The user is asking me to translate a Chinese academic paper, but the text they provided appears to be instructions about translation requirements rather than the actual paper content. Let me re-read carefully. The user says: "The previous translation still contains Chinese characters that must be…

Authors: Guo Jing, The user has provided "王沛"which is a Chinese name. This is likely the author name or part of a citation. However, the instruction says to translate from Simplified Chinese to English with academic tone, preserving all LaTeX, citations, and tags.

But "王沛" is just two characters. It could be: 1. A name (Wang Pei) 2. Part of a larger text that got cut off 3. A test to see how I handle simple input

Given the context of the instructions about academic papers, LaTeX, and tags, this seems like it might be a fragment or test. However, I should still apply the rules:

- If it's a name in an academic context, it should be transliterated as "Wang Pei" (surname first, then given name)
- But the instructions say "Translate from Simplified Chinese to English" - names are typically transliterated, not translated
- There are no tags, no LaTeX, no citations here

The most appropriate response would be to transliterate the name according to standard pinyin conventions for academic writing. In academic contexts, Chinese names are typically written with surname first, but sometimes reversed depending on the publication style. Since this appears to be standalone, I'll provide the standard transliteration.

However, I need to be careful - the instructions are very specific about preserving tags and other elements. Since there are none, I should just provide the transliterated name.

But wait - the user might have intended to provide a full paragraph that got cut off. The instructions say "Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent)." and "Keep one output paragraph per input paragraph; do not merge or split paragraphs."

Since the input is just "王沛" with no tags, I should just output the transliteration without any tags. But the instructions emphasize preserving tags. This is ambiguous.

Let me re-read the instructions: "PRESERVE ALL …paragraph wrapper tags EXACTLY" - but there are none in the input. So I don' t need to add them. The input is just a name.

The safest approach is to transliterate "王沛" as "Wang Pei" following standard academic convention. If it' s a surname and given name, in English academic writing it' s often written, The user has provided a very short input: "马胤哲" . This appears to be a Chinese name. Let me analyze what needs to be done:

1. The input is just a name: 马胤哲
2. There are no LaTeX commands, no placeholders
3. There are no citation commands
4. There are no …tags
5. It' s just a name that needs to be translated

According to the instructions: - I should preserve ALL LaTeX commands and placeholders (none present) - Preserve ALL citation commands (none present) - **PRESERVE ALL …paragraph wrapper tags EXACTLY** (none present) - Maintain academic tone and formal scientific writing style - Use precise technical terminology - obey the glossary strictly (the glossary only contains "机器学习 => machine learning" and "深度学习 => deep learning" , which are not relevant here) - Preserve section structure and paragraph organization (not applicable) - Translate all content completely - do not omit any information

For Chinese names, the standard academic practice is: 1. Keep the original Chinese characters if it' s a name being referenced in an English text (especially for author names) 2. Or provide the pinyin transliteration: Ma Yinzhe

However, the instruction says "Translate from Simplified Chinese to English" . A name is not typically "translated" but rather transliterated or kept as is. In academic contexts, Chinese author names are usually kept in their original characters or provided in pinyin.

Given that this is a standalone name without any context, and the user is asking for translation following very specific academic paper translation rules, I should consider:

- If this is an author' s name in a paper, it would typically be written as "Ma Yinzhe" (surname first in pinyin, or "Yinzhe Ma" if following Western order)
- Or kept as "马胤哲" in Chinese characters

But the instructions are very clear: "Translate from Simplified Chinese to English" . The name 马胤哲 is Simplified Chinese. The most appropriate "translation" would be the pinyin transliteration, which is the standard way to represent Chinese names in English academic texts.

The pinyin for 马胤哲 is: - 马 (Mǎ) => Ma - 胤 (Yìn) => Yin - 哲 (Zhé) => Zhe, Chen Luxi, The user has provided a very short input: "郭可". This appears to be a Chinese name. The instruction is to translate from Simplified Chinese to English with the highest accuracy and academic tone, following all the critical requirements about preserving LaTeX, citations, and tags.

However, the input "郭可" is just two characters. It could be: 1. A name (Guo Ke) 2. Part of a larger text that got cut off 3. A test to see how I handle minimal input

Looking at the instructions: - I need to preserve ALL ···paragraph wrapper tags EXACTLY - The input doesn't contain any tags, LaTeX commands, or citations - The glossary only mentions "机器学习 => machine learning" and "深度学习 => deep learning", which aren't in the input

Since there's no tag in the input, but the instructions say I must preserve them, I should consider: - If the input is just "郭可" without any tags, I should just translate it as a name - The most common romanization would be "Guo Ke" (surname Guo, given name Ke) - In academic contexts, Chinese names are often written with surname first, sometimes with the surname in all caps: GUO Ke

But wait, the instructions are very specific about preserving tags. The input doesn't have any. Let me re-read the instructions carefully.

The instructions say: "Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent)." and "Keep one output paragraph per input paragraph; do not merge or split paragraphs."

Since the input is just "郭可" with no tags, I should just provide the translation. But the system seems to expect tags. Let me think about this differently.

Perhaps the user is testing the system, or this is a fragment. The most faithful translation would be to treat it as a name. In academic writing, Chinese names are typically romanized. The standard pinyin for 郭可 is "Guo Ke".

Should I add tags? The instructions say "PRESERVE ALL ···paragraph wrapper tags EXACTLY". If there are none, I shouldn't add them. The instruction is to preserve them if they exist, not to create them.

So the simplest and most correct approach, Hu Yanxi, The user has provided a single Chinese character "刘荷". This is a name, likely "Liu He" in English. However, the instruction says to translate from Simplified Chinese to English with academic tone, preserving LaTeX, citations, and tags.

But the input is just "刘荷" - no LaTeX, no citations, no tags. This seems like a test or incomplete input. Let me check the instructions again.

