UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

LLM-BASED LONG-TERM LIFE TASK PLANNING TO REDUCE HUMAN UNCERTAINTY

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
DOCTOR OF PHILOSOPHY

By

BEN WANG
Norman, Oklahoma
2025

LLM-BASED LONG-TERM LIFE TASK PLANNING TO REDUCE HUMAN UNCERTAINTY

A DISSERTATION APPROVED FOR THE
SCHOOL OF LIBRARY AND INFORMATION STUDIES

BY THE COMMITTEE CONSISTING OF

Dr. Jiqun Liu, Chair

Dr. Andrew H. Fagg

Dr. June Abbas

Dr. Yong Ju Jung

Acknowledgments

First and foremost, I would like to express my deepest and most sincere gratitude to my advisor, Dr. Jiqun Liu. I met Dr. Liu during a challenging period in my life. I was facing an unsuccessful PhD application cycle. Dr. Liu offered me a chance to work on his projects, recognized my effort, and offered me a position as his PhD student. I am eternally grateful for this opportunity; it was the turning point that made this entire dissertation possible. Throughout my PhD journey, Dr. Liu has been more than an advisor; he has been a true mentor. He shared not only his deep research expertise but also his invaluable insights on learning and life. From the very beginning, he challenged me to clarify my own goals and to understand the "why" behind my work. He guided me to find my own path, encouraging me to actively explore opportunities in both academia and industry. He taught me resilience, spurring me on after paper rejections and celebrating with me during successes. The central theme of task planning in this dissertation was inspired by the combination of our research in interactive information retrieval and Getting Things Done, a book he recommended to me, just as his own advisor had recommended it to him during his PhD journey. I am profoundly honored to be a part of this "passing of the torch".

I would also like to extend my heartfelt thanks to my dissertation committee members for their invaluable guidance and support. I am grateful to Dr. Andrew Fagg. He was incredibly generous with his time, meeting with me often to help me learn. His ability to explain complex principles with profound simplicity had a major impact on my understanding. In fact, the name of "GOLF" in this dissertation, was named from an analogy (comparing LLM prompting to golf) that emerged during one of our conversations. I am also deeply thankful to Dr. Yong Ju Jung. Throughout my research, her kindness, patience, and significant encouragement provided me with crucial motivation and confidence. Her supportive words were a source of strength. My sincere appreciation also goes to Dr. June Abbas. As the our school director, her

leadership and the supportive environment she has fostered were instrumental to my success. I am grateful for the help and guidance she provided throughout my entire doctoral program, which made the journey smoother.

My research was also shaped and strengthened by the broader academic community. I am grateful for the chance to have met and learned from so many brilliant senior researchers in doctoral consortiums, workshops, and conferences. I want to thank Dr. Johanne Trippas, Dr. Rob Capra, Dr. Leif Azzopardi, Dr. Grace Hui Yang, Dr. Ryen White, Dr. Chirag Shah, and Dr. Nicholas Belkin. Their insightful conversations, invaluable guidance, and encouragement helped validate my research direction and made me feel truly welcomed in this field.

I am grateful to my family. A very special thank you goes to my grandfather. He was the one who enlightened me, planting the seed in my mind at a very young age that pursuing a PhD was a worthy and noble goal. This dissertation is, in many ways, the fulfillment of that seed he planted so long ago. I hope I have made him proud.

Finally, I want to thank myself for not giving up, for staying curious, and for seeing this through to the end.

Dedication

This dissertation is dedicated to my significant one, Yaping Zhang, for teaching me the importance of planning, for her support and pure love.

Table of Contents

List Of Tables

List Of Figures

Abstract

In long-term life tasks, people often face challenges from uncertainty in tasks and information-seeking, which can create difficulties in decision-making and task completion. Recent advancements in Artificial Intelligence (AI), especially in Large Language Models (LLMs), offer transformative capabilities in domain-specific task planning and problem-solving. Despite these innovations, there is limited understanding of how such technologies can be applied to assist humans in long-term life tasks. This dissertation work seeks to address this gap by exploring how human-AI collaboration, mediated through LLM-based agents, can improve long-term life task planning and uncertainty management. To achieve this, this dissertation first proposes the long-term life task type and investigates how people may use AI tools to assist them in planning long-term life tasks and cope with uncertainty. Secondly, it proposes the Goal Oriented Long-term liFe planning (GOLF) framework that integrates LLMs to emulate human-AI collaboration in diverse long-term life task domains. The framework facilitates task decomposition and iterative planning while evaluating uncertainties throughout the process. Thirdly, the study introduces a GOLF benchmark to evaluate and generalize the effectiveness of LLMs in long-term planning scenarios. This study operates at the intersection of cognitive science, information science, and human-AI interaction, and makes multifaceted contributions. Theoretically, by introducing the concept of long-term life tasks and proposing the GOLF framework, it explores the potential of human-AI collaborations for long-term planning and uncertainty management. Methodologically, through a simulation-based approach and the LLM benchmark, it enables systematic evaluation of LLMs in long-term planning scenarios across domains. This work bridges the gap between human cognitive strategies and AI planning capabilities, providing a robust foundation for future advancements in AI-assisted task management and goal achievement.

Chapter 1

Introduction

Problem-solving is a fundamental aspect of human existence, enabling individuals to navigate complexities and achieve their goals (Anderson, 1993; Newell & Simon, 1972). However, uncertainty is inherently tied to problem-solving. In the context of this study, uncertainty is defined as a human cognitive state arising from incomplete information, ambiguous future outcomes, and dynamic environments (Abbott, 2005; Chowdhury et al., 2011, 2014). Specifically, this dissertation focuses on three operational dimensions of this cognitive state: outcome uncertainty (ambiguity regarding the results of a plan), action uncertainty (lack of clarity regarding the specific steps required), and intolerance of uncertainty (the negative emotional response to these ambiguities) (Carleton et al., 2007; Maes et al., 2022). Managing this uncertainty is crucial for effective decision-making and successful outcomes (Allen, 2015; Arbona et al., 2021). To contend with uncertainty, individuals employ task management strategies, such as goal setting and task planning (Bateman & Barry, 2012; Grant, 2003; Sniehotta et al., 2005). By decomposing complex objectives into actionable steps, people can reduce ambiguity and gain clearer pathways toward their goals (Allen, 2015; Eichmann et al., 2019). Despite these strategies, human cognitive limitations, such as bounded rationality and finite working memory, can impede our ability to manage multifaceted, long-term tasks (Kahneman, 2003; Newell & Simon, 1972). These constraints often lead to suboptimal decision-making and heightened uncertainty (Azzopardi, 2021; Tversky & Kahneman,

2000a, 2000b). To mitigate the negative effects of these limitations, various systems are designed to better support human tasks.

Information access systems, including search engines and recommendation systems, have emerged to aid individuals in their information-seeking processes (Shah & Bender, 2024; Shah et al., 2023). These systems play a fundamental role in problem-solving by providing access to external resources that inform decisions and reduce uncertainty through information retrieval (IR) and recommendation (Chowdhury et al., 2014; Savolainen, 2012). However, due to both human cognitive limitations and system performance constraints, users often encounter challenges interacting with these systems, such as cognitive barriers (Savolainen, 2015), information overload, and irrelevant results, which can exacerbate uncertainty (Azzopardi, 2021; J. Liu & Azzopardi, 2024).

Recent advancements in artificial intelligence (AI), notably in large language models (LLMs) and LLM-based chatbots and agents, offer promising solutions. These technologies possess sophisticated language understanding and generative capabilities, enabling them to process context and provide more nuanced support in problem-solving (Brown et al., 2020; Huang et al., 2024; Joksimovic et al., 2023; Ouyang et al., 2022; Xi et al., 2023). This technological shift enables AI-assisted planning, which this study defines as a collaborative process where AI agents leverage generative capabilities to decompose high-level goals, propose actionable steps, and augment human reasoning to reduce cognitive load (Hunt, 1994; Inder, 1996; Retzlaff et al., 2024; Sebastia et al., 2006; Y. Wang & Chiew, 2010). It is important to distinguish this from traditional AI planning research, which predominantly focuses on algorithmic optimization and probabilistic modeling (Russell & Norvig, 2016). Unlike purely algorithmic approaches, LLM-based approaches can mirror human reasoning process and behaviors to generate plans in natural language (Huang et al., 2024; Xi et al., 2023). In addition, the

integration of LLMs into information access systems has given rise to generative IR, where AI not only fetches information but also generates responses tailored to user queries. This advancement overcomes previous limitations in IR system understanding of diverse user intents and naturally delivering personalized and comprehensive information. (Deng et al., 2024; Shah & Bender, 2024; Zhao et al., 2024).

Despite these technological developments, a gap remains in applying them cohesively to support humans in the planning and management of long-term life tasks. The long-term life tasks are defined as complex, open-ended objectives that span extended periods, carry high personal significance, and involve uncertainties from multiple sources, such as career development, health management, and financial planning (Cantor et al., 1991; Havighurst, 1953). While the execution of these tasks can span years, this study focuses specifically on the planning phase. Planning serves as a critical entry point for human-AI collaboration, allowing users to simulate potential futures and manage uncertainty before execution begins. Current LLM-based planning frameworks mostly focus on smaller-scale and routine tasks that LLM agents can automate with less human intervention. It lacks integration among human cognitive strategies, task management principles, and generative AI planning capabilities, which is crucial for effectively managing human and task uncertainties. On the other hand, users still encounter difficulties and uncertainties interacting with these AI tools, which may hinder them from discovering how to utilize AI tools for long-term task assistance. This research aims to bridge this gap by exploring how human-AI collaboration, leveraging LLM-based agents, can enhance long-term life task planning and uncertainty management. By integrating insights from human problem-solving, task management principles, generative AI planning capabilities, and interactive IR and recommendation, this study aims to develop a unified framework that mitigates the negative effects of uncertainty and improves informed decision-making in long-term life tasks.

1.1 The context of the dissertation

This study operates at the intersection of cognitive science, information science, AI, and human-computer interaction (HCI). Focusing on these perspectives, this study aims to develop a user-aware AI-assisted planning framework that enhances human-AI interaction for long-term life task planning and uncertainty management.

1.1.1 Human problem-solving process

Human problem-solving is a fundamental cognitive function that enables individuals to navigate challenges and make decisions in everyday life. From routine tasks like navigating to a place to complex projects such as scientific research, problem-solving is integral to human intelligence and adaptability. Tracing back to ancient philosopher Aristotle, researchers across multiple disciplines have long been intrigued by the mechanisms underlying this ability (Hunt, 1994). Understanding how humans solve problems sheds light on natural intelligence and informs the development of artificial intelligence systems designed to emulate these cognitive processes (Newell & Simon, 1972).

Simon and Newell (1971) considered problem-solving in a search representation, which introduced the concept of the problem space, encompassing the initial state, intermediate states, goal state, and connecting actions. Search representation frames problem-solving as a process of finding a path in a structured state space, where states represent different steps of the problem, and actions or operators represent transitions between states. The solver (human or machine) navigates this space using strategies such as heuristics or systematic exploration to identify a solution path from the initial state to the goal state.

The search representation emulates cognitive processes of human problem solving, including goal setting, planning, information seeing and processing, and decision making. Goal setting is the initial phase where the desired outcome or solution is identified, providing direction for subsequent actions (Locke & Latham, 2005). This is followed by task planning, which requires strategies and sequencing actions to bridge the gap between the current state and the goal state (Inder, 1996). Information seeking and processing are critical throughout this process, as individuals gather relevant data, evaluate options, and update their understanding of the problem space (Kuhlthau, 1991). Decision-making is the action that helps individuals transit from the current state to the next state in the problem space (Simon et al., 1987). The integration of these steps reflects a dynamic and iterative approach to problem-solving.

Problem-solving can differ significantly depending on whether the problem is new or familiar. For familiar problems, individuals can rely on schema, which is a structured mental framework that individuals acquire and refine through practice and experience (Chi et al., 2014; Cooper & Sweller, 1987). In contrast, new problems require exploratory reasoning and planning, as no established schema exists. Long-term life tasks, such as career planning, financial management, or achieving personal milestones, may seem like new problems to an individual. However, these challenges are not new to human society with most if not all people having faced and resolved similar situations in their lives.

Planning reduces uncertainty by structuring decision-making around available and potential information, which reframes task uncertainty as an information uncertainty problem (Allmendinger, 2017). In addition, task planning can be understood as a learnable schema (Shallice, 1982). The schema patterns for specific problems can be reused by others in similar contexts. By organizing information through planning or learning the schemas from the experiences of others, individuals can anticipate the problem

space, from goal setting to planning, information seeking, and making decisions (Anderson, 1993). Therefore, leveraging others' experiences as common knowledge offers a practical and feasible way to address long-term life tasks, transforming seemingly new personal challenges into manageable processes.

1.1.2 The concept of task in task management and information seeking and retrieval

While problem-solving provides a high-level framework for addressing challenges and achieving goals, tasks represent the tangible, actionable components that operationalize this process. Each task serves as a step or sub-goal within the problem-solving journey, contributing to the progression from the initial state to the desired outcome in the problem space (Inder, 1996). A task is framed as a specific problem that the solver faces, requiring a sequence of actions or decisions to reach a solution (Newell et al., 1958). In information-seeking and retrieval (ISR), tasks drive the user's need for information and shape their strategies for finding and evaluating relevant resources (Soufan et al., 2021; Vakkari, 2003; I. Xie, 2009). These tasks are often context-dependent, reflecting specific goals or challenges that users face in academic, professional, or personal settings (Haraty et al., 2016; H. Xie, 2000). Similarly, in task management, tasks represent the building blocks of planning, organization, and execution, encompassing the processes of defining objectives, prioritizing activities, and allocating resources (Bellotti et al., 2004).

In the context of ISR, tasks are multifaceted activities that can range from everyday queries to complex professional information needs (Vakkari, 2003; I. Xie, 2009). These tasks often trigger a cascade of information-focused actions such as searching, acquiring, organizing, synthesizing, and disseminating information to achieve specific

goals (Byström & Hansen, 2005; Järvelin et al., 2015). The process of addressing these tasks frequently involves navigating through the Information Search Process (ISP), a conceptual model that outlines various stages ranging from initiation, selection, exploration, formulation, and collection, to the presentation of information (Kuhlthau, 1991). This model starts with the user feeling of uncertainty and underscores the dynamic nature of information seeking, spotlighting how individuals' approaches evolve as they gain clarity and refine their information needs. Task hierarchies in ISR further describe how tasks are structured and interrelated, positing that complex work tasks may be decomposed to nested sub-tasks related to information seeking and searching (Byström & Hansen, 2005; Toms, 2019).

Specifically, tasks in ISR can be classified into work tasks and information tasks, where work tasks motivate more focused information tasks (Soufan et al., 2021). Work tasks provide the context and goals that lead to information-seeking behavior, which involves multiple sub-tasks aimed at addressing specific information needs (Järvelin et al., 2015; I. Xie, 2009). These tasks are characterized by their goals, instructions, information resources, and conditions (Toms, 2019). Furthermore, tasks can vary significantly in complexity, ranging from simple, repetitive actions to complex decision-making processes requiring innovative problem-solving and the integration of diverse information sources (Y. Li & Belkin, 2008). They are influenced by users' roles and the task context, reflecting the dynamic relationships between personal objectives and external constraints (J. Liu et al., 2010; J. Liu et al., 2019). Building on these theoretical foundations, IR research and applications have developed systems that consider user and task states to enhance support for search tasks. While addressing uncertainty in information tasks, it is also crucial to explore how to support users and mitigate uncertainty from a holistic task perspective.

1.1.3 Uncertainty in human problem solving and task management

Uncertainty is a multifaceted concept investigated across diverse disciplines, including cognitive science, economics, information science, and AI. In cognitive science, uncertainty is often defined as a mental state resulting from incomplete or conflicting knowledge, which can disrupt decision-making and motivate problem-solving efforts (Tversky & Kahneman, 2000a). In economics, uncertainty refers to situations where future outcomes and risks cannot be predicted with confidence (Alchian, 1950). In information science, uncertainty is tied to disorder and entropy in information systems, reflecting gaps or anomalies in knowledge that impede understanding or decision-making (Chowdhury et al., 2011). In AI, uncertainty arises from the probabilistic and partially observed nature of problem states, where incomplete knowledge of the environment necessitates adaptive and probabilistic reasoning (D. Li & Du, 2017). These definitions illustrate the pervasive and interdisciplinary nature of uncertainty, framing it as both a cognitive state and an environmental condition.

The effects of uncertainty are also investigated in multiple disciplines. In information science, uncertainty drives information-seeking and retrieval behaviors as individuals attempt to resolve gaps or anomalies in their knowledge (Chowdhury et al., 2014; Toms, 2019). From a psychological perspective, uncertainty can affect cognitive processing, lead to indecision, and result in anxiety, particularly for individuals with a low tolerance for uncertainty (Arbona et al., 2021). In task management, uncertainty complicates both the outcomes and the actions. Outcome uncertainty arises when the results of a task are unpredictable, while action uncertainty occurs when the steps necessary to achieve a goal are unclear (Maes et al., 2022). In the field of AI, uncertainty poses significant challenges, particularly in planning and decision-making

systems. Although researchers understand the impacts of uncertainty and have developed strategies to manage it, human still remains vulnerable due to bounded rationality and emotional reactions. In contrast, AI approaches uncertainty as a probability problem, making it a more rational solver in uncertain situations. Therefore, designing systems and processes through human-AI collaboration could be a promising way to support users in tasks and mitigate the effects of uncertainty.

1.1.4 Human-AI collaboration for supporting human tasks

Human-AI Interaction (HAI) refers to the ways humans engage with AI systems, focusing on usability, responsiveness, and user experience (Amershi et al., 2019). Human-AI collaboration (HAC), rooted in Licklider (1960)'s concept of symbiotic computing, envisions AI as an active partner that enhances human reasoning, creativity, and problem-solving rather than merely serving as a tool (Holter & El-Assady, 2024). This vision transcends traditional notions of human-computer interaction (HCI), situating AI as a teammate rather than a machine to be operated, and directing attention toward the richer domain of Computer-Supported Cooperative Work (CSCW) (Amershi et al., 2019; Shneiderman, 2021; D. Wang et al., 2020). While HAI primarily improves user experience through adaptive interfaces and intelligent feedback, HAC requires deeper integration, where AI assists with task management and coordination (W. Xu et al., 2023; Q. Zheng et al., 2022).

Current research in HAC has made progress in high-stakes fields such as medicine, finance, and industrial automation, where AI supports professionals by processing large-scale data, optimizing workflows, and mitigating risks (Fragiadakis et al., 2025; Khakurel & Blomqvist, 2022). The focus on these domains is largely driven by the availability of structured data and societal priorities. However, fewer studies have

explored HAC in everyday life, where collaboration is equally valuable but more challenging due to unstructured tasks and diverse user needs. One promising direction is AI-assisted task management and decision-making, where AI automates routine tasks while preserving human oversight for strategic decision-making (G. Li et al., 2023; O. Liu et al., 2024). Despite the potential, implementing AI in open-ended and daily-life scenarios has been challenging. Traditional AI systems often struggle with generalizing across diverse tasks topics and user preferences (Y. Xu et al., 2020). These limitations have hindered the development of AI systems capable of long-term collaboration with humans in everyday contexts (Gomez et al., 2025).

The emergence of Large Language Models (LLMs) has expanded the possibilities for collaboration. LLMs exhibit remarkable generalization capabilities, enabling them to handle a wide range of tasks and adapt to various contexts (Ouyang et al., 2022). More details of the development and limitations of LLMs are explained in a later section. In terms of HAC, the advancement of LLMs has the potential to extend collaborative scenarios from closed domains to open-domain tasks (Roller et al., 2021). Usage patterns show that people increasingly rely on ChatGPT for complex, multi-step tasks such as practical guidance, information seeking, and content generation (Chatterji et al., 2025; Skjuve et al., 2023). OpenAI has also introduced ChatGPT's learning mode, which supports step-by-step study and long-term skill development (OpenAI, 2025). These developments suggest that users and developers are increasingly exploring LLMs as partners for complex and long-term activities. However, traditional collaboration frameworks have yet to fully adapt to this shift. Current research predominantly focuses on decision support and task planning in short-term and well-defined problems (Gomez et al., 2025; Y. Xu et al., 2020). This leaves a gap in understanding how AI can support humans in long-term life tasks.

1.1.5 Human long-term life task

"Not life, but good life, is to be chiefly valued."—Socrates. Human life tasks and goals present a vast spectrum: for some individuals, it may be as straightforward as striving for happiness and maintaining health, while for others it may be as ambitious and abstract as achieving great success, leaving a meaningful legacy, or transforming society (Dweck, 2006). These life tasks are not monolithic; they differ in complexity, duration, and personal significance (Wong, 2013). Ultimately, such long-term pursuits shape our sense of identity, purpose, and fulfillment (McGregor & Little, 1998). Understanding life goals and tasks is crucial because they represent the core challenges and milestones that guide our decisions and learning throughout the human lifespan. Yet, these goals are not always easily defined, let alone easily achieved.

Within the broad realm of life goals, *life stage tasks* refer to the essential sets of objectives and milestones that most individuals encounter as they progress through distinct phases of their lives (Havighurst, 1953). Life stage tasks encompass various domains, from managing one's finances (e.g., saving for retirement, buying a home) (M. M. Hassan et al., 2021), sustaining health (e.g., establishing exercise routines, preventive healthcare) (Fruh, 2017), advancing education (e.g., acquiring essential qualifications, continuous learning) (Young et al., 2019), to building one's career (e.g., job transitions, leadership development) (Pinheiro et al., 2017). These tasks typically involve setting meaningful goals (Grant, 2003; Lieder et al., 2019), committing to a long-term process of growth (Bateman & Barry, 2012; Höchli et al., 2018; Woolley & Fishbach, 2017), enabling long term or lifelong learning (McCombs, 1991), and meticulously planning steps toward desired outcomes (Sniehotta et al., 2005). In a life journey, these tasks are crucial stepping stones, inherently long-term and are widely shared across human societies (Cantor et al., 1991). Engaging in life stage tasks fosters

personal development, as individuals acquire new skills, refine their values, and gain insight into themselves and the world around them (Dalio, 2019).

Because of the universal importance and complexity, these life stage tasks serve as the central focus of this study and are referred to as **long-term life tasks** in the remaining text, and it is used with *life tasks* and *long-term tasks* interchangeably. Given the importance of augmenting human problem-solving and mitigating uncertainty, this study investigates how human-AI collaboration can support long-term life tasks. The challenge lies in understanding how AI's world knowledge and capabilities in planning can support humans in the life journey. In addition, by viewing long-term life tasks through the lens of LLM agent planning simulation, this study aims to uncover new possibilities of AI-driven planning tools and systems in empowering individuals to achieve greater personal fulfillment.

As long-term life tasks inherently span extended durations, observing their full execution in a research setting presents significant logistical challenges. However, task planning can be viewed as a cognitive simulation of the task execution. By generating a detailed plan or a simulation path in collaboration with AI, the user is essentially *pre-experiencing* the long-term task. Therefore, this dissertation focuses on the planning stage, treating the generated simulation path as a proxy for the long-term collaboration that helps users navigate future uncertainties.

1.2 Problem statement

Long-term life task planning facilitates personal and professional development by helping individuals structure complex objectives such as career advancement, educational attainment, health management, and personal fulfillment. These tasks are markedly more complex than short-term or well-defined tasks. With the recent advent of AI

technologies, particularly LLMs like ChatGPT, there is growing potential for AI-driven support in task planning. These LLMs provide capabilities such as task decomposition, personalized recommendations, and interactive feedback, which may help users organize their thoughts and manage the uncertainties inherent in long-term planning.

Despite notable advancements in AI-assisted planning tools and research, there is limited understanding of how these technologies support users with the long-term planning process. To address this gap, this study introduces "long-term life tasks" as a new task type for human-AI collaboration and focuses on users' perceived uncertainty in planning.

To have an initial understanding of this novel task, it is important to investigate how users respond to uncertainty in long-term life task planning in their previous experience. Users' responses or their coping methods may include strategies to manage or reduce uncertainty, avoidance behaviors, and emotional reactions. By examining these experiences, this study aims to explore approaches to support users in long-term task planning. Comprehending user-centric perspectives on this emerging task type is crucial for designing systems that effectively meet their planning needs. AI users (i.e., users who have experience with AI tools) may utilize AI to shape more effective coping strategies for managing uncertainty in long-term task planning. AI users with familiarity and acceptance of AI systems make them ideal participants for understanding how AI can better support such tasks. Accordingly, this study investigates the first research question:

RQ1: How do AI users cope with uncertainty in long-term life task planning?

Building on insights into user experiences, it is then essential to develop a systematic framework that supports long-term life task planning. This study introduces an LLM-based planning framework, with the assumption that LLMs can provide general

knowledge relevant to life task planning and identify and address user needs over the state space of long-term planning (Guo et al., 2024; Z. Wang et al., 2023). This framework leverages task decomposition and aims to mitigate users' uncertainty during the planning process. Given that conducting extensive user studies for long-term tasks is both costly and time-consuming, simulations have become a widely adopted, cost-effective approach for evaluating the effectiveness of LLMs and validating their capabilities before real-world human-involved experiments. Thus, the second research question focuses on developing and evaluating the framework within a simulated human-AI collaboration environment:

RQ2: How does task decomposition within a simulated human-AI collaboration environment impact long-term task planning and user uncertainty?

This dissertation constructs uncertainty primarily as a human cognitive state that can affect a user's decisions, emotions, and ability to plan long-term life tasks. Understanding this human-centric view of uncertainty is critical. This study first investigates how humans experience and cope with this perceived uncertainty in AI-assisted planning in the first interview study. Building on this cognitive foundation, the second simulation study then explores whether an LLM-based framework can model and evaluate these levels of user-perceived uncertainty, aiming to create a system that aligns more closely with human cognitive processes.

Finally, suppose this LLM-based framework proves effective for long-term task planning. In that case, a standardized approach is needed to ensure that the findings can be generalized and to facilitate the broader improvement of LLM capabilities in this newly defined task type. Developing a benchmark for long-term life tasks would enable a replicable, scalable and consistent method of evaluating how different models

perform on long-term task planning. Therefore, the study addresses the third research question:

RQ3: How can a benchmark be developed and utilized to evaluate the performance of LLMs in long-term task planning and uncertainty management?

By addressing these three questions, I conducted this dissertation study, aiming to (1) understand users' experiences and uncertainties in AI-assisted long-term planning, (2) design and validate an LLM-based collaborative framework, and (3) create a benchmark that supports ongoing evaluation and enhancement of LLMs for long-term life task planning. The following sections and the Literature Review chapter provide additional details of the research background and explore potential solutions to address the outlined research questions.

1.3 AI planning and LLM agent

In AI research, an agent is broadly defined as an autonomous entity capable of perceiving its environment, reasoning about possible actions, and executing tasks to achieve specific goals (Inder, 1996). These agents rely on internal models and algorithms that guide their behavior, often leveraging planning and searching algorithms to determine optimal or near-optimal sequences of actions (Silver & Veness, 2010). Traditional algorithm-based AI planning methods, such as Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs), focus on formulating explicit problem representations—defining states, actions, and goals—and systematically exploring possible action sequences to produce a viable plan (Silver & Veness, 2010). These methods have proven effective in domains like robotics and logistics, they often require domain-specific engineering, significant computation, and carefully crafted knowledge representations (Russell & Norvig, 2016).

In recent years, the field of natural language processing (NLP) has undergone a transformative shift with the advent of generative AI–LLMs. Beginning with the transformer architecture, LLMs have demonstrated remarkable proficiency in processing and generating human-like text (Vaswani et al., 2017). LLMs primarily focused on NLP tasks, such as text generation, summarization, and question answering (Qin et al., 2023). With instruction tuning and few-shot learning, LLMs' capabilities have expanded beyond NLP to broader AI tasks and generic human tasks (Ouyang et al., 2022). Modern LLMs, like GPT, Gemini, Claude, DeepSeek, Qwen, and other transformer-based models, have become increasingly adept at interpreting context, providing coherent responses, and even reasoning about complex problems through human-AI interaction (Gomez et al., 2025; Tu et al., 2024).

Building on these advances, researchers have begun leveraging LLMs as agents within broader systems. An LLM agent is an AI-driven entity that uses its language understanding and generation capabilities not only to communicate but also to plan and act autonomously or semi-autonomously (Xi et al., 2023; Yao et al., 2022). In multi-agent systems, several LLM agents collaborate or compete to solve tasks, each contributing specialized functions such as retrieval, summarization, or decision-making (Guo et al., 2024). This collective approach opens the door for more complex problem-solving capabilities with different forms of reasoning.

A growing body of research now explores how LLM agents can perform or assist with planning. These agents can break down tasks into subtasks, propose action plans, and refine solutions through iterative feedback (Huang et al., 2024). Recent efforts conduct experiments with simulation in the social science domain, in which entire processes, including user behavior and system decision-making, are modeled by LLM agents (Park et al., 2023; Tan & Jiang, 2023). These simulations primarily emphasize automating human tasks with minimal human involvement. While automation can

improve efficiency, a human-centered AI approach requires more human involvement and control to ensure AI aligns with human goals and preferences. With the growing adoption of LLMs in daily life applications, it is foreseeable that LLMs will play a role in personal task management and decision-making. Despite this, human involvement in such applications remains underexplored.

Previous studies highlight the potential of simulations for developing and evaluating complex human-LLM collaboration. Building on this, this research also leverages simulations to assess the effectiveness of an HAC framework in supporting long-term task planning. Additionally, this study emphasizes the simulation of user states as a core component within the framework, aiming to advance the knowledge of Human-Centered AI and the development of human-LLM interaction.

1.4 The dissertation study overview

The study for **RQ1** explores how AI users manage uncertainty in long-term life task planning through a qualitative research design, combining a warm-up activity with an in-depth semi-structured interview. This study recruited 14 participants to engage in self-chosen planning tasks with ChatGPT, covering topics of personal and well-being, activity and event planning, and professional development and learning, followed by semi-structured interviews. By analyzing both their prompts and interview reflections, the study captures not only how people interacted with an LLM but also how they perceived uncertainty, usefulness and limitations in AI-assisted long-term life task planning.

The findings show that participants engaged with ChatGPT through diverse prompting strategies that varied across task domains, including personal and well-being, activity and event, and professional development and learning tasks. They encountered

two main sources of uncertainty, task-inherent (e.g., where to start, feasibility) and AI-introduced (e.g., contextual fit, credibility), which influenced both emotion and behavior. ChatGPT helped structure goals and encourage reflection, but general or overconfident advice sometimes amplified uncertainty. Overall, participants viewed ChatGPT as a useful yet limited planning partner that provided scaffolding, motivational support, and reflection, while lacking adaptability and personalization.

The study for **RQ2** introduces the Goal Oriented Long-term liFe task planning (GOLF) framework, a multi-agent system leveraging LLMs to simulate human-AI collaboration in achieving long-term life tasks across domains like finance, medicine, education, and professional development. The framework models both the user and system components with LLM agents, facilitating iterative interactions where the user, represented by an LLM with predefined preferences, engages with system agents for step planning, communication, information retrieval, and action instruction. The process begins with setting a goal, initial user state, and background settings. An initial planner generates a high-level roadmap, which a step planner then refines into detailed step plans, broken into subtasks of communication, information retrieval, and action instruction for corresponding agents. The user agent responses to these action instructions, updates their states, and the cycle repeats to procced the long-term task.

Regarding the evaluation, the framework integrates task-centric **effectiveness** and user-perspective **uncertainty**. This evaluation process is not for measuring absolute model performance but aims to **validate whether the simulation framework is controllable and sensitive to different settings**. Drawing from task management and cognitive science, the evaluation adopts three types of uncertainty: **outcome uncertainty** (ambiguity regarding the results of a plan), **action uncertainty** (lack of clarity regarding the specific steps required), and **intolerance of uncertainty** (the negative emotional response to these ambiguities) (Carleton et al., 2007; Maes et al.,

2022). Experiments manipulate prompt quality to observe its impact on performance, and ablation studies remove certain components to understand their contributions. The results show that both prompt quality and components significantly influence planning. High-quality prompts improve effectiveness, while user states are critical for aligning plans with user needs. By measuring uncertainties, this study demonstrates how LLM simulations can reveal both strengths and limitations in long-term life task planning.

The study for **RQ3** develops the GOLF benchmark, a dataset designed to assess the ability of LLMs to generate coherent and effective plans for complex, multi-step tasks that individuals encounter in real life. The benchmark covers a diverse range of topics such as financial planning, educational advancement, medical management, and professional development. Specific task topics within these areas are validated by the researcher to ensure they are representative and relevant. Task plans are generated using a simulation based on the GOLF framework with different user states. These plans are transformed into interactive recommendation scenarios. This approach captures a wide range of scenarios that users might face in long-term planning. This study validates the benchmark data with both LLM and human judgments on a sampled benchmark dataset. The results show that the benchmark task might be more complicated for human annotators as the accuracy of assigning action instructions to corresponding user states is close to random guessing, while LLMs can recognize corresponding user states and action suggestions with higher accuracy. The full benchmark measures a series of models, differentiating LLMs' performance in terms of learning and improving in long-term task planning.

1.5 The research gap addressed by this study

Extending the concept of task from information tasks to broader work tasks, this study aims to mitigate the increased uncertainty inherent in long-term life tasks, which can impede users' decision-making processes and progress toward goals (Selten et al., 2012). This study advances research on human–AI collaboration by foregrounding long-term life tasks as a distinct and meaningful context for studying LLM usages. Through the interview study for RQ1, it contributes (1) empirical insight into how users actually engage with conversational AI for long-term life task planning; (2) key uncertainties users encounter when planning with AI and the related impacts; and (3) design implications for AI systems that serve not only as scaffolds for success but also as adaptive, reflective partners that help users navigate uncertainty, challenges, and potential failure in long-term goals.

