnesota：Association for Computational Linguistics：1702-1723.［DOI：10.18653/v1/N19-1170］

Sweller J. 2011. Chapter two - cognitive load theory//Mestre J P and Ross B H（eds.）. Psychology of Learning and Motivation. Vol. 55. Academic Press：37 - 76［DOI：10.1016/B978-0-12-387691-1.00002-8］

Su Z D and Sheng W H. 2025. Context-aware proactive and adaptive conversation for human - robot interaction. Robotics and Autonomous Systems，195：105207［DOI：10.1016/j.robot.2025.105207］

Sun N，Mao B，Li Y C，Guo D and Liu H P. 2024. Assistantx：an llm-powered proactive assistant in collaborative human-populated environment［EB/OL］.［2025-06-19］.
https://arxiv.org/pdf/2409.17655.pdf

Tang J H，Zhao T C，Xiong C Y，Liang X D，Xing E and Hu Z T. 2019. Target-guided open-domain conversation//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence，Italy：Association for Computational Linguistics：5624 - 5634［DOI：10.18653/v1/P19-1565］

Tang Y B，Tang H Z，Cao T Y，Nguyen L，Zhang A P，Cao X W，et al. 2026. ProAgentBench：evaluating llm agents for proactive assistance with real-world data［EB/OL］.［2026-02-09］.
https://arxiv.org/pdf/2602.04482.pdf

Wang J，Cheng Y，Lin D D，Leong C and Li W J. 2023. Target-oriented proactive dialogue systems with personalization：problem formulation and dataset curation//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore：Association for Computational Linguistics：1132 - 1143［DOI：10.18653/v1/2023.emnlp-main.72］

Wang J，Lin D D，and Li W J. 2024. Target-constrained bidirectional planning for generation of target-oriented proactive dialogue. ACM Transactions on Information Systems，42(5)，1-27［DOI：10.1145/3652598］

Wang S Y，Fu J L，Liu F H，He X Z，Wu H X，Shi J H，et al. 2025a. Roboomni：proactive robot manipulation in omni-modal context［EB/OL］.［2025-11-01］.
https://arxiv.org/pdf/2510.23763

Wang X J，Sharma T，Kulshrestha A，Meka A，Purohit A and Manocha D. 2025b. Egosocial：benchmarking proactive intervention ability of omnimodal llms via egocentric social interaction perception［EB/OL］.［2025-10-15］.
https://arxiv.org/pdf/2510.13105.pdf

Wang X W，Shi W Y，Kim R，Oh Y J，Yang S J，Zhang J W，et al. 2019. Persuasion for good：towards a personalized persuasive dialogue system for social good//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence，Italy：Association for Computational Linguistics：5635 - 5649［DOI：10.18653/v1/P19-1566］

Wang Y F，Hu J W，Huang Z T，Lin K Y，Zhang Z T，Chen P H，et al. 2025c. Enhancing user-oriented proactivity in open-domain dialogues with critic guidance［EB/OL］.［2025-05-18］.
https://arxiv.org/pdf/2505.12334.pdf

Wang Y Q，Meng X J，Wang Y F，Zhang H S and Zhao D Y. 2025d. Proactivevideoqa：a comprehensive benchmark evaluating proactive interactions in video large language models［EB/OL］.［2025-07-15］.
https://arxiv.org/pdf/2507.09313.pdf

Wei F，Chen D Y，Wang C，Huang Y L，Chen Y S，Pan X C，et al. 2025. Grounded in reality：learning and deploying proactive llm from offline logs［EB/OL］.［2025-10-29］.
https://arxiv.org/pdf/2510.25441.pdf

Wen X H，Liu W，Yue X D and Chen Y F. 2026. Survey of trustworthy medical image analysis methods on the basis of evidential deep learning. Journal of Image and Graphics，31(7)：2505-2525（文昕颢，刘伟，岳晓冬，陈宇飞. 2026. 应用证据深度学习的可信医学影像分析方法综述. 中国图象图形学报，31(7)：2505-2525）［DOI：10.11834/jig.250475］

Wen Z C，Wang Y Y，Liao C F，Yang B X，Li J X，Liu W F，et al.

2025. Ai for service: proactive assistance with ai glasses [EB/OL]. [2025-10-16]. https://arxiv.org/pdf/2510.14359.pdf

Wu B W, Wang W Q, Li H R, Li Y, Yu J S and Wang B X. 2025. Interpersonal memory matters: a new task for proactive dialogue utilizing conversational history//Proceedings of the 29th Conference on Computational Natural Language Learning. Vienna, Austria: Association for Computational Linguistics: 47 - 67 [DOI: 10.18653/v1/2025.conll-1.4]

Wu Z Q, Parish R, Cheng H, Min S, Ammanabrolu P, Ostendorf M, et al. 2023. Inscit: information-seeking conversations with mixed-initiative interactions. Transactions of the Association for Computational Linguistics, 11: 453 - 468 [DOI: 10.1162/ tacl_a_00559]

Xie Z F, Hu Z Z, Zhang G B, Zhang J L, Liao Y, Miao C Y, et al. 2025. Pask: providing answer before asking toward proactive ai agent//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: Association for Computing Machinery: 13552 - 13554 [DOI: 10.1145/3746027.3754487]