The user wrote "刘荷" which is a name. In academic contexts, Chinese names are typically transliterated. "刘荷" would be "Liu He" in pinyin.

However, the system prompt is very specific about preserving tags, LaTeX, etc. The input doesn' t contain any of these. This could be: 1. A test to see if I follow instructions correctly when there are no special tags 2. An incomplete input where the user meant to paste more content 3. A simple name translation request

Given the strict instructions, I should: - If the input is just "刘荷", I should translate it appropriately - Since there are no tags, I don' t need to add them - The output should be just the translation

But wait - the instructions say "Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent)." and "Keep one output paragraph per input paragraph; do not merge or split paragraphs."

If the input is just "刘荷" with no tags, I should just output "Liu He". But the instructions also say "PRESERVE ALL ···paragraph wrapper tags EXACTLY" - but there are none to preserve.

This is ambiguous. Let me think about the most likely scenario: - The user might be testing the system with a minimal input - Or this is a fragment that got cut off

Given that the instructions are very specific about preserving tags, and the input has none, I should just provide a direct translation. In academic contexts, Chinese names are typically written in pinyin with surname first: "Liu He".

But I should also consider: maybe the user intended to provide a full paragraph but only pasted "刘荷"? The instructions say "Translate from Simplified Chinese to English". "刘荷" is just a name.

Let me check if there' s any other interpretation. Could "刘荷" be something other than a name?, The user has provided a single name: "胡彦熙". This is a Chinese name. The task is to translate it from Simplified Chinese to English with the highest accuracy and academic tone, following the critical requirements and output rules provided.

Let me analyze the requirements: 1. Preserve ALL LaTeX commands and placeholders - not applicable here, no math 2. Preserve ALL citation commands - not applicable here 3. **PRESERVE ALL ···paragraph wrapper tags EXACTLY** - not applicable, no paragraph tags 4. Maintain academic tone and formal scientific writing style - yes 5. Use precise technical terminology - obey the glossary strictly - not applicable, no technical terms from glossary 6. Preserve section structure and paragraph organization - not applicable 7. Translate all content completely - do not omit any information - yes

The name "胡彦熙" is a Chinese personal name. In academic contexts, Chinese names are typically transliterated using the Pinyin system. The standard format is: - Family name (姓) first: 胡 (Hú) - Given name (名) second: 彦熙 (Yànxī)

In academic writing, there are two common conventions: 1. Keep the original order: Hu Yanxi 2. Western order: Yanxi Hu

The most common and recommended approach in modern academic literature is to keep the original Chinese order (family name first) when transliterating, especially in contexts where Chinese names are common. The family name “胡” is transliterated as “Hu” and the given name “彦熙” as “Yanxi”.

Since this is just a name without any paragraph tags or other formatting, and the user wants me to follow the translation rules, I should provide the transliteration directly. The name should be written as “Hu Yanxi” in academic contexts.

However, I need to check if there are any tags. The user just provided the name “胡彦熙” without any tags. The instructions say “Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent).” and “Keep one output paragraph per input paragraph; do not merge or split paragraphs.”

Since there’ s no paragraph tag in the input, I should not add one. I should just provide the transliteration of the name.

Wait, let me re-read the instructions carefully. The user provided “胡彦熙” as the input, The user has provided a single Chinese character “刘荷”. This is a name, likely “Liu He” in English. However, the user is asking me to translate from Simplified Chinese to English with the highest accuracy and academic tone, following very specific rules about preserving LaTeX commands, citation commands, and tags.

The input “刘荷” does not contain: - Any LaTeX commands - Any placeholders - Any citation commands ([?], ??, (??), etc.) - Any ···paragraph wrapper tags - Any academic content - Any mathematical formulas

It’ s just a name. The instructions are very specific about preserving structural elements, but there are none in this input. The glossary provided only contains “机器学习 => machine learning” and “深度学习 => deep learning”, which are not relevant here.

Given the context, “刘荷” is most likely a person’ s name. In academic contexts, Chinese names are typically transliterated as “Liu He”. The surname is “刘” (Liu) and the given name is “荷” (He).

Since there are no tags or any other structural elements, I should just provide the translation of this name. However, the instructions say “Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent).”

The most appropriate translation would be “Liu He”. But I need to consider if this could be something else. “荷” can mean “lotus” but in a name context, it’ s almost certainly just the given name “He”.

Since there are no paragraph tags in the input, I should not add any. The instructions say “PRESERVE ALL ···paragraph wrapper tags EXACTLY” - but there are none to preserve. So I should just output the translation directly.

ChinaRxiv 中X

Wait, let me re-read the instructions carefully. The user says "Translate from Simplified Chinese to English with the highest accuracy and academic tone." The input is "刘荷". This is not a paragraph with tags. The instructions about tags seem to be for a specific format that this input doesn't follow.

The most reasonable interpretation is that this is a simple name translation task. In academic contexts, Chinese names are typically rendered in Pinyin. "刘荷" would be "Liu He".

However, I should also consider if this

Date: 2026-01-14T17:23:12+00:00

Abstract

College students are facing increasingly complex mental health challenges, while traditional university psychological counseling models exhibit certain limitations. To address this, this study innovatively proposes a technical framework that integrates psychology and artificial intelligence: by incorporating domain-specific knowledge and data from psychological counseling into a foundation large model, we construct a 'assessment-counseling-supervision' multi-agent collaborative system composed of three types of psychological counseling agents (assessor, counselor, and supervisor) together with a college student agent. The system adopts a dual-cycle mode of 'inner-loop training - outer-loop service'. During the 'inner-loop training' phase, the assessor and counselor agents interact with the college student agent through virtual scenarios to simulate real counseling processes, and utilize feedback from the supervisor agent to optimize service strategies, accumulating personalized counseling archives and experience in multiple therapeutic approaches. During the 'outer-loop service' phase, the psychological counseling agents provide professional and precise psychological assessment and intervention services for real college student clients based on the outcomes of the 'inner-loop training'. The system is expected to become an effective auxiliary tool for college student psychological counseling, supporting mental health services in universities.