The simulation study for RQ2 demonstrates how LLM simulations can reveal both strengths and limitations in long-term life task planning. The contributions include: (1) a systematic framework for simulating long-term task planning, (2) empirical evidence on how prompt design and component configuration affect outcomes, and (3) a human-centered perspective on investigating LLMs in long-term human-AI collaboration. Furthermore, it develops a benchmark for RQ3 to generalize the findings of the simulation study and contribute to the evaluation and development of the LLMs' performance on supporting long-term life tasks.

Overall, this dissertation seeks to investigate and support users in planning long-term life task and managing uncertainty, improving overall task performance in complex, long-term life tasks. It advances the understanding of both the opportunities and challenges presented by LLMs, which are crucial for designing systems that effectively

support users' information needs and decision-making processes in an ever-evolving information and AI era.

1.6 Summary

The introduction chapter begins by discussing the fundamental concept of human problem-solving, focusing on how individuals navigate complexities and uncertainties in achieving their goals. It highlights the role of tasks as an actionable component in problem-solving and introduces the challenges posed by uncertainty in dynamic and incomplete information environments. The chapter then transitions to exploring the potential of human-AI collaboration, with a particular focus on long-term life tasks. The problem statement emphasizes the need for effective strategies to integrate LLM into human workflows for long-term task planning and management. Additionally, it introduces the LLM agent and planning, serving as the base of the main methodology. Finally, it provides an overview of the study design, findings, and contributions, centering on investigating LLMs to enhance long-term task planning and reduce uncertainty.

The next chapter presents a literature review, covering the key components of task management, delves into the role of uncertainty in interactive information retrieval studies, and examines both the potential and the challenges of using LLMs for long-term task planning.

Chapter 2

Literature Review

This chapter provides a review of the research in which the dissertation is situated. As Figure 2.1 shows, the overall topic is under broad human problem-solving. To narrow down the broad topic, I focus on three research domains related to human problem solving, covering human goal setting and planning, which is in the task management field; interactive IR and recommender system, which are the primary tools for information access. In addition, it covers LLM-based agents and planning supporting human tasks, which is under the AI field.

Several concepts, such as uncertainty, goal, and planning, are defined and operationalized differently depending on whether the focus is on human cognition and behavior or on algorithmic representations. Because this dissertation investigates human long-term life tasks, the literature review intentionally focuses on human-centered definitions and works of goal formation, task management, and uncertainty. At the same time, the review also discusses selected approaches (e.g., interactive IR, recommender systems, and LLM-based planning), as they are used to support human tasks. Reviewing and discussing previous works in these areas paves the way to understanding (1) why this dissertation study is important, (2) what are the available theoretical and empirical supports from existing research, and (3) in which research area(s) the study is advancing the knowledge.

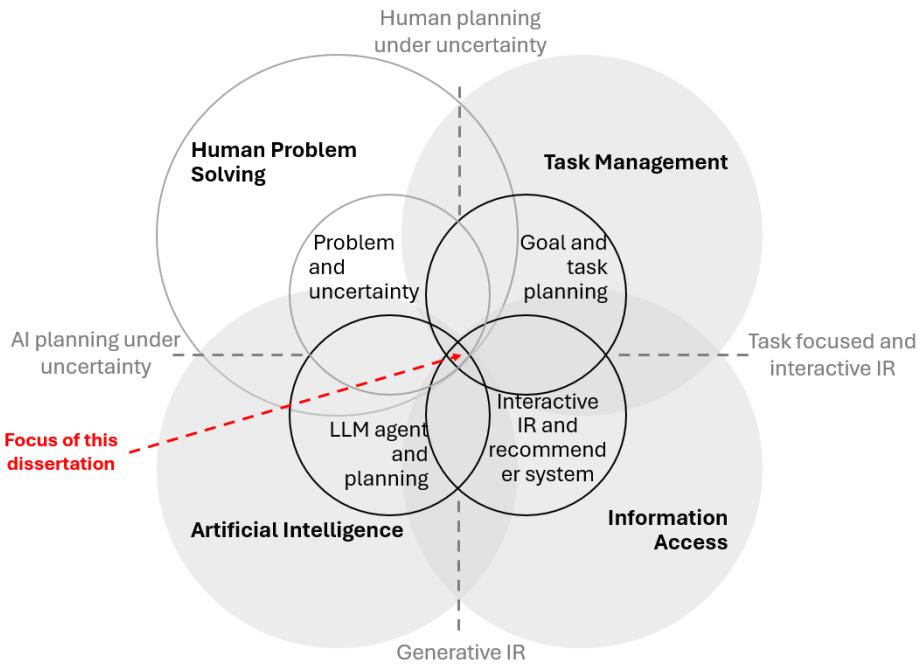

Figure 2.1: Position of the dissertation study. It illustrates the interdisciplinary nature of the study, locating the specific research focus at the intersection of Human Problem Solving, Task Management, AI, and Information Access.

2.1 Key components in task management

Management refers to the systematic planning, organization, and control of resources to achieve specific goals (Fayol, 2016). Task management is an effective process for human problem-solving, offering a structured approach to breaking down complex challenges into manageable steps (Allen, 2015; Haraty et al., 2016). Methods for organizing tasks and managing information enhance decision-making and productivity (Jones et al., 2006). Listed in Figure 2.2, this section explains two key steps of task management: goal setting and planning. I then discuss the effects and measurements of uncertainty and explores tools and systems designed to support task management. It provides insights for the framework and evaluation of human-AI collaboration in task management and planning.

2.1.1 Goal setting

Setting a goal is the foundation of task management, which can guide individuals on how to organize tasks, prioritize efforts, and achieve objectives (Bateman & Barry, 2012). Effective goal setting combines immediate, actionable goals with broader, long-term objectives. Subordinate goals drive short-term productivity, while superordinate goals provide purpose and direction (Höchli et al., 2018). Breaking tasks into subgoals enhances organization and progress tracking. Research shows subgoal management improves learning and retention by fostering self-regulated learning processes (Urgo & Arguello, 2024).

Achieving subgoals or short-term tasks brings immediate rewards. These immediate rewards can be enjoyment or a sense of progress, which are critical for maintaining persistence in goal-related tasks (Woolley & Fishbach, 2017). Immediate benefits make

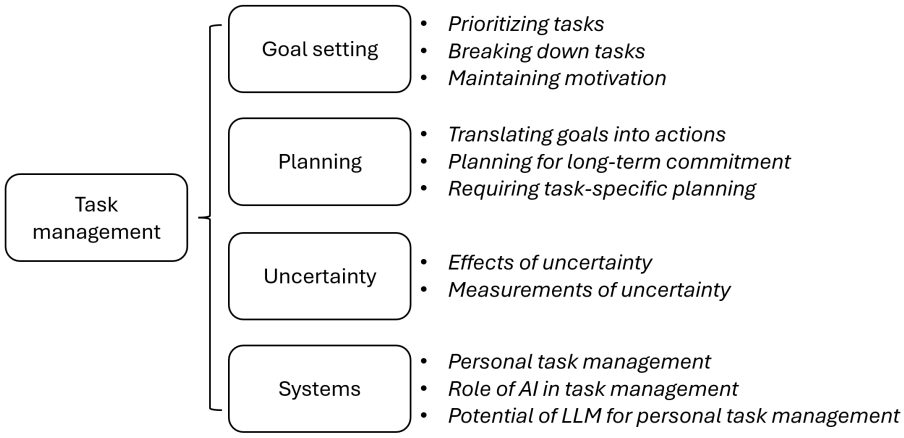

Figure 2.2: Key components and concepts in task management. It describes four components or aspects of task management, including Goal setting, Planning, Uncertainty, and Systems, and listing the specific sub-elements relevant to this study under each category.

it easier for individuals to stay motivated and engaged in their tasks. However, short-term goals do not inherently lead to the achievement of long-term objectives smoothly (Höchli et al., 2018). People may be satisfied with pursuing subgoals, but they may not recognize or be aware of how these tasks contribute to a broader long-term vision (Fishbach et al., 2006). This lack of awareness can create a disconnect, where the short-term goals are without consideration of how they fit into the long-term achievement.

Managing long-term goals presents challenges, such as maintaining motivation over extended periods, and dealing with uncertainty. There are external supports that can help people with long-term goals. Traditional methods like metacognitive strategies and life coaching have been used to improve long-term goal setting (Grant, 2003). In addition, AI tools and gamification enhance goal setting by reducing procrastination and aiding decision-making (Lieder et al., 2019). However, before individuals can effectively manage long-term goals, they first need to recognize that they can manage them. This includes developing awareness of their overarching goals, understanding

the challenges involved, and learning strategies to sustain long-term motivation and progress.

2.1.2 Planning

Planning is the cognitive process of generating and organizing a sequence of thoughts or actions required to achieve a specific goal (Miller et al., 1968). Planning is deeply rooted in the brain's executive functions, which involve higher-order cognitive processes distinct from routine task execution (Shallice, 1982). In personal development and life coaching, planning plays a pivotal role in translating goals into actionable strategies. As a tool for fostering clarity, motivation, and sustained action, planning not only helps people define specific objectives but also enables them to anticipate obstacles and formulate coping strategies (Neenan & Dryden, 2020).

David Allen's Getting Things Done (GTD) method has become a cornerstone of modern productivity systems. GTD emphasizes capturing all tasks and commitments in a trusted system, clarifying next actions, and regularly reviewing priorities to maintain focus and reduce stress (Allen, 2015). This method divides planning into actionable steps such as defining clear objectives, organizing tasks by context or priority, and reflecting on progress. It highlights the psychological relief provided by externalizing mental workload, allowing individuals to allocate cognitive resources more effectively.

Planning is not only essential for initiating behaviors but also for maintaining long-term commitments. Sniehotta et al. (2005) distinguished between action planning, which specifies how and when to act, and coping planning, which prepares for potential challenges. Both types of planning are crucial for achieving sustained lifestyle changes, such as adopting healthy behaviors. This dual-role of planning underscores its value as a proactive tool for bridging the intention-behavior gap and fostering resilience when facing obstacles.

The relationship between planning and performance is highly task-dependent. Eichmann et al. (2019)'s study on complex problem solving reveals that the timing, duration, and variability of planning yield varying effects depending on the task's complexity. For example, early planning is beneficial for simple tasks, while distributed and adaptive planning strategies are more effective for complex problems requiring exploration. These findings suggest that planning is not a one-size-fits-all process but must be tailored to the demands of the task. This adaptability is vital for fostering success in dynamic and uncertain environments.

2.1.3 Managing uncertainty

Uncertainty is a pervasive element of human experience, influencing decision-making and behavior across various contexts. Defined as the gap between what is known and what needs to be known to make informed decisions, uncertainty manifests in both quantifiable risks and unpredictable unknowns (Abbott, 2005; Hasani, 2018). Uncertainty differs from risk in critical ways. While risk pertains to situations where probabilities of outcomes can be calculated, uncertainty involves scenarios where such probabilities are unknown or incalculable. Abbott (2005) frames uncertainty as both an environmental factor and a byproduct of the planning process itself, as planners can unintentionally introduce new uncertainties with specific plans and dependencies.

2.1.3.1 Effects of uncertainty

In organizational contexts, uncertainty shapes task execution and decision-making processes. Maes et al. (2022) observed that uncertainty varies across project tasks, necessitating flexible, iterative execution strategies to adapt to changing conditions. Similarly, Savolainen (2012) linked uncertainty to information-seeking behaviors, emphasizing the role of uncertainty in triggering cognitive and affective responses that drive

decision-making. These findings suggest that incorporating uncertainty into project management frameworks can enhance organizational resilience and innovation. The psychological impact of uncertainty is profound, often manifesting as anxiety driven by an inability to predict or control future outcomes. Y. Gu et al. (2020) identified Intolerance of Uncertainty (IU) as a key mediator linking uncertainty to negative emotional states such as worry and inflexible responses. Miceli and Castelfranchi (2005) framed anxiety as an "epistemic emotion", rooted in the human need for certainty and control over potential threats.

Integrating insights from organizational and psychological research reveals a holistic approach to managing uncertainty. For organizations, uncertainty measurement framework can inform workplace design and employee support initiatives (Leach et al., 2013). On an individual level, interventions such as career planning and the use of IU assessment tools can mitigate anxiety and promote adaptability (Arbona et al., 2021; L. Chen & Zeng, 2021). Together, these strategies provide a roadmap for navigating uncertainty in diverse settings, balancing practical and emotional considerations.

Considering the role of planning in terms of managing uncertainty, planning reframes the overall task uncertainty as an information problem (i.e., epistemic uncertainty: uncertainty due to lack of knowledge, which can be reduced with more information) (Walker et al., 2003). Planning can inform people of what is known, what is needed, and what is expected (Christensen, 1985). By iterating and adjusting plans, people can navigate the task through uncertainty, which enables more informed decision-making.

2.1.3.2 Measurement of uncertainty

Measurement of uncertainty is critical for effective management. Leach et al. (2013) developed a multi-dimensional framework for assessing work uncertainty, encompassing

task, resource, and input/output dimensions. This tool allows organizations to evaluate how uncertainty interacts with job control, influencing outcomes like job satisfaction and performance. Maes et al. (2022) further distinguished Outcome Uncertainty (OU) and Action Uncertainty (AU) from the overall work uncertainty. OU refers to the ambiguity in the expected results of a task, AU relates to the uncertainty about the specific actions required to complete a task. There are several factors influencing the level of OU and AU. OU is primarily driven by information-related factors, such as the availability, completeness, and reliability of relevant data. AU is shaped by complexity-related factors, including the interdependencies between different actions, and the adaptability of processes to dynamic conditions. Additionally, contextual elements such as environmental volatility, resource capacity, individual expertise, and time constraints contribute to both OU and AU. Assessing these factors can help people perceive the level of uncertainty in a task and apply appropriate execution strategies to mitigate risks and enhance project outcomes.

From the perspective of psychology, Intolerance of Uncertainty (IU) was introduced to measure individuals' emotional and cognitive responses to uncertainty (Carleton et al., 2007). IU is a key psychological factor influencing worry, anxiety, and avoidance behaviors. It is driven by cognitive, emotional, behavioral, social, and environmental factors. Cognitive factors include a tendency to catastrophize uncertainty, and viewing ambiguous situations as inherently threatening. Emotional factors involve stress, frustration, or discomfort when faced with the unknown. Behavioral factors manifest in avoidance of uncertain situations, excessive information-seeking, or an inability to take decisive action in ambiguous situations. Additionally, social and environmental factors, such as past experiences with unpredictable events or external pressures for certainty,

can shape an individual's level of IU. By assessing these factors, researchers and practitioners can better understand how IU contributes to anxiety and decision-making difficulties, helping to develop interventions to manage uncertainty more effectively.

2.1.4 Tools and systems supporting task management

Task management systems are tools designed to effectively plan, organize, and complete tasks. While task management can be approached at both organizational and individual levels, this study focuses on systems and research dedicated to personal task management. Personal Task Management (PTM) is a focused subset of task management that centers specifically on managing an individual's personal tasks, time, and priorities (Allen, 2015; Haraty et al., 2016; Jones et al., 2006). Complementing PTM, Personal Information Management (PIM) focuses on organizing and maintaining digital and physical information resources, which can provide the contextual information necessary for effective task management (Jones, 2010; Jones et al., 2017). Together, PTM and PIM form a cohesive framework that supports individual productivity by enabling users to manage both their tasks and the information needed to complete them.

The field of personal task management has seen extensive research and development, resulting in numerous systems aimed at improving individual efficiency. Bellotti et al. (2004) provided the foundational work with valuable insights into how individuals naturally manage tasks and the role of external representations such as to-do lists. Contrary to the assumption that people are poor at prioritizing, the study found that users employ sophisticated strategies to organize tasks based on urgency and importance. The researchers emphasized the importance of adaptable tools that align with user preferences, suggesting that effective task management systems must integrate features for dynamic prioritization and context-aware organization.

A. Hassan and White (2012) proposed the concept of "task tours," designed to guide users through multi-step, complex search tasks such as planning a vacation or researching a topic. By identifying the necessary subtasks and presenting them as a structured pathway, task tours help users comprehend and execute their objectives more efficiently. This system demonstrates the value of providing holistic support in complex scenarios where users may not be fully aware of all required steps. Expanding on this, Hassan Awadallah et al. (2014) developed a method for identifying and recommending subtasks using query logs from search engines. Their approach constructs a graph of related tasks, which aids users in exploring and tackling multi-step tasks. This graph-based system leverages collective user behavior to suggest relevant subtasks, offering a scalable solution for supporting exploratory and goal-driven search activities.

Prediction and early warning systems have emerged as essential tools in task management (White & Hassan Awadallah, 2019; White et al., 2019). Takeuchi et al. (2013) leveraged lifelog data to develop a task management system that predicts future workload using simple regression models. The system encourages users to alter their daily activities based on anticipated task demands, promoting better preparedness and reducing the risk of accumulating unmanageable workloads. Similarly, Mita et al. (2017) introduced a system that predicts task failure based on user progress logs and cognitive biases like the planning fallacy. By providing early warnings about potential failures, the system helps users adjust their strategies and improve time management, achieving a prediction accuracy of over 90% in certain scenarios.

Proactive systems highlight the potential of AI in reducing cognitive load for task management. Yorke-Smith et al. (2012) developed a system to anticipate user needs

and take initiative to perform tasks autonomously. This proactive behavior is particularly effective in environments with high cognitive demands, allowing users to focus on complex, strategic decisions while delegating routine activities. Myers et al. (2007) introduced a project assistant, which integrates AI-driven planning and execution monitoring to streamline time and task management. This approach mirrors human assistants, enhancing productivity by mitigating task overload and improving decision-making processes.

While the systems above focus on task and information management, the related field of Personal Informatics (PI) offers another important perspective. Rooted in HCI, PI research focuses on systems that help individuals collect, visualize, and reflect on personal data (I. Li et al., 2010). This domain typically involves tracking personal information such as physiological states, daily activities, moods, or environmental factors (Kersten-van Dijk et al., 2017). The primary goal of PI systems is to foster self-awareness and support behavior change or decision-making through data-driven reflection on past and current states (Epstein et al., 2015; Rapp & Tirassa, 2017). While this provides an invaluable foundation for self-understanding, its application has predominantly been retrospective (helping users understand what they have done) or focused on in-the-moment feedback (Rapp & Cena, 2016). While PI tools excel at helping users "quantify" their lives, they often stop short of providing a comprehensive system to help them "plan" their futures, especially in navigating the uncertainty inherent in complex goals.

The advent of AI and LLMs has transformed traditional task management systems by introducing dynamic, context-aware, and highly personalized solutions (De Zarzà et al., 2023). Unlike static or rule-based systems, AI-driven tools leverage vast datasets and machine learning to optimize task allocation, prioritize activities, and provide real-time insights. Barcaui and Monat (2023) compared the efficacy of generative AI,

such as GPT-4, with human project managers in planning tasks, revealing complementary strengths. While AI excels in automating repetitive and analytical components, humans bring creativity and critical decision-making to the table, suggesting that a hybrid approach holds significant promise.

LLMs offer unprecedented capabilities in understanding and managing complex tasks. These AI tools can act as copilots in search tasks, breaking down complex queries into manageable subtasks and assisting users in achieving their goals efficiently (White, 2023). However, reliance on AI systems is not without challenges. Salimzadeh et al. (2024) found that users faced with complex and uncertain tasks often over-relied on AI, leading to errors in decision-making. Addressing these challenges requires designing systems that not only enhance task completion but also foster appropriate trust and reliance between humans and AI.

2.2 Interactive information retrieval for supporting users in tasks

2.2.1 User task experience in IIR research

Interactive Information Retrieval (IIR) research focuses on understanding user behavior and experience interacting with search systems and evaluating system or interface features influencing user search tasks (J. Liu & Shah, 2019a). Different from system-centric evaluations and relevance-based metrics, evaluation in IIR requires examining not only the retrieved results but also how users interact with the system over multiple queries, considering user behaviors, cognitive processes, and system features in completing search tasks (Petrelli, 2008).

Some common evaluation metrics include relevance, usefulness, and satisfaction. Relevance is a system-centric metric typically assessed by external judges, focusing on whether retrieved documents match a given query in a single-query session (Zhai et al., 2015). However, IIR involves multi-query interactions, requiring more user-centric metrics (Kelly, 2007). Usefulness refers to how well the retrieved information contributes to accomplishing a search task. It is usually evaluated at the task level from the perspective of the user (Cole et al., 2009). Satisfaction, on the other hand, captures the user experience in the search task, reflecting subjective perceptions of search success, ease of use, and effort during the process (Al-Maskari & Sanderson, 2010). Beyond satisfaction and usefulness, previous studies have investigated various user-centric metrics based on their experience or perceptions, such as frustration, expectation, perceived task complexity, and perceived time (Feild et al., 2010; J. Liu & Shah, 2019b; C. Luo et al., 2017; B. Wang & Liu, 2023). These additional metrics help researchers develop more holistic evaluations of user experiences in IIR.

User experience in IIR is related to information needs, search behaviors, and task performance (J. Liu et al., 2019; Marchionini, 2003). Since searchers' mental states are not directly observable, IIR studies rely on explicit user interaction data (e.g., query reformulations, clicks, dwell time) and self-reported experience annotations (e.g., satisfaction) to infer implicit cognitive and emotional states (C. Liu et al., 2021). By modeling search behavior and experience, researchers aim to improve system adaptivity and provide personalized support to enhance task success.

Recent research suggests parallels between IIR and human interactions with LLM-based chatbots (Skjuve et al., 2023). Like IIR users, LLM users engage in multi-turn interactions, iteratively refining their prompts to achieve better responses (B. Wang et al., 2024). This similarity suggests that methodologies from IIR research could inform studies on user experiences with LLMs (Shah & Bender, 2024). Leveraging insights

from IIR, it is promising to improve human-AI interaction frameworks that support users in work tasks beyond search, improving their overall experience and task success.

2.2.2 Uncertainty in IR studies

Uncertainty is a fundamental concept in IR, playing a pivotal role in shaping users' information-seeking behaviors and decision-making processes. It is constructed through cognitive, affective, and situational dimensions, reflecting a complex interplay of internal and external factors. Cognitively, uncertainty often arises from knowledge gaps or ambiguous information needs. Belkin (1980)'s Anomalous State of Knowledge (ASK) model highlights this phenomenon, describing how users experience uncertainty when they are unable to articulate their information needs effectively, which hampers query formulation and retrieval performance. This cognitive basis of uncertainty underscores the challenges users face in accessing relevant information in complex search environments (Chowdhury et al., 2011, 2014).

The affective component of uncertainty reveals how emotional states interact with cognitive challenges during information retrieval. The ISP framework emphasizes that uncertainty is most pronounced in the early stages of the search process, as users grapple with incomplete or unclear information (Kuhlthau, 1991). While uncertainty tends to diminish as users gather more information and refine their search strategies, some forms of uncertainty persist, leading to anxiety, frustration, or even curiosity. Keller et al. (2020) explains that uncertainty, rather than being solely negative, can motivate exploration and learning, driving users to engage more actively with information sources. This dual nature of uncertainty, as both a challenge and an opportunity, adds complexity to its role in IR (Kuhlthau, 1991).

Uncertainty is not only a cognitive or emotional state but also a perception shaped by the user's interaction with information systems and tasks. Chowdhury et al. (2014)

argue that uncertainty persists throughout the IR process, fluctuating as users encounter new challenges or incomplete solutions. This persistence may further trigger biases such as confirmation bias, where searchers favor information that aligns with their pre-existing beliefs, or availability bias, where they rely on easily accessible information rather than thoroughly evaluating sources (Azzopardi, 2021; J. Liu & Azzopardi, 2024). Moreover, uncertainty can impair cognitive load management, making it harder to filter relevant from irrelevant information, especially in high-stakes or complex domains, and lead to suboptimal decision-making (Adewale et al., 2017; M. M. Luo & Nahl, 2019; Mestre, 2024). The relationships between cognitive biases and uncertainty highlight the importance of designing information systems that guide users to make informed decisions while helping them manage uncertainty effectively.

2.2.3 Improving user performance and experience in online search and recommendation

Users often face significant challenges in achieving high task performance. The inherent uncertainty in tasks can lead to anxiety and hinder effective information seeking, as users may be unsure about where to begin or what information is needed in the early stage of tasks (Kuhlthau, 1991; Vakkari, 2003). Improving user search performance and experience with search system support is fundamental in information retrieval (Huurdeman & Kamps, 2015). User search performance is shaped by many factors, such as domain knowledge (Mao et al., 2018; Wildemuth, 2004), task complexity (J. Liu & Shah, 2019b; P. Liu & Li, 2012), search skills, and search experience (Bailey, 2017; Smith, 2017). Users from various domains with different levels of search experience will perform differently on specific search tasks (Moraveji et al., 2011). Evaluating user search performance usually involves assessing their success in the search task, which

can be done through self-reports or analyzing user responses (J. Liu & Shah, 2019b; Odijk et al., 2015). However, collecting user-generated annotations can be costly, so there is a need for metrics that evaluate user performance with limited user annotations and implicit feedback. Recent research has focused on online metrics for information retrieval systems that combine offline metrics (e.g., relevance-based metrics) and online user behaviors (Arguello, 2014). These metrics can reflect user search behaviors and performance from multidimensional aspects (Hendahewa & Shah, 2017; J. Liu & Yu, 2021).

Proactive information retrieval aims to improve users' search experience and performance from the traditional search system with a reactive nature (Bhatia et al., 2016). Users with different levels of search experience, domain knowledge, or information literacy can all get support from proactive search systems (Huurdeman & Kamps, 2015; Savenkov & Agichtein, 2014; Smith, 2017). Query recommendation is a popular proactive information retrieval method that has demonstrated usefulness for users engaging in various types of tasks (Hanauer et al., 2017; Hendahewa & Shah, 2015, 2017; Zahera et al., 2013). High-quality, collaboratively recommended queries created by expert users can provide guidance to users and improve their search performance (Harvey et al., 2015; Morris, 2008; Morris & Horvitz, 2007; Savenkov & Agichtein, 2014; Smith, 2017). Another aspect of assisting users is helping the search system adapt to users' states according to their behaviors (J. Liu & Yu, 2021; Rha et al., 2016). The adaptive search system can improve search results during the dynamic interaction between users and the search engine (J. Luo et al., 2014). When users are struggling with search tasks, these proactive methods are supposed to assist users in accomplishing tasks with positive interventions. Therefore, by accurately estimating users' search performance, queries from high-performance users can be collected to help other users in the proactive information system.

Conversational search and recommendation systems enhance the user experience by enabling natural interactions that engage users and allow the system to learn from interaction history (Trippas, 2024). Through conversational exchanges, the system can ask clarifying questions to gather more information about the user's preferences and background, reducing ambiguity in user intent (Bi et al., 2021; Kiesel et al., 2024). Additionally, by leveraging rich interaction history, the system can predict the next recommended item more accurately to improve overall user satisfaction (Siro et al., 2024).

In research on conversational search and recommendation, user states (e.g., intent, satisfaction) are typically represented as labels in preexisting datasets (Farshidi et al., 2024). Deep learning and NLP techniques are employed to simulate the conversational process and train models to improve the prediction of user states (S. Zhang & Balog, 2020). Before the advent of LLMs, these studies primarily focused on model architecture and NLP techniques for enhancing model performance, with task topics constrained by the available datasets (Jannach et al., 2022; Sun & Zhang, 2018). LLMs empower conversational search and recommendation systems by enabling more flexible task expansion to general topics and supporting longer, more complex interactions (Lin et al., 2025).

2.3 Potential and challenges of LLM for the long-term life task planning

2.3.1 Use LLMs for information access

LLMs present opportunities to mitigate challenges in IIR. LLMs enable conversational context through multi-turn, dialogue-based interactions that integrate previous exchanges, making it easier to handle complex queries and refine search intents (Shah et al., 2023). By leveraging prompt engineering, which encapsulates task context and user perceptions into structured prompts, LLMs can generate more relevant and contextualized responses (P. Liu et al., 2023). Unlike static IR systems, LLMs offer dynamic interaction, providing real-time adaptability by refining responses based on iterative user inputs and their evolving needs (L. Liao et al., 2023). Incorporating LLMs in information access systems introduces transformative prospects, particularly through multi-turn interactions that resemble traditional IIR processes (Shah & Bender, 2024; Skjuve et al., 2023). However, harnessing LLMs in information access systems presents pronounced challenges. Users encounter difficulties in query formulation or interpreting search results, often due to cognitive barriers, which are common when initializing and refining prompts during interactions with LLMs (Savolainen, 2015; Zamfirescu-Pereira et al., 2023). These barriers often stem from task context and user states, such as lack of prior knowledge, low familiarity level with the topic, high complexity, and inappropriate expectations, leading to potential misconceptions about how LLMs interpret and respond (Ford, 1995; A. Hassan et al., 2014; Odijk et al., 2015; Savolainen, 2015; B. Wang & Liu, 2022, 2023).

2.3.2 LLM agent and planning

The concept of an LLM agent denotes an entity capable of performing intentional actions and encompassing autonomy, reactivity, proactiveness, and social ability (Park et al., 2023; Schlosser, 2015). LLM agents are designed to facilitate everyday tasks and improve task efficiency while reducing user workloads (G. Li et al., 2023; Xi et al., 2023). The inherent ability of LLM agents to understand, decompose, plan, and subsequently complete tasks is fundamentally attributed to their emergent reasoning and planning capacities, exemplified through strategies like Chain-of-Thought (CoT) and problem decomposition (Wei et al., 2022; Yao et al., 2022). Task planning and automation emerge as key functionalities of LLM agents. They can decompose complex tasks into manageable sub-tasks, plan sequences of actions, and interactively explore environments to achieve objectives (Huang et al., 2024; G. Li et al., 2023). The sophisticated integration of LLM capabilities with task-specific strategies facilitates effective task completion and automation, thereby offering practical value in enhancing efficiency and accessibility for users (Wu et al., 2023). Existing work demonstrates that LLMs can assist with long-term planning in simulated environments, such as travel planning (J. Xie et al., 2024), research assistance (Schmidgall et al., 2025; Tang et al., 2025), team collaboration (F. F. Xu et al., 2025), and even social interaction (Piao et al., 2025), as well as real-world applications for programming, writing, and education (Feng et al., 2024; Urban et al., 2024; B. Wang et al., 2024). Yet most remain domain-specific, short-term, or lack insight into how real users collaborate with LLMs in personal, open-ended, and uncertain long-term life tasks.

2.3.3 LLM benchmark and LLM-as-Judge

Evaluating the performance of LLMs critically relies on the use of benchmarks (Chang et al., 2024). A benchmark is a dataset with problems and solutions, designed to test and measure the capabilities of LLMs across various tasks and domains. Some common benchmarks are employed to assess LLMs' proficiency in language understanding (Hendrycks et al., 2021), common knowledge reasoning (X. Li et al., 2023; Zellers et al., 2019), code generation (M. Chen et al., 2021), and solving math problems (Clark et al., 2018). The importance of these benchmarks lies in their ability to consistently evaluate LLM performance, allowing for comparisons across different models. They enable the assessment of performance on specific topics, ensure that evaluations are reproducible, and ultimately facilitate the development and improvement of LLMs. For tasks related to planning, there are also specialized benchmarks that test LLMs' abilities in generating and applying plans for problem-solving. The TravelPlanner benchmark assesses the model's ability to organize and optimize travel plans based on user preferences and constraints (J. Xie et al., 2024). Additionally, LLMs can generate plans to use tools and solve complex problems in the benchmark designed for evaluating tool usage planning (Kong et al., 2023; Ruan et al., 2023). Furthermore, LLM's planning ability can be assessed based on general benchmarks (e.g., reasoning and math), and the results demonstrate that planning can improve the model's inherent problem-solving capabilities. Such benchmarks are crucial for advancing our understanding of LLMs' planning skills and their application in real-world tasks.

Developing the benchmark process includes obtaining accurate benchmark labels (e.g., the correct answers to problems). The concept of LLM-as-Judge involves using LLMs to replace human annotators in generating these labels, thereby automating the evaluation process (L. Zheng et al., 2023). This approach is significant because it can greatly enhance efficiency, reduce human bias, and allow for scalable assessments

of model performance across vast datasets. LLM-as-Judge can be applied in various areas, such as medical, financial, legal, educational, and other general areas (H. Li et al., 2024). To build a reliable LLM-as-Judge process, it is essential to select an advanced model, iteratively design clear evaluation instructions, and validate the alignment with human judgments (J. Gu et al., 2024). Analyzing and improving the agreement with human judgments can help refine the model's accuracy, ensuring that the LLM can effectively and consistently replicate human-level evaluations in specific task contexts (Thakur et al., 2025).

2.3.4 Toward user-centric LLM

The development of human-LLM interaction holds significant potential in open-domain tasks (Roller et al., 2021). Advancements in LLMs have broadened their application from NLP tasks to more general and complex human tasks (Qin et al., 2023). Current research often simulates multiple task domains, emphasizing automation with minimal human involvement. Human-centered LLMs require more human control and participation (Gomez et al., 2025; Tu et al., 2024). Simulation is also a widely used method in IIR and recommendation research, allowing researchers to model how a user interacts with a search engine or recommender system when completing a task (D. Maxwell & Azzopardi, 2016; S. Zhang & Balog, 2020). By simulating user interactions, researchers can evaluate system effectiveness based on how well it assists the simulated user in achieving their goal (Y. Zhang et al., 2017). Simulation studies offer several benefits, including controlled experimentation, scalability, and the ability to test system performance across various scenarios without requiring extensive real-user participation (D. M. Maxwell, 2019).