Xu K S, Cheng Y, Hou W J, Tan Q Y, Li W J. 2024a. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics: 6796-6814. [DOI: 10.18653/v1/2024.findings-acl.406]

Xu S C, Chen C, Liu Z C, Jin X F, Yuan L P, Yan Y K, et al. 2024b. Memory reviver: supporting photo-collection reminiscence for people with visual impairment via a proactive chatbot//Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. Pittsburgh, PA, USA: Association for Computing Machinery: 88 [DOI: 10.1145/3654777.3676336]

Yamasaki K, Tanaka S, Yuguchi A, Kawano S and Yoshino K. 2025. Multi-step or direct: a proactive home-assistant system based on commonsense reasoning//Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Avignon, France: Association for Computational Linguistics: 561 - 572.

Yang B F, Guo Y Q, Xu L L, Yan Z Y, Chen H K, Xing G L, et al. 2024. Socialmind: llm-based proactive ar social assistive system with human-like perception for in-situ live interactions[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2025, 9(1): 1-30 [DOI:10.1145/ 3712286]

Yang B F, Xu L L, Zeng L K, Liu K W, Jiang S Y, Lu W R, et al. 2025. Contextagent: context-aware proactive llm agents with open-world sensory perceptions[EB/OL].[2025-10-27]. https://arxiv.org/pdf/2505.14668.pdf

Zang X X, Rastogi A, Sunkara S, Gupta R, Zhang J G and Chen J D. 2020. MultiWOZ 2.2: a dialogue dataset with additional annotation corrections and state tracking baselines//Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI. Online: Association for Computational Linguistics: 109 - 117 [DOI: 10.18653/v1/2020.nlp4convai-1.13]

Zhan H L, Li Z, Wang Y F, Luo L H, Feng T, Kang X X, et al. 2023. SocialDial: A benchmark for socially-aware dialogue systems//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, Taiwan: Association for Computing Machinery: 2712-2722 [DOI: 10.1145/3539618.3591877]

Zhang D D, Fan Y X, Li P F, and Zhu Q M. 2025a. Enhancing goal-oriented proactive dialogue systems via consistency reflection and correction[EB/OL].[2025-06-18]. https://arxiv.org/pdf/2506.13366.pdf

Zhang T Y, Kishore V, Wu F, Weinberger K Q and Artzi Y. 2020. BERTScore: evaluating text generation with BERT//International Conference on Learning Representations.

Zhang X, Deng Y, Ren Z F, Ng S K and Chua T S. 2024. Ask-before-plan: proactive language agents for real-world planning//Findings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida: Association for Computational Linguistics: 10836 - 10863.

Zhang X T, Wang Y, Chen R Z, Wang Z Y, Hou R C and Liu Z Z. 2025b. Towards proactive personalization through profile customization for individual users in dialogues[EB/OL].[2025-12-17]. https://arxiv.org/pdf/2512.15302

Zhang Y C, Dong X L, Lin Z J, Madotto A, Kumar A, Damavandi B, et al. 2025c. Proactive assistant dialogue generation from streaming egocentric videos[EB/OL].[2025-06-06]. https://arxiv.org/pdf/2506.05904

Zhao H Q, Li L Y, Chen S S, Kong S Q, Wang J A, Huang K X, et al. 2024. Esc-eval: evaluating emotion support conversations in large language models//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics: 15785 - 15810 [DOI: 10.18653/v1/2024.emnlp-main.883]

Zhao Y Y, Shi W T, Feng F L and He X N. 2025. AppAgent-Pro: a proactive gui agent system for multidomain information integration and user assistance//Proceedings of the 34th ACM International Conference on Information and Knowledge Management. Seoul, Republic of Korea: Association for Computing Machinery: 6767 - 6771 [DOI: 10.1145/3746252.3761473]

Zheng L M, Chiang W L, Sheng Y, Zhuang S Y, Wu Z H, Zhuang Y H, et al. 2023a. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena//Advances in Neural Information Processing Systems. Red Hook, New York, USA: Curran Associates, Inc.: 46595-46623. [DOI: 10.52202/075280-2020]

Zheng Z H, Liao L Z, Deng Y and Nie L Q. 2023b. Building emotional support chatbots in the era of llms[EB/OL].[2023-08-17]. https://arxiv.org/pdf/2308.11584.pdf

Zulfikar W D, Chan S and Maes P. 2024. Memoro: using large language models to realize a concise interface for real-time memory augmen-

tation//Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Honolulu, HI, USA: Association for Computing Machinery: 450 [DOI: 10.1145/3613904.3642450]

作者简介

刘嘉玲，女，博士研究生，主要研究方向为大模型主动对话、情感计算。E-mail：jialingliu1@ruc.edu.cn

张鑫洁，女，博士研究生，主要研究方向为数字心理健康、大模型、情感计算。E-mail：zhangxinjie827@ruc.edu.cn

金琴，通信作者，女，教授，主要研究方向为多模态大模型、情感计算、具身智能。E-mail：qjin@ruc.edu.cn