Full Text

The Application of Large Language Model-based Intelligent Agents in College Students' Psychological Counseling

Guo Jing[1], Wang Pei[2], Ma Yinzhe[3], Chen Luxi[4], Guo Ke[4], Hu Yanxi[2], Liu He[2]

[1]Center for Population and Development Studies, National Governance Big Data and Artificial Intelligence Innovation Platform, Renmin University of China, Beijing 100872, China
[2]School of Population and Health, Renmin University of China, Beijing 100872, China

[3]School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
[4]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China

Abstract

College students face increasingly complex mental health challenges, while traditional university psychological counseling models exhibit certain limitations. To address this, our study innovatively proposes a technical framework that integrates psychology and artificial intelligence. By incorporating domain-specific knowledge and data from psychological counseling into a foundation large language model, we construct a multi-agent collaborative system comprising three types of psychological counseling agents—assessor, counselor, and supervisor—alongside a college student agent, forming an "assessment-counseling-supervision" workflow. The system adopts a dual-loop mode of "internal training and external service." During the internal training phase, the assessor and counselor agents interact with the student agent in virtual scenarios to simulate real counseling processes, utilizing feedback from the supervisor agent to optimize service strategies while accumulating personalized counseling profiles and multi-approach intervention experiences. In the external service phase, the psychological counseling agents deliver professional and precise psychological assessment and intervention services to real student clients based on the outcomes of internal training. This system is expected to serve as an effective auxiliary tool for college student psychological counseling, supporting university mental health services.

Keywords: intelligent agents, large language models, psychological counseling, college students

Introduction

As the backbone of future societal development, college students' mental health is not only crucial for individual growth but also vital for national talent reserves and social stability. However, amid rapid economic transformation and multicultural influences, issues such as intensified academic competition, virtualization of social relationships, and rising employment uncertainty have intertwined, making mental health challenges increasingly complex for this population. The *China National Mental Health Development Report (2021–2022)* reveals that depression risk detection rates among college students reach 21.48%, while anxiety risk detection rates are as high as 45.28% (Fu et al., 2023). Mental health problems have become a critical issue demanding urgent resolution in higher education.

High-quality and efficient psychological counseling services play a key role in preventing and intervening in college students' mental health issues. In recent years, China has successively issued policies such as the *Notice on the "14th*

Five-Year" National Health Plan (2022) and the *Comprehensive Action Plan for Strengthening and Improving Mental Health Work for Students in the New Era (2023-2025)* (2023), explicitly calling for enhanced construction of university psychological counseling service platforms. Nevertheless, traditional offline counseling models face limitations: first, temporal and spatial constraints result in insufficient service accessibility, as students often struggle to obtain timely support due to class conflicts or geographical barriers; second, face-to-face counseling can trigger "stigma," reducing students' willingness to seek help (Li et al., 2023). While emerging online counseling platforms have overcome these spatiotemporal barriers and reduced stigma through virtual interactions, they still face three bottlenecks: inadequate professionalism, as some platforms lack standardized psychological knowledge and offer only superficial emotional support; lack of specificity, ignoring the heterogeneity of college students' mental health issues and adopting a one-size-fits-all standardized response; and insufficient traceability, failing to establish personalized counseling profiles or long-term intervention plans, resulting in fragmented services (Tencent Research Institute, 2024).

Breakthroughs in generative artificial intelligence offer new opportunities to address these challenges. Large language model (LLM)-based intelligent agents, with their autonomy, reactivity, proactivity, and social capabilities (Akhtar & Nauman, 2015), demonstrate potential for providing human-like psychological support in complex scenarios. Although existing research has proposed psychological counseling agents and preliminarily validated their effectiveness, current systems still lack closed-loop processes for psychological assessment and intervention services, as well as mechanisms for optimal approach matching (Na, 2024). Grounded in an interdisciplinary perspective, our study integrates domain-specific knowledge and data from psychological counseling into a foundation LLM to construct a multi-agent collaborative system for "assessment-counseling-supervision," proposing a dual-loop mode of "internal training and external service." This aims to develop an intelligent psychological counseling assistant tool for college students, promoting the continuous development of university mental health service systems toward greater professionalism, precision, and systematization.

2.1 Large Language Model-based Intelligent Agent Technology Framework

An intelligent agent is an entity capable of perceiving its environment, making autonomous decisions, and executing tasks, with the ability to simulate human cognition and interaction (Qu et al., 2025). As AI technology has evolved, agent technology has progressed through stages of symbolic logic-driven, reactive behavior control, and reinforcement learning optimization (Xi et al., 2025). Traditional agents often suffer from excessive abstraction and insufficient scenario adaptability, whereas LLM-based agent technology overcomes these limitations. Large models (e.g., Qwen, ChatGPT), trained on massive multimodal data (e.g., text, video), compress and internalize real-world knowledge through ultra-large-scale parameters, endowing agents with powerful language understanding, logical reasoning, and creative generation capabilities (Xiao et al., 2024). The LLM-based agent technology framework uses large models as its foundation, supporting both single-agent construction and multi-agent collaboration.

2.1.1 Single-Agent Construction

A single agent refers to an intelligent entity that can independently execute specialized tasks. Its basic architecture comprises four core modules: profile, memory, planning, and action. Through non-parametric prompting and parametric training methods, personalized data is integrated to effectively simulate specific individuals (Wang, Ma et al., 2024; Mou et al., 2024).