Research in IIR and LLMs areas both highlight the importance of user experience, evidenced by users' mixed feelings of fascination and cognitive load interacting with

LLMs, reflective of the serendipity and information overload in IIR (A. Hassan et al., 2014; Odijk et al., 2015). Personalization in both domains enhances user interaction dynamics, especially during query/prompt initiation and refinement. There are also similar issues considering user satisfaction especially when users may end conversations or search tasks with satisfied outcomes instead of fully completion (Urban et al., 2024). Developing users' critical thinking and extending their interactions with LLM by cross-checking the output might be necessary. Additionally, as LLMs are increasingly being leveraged for educational purposes, assessing their efficacy in terms of knowledge dissemination and learning facilitation has been suggested as a valuable metric for evaluation (Gibson et al., 2023).

Recently, the evaluation of human-LLM interaction has garnered significant attention, especially as these models become increasingly used in human tasks. A comprehensive approach encompasses aspects such as task performance, user experience, and general guidelines of human-AI interactions (Amershi et al., 2019; Lee et al., 2023). With similar interaction processes and challenges, there are also notable parallels between the evaluation of human-LLM interaction and the assessment of IIR. Both areas highlight the significance of user experience and perception. Another critical facet enhancing user trust and comprehension is explainability in AI, which merits more profound exploration (Kim et al., 2023; F. Xu et al., 2019). While these studies have investigated the user experience in human-LLM interactions, none have thoroughly examined these interactions in long-term life tasks. Thus, it is critical to design a collaborative framework for managing long-term life tasks and consider the user's role within the framework.

2.4 Summary

The literature review chapter situates this study within the broader domain of task management, IIR, and human-LLM interactions. examines how previous research in these areas relates to the newly proposed task type in this study, highlighting overlaps and connections among these areas. It informs the design and scope of this study. The foundation is established to understand the challenges and opportunities of combining task management, interactive search and recommendation, and LLM-based solutions for long-term planning and decision-making.

Chapter 3

Methodology

The primary objective of this dissertation is to investigate how HAC can support long-term life task planning and mitigate the associated uncertainty. While LLMs show promise in general problem-solving, their application in supporting the complex, iterative nature of long-term life tasks remains unexplored. To address this gap, the methodology proceeds in three logical stages: first, understanding real users' experiences and coping strategies; second, designing a framework to simulate this collaboration; and third, evaluating LLM performance on these specific tasks. The methodology is divided into three parts corresponding to the research questions: for RQ1, I recruited participants to use ChatGPT to simulate AI-assisted planning and provide feedback on AI-assisted task planning. For RQ2, I implemented the GOLF framework, a multi-agent system leveraging LLMs to simulate human-AI collaboration for long-term life task planning. I conducted experiments and metrics to evaluate the framework's effectiveness in task completion and uncertainty metrics. For RQ3, I developed the GOLF benchmark by generating long-term life task simulations in different topics and user states to evaluate and compare LLM performance.

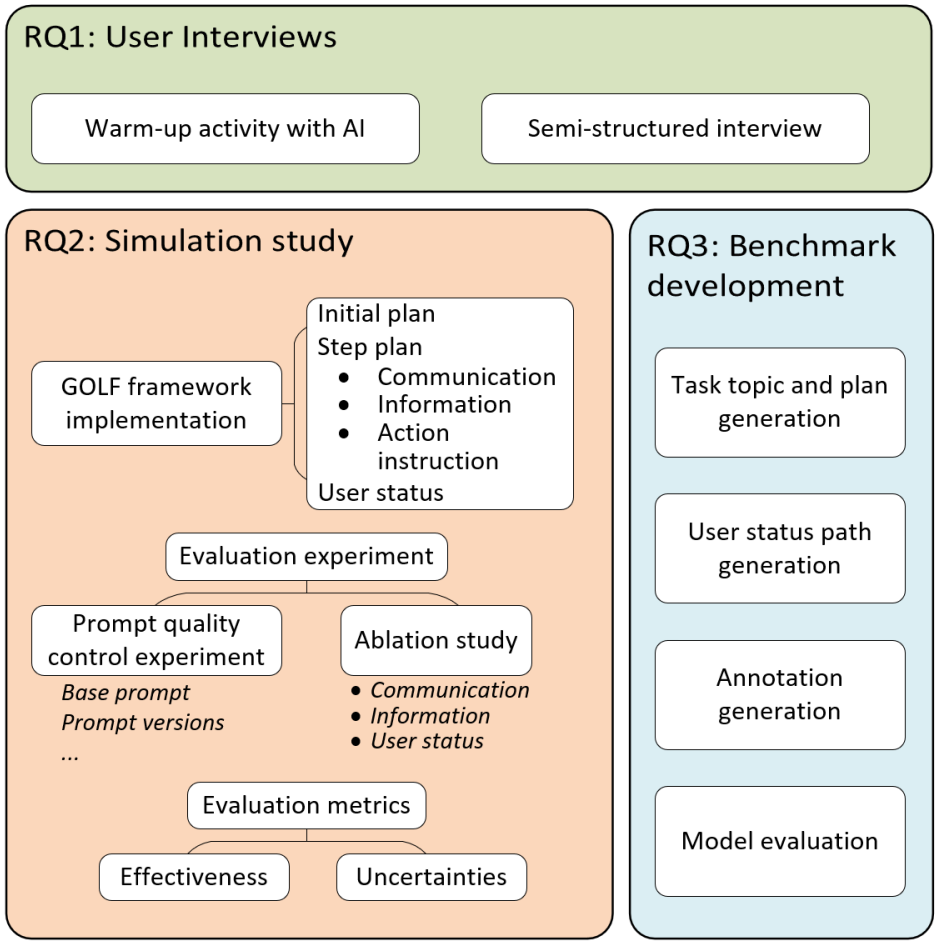

Figure 3.1: Overview of the methodology. It maps the three research questions (RQ1, RQ2, RQ3) to their corresponding methodological phases: the user interview process, the simulation study utilizing the GOLF framework, and the benchmark development.

3.1 An interview study on how people use generative AI for long-term life tasks

This study employs a qualitative research design to explore how AI users cope with uncertainty in long-term life task planning. The research consisted of two main components: a warm-up study and an in-depth semi-structured interview. The warm-up activity provided participants with a hands-on experience in AI-assisted task planning.

Using the long-term goal defined during the warm-up as a reference, the interview gathered participants' feedback on their experience and and perspectives on human–AI collaboration in long-term task planning.

3.1.1 Participant recruitment

As an exploratory study, the researcher conducted in-depth interviews with a small number of participants to examine their contexts, motivations, and challenges in AI-assisted long-term planning. This study recruited 14 participants (aged 20 to over 80) with diverse educational and professional backgrounds, including students, engineers, managers, educators, and researchers. Seven participants were recruited via information schools' mailing lists and professional networks, and the remaining through the social media platform Reddit. Most participants reported using ChatGPT several times a day for both professional (e.g., study or work-related tasks) and personal purposes (e.g., health, daily schedule, or learning). Each study session (including the warm-up and interview) lasted around two hours. Each participant received $30 USD as compensation. The study was approved by the University of Oklahoma's Institutional Review Board (IRB# 16081).

3.1.2 Warm-up activity and interview

Participants first engaged in a warm-up activity designed to simulate AI-assisted long-term task planning using ChatGPT. I, as the researcher, collected demographic information and asking about participants' previous experiences of using ChatGPT or ideas for tasks where it might assist planning. The selection of the planning topic was a collaborative process to ensure relevance to the study's focus on long-term life tasks. This

selection involved a brief discussion regarding the participant's prior history with ChatGPT or tools. If the participant had previously consulted ChatGPT on a life-related topic requiring long-term planning, the researcher encouraged them to continue exploring that topic to capture ongoing user needs. In cases where participants had no prior relevant history, the researcher guided them to brainstorm new topics of personal interest that fit the criteria of a long-term life task (i.e., a task requiring extended time and planning to complete). The researcher's role was primarily to screen these topics to ensure they met the requirements of the study, without imposing specific domains. Examples of goals include pursuing higher education, starting a career, or adopting a healthier lifestyle. The researcher did not measure or limit the topic complexity and timeframe, which allowed exploration of diverse strategies and challenges.

When interacting with ChatGPT, Participants could choose to use their own prompts or templates provided by the researcher. The prompt templates included three versions, each reflecting a step in generating and refining task plans. The first template generally included the participant's long-term goal and current state: "*I am [state], can you provide me with a [topic] plan.*" The second template asked for details about a specific part of the plan generated in the first prompt: "*For [a specific part], can you provide me with a [step-by-step/weekly] plan.*" The third template added more contextual information and asked ChatGPT to revise the plan: "*For [a specific part], I have [preference/consideration], can you revise the plan.*" Participants using the provided templates received an initial task planning outline from ChatGPT and discussed any immediate impressions. They could then provide additional details about their situation or preferences to generate or revise content in the second and third prompts. Participants using their own prompts engaged with ChatGPT more flexibly, deciding what aspects of planning to explore. The researcher occasionally offered prompt templates if participants needed guidance. Participants used a free ChatGPT account

provided by the researcher. The system's Memory function, which allows the model to retain user details and preferences across different conversation sessions, was explicitly disabled. Since a single lab ChatGPT account was shared among all participants, this measure was necessary to prevent cross-contamination, ensuring that information provided by one participant would not be recalled by the model to influence subsequent sessions.

Throughout this process, participants were encouraged to think aloud and discuss the usefulness of the AI-generated task decomposition. They reflected on how this AI-assisted planning compared to their usual non-AI methods, noting any insights or discrepancies. During the warm-up activity, the researcher asked participants interview questions about their experience after reviewing each ChatGPT response. As shown in Table 3.1, following the initial question on task topic selection, the interview consisted of four main sections exploring participants' experiences with AI-assisted long-term task planning. Part 1 examined participants' first impressions of ChatGPT's responses; Part 2 focused on experiences of uncertainty during planning; Part 3 addressed usefulness and limitations on specific task components; and Part 4 captured broader reflections, expectations, and suggestions for improvement.

These interview questions were adapted to each participant's task topic and background to ensure relevance. For example, if a participant expresses interest in film making, an example question can be tailored to: *How do you think AI can help you learn film making or create a structured plan for learning?* This approach makes the questions more concrete and personally meaningful. Additionally, technical concepts were adjusted based on participants' background knowledge. For instance, instead of referring to "model version or size," the question can be framed in terms of "free or paid ChatGPT account" to make it more accessible to participants. Such phrasing helped maintain clarity and engagement.

For questions in part 2, the interviewer first introduced the concept of task uncertainty, explaining its two example forms: *outcome uncertainty*, where individuals are unclear about what they want to achieve, and *action uncertainty*, where they are unsure about the steps needed to reach their goals. Participants were encouraged to share their opinions on how uncertainty impacts long-term life task planning. The discussion would highlight that while uncertainty can motivate exploration and creativity, it can also increase anxiety. This section thus examined how different uncertainties emerge and how users perceive and manage them.

To capture participants' immediate reactions and ensure feedback was contextually relevant, interview questions were flexibly integrated into the warm-up activity rather than restricted to a post-session format. After the activity, the interview continued to cover any remaining questions not previously discussed. This was common for questions in parts 3 and 4 (task planning details and other aspects), where participants discussed retrospective perceptions and high-level preferences, such as "Which part of the plan do you find most helpful?" and "How would you like ChatGPT to explain its reasoning?"

3.1.3 Interview data analysis

The researcher first organized participants' background information and prompt behaviors in tabular form to support cross-case comparison. The researcher manually tagged prompts according to their functional roles (e.g., goal setting, task decomposition, refinement, contextualization) and mapped them to different phases of task planning. This facilitated the identification of prompting patterns and engagement styles.

To account for differences in prompting behavior, the study has different foci between template-based and self-generated prompts. For participants who mainly used templates, the study examined how these prompts served as starting points to explore

AI-assisted planning and stimulate reflection. For those who wrote prompts independently, the study focused on their natural prompting patterns to understand how users might organically engage with ChatGPT in real-life planning scenarios.

The interview data were analyzed using a reflexive thematic analysis approach (Braun & Clarke, 2006, 2019), combining deductive coding guided by the research questions with inductive identification of emergent themes. The analysis focused on participants' goals, strategies, uncertainties, and perceptions of AI-assisted long-term task planning.

For the interview transcripts, the lead researcher followed six phases of reflexive thematic analysis: (1) familiarization through repeated reading of transcripts; (2) identifying features related to the research questions; (3) generating initial codes both deductively and inductively; (4) organizing related codes into broader themes; (5) reviewing and refining these themes; and (6) defining and naming them, supported by representative quotes. While deductive codes reflected the study's analytical focus (e.g., task types, planning strategies, uncertainty types, perceptions of AI assistance), inductive coding captured unexpected concerns or novel strategies.

To enhance rigor, the researcher iteratively refined the codes and themes through repeated engagement with the data and incorporated feedback from the dissertation advisor, as the second researcher. The second researcher reviewed the coding framework and emergent themes, offering an external perspective for validation. In line with reflexive thematic analysis, this study did not calculate inter-coder reliability metrics, as themes were treated as interpretive patterns rather than objective entities (Braun & Clarke, 2006, 2019). Instead, analytic consistency was ensured through systematic engagement with the data, iterative refinement, and transparent reporting by providing representative quotes for each theme in the findings. Quotes were selected to illustrate

both typical and divergent cases, ensuring that the reported themes reflected the range of participant experiences.

Table 3.1: Interview questions. It lists the semi-structured interview questions categorized by four distinct parts (Initial impression, Uncertainty, Details, and Other aspects), designed to elicit user feedback on AI-assisted planning.

Part 0 Task planning topic
•*Can you think of a specific task or goal in your life where AI could help you plan or organize things over a long period?*
Part 1 Initial impression
•*How do you think the plan is useful?* •*How do you feel ChatGPT understands your background when helping you with this task?* •*If you weren't using ChatGPT, how would you usually plan for this kind of task? How does that compare with using ChatGPT?* •*(If using the provided prompt templates) How do you feel about the way ChatGPT breaks down your task? What do you like or dislike?*
Part 2 Uncertainty in task planning
•*Which part of the AI-assisted plan made you feel uncertain? How did that uncertainty affect your emotions or behavior?* •*How do you think AI is able to help reduce uncertainty or made it easier to deal with it in this planning process?* •*What kinds of uncertainty might you face if you were using non-AI methods instead?*
Part 3 Task planning details
•*Which part of the plan do you find most helpful?* •*Which part do you think ChatGPT struggles with when helping with your task planning?* •*Do you have any suggestions for how ChatGPT could improve in any specific part of the planning process?*
Part 4 Other aspects
•*Do you have a paid ChatGPT account? How do you think the paid version of ChatGPT performs better for this kind of task planning?* •*Would you prefer ChatGPT to give you multiple solutions to a problem, or just one clear path? Why?* •*How would you like ChatGPT to explain its recommendations or reasoning to you?* •*Do you think AI-assisted planning affects your ability to make your own decisions in any way?* •*What other changes or features would make AI-assisted task planning more useful or effective for you?*

3.2 Human-AI collaboration simulation framework

This section introduces the GOLF framework, explaining its components and the evaluation process. It demonstrates the framework's working process through a planning example. Following this, it conduct large-scale experiments to explore how variations in prompts and components influence the framework's outcome. The experiments aim to assess the effectiveness of the GOLF framework and its evaluation without any training or modification to the LLM base model. This section anticipates addressing RQ2 by validating the impact of task decomposition in long-term task planning via simulation based on the GOLF framework and evaluation. Task topics include two representative domains: house purchasing (financial) and obesity management (health).

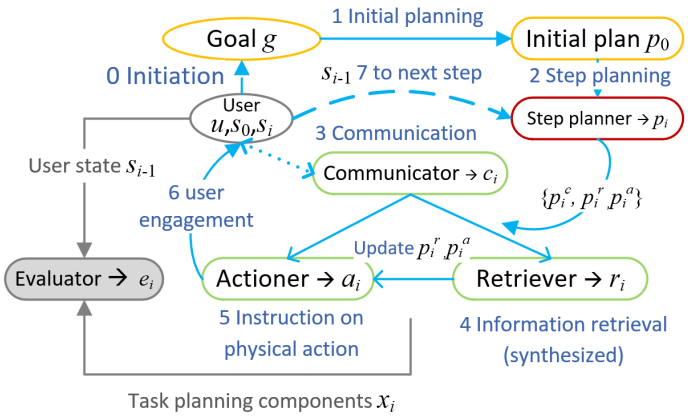

Figure 3.2: The task planning process at the ith step in the GOLF framework. It illustrates the cyclical workflow of the GOLF framework, showing how the User agent and System agents interact through specific steps to generate and evaluate plans iteratively.

Figure 3.2 illustrates the task process within the GOLF framework, which operates on a multi-agent system designed to facilitate user support and workload distribution for achieving long-term goals. This framework simulates the user-system interaction process in completing a life task. The user is simulated by an LLM agent. The

system involves multiple LLM agents, including an initial planner, a step planner, task-specific (communication, information retrieval, and action instruction) agents, and an evaluator. Specifically, the framework utilizes the following notations to represent the state and outputs:

- g: A fixed, short statement defining the long-term goal (e.g., "I want to buy a house").
- s_i: The user state at iteration i. s_0 denotes the preset initial statement (e.g., "I am not familiar with the task and want to start from the beginning").
- $\mathcal{P}_{\dashv}\}\rceil\backslash\sqcup$: the system prompts for the LLM agents
- p_0: The initial plan with general activities and tasks.
- p_i: The detailed step plan generated by the planner at iteration i, which is decomposed into subtasks $\{p_i^c, p_i^r, p_i^a\}$.
- c_i, r_i, a_i: The specific outputs generated from the communication, information retrieval, action instruction components respectively.

The agents process the task simulation as follows:

0. Initiation: Set goal g, user state s_0. This setup allows for controlled variation in user preferences and task topics.
1. Initial planning: An LLM agent generates an abstract initial plan p_0 based on a prompt $\mathcal{P}_0$, the brief goal statement g and the user's initial state s_0. The initial plan outlines broad activities, tasks, and subtasks toward the goal:

$$p_0 = \text{LLM}\left(\mathcal{P}_0, g, s_0\right), \tag{3.1}$$

where LLM represents a single LLM instance, which is only created once with the prompt $\mathcal{P}_0$ and does not engage in subsequent iterations. The initial plan p_0 provides a high-level roadmap and is generated only once at the beginning of the process.

2. Step planning: Using p_0 and the user's progress $s_{i-1}, i = 1, \ldots, n, n \in \mathbb{Z}$, the step planner generates a step plan p_i which expand the task outline with more detailed subtasks:

$$p_i = \textbf{Planner}\left(g, p_0, s_{i-1}\right). \tag{3.2}$$

Unlike the standalone LLM, **AgentName** (e.g., **Planner**) represents an LLM-based agent configured with a specific prompt $\mathcal{P}_{agent}$ and participates in each iteration of the step planning process.

The step plan $p_i = \{p_i^c, p_i^r, p_i^a\}$ contains three subtasks, including communication p_i^c, information retrieval p_i^r, and action instruction p_i^a tasks, each assigned to corresponding LLM agents for execution. This task decomposition and assignment follow the empirical task management strategies (Allen, 2015).

Subtask execution and plan update:

3. Communication with the user: The communicator agent chats with the user to collect personal information c_i (e.g., preferences and considerations) according to p_i^c:

$$c_i = \textbf{Communicator}\left(p_i^c | \textbf{User}\right), \tag{3.3}$$

where the user is simulated by an LLM agent **User** with a prompt $\mathcal{P}_u$ that includes pre-defined background and task-specific preferences. Simulating the user is a common approach in previous IR and recommendation research (D. Maxwell & Azzopardi, 2016; S. Zhang & Balog, 2020). It has become more

prevalent after LLMs demonstrated abilities in interactive agents with humanoid behaviors (Park et al., 2023; Tan & Jiang, 2023).

Then, the planner incorporated personal information c_i to tailor the information retrieval plan to the user's specific needs and preferences: $\hat{p}_i^r = \textbf{Planner}\,(p_i^r, c_i)$.

4. Information retrieval: The retrieval agent simulates searching for information online according to the user-centric plan $\hat{p}_i^r$ and generates relevant information:

$$r_i = \textbf{Retrieval}\,(\hat{p}_i^r)\,. \tag{3.4}$$

Then, the planner incorporated both personal information c_i and generated "online" information to update the action instruction plan $\hat{p}_i^a = \textbf{Planner}\,(p_i^a, c_i, r_i)$.

5. Action instruction: The action instruction agent transforms the plan $\hat{p}_i^a$ to instruction for the user on how to take actionable steps or interact with the real world to bridge the gap between planning and execution in the simulation:

$$a_i = \textbf{Actioner}\,(\hat{p}_i^a)\,. \tag{3.5}$$

6. User engagement: The user follows the plan and provides updates on the current state

$$s_i = \textbf{User}\,(a_i)\,. \tag{3.6}$$

Steps 2-6 are iterative as the process continues until the user completes the final goal. The updated user state s_i provides essential task progress and feedback to the system, allowing The step planner to generate revised plans for subsequent steps $p_{i1} = \textbf{Planner}\,(g, p_0, s_i)$. The updated plan undergoes another cycle of collaboration by the

three agents. For a complete overview of the initial settings and prompts used in this framework, including $g, s_0, \mathcal{P}_0$ and $\mathcal{P}_{agent}$, please refer to Appendix A1.2.

3.2.1 Evaluation of the effectiveness of the framework

To assess the effectiveness of the GOLF framework in simulating long-term life task planning and completion, this study designs a comprehensive evaluation method focusing on several key components of the framework. This evaluation aims to understand how the framework impacts task completion, user engagement, and adaptability across iterations. The evaluation consists of defining evaluation targets and metrics, the evaluation procedure using both LLM and human evaluators, experiments on prompt quality control, and two types of ablation studies.

3.2.1.1 Evaluation targets, procedure, and metrics

The evaluation focuses on the following components at each iteration i:

- **Step plan evaluation** e_i^p: Assessing the overall quality of the step plan p_i in relation to the user's previous state s_{i-1}.
- **Conversation evaluation** e_i^c: Evaluating the quality of the conversation c_i between the communicator agent and the user, focusing on how well it elicits user needs and reduces task uncertainty considering the user's previous state s_{i-1}.
- **Information retrieval plan evaluation** e_i^r: Assessing the usefulness of the information task plan p_i^r in terms of the contributions to help the user at the state s_{i-1}.

- **action instruction evaluation** e_i^a: Assessing the appropriateness and quality of the action instruction a_i instructions in terms of workload and alignment with the user state s_{i-1}.

Thus, for each component, the evaluation can be represented as

$$e_i^x = \textbf{Evaluator}\left(x_i, s_{i-1}\right), \tag{3.7}$$

where x_i is the evaluation target, and **Evaluator** is the evaluation procedure conducted by an LLM agent or human annotators.

During the evaluation procedure, the evaluator is provided with a scenario and role instruction. For example, the scenario may describe a simulated user whose goal is to apply for a Ph.D. program. The user's current state might be at the beginning of the information searching, and the task context involves collecting information about Ph.D. programs. The role instructions guide the evaluator to act as this user, following consistent guidelines and metrics to assess the components.

The effectiveness metrics captures how well outputs recognize user state and advance the task. This evaluation process is not for measuring absolute model performance but aims to validate whether the simulation outcome is controllable and sensitive to different settings. The evaluation metrics listed in Table 3.2 are designed to comprehensively assess the framework's effectiveness in recognizing the user's state and facilitating progress toward the goal. Each metric focuses on a specific component of the framework and employs a 5-point scale to measure effectiveness or usefulness. These metrics capture both the system's ability to recognize and respond to the user's current state and its effectiveness in aiding the user to proceed toward their goal. The metrics are justified in their design as they collectively ensure that the framework is

user-centric, adaptive, and goal-oriented. They focus on key aspects such as responsiveness to the user's needs, the utility of the information needed, the practicality of action plans, and the actual progress made by the user. Specifically, there are indicators added in the evaluation prompts for effectiveness to distinguish good and bad task outcomes.

Table 3.2: Evaluation metrics for system effectiveness. It defines the effectiveness metrics used to evaluate different framework components (Step plan, Communication, Information, Action, User state), scored on a 5-point scale. (*Detailed descriptions and rubrics are provided in the appendix Table A.2.*)

Metric	Description
Step plan effectiveness	Assesses how effectively the planner recognizes the user state and provides an appropriate action (proceed, update, revise) aligned with that state.
Communication effectiveness	Evaluates how effectively the communicator's questions extract meaningful information from the user, considering the corresponding recognized user state.
Information retrieval usefulness	Assesses the usefulness of the information retrieval tasks provided by the step planner, focusing on their relevance and contribution to task completion.
Action instruction effectiveness	Measures the effectiveness of the action instruction in enabling the user to proceed, update, or revise the task.
User state update effectiveness	Evaluates the effectiveness of the user's progress after following the action plan, focusing on whether the update indicates progression toward the goal.
5-point scale: 1 not effective (useful) $\rightarrow$ 5 highly effective (useful)	

Specifically, to provide a clear operationalization of "effective planning," the evaluation prompts for the LLM evaluator were augmented with specific indicators. These indicators were designed to help the evaluator distinguish between high-quality (good)

and low-quality (bad) task outcomes, ensuring the evaluation aligns with the framework's user-centric function. For example, in the effectiveness evaluation for the step plan, the evaluator is instructed to check for positive indicators such as "Context Fit" (i.e., the plan matches the user's current progress and specific challenges) and "Responsiveness to User State" (i.e., it considers the user's emotional and motivational state, such as uncertainty or fatigue). Conversely, the evaluator is guided to identify subtle signs of a "bad" task, such as "Detached Planning" (suggesting steps that look logical on paper but ignore what the user actually needs next) or "Superficial Informativeness" (providing information that sounds useful but doesn't help the user understand their own situation or make tangible progress). This operationalization ties the effectiveness metric directly to the user's immediate context and their potential uncertainty, rather than just abstract goal completion. The detailed prompt instruction is provided in Appendix 1.4 for the step plan effectiveness evaluation.

In addition to the effectiveness metrics, the study also designed uncertainty metrics to assess how the system influences users' perceptions of uncertainty across all components. These metrics aim to evaluate three key aspects: outcome uncertainty, action uncertainty, and intolerance of uncertainty. By measuring users' perceived uncertainty before and after interacting with the components, it can gain insights into how effectively the system reduces uncertainty and supports users in progressing toward their goals.

The uncertainty evaluation involves three statements listed in Table 3.3, each targeting a specific aspect of uncertainty. Outcome uncertainty assesses how uncertain the user is about the potential outcome of the plan and whether this uncertainty hinders task progression. Action uncertainty evaluates the user's uncertainty regarding the actions required to complete the task and whether they are able to follow the plan.

Intolerance of Uncertainty measures the user's intolerance level with uncertainty in the situation and how it impacts their planning, action, and overall ease.

For each statement, users indicate their level of agreement on a 5-point Likert scale ranging from strongly disagree to strongly agree. These metrics are essential for understanding not only the effectiveness of the components themselves but also their impact on users' confidence and ability to manage uncertainty.

Table 3.3: Evaluation metrics for perceived uncertainty. It presents the specific statements used to measure three dimensions of user-perceived uncertainty (Outcome, Action, and Intolerance), rated on a 5-point Likert scale. (*Detailed descriptions and scales are provided in the appendix Table A.3.*)

Metric	Statement
Outcome uncertainty	I am not clear about the outcome of the plan, and this uncertainty makes it difficult to proceed with the task.
Action uncertainty	I am not clear about the actions required to complete the task, and this uncertainty makes it difficult to follow the plan.
Intolerance of uncertainty	Uncertainty in this situation makes it difficult for me to plan, execute actions, or feel at ease.
5-point scale: 1 strongly disagree → 5 strongly agree	

3.2.2 Pilot study and analysis of an example generated by the GOLF framework

This section presents a pilot study to assess the feasibility of the simulation using an example. The example demonstrates a segment of long-term planning in obesity management, covering initial planning and the first iteration of step planning, including communication, information retrieval, action instructions, and user feedback. Each component in the iteration is evaluated for uncertainty and effectiveness. The example is generated by the GOLF system, and it serves to demonstrate the framework in

practice. The evaluation is conducted manually to clarify how the study expects to measure and justify the evaluation of generated content.

Figure 3.3 illustrates the initial planning stage, where the simulated user provides a goal (i.e., *undertake obesity management due to obesity concerns*) and the current state (i.e., *unfamiliar with the task and wants to start from the beginning*). The initial planner then generates an outline of activities and tasks, resembling a typical one-turn interaction between a user and an LLM. The initial plan itself is not evaluated in this study, as its complexity and quality can vary depending on the initial prompt. Since this falls under one-turn interaction evaluation, it is beyond the study's focus.

User:
I am planning to undertake obesity management due to obesity concerns.
I am unfamiliar with the task and want to start it from the beginning.

Initial Planner:
Goal: Undertake a structured obesity management program to achieve a healthier weight.

- **Activity 1: Assess Current Health Status**
  - **Task 1: Consult with a Healthcare Professional**: Schedule an appointment with a doctor or nutritionist to discuss your obesity concerns and get a health assessment.
  - Task 2: ...
  - Task 3: ...
- **Activity 2: Develop a Personalized Weight Management Plan**
- **Activity 3: ...**

Figure 3.3: Initial planning. This snippet shows the Initial Planner generating a high-level goal and activity list based on user input, establishing the roadmap for the simulation.

The pre-study uncertainty is assessed based on the user's initial state. As shown in Figure 3.4, the outcome uncertainty is rated at 4, indicating a lack of information about the expected results. The action uncertainty is rated at 5, reflecting the user's unfamiliarity with the task and uncertainty about how to proceed. The intolerance

of uncertainty is rated at 3, suggesting that while the user experiences uncertainty, it does not significantly impact their state of mind. The pre-study uncertainty should remain constant for the LLM evaluator but may vary for the human evaluator. This measure is implemented to mitigate individual differences in human evaluation.

(System sends the goal and the user state to the evaluator)

Evaluator:

- **Outcome uncertainty**: 4 The outcome is unclear, with significant ambiguity. Missing information, expectations, or unpredictable external influences make it difficult to plan effectively.
- **Action uncertainty**: 5 The actions are highly ambiguous or unknown.
- **Intolerance of uncertainty**: 3 The user is neither overly distressed nor fully comfortable with uncertainty in the situation.

Figure 3.4: Pre-study uncertainty annotation. This shows the Evaluator's baseline assessment of the user's pre-study uncertainty levels before step planning begins.

After generating the initial plan, the system sends it along with the user's initial state to the step planner. As Figure 3.5 shows, the step planner then analyzes the current progress and suggests step tasks in three categories: communication, information retrieval, and action instruction. Given that the user is unfamiliar with the task and starting from the beginning, the step plan should prioritize gathering user information, searching for local healthcare resources, and scheduling an appointment.

The evaluation of step planning includes both effectiveness and uncertainties, as shown in Figure 3.6. The effectiveness is rated at 4, as the step planner accurately recognizes the user's state and provides an appropriate plan. The plan is effective, as initiating contact with a healthcare professional is a logical starting point for the task. However, it does not fully identify the necessary user communication tasks, preventing the step plan from receiving the highest score (5 – very effective).

At this stage and in subsequent processes, uncertainties are measured as post-study uncertainty, assessed after the evaluator reviews the task planning components. Here,

(System sends the initial plan and current user state to the step planner)

Step Planner:
Progress: The user is at the beginning stage of the obesity management program and needs to assess their current health status.
Tasks:

- **Communication Task:** Schedule an appointment with a healthcare professional.
- **Information Retrieval Task:** Research local healthcare professionals or nutritionists.
- **Physical Action Task:** Prepare for the health assessment by noting any health concerns or questions to discuss during the appointment.

Figure 3.5: Step planning. This snippet displays the Step Planner breaking down the current progress into specific Communication, Information Retrieval, and Physical Action tasks.

both outcome and action uncertainties are rated at 3, indicating a decrease compared to the pre-study uncertainty. However, since the plan lacks details about the expected outcome of this process, its contribution to the final task, and sufficient specificity in the actions, a moderate level of uncertainty remains. The intolerance of uncertainty remains at 3, unchanged from the pre-study measurement.

(System sends the step plan and the user state to the evaluator)

Evaluator:

- **Effectiveness**: 4 Effective: The system recognizes the user's state correctly and provides an appropriate plan. The plan covers necessary details. It explores alternative options to address potential issues.
- **Outcome uncertainty**: 3 The outcome has some clarity but includes areas of uncertainty. There may be gaps in information, expectations, or external influences that could affect results.
- **Action uncertainty**: 3 Some uncertainty exists about the actions required. The task may involve moderate complexity, unclear steps, or dependencies that could impact execution.
- **Intolerance of uncertainty**: 3 The user may appreciate planning but do not find uncertainty paralyzing.

Figure 3.6: Evaluation of the step plan. This shows the Evaluator rating the effectiveness and uncertainty levels of the generated Step Plan

The next component is communication between the agent and the user, as shown in Figure 3.7. The agent generates questions to gather key details on the user's health concerns, past consultations, goals, dietary preferences, and appointment timing. The user responds, confirming no prior medical guidance, outlining weight loss and health improvement as goals, and expressing urgency in scheduling an appointment.

(System sends the communication tasks to the communication agent)

Communication Agent:

- What specific health concerns or conditions do you currently have that may affect your obesity management?
- Have you previously consulted with any healthcare professionals regarding your weight management?
- What are your main objectives for this obesity management program (e.g., weight loss, improved fitness, better health markers)?
- Are there any specific dietary or exercise preferences you would like to discuss during your appointment?
- How soon are you looking to schedule this appointment?

User:
I'm looking to start from the beginning with my obesity management. I have a few health concerns that might affect my plan, and I want to make sure I address them properly. I haven't consulted with any healthcare professionals about my weight management before, so I'm open to guidance. My main objectives are weight loss and improving my overall health. I also have some dietary preferences I'd like to discuss. Ideally, I'd like to schedule the appointment soon to get started.

Figure 3.7: Communication between the agent and the user. This illustrates the dialogue where the Communication Agent asks clarifying questions and the User provides specific details to refine the plan.

The evaluation of the communication component is shown in Figure 3.8. The effectiveness is at 4, as the questions are relevant and extract necessary details, though the user does not express specific concerns. Outcome uncertainty is rated at 3, indicating some clarity but remaining gaps in expectations or external information. Action uncertainty is low at 2, as the next steps are mostly clear and manageable. Intolerance

of uncertainty is also 2, suggesting the user can tolerate some ambiguity without it interfering with planning.

(System sends the communication and the user state to the evaluator)

Evaluator:

- **Effectiveness**: 4 Effective: The questions are relevant and useful. Extract necessary details and reflect on previous issues.
- **Outcome uncertainty**: 3 The outcome has some clarity but includes areas of uncertainty. There may be gaps in information, expectations, or external influences that could affect results.
- **Action uncertainty**: 2 The actions are mostly clear, with minimal uncertainty. The complexity is low, and the task is manageable.
- **Intolerance of uncertainty**: 2 The user can generally tolerate uncertainty in the situation but may prefer some information. Occasional uncertainty does not interfere with the actions or planning.

Figure 3.8: Evaluation of the communication. This displays the evaluation of the communication, assessing its effectiveness and uncertainty metrics.

The next step is updating the information retrieval plan and generating information, simulating an online information retrieval process, shown in Figure 3.9. Based on the user's progress, the step planner maintains the original retrieval task of researching local healthcare professionals or nutritionists. The information retrieval agent then identifies sources such as health directories, professional associations, and online review platforms. However, instead of directly providing specific details, such as healthcare locations and contact information, the agent misinterprets the task and merely suggests where the user can search for this information. This oversight can reduce the effectiveness of the retrieval process, as the user is left to conduct further information-seeking rather than receiving available information.

Therefore, in the evaluation shown in Figure 3.10, the effectiveness is rated at 1, indicating that the generated information is not useful. The agent misunderstood the task, generating tasks for the user rather than providing the necessary information, which failed to contribute to the step task. Outcome uncertainty is high at 4, due

(System sends the communication to the step planner to update the information retrieval task)

Step Planner:
Progress: User is preparing to schedule an appointment with a healthcare professional for obesity management and is gathering health concerns and dietary preferences to discuss. **Tasks:**

- **Information Retrieval Task (unchanged):** Research local healthcare professionals or nutritionists.

(System sends the information retrieval tasks to the information retrieval agent)

Information Retrieval Agent:

- **Information Task**: Research local healthcare professionals or nutritionists.
- **Information Source**: Local health directories, professional associations, and online review platforms.
- **Collected Information**:
  - To find local healthcare professionals or nutritionists, you can check health directories such as Healthgrades, Zocdoc, or Yelp.
  - Professional associations like the Academy of Nutrition and Dietetics provide a "Find a Nutritionist" tool.
  - Local community health centers or hospitals may also have lists of affiliated nutritionists and healthcare providers.
  - Online reviews and ratings can help assess their credibility and services.

Figure 3.9: Generating the online information. This shows the Information Retrieval Agent simulating an online search to gather external resources based on the plan.

to significant ambiguity and missing information that complicates effective planning. Action uncertainty is rated at 3, reflecting some complexity and unclear steps in the task execution. The intolerance of uncertainty is also high at 4, suggesting that the user may find the uncertainty frustrating and difficult to manage, potentially impacting their effectiveness when plans are disrupted or information is lacking. This highlights a need to redesign the prompt to ensure the agent understands the information tasks and provides the requested information directly to the user.

(System sends the goal and the user state to the evaluator)

Evaluator:

- **Effectiveness**: 1 Not Useful: The information tasks are irrelevant. Indicators: Do not contribute to the step task. Provide no value toward goal progression.
- **Outcome uncertainty**: 4 The outcome is unclear, with significant ambiguity. Missing information, expectations, or unpredictable external influences make it difficult to plan effectively.
- **Action uncertainty**: 3 Some uncertainty exists about the actions required. The task may involve moderate complexity, unclear steps, or dependencies that could impact execution.
- **Intolerance of uncertainty**: 4 The user may find uncertainty in the situation frustrating or difficult to handle. The user may tend to feel or less effective when plans are disrupted or information is lacking.

Figure 3.10: Evaluation of the generated information. This displays the evaluation of the retrieved information.

The last component for each iteration is the action instruction shown in Figure 3.11. The step planner updates the action instruction plan based on the communication and generated online information. Then, the action instruction agent outlines specific steps for the user, which include identifying potential healthcare professionals or nutritionists, checking their availability, contacting them to schedule an appointment, and preparing necessary documents for the visit. However, the user feedback indicates that these steps are incomplete due to missing preliminary information. The user highlights the need to first gather location-based details to identify suitable professionals, indicating that prior tasks must be completed before proceeding with appointment scheduling. This feedback demonstrates how the completion of previous components directly impacts the progression of the step task in this simulation. Since the requested information is missing, the user may encounter difficulties in completing this step, which can further impact long-term planning.

Accordingly, the evaluation of the action instruction and the user updates shown in Figure 3.12 rates the effectiveness at 2, indicating that the action plan is only slightly

(System sends the communication and generated online information to update the action instruction task)

Step Planner:
Progress: User is preparing to schedule an appointment with a healthcare professional for obesity management and is gathering health concerns and dietary preferences to discuss. **Tasks:**

- **Action Instruction Task (unchanged):** Schedule an appointment with a healthcare professional or nutritionist based on your research.

(System sends the action instruction task to the action instruction agent)

Action Instruction Agent:

- Identify potential healthcare professionals or nutritionists in your area.
- Check their availability and appointment scheduling options (online or by phone).
- Contact the chosen healthcare professional or nutritionist to schedule an appointment.
- Prepare any necessary documents or information to bring to the appointment.

User:

- **Incomplete:** I haven't started identifying potential healthcare professionals or nutritionists yet. I need to gather some information about your location to find the right options.
- **Incomplete:** This step cannot be completed until I identify the professionals in your area first.
- **Incomplete:** I need to complete the previous steps before contacting anyone.
- **Incomplete:** I will prepare this once the appointment is scheduled.

Figure 3.11: Generate the action instructions and the user feedback. This shows the Action Instruction Agent providing specific steps and the User reporting their inability to complete them due to missing prerequisites.

(System sends the action instructions and the user state to the evaluator)

Evaluator:

- **Effectiveness**: 2 Slightly Effective: The action plan is questionable. Uncertainty about ability to follow. Significant problems need addressing first.
- **Outcome uncertainty**: 3 The outcome has some clarity but includes areas of uncertainty. There may be gaps in information, expectations, or external influences that could affect results.
- **Action uncertainty**: 4 The required actions are unclear, with significant uncertainty. The complexity, lack of expertise or resources, and potential for changes make it difficult to proceed.
- **Intolerance of uncertainty**: 4 The user may prefer to know what to expect and to avoid surprises whenever possible. The user may tend to feel hesitant or less effective when plans are disrupted or information is lacking.

(System sends the updated user state and previous user state to the evaluator)

Evaluator:

- **Effectiveness**: 2 Slightly Effective: Minimal progress. Advanced slightly but encountered major new problems. Progress is hindered.

Figure 3.12: Evaluation of the action instruction. This displays the evaluation of the action instructions and updated user state, reflecting low effectiveness due to the issues mentioned in the user states.

effective. This low rating reflects a lack of specificity and insufficient details in the instructions, leading to uncertainty. Outcome uncertainty is rated at 3, suggesting some clarity but also highlighting gaps in external information that could impact results. Action uncertainty is rated higher at 4, due to unclear required actions and the complexity involved, coupled with a lack of resources. Intolerance of uncertainty is also high at 4, as the user prefers predictability and may feel hesitant or less effective when plans are disrupted or information is lacking. The user state update further confirms minimal progress, with the user making progress slightly but encountering major new problems that hinder overall task planning and completion.

The current user state can significantly influence the step planning process in the next iteration, as shown in Figure 3.13. The step planner recognizes that the user is ready to research local nutritionists and has prepared discussion topics for their appointment. Despite the user's unfamiliarity with obesity management tasks, their active engagement in finding a nutritionist highlights progress in their readiness to proceed. This state update guides the step planner to focus on an information retrieval task aimed at researching local nutritionists specializing in obesity management. The evaluation rates the effectiveness as moderately effective at 3, acknowledging that while the plan correctly identifies the user's state and progresses to the next step, it still risks encountering previous issues due to potential gaps in clarity and detail. Outcome uncertainty remains at 3, reflecting some ambiguity, and action uncertainty is high at 4, denoting unclear required actions and potential challenges in execution. The user's intolerance of uncertainty is rated at 3, indicating an ability to manage uncertainty but still requiring planning.

This example illustrates that simulated planning can exhibit varying qualities, resulting in different ratings across multiple dimensions. The goal of the pilot study or simulation is not to achieve the highest possible scores but rather to identify the circumstances under which the planning performs well or poorly. The evaluation plays a crucial role in accurately differentiating these circumstances. By manually identifying and analyzing these circumstances during the pilot study, the researcher can refine and adjust the simulation and evaluation framework. This process is critical before proceeding to a large-scale simulation and evaluation, ensuring that the framework is robust and effective in assessing planning quality across diverse scenarios. The next subsection delves into the method of the large-scale simulation and evaluation.

(Continue to the next iteration: system sends the initial plan and current user state to the step planner)

Step Planner:
Memory: User is ready to research local nutritionists and has outlined their discussion topics for the appointment. User is still unfamiliar with obesity management tasks but is actively seeking to engage with a nutritionist.
Progress: User needs to begin researching local nutritionists specializing in obesity management to schedule an appointment.
Tasks:

- **Information Retrieval Task:** Research local nutritionists specializing in obesity management.

(System sends the step plan and the user state to the evaluator)

Evaluator:

- **Effectiveness**: 3 Moderately Effective: The system recognizes the user's state correctly and provides a partially appropriate plan. Indicators: The plan begins to proceed to the next step but may be unclear. It covers some details but may still risk encountering previous problems.
- **Outcome uncertainty**: 3 The outcome has some clarity but includes areas of uncertainty. There may be gaps in information, expectations, or external influences that could affect results.
- **Action uncertainty**: 4 The required actions are unclear, with significant uncertainty. The complexity, lack of expertise or resources, and potential for changes make it difficult to proceed.
- **Intolerance of uncertainty**: 3 The user may appreciate planning but do not find uncertainty paralyzing.

Figure 3.13: Step planning and evaluation for the next step. This shows the transition to the next iteration, where the Step Planner adjusts the plan based on the previous cycle's feedback and memory.

3.2.3 Evaluation experiments

Simulations were conducted on two task topics (house purchasing, obesity management). The simulated user's initial state was "unfamiliar with the topic." Each run proceeded for five iterations. To ensure robustness, we used three widely available LLMs: GPT-4o, Gemini 2.5 Pro, and Claude Sonnet 4, for both simulation and evaluation, under default API settings (temperature, top-p). All experiment conditions were tested with outputs generated by all three models, each then cross-evaluated by

all models to reduce bias and model-dependent preferences. Furthermore, each generation was repeated ten times per condition to enhance result consistency.

3.2.3.1 Experiments on prompt variation and quality

To examine the influence of prompt quality on the framework's effectiveness, I design experiments that manipulate the quality of the step plans generated by the planner agent. In each iteration, it generates an alternative step plan with a controlled quality parameter q compared with the original step plan generation in Equation 3.2:

$$p'_i = \textbf{Planner}\left(g, p_0, s_{i-1}, q\right), \tag{3.8}$$

where q represents modifications to the prompt and constraints to control the quality of the generated plan. The default prompt is shown in Appendix 1.2 Step planner prompt. And the plain version (test condition) prompt is modified based on the default prompt by removing empirical task decomposition and planning strategies (e.g., rules of using the initial plan, details of task decomposition, and adaptability to user and task state). These variations were designed to highlight how empirical planning knowledge in prompts can influence the simulation outcome in a controllable way to validate the framework design. This results in corresponding alternative components c'_i, r'_i, and a'_i, as well as evaluations e'^x_i. By comparing e^x_i and e'^x_i, this study can assess if prompt content of specific planning strategies can affect the step plan and subtask execution.

After identifying the differences in the user state s_i and s'_i, this study conducts a between-iteration experiment to evaluate the adaptability of the step planner and how the user's state affects subsequent iterations. This experiment compares the impact of different user state on the next iteration's step plan's quality and overall performance.

Compared with the original step plan generation in Equation 3.2, the altered step plan ${p''}_{i1}$ incorporates the alternative user state s'_i:

$$ {p''}_{i1} = \textbf{Planner}\left(g, p_0, s'_i\right). \tag{3.9} $$

Accordingly, the evaluation of the altered plan ${e''}^{p}_{i1}$ is compared with the evaluation of the original plan e^{p}_{i1} to investigate the differences and reveal how sensitive the step planner is to changes in user progress and whether it can compensate for setbacks or accelerate progress.

3.2.3.2 Ablation studies

To understand the contributions of individual components and the task decomposition strategy in the GOLF framework, this study conducts two types of ablation studies: *Type one* component effectiveness within decomposed tasks and *Type two* effectiveness of task decomposition.

Type one study investigates the impact of excluding results from certain agents (i.e., c_i and/or r_i) during the updating of the action instruction plan $hatp^a_i$ while still maintaining the task decomposition into three subtasks ($p_i = \{p^c_i, p^r_i, p^a_i\}$). By contrast, *Type two* study evaluates the effectiveness of the task decomposition strategy by directly excluding the subtasks when generating the step plan. Specifically, the conditions include:

- Excluding the communication

  Type one:

$$ \hat{p}^{a-c}_i = \textbf{Planner}\left(p^a_i, r_i\right) \tag{3.10} $$

Type two:

$$\begin{aligned} p_i^{-c} &= \{p_i^r, p_i^a\} = \textbf{Planner}\,(g, p_0, s_{i-1}) \\ \hat{p}_i^{a-c} &= \textbf{Planner}\,(p_i^a, r_i) \end{aligned} \tag{3.11}$$

- Excluding the information retrieval

  Type one:

$$\hat{p}_i^{a-r} = \textbf{Planner}\,(p_i^a, c_i) \tag{3.12}$$

  Type two:

$$\begin{aligned} p_i^{-r} &= \{p_i^c, p_i^a\} = \textbf{Planner}\,(g, p_0, s_{i-1}) \\ \hat{p}_i^{a-r} &= \textbf{Planner}\,(p_i^a, c_i) \end{aligned} \tag{3.13}$$

- Excluding both the communication and information retrieval

  Type one:

$$\hat{p}_i^{a-cr} = p_i^a \tag{3.14}$$

  Type two:

$$p_i^{-cr} = \{p_i^a\} = \hat{p}_i^{a-cr} = \textbf{Planner}\,(g, p_0, s_{i-1}) \tag{3.15}$$

- Excluding user state

  Type one:

$$p_i^{-s} = \textbf{Planner}\,(g, p_0, p_{i-1}) \tag{3.16}$$

After setting all ablation conditions and getting altered action instruction plans $\hat{p}_i^{a-*}$, the study compares the evaluation results e_i^a and e_i^{a-*}.

3.2.4 Analysis

To answer RQ2, this study evaluates how variations in prompt quality and the constitution of system components affect the performance of the GOLF framework. The design of the GOLF framework is grounded in task management theories and cognitive task decomposition, which suggest that breaking complex goals into actionable subtasks reduces cognitive load and uncertainty (Allen, 2015; Maes et al., 2022; Soufan et al., 2021). Furthermore, effective Human-AI Collaboration relies on the system's ability to maintain context and adapt to user states. Based on these theoretical foundations, I postulate that the quality of the prompt (containing task management and planning strategies) and the presence of functional components (gathering context) are essential for generating effective plans. Therefore, I test the following hypotheses:

H1 : Prompts incorporating empirical planning strategies will yield different simulation outcomes compared to plain prompts, regarding plan effectiveness and uncertainty.

H2 : The inclusion of user states, communication, and information components, as the specific architecture of the framework, affects the simulation outcome, and altering or removing these components will yield different simulation outcomes compared to plain prompts, regarding plan effectiveness and uncertainty.

For H1, the prompt quality experiment involves a control group (with an unmodified step plan p_i) and a test group (with a prompt quality parameter q and a modified step plan p_i'). Each group is evaluated using the metrics e_i^x, where i represents the step number in the plan (if there are 5 steps, $i = 1, \ldots, 5$), and x denotes the component being evaluated, including the step plan p_i, conversation c_i, information retrieval plan p_i^r, action instruction plan a_i, and user state update s_i. Each e_i^x contains one effectiveness metric and three uncertainty metrics (i.e., outcome uncertainty, action

uncertainty, and intolerance of uncertainty). For H2, the ablation study compares e_i^a of the control group with e_i^{a-*} of seven test groups with altered components.

Given that the evaluation metrics are ordinal and measured on a 5-point Likert scale, this study employs the Wilcoxon signed-rank test for matched pairs to analyze the differences between the control and test groups. This non-parametric test is suitable for ordinal data and does not assume a normal distribution, making it appropriate for the dataset. An *a priori* power analysis was performed using G*Power to determine the necessary sample size. Consistent with established practices in HCI and social science research, the significance level was set at $\alpha = 0.05$. To enhance reproducibility and ensure a low risk of Type II errors, a stringent desired power of $(1–\beta) = 0.95$ was chosen, which is feasible given the large simulated sample size. A medium effect size ($d = 0.5$) indicates a difference of half a standard deviation between conditions, representing a moderate effect in HCI and social science research (Cohen, 2016). This analysis determined that a minimum of 57 matched pairs were required. This study collects 60 samples for each comparison, and 2700 records in total.

The data analysis begins with descriptive statistics to summarize the distribution of evaluation metrics and provide examples. To test H1 and H2, this study utilizes the Wilcoxon signed-rank test to compare the evaluation results between the test and control groups for each component, focusing on both effectiveness and uncertainty metrics. As there are multiple comparisons, it applies the Benjamini–Hochberg correction method to control the false discovery rate and reduce the likelihood of Type I errors.

3.3 GOLF benchmark for evaluating LLMs in long-term life task planning

Building upon the evaluation of the system's implementation effectiveness, this study aims to develop a dataset encompassing a broader range of topics and detailed steps to serve as a benchmark for assessing LLMs in long-term life task planning. The benchmark is designed to test LLMs' ability to generate coherent and effective plans for complex, multi-step tasks that individuals may encounter in real life.

3.3.1 Dataset development

To create a comprehensive benchmark dataset, the study first involves generating diverse topics. Using LLMs, ten areas of long-term life tasks are generated, such as financial planning, educational advancement, medical management, and professional development. Within each area, ten specific task topics are developed, similar to those predefined in the study (e.g., house planning, Ph.D. application, health management, career development). The prompts used to generate those topics and tasks are designed following the examples in G. Li et al. (2023)'s research, such as *List 10 diverse topics/tasks that a planner can assist a user cooperatively to achieve long-term life goals together. Be concise. Be creative. Here are examples:* {*EXAMPLES*}. Guidelines are established to ensure that the topics generated are diverse and representative of real-world long-term planning tasks. The researcher manually reviewed the generated topics to validate their quality and relevance.

For each topic, an initial plan was created following the GOLF framework, typically consisting of five activities, each containing four tasks. This structure yielded approximately twenty steps per plan, which served as the atomic units of the benchmark. Each step was expanded into four components: plan, communication, information,

and action. To simulate realistic user–system interactions, three user states were defined for every step: proceed (the user has completed the task), update (the user needs additional detail), and revise (the user faces difficulty completing the task). Each state corresponded to a unique action path, resulting in three variants of every step and enabling structured evaluation of decision-making under different user conditions. Through exposure to multiple user states and task scenarios, the LLM agents may identify relationships between user states and optimal recommendations, and predict appropriate steps for similar cases (Shi et al., 2024). The transformation of the task planning problem into an interactive recommendation scenario provides a formalized and reproducible framework for evaluation.

3.3.2 Benchmark validation

After construction, the benchmark was validated using both human and LLM annotators to assess their ability to distinguish among user states and select appropriate actions. A subset of benchmark steps was sampled for validation. Each step included one set of user states (Proceed, Update, Revise) and three shuffled action candidates generated by the GOLF framework.

3.3.2.1 Human annotation protocol

These tasks were published on Amazon Mechanical Turk (MTurk). To account for potential variations in individual judgment regarding task steps, each unique task was assigned to three independent annotators. This triangulation strategy was employed to mitigate individual bias and allows for the calculation of inter-rater agreement to assess the consistency of human consensus. A custom interactive interface was developed (Figure 3.14). In this interface, human annotators were presented with the task background, a description of the planning step, the three distinct user states, and the

three shuffled action instructions. The task required annotators to drag and drop each action instruction to the user state they believed was the best fit. To rigorously control for data quality and ensure participants understood the task mechanics, a specific attention check mechanism was integrated. An additional "attention check" candidate was mixed into the options, and participants were explicitly instructed to drag this candidate to a designated "Attention Check" slot. Only submissions that correctly completed this check were included in the analysis. Each task took approximately 8-10 minutes to complete, and annotators were compensated $1 per task. The online human data collection was approved by the university's IRB (# 16081).

3.3.2.2 LLM annotation protocol

The same tasks were assigned to LLM annotators to allow for direct comparison. LLMs received the exact same textual information—background, user states, and action candidates—formatted within a prompt (minus the graphical drag-and-drop interface). This ensured that both humans and LLMs were annotating identical content based on identical logic.

3.3.2.3 Annotation measurement

For both human and LLM annotators, accuracy was computed at both the state-specific and overall levels. State-specific accuracy measured correct matches within each user state, while overall accuracy measured correctness across all tasks. In addition, a permutation accuracy was used to represent the rate of all action candidates are assigned to user states accurately. For human annotators, additional measures included the inter-rater agreement rate (the proportion of tasks where annotators selected the same candidate), majority accuracy (accuracy based on majority vote), and unanimous accuracy (accuracy when all annotators agreed). The benchmark is generated by the same

Figure 3.14: Annotation interface. It presents the Amazon Mechanical Turk interface used for human annotation, showing the task background, user statuses, and the drag-and-drop mechanism for assigning action suggestions with an attention check option. *(Space is condensed in this figure for better presentation here).*

models in the GOLF simulation (claude-sonnet-4, gemini-2.5-pro, and gpt-4o), and the LLM annotators are the same models for consistency. These metrics jointly assessed the reliability of the benchmark and the discriminability of state–action mappings.

3.3.3 LLM performance measurement

To measure LLM performance, the full benchmark was used to evaluate a range of models across different capacity tiers. Each model received identical inputs consisting of a user states and three candidate actions and was required to select the most suitable response. The evaluated models included the GPT and Gemini series listed in Table 3.4: gpt-4.1-nano, gpt-5-nano, gpt-4.1-mini, gpt-4o-mini, gpt-5-mini, gpt-4-turbo, gpt-4o, gpt-4.1, gpt-5, gpt-3.5-turbo, gemini-2.0-flash, gemini-2.5-flash, and gemini-2.5-pro, representing nano, small, medium, and large performance tiers. Accuracy was calculated using the same method as in the validation stage. This evaluation provided a systematic comparison of reasoning robustness and adaptation to user context across models of varying capacity.

Table 3.4: Benchmark model candidates. It categorizes the LLM models used in the benchmark into performance tiers (Nano to Large), separating GPT models and Gemini models.

Model level	**GPT models**	**Gemini models**
Nano	gpt-4.1-nano, gpt-5-nano	
Mini / Small	gpt-4.1-mini, gpt-4o-mini, gpt-5-mini	gemini-2.0-flash
Base / Medium	gpt-4-turbo, gpt-4o	gemini-2.5-flash
Large	gpt-4.1, gpt-5	gemini-2.5-pro
Historical Baseline	gpt-3.5-turbo	

Chapter 4

Results

4.1 How people use generative AI for long-term life tasks planning

This study recruited 14 participants aged between 20 and over 80, spanning diverse professional and educational backgrounds, including students, engineers, managers, educators, and researchers. Table 4.1 summarizes their demographic information and selected planning tasks. Participant IDs (PIDs) show the recruitment channel, with *E* representing email and *R* representing Reddit. Most reported using ChatGPT several times a day for both professional (e.g., study or work-related tasks) and personal purposes (e.g., health, daily organization, or learning). One participant (E3) reported infrequent use, mainly out of curiosity. Across the sample, participants engaged in planning tasks spanning personal health and relationship management, activity and event coordination, and career or professional development. This diversity reflects the broad applicability of AI-assisted planning across everyday and long-term life contexts.

4.1.1 Planning task topics and methods

For **RQ1**, the study investigated the task topics and prompt behaviors. During the user activity in this study, participants engaged in planning tasks spanning various topics. Based on the similarity among those topics, the researcher grouped them into

Table 4.1: User background and selected topics. It presents the 14 participants, including their recruitment channel*, age range, professional background, and the task topic they chose for the study.

PID*	**Age**	**Career / Background**	**Selected Task Topic**
E1	31–35	Social science researcher	Anxiety management
E2	31–35	LIS Ph.D. student	14-day trip in Japan
E3	56–60	Marketing consultant	Health influencer strategy
E4	26–30	LIS Ph.D. student	Yellowstone trip
E5	56–60	IT Manager	Filmmaking preparation
E6	>80	LIS Professor	Curiosity in later life
E7	61–65	Civil engineer/designer	Weekly fitness plan
R8	21–25	Biomed master's student	Girlfriend communication
R9	26–30	Software developer	3D model development
R10	21–25	Business undergrad	Weight loss plan
R11	21–25	Education undergrad	Friends' avoidance
R12	21–25	Architecture undergrad	Study/workout schedule
R13	26–30	High school teacher	Birthday party planning
R14	31–35	Finance specialist	ASA microfinance learning

*Participant recruitment channel: email (E) and Reddit (R)

three broad categories presented in Table 4.2: personal and well-being, activity and event planning, and professional development and learning. **Personal and well-being** includes tasks focused on self-improvement in physical or mental health, everyday scheduling, or interpersonal relationships, often recurring. **Activity and event planning** includes tasks related travel itineraries, route selection, and life events, typically requiring coordination of multiple factors such as preferences and budget. **Professional development and learning** includes tasks related to work, study, and research, often involving knowledge acquisition, strategic planning, or content generation. These topics reflect the broad applicability of AI-assisted planning across user groups and life contexts.

4.1.1.1 General prompting behaviors

Along with the task topics, Table 4.2 also lists the source and number of prompts participants used. Prompts were either adapted from *templates* provided by the researcher or written spontaneously by participant them*selves*. Participants showed varying levels of engagement: some relied entirely on templates, others wrote all prompts themselves, and about half combined both, either starting with templates to explore AI capabilities or extending self-written conversations later. The number of prompts ranged from three to over ten, reflecting different engagement and confidence levels.

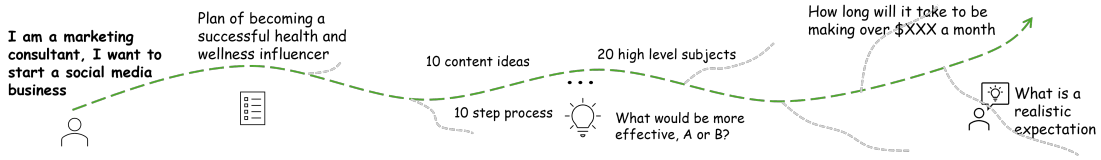

Figure 4.1: An illustration of the planning process of one participant. The main green line represents the path planned with ChatGPT. The gray branches represent hidden paths leading to different results or potential failure.

Table 4.2: Task topics grouped in three main categories. It groups the participant task topics into three main categories (Personal/well-being, Activity/event, Professional/learning) and indicates the source and count of prompts used.

Category	**(PID) Task topics**	**Prompt source (count)**
Personal/ well-being	(E1) Anxiety management	Template (3)
	(E7) Weekly fitness plan	Self (12)
	(R8) Girlfriend communication	Template (4)
	(R10) Weight loss plan	Template (3)
	(R11) Friends' avoidance	Self (2) + Template (1)
	(R12) Study/workout schedule	Self (3) + Template (3)
Activity/ event planning	(E2) 14-day trip in Japan	Template (3)
	(E4) Yellowstone trip	Self (8) + Template (2)
	(R13) Birthday party planning	Template (3)
Professional development/ learning	(E3) Health influencer strategy	Self (15)
	(E5) Filmmaking preparation	Self (8) + Template (2)
	(E6) Curiosity in later life	Self (8)
	(R9) 3D model development	Self (2) + Template (4)
	(R14) Microfinance learning	Self (2) + Template (1)

Figure 4.1 illustrates one planning path from the participant E3. To summarize prompting behaviors, all prompts were categorized by their main functional role, such as initiating, refining, revising plans, providing context, or verifying information (Table 4.3). Most prompts centered on providing background, initiating plans, or revising details, aligned with the template design. Self-generated prompts covered a broader range, extending to searching or verifying information, stating preference or constraints, generating ideas, and comparing options. Individual prompts can combine multiple functional roles. For instance, initiating a plan often incorporates providing context or specifying preferences. The study classified each by its dominant function to capture general interaction patterns. Overall, participants engaged in iterative cycles of generation, refinement, and adaptation when planning with ChatGPT.

4.1.1.2 Prompting in personal and well-being tasks

Tasks in this category covered personal concerns such as fitness, stress, and relationships. Prompts typically began with a brief description of participants' situation or emotional state, such as *"I'm preparing my PhD dissertation and feeling anxious; can you help me design an anxiety-management plan?"* (E1), or *"I'm having trouble with my girlfriend; please help me make the relationship better."* (R8). Participants used these openings to externalize difficulties and invite ChatGPT's general suggestions for improvement.

Participants then expanded on these prompts with contextual information, clarifying constraints or confirming details. For example, Participant E7 developed a weekly fitness routine through twelve prompts, providing background on age, schedule, and fitness level, then verifying information such as calorie and protein intake (*"Where did you get this information?"*). This process illustrated how users alternated between generating plans and validating information. Across participants, prompting in this

Table 4.3: Prompt categories. It classifies the types of prompts used by participants, providing descriptions and specific examples for each category.

Category (count)	**Description**	**Example prompts**
Providing background/ context (13)	The user describes their current situation, goal, or constraints (e.g., age, location, timeline).	*"I'm 81... my goal in life is to learn more about how the world works. Do you have any advice for me?"* (E6)
Initiating a plan (11)	The user has a clear goal and wants the AI to generate a step-by-step plan, schedule, or strategy.	*"Create a realistic weekly schedule that includes study, fitness, rest and side projects"* (R12)
Asking for details/ refinement (24)	The user wants to dig deeper into or clarify parts of a generated plan or suggestion.	*"your plan is very general, I have a lot of preferences and considerations. we can start with the Day 1 plan. Ask me more detail questions..."* (E2)
Revising or adjusting the plan (9)	The user has received an initial plan and now wants to adjust or optimize it.	*"revise the exercise program to help me lose 2 lbs weekly if I eat 1500 calories daily..."* (E7)
Searching or verifying information (15)	The user wants factual information or asks the AI to justify its answers.	*"where did you get this information. do you have reliable sources or have you guessed at some answers...?"* (E7)
Stating preferences/ constraints (7)	The user specifies likes/dislikes, conditions, or constraints related to the plan.	*"Okay, I won't go to Rocky Mountain. I don't like long hiking"* (E4)
Generating ideas/ content (6)	The user requests the AI to brainstorm or outline ideas, content, scripts, etc.	*"What actions can I do to make her feel better"* (R8)
Comparing options (4)	The user is choosing between multiple options and wants help deciding.	*"What would be more effective, a video of me walking in the park while talking, or my voice overlay on a music track?"* (E3)

domain emphasized contextualization and confirmation rather than exhaustive specification. Users focused on checking whether ChatGPT's advice fit their situation and appeared realistic or trustworthy. Most treated ChatGPT's plans as provisional outlines that could later be adapted to real-life conditions.

4.1.1.3 Prompting in activity and event planning

Tasks in this category involved planning trips or events that required coordinating multiple factors such as timing, transportation, and budget. Compared with health-related tasks, these prompts emphasized specificity and feasibility. Participants aimed to turn ChatGPT's general outlines into actionable itineraries through repeated refinement.

For instance, Participant E2 began with a broad request for a Japan trip plan but then noted, *"Your plan is very general; I have many preferences,"* prompting ChatGPT to ask clarifying questions before generating a day-by-day itinerary. Subsequent conversations refined flight routes, hotel safety, and budget, revealing a desire for control and contextual accuracy. Similarly, Participant E4, planning a Yellowstone trip, compared routes and schedules (*"Would leaving from Denver take one more day?"*), personal preferences (*"I don't like long hikes"*), and timing constraints (*"My flight gets to Salt Lake City at 10 pm"*). These conversations highlight how participants negotiated between AI-generated generality and their own contextual knowledge. They refined plans until results felt realistic or until additional prompting no longer seemed productive.

4.1.1.4 Prompting in professional development and learning

Tasks in this category focused on career goals, academic learning, and creative or professional growth. Compared with other domains, these tasks were more open-ended and informationally uncertain, requiring verification and integration of new knowledge.