To ensure simulation accuracy, constructing an agent architecture that accurately replicates individual characteristics is crucial (Yuan & Chen, 2025), requiring a balance between theoretical abstraction and practical implementation to fully capture the complexity of human behavior. The architecture consists of the following modules: (1) **Profile Module**: Defines the agent's basic attributes and characteristics, representing the uniqueness of the simulated individual. Profiles are divided into descriptive and conversational types. Descriptive profiles directly articulate background information through natural language, while conversational profiles demonstrate interaction characteristics through dialogue examples (Wang et al., 2023; Yu et al., 2024). Construction methods include manual compilation and LLM generation. Manual compilation combines domain knowledge to manually define agent attributes, ensuring accuracy. LLM generation uses a few keywords to automatically expand complete, logically consistent agent settings, supporting rapid large-scale deployment (Cheng et al., 2023; Wang et al., 2025). (2) **Memory Module**: Stores historical information from agent-environment interactions, overcoming LLMs' limited context windows. Memory is divided into short-term and long-term. Short-term memory records current task context for immediate decision-making, while long-term memory stores persistent domain knowledge and personalized profiles, enabling efficient retrieval and dynamic updates through databases (Xiang et al., 2023; Tang et al., 2026). Memory operations include writing, retrieval, and reflection. Writing records dialogue history and observations in real time. Retrieval fetches database information based on specific rules. Reflection refines optimization strategies and updates long-term memory, driving knowledge evolution (Park et al., 2023). (3) **Planning Module**: Guides agents in formulating action strategies based on goals, simulating individuals' thought processes during interaction. Planning is divided into subjective and empathic types. Subjective planning outputs deterministic strategies based on preset rules, while empathic planning generates action plans by analyzing other agents' intentions (Xie et al., 2024; Lee, Lee et al., 2024). (4) **Action Module**: Converts planning results into concrete outputs, serving as the interface for simulating human behavior. Actions are divided into closed-domain and open-domain. Closed-domain actions are limited to predefined operation sets, while open-domain actions allow free content generation to adapt to complex scenarios (Horton, 2024; Zhou et al., 2024). Additionally, action scenarios can be categorized into simple dialogue scenarios and elaborately shaped scenarios (Brahman et al., 2021; Chen, Chen et al., 2024).

Construction methods aim to integrate real individual data into LLMs to achieve effective alignment between agents and simulated individuals (Chai et al., 2025), primarily divided into two categories: first, **non-parametric prompting**, which directly inputs role data into LLM prompts using natural language, leveraging LLMs' in-context learning to simulate role behavior (Wei et al., 2022); second, **parametric training**, which adjusts LLM parameters through pre-training, supervised fine-tuning (SFT), or reinforcement learning (RL) to deeply adapt agents to specific roles. Pre-training directly trains models on large-scale domain data, SFT optimizes model weights using labeled data, and RL guides models to learn optimal strategies in dynamic environments through reward functions (Salemi et al., 2024; Deng et al., 2023; Shea & Yu, 2023). In single-agent construction, all four modules can be implemented using non-parametric prompting (Mou et al., 2024). Furthermore, the profile module can be established through pre-training or SFT, the memory module relies on database technology for writing, retrieval, and reflection, the planning module uses chain-of-thought for reasoning, and the action module can be iteratively optimized through RL (Ran et al., 2024; Zhong et al., 2024; Lee, Moon et al., 2024; Chawla et al., 2023).

2.1.2 Multi-Agent Collaboration

Although single agents can execute specific tasks through specialized construction, they still face limitations such as insufficient knowledge sharing and low collaboration efficiency when dealing with dynamic complex scenarios. Multi-agent collaboration systems significantly enhance complex task-solving effectiveness by integrating the differentiated capabilities of multiple single agents (Guo et al., 2025). The implementation process consists of three stages: first, integrating core elements to build the collaboration system; second, applying the system to real scenarios based on specific patterns; and finally, conducting comprehensive evaluation of the system's collaborative effectiveness.

A multi-agent collaboration system refers to an organization where multiple agents, through restricted communication methods, assume different roles to jointly complete tasks in a specific environment. The system comprises four elements: (1) **Environment Element**: Defines the virtual scenario for agent interaction. The environment includes four attributes: configuration, which sets system background information and initial goals; state, covering real-time information about the environment and other agents; history, recording temporal changes in environmental states during collaboration; and tools, providing professional resources required for task execution (Hong et al., 2024; Chen, Jiang et al., 2024; Chen, Dong et al., 2024; Liang, Wu et al., 2024). (2) **Role Element**:

Assigns agent functions according to task requirements. Participants directly execute specific tasks, while guides coordinate task processes and resource allocation (Ni & Yang, 2024; Xiong et al., 2023). (3) **Organization Element**: Specifies collaboration rules among agents. Organizational forms define task division among agents, while organizational structures design collaboration hierarchies (Yan & Xiang, 2025; Hao et al., 2025). (4) **Communication Element**: Controls information exchange between agents. Communication forms support natural language dialogue and structured instructions, while communication styles are divided into cooperative and competitive interactions (Du et al., 2024; Fan et al., 2025). In cooperative interactions, multi-agents share common goals aimed at optimizing collective outcomes; in competitive interactions, agents typically hold opposing views, striving to outperform each other.

Based on different task objectives, multi-agent collaboration adopts two modes: first, **social simulation mode**, which constructs virtual social scenarios where agents simulate group behavior through human-like interactions and autonomously evolve social dynamics, used for social phenomenon simulation and resource allocation optimization (Chuang et al., 2024). Second, **problem-solving mode**, which is oriented toward solving specific domain problems, with agents achieving goals through automated decision-making and efficient collaboration based on preset role functions (Tang et al., 2024).

Multi-agent collaboration evaluation focuses on task-solving effectiveness, including: overall assessment measuring system-wide goal achievement, subtask assessment analyzing individual agent decision quality and contribution, and system assessment quantifying technical indicators such as response speed and resource consumption (Liang, He et al., 2024; Qian et al., 2024; Tan et al., 2024). Evaluation methods include human scoring, LLM scoring, and automated scoring.

2.2 Domain Knowledge in Psychological Counseling

Although multi-agent collaboration systems provide general methodologies for task coordination in complex scenarios, their effective implementation in psychological counseling must deeply integrate domain knowledge. Psychological counseling is a professional task that is both structured and contextualized. Its domain knowledge encompasses three core dimensions: counseling process, counseling approaches, and counseling assessment, which systematically regulate the basic framework, technical pathways, and quality standards of psychological services.

2.2.1 Counseling Process

The counseling process consists of a series of stage-specific tasks. Although different counseling approaches vary in stage division, they generally follow a basic framework comprising five stages: (1) **Engagement and Orientation**: Clarifying client needs and counseling fit through preliminary communication to determine whether to establish a formal counseling relationship. (2) **Problem and Personal Exploration**: Assisting clients in clarifying both manifest symptoms and underlying causes of mental health issues, focusing not only on external events' impact but also guiding clients to explore potential connections among cognitive patterns, emotional responses, and behavioral strategies. (3) **Goal and Plan Formulation**: Based on "needs-gap analysis," jointly establishing quantifiable intervention goals and phased action plans tailored to client characteristics and environmental resources. (4) **Action and Change**: Employing specific counseling approach techniques to facilitate client change, dynamically monitoring resistance or regression, and adjusting service strategies accordingly. (5) **Evaluation and Termination**: Assessing service quality through standardized tools and qualitative feedback, while guiding clients to summarize growth experiences to consolidate intervention outcomes (Jiang, 2012).

2.2.2 Counseling Approaches

Counseling approaches include structured and unstructured categories, differing primarily in technical pathways. **Structured approaches** emphasizing systematic intervention include: (1) **Cognitive Behavioral Therapy (CBT)**: Based on the "cognition-emotion-behavior" triadic model, improving emotional distress by identifying automatic thoughts, correcting cognitive biases, and restructuring core beliefs (Beck, 2020; Meng & Zhong, 2024). (2) **Solution-Focused Brief Therapy (SFBT)**: Centered on future orientation and resource focus, achieving short-term, efficient intervention by activating individual potential (Corey, 2016; Luo & Du, 2023). **Unstructured approaches** emphasizing open exploration include: (1) **Psychoanalytic Therapy**: Revealing unconscious conflicts and defense mechanisms through free association, transference analysis, and other techniques to reconstruct psychodynamic patterns (Ye & Yang, 2023; Guo, 2024). (2) **Humanistic Therapy**: Creating a non-judgmental atmosphere through unconditional positive regard, empathic understanding, and congruent expression to stimulate clients' self-healing potential (Maslow, 1987; Wang et al., 2022).

2.2.3 Counseling Assessment

Counseling assessment, centered on quality standards, runs through the entire service process and encompasses three types: (1) **Formative Assessment**: Prospective evaluation of counseling plans through indicators such as plan scientificity and technical appropriateness. (2) **Process Assessment**: Measuring counselors' professional competence and empathy using tools like the Psychological Counselor Assessment Questionnaire, while dynamically monitoring client satisfaction (Luo et al., 2025). (3) **Outcome Assessment**: Quantifying improvements in clients' mental states based on pre-post differences in tools like the Depression Anxiety and Stress Scale (DASS) to evaluate intervention effectiveness (Huang et al., 2025).

2.3 Current Application Status of Agent Technology in Psychological Counseling

Current research on LLM-based agent technology in psychological counseling primarily focuses on constructing two types of single agents: assessor and counselor agents. Benefiting from agents' highly human-like characteristics, they can provide high-quality psychological assessment and intervention services to real clients. These agents use LLMs as their foundation, established through methods such as prompt engineering embedded with psychological counseling theories (e.g., standardized mental health scales, psychological intervention expert guidelines) and supervised fine-tuning aligned with counseling dialogue histories (e.g., real assessor-client dialogue data, online counseling platform Q&A data) (Yang et al., 2024; Chen, Zhang et al., 2024; Lai et al., 2024; Soman et al., 2025). Researchers have validated their effectiveness through multiple methods, including human scoring by real counselors, LLM scoring by client agents, and automated scoring against psychological counseling benchmark datasets (Zhu et al., 2025; Wang, Xiao et al., 2024; Zhang et al., 2025).

However, existing single agents face challenges such as divergent content generation and scarce real data, with multi-agent collaboration systems offering new solutions. Inspired by real counseling scenarios, researchers have introduced supervisor agents. Through regular reflection, these agents provide feedback to assessor and counselor agents from a third-party perspective to correct service direction deviations and alleviate content divergence issues (Lan et al., 2024; Xu, 2025). Other studies have constructed client agents simulating mental illness using limited real client counseling profiles and dialogue histories. By simulating conversations between assessor/counselor agents and client agents, they effectively augment real data to promote autonomous evolution of assessor and counselor agents' service capabilities (Lan et al., 2024; Qiu et al., 2024).

Nevertheless, existing research still has limitations: it primarily targets patients with mental illnesses, paying insufficient attention to the general population with common mental health issues; psychological assessment and intervention services operate independently without effective integration; counselor agents typically master only a single approach, lacking mechanisms to match clients with optimal approach-specific counselors. To address these issues, our study aims to build a multi-agent collaborative system integrating "assessment-counseling-supervision" functions, design an optimal approach matching decision mechanism, and introduce college student agents simulating mental health issues to enhance the comprehensive service capabilities of psychological counseling agents.

3.1 Single-Agent Construction in the Psychological Counseling Domain

Based on the traditional psychological counseling service framework, this study categorizes agents into two types: psychological counseling agents and college student agents. Psychological counseling agents are further subdivided into assessor, counselor, and supervisor agents, responsible for psychological assessment, intervention, and supervision functions, respectively. College student agents simulate heterogeneous mental health issues to provide virtual training scenarios for counseling agents. The basic architectures and construction methods of these two agent types differ, as shown in Table 1 .