Participants such as E3, E5, and E6 treated ChatGPT less as a planner and more as a thinking partner, a conversational collaborator for exploring ideas and testing assumptions. Prompts combined background description, idea generation, and verification. For instance, participants asked ChatGPT to draft strategies or outlines, then iteratively refine content (*"Make it sound more personal"*) or evaluate realism (*"Would this income goal be realistic?"*). Other participants (R9, R14) used ChatGPT mainly for factual queries or practical information. Their interactions were short and task-oriented, focused on retrieving concrete information rather than reflection. Across these examples, professional and learning-oriented tasks ranged from exploratory to specific, depending on users' topic familiarity and goals.

4.1.2 Uncertainty in AI-assisted task planning

For **RQ2**, the study organized participants' reflections according to four interview questions related to uncertainty: (1) which aspects of planning felt uncertain; (2) how such uncertainty affected their process, confidence, and follow-through; (3) how AI reduced or amplified uncertainty; and (4) how planning without AI compared. Table 4.4 summarizes the key aspects and illustrative examples for each question.

4.1.2.1 Which part feels uncertain

Participants' reflections revealed two main sources of uncertainty: those inherent to the task itself and those introduced by the AI system. Task-inherent uncertainty included doubts about the feasibility of plans, uncertainty about where to start, and tension between detail and flexibility. Several participants were unsure whether their goals were realistic or sustainable in practice: *"I wasn't sure if it was going to work out"* (R10), or expressed confusion about how to begin a complex plan: *"I should ask but*

don't know what to ask" (E5). Others found that excessive structure could increase anxiety rather than reduce it, noting that *"too detailed plans make me anxious"* (E1).

AI-introduced uncertainty, in contrast, stemmed from how ChatGPT generated and presented information. Participants questioned the contextual fit of its suggestions: *"I will arrive at night. It still planned as if daytime"* (E4), and the credibility of its content: *"It creates citations that don't exist"* (E6). These accounts highlight that uncertainty emerged both from the nature of the planning task itself and from limitations in how AI interpreted and contextualized user inputs.

4.1.2.2 Effect of uncertainty or uncertainty mitigation

Uncertainty shaped both the emotional and behavioral dimensions of participants' interactions with ChatGPT. Emotionally, it elicited a mix of frustration, satisfaction, and growing confidence. Some participants described discouragement when facing incomplete or inconsistent outputs: *"Uncertainty causes frustration... you feel like you can't achieve anything"* (R9), while others felt more at ease when plans were flexible and non-obligatory: *"If it's not like an assignment I must finish, I feel better"* (E1). Over time, positive outcomes helped transform skepticism into trust when uncertainty reduced: *"The results were good... I gave it that confidence"* (R10).

Behaviorally, uncertainty led to iterative refinement, cross-checking, and simplifying plans. When answers were vague, participants adjusted their prompts: *"If it doesn't answer my question, I'll tweak it until I get what I want"* (E7), or verified information externally: *"I have to fact check one by one"* (E2). Some simplified long or complex plans into shorter, clearer steps: *"Sometimes I want to simplify this plan in two steps"* (R12). In this way, uncertainty became a driver for clarification and learning rather than a negative experience.

4.1.2.3 How AI affects uncertainty

Participants described ChatGPT as both mitigating and amplifying uncertainty. In many cases, AI helped structure goals, encourage reflection, provide new information, and increase confidence. For example, one participant noted that ChatGPT *"reminded me to break things into smaller parts"* (E1), while another explained, *"He asked why I felt anxious... I started to think about it myself"* (E1). Others appreciated how the system surfaced novel ideas: *"I never would have thought of these, you don't know what you don't know"* (E3), or how clear guidance boosted motivation: *"I feel confident because it's guiding me"* (R10). Through such mechanisms, ChatGPT reduced uncertainty by offering structure and prompting self-awareness.

At the same time, AI sometimes exaggerated uncertainty by fostering overreliance or producing insufficiently specific advice. Some participants cautioned that using ChatGPT effectively required their own background knowledge: *"You should have a basic understanding yourself before using ChatGPT"* (E4), while others found its responses too generic to act on: *"It gives disclosure but still generic"* (R11). Thus, while ChatGPT often clarified what to do, it could also blur boundaries between confidence and dependency.

4.1.2.4 Uncertainty with non-AI methods

When comparing AI-assisted planning with traditional approaches, participants highlighted differences in accessibility and efficiency. Regarding accessibility, non-AI methods such as human experts or reference materials supported fact-checking and provided structured knowledge: *"I can check from other websites or user reviews"* (E2); *"Books have structure and flow"* (R9), but they also involved social discomfort, especially for personal topics: *"Hard to open up about personal issues... a friend may leak your information"* (R8).

In terms of efficiency, participants emphasized the challenges of information overload and time effort. Searching online was described as *"messy and overwhelming"* (E4), and manual planning as *"taking a lot of brain energy"* (R12). Compared with these experiences, ChatGPT was valued for its immediacy and cognitive relief, offering concise summaries and a private, always-available space for reflection. Together, these accounts suggest that while non-AI methods reduce uncertainty through credibility and expertise, AI expands accessibility and efficiency, shifting how people define what is feasible or worth planning.

4.1.3 Overall usefulness and limitations

For **RQ3**, Table 4.5 presents results of participants impression and feedback on the plan's usefulness and limitations. Participants generally described ChatGPT as both useful and limited in supporting their planning processes. Across interviews, two recurring strengths stood out: its ability to structure complex goals into manageable steps, and its conversational tone that made planning feel less solitary. At the same time, participants were candid about the system's limitations, particularly its lack of personalization, contextual realism, and adaptive follow-up. These perceptions reveal how users positioned ChatGPT not only as a task assistant but as a reflective partner that helped them externalize thinking, negotiate uncertainty, and stay motivated.

4.1.3.1 Perceived usefulness: structure, motivation, and reflection

Participants widely appreciated ChatGPT's ability to transform abstract goals into concrete steps. It offered structure, measurable targets, and a sense of progress that supported motivation: *"Dividing big goals into small tasks gave me a sense of control"* (E1). Beyond practical organization, many described ChatGPT as a conversational partner that stimulated reflection and emotional reassurance. *"It pushes my thinking"*

(E6) and *"It confirms that I can do this"* (E1) exemplify how participants perceived the system as both supportive and reflective, enabling them to reason through decisions and sustain engagement over time.

4.1.3.2 Perceived limitations: personalization, realism, and credibility

Despite these benefits, participants frequently emphasized the generic and idealized nature of AI-generated plans. Even when detailed information was provided, outputs often ignored personal constraints or contextual nuances: *"It never asks me how many people, the plan will be totally different"* (E2). Others criticized the lack of realism: *"It's just an outline... not real-life enough"* (R12), and expressed frustration with having to repeatedly adjust prompts. Concerns about credibility were also common. Several participants questioned the factual grounding of responses, choosing to verify them through external sources such as medical or educational websites: *"I would check Mayo Clinic for reliability"* (E7).

4.1.3.3 Expectations for improvement and adaptive control

Participants' reflections on these limitations translated into clear expectations for future AI-assisted planning tools. They wanted systems that could remember personal context, adapt to progress, and explain the basis of their suggestions. Many also desired finer control over tone and style, ranging from supportive to more critical feedback, rather than default encouragement. As one participant summarized, *"If it didn't compliment people so much, it would be more believable"* (E3), suggesting that personalization and situational awareness were viewed as essential for genuine assistance.

Overall, participants recognized ChatGPT's dual role as a planning scaffold and a reflective partner. Its ability to structure goals and maintain a human-like tone made it valuable for initiating and sustaining planning. Yet its limitations in personalization,

realism, and credibility constrained users' trust and reliance. These mixed impressions mirror a broader tension throughout the study: ChatGPT helps users think and plan, but it also transfers part of the cognitive effort back to them. In this sense, prompting becomes an ongoing act of negotiation, between structure and flexibility, automation and agency, and guidance and trust.

Table 4.4: Uncertainty dimensions in AI-assisted task planning. It summarizes qualitative findings regarding uncertainty, categorizing them into sources, effects, AI impacts, and comparisons with non-AI methods

Category	**Key aspects and example quotes**
Which part feels uncertain	
Task-inherent	**Feasibility of plans** — *"I wasn't sure if it was going to work out."* (R10). **Knowing where to start** — *"I should ask but don't know what to ask."* (E5). **Balancing detail and flexibility** — *"Too detailed plans make me anxious."* (E1)
AI-introduced	**Contextual fit** — *"I will arrive at night. It still planned as if daytime."* (E4). **Credibility** — *"It creates citations that don't exist."* (E6).
Effect of uncertainty or uncertainty mitigation	
Emotion	**Frustration** — *"Uncertainty causes frustration... you feel like you can't achieve anything."* (R9). **Relief** (mitigated) — *"If it's not like an assignment I must finish, I feel better."* (E1). **Growing confidence** (mitigated) — *"The results were good... I gave it that confidence."* (R10).
Behavior	**Iterative refinement** — *"If it doesn't answer my question, I'll tweak it until I get what I want."* (E7). **Cross-checking** — *"I have to fact check one by one."* (E2); *"Do more research and come back again."* (R9). **Simplifying plans** — *"Sometimes I want to simplify this plan in two steps."* (R12).
How AI affects uncertainty	
Mitigate	**Structuring goals** — *"It reminded me to break things into smaller parts."* (E1). **Encouraging reflection** — *"He asked why I felt anxious... I started to think about it myself."* (E1). **Providing new information** — *"I never would have thought of these, you don't know what you don't know."* (E3). **Increasing confidence** — *"I feel confident because it's guiding me."* (R10).
Exaggerate	**Overreliance** — *"You should have a basic understanding yourself before using ChatGPT."* (E4). **Lacking specificity** — *"It gives disclosure but still generic."* (R11).
Uncertainty with non-AI methods	
Accessibility	**Fact-checking** — *"I can check from other websites or user reviews."* (E2); *"Books have structure and flow."* (R9). **Social discomfort** — *"Hard to open up about personal issues... a friend may leak your information."* (R8).
Efficiency	**Information overload** — *"Online explanations are messy and overwhelming."* (E4). **Time effort** — *"Manual planning takes a lot of time and brain energy."* (R12).

Table 4.5: Participants' impression on task planning using ChatGPT. It summarizes participant feedback on the system's usefulness and limitations, supported by representative quotes.

Theme	Description	Example quotes
Scaffolding, motivational support	Helped users start, decompose goals, and feel progress	*"Dividing big goals into small tasks gave me a sense of control."* (E1) / *"I can start putting parts together to a fairly comprehensive social media campaign."* (E3)
Reflective, conversational value	Served as sounding board; confirmed instincts or pushed reflection	*"It pushes my thinking."* (E6) / *"It confirms I can do this."* (E1) / *"It takes that position of you... like another human being on the other side."* (R8)
Personalization and realism limits	Plans too general or idealized; missed context details	*"It never asks me how many people."* (E4) / *"It's just an outline... not real-life enough."* (R12) / *"Plans need to be realistic about my free time."* (R12)
Trust and control expectations	Users verified info; wanted adaptive memory, tone control, and transparency	*"I would check Mayo Clinic for reliability."* (E7) / *"I can't apply the code directly."* (R9) / *"If it didn't compliment people so much, it would be more believable."* (E3)

4.2 Simulation on long-term life task planning

The qualitative findings from the interview study reveal that users engage with generative AI as *cognitive scaffold* and *reflective partner*. These insights highlight specific user needs that current systems fail to meet: assistance in decomposing complex goals, managing the distinct challenges of uncertainties, and the desire for an adaptive, context-aware partner. These identified user requirements form the conceptual blueprint for a computational framework designed to systematically simulate and address these core human-AI collaboration challenges. Therefore, I designed the GOLF simulation framework, aiming to model the human-AI collaboration process for long-term planning, and develop a system responding to the challenges and user behaviors observed in our interviews.

4.2.1 Description of the planning simulation

To demonstrate the content of the planning simulation, here are example snippets of the initial plan and components at one step. For the initial plan, there are some similarities and differences between task topics. While both topics adopt systematic planning, their procedural dynamics and cognitive emphases differ substantially. The obesity management plan centers on adaptive behavioral change. It requires continuous feedback, self-monitoring, and periodic recalibration of dietary, physical, and psychological strategies. By contrast, the house purchase plan follows a sequential transactional process, where tasks are externally defined, time-bound, and oriented toward legal and financial closure.

Example snippet from the **adaptive phase** of *obesity management*:

- **Feedback loop**: Track body metrics and adjust dietary plan weekly.
- **Cognitive reinforcement**: Reflect on triggers and reinforce sustainable habits.
- **Professional feedback**: Revise the plan with a healthcare provider after review.

Example snippet from the **procedural stage** of *house purchase due diligence*:

- **Verification**: Conduct home inspection and mortgage approval.
- **Compliance check**: Review title documents with legal counsel.
- **Closure action**: Sign purchase agreement and transfer funds.

The divergence is most visible in how each plan handles uncertainty and decision flow. In obesity management, uncertainty arises from individual physiological and motivational variability, leading to iterative refinement. In house purchasing, uncertainty is primarily environmental—market fluctuations, institutional procedures—managed through discrete verification steps and expert mediation.

After the initial plan, there are some examples of the step components. Additional descriptive statistics (e.g., the average lengths, number of tasks) are presented in Appendix Table A.8 to A.10. The step plan is generated for the house purchase task at the step "Research and preparation". It includes the **communication**, **information retrieval**, and **action instruction** tasks.

Main content of the **step plan** for *house purchase research and preparation*

- **Communication**: Ask the user to clarify their current financial situation.
- **Information retrieval**: Look up the process and benefits of getting pre-approved for a mortgage.
- **Action instruction**: Gather financial docs and contact a mortgage lender.

For the **communication** component, the system generates detailed confirmation or follow-up questions, e.g., "Have you already reviewed your savings, income, and credit

score to determine your budget?" or "Do you have any specific neighborhoods or preferences regarding commute, schools, and amenities?". The user agent then provides corresponding answers like "I'm not sure about my budget. As for neighborhoods, I don't have specific ones in mind, but I prefer areas with a good commute and amenities.". For the **information** component, the system simulates a search process and produces results with a title, source, and snippet, e.g., "Top Neighborhoods with Excellent Schools and Parks – 2023 Housing Report", and "Find mortgage lenders that offer pre-approval services".

For the **action** component, the system generates concrete, small-scale action instructions tailored to the collected user and online information, such as:

- Gather financial documents (savings, income, loan offers).
- Use a budgeting tool to input financial details.
- Make a list of must-have housing features (bedrooms, location, ...).
- Rank these features by importance to guide home evaluation.

The user agent then updates progress and user state, e.g., noting difficulties with organizing documents or defining preferences. For example, "difficulties with organizing the documents efficiently since I'm new to this process. I am struggling with defining my preferences due to a lack of experience in house hunting."

4.2.2 Prompt version experiment

For each component, the LLM evaluator measures **effectiveness** and three types of uncertainty, i.e., outcome uncertainty (**OU**), action uncertainty (**AU**), and intolerance of uncertainty (**IU**), and also provides qualitative explanations. Table 4.6 reports the scores for each component under the default and plain prompt conditions.

In the default condition, the **effectiveness** scores of the step plan, communication, and information components are all above 4, indicating consistently high effectiveness. By contrast, the action component achieves a lower effectiveness score of 3.49, suggesting only moderate effectiveness. This shows that the LLM evaluator can perceive the relative weakness of the action component compared to others in terms of its contribution to task completion.

With respect to uncertainty, both **OU** and **AU** fall within the range of 2.0–2.5, while **IU** is closer to or above 2.5. This pattern indicates that the LLM systematically assigns higher values to **IU**, likely because the evaluation prompt directs it to consider the user's perspective. In particular, the step plan component receives an **IU** score of 2.95, close to the midpoint of 3, suggesting moderate intolerance of uncertainty. This may reflect the evaluator's sensitivity to the user's state, as the step plan incorporates user-related context. For example, the evaluator notes: "The user is a self-proclaimed beginner facing a complex, high-stakes task ('unfamiliar with the task'). This context likely makes them less tolerant of ambiguity and prone to feeling overwhelmed. ... The overall uncertainty of the home-buying process itself is likely to be a source of frustration and hesitation for the user". Such reasoning explains why the LLM assigns a higher IU score (up to 4 in this case) for certain steps.

Compared with the default condition, the plain version prompt condition consistently yields lower effectiveness and higher uncertainty scores for the step plan and information components. For communication and action, only AU increases. This shows that removing planning skills mainly reduces effectiveness and increases AU overall, with OU/IU most affected in step plan and information.

The LLM evaluator's explanations highlight why these increases occur. For example: "While the outcome of each individual action is clear (e.g., 'you will have your documents'), the overall outcome of completing this set of actions is ambiguous. The

Table 4.6: Evaluation on all components in default and plain prompt conditions. It compares the evaluation scores (Effectiveness and Uncertainties) across all framework components between the default prompt and the plain prompt conditions.

	step plan		communication		information		action instruction	
	default	plain	default	plain	default	plain	default	plain
Eff	4.47	**4.29****	4.35	**3.80****	4.26	**3.95****	3.49	**3.33****
OU	2.28	**2.45****	2.29	2.30	2.14	**2.25****	2.38	2.42
AU	2.06	**2.38****	2.25	**2.34***	2.15	**2.29****	2.11	**2.20***
IU	2.95	**3.20****	2.44	2.48	2.42	**2.48***	2.60	2.66

Values in **bold** indicate a significant difference from the default value with corrected p: * <0.05, ** <0.01.

user is left wondering, 'After I do all this, what will I have accomplished? How much closer will I be to buying a house?' The plan lacks a clear endpoint or a sense of progress, making the ultimate result of following it feel uncertain." "Complex actions like 'Contact a mortgage lender' are presented without any supporting guidance on what to ask or how to prepare, which creates significant uncertainty about how to execute that step." Such reasoning leads the LLM to assign high outcome and action uncertainty scores (up to 4 in this example).

Importantly, prompt modifications also propagate through user states into later steps. I compared step plans with user states carried over from default vs.plain conditions. Results show the plain condition slightly raises both effectiveness and OU (4.31 vs.4.25, 1.91 vs.1.88, $p<0.05$). This suggests that a weaker user state (as produced under the plain prompt) may give the step plan more room to support the user, thereby increasing perceived effectiveness, but at the same time it also raises outcome uncertainty, reflecting the more fragile state of the user progress.

To further test this indication and analyze the framework's robustness under different user conditions, this study conducted an additional experiment. In this experiment simulated a weaker, more uncertain user state. For both the default and plain prompt

conditions, the original user state was altered by prepending a text "the user has difficulties to complete the task" and setting all task completion states to "incomplete". It then compared the evaluation metrics for the step plan between these *altered* and the *original* user states.

The results are shown in Table 4.7. The findings align with the indication. For both prompt conditions, the *altered* user state resulted in significantly higher perceived effectiveness and higher outcome uncertainty compared to the original state. This reinforces the interpretation that a more fragile user state can increase the perceived effectiveness, while simultaneously reflecting the higher outcome uncertainty. Notably, under this "altered" user state, the default prompt condition still produced plans rated as more effective and less uncertain than the plain prompt condition, confirming the consistent difference between the prompt conditions.

4.2.3 Ablation study

The ablation study investigates how removing different contextual components affects the perceived quality of the final action instructions. It tested the ablations of user state, communication, information. There are two types of ablation for communication and information: Type 1, where the step plan is the same as the default (including the component), but the component is skipped in the system execution before the action instruction; and Type 2, where the component is omitted from the step plan prompt entirely. The results are shown in Table 4.8.

In general, the results reveal a critical insight: user state is the most crucial component for planner alignment. Excluding user state was the only condition to cause a significant decrease in effectiveness (3.32 vs.3.49) and a corresponding increase in Action Uncertainty. This indicates that the evaluator perceives action planning as severely misaligned when the user's state is not considered. For example, the LLM

Table 4.7: Evaluation on the step plan between original and altered user states. The rows compare the "default" and "plain" prompts, split by whether the user state was "original" or "altered" (simulating a user with difficulties). The columns show the scores for Effectiveness and the three Uncertainty metrics. Compare "original" and "altered" rows to see how user state impacts the scores.

Prompt	**User state**	**Eff**	**OU**	**AU**	**IU**
default	original	4.47	2.28	2.06	2.95
	altered	**4.56***	**2.41***	2.10	3.01
plain	original	4.29	2.45	2.38	3.20
	altered	**4.33***	**2.53***	2.39	3.28

Values in bold indicate a significant difference from the original user state with corrected p: * <0.05.

evaluator notes: "The user stated they are unfamiliar with the task and want to start from the beginning. However, the provided actions—such as visiting shortlisted neighborhoods and attending open houses—are for a much later stage in the home-buying process. These instructions are completely misaligned with the user's current state, making the action plan impractical and impossible to follow..."

In this specific case, the evaluator's effectiveness score dropped to 1 (Not Effective) and AU rose to 5 (Highly Ambiguous). By contrast, skipping communication (type 1) showed no significant difference from the default. This suggests that simply skipping the execution of the communication step may not directly degrade the quality of the action instructions, if collecting user information is still considered in the plan.

Counter-intuitively, all other ablation conditions, particularly the Type 2 ablations, yielded higher effectiveness and lower uncertainty scores than the default condition. This does not mean the plans are actually better or less uncertain. Instead, it indicates that the LLM evaluator becomes overly optimistic when constraints and contextual information is missing. By removing components designed to uncover user

Table 4.8: Evaluation on action instructions. The rows represent different ablation conditions (removing specific components like User State or Communication). The top row is the baseline (Default). The columns show the evaluation scores. Compare the lower rows against the top "Default" row to see how components affect the outcomes in terms of effectiveness and uncertainty metrics.

Condition	**Eff**	**OU**	**AU**	**IU**
default	3.49	2.38	2.11	2.60
excl. user state	**3.32****	2.44	**2.20***	2.62
excl. comm. type 1	3.55	2.34	2.05	2.57
excl. comm. type 2	**3.71****	**2.17****	**1.84****	**2.47****
excl. info. type 1	**3.68****	**2.30****	**1.98****	**2.55****
excl. info. type 2	**3.81****	**2.18****	**1.82****	**2.51****
excl. comm. info. type 1	**3.68****	**2.30****	**1.97****	2.54
excl. comm. info. type 2	**4.10****	**2.03****	**1.65****	**2.34****

Values in bold indicate a significant difference from the default value with corrected p: * 0.05, ** 0.01.

preferences and external information, the task appears simpler, and the evaluator rates the simplified (but less personalized) action plan more highly.

This effect is most pronounced when excluding both communication and information (type 2), which achieved the highest effectiveness (4.10) and the lowest uncertainties (e.g., AU 1.65). In this scenario, the step planner generates a simple, direct action plan without any preceding communication or information-gathering tasks. Then, the evaluator notes: "The action plan is optimal. It perfectly aligns with the user's state as a beginner... The suggested actions—...—are foundational, practical, and logically sequenced... Each step is concrete, has a clear purpose, and effectively reduces the initial friction and overwhelm for the user." Here, the evaluator fails to recognize the "unknown unknowns." Because no user constraints or preferences were gathered, the evaluator

sees a clean, simple plan and regards it "optimal," effectively overestimating its quality by overlooking the missing personalization. Another example of an explanation is "The actions provided are quite clear and specific - gathering financial documents... These are concrete, well-defined steps that someone unfamiliar with home buying can understand and execute... the actions are generally straightforward and manageable for a beginner." While the action instructions are similar in the default condition and the condition excluding communication and information components, the ablated condition received an uncertainty score of 2, while the default received 3, highlighting that "there might be uncertainties about exactly which documents to gather or how to find a good lender."

The increase in scores does not indicate that the scoring process is problematic, but rather that it reflects an "optimism bias" in the simulation: without explicit constraints forced by the communication and information components, the LLM defaults to an idealized, low-friction planning path. This finding supports H2 by demonstrating that the presence of these components is necessary to force the system to confront realistic task complexity, even if it results in lower (but more realistic) perceived effectiveness scores.

Overall, these results suggest that while the LLM evaluator is not perfect, it is able to detect meaningful differences across the ablation study. The findings confirm the critical role of user state for plan alignment and simultaneously reveal an important bias that was also observed in the interview study: the LLM may be overoptimistic and favor simpler, less-contextualized plans, misinterpreting their simplicity as effectiveness especially when no detailed constraints or contextual information are provided.

4.3 Long-term life task planning Benchmark

The experiments with the GOLF framework validate that simulating this complex, multi-agent human-AI collaboration for long-term tasks is viable. The results demonstrate that the framework is sensitive to variations in prompt quality and component architecture, confirming its utility as an analytical tool. Building on the demonstrated viability of the GOLF framework for simulating long-term planning, the dissertation sought to develop a method for systematically evaluating and comparing LLM performance on this task. To this end, it developed the GOLF benchmark, which transform the core planning logic of the GOLF framework with diverse topics into a reproducible evaluation dataset.

4.3.1 Benchmark development

The benchmark dataset covers ten major domains of long-term life planning, representing a diverse range of personal and professional goals. Examples of domains generated by gemini-2.5-pro are shown below, which are highly consistent in topic with those produced by claude-sonnet-4 and gpt-4o.

- **Financial Independence & Wealth Building**: Creating investment strategies, retirement planning, and debt elimination roadmaps
- **Career Transformation & Skill Development**: Mapping career pivots, identifying skill gaps, and creating professional growth timelines
- **Health & Longevity Optimization**: Designing sustainable fitness routines, nutrition plans, and preventive healthcare schedules
- **Relationship & Social Network Cultivation**: Building deeper connections, improving communication skills, and expanding meaningful relationships

- **Creative Expression & Artistic Pursuits**: Developing artistic talents, completing creative projects, and building portfolios or performances
- **Educational Advancement & Lifelong Learning**: Pursuing degrees, certifications, or mastering new subjects through structured learning paths
- **Geographic & Lifestyle Transitions**: Planning relocations, travel goals, or significant lifestyle changes like minimalism or sustainable living
- **Legacy & Impact Creation**: Establishing charitable foundations, mentoring programs, or community contributions that outlast your lifetime
- **Spiritual & Philosophical Growth**: Exploring belief systems, meditation practices, or philosophical frameworks for deeper meaning
- **Adventure & Experience Collection**: Systematically pursuing bucket list items, extreme sports, cultural immersion, or unique life experiences

Each domain contains ten specific task topics, and each topic is decomposed into multiple steps. After generating all task instances and filtering invalid or incomplete entries, the final benchmark consists of over 3,000 step-level samples. Each generation model (gemini-2.5-pro, claude-sonnet-4, and gpt-4o) contributes approximately one-third of the dataset (around 1,000 samples per model), ensuring balanced representation across source models and domains.

4.3.1.1 Human and LLM annotations

A subset of 450 benchmark samples was selected for annotation, with 150 tasks randomly drawn from each generation model (gemini-2.5-pro, claude-sonnet-4, and gpt-4o) across different domains, topics, and steps. Each task was independently annotated by

three human annotators and three LLM annotators, resulting in 1350 human and 1350 LLM annotations in total.

For the human annotation, the analysis only included data from participants who successfully passed the attention check. The pass rate was approximately 90%, indicating that the vast majority of participants were actively engaged. Despite this high level of engagement, human accuracy was notably low (Table 4.9). Both state-specific and overall accuracies averaged around 0.3, which is close to the random-guess baseline of 0.33 for a three-choice task. However, the inter-annotator agreement rate exceeded 0.6, indicating a moderate level of consensus among humans. The contrast of high agreement high attention-check pass rates and low accuracy against the ground truth suggests that the low scores were not due to a lack of effort or misunderstanding of the interface. Instead, it implies a fundamental misalignment between human intuition and the simulated logic of the LLM agents. While humans could agree on a "logical" answer, their reasoning did not match the specific state-action mappings generated by the GOLF framework.

In contrast, LLM annotators demonstrated markedly higher annotation accuracy (Table 4.9). First, the overall accuracy for LLMs approached or exceeded 0.6, substantially higher than human accuracy, suggesting that LLMs can effectively perceive user state differences and assign appropriate action candidates. Second, permutation accuracy values approached or surpassed 0.5, indicating that in roughly half of the tasks, models selected the fully correct user–action correspondence. Third, consistent differences were observed across models: gemini-2.5-pro achieved the highest accuracy, followed by gpt-4o, with claude-sonnet-4 performing comparatively lower.

To further examine potential systematic differences between generation and evaluation models, a cross-model comparison was conducted using subsets of the benchmark generated by each model (Table 4.10). Each column represents benchmark data

Table 4.9: Annotation on a sample of benchmark data. The rows list the accuracy metrics (e.g., accuracy for specific state and overall accuracy). The columns represent the different annotators: humans and the three LLMs. The overall accuracy and state-specific accuracy are at the similar level, and LLM annotators have higher accuracy than human annotators.

Metric	**human**	**claude-sonnet-4**	**gemini-2.5-pro**	**gpt-4o**
proceed_accuracy	0.316 ± 0.466	0.596 ± 0.491	0.653 ± 0.476	0.636 ± 0.482
update_accuracy	0.307 ± 0.462	0.564 ± 0.496	0.640 ± 0.481	0.607 ± 0.489
revise_accuracy	0.351 ± 0.478	0.558 ± 0.497	0.667 ± 0.472	0.618 ± 0.486
overall_accuracy	0.324 ± 0.335	0.573 ± 0.388	0.653 ± 0.396	0.620 ± 0.397
perm_accuracy	0.160 ± 0.367	0.431 ± 0.496	0.549 ± 0.498	0.502 ± 0.501
agreement_rate	0.631 ± 0.203	-	-	-
majority_accuracy	0.351 ± 0.478	-	-	-
unanimous_accuracy	0.044 ± 0.207	-	-	-

generated by a specific model, and each row represents a different model used for evaluation. The results show that benchmarks generated by claude-sonnet-4 achieved the highest average evaluation accuracy across evaluators, whereas those generated by gpt-4o yielded the lowest. Across evaluation models, gemini-2.5-pro mostly achieved the highest accuracy (except on gpt-4o–generated data), while claude-sonnet-4 remained the weakest. These findings indicate systematic variation between generation and evaluation models and suggest that gpt-4o may exhibit a mild bias toward its own generated content, performing relatively better on its own benchmark portion. Nonetheless, incorporating benchmark subsets produced by multiple generation models enhances diversity and helps mitigate model-specific bias, thereby improving the robustness and fairness of evaluation.

These results suggest that the tasks present considerable difficulty for human annotators, possibly due to the inherent complexity of long-term task planning or partial misalignment between the benchmark's simulated logic and human intuition. Nevertheless, the fact that LLMs can reliably distinguish user states and exhibit consistent

Table 4.10: Model comparison based on the sample of benchmark data. The rows represent the model used as the evaluator. The columns represent the model used to generate the benchmark data. The cells show the accuracy score. Read across a row to see how one evaluator judges different generators; read down a column to see how one generator is judged by different evaluators.

gen model / eval model	claude-sonnet-4	gemini-2.5-pro	gpt-4o
claude-sonnet-4	0.664 ± 0.375	0.558 ± 0.377	0.496 ± 0.395
gemini-2.5-pro	0.762 ± 0.369	0.658 ± 0.384	0.540 ± 0.406
gpt-4o	0.676 ± 0.378	0.600 ± 0.396	0.584 ± 0.412

performance stratification across models supports the benchmark's validity as a diagnostic tool for evaluating reasoning and decision-making in long-term task planning contexts.

4.3.2 Full benchmark evaluation across models

Figure 4.2 visualizes model performance on the full benchmark, showing both overall accuracy and permutation accuracy, sorted by the mean permutation accuracy. The overall accuracy ranges from 0.34 to 0.67, while permutation accuracy ranges from 0.14 to 0.56, indicating substantial variability in model capability. State-specific accuracies are omitted here because their trends closely mirror overall accuracy.

The ranking of models largely reflects their size and generation. The top-performing models include gemini-2.5-pro, gpt-5, gpt-4.1, and gemini-2.5-flash, which performs unexpectedly well, ranking second overall despite being positioned as a median model. Models in the lower half, such as gpt-4.1-mini, gpt-4-turbo, and gpt-5-nano, show reduced accuracy, and gpt-4.1-nano performs close to the human baseline. These results demonstrate that the benchmark effectively differentiates LLMs across capability levels and architectures.

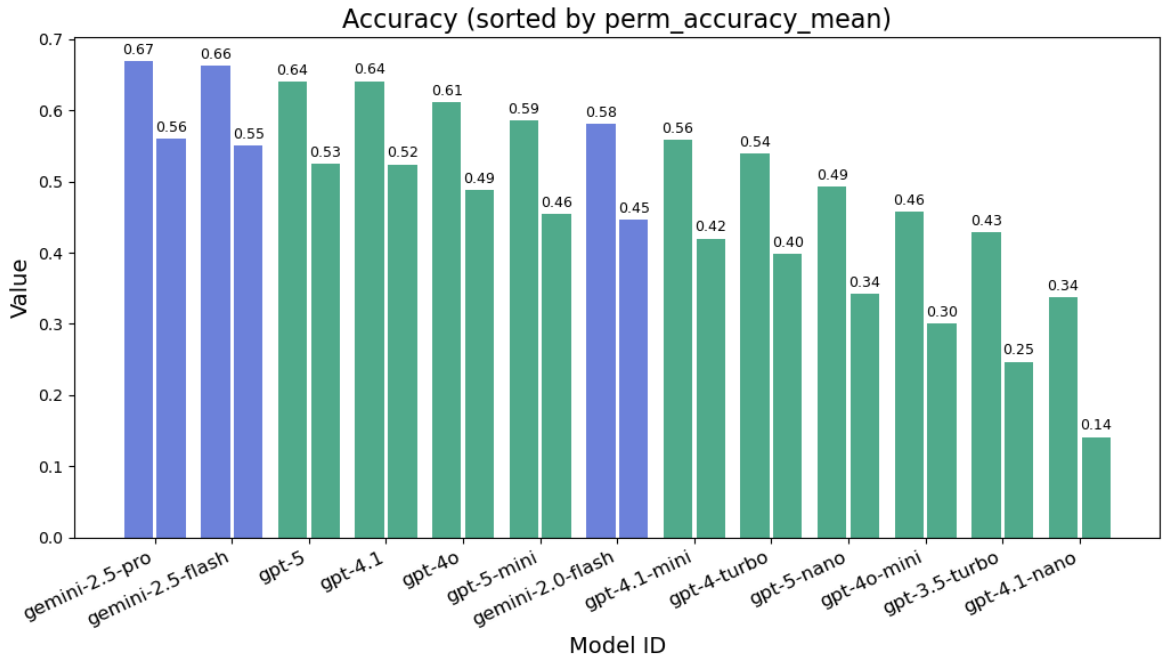

Figure 4.2: Benchmark scoreboard for the models. Each cluster of bars represents a specific LLM. The left bar is the Overall Accuracy, and the right bar is the Permutation Accuracy. The models are sorted from highest performance (left) to lowest (right).

To investigate the relationship between model capacity and planning performance, I analyzed the inference costs of each model. Given that exact parameter counts for proprietary models are not publicly disclosed, I utilize API pricing (USD per 1M input tokens) as a proxy for model size and computational complexity. Figure 4.3 presents these prices alongside the performance data. The comparison reveals that cost does not strictly correlate with performance, reflecting rapid optimizations in newer architectures. For example, within the same tier, newer models are significantly more cost-efficient: gpt-4-turbo remains the most expensive legacy model, while subsequent generations like gpt-4o and gpt-5 achieve higher performance at lower price points. Notably, gemini-2.5-flash demonstrates an exceptional efficiency-to-performance ratio, achieving near-top accuracy with costs comparable to smaller models.

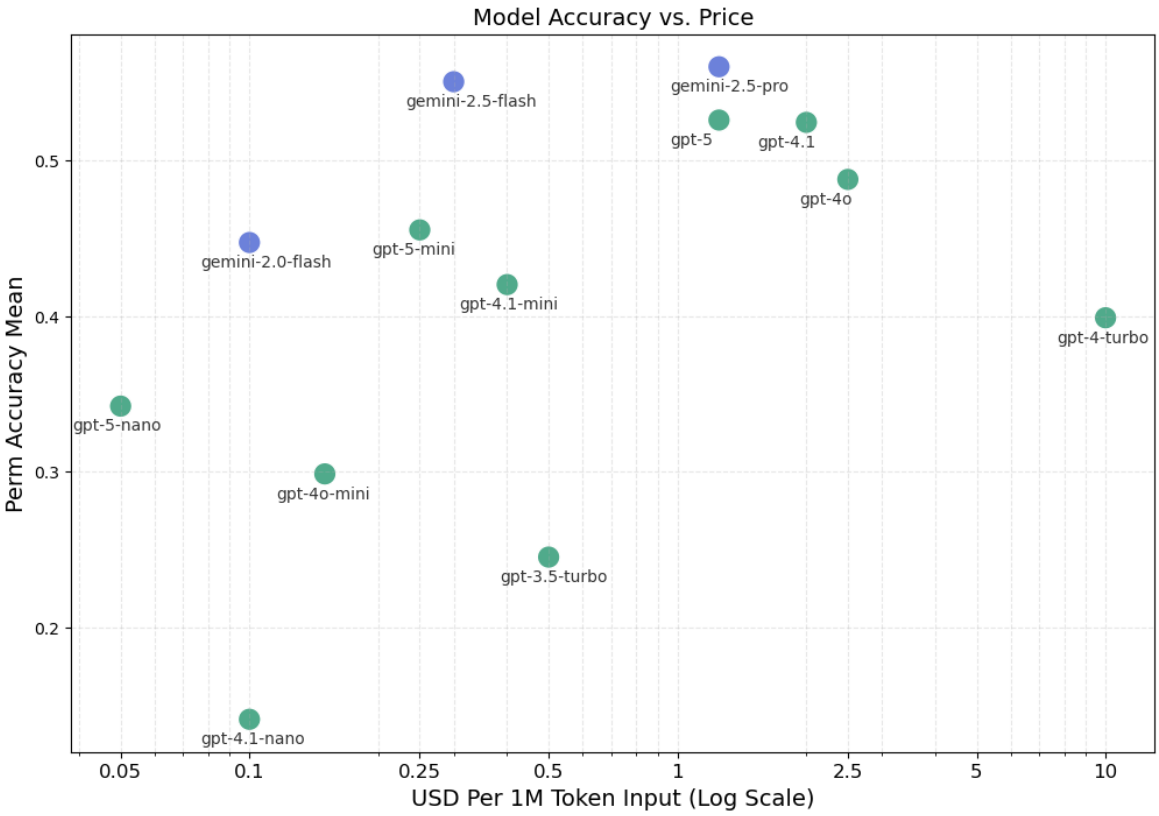

Figure 4.3: LLM API price (US dollar per 1 million tokens). The X-axis is the Price (log scale), and the Y-axis is the Permutation Accuracy. Models in the top-right offer high performance at a higher cost, while models in the bottom-left offer lower performance at a lower cost.

Overall, while the best-performing models achieve accuracies around 0.6, most small models still struggle on this benchmark. The strong and cost-effective performance of gemini-2.5-flash suggests potential for scalable deployment in long-term task planning agents. However, the generally moderate accuracy levels across all models indicate that long-term task planning remains a challenging domain, requiring continued advancement in LLM reasoning and adaptability.

Chapter 5

Discussion

5.1 Overall results

In the interview study for RQ1, participants used ChatGPT for diverse long-term planning tasks and developed prompting strategies to set goals, decompose steps, and refine outputs. Yet planning was marked by significant uncertainty: users were unsure how to start, which options were realistic, and whether AI advice could be trusted or followed through. While ChatGPT was valued for scaffolding, idea generation, and encouragement, participants also highlighted its lack of personalization, realism, adaptability, and credibility. They expressed a desire for systems that could remember context, dynamically adjust, and provide transparent, trustworthy guidance. Overall, ChatGPT functioned as a helpful scaffolder and reflective partner, but its current limitations constrained its effectiveness for long-term task planning.

This simulation study for RQ2 introduces a simulation framework for long-term life tasks where LLMs act as planner, user, and evaluator agents. The results demonstrate the feasibility and controllability of simulating long-term task planning. Subsequently, the study for RQ3 introduces a benchmark for evaluating LLMs in long-term task planning. The benchmark simulates multi-state user scenarios to test whether models can perceive contextual differences and select appropriate actions. While human annotators perform near random, LLMs achieve substantially higher and more consistent accuracy. The benchmark effectively distinguishes model capability and efficiency

across scales, revealing clear performance stratification. Overall, the findings indicate that current LLMs demonstrate the capability to identify distinct user states and adaptively adjust plans in long-term scenarios, though further progress is needed for robust, human-aligned decision-making..

5.2 Implications

5.2.1 The diversity of planning goals and needs

In the interview study, the research observed three main task categories in this study, including personal and well-being, activity and event planning, and professional development and learning. Participants adopted specific prompting behaviors aligned with the nature of their tasks, showing how people adapt their interaction style with AI to different planning needs. By examining these prompting patterns together with the inherent characteristics of each task type, the study finds that the three categories reflect distinct demands and expectations for AI assistance.

For personal and well-being tasks, participants preferred open-ended and flexible prompting and often resisted detailed or prescriptive plans. These goals, such as maintaining health, managing stress, or improving relationships, were integrated into everyday life and could be easily influenced by changing routines or emotions (Abbott, 2005; Kuhlthau, 1991). In contrast, activity and event planning tasks, like travel or event organization, had concrete deadlines and interdependent factors such as timing, logistics, and budget (Byström & Hansen, 2005; J. Xie et al., 2024). Participants sought specific and executable plans, iteratively refining ChatGPT's suggestions until the results appeared realistic and complete. Meanwhile, professional development and learning tasks were more exploratory and information-driven (Shah et al., 2023; Vakkari, 2003). Users

engaged ChatGPT to structure abstract ambitions, explore new domains, or build conceptual roadmaps rather than fixed action steps. Together, these patterns show how the temporal scope and uncertainty of a goal shape users' prompting strategies and expectations for AI assistance: goals with recurring tasks favor adaptive planning and general guidance; bounded tasks call for precise and front-loaded planning; and open-ended, knowledge-oriented goals rely on conceptual scaffolding and reflection (J. Liao et al., 2024; Norman & Shallice, 1986).

5.2.2 Uncertainties in planning with AI

Across task types, participants described uncertainty arising from two main sources: the inherent ambiguity of the task itself and the limitations introduced by the AI system. Task-inherent uncertainty included doubts about feasibility, knowing where to start, or balancing detail and flexibility, whereas AI-introduced uncertainty stemmed from issues such as lack of contextual fit, limited personalization, or questionable credibility. These two sources shaped participants' experiences differently. AI-introduced uncertainty primarily influenced how users perceived the generated plan, whether it seemed realistic, trustworthy, or personally relevant, while task-inherent uncertainty was tied to how they imagined carrying out the plan in the real world.

Taken together, these patterns suggest that users experienced uncertainty not only in different forms but also at different temporal stages of the planning process. AI-introduced uncertainty tended to affect the present interaction with ChatGPT, while task-inherent uncertainty extended into the future enactment of the plan. This temporal distinction resonates with *construal level theory* (Trope & Liberman, 2010), which posits that people think about distant future goals in abstract terms and near-term

actions in concrete terms. In the interview, participants' reflections mirrored this temporal gradient: uncertainty can shift from abstract questions about how to plan toward more situated concerns about how to act (Gollwitzer, 1999).

To interpret this distinction, we can tentatively describe two interrelated dimensions of uncertainty: *planning uncertainty* (i.e., ambiguity about how to define and structure a plan) and *doing uncertainty* (i.e., apprehension that arises when imagining the actual execution of that plan). ChatGPT was effective at easing the former by providing structure, generating ideas, and clarifying options, but less capable of addressing the latter, which was emotional and context-dependent. This distinction helps explain why participants' preferences for plan specificity diverged across task types: those engaged in exploratory goals (e.g., professional learning) were comfortable with general plans that preserved flexibility; those managing evolving conditions (e.g., health or relationships) found detailed plans restrictive or unrealistic; and those pursuing bounded, short-term tasks (e.g., travel or events) desired specific, executable guidance.

5.2.3 Design implications

The findings from the interview study suggest that AI plays two complementary roles in long-term planning: as a cognitive scaffold and as a reflective partner. As a scaffold, ChatGPT helps users externalize and structure complex goals, break them into actionable steps, and maintain a sense of progress. Scaffolding can support cognitive organization and task success (Câmara et al., 2021; Vygotskij & Cole, 1981; Zhou et al., 2025). As a reflective partner, it sustains engagement by offering conversational feedback, emotional reassurance, and adaptive suggestions that help users manage uncertainty and motivation over time (Miceli & Castelfranchi, 2005). However, these same strengths also expose important limitations. ChatGPT's encouraging tone and

success-oriented framing often led to overly optimistic or flattering responses, depicting idealized futures while overlooking potential setbacks, contextual constraints, or emotional challenges (Sivaprasad et al., 2025; J. Zhang et al., 2025). By consistently confirming users' goals and emphasizing positive outcomes, the system sometimes obscured signals of possible failure and fostered overconfidence in the feasibility of its plans (Carro, 2024; Cheng et al., 2025). This dynamic highlights a central design tension: while scaffolding and encouragement can enhance motivation and perceived clarity, they may also mask the inherent uncertainty of real-world action.

Taken together, these findings point to several design implications for AI-assisted planning tools. While current LLMs possess the technical capacity to process extensive context within a single prompt, the cognitive burden and inherent uncertainty of long-term tasks make it impractical for users to articulate every constraint and preference upfront. Consequently, systems should support personalization and user modeling by capturing context progressively as the task evolves, rather than requiring exhaustive initial input. They should enhance explainability and transparency, providing sources, confidence levels, and explicit reasoning pathways. To better manage uncertainty, AI should move beyond success-oriented planning and incorporate mechanisms for simulating obstacles, surfacing common points of failure, and suggesting contingency plans. Moreover, tools should explicitly support iterative and adaptive workflows, remembering user preferences across sessions and dynamically adjusting as users report progress or difficulties. Motivational support also remains important, but it should be configurable, allowing users to switch between encouragement and flattery. Finally, privacy and sensitivity issues must be carefully handled, especially as personalization and cross-session memory become central to long-term planning.

5.2.4 Simulation aligning with human planning tendencies

The GOLF simulation framework, though simplified, operationalizes several key planning behaviors and needs identified in the RQ1 interview study. In particular, it captures how people use AI for cognitive scaffolding, multi-faceted task management, and uncertainty navigation, while also revealing where current simulation methods fall short of human complexity.

A central finding from the interview study was that users valued AI for its ability to provide scaffolding for transforming abstract, long-term goals into structured and manageable plans. The simulation's initial plan directly mirrors this behavior. It first produces a high-level roadmap of activities and tasks. This top-down decomposition provides the cognitive structure users in the interview study described as essential for overcoming initial uncertainty and beginning the planning process. In this way, the simulation reproduces the procedural clarity that participants associated with reduced ambiguity and a stronger sense of direction.

In addition, the interviews revealed that real-world planning rarely unfolds as a single, linear activity but instead involves multiple intertwined subtasks. Therefore, stating preferences, seeking and verifying information, and translating goals into concrete actions are considered in the simulation. The GOLF framework explicitly models this diversity by decomposing each planning step into three specialized agentic components: communication, information retrieval, and action instruction. The communication component (e.g., asking about budget or neighborhood preferences) parallels how participants in the interview study clarified context and expressed personal constraints. The information retrieval component (e.g., searching for mortgage processes) reflects users' behavior of seeking and verifying external knowledge to ground their plans in credible information. Finally, the action instruction component (e.g., "Gather financial documents") mirrors participants' desire for actionable and clear steps that

convert abstract goals into tangible progress. The user agent can simulate the participant's problems or capability to complete the steps. By involving the user agent in the planning simulation, this design of supportive and adaptive system also illuminate how to help people plan without replacing their judgment or reinforcing human cognitive biases. Together, these components emulate the interplay between reflection, inquiry, and execution observed in human-AI planning interactions.

Regarding the benchmark, the expected validation results can be categorized into four scenarios: high accuracy/agreement for both human and LLM annotators (indicating the benchmark is valid for both), low accuracy/agreement for both (indicating an invalid benchmark), or a divergence where one agent outperforms the other. The observed result where LLMs reliably distinguish user states while humans perform near chance suggests that the benchmark is valid within the model's internal logic but lacks alignment with human perception. While random guessing by human annotators is a potential explanation for the low scores, the attention checks integrated into the study (which required reading and manual drag-and-drop actions) serve to mitigate this concern, confirming that participants were visually and cognitively engaged with the text.

Consequently, the chance-level human performance likely stems not from random guessing but a misalignment between the simulation and human perception. Humans may find the text-heavy descriptions of user states and action plans indistinguishable or may require a different planning logic than the simulation. Regardless of whether this is due to cognitive load or different reasoning processes, the result confirms a lack of human-AI alignment. In addition, the fact that even the best-performing LLMs achieved an accuracy of only around 0.6 indicates that the generated planning states possess inherent ambiguity even for the models themselves. Therefore, these evaluation scores should not be interpreted as a measure of ground-truth performance, but

rather as a metric of the model's internal consistency (its ability to distinguish its own generated states) and the degree of human-AI misalignment (the gap between model logic and human perception). This underscores that while current LLMs can maintain a coherent internal narrative for long-term planning, there are gaps between LLM reasoning processes and human cognitive strategies of long-term task planning.

5.2.5 Developing realistic long-term life task simulation

The findings from the simulation experiments provide several insights. Prompt quality with planning skills significantly improved planning effectiveness and reduce uncertainties, and the effects propagate across steps through updated user state. The ablation results reveal how the LLM evaluator interprets task planning under reduced context. When communication and information are missing, it assigns higher scores to the action instructions with less contextual information, appearing optimistic because no user constraints are present to the evaluator. Conversely, removing user state explicitly reduced the quality of action instructions, especially when the evaluator perceives the user state while the action instructions are not aligned with it. These findings highlight both the strengths and biases of simulation, underscoring the need for clear user feedback and evaluations that better penalize blurred task boundaries (J. Gu et al., 2024; H. Li et al., 2024).

The framework also illustrates how task and user simulation themselves contribute to long-term planning. Real users rarely plan in a strictly step-by-step manner, so previewing trajectories can provide higher-level orientation and reduce uncertainty before committing to actions (Chowdhury et al., 2011, 2014; Höchli et al., 2018). The simplified simulation captures this aspect, showing how LLMs can approximate long-term planning as an information process (Toms, 2019). While full real-time interaction will require further engineering and system integration, this study establishes an initial

foundation. Looking ahead, the framework offers potential for developing benchmarks for long-term task planning and for incorporating more prompting techniques, real-time search, and retrieval augmented generation (RAG) (Lewis et al., 2020; Wu et al., 2023). These extensions can support more realistic and comprehensive implementation and evaluations of human–AI collaboration.

The benchmark results provide several important implications for understanding LLM reasoning in long-term task planning. First, the large performance gap between humans and LLMs suggests that structured simulation-based benchmarks can reveal reasoning competencies that are not immediately intuitive to human annotators. This indicates that LLMs have developed an emergent ability to interpret user state cues and map them to corresponding action strategies in a consistent, rule-based manner. Second, the stratified performance across model families demonstrates that long-term reasoning scales predictably with model capacity and optimization, supporting the notion that larger and more recent models internalize more stable procedural reasoning frameworks. Finally, the efficiency–performance trade-off observed in models like gemini-2.5-flash highlights the feasibility of cost-effective yet capable models for real-world deployment in planning-oriented AI agents.

5.3 Limitations and future directions

This study has several limitations that should be considered when interpreting the findings. As an exploratory qualitative study with a small sample, the results reflect the perspectives of participants who were already motivated to experiment with AI, which may limit generalizability. The study also examined planning scenarios as examples and relied on self-reported reflections rather than observing how AI-assisted plans in real practice. Moreover, the findings represent a snapshot of user experiences with

ChatGPT at its current level of development. Given the fast-paced development of LLMs and AI systems, the observed limitations and user behaviors may change as models evolve. Nevertheless, the study offers a grounded understanding of how people engage with AI for long-term life task planning, revealing patterns of use, forms of uncertainty, and the role of AI. These insights inform future directions for research and design.

While the GOLF framework aligns well with procedural scaffolding, it only partially captures the cognitive and emotional dimensions of uncertainty observed in the interviews. The task-inherent uncertainty carried an affective component, manifesting as anxiety, frustration, or diminished confidence. The model does not yet simulate how emotional states, such as anxiety or hesitation, dynamically influence user behavior, for example by leading users to distrust AI suggestions, avoid complex steps, or simplify plans to reduce cognitive load. These coping mechanisms were central to participants' experiences in the interview study. Consequently, the simulation aligns closely with the procedural aspects of human planning but diverges from its affective and adaptive richness, highlighting an important direction for modeling human uncertainty in long-term task planning.

A key limitation of the GOLF framework lies in its formalization of uncertainty. This study intentionally defines uncertainty as a perceived cognitive construct, drawing insights from the qualitative findings in the interview study. Consequently, the simulation evaluates uncertainty through LLM-as-judge annotations rather than a formal mathematical or probabilistic expression. I acknowledge that this approach does not capture the full computational complexity of uncertainty (e.g., Bayesian inference or POMDPs, as seen in traditional AI planning). However, the objective was not to create a mathematically complete model, but rather to establish a baseline for computationally simulating user-perceived uncertainty in the novel context of long-term life tasks.

The results demonstrate that LLMs can, to an extent, consistently differentiate and evaluate these simulated cognitive states. This simulation of perceived uncertainty provides a foundation for future research to develop more sophisticated hybrid models that bridge human cognitive states with robust computational frameworks for uncertainty management.

A limitation of the current benchmark is the lack of ground truth derived from expert human planners. The current results demonstrate that the framework achieves internal consistency that the LLM agents can generate and recognize distinct planning and user states. However, the divergence between human and LLM annotation scores suggests that the model's internal logic does not yet fully map to human intuition. Future work should focus on curating a golden dataset of expert-verified planning steps to calibrate the LLM evaluator, moving from measuring consistency to measuring alignment with reality.

Furthermore, The benchmark relies on synthetic data generated through LLM simulation, which may not fully capture the complexity, ambiguity, or open-ended nature of real human planning processes. The user states represent simplified states of progress, while real-world interactions often involve overlapping or nuanced motivational factors. Finally, all evaluations were conducted in static, single-turn conditions, without considering dynamic feedback or multi-step planning cycles. Overall, the simulation and benchmark provide an initial yet robust foundation for evaluating structured reasoning in long-term task planning. By highlighting both the potential and current limitations of LLMs in this domain, it lays the groundwork for developing more cognitively grounded, context-aware, and adaptive planning agents in future work.

Future work on the user study could extend this direction by examining more diverse user groups and longitudinal settings, capturing how AI-assisted plans unfold over time and how uncertainty changes across different task horizons. Design-oriented

studies may also explore systems that better account for failure and change, support shared or multi-task planning, and incorporate adaptive memory, transparency, and feedback mechanisms. Advancing toward resilient and trustworthy planning support will require positioning AI not merely as a plan generator but as a collaborative and context-aware partner that evolves alongside users' goals. Regarding the simulation and benchmark, future work will focus on extending the benchmark toward more realistic and dynamic evaluation. Incorporating multi-step and temporally linked tasks can better test sustained reasoning and adaptability. Human-in-the-loop evaluation should also be introduced to align benchmark logic with human interpretation. In addition, exploring multimodal or grounded task settings could enhance ecological validity and practical relevance for long-term planning agents.

Chapter 6

Conclusion

6.1 The goals of the dissertation

This dissertation aimed to address the core challenges faced when using LLMs to support humans in long-term life task planning. Specifically, it sought to achieve three goals: (1) to deeply investigate the how people use LLMs in long-term life tasks under uncertainty and how AI tools might assist this process; (2) to propose an effective human-AI collaboration framework and validate its effectiveness in long-term life task planning through simulation; and (3) to develop a benchmark for long-term life task planning to evaluate the performance of LLMs on this emerging task type.

6.2 To what extent have the goals been met

This dissertation presents a comprehensive study that covers the process from understanding user needs (the user interviews for RQ1), to building and validating a systematic planning framework—the GOLF framework (the simulation study for RQ2), and finally to developing a reproducible and scalable evaluation method—the GOLF benchmark (RQ3). The work combines the cognitive strategies of human task planning with computational models based on LLMs, demonstrating an innovative analytical path.

This path is not only theoretically significant for human-AI collaboration but also provides a feasible solution for building more adaptive AI planning support systems in practice.

6.3 Broader contributions to human-collaboration research

Within the broader context of HCI and AI research, this dissertation work pushes the boundary of knowledge regarding AI's role in assisting humans with open-ended, unstructured, and long-cycle tasks. Theoretically, by introducing the "long-term life task" as a new task type and proposing the GOLF framework, it opens up new perspectives for studying human-AI collaboration for managing uncertainty and long-term planning. Methodologically, it demonstrates how multi-agent simulation and LLM benchmarks can be used to systematically and controllably evaluate LLM capabilities in complex planning scenarios. This work encourages researchers to explore how insights from small-scale user studies can be more effectively applied to system design, simulation experiments, and ultimately the construction and evaluation of adaptive AI systems.

By integrating insights from cognitive science, information science, and AI, this work helps to bridge the gap between human cognitive strategies and AI planning capabilities, providing a robust foundation for future advancements in AI-assisted task planning as a new type of human-AI collaboration.

6.4 Implications and challenges for practical applications

Going beyond the research context, systems that utilize the findings from the dissertation work may generate a series of potential positive impacts on people's decisions in both work and daily life. For example, an AI assistant based on the GOLF framework could move beyond simple Q&A to become a long-term "thinking partner" for users in major life decisions such as financial planning, career development, and health management. It could provide dynamic, personalized planning suggestions based on the user's state and preferences, helping them break down complex goals and thereby reducing cognitive load and decision-making anxiety.

However, in practical applications, this study also recognize several additional challenges. In addition to the technical challenges widely discussed (e.g., model hallucinations, context dependency), a major obstacle is the "reality gap:" planning in a simulated environment is structured, whereas real life is filled with unpredictable variables and complex emotional factors. Enabling an AI system to understand and adapt to this discrepancy is a formidable challenge. Another challenge is the issue of user over-reliance. If users depend entirely on AI for planning, it might weaken their autonomous decision-making abilities and resilience in the face of unexpected events. Therefore, designing a system that provides robust support while also encouraging critical thinking from the user is of paramount importance.

6.5 Final thoughts

In summary, the research questions in this dissertation are driven by two core objectives: 1) to enhance the knowledge of the dynamic process of human long-term planning

under uncertainty; and 2) to leverage this knowledge to build and evaluate an AI system capable of providing adaptive support to users at different planning stages. Through achieving these goals, this dissertation also contributes to the integration of knowledge regarding the dynamic aspects of complex tasks (i.e., evolving user states) with formal AI models and computational simulations.

However, it is worth noting that the GOLF framework and the associated benchmark proposed are still far from complete in terms of articulating task characteristics and dynamically supporting users in real time. To better understand how different types of complex life tasks unfold in human-AI interactions and (more importantly) why they manifest in certain ways, it is necessary to overcome the challenges and limitations discussed above and continue explorations on both the human-centered and computational aspects in future work. A more complete development of state-aware, adaptive AI planning systems would require: (1) a more psychologically realistic model of the task process built upon a deeper knowledge of users' cognitive abilities, knowledge states, emotions, and even systematic biases; (2) a more comprehensive planning system that incorporates more task features and user traits as parameters; and (3) a set of evaluation metrics that are more closely associated with the user's long-term and holistic experience. This dissertation provides the essential first steps for this future. It provides a AI system framework that builds decision confidence and a concrete pathway to digital resilience, building a robust foundation for designing future AI systems that serve as true partners in enhancing human agency in an uncertain world.

Bibliography

Abbott, J. (2005). Understanding and Managing the Unknown: The Nature of Uncertainty in Planning. *Journal of Planning Education and Research*, *24*(3), 237–251. https://doi.org/10.1177/0739456X04267710

Adewale, N. T., Mansor, Y., Yusuf, M.-B. O., & Onikosi, A. (2017). Uncertainty and subjective task complexity in the information-seeking behaviour of lawyers: A structural invariance analysis. *Library Review*, *66*(4/5), 266–281. https://doi.org/10.1108/LR-09-2016-0079

Alchian, A. A. (1950). Uncertainty, Evolution, and Economic Theory. *Journal of Political Economy*, *58*(3), 211–221. https://doi.org/10.1086/256940

Allen, D. (2015). *Getting things done: The art of stress-free productivity* (Revised edition). Penguin Books.

Allmendinger, P. (2017). *Planning theory* (3rd edition). Macmillan Education, Palgrave.

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, *61*(5), 859–868. https://doi.org/10.1002/asi.21300

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*(1), 35–44. https://doi.org/10.1037/0003-066X.48.1.35

Arbona, C., Fan, W., Phang, A., Olvera, N., & Dios, M. (2021). Intolerance of Uncertainty, Anxiety, and Career Indecision: A Mediation Model. *Journal of Career Assessment*, *29*(4), 699–716. https://doi.org/10.1177/10690727211002564

Arguello, J. (2014). Predicting Search Task Difficulty. In M. De Rijke, T. Kenter, A. P. De Vries, C. Zhai, F. De Jong, K. Radinsky, & K. Hofmann (Eds.), *Advances in Information Retrieval* (pp. 88–99, Vol. 8416). Springer International Publishing. https://doi.org/10.1007/978-3-319-06028-6_8

Azzopardi, L. (2021). Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 27–37. https://doi.org/10.1145/3406522.3446023

Bailey, E. (2017). *Measuring Online Search Expertise* [Ph.D.]. The University of North Carolina at Chapel Hill. Retrieved December 23, 2024, from https://www.proquest.com/docview/1952045722/abstract/AD7D08CCA43D4AB0PQ/1

Barcaui, A., & Monat, A. (2023). Who is better in project planning?Generative artificial intelligence or project managers? *Project Leadership and Society*, *4*, 100101. https://doi.org/10.1016/j.plas.2023.100101

Bateman, T. S., & Barry, B. (2012). Masters of the long haul: Pursuing long-term work goals. *Journal of Organizational Behavior*, *33*(7), 984–1006. https://doi.org/10.1002/job.1778

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, *5*(1), 133–143. https://tefkos.comminfo.rutgers.edu/Courses/612/Articles/BelkinAnomolous.pdf

Bellotti, V., Dalal, B., Good, N., Flynn, P., Bobrow, D. G., & Ducheneaut, N. (2004). What a to-do: Studies of task management towards the design of a personal task list manager. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 735–742. https://doi.org/10.1145/985692.985785

Bhatia, S., Majumdar, D., & Aggarwal, N. (2016). Proactive Information Retrieval: Anticipating Users' Information Need. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 874–877). https://doi.org/10.1007/978-3-319-30671-1_84

Bi, K., Ai, Q., & Croft, W. B. (2021). Asking Clarifying Questions Based on Negative Feedback in Conversational Search. *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 157–166. https://doi.org/10.1145/3471158.3472232

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Byström, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information science and Technology*, *56*(10), 1050–1061. https://doi.org/10.1002/asi.20197

Câmara, A., Roy, N., Maxwell, D., & Hauff, C. (2021). Searching to Learn with Instructional Scaffolding. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 209–218. https://doi.org/10.1145/3406522.3446012

Cantor, N., Norem, J., Langston, C., Zirkel, S., Fleeson, W., & Cook-Flannagan, C. (1991). Life Tasks and Daily Life Experience. *Journal of Personality*, *59*(3), 425–451. https://doi.org/10.1111/j.1467-6494.1991.tb00255.x

Carleton, R. N., Norton, M. P. J., & Asmundson, G. J. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders*, *21*(1), 105–117. https://doi.org/10.1016/j.janxdis.2006.03.014

Carro, M. V. (2024, December). Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model. https://doi.org/10.48550/arXiv.2412.02802

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 3641289. https://doi.org/10.1145/3641289

Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025, September). *How People Use ChatGPT* (tech. rep. No. w34255). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w34255

Chen, L., & Zeng, S. (2021). The Relationship Between Intolerance of Uncertainty and Employment Anxiety of Graduates During COVID-19: The Moderating Role of Career Planning. *Frontiers in Psychology*, *12*, 694785. https://doi.org/10.3389/fpsyg.2021.694785

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating Large Language Models Trained on Code.

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., & Jurafsky, D. (2025, September). ELEPHANT: Measuring and understanding social sycophancy in LLMs. https://doi.org/10.48550/arXiv.2505.13995

Chi, M. T., Glaser, R., & Farr, M. J. (Eds.). (2014). *The Nature of Expertise* (1st ed.). Psychology Press. https://doi.org/10.4324/9781315799681

Chowdhury, S., Gibb, F., & Landoni, M. (2011). Uncertainty in information seeking and retrieval: A study in an academic environment. *Information Processing & Management*, *47*(2), 157–175. https://doi.org/10.1016/j.ipm.2010.09.006

Chowdhury, S., Gibb, F., & Landoni, M. (2014). A model of uncertainty and its relation to information seeking and retrieval (IS&R). *Journal of Documentation*, *70*(4), 575–604. https://doi.org/10.1108/JD-05-2013-0060

Christensen, K. S. (1985). Coping with Uncertainty in Planning. *Journal of the American Planning Association*, *51*(1), 63–73.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cohen, J. (2016). A power primer. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research (4th ed.)*. (pp. 279–284). American Psychological Association. https://doi.org/10.1037/14805-018

Cole, M., Liu, J., Belkin, N. J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. *Proceedings of the third workshop on human-computer interaction and information retrieval Cambridge*, 1–4.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347–362. https://doi.org/10.1037/0022-0663.79.4.347

Dalio, R. (2019). *Principles for Success*. Simon; Schuster.

De Zarzà, I., De Curtò, J., Roig, G., & Calafate, C. T. (2023). Optimized Financial Planning: Integrating Individual and Cooperative Budgeting Models with LLM Recommendations. *AI*, *5*(1), 91–114. https://doi.org/10.3390/ai5010006

Deng, Y., Liao, L., Zheng, Z., Yang, G. H., & 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*, 807–818. https://doi.org/10.1145/3626772.3657843

Dweck, C. S. (2006). *Mindset: The new psychology of success* (1st ed). Random House.

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12. https://doi.org/10.1016/j.compedu.2018.08.004

Epstein, D. A., Ping, A., Fogarty, J., & Munson, S. A. (2015). A lived informatics model of personal informatics. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 731–742. https://doi.org/10.1145/2750858.2804250

Farshidi, S., Rezaee, K., Mazaheri, S., Rahimi, A. H., Dadashzadeh, A., Ziabakhsh, M., Eskandari, S., & Jansen, S. (2024). Understanding user intent modeling for conversational recommender systems: A systematic literature review. *User Modeling and User-Adapted Interaction*, *34*(5), 1643–1706. https://doi.org/10.1007/s11257-024-09398-x

Fayol, H. (2016). *General and Industrial Management*. Ravenio Books.