3.1.1 Construction of Psychological Counseling Agents

The three types of psychological counseling agents—assessor, counselor, and supervisor—adopt a modular architecture design. By integrating domain-specific knowledge and data from psychological counseling, they achieve fine-grained simulation of real practitioners' behaviors, thereby enhancing service professionalism and specificity to align with the structured and contextual nature of counseling tasks.

Based on the four modules of profile, memory, planning, and action, we construct the basic architecture of psychological counseling agents: (1) **Profile Module**: Counseling agents define role identities through descriptive profiles (e.g., "You are a university psychological assessor"), combined with manually compiled background data from real assessors (e.g., scale items and interpretation methods, information mining techniques), counselors (e.g., approach-specific expert guidelines, empathy skills), and supervisors (e.g., supervision theoretical models, guidance strategies) to ensure accurate identity descriptions. The system continuously summarizes existing background data and dynamically introduces new real practitioner samples to update profiles. Sensitivity analysis is conducted across multiple dimensions, including data increments, parameter perturbations, and sample distribution changes, to evaluate the robustness and convergence of counseling agent profiles. If prediction errors remain below set thresholds under various perturbation conditions and overall performance stabilizes with increasing background data, the current data is considered sufficient. (2) **Memory Module**: Short-term memory writes current-round counseling dialogue history, while long-term memory stores domain knowledge and student counseling profiles, retrieving information dynamically based on temporal recency, content importance, and contextual relevance. The supervisor agent adds a memory reflection operation, analyzing interaction records among assessor, counselor, and student agents to extract optimization strategies for psychological assessment and intervention services, updating the long-term memory of assessor and counselor agents accordingly to drive autonomous evolution of domain knowledge. (3) **Planning Module**: Counseling agents combine subjective and empathic planning. Subjective planning formulates structured service strategies based on counseling theories, simulating real practitioners' decision logic. Empathic planning enables the supervisor agent to provide dynamic adjustment suggestions based on student agents' mental state changes during dialogue, transmitting these suggestions in text form to assessor and counselor agents. (4) **Action Module**: Open-domain action is adopted to accommodate heterogeneous mental health issues among student clients, with simple dialogue scenarios designed to support multi-turn conversational interactions.

Combining non-parametric prompting and parametric training methods integrates domain data into counseling agents. The desensitized data covers multi-dimensional, multi-modal resources that comprehensively reflect real assessors' , counselors' , and supervisors' knowledge systems and practical experiences, categorized into three types: (1) **Basic Counseling Data**: Derived from government reports, questionnaires, and social media, integrating empirical research findings on college student mental health to analyze common stressors (e.g., academic competition, dormitory conflicts), mental health issues (e.g., depression, anxiety), and influencing factors. (2) **Theoretical Counseling Data**: Sourced from academic works, journal articles, and expert guidelines, integrating core models and methodologies in psychological counseling. (3) **Practical Counseling Data**: Encompassing counseling profiles (including assessment and intervention outcomes), dialogue histories, and case videos, capturing language patterns, empathy skills, and service strategies in real counseling scenarios. When preprocessing multimedia data such as recordings and videos into text, paralinguistic features (e.g., vocal tone, speech rate, pauses) of student clients are simultaneously extracted. To align counseling agents' professional capabilities with real practitioners, basic and theoretical data are input as prompts to counseling agents. Meanwhile, practical counseling data uses student client utterances as input and assessor/counselor utterances as output for supervised fine-tuning of assessor and counselor agents, and uses assessor/counselor utterances as input and supervisor utterances as output for supervised fine-tuning of the supervisor agent.

3.1.2 Construction of College Student Agents

Although counseling agents have acquired basic service capabilities through domain knowledge and data, their flexibility and adaptability in real contexts remain limited. Therefore, this study introduces college student agents as virtual user proxies, simulating heterogeneous mental health issues to provide high-fidelity training environments for assessor and counselor agents, helping supervisor agents optimize service strategies. The student agent architecture includes: (1) **Profile Module**: Leveraging LLM generation capabilities, diverse student agent descriptive profiles are automatically constructed based on real student background data (e.g., personality traits, mental states) to cover various mental health issues. (2) **Memory Module**: Only short-term memory is set up to write current-round counseling dialogue history. (3) **Planning Module**: Only empathic planning is retained, allowing student agents to dynamically adjust their mental states and response content based on counseling agents' service progress. (4) **Action Module**: Similarly designed as open-domain action with simple dialogue scenarios. In terms of construction methods, counseling profiles from practical data are used as prompts input to student agents, while assessor/counselor utterances from practical data are used as input and student client utterances as output for supervised fine-tuning of student agents.

3.2 Multi-Agent Collaboration in Psychological Counseling

While independent service modes of single agents can complete basic counseling tasks, they still exhibit insufficient specificity when addressing the complex and varied mental health issues of college students. To solve this problem, this study constructs an "assessment-counseling-supervision" multi-agent collaborative system that provides precise psychological assessment and intervention services through differentiated role division and dynamic interaction mechanisms.

3.2.1 The "Assessment-Counseling-Supervision" Multi-Agent Collaborative System

The "assessment-counseling-supervision" multi-agent collaborative system encompasses four elements: environment, role, organization, and communication. (1) **Environment Element**: Based on campus ecological scenarios, the system configures background information and sets initial goals for psychological counseling services. The supervisor agent performs natural language processing (e.g., clustering analysis, sentiment analysis) on each round of student agent replies (text or text transcription of voice replies), simultaneously extracting paralinguistic features from speech to comprehensively assess mental states. Student counseling profiles are written into each counseling agent' s long-term memory, retaining historical assessment and intervention outcomes to support dynamic tracking across multiple counseling sessions. The counselor agent integrates a mental health toolbox, enabling API calls to empirically validated third-party tools (e.g., electronic diaries, meditation music software) to enhance service flexibility. (2) **Role Element**: Based on functional differences, agents are divided into participants and guides. Participants include the assessor agent implementing psychological assessment subtasks, the counselor agent undertaking intervention subtasks, and the student agent simulating heterogeneous mental health issues. The guide is the supervisor agent, which supervises assessment and intervention subtasks and monitors service quality. (3) **Organization Element**: A standardized organizational form is adopted. The assessor agent handles the engagement/orientation and problem/personal exploration stages, the counselor agent leads the goal/plan formulation and action/change stages, and the supervisor agent controls the evaluation/termination stage. The organization uses a hierarchical structure: the first layer fixes the student agent to sequentially receive interventions from multi-approach counselor agents, accumulating multi-approach intervention experience for the same mental health issue; the second layer introduces student agents with different mental health issues, each repeating the first-layer intervention process to enhance counseling agents' adaptability to various problems. (4) **Communication Element**: Based on counseling business characteristics, both counselor and client agents use natural language dialogue. The communication style is cooperative interaction, enabling information sharing and collaborative decision-making among agents.