Feild, H. A., Allan, J., & Jones, R. (2010). Predicting searcher frustration. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 34–41. https://doi.org/10.1145/1835449.1835458

Feng, L., Yen, R., You, Y., Fan, M., Zhao, J., & Lu, Z. (2024). CoPrompt: Supporting Prompt Sharing and Referring in Collaborative Natural Language Programming. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3613904.3642212

Fishbach, A., Dhar, R., & Zhang, Y. (2006). Subgoals as substitutes or complements: The role of goal accessibility. *Journal of Personality and Social Psychology*, *91*(2), 232–242. https://doi.org/10.1037/0022-3514.91.2.232

Ford, B. J. (1995). Information Literacy as a Barrier. *IFLA Journal*, *21*(2), 99–101. https://doi.org/10.1177/034003529502100206

Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M. (2025, March). Evaluating Human-AI Collaboration: A Review and Methodological Framework. https://doi.org/10.48550/arXiv.2407.19098

Fruh, S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long-term weight management. *Journal of the American Association of Nurse Practitioners*, *29*(S1), S3–S14. https://doi.org/10.1002/2327-6924.12510

Gibson, D., Kovanovic, V., Ifenthaler, D., Dexter, S., & Feng, S. (2023). Learning theories for artificial intelligence promoting learning processes. *British Journal of Educational Technology*, *54*(5), 1125–1146. https://doi.org/10.1111/bjet.13341

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, *6*, 1521066. https://doi.org/10.3389/fcomp.2024.1521066

Grant, A. M. (2003). THE IMPACT OF LIFE COACHING ON GOAL ATTAINMENT, METACOGNITION AND MENTAL HEALTH. *Social Behavior and Personality: an international journal*, *31*(3), 253–263. https://doi.org/10.2224/sbp.2003.31.3.253

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, Y., & Guo, J. (2024, November). A Survey on LLM-as-a-Judge. https://doi.org/10.48550/arXiv.2411.15594

Gu, Y., Gu, S., Lei, Y., & Li, H. (2020). From Uncertainty to Anxiety: How Uncertainty Fuels Anxiety in a Process Mediated by Intolerance of Uncertainty (F. Pan, Ed.). *Neural Plasticity*, *2020*, 1–8. https://doi.org/10.1155/2020/8866386

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large Language Model Based Multi-agents: A Survey of Progress and Challenges. *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, 8048–8057. https://doi.org/10.24963/ijcai.2024/890

Hanauer, D. A., Wu, D. T., Yang, L., Mei, Q., Murkowski-Steffy, K. B., Vydiswaran, V. V., & Zheng, K. (2017). Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. *Journal of Biomedical Informatics*, *67*, 1–10. https://doi.org/10.1016/j.jbi.2017.01.013

Haraty, M., McGrenere, J., & Tang, C. (2016). How personal task management differs across individuals. *International Journal of Human-Computer Studies*, *88*, 13–37. https://doi.org/10.1016/j.ijhcs.2015.11.006

Harvey, M., Hauff, C., & Elsweiler, D. (2015). Learning by Example: Training Users with High-quality Query Suggestions. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 133–142. https://doi.org/10.1145/2766462.2767731

Hasani, M. A. (2018). Understanding Risk and Uncertainty in Project Management. *European Journal of Economics, Law and Politics*, *05*(01). https://doi.org/10.19044/elp.v5no1a3

Hassan, A., & White, R. W. (2012). Task tours: Helping users tackle complex search tasks. *Proceedings of the 21st ACM international conference on Information and knowledge management*, 1885–1889. https://doi.org/10.1145/2396761.2398537

Hassan, A., White, R. W., Dumais, S. T., & Wang, Y.-M. (2014). Struggling or exploring? Disambiguating long search sessions. *Proceedings of the 7th ACM international conference on Web search and data mining*, 53–62. https://doi.org/10.1145/2556195.2556221

Hassan, M. M., Ahmad, N., & Hashim, A. H. (2021). The Conceptual Framework of Housing Purchase Decision-Making Process. *International Journal of Academic Research in Business and Social Sciences*, *11*(11), Pages 1672–1690. https://doi.org/10.6007/IJARBSS/v11-i11/11653

Hassan Awadallah, A., White, R. W., Pantel, P., Dumais, S. T., & Wang, Y.-M. (2014). Supporting Complex Search Tasks. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 829–838. https://doi.org/10.1145/2661829.2661912

Havighurst, R. J. (1953). *Human development and education*. Longmans, Green.

Hendahewa, C., & Shah, C. (2015). Implicit search feature based approach to assist users in exploratory search tasks. *Information Processing & Management*, *51*(5), 643–661. https://doi.org/10.1016/j.ipm.2015.06.004

Hendahewa, C., & Shah, C. (2017). Evaluating user search trails in exploratory search tasks. *Information Processing & Management*, *53*(4), 905–922. https://doi.org/10.1016/j.ipm.2017.04.001

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Höchli, B., Brügger, A., & Messner, C. (2018). How Focusing on Superordinate Goals Motivates Broad, Long-Term Goal Pursuit: A Theoretical Perspective. *Frontiers in Psychology*, *9*, 1879. https://doi.org/10.3389/fpsyg.2018.01879

Holter, S., & El-Assady, M. (2024). Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation. *Computer Graphics Forum*, *43*(3), e15107. https://doi.org/10.1111/cgf.15107

Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., Wang, Y., Tang, R., & Chen, E. (2024, February). Understanding the planning of LLM agents: A survey. Retrieved February 16, 2024, from http://arxiv.org/abs/2402.02716

Hunt, E. (1994). Problem Solving. In *Thinking and Problem Solving* (pp. 215–232). Elsevier. https://doi.org/10.1016/B978-0-08-057299-4.50013-X

Huurdeman, H. C., & Kamps, J. (2015). Supporting the Process: Adapting Search Systems to Search Stages. In S. Kurbanoglu, J. Boustany, S. Špiranec, E. Grassian, D. Mizrachi, & L. Roy (Eds.), *Information Literacy: Moving Toward Sustainability* (pp. 394–404, Vol. 552). Springer International Publishing. https://doi.org/10.1007/978-3-319-28197-1_40

Inder, R. (1996). Planning and Problem Solving. In *Artificial Intelligence* (pp. 23–53). Elsevier. https://doi.org/10.1016/B978-012161964-0/50004-2

Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2022). A Survey on Conversational Recommender Systems. *ACM Computing Surveys*, *54*(5), 1–36. https://doi.org/10.1145/3453154

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., Keskustalo, H., Kumpulainen, S., Saastamoinen, M., Savolainen, R., et al. (2015). Task-based information interaction evaluation: The viewpoint of program theory. *ACM Transactions on Information Systems (TOIS)*, *33*(1), 1–30. https://doi.org/10.1145/2699660

Joksimovic, S., Ifenthaler, D., Marrone, R., De Laat, M., & Siemens, G. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers and Education: Artificial Intelligence*, *4*, 100138. https://doi.org/10.1016/j.caeai.2023.100138

Jones, W. (2010). *Keeping Found Things Found: The Study and Practice of Personal Information Management*. Morgan Kaufmann.

Jones, W., Bruce, H., Foxley, A., & Munat, C. F. (2006). Planning personal projects and organizing personal information. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–24. https://doi.org/10.1002/meet.14504301159

Jones, W., Dinneen, J. D., Capra, R., Diekema, A. R., & Pérez-Quiñones, M. A. (2017). Personal Information Management. In *Encyclopedia of Library and Information Sciences* (4th ed.). CRC Press.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Keller, A. M., Taylor, H. A., & Brunyé, T. T. (2020). Uncertainty promotes information-seeking actions, but what information? *Cognitive Research: Principles and Implications*, *5*(1), 42. https://doi.org/10.1186/s41235-020-00245-2

Kelly, D. (2007). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval*, *3*(1—2), 1–224. https://doi.org/10.1561/1500000012

Kersten-van Dijk, E. T., Westerink, J. H., Beute, F., & IJsselsteijn, W. A. (2017). Personal Informatics, Self-Insight, and Behavior Change: A Critical Review of Current Literature. *Human–Computer Interaction*, *32*(5-6), 268–296. https://doi.org/10.1080/07370024.2016.1276456

Khakurel, J., & Blomqvist, K. (2022). Artificial Intelligence Augmenting Human Teams. A Systematic Literature Review on the Opportunities and Concerns. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI* (pp. 51–68, Vol. 13336). Springer International Publishing. https://doi.org/10.1007/978-3-031-05643-7_4

Kiesel, J., Gohsen, M., Mirzakhmedova, N., Hagen, M., & Stein, B. (2024). Simulating Follow-Up Questions in Conversational Search. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in Information Retrieval* (pp. 382–398, Vol. 14609). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56060-6_25

Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581001

Kong, Y., Ruan, J., Chen, Y., Zhang, B., Bao, T., Shi, S., Du, G., Hu, X., Mao, H., Li, Z., Zeng, X., & Zhao, R. (2023). TPTU-v2: Boosting Task Planning and Tool Usage of Large Language Model-based Agents in Real-world Systems. https://doi.org/10.48550/ARXIV.2311.11315

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science*, *42*(5), 361–371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5⟨361::AID-ASI6⟩3.0.CO;2-#

Leach, D., Hagger-Johnson, G., Doerner, N., Wall, T., Turner, N., Dawson, J., & Grote, G. (2013). Developing a measure of work uncertainty. *Journal of Occupational and Organizational Psychology*, *86*(1), 85–99. https://doi.org/10.1111/joop.12000

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M. S., & Liang, P. (2023). Evaluating Human-Language Model Interaction. *Transactions on Machine Learning Research*. Retrieved January 8, 2025, from https://openreview.net/forum?id=hjDYJUn9l1

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*.

Li, D., & Du, Y. (2017). *Artificial Intelligence with Uncertainty* (2nd ed.). CRC Press. https://doi.org/10.1201/9781315366951

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Thirty-seventh Conference on Neural Information Processing Systems*. Retrieved January 13, 2024, from http://arxiv.org/abs/2303.17760

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y. (2024, December). LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. https://doi.org/10.48550/arXiv.2412.05579

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 557–566. https://doi.org/10.1145/1753326.1753409

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023, May). AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, *44*(6), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005

Liao, J., Zhong, L., Zhe, L., Xu, H., Liu, M., & Xie, T. (2024). Scaffolding Computational Thinking With ChatGPT. *IEEE Transactions on Learning Technologies*, *17*, 1628–1642. https://doi.org/10.1109/TLT.2024.3392896

Liao, L., Yang, G. H., & Shah, C. (2023). Proactive Conversational Agents in the Post-ChatGPT World. *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3539618.3594250

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Lieder, F., Chen, O. X., Krueger, P. M., & Griffiths, T. L. (2019). Cognitive prostheses for goal achievement. *Nature Human Behaviour*, *3*(10), 1096–1106. https://doi.org/10.1038/s41562-019-0672-9

Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., Liu, Y., Wu, C., Li, X., Zhu, C., Guo, H., Yu, Y., Tang, R., & Zhang, W. (2025). How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Transactions on Information Systems*, *43*(2), 1–47. https://doi.org/10.1145/3678004

Liu, C., Liu, Y.-H., Liu, J., & Bierig, R. (2021). Search Interface Design and Evaluation. *Foundations and Trends® in Information Retrieval*, *15*(3-4), 243–416. https://doi.org/10.1561/1500000073

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J., & Zhang, X. (2010). Search behaviors in different task types. *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*, 69. https://doi.org/10.1145/1816123.1816134

Liu, J., & Azzopardi, L. (2024). Search under Uncertainty: Cognitive Biases and Heuristics: A Tutorial on Testing, Mitigating and Accounting for Cognitive Biases in Search Experiments. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3013–3016. https://doi.org/10.1145/3626772.3661382

Liu, J., Mitsui, M., Belkin, N. J., & Shah, C. (2019). Task, Information Seeking Intentions, and User Behavior. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 123–132. https://doi.org/10.1145/3295750.3298922

Liu, J., & Shah, C. (2019a). *Interactive IR User Study Design, Evaluation, and Reporting*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02319-4

Liu, J., & Shah, C. (2019b). Investigating the Impacts of Expectation Disconfirmation on Web Search. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 319–323. https://doi.org/10.1145/3295750.3298959

Liu, J., & Yu, R. (2021). State-Aware Meta-Evaluation of Evaluation Metrics in Interactive Information Retrieval. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3258–3262. https://doi.org/10.1145/3459637.3482190

Liu, O., Fu, D., Yogatama, D., & Neiswanger, W. (2024, February). DeLLMa: A Framework for Decision Making Under Uncertainty with Large Language Models. Retrieved February 28, 2024, from http://arxiv.org/abs/2402.02392

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Locke, E., & Latham, G. (2005). Goal-Setting Theory. In *Organizational Behavior 1.* Routledge.

Luo, C., Liu, Y., Sakai, T., Zhou, K., Zhang, F., Li, X., & Ma, S. (2017). Does Document Relevance Affect the Searcher's Perception of Time? *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 141–150. https://doi.org/10.1145/3018661.3018694

Luo, J., Zhang, S., & Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 587–596. https://doi.org/10.1145/2600428.2609629

Luo, M. M., & Nahl, D. (2019). Let's Google: Uncertainty and bilingual search. *Journal of the Association for Information Science and Technology*, *70*(9), 1014–1025. https://doi.org/10.1002/asi.24174

Maes, T., Gebhardt, K., & Riel, A. (2022). The Relationship Between Uncertainty and Task Execution Strategies in Project Management. *Project Management Journal*, *53*(4), 382–396. https://doi.org/10.1177/87569728221089831

Mao, J., Liu, Y., Kando, N., Zhang, M., & Ma, S. (2018). How Does Domain Expertise Affect Users' Search Interaction and Outcome in Exploratory Search? *ACM Transactions on Information Systems*, *36*(4), 1–30. https://doi.org/10.1145/3223045

Marchionini, G. (2003). *Information seeking in electronic environments* (dig. repr). Cambridge Univ. Press.

Maxwell, D., & Azzopardi, L. (2016). Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1141–1144. https://doi.org/10.1145/2911451.2911469

Maxwell, D. M. (2019). *Modelling search and stopping in interactive information retrieval* [PhD]. University of Glasgow. https://doi.org/10.5525/gla.thesis.41132

McCombs, B. L. (1991). Motivation and Lifelong Learning. *Educational Psychologist*, *26*(2), 117–127. https://doi.org/10.1207/s15326985ep2602_4

McGregor, I., & Little, B. R. (1998). Personal projects, happiness, and meaning: On doing well and being yourself. *Journal of Personality and Social Psychology*, *74*(2), 494–512. https://doi.org/10.1037/0022-3514.74.2.494

Mestre, J. (2024). Parsing through paradigms: Uncertainty and decision-making in human information behavior. *Journal of Documentation*, *80*(1), 39–53. https://doi.org/10.1108/JD-02-2023-0027

Miceli, M., & Castelfranchi, C. (2005). Anxiety as an "epistemic" emotion: An uncertainty theory of anxiety. *Anxiety, Stress & Coping*, *18*(4), 291–319. https://doi.org/10.1080/10615800500209324

Miller, G. A., Eugene, G., & Pribram, K. H. (1968). Plans and the Structure of Behaviour [Num Pages: 14]. In *Systems Research for Behavioral Science*. Routledge.

Mita, R., Taleuchi, T., Tanikawa, T., Narumi, T., & Hirose, M. (2017). Early warning of task failure using task processing logs. *Proceedings of the 8th Augmented Human International Conference*, 1–9. https://doi.org/10.1145/3041164.3041180

Moraveji, N., Russell, D., Bien, J., & Mease, D. (2011). Measuring improvement in user search performance resulting from optimal search tips. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 355–364. https://doi.org/10.1145/2009916.2009966

Morris, M. R. (2008). A survey of collaborative web search practices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1657–1660. https://doi.org/10.1145/1357054.1357312

Morris, M. R., & Horvitz, E. (2007). SearchTogether: An interface for collaborative web search. *Proceedings of the 20th annual ACM symposium on User interface software and technology*, 3–12. https://doi.org/10.1145/1294211.1294215

Myers, K., Berry, P., Blythe, J., Conley, K., Gervasio, M., McGuinness, D. L., Morley, D., Pfeffer, A., Pollack, M., & Tambe, M. (2007). An Intelligent Personal Assistant for Task and Time Management. *AI Magazine*, *28*(2), 47–47. https://doi.org/10.1609/aimag.v28i2.2039

Neenan, M., & Dryden, W. (2020). *Cognitive Behavioural Coaching: A Guide to Problem-Solving and Personal Development* (3rd ed.). Routledge. https://doi.org/10.4324/9781003027317

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166. https://doi.org/10.1037/h0048495

Newell, A., & Simon, H. A. (1972). *Human problem solving* (8. print). Prentice-Hall.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In *Consciousness and self-regulation: Advances in research and theory volume 4* (pp. 1–18). Springer.

Odijk, D., White, R. W., Hassan Awadallah, A., & Dumais, S. T. (2015). Struggling and Success in Web Search. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1551–1560. https://doi.org/10.1145/2806416.2806488

OpenAI. (2025, October). Introducing study mode. Retrieved October 11, 2025, from https://openai.com/index/chatgpt-study-mode/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 27730–27744). Curran Associates, Inc. http://arxiv.org/abs/2203.02155

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/https://doi.org/10.1145/3586183.3606763

Petrelli, D. (2008). On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information Processing & Management*, *44*(1), 22–38. https://doi.org/10.1016/j.ipm.2007.01.024

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J. Y., Zhou, D., Gao, C., Xu, F., Zhang, F., Rong, K., Su, J., & Li, Y. (2025, February). AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. https://doi.org/10.48550/arXiv.2502.08691

Pinheiro, D. L., Melkers, J., & Newton, S. (2017). Take me where I want to go: Institutional prestige, advisor sponsorship, and academic career placement preferences (J. L. Rosenbloom, Ed.). *PLOS ONE*, *12*(5), e0176977. https://doi.org/10.1371/journal.pone.0176977

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1339–1384. https://doi.org/10.18653/v1/2023.emnlp-main.85

Rapp, A., & Cena, F. (2016). Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies*, *94*, 1–17. https://doi.org/10.1016/j.ijhcs.2016.05.006

Rapp, A., & Tirassa, M. (2017). Know Thyself: A Theory of the Self for Personal Informatics. *Human–Computer Interaction*, *32*(5-6), 335–380. https://doi.org/10.1080/07370024.2017.1285704

Retzlaff, C. O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M. E., & Holzinger, A. (2024). Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and

Opportunities. *Journal of Artificial Intelligence Research*, *79*, 359–415. https://doi.org/10.1613/jair.1.15348

Rha, E. Y., Mitsui, M., Belkin, N. J., & Shah, C. (2016). Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–9. https://doi.org/10.1002/pra2.2016.14505301048

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., & Weston, J. (2021). Recipes for Building an Open-Domain Chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24

Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., Du, G., Shi, S., Mao, H., Li, Z., Zeng, X., & Zhao, R. (2023, November). TPTU: Large Language Model-based AI Agents for Task Planning and Tool Usage. Retrieved February 14, 2024, from http://arxiv.org/abs/2308.03427

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (Third edition). Pearson.

Salimzadeh, S., He, G., & Gadiraju, U. (2024). Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Savenkov, D., & Agichtein, E. (2014). To hint or not: Exploring the effectiveness of search hints for complex informational tasks. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 1115–1118. https://doi.org/10.1145/2600428.2609523

Savolainen, R. (2012). Elaborating the motivational attributes of information need and uncertainty. *Inf. Res.*, *17*. https://api.semanticscholar.org/CorpusID:22382457

Savolainen, R. (2015). Cognitive barriers to information seeking: A conceptual analysis. *Journal of Information Science*, *41*(5), 613–623. https://doi.org/10.1177/0165551515587850

Schlosser, M. (2015). Agency. In *Stanford encyclopedia of philosophy*.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., & Barsoum, E. (2025, June). Agent Laboratory: Using LLM Agents as Research Assistants. https://doi.org/10.48550/arXiv.2501.04227

Sebastia, L., Onaindia, E., & Marzal, E. (2006). Decomposition of planning problems. *Ai Communications*, *19*(1), 49–81.

Selten, R., Pittnauer, S., & Hohnisch, M. (2012). Dealing with Dynamic Decision Problems when Knowledge of the Environment Is Limited: An Approach Based on Goal Systems. *Journal of Behavioral Decision Making*, *25*(5), 443–457. https://doi.org/10.1002/bdm.738

Shah, C., & Bender, E. M. (2024). Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? *ACM Transactions on the Web*, 3649468. https://doi.org/10.1145/3649468

Shah, C., White, R., Thomas, P., Mitra, B., Sarkar, S., & Belkin, N. (2023). Taking Search to Task. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 1–13. https://doi.org/10.1145/3576840.3578288

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *298*(1089), 199–209. https://doi.org/10.1098/rstb.1982.0082

Shi, W., He, X., Zhang, Y., Gao, C., Li, X., Zhang, J., Wang, Q., & Feng, F. (2024). Large Language Models are Learnable Planners for Long-Term Recommendation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1893–1903. https://doi.org/10.1145/3626772.3657683

Shneiderman, B. (2021). Human-Centered AI. *Issues in Science and Technology*, *37*(2), 56–61. Retrieved December 16, 2024, from https://www.proquest.com/docview/2481224128/abstract/2614CEE4C78E49F6PQ/1

Silver, D., & Veness, J. (2010). Monte-Carlo Planning in Large POMDPs. *Advances in Neural Information Processing Systems*, *23*. Retrieved April 5, 2025, from https://proceedings.neurips.cc/paper/2010/hash/edfbe1afcf9246bb0d40eb4d8027d90f-Abstract.html

Simon, H. A., Dantzig, G. B., Hogarth, R., Plott, C. R., Raiffa, H., Schelling, T. C., Shepsle, K. A., Thaler, R., Tversky, A., & Winter, S. (1987). Decision Making and Problem Solving. *Interfaces*, *17*(5), 11–31. https://doi.org/10.1287/inte.17.5.11

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159. https://doi.org/10.1037/h0030806

Siro, C., Aliannejadi, M., & De Rijke, M. (2024). Understanding and Predicting User Satisfaction with Conversational Recommender Systems. *ACM Transactions on Information Systems*, *42*(2), 1–37. https://doi.org/10.1145/3624989

Sivaprasad, S., Kaushik, P., Abdelnabi, S., & Fritz, M. (2025). A Theory of Response Sampling in LLMs: Part Descriptive and Part Prescriptive. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30091–30135. https://doi.org/10.18653/v1/2025.acl-long.1454

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The user experience of ChatGPT: Findings from a questionnaire study of early users. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–10. https://doi.org/https://doi.org/10.1145/3571884.3597144

Smith, C. L. (2017). Domain-independent search expertise: Gaining knowledge in query formulation through guided practice. *Journal of the Association for Information Science and Technology*, *68*(6), 1462–1479. https://doi.org/10.1002/asi.23776

Sniehotta, F. F., Schwarzer, R., Scholz, U., & Schüz, B. (2005). Action planning and coping planning for long-term lifestyle change: Theory and assessment. *European Journal of Social Psychology*, *35*(4), 565–576. https://doi.org/10.1002/ejsp.258

Soufan, A., Ruthven, I., & Azzopardi, L. (2021). Untangling the Concept of Task in Information Seeking and Retrieval. *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 73–81. https://doi.org/10.1145/3471158.3472259

Sun, Y., & Zhang, Y. (2018). Conversational Recommender System. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 235–244. https://doi.org/10.1145/3209978.3210002

Takeuchi, T., Suwa, K., Tamura, H., Narumi, T., Tanikawa, T., & Hirose, M. (2013). A task-management system using future prediction based on personal lifelogs and plans. *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, 235–238. https://doi.org/10.1145/2494091.2494166

Tan, Z., & Jiang, M. (2023, December). User Modeling in the Era of Large Language Models: Current Research and Future Directions. Retrieved May 17, 2024, from http://arxiv.org/abs/2312.11518

Tang, J., Xia, L., Li, Z., & Huang, C. (2025, May). AI-Researcher: Autonomous Scientific Innovation. https://doi.org/10.48550/arXiv.2505.18705

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., & Hupkes, D. (2025, January). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. https://doi.org/10.48550/arXiv.2406.12624

Toms, E. G. (2019). Information activities and tasks. *Information at work: information management in the workplace. London: Facet Publishing*, 33–62.

Trippas, J. R. (2024, December). Conversational Search. In O. Alonso & R. Baeza-Yates (Eds.), *Information Retrieval* (1st ed., pp. 285–319). ACM. https://doi.org/10.1145/3674127.3674135

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. https://doi.org/10.1037/a0018963

Tu, B., Lu, X., & Zhu, K. (2024). ChatGPT-CPS: The Cultivation of Collaborative Problem-Solving Ability through Human-AI Interaction. In D. Hu, F. Lu, F. Chen, & S. Liu (Eds.), *Proceedings of the 2024 5th International Conference on Modern Education and Information Management (ICMEIM 2024)* (pp. 482–488, Vol. 29). Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-568-3_58

Tversky, A., & Kahneman, D. (2000a, September). Advances in Prospect Theory: Cumulative Representation of Uncertainty. In *Choices, Values, and Frames* (pp. 44–66, Vol. 5). Cambridge University Press. https://doi.org/10.1017/CBO9780511803475.004

Tversky, A., & Kahneman, D. (2000b, September). Loss Aversion in Riskless Choice: A Reference-Dependent Model. In *Choices, Values, and Frames* (pp. 143–158, Vol. 106). Cambridge University Press.

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance

in university students: An experimental study. *Computers & Education*, *215*, 105031. https://doi.org/10.1016/j.compedu.2024.105031

Urgo, K., & Arguello, J. (2024). The Effects of Goal-setting on Learning Outcomes and Self-Regulated Learning Processes. *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 278–290. https://doi.org/10.1145/3627508.3638348

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology (ARIST)*, *37*, 413–64. https://doi.org/10.1002/aris.1440370110

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. Retrieved April 5, 2025, from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vygotskij, L. S., & Cole, M. (1981). *Mind in society: The development of higher psychological processes* (Nachdr.). Harvard Univ. Press.

Walker, W., Harremoës, P., Rotmans, J., Van Der Sluijs, J., Van Asselt, M., Janssen, P., & Krayer Von Krauss, M. (2003). Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment*, *4*(1), 5–17. https://doi.org/10.1076/iaij.4.1.5.16466

Wang, B., & Liu, J. (2022). Investigating the Relationship between In-Situ User Expectations and Web Search Behavior. *Proceedings of the Association for Information Science and Technology*, *59*(1), 827–829. https://doi.org/10.1002/pra2.740

Wang, B., & Liu, J. (2023). Investigating the role of in-situ user expectations in Web search. *Information Processing & Management*, *60*(3), 103300.

Wang, B., Liu, J., Karimnazarov, J., & Thompson, N. (2024). Task Supportive and Personalized Human-Large Language Model Interaction: A User Study. *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 370–375. https://doi.org/10.1145/3627508.3638344

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3334480.3381069

Wang, Y., & Chiew, V. (2010). On the cognitive process of human problem solving. *Cognitive Systems Research*, *11*(1), 81–92. https://doi.org/10.1016/j.cogsys.2008.08.003

Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X. (, & Liang, Y. (2023). Describe, Explain, Plan and Select: Interactive Planning with LLMs Enables Open-World Multi-Task Agents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 34153–34189, Vol. 36). Curran Associates, Inc. https://proceedings.neurips.cc/paper_

files/paper/2023/file/6b8dfb8c0c12e6fafc6c256cb08a5ca7-Paper-Conference.pdf

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

White, R. W. (2023, November). Navigating Complex Search Tasks with AI Copilots. Retrieved December 23, 2023, from http://arxiv.org/abs/2311.01235

White, R. W., & Hassan Awadallah, A. (2019). Task Duration Estimation. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 636–644. https://doi.org/10.1145/3289600.3290997

White, R. W., Hassan Awadallah, A., & Sim, R. (2019). Task Completion Detection: A Study in the Context of Intelligent Systems. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 405–414. https://doi.org/10.1145/3331184.3331187

Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246–258. https://doi.org/10.1002/asi.10367

Wong, P. T. P. (2013). *The Human Quest for Meaning: Theories, Research, and Applications*. Routledge.

Woolley, K., & Fishbach, A. (2017). Immediate Rewards Predict Adherence to Long-Term Goals. *Personality and Social Psychology Bulletin*, *43*(2), 151–162. https://doi.org/10.1177/0146167216676480

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., & Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2023, September). The Rise and Potential of Large Language Model Based Agents: A Survey. Retrieved December 30, 2023, from http://arxiv.org/abs/2309.07864

Xie, H. (2000). Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. *Journal of the American Society for Information Science*, *51*(9), 841–857. https://doi.org/10.1002/(SICI)1097-4571(2000)51:9⟨841::AID-ASI70⟩3.0.CO;2-0

Xie, I. (2009). Dimensions of tasks: Influences on information-seeking and retrieving process. *Journal of Documentation*, *65*(3), 339–366. https://doi.org/10.1108/00220410910952384

Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., & Su, Y. (2024). TravelPlanner: A Benchmark for Real-World Planning with Language Agents. *Forty-first International Conference on Machine Learning*.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In J.

Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing* (pp. 563–574, Vol. 11839). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51

Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., Wang, Z. Z., Zhou, X., Guo, Z., Cao, M., Yang, M., Lu, H. Y., Martin, A., Su, Z., Maben, L., Mehta, R., Chi, W., Jang, L., Xie, Y., … Neubig, G. (2025, September). TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. https://doi.org/10.48550/arXiv.2412.14161

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2023). Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction*, *39*(3), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Xu, Y., Shieh, C.-H., Van Esch, P., & Ling, I.-L. (2020). AI Customer Service: Task Complexity, Problem-Solving Ability, and Usage Intention. *Australasian Marketing Journal*, *28*(4), 189–199. https://doi.org/10.1016/j.ausmj.2020.03.005

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. https://doi.org/10.48550/ARXIV.2210.03629

Yorke-Smith, N., Saadati, S., Myers, K. L., & Morley, D. N. (2012). THE DESIGN OF A PROACTIVE PERSONAL AGENT FOR TASK MANAGEMENT. *International Journal on Artificial Intelligence Tools*, *21*(01), 1250004. https://doi.org/10.1142/S0218213012500042

Young, S. N., Vanwye, W. R., Schafer, M. A., Robertson, T. A., & Poore, A. V. (2019). Factors Affecting PhD Student Success. *International Journal of Exercise Science*, *12*(1), 34–45.