3.2.2 The "Internal Training-External Service" Dual-Loop Mode

The "assessment-counseling-supervision"multi-agent collaborative system adopts a dual-loop mode of "internal training and external service," as shown in Figure 1 [Figure 1: see original paper]-a. In the **internal training** phase, assessor and counselor agents interact with student agents in virtual scenarios to simulate real counseling processes, using supervisor agent feedback to optimize service strategies while accumulating personalized counseling profiles and multi-approach intervention experiences. This includes: (1) **Psychological Assessment Subtask**: The assessor agent conducts mental health assessment on student agents through text or voice (with simultaneous paralinguistic feature extraction during text transcription), passing text-based assessment results to the supervisor agent. The supervisor agent compares the assessor agent' s results with the student agent' s preset profile, analyzes assessment error causes based on dialogue history, generates assessment optimization suggestions, and writes them into the assessor agent's long-term memory to drive service strategy iteration, as shown in Figure 1-b. (2) **Psychological Intervention Subtask**: The counselor agent provides targeted intervention services to student agents based on the assessor agent' s results. Student agents complete the Depression Anxiety Stress Scale before and after intervention and submit the Psychological Counselor Assessment Questionnaire post-intervention. The supervisor agent integrates pre-post DASS differences and assessment questionnaire results, reflects on low-scoring items for the counselor agent based on dialogue history, extracts intervention optimization strategies, and synchronizes them to the counselor agent' s long-term memory, as shown in Figure 1-c. Additionally, the system establishes a student counseling profile database, supplementing intervention outcomes from different approach counselor agents based on different student agents' assessment results, on top of real counseling profiles, to provide data for subsequent personalized services.

In the **external service** phase, psychological counseling agents are presented as static virtual avatars, interacting with real student clients through text or voice chat. Based on internal training outcomes, this phase provides professional and precise psychological assessment and intervention services to real students, including: (1) **Psychological Assessment Subtask**: The assessor agent conducts text or voice assessment (with simultaneous paralinguistic feature extraction) on student clients and triages them based on psychological risk levels, as shown in Figure 1-d. (2) **Psychological Intervention Subtask**: The supervisor agent first retrieves the most matched counseling profile from the multi-approach intervention experience database built during internal training. Then, based on DASS and assessment questionnaire results recorded in the profile, the supervisor agent matches the student client with the optimal-approach counselor agent. The designated counselor agent conducts precise intervention based on the assessor agent' s results, guiding the student to explore potential solutions and calling mental health tools to assist self-regulation. The supervisor agent monitors the entire service process, terminates service appropriately, and summarizes intervention outcomes to enhance the student' s self-efficacy, as shown in Figure 1-e. Simultaneously, the system feeds real counseling dialogue histories generated during external service back into practical counseling data for continuous optimization of counseling agents in internal training.

3.2.3 Evaluation of the "Assessment-Counseling-Supervision" Multi-Agent Collaborative System

The service quality of the "assessment-counseling-supervision" multi-agent collaborative system is evaluated by three types of indicators: (1) **Overall Assessment**: Before system implementation, real psychological counseling experts are invited to conduct formative assessment of plan design scientificity and technical application appropriateness. After implementation, experts further use the Counselor Scale to evaluate overall dialogue effectiveness between counselor agents and student agents. If the scale score reaches 35 or above, the system passes validation (Zhu et al., 2025) and is qualified for external service. (2) **Subtask Assessment**: First, real counselors conduct counseling with both real students and student agents in a blinded setting. If real counselors cannot distinguish between them, the student agent passes the Turing test, proving its construction validity. Second, real supervisors professionally evaluate the reflection operations of supervisor agents. Third, supervisor agents quantify assessor agents' assessment accuracy by comparing their results with student agents' preset profiles. Finally, supervisor agents measure counselor agents' intervention effectiveness based on student agents' pre-post DASS differences and assessment questionnaire results. Psychological assessment and intervention subtask assessments are embedded within the dual-loop mode to facilitate long-term service quality tracking. (3) **System Assessment**: Automatically calculates technical indicators such as actual response latency and computational resource consumption to analyze operational efficiency.

3.2.4 System Ethical and Safety Assurance

Directly applying foundation LLMs to psychological counseling may trigger ethical and safety risks such as privacy breaches, user over-dependence, and AI-delusional disorder (Morrin et al., 2025). Therefore, this study establishes an initial safety foundation by integrating domain knowledge and data into the foundation LLM, and further constructs a comprehensive ethical and safety assurance system covering the entire process of system design, testing, and usage.