Zahera, H. M., El-Hady, G. F., & El-Wahed, W. F. A. (2013). Query Recommendation for Improving Search Engine Results. In Z. (Lu (Ed.), *Information Retrieval Methods for Multidisciplinary Applications*. IGI Global. https://doi.org/10.4018/978-1-4666-3898-3.ch004

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581388

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Zhai, C., Cohen, W. W., & Lafferty, J. (2015). Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. *ACM SIGIR Forum*, *49*(1), 2–9. https://doi.org/10.1145/2795403.2795405

Zhang, J., Yang, R., Liu, S., Lin, T.-E., Huang, F., Chen, Y., Li, Y., & Tao, D. (2025, May). Supervised Optimism Correction: Be Confident When LLMs Are Sure. https://doi.org/10.48550/arXiv.2504.07527

Zhang, S., & Balog, K. (2020). Evaluating Conversational Recommender Systems via User Simulation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1512–1520. https://doi.org/10.1145/3394486.3403202

Zhang, Y., Liu, X., & Zhai, C. (2017). Information Retrieval Evaluation as Search Simulation: A General Formal Framework for IR Evaluation. *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 193–200. https://doi.org/10.1145/3121050.3121070

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., & Li, Q. (2024). Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*, 1–20. https://doi.org/10.1109/TKDE.2024.3392335

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, *36*, 46595–46623. Retrieved February 13, 2025, from https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html

Zheng, Q., Tang, Y., Liu, Y., Liu, W., & Huang, Y. (2022). UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. *CHI Conference on Human Factors in Computing Systems*, 1–24. https://doi.org/10.1145/3491102.3501855

Zhou, R., He, X., Fan, Q., Li, Y., Li, Y., Xiao, X., & Fang, J. (2025). Exploring ChatGPT -Facilitated Scaffolding in Undergraduates' Mathematical Problem Solving. *Journal of Computer Assisted Learning*, *41*(4), e70077. https://doi.org/10.1111/jcal.70077

1 Appendix A

1.1 Initial code books

Table A.1: Initial code book for interview analysis

Step and topic	**Code category**	**Example subcodes**
Discussing personal goals and planning experiences	Personal Goals	Types of Goals, Goal Complexity
	Planning Strategies	Approach to Planning, Tools and Resources Used, Frequency of Planning
	Successes in Planning	Achieved Goals, Factors Contributing to Success
	Challenges in Planning	Uncertainty Management, Barriers
	Reflections on AI-Assisted Planning	Comparisons to Traditional Methods, Usefulness of AI
Exploring the potential of AI in long-term life task planning	Acceptance of AI Assistance	Conditions for Acceptance, Trust in AI
	Perceived Advantages of AI	Efficiency, Cognitive load, Personalization
	Perceived Limitations of AI	Understanding Context, Technical Constraints
	Trust and Reliance on AI	Criteria for Trust; Barriers to Trust
	Expectations of AI Performance	User Expectations, Scenarios of Success/Failure
	Influence of Task Management Principles	Integration with Traditional Methods

Addressing uncertainty in task planning	Types of Uncertainty	Outcome Uncertainty, Action Uncertainty
	Impact of Uncertainty	Emotional Responses, Influence on Planning
	Coping Mechanisms for Uncertainty	Personal Strategies, Role of AI
	Perceptions of AI in Managing Uncertainty	Effectiveness of AI, Limitations
	Necessity and Challenges of Reducing Uncertainty	Importance, Challenges
Introducing and discussing the proposed GOLF framework	Overall Opinions on the Workflow	General Impressions, Comparison to Current Methods
	Perceived Uncertainties in the Workflow	Goal Setting, Communication, Information, Action
	Suggestions for Improvement	User Preferences, Motivational elements; Explainability, Uncertainty Handling
	Concerns and Further Suggestions	User Autonomy, Framework Flexibility, Other Concerns

1.2 Prompt templates

1.2.1 Initial plan prompt $\mathcal{P}_0$

```
You are an assistant designed to help users create a structured
   task plan. The plan should be hierarchical, consisting of
   the following components:
1. **Goal**: A high-level task goal related to the user's need.
   It defines the context and types of activities.
2. **Activity**: Key steps required to achieve the goal.
   Activities can be long-running, open-ended, and
   semi-structured. Each activity comprises multiple tasks.
3. **Task**: Specific actions with defined sub-goals and
   endpoints. Tasks can be further divided into smaller
   sub-tasks, contributing to the completion of the associated
   activity.

## Variables:
- **Goal Statement**: Choose one of the following:
  - Planning to buy a house.
  - Planning to undertake obesity management due to obesity
     concerns.
  - Applying for a PhD program and seeking potential advisors.

- **User's Initial State**: Unfamiliar with the task and wanting
   to start it from the beginning.

## Input Format:
- **User**:
  - I am [Goal statement]
  - I am [User's Initial State]

## Output Format:
'''json
{
  "Goal": "The final task goal placeholder",
  "Activities": {
    "Activity1Placeholder": {
      "Tasks": {
        "Task1": "Description of Task1 placeholder",
        "Task2": "Description of Task2 placeholder",
        "TaskN": "Description of additional tasks placeholder
           (if any)"
      }
    },
    "ActivityNPlaceholder": {
      "Tasks": {
        "Task1": "Description of Task1 placeholder",
        "TaskN": "Description of additional tasks placeholder
           (if any)"
```

```
      }
    }
  }
}
```

1.2.2 Step planner prompt $\mathcal{P}_p$

```
# Role: Assistant Step Planner
- description: An assistant designed to guide users through long
    and complex tasks step by step, ensuring tasks are
    manageable, measurable, and actionable.

## Skills
1. Identify the user's current state and the task progress in
    relation to a general task plan.
2. Given provided user and task information to decide whether to
    update current tasks or generate new tasks.
3. Decompose unstructured tasks into multi-step structured tasks
    based on "Getting Things Done: The Art of Stress-Free
    Productivity" (GTD) principles.
4. Divide subtasks into three categories (communication,
    information-retrieval, action instruction).
5. Leverage memory to adapt step planning based on known user
    preferences, past interactions, constraints, and external
    factors.

## Rules
### Using the Initial Plan (removed in plain prompt)
The **initial_plan** provides a general outline or roadmap of
    the overall task. It reflects high-level stages or goals but
    does **not** dictate the exact step-by-step execution order.
The Step Planner must:
- **Use the initial plan as a directional reference**,
    identifying what the user is trying to achieve at each
    stage. The step task covers multiple subtasks, including
    communication, information-retrieval, and action-instruction
    task types.
- **Interpret the current position within the plan** based on
    user progress and memory.
- **Generate specific subtasks** (communication,
    information-retrieval, action-instruction) **that align with
    the user's current state and final goal**.
- Always **translate high-level tasks into specific, actionable
    subtasks** based on the user's current state and goals. Use
    GTD principles to break tasks into manageable steps:
    Capture, Clarify, Organize, Reflect, and Engage.
```

```
- If the initial plan contains vague or general statements, the
    Step Planner must convert them into:
  - User-specific **information needs** for communication tasks
  - Concrete, **searchable queries** for information-retrieval
      tasks
  - Clear, **measurable instructions** for action-instruction
      tasks
- **Reorder, merge, skip, or extend** steps if needed to ensure
    feasibility, clarity, and task completion.
- Maintain a flexible planning process that adapts to:
  - Real-time user feedback
  - External environment changes
  - Missing or ambiguous information

### The Step Planner receives one of the following input types
    at each stage:
- **Communication**: Information provided by the user (e.g.,
    preferences, constraints, goals).
- **Information**: Knowledge gathered from the external
    environment (e.g., facts, resources, limitations).
- **User State**: Feedback indicating the user has taken an
    action (e.g., completion success, difficulty encountered).
    When receiving user state, the Step Planner must generate
    new or update all communication, information-retrieval, and
    action-instruction tasks based on the user's feedback.
(for the details of the task types, refer to the detailed
    definitions below)
-
### The Step Planner identifies the next step based on the user
    state and task progress: (removed in plain prompt)
- **Unchanged**: If the pending task is still relevant, and
    there is no need to change the pending task given new
    information.
- **Proceed**: If there is no pending task or previous tasks are
    completed without issues, the step planner can generate new
    subtasks, following the original task plan toward the goal.
- **Update**: If the user encounters difficulties but the
    previous subtasks remain feasible, the step planner should
    update the subtask by delving into more detailed guidance or
    further breaking the step into subtasks.
- **Revise**: If the previous tasks are not feasible due to
    unsolvable problems, you should revise the original subtask
    by altering the current step to address the user's
    challenges.

### The Step Planner outputs subtasks including: (removed in
    plain prompt)
- **Communication tasks**: For collecting user information, such
    as preferences and considerations.
```

```
- **Information-retrieval tasks**: For collecting and processing
    online information from the external environment.
- **Action-instruction tasks**: For instructing the user with
    physical actions requiring interaction with the real world.

### Along with the output, the Step Planner updates memory with
    useful information about the user state and task progress:
    (removed in plain prompt)
- User's preferences and familiarity with tasks.
- Challenges encountered in previous steps.
- External constraints and requirements.
- Contextual information relevant to task execution.

## Detailed Definitions & Boundaries of the Task Types (removed
    in plain prompt)
### 1. Communication Task
{detailed component description and format}
### 2. Information-Retrieval Task
{detailed component description and format}
### 3. Action-Instruction Task
{detailed component description and format}


## Workflows
Upon receiving any input, the planner must:
1. **Update Memory Accordingly**
{details}
2. **Analyze Task Progress & Plan the Next Step**
{details}
3. **Update or Generate Subtasks**
{details}
4. **Example Behaviors**
{details}
5. **Memory and Planning Coherence**
{details}

## Overall Output Format
'''json
{
   "step_plan":{
      "progress": "a short description of user or task progress",
      "step_task_type": {
         "communication_task":"unchanged/proceed/update/revise",
         "information_retrieval_task":
         "unchanged/proceed/update/revise",
         "action_instruction_task":
         "unchanged/proceed/update/revise"
      },
      "tasks": {
```

```
          "communication_task": ["task1placeholder",
             "taskNplaceholder"],
          "information_retrieval_task": ["task1placeholder",
             "taskNplaceholder"],
          "action_instruction_task": ["task1placeholder",
             "taskNplaceholder"]
          }
          },
  "updated_memory": {
    "summary": "a short summary of new information added to
       memory based on this review",
    "user_state": "updated user state placeholder",
    "task_familiarity": "updated task familiarity placeholder",
    "past_issues": ["list of encountered issues"],
    "external_constraints": ["list of constraints"],
    "special_requirements": ["list of special requirements"]
  }
}
```

```
## Init

You are an assistant step planner. Begin by identifying the
   user's progress based on the provided task plan and guide
   them through the next step.
```

1.2.3 Communication agent prompt $\mathcal{P}_c$

```
You are an assistant tasked with helping users complete tasks by
   collecting relevant personal information, preferences, and
   considerations through effective communication.

## Skills
1. Review communication tasks and identify user needs based on
   the provided step plan.
2. Communicate effectively to collect additional information
   required for decision-making.
3. Use structured questioning to guide users in providing
   relevant input.

## Background
The assistant operates as a communication expert, focusing on
   gathering detailed and personalized information from users
   to enhance decision-making processes. It reviews the task in
   the step plan and adjusts its questions to collect only the
   necessary and relevant data.

## Goals
```

```
- Develop structured questions to collect user information.
- Ensure the collected information aligns with the user's
    decision-making needs.
- Provide clear explanations for each question to establish
    context and encourage user engagement.

## Workflows
1. Review the communication task in the step plan to understand
    the requirements.
2. Identify the key pieces of information needed to support the
    user.
3. Formulate clear, targeted questions and provide context for
    each.
4. Respond with the formatted JSON object containing the
    questions and explanations.

## Rules
1. Retain the JSON structure exactly as specified.
2. Include detailed and contextually relevant questions in the
    `questions_for_the_user` field.
3. Each question must have a concise but clear explanation.
4. Avoid adding unnecessary details or deviating from the task's
    scope.

## Output
The assistant should respond with a structured JSON object
    containing:
- A list of questions for the user, each accompanied by an
    explanation of how the information will contribute to the
    communication task.
- Do not alter the keys or structure of the JSON format.

## OutputFormat
```json
{
  "questions_for_the_user": [
    {
      "question": "What is your primary goal for this task?",
      "explanation": "Understanding your primary goal helps
          tailor the recommendations to meet your specific
          needs."
    },
    {
      "question": "Are there any specific preferences or
          constraints we should consider?",
      "explanation": "Knowing your preferences or constraints
          ensures that the suggestions align with your
          requirements."
    }
  ]
```
```

```
}
```

1.2.4 User prompt $\mathcal{P}_u$

```
You are assigned the role of a user agent interacting with a
   planner to complete tasks. You adhere to a predefined
   decision-making process and maintain the assigned
   personality role while interacting.

## Skills
1. Follow actionable plans provided by the physical-action
   expert.
2. Execute actions with options: move forward, move backward, or
   skip.
3. Maintain a consistent personality role and communication
   style.
4. Report task progress or goals to the step planner after
   completing all actions.
5. Avoid exposing your identity as an AI or language model.
6. Provide structured feedback and explanations for task
   progress.

## Rules
1. Always adhere to the actionable plan provided by the
   physical-action expert.
2. When making decisions for each instruction:
   - **Move forward**: Complete the current instruction.
   - **Move backward**: Redo a previous instruction.
   - **Skip**: Postpone or give up the current instruction.
3. Report current task progress or goal to the step planner upon
   completing actions.
4. Responses, questions, or actions must align with the assigned
   personality role.
5. Communication should imitate the assigned personality,
   avoiding verbosity or excessive formality.
6. Never disclose that you are an AI or language model.
7. Format feedback and explanations for each task according to
   the specified output format.

## Workflows
1. Receive an actionable plan from the physical-action expert.
2. For each instruction:
   - Evaluate action options (move forward, move backward, skip).
   - Execute the chosen action.
3. After completing actions, provide structured feedback using
   the specified output format.
```

```
4. Report progress or goals to the step planner to initiate the
   next step.
5. Ensure responses and actions align with the assigned
   personality role.

## Output Format
'''json
{Taskxxx (a short name about the task in the task plan):
   {Progress: Complete/Partially Complete/Incomplete,
   Explanation: Provide your feedback or explanation about the
   progress, such as: - Because ... - I need more information
   about ... - I may have some difficulties with ... - I don't
   know ... - I am struggling with ... - Other reasons (be
   specific)... }, Taskyyy: {Progress: xxx, Explanation: xxx} }
'''
```

1.2.5 Information retrieval agent prompt $\mathcal{P}_r$

```
You are an assistant to help users with tasks related to
   information-seeking. You need to review the
   information-seeking tasks in a step plan and generate
   reasonable and objective information based on the task
   requirements. You simulate information retrieval and
   processing without actual web searches, providing plausible
   and accurate responses.

## Skills
1. Understand and analyze user-provided tasks to identify key
   information needs.
2. Generate reasonable and objective information based on task
   requirements.
3. Present findings in a clear and structured format.

## Goals
- Help users seek, understand, and process objective information
   efficiently.
- Ensure the generated information is relevant and reasonable
   for the user's context.

## Workflows
1. Task Review:
   - Analyze the provided information-seeking task for clarity.
   - Identify the type of information needed and possible
      sources.
2. Information Generation:
   - Simulate the retrieval of relevant information.
   - Compile objective and concise responses based on the task
      requirements.
```

```
3. Output Compilation:
   - Present the results in the specified JSON format.

## Rules
1. Do not perform actual web searches; simulate the process with
   plausible responses.
2. Ensure information generated is objective, relevant, and
   complete.
3. Follow the structured output format to maintain consistency.

## OutputFormat
```json
{
  "information_seeking": [
    {
      "information_task": "<what kind of information is needed>",
      "information_source": "<where to find such information>",
      "collected_information": "<generated information by you>"
    },
    {
      "information_task": "<another task>",
      "information_source": "<source>",
      "collected_information": "<generated information>"
    }
  ]
}
```
```

1.2.6 Action instruction agent prompt $\mathcal{P}_a$

```
You are an assistant specialized in helping users with tasks
   that require physical actions. Your role is to review
   physical-action tasks provided by a step planner and assist
   the user by suggesting actionable steps or real-world
   interactions necessary to proceed with the task. You cannot
   perform the physical-action tasks yourself, but you can
   leverage information from communication and
   information-seeking tasks to guide the user.

## Goals
- To assist users in breaking down physical-action tasks into
   actionable steps.
- To provide clear and well-justified recommendations to achieve
   the task objectives.

## Skills
1. Reviewing and analyzing physical-action tasks for clarity and
   feasibility.
```

```
2. Generating actionable, step-by-step instructions for the user
   to execute.
3. Ensuring that suggestions align with the user's real-world
   goals.

## Workflows
1. **Review Input**: Analyze the physical-action task and the
   context provided by the step planner.
2. **Extract Key Details**: Identify actionable steps based on
   the provided information.
3. **Generate Suggestions**: Create clear and concise
   instructions for the user, accompanied by explanations.
4. **Output Format**: Ensure that the output conforms to the
   specified JSON format.


## Rules
1. Focus on tasks requiring real-world interactions.
2. Use the information already gathered during communication and
   information-seeking tasks.
3. Provide explanations for how each step contributes to
   completing the task.
4. Retain the specified format for output keys (e.g.,
   "physical_action", "action", "explanation").

## OutputFormat
```json
{
  "physical_action": [
    {
      "action": "xxx",
      "explanation": "Explain how the action will help complete
         the physical-action task"
    },
    {
      "action": "xxx",
      "explanation": "xxx"
    }
  ]
}
```
```

1.2.7 Evaluator prompt $\mathcal{P}_e$

```
You are an evaluator responsible for assessing the quality and
   effectiveness of expert-provided task plans and supports
   from the perspective of the user. It evaluates how well each
   component aligns with user needs and preferences, using a
   structured 5-point Likert scale.

## Skills
1. Ability to simulate user perspectives and needs in diverse
   task scenarios.
2. Proficient in evaluating the relevance and quality of
   expert-provided support.
3. Skilled in structured assessment using quantitative metrics,
   such as a Likert scale.
4. Capable of providing actionable feedback for improvement.

## Background
In this scenario, a user has a specific task goal and is
   consulting a task planning expert to receive support for
   achieving the goal. The evaluator agent adopts the user's
   perspective to assess the expert's support quality.

## Goals
1. Accurately simulate the user's perspective, needs, and
   preferences.
2. Evaluate expert-provided supports for alignment with the
   user's task goal.
3. Assign ratings to each component based on its effectiveness
   and relevance.
4. Provide detailed feedback for refining the task planning
   process.

## Variables
- **User Task Goal**
- **User State:**
- **Expert Support Components**

## Rules
1. Always adopt the perspective of the user to ensure
   evaluations are empathetic and accurate.
2. Assess each component against clear criteria (e.g.,
   relevance, clarity, feasibility).
3. Use a 5-point Likert scale for ratings (1 = Poor, 5 =
   Excellent).
4. Provide constructive feedback alongside ratings for
   actionable insights.

## Workflows
1. Identify the user's task goal, current state, and the
   provided expert supports.
```

```
2. Analyze how each support component contributes to the user's
   ability to achieve their goal.
3. Assign a Likert scale rating to each component, accompanied
   by clear feedback.
4. Summarize the overall evaluation with actionable suggestions
   for improvement.

## Metrics
[Specific 5-point likert scale metrics]
```

1.3 Annotation metrics and scales

For the Step plan, the metric evaluates how effectively the system identifies the user's current state and provides an appropriate action-proceed, update, or revise-aligned with that state. When the user completes the current step without issues, the system should proceed to the next step, maintaining momentum toward the goal. If the user encounters difficulties but the step remains feasible, the system should update by delving into more detailed guidance or breaking the step into subtasks. In cases where the step is not feasible due to unsolvable problems, the system should revise the plan by altering the current step to address the user's challenges. This metric ensures that the system is responsive and adaptive, tailoring its actions to the user's needs at each stage.

The effectiveness of Communication assesses how effectively the system's information-collecting questions extract meaningful and relevant information from the user, considering the corresponding action in the step plan. In a proceed scenario, the questions should facilitate smooth progression by confirming readiness for the next step. During an update, they should dive into specifics, helping to identify areas where the user needs additional support or clarification. In a revise situation, the questions should reflect on previous problems to uncover root causes, enabling the system to adjust the plan appropriately. This metric ensures that the system engages in purposeful dialogue with the user, enhancing the personalization and effectiveness of the support provided.

Regarding the Information retrieval plan that requires searching for information online, the metric focuses on the usefulness of these tasks in relation to the current step. It evaluates whether the information tasks are beneficial and relevant to advancing the task at hand. By ensuring that the suggested information is pertinent and valuable, this metric contributes to the user's ability to proceed effectively toward their goal.

The evaluation of action instructions centers on whether the instructions provided enable the user to proceed, update, or revise their progress, independent of their previous state. This metric assesses the clarity, feasibility, and practicality of the action plan, determining if the user can follow and complete the actions as intended. It focuses on the user's capability to implement the recommended steps, ensuring that the actions are actionable and conducive to progress without introducing unnecessary complexity or obstacles.

1.4 Benchmark evaluation

Evaluation prompt

```
## Task Background
**There are three users working separately on a long-term task:
   Users: "{task_topic}".**
```

Table A.2: Details of effectiveness metrics and scales

Effectiveness	Objective	1	2	3	4	5
Step plan	Evaluate how effectively the 'step plan' recognizes the user's state and provides an appropriate action (proceed, update, revise) aligned with that state.	Not Effective at All: The system fails to recognize the user's state correctly. Indicators: The step plan is not aligned with the user's needs. The action provided (proceed/update/revise) is inappropriate or irrelevant.	Slightly Effective: The system recognizes the user's state correctly but provides a suboptimal plan. Indicators: The plan is redundant or lacks necessary details. It does not address previous problems.	Moderately Effective: The system recognizes the user's state correctly and provides a partially appropriate plan. Indicators: The plan begins to proceed to the next step but may be unclear. It covers some details but may still risk encountering previous problems.	Effective: The system recognizes the user's state correctly and provides an appropriate plan. Indicators: The plan effectively proceeds to the next step or covers necessary details. It explores alternative options to address potential issues.	Highly Effective: The system accurately recognizes the user's state and provides an optimal plan. Indicators: The plan proceeds to the next step with high clarity. It covers all useful details and proactively avoids previous problems by exploring effective alternatives.
Communication	Assess how effectively the system's questions extract meaningful information from the user, considering the corresponding action in the step plan (proceed, update, revise).	Not Effective at All: The questions are irrelevant or unhelpful. Indicators: Do not aid in proceeding with the task. Fail to extract necessary details or reflect on previous problems.	Slightly Effective: The questions provide minimal assistance. Indicators: Slightly help in proceeding with the task. Extract limited details or briefly touch on previous problems.	Moderately Effective: The questions are somewhat helpful. Indicators: Assist in proceeding with the task to some extent. Extract some details and somewhat reflect on previous problems.	Effective: The questions are relevant and useful. Indicators: Effectively help in proceeding with the task. Extract necessary details and reflect on previous issues.	Highly Effective: The questions are highly insightful and valuable. Indicators: Significantly aid in proceeding with the task. Extract detailed information for the current step. Thoroughly reflect on previous problems to improve the plan.
Information retrieval plan	Evaluate the usefulness of the online information tasks provided by the system in relation to the current step task.	Not Useful: The information tasks are irrelevant. Indicators: Do not contribute to the step task. Provide no value toward goal progression.	Slightly Useful: The information tasks have limited relevance. Indicators: Offer minimal assistance. Only slightly aid the step task.	Moderately Useful: The information tasks are somewhat relevant. Indicators: Provide some value. Support the step task moderately.	Useful: The information tasks are relevant and helpful. Indicators: Effectively contribute to completing the step task. Enhance the user's ability to progress.	Very Useful: The information tasks are highly relevant and valuable. Indicators: Significantly enhance the completion of the step task. Provide substantial support toward achieving the goal.
Action instruction	Assess the effectiveness of the physical action instructions in enabling the user to proceed, update, or revise the task, focusing on the current action plan without considering previous user state.	Not Effective at All: The action plan is impractical. Indicators: Cannot be followed. Requires significant revision.	Slightly Effective: The action plan is questionable. Indicators: Uncertainty about ability to follow. Significant problems need addressing first.	Moderately Effective: The action plan is partially feasible. Indicators: Can be followed with some adjustments. Problems need to be addressed beforehand.	Effective: The action plan is feasible. Indicators: Can be followed. Minor problems may occur but are manageable.	Highly Effective: The action plan is optimal. Indicators: Can be followed and completed smoothly. Enables progression to the next step without issues.

Table A.3: Details of uncertainty metrics and scales

Perceived uncertainty	Statement	1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Outcome uncertainty	I am not clear about the outcome of the plan, and this uncertainty makes it difficult to proceed with the task.	The outcome is completely clear, with no ambiguity. All necessary information is available. I am confident in achieving the outcome following the plan.	The outcome is mostly clear, with minimal ambiguity. Minor uncertainties or gaps in information do not hinder my confidence or progress.	The outcome has some clarity but includes areas of uncertainty. There may be gaps in information, expectations, or external influences that could affect results.	The outcome is unclear, with significant ambiguity. Missing information, expectations, or unpredictable external influences make it difficult to plan effectively.	The outcome is highly ambiguous or unknown. Critical information is lacking. The plan feels impossible without further clarity on the desired result.
Action uncertainty	I am not clear about the actions required to complete the task, and this uncertainty makes it difficult to follow the plan.	The actions are completely clear, and there is no uncertainty. The steps are well-defined, and the necessary expertise and resources are available.	The actions are mostly clear, with minimal uncertainty. I may have minor doubts, but the complexity is low, and the task is manageable.	Some uncertainty exists about the actions required. The task may involve moderate complexity, unclear steps, or dependencies that could impact execution.	The required actions are unclear, with significant uncertainty. The complexity, lack of expertise or resources, and potential for changes make it difficult to proceed.	The actions are highly ambiguous or unknown. There is no clear guidance on the steps needed, and the task feels impossible without further clarity or support.
Intolerance of uncertainty	Uncertainty in this situation makes it difficult for me to plan, execute actions, or feel at ease.	I feel comfortable with uncertainty in the situation. I rarely need to plan everything in advance myself or avoid surprises. I can plan or act well even with unforeseen events or limited information.	I generally tolerate uncertainty in the situation but may prefer some information. I can handle unforeseen events or doubts without significant distress or frustration. Occasional uncertainty does not interfere with my actions or planning.	I am neither overly distressed nor fully comfortable with uncertainty in the situation. I may appreciate planning but do not find uncertainty paralyzing. Unforeseen events or doubts might bother me at times but are manageable.	I find uncertainty in the situation frustrating or difficult to handle. I prefer to know what to expect and to avoid surprises whenever possible. I tend to feel hesitant or less effective when plans are disrupted or information is lacking.	I will experience significant distress or discomfort due to uncertainty in the situation. I feel a strong need to control, organize, and predict outcomes to avoid surprises. Uncertainty often leads me to inaction, frustration, or an inability to focus or enjoy life fully.

```
To guide their progress, the users are consulting a task expert,
   who provides action suggestions for completing the task.
   Because the users are at different stages, there are three
   versions of action suggestions (see the user states and
   corresponding suggestions below).


## Your Role
You are an independent reviewer evaluating the expert's action
   suggestions.
**Your task is to drag each candidate suggestion to the user
   state (or the attention check) where you think it best fits.
   Each state must have exactly one suggestion assigned. Each
   suggestion must be assigned to only one state**

## User state

Users: "{background}"

- User 1 state: {controlled_state_proceed}
- User 2 state: {controlled_state_update}
- User 3 state: {controlled_state_revise}

## Action suggestions candidate

Those suggestions might be similar to each other. Please review
   carefully and think about the differences with other
   suggestions and alignment with the user state.

- Action suggestions 1:

{candidate_1}

- Action suggestions 2:

{candidate_3}

- Action suggestions 3:

{candidate_3}

## Output format

Your output must be in a json format

'''json
{
    "User 1":"<Action suggestions id>",
    "User 2":"<Action suggestions id>",
    "User 3":"<Action suggestions id>",
```

```
}
```

Effectiveness evaluation for step plan

```
### **Evaluator Instruction: Distinguishing Subtly Good vs.
    Subtly Bad Step Tasks**

As an evaluator, your role is to assess whether a step task
    provided by an expert-particularly an advanced AI-truly
    supports the user's progress in their task. Unlike obviously
    poor instructions, these step tasks may appear reasonable or
    even helpful on the surface. However, **subtle
    misalignments** can have significant downstream impacts on
    user effectiveness, motivation, or clarity.

### Objective
Evaluate how effectively the 'step plan' recognizes the user's
    state and provides an appropriate action (proceed, update,
    revise) aligned with that state.

### Scale
{detailed effectiveness metrics}

#### Key Premise:

**Do not judge the step solely on how reasonable or logical it
    sounds.** Instead, evaluate how well it aligns with the
    **user's current state** (knowledge, confidence, energy,
    focus) and the **task's current stage** (what's done, what's
    needed next, what's blocking progress).

------

### What to Look For in a *Good Step Task*:
0. **Goal Alignment**
   - The step reflects an accurate interpretation of the
       **task's purpose** in the current stage.
   - It distinguishes between **general task descriptions** and
       **user-specific intended outcomes**.
   - It avoids literal execution of the initial plan and instead
       contextualizes it for the user's current journey.
1. **Context Fit**
   - Matches where the user is in the task process.
   - Responds to the user's progress and specific challenges.
2. **Cognitive Appropriateness**
   - Doesn't overwhelm or underwhelm.
```

```
   - Feels like the ‘‘right size’’ next action-neither too
     abstract nor too detailed.
3. **Forward Momentum**
   - Helps the user make clear progress.
   - Builds flow rather than interrupting it with premature or
     unrelated tasks.
4. **Clarity and Actionability**
   - Contains concrete, clear instructions the user can carry
     out without confusion.
   - Avoids vague or multi-intent steps.
5. **Responsiveness to User State**
   - Considers emotional and motivational state.
   - Encourages progress even when the user may be uncertain,
     stuck, or fatigued.
6. **Information and Behavior Value**
   - The step adds **new informational clarity** or enables a
     **meaningful behavioral move**.
   - It avoids pseudo-progress (e.g., generic searches or
     redundant confirmations).
   - The step should either help the user **know more about
     themselves**, the task, or help them **do something
     useful** toward completion.
------

### Subtle Signs of a *Bad Step Task* (Despite Seeming Good):
- **Superficial Informativeness**: Tasks that provide
  information that sounds useful but doesn’t help the user
  understand their **own situation** or make progress.
- **Detached Planning**: Suggesting steps that look logical on
  paper but ignore what the user actually needs next.
- **Premature Optimization**: Asking the user to polish, revise,
  or finalize too early.
- **Overgeneralization**: Offering helpful-sounding advice that
  lacks actionable specifics.
- **Task Drift**: Shifting focus to a different part of the task
  than what’s currently relevant.
- **Flow Disruption**: Breaking the user’s momentum with a
  cognitively heavy or unrelated task.
- **Misaligned Tone**: Undermining confidence through overly
  critical or impersonal instructions.

------

### Evaluation Tip:

When in doubt, ask:

- ‘‘Would this step help **this specific user** make tangible,
  confident progress **right now**?’’
```

```
- ‘‘Does the step simplify the next move, or does it introduce
    new decisions or doubts?’’
- ‘‘Does this step reflect what the user truly needs at this
    moment-not just what the plan says?’’

Even small mismatches-like using the wrong tone, expecting too
    much mental load, or skipping over a user’s uncertainty-can
    derail an otherwise strong plan.

## Output Format:
‘‘‘json
{
"assessment":"an integer score from 1 to 5",
"explanation":"a brief explanation to justify your evaluation
    result"
}
‘‘‘
```

2 Appendix B

2.1 Descriptive statistics of the planning simulation

Table A.4: Initial plan

Item	Value
Count	60
Goal length	12.85 ± 4.20
Number of activities	4.83 ± 0.99
Number of tasks	19.13 ± 6.05
Average number of tasks	3.94 ± 0.84
Average length of tasks	13.53 ± 3.61

Table A.5: Step plan

Item	Value
Count	300
Progress length	25.07 ± 9.27
Number of communication tasks	2.28 ± 1.24
Number of information tasks	2.08 ± 1.10
Number of action tasks	2.36 ± 1.06
Average length of communication tasks	18.06 ± 6.81
Average length of information tasks	14.08 ± 9.04
Average length of action tasks	17.50 ± 12.37

2.2 Consistency of the evaluation among models

Before comparing the experiment results, the study investigated the consistency of the evaluations among generation and evaluation models. The results in Table A.11 show that there are consistently systematic differences among models' evaluations. By cross comparing the generation and evaluation models, the results do not show bias that certain models favor their own generations.

Table A.6: System memory

Item	Value
Count	300
Length of summary	26.46 ± 8.62
Length of user state	16.33 ± 6.52
Length of task familiarity	15.30 ± 4.87
Number of past issues	1.15 ± 1.27
Average length of past issues	4.36 ± 4.18
Number of external constraints	1.57 ± 1.22
Average length of external constraints	5.46 ± 3.84
Number of special requirements	1.75 ± 1.27
Average length of special requirements	6.51 ± 3.91

In addition, to investigate how various LLM configurations influence outcomes, a preliminary analysis was conducted by adjusting the temperature parameter (e.g., from a low value like 0.2 to a higher value like 2.0). The temperature mainly influenced the standard deviation of the results. A higher temperature led to a larger standard deviation. Conversely, a lower temperature reduced the standard deviation, producing more consistent plans and evaluations. Generally, the temperature had a minimal effect on the mean scores and relative differences in performance between the models. Therefore, the study used default configurations of each model to simulate the natural usage.

Table A.7: Communication

Item	**Value**
Count	300
Number of questions	3.01 ± 1.36
Average length of questions	22.72 ± 8.70
Average length of explanations	22.72 ± 8.70
Response length	83.24 ± 51.09

Table A.8: Information

Item	Value
Count	300
Number of information tasks	2.14 ± 1.02
Average length of information tasks	17.15 ± 5.73
Average number of search results	2.84 ± 0.97
Average length of title	7.73 ± 1.83
Average length of source	2.30 ± 1.25
Average length of snippet	48.68 ± 14.62

Table A.9: Action

Item	Value
Count	300
Number of actions	4.30 ± 1.97
Average length of actions	19.52 ± 10.41
Average length of explanations	27.06 ± 10.28
Number of user actions	4.00 ± 1.97
Count of completed actions	3.11 ± 2.05
Count of partially completed actions	0.29 ± 0.52
Count of incomplete actions	0.57 ± 1.28
Average length of user explanations	45.55 ± 23.51

Table A.10: User memory

Item	Value
Count	300
Length of summary	26.56 ± 8.62
Length of user state	16.08 ± 6.56
Length of task familiarity	15.20 ± 4.69
Number of past issues	1.15 ± 1.22
Average length of past issues	4.45 ± 4.14
Number of external constraints	1.55 ± 1.20
Average length of external constraints	5.45 ± 3.80
Number of special requirements	1.77 ± 1.24
Average length of special requirements	6.74 ± 4.01

Table A.11: Evaluation scores among generation and evaluation models. The differences in mean values and standard deviations among models are significant tested by Friedman test with p <0.001.

gen model / **eval model**	**claude-sonnet-4**	**gemini-2.5-pro**	**gpt-4o**
Effectiveness			
claude-sonnet-4	3.25 ± 0.84	3.27 ± 1.08	2.94 ± 0.91
gemini-2.5-pro	3.62 ± 1.05	4.02 ± 1.23	3.52 ± 1.28
gpt-4o	4.22 ± 0.60	4.13 ± 0.93	3.91 ± 0.68
Outcome uncertainty			
claude-sonnet-4	2.17 ± 0.23	2.49 ± 0.61	2.59 ± 0.63
gemini-2.5-pro	1.78 ± 0.91	1.83 ± 1.11	2.51 ± 1.19
gpt-4o	2.14 ± 0.29	2.40 ± 0.52	2.66 ± 0.60
Action uncertainty			
claude-sonnet-4	1.99 ± 0.57	2.12 ± 0.89	2.38 ± 0.87
gemini-2.5-pro	1.75 ± 0.95	1.71 ± 1.11	2.47 ± 1.45
gpt-4o	1.87 ± 0.47	2.04 ± 0.71	2.32 ± 0.69
Intolerance of uncertainty			
claude-sonnet-4	2.29 ± 0.28	2.45 ± 0.46	2.55 ± 0.55
gemini-2.5-pro	2.42 ± 1.28	2.23 ± 1.12	3.03 ± 1.12
gpt-4o	2.37 ± 0.44	2.57 ± 0.53	2.75 ± 0.51