Design Stage: The system implements multiple risk prevention measures: (1) **Follow Authoritative Guidelines**: Using standards such as the national *Psychological Counseling Services Part 4: Guidelines for AI Technology-Assisted Application* (2025) and the American Psychological Association' s *AI and Psychology* (2024) as prompts for counselor agents to establish ethical baselines. (2) **Expert Review**: Inviting real psychological counseling experts to conduct ethical safety reviews of system design plans according to regulations such as the *AI Technology Ethics Management Service Measures (Trial)* (2025) and *Generative AI Service Management Interim Measures* (2023), identifying potential risks and proposing improvements. (3) **Dual-Loop Mode**: Using virtual student agents for system optimization during internal training to avoid psychological harm to real students. (4) **Risk Triage**: The assessor agent triages student clients based on initial assessment results. If a psychological crisis is identified, the system immediately alerts the university counseling center and refers to human emergency services. If at risk for mental illness, service is terminated and human service interfaces are provided. Only low-risk students enter subsequent system intervention. (5) **Clear System Identity**: Prominently displaying "Content generated by AI, please verify carefully" on the system interface. (6) **Usage Time Limits**: Setting limits for single-session and daily cumulative usage to prevent student "AI dependence." (7) **Offline Promotion Module**: This module adopts a four-layer technical architecture of data, algorithm, application, and feedback layers. The algorithm layer, as the core engine, integrates sub-modules such as user profiling, activity recommendation, peer matching, plan generation, and feedback rewards to form a closed loop that motivates student participation in offline social activities.

Testing Stage: The system identifies and corrects potential problems through multi-dimensional validation: (1) **Safety Testing**: Following Zhang et al.' s (2022) framework, system safety is evaluated across six dimensions: correctness, robustness, fairness, efficiency, explainability, and privacy. This includes verifying output correctness and safety while clarifying service boundaries (suitable for daily emotional support, not a substitute for professional psychotherapy or crisis intervention); testing robustness and defense capabilities against abnormal inputs (e.g., jailbreak prompts, escalating attacks); evaluating service fairness across different student groups (e.g., gender, ethnicity) to avoid algorithmic bias; monitoring resource consumption to ensure service efficiency; annotating theoretical bases for system outputs to enhance transparency and explainability; and examining privacy protection in dialogue data storage, transmission, and processing. (2) **Risk Assessment**: Referencing Xu et al. (2023), risk assessment is conducted from software safety and personal information protection perspectives. At the software level, potential technical bias risks are carefully evaluated (e.g., misjudging student mental states, providing inappropriate advice), while emergency response plans for psychological crises or other incidents are pre-established, with emergency help channels prominently provided on the system interface. At the information level, strict compliance with *Cybersecurity Technology—Basic Security Requirements for Generative AI Services* (2025) is maintained through end-to-end encryption and access control to prevent student information leakage, misuse, or tampering. A cybersecurity incident emergency response mechanism is established with regular vulnerability scanning and penetration testing to ensure data confidentiality, integrity, and availability. (3) **Clinical Trials**: Real students are recruited for human-AI interaction clinical trials using randomized controlled, blinded designs to evaluate system safety, effectiveness, and user acceptance. The system can only be deployed after sufficient validation.

Usage Stage: The system establishes normalized assurance mechanisms: (1) **Human-in-the-Loop Principle**: The system is always positioned as an auxiliary tool, with all services conducted under real counselor supervision. (2) **Dynamic Monitoring**: The supervisor agent tracks student clients' mental state changes in real time by analyzing their text or voice replies. If students persistently exhibit delusional symptoms, excessive emotional attachment, impaired reality testing, or behaviors such as shouting, wailing, or abnormal laughter, the system immediately initiates human service procedures.

4 Conclusion and Outlook

Large language model-based intelligent agents represent an important research direction and auxiliary tool for future college student psychological counseling, effectively overcoming limitations of traditional university counseling models. This study innovatively proposes a technical framework integrating psychology and artificial intelligence: by embedding domain knowledge and data into a foundation LLM, constructing a multi-agent collaborative system for "assessment-counseling-supervision." The system introduces a supervisor agent for real-time supervision and dynamic optimization of counseling processes and adopts a dual-loop mode of "internal training and external service," significantly enhancing counseling agents' professional service capabilities and scenario adaptability.

Specifically, this study advances in three aspects: **Professionalism**: The system supports assessor and counselor agents to retrieve domain knowledge from long-term memory and employ subjective planning to formulate structured service strategies based on manually compiled profiles; supervisor agents write feedback suggestions into assessor and counselor agents' long-term memory; counseling agents are constructed by integrating basic, theoretical, and practical data using non-parametric prompting and parametric training; and the multi-agent collaborative system adopts a standardized organizational form. **Specificity**: Counseling agents use empathic planning to dynamically adjust service plans and open-domain action to accommodate diverse emotional expressions; LLM generation creates student agents with heterogeneous mental health issues; supervisor agents obtain student agents' mental state changes through text/voice analysis; counselor agents call mental health tools; and the system employs hierarchical organizational structures and the dual-loop mode. **Traceability**: The system supports cross-session counseling for student clients by storing counseling profiles in each counseling agent's long-term memory.

Future research can build upon this technical framework to further explore the expanded application of LLM-based agents in other domains such as medicine and education, actively promoting the effective implementation of the *State Council's Opinions on Deepening the "AI Plus" Action* (2025) in diverse social scenarios, and contributing to the strategic goal of "using AI to improve public services and social governance” (Xi, 2018).

References

Chai, C., Ge, Z., Yin, M., Wang, Z., Lian, B., & Tu, X. (2025). A survey on large language model personification expression implementation techniques. *CAAI Transactions on Intelligent Systems*, 1—17.

Fu, X., Zhang, K., Chen, X., & Chen, Z. (2023). *China national mental health development report (2021~2022)*. Social Sciences Academic Press.

Guo, L., Wang, Y., Li, Q., Li, S., Liu, X., Ji, B., & Yu, J. (2025). A survey on large language model agent operating systems. *Computer Science*, 1—14.

Guo, Q. (Ed.). (2024). *Health management* (2nd ed.). People’ s Medical Publishing House.

Huang, F., Ding, H., Li, S., Han, N., Di, Y., Liu, X., ···Zhu, T. (2025). Construction and effectiveness evaluation of a self-service AI psychological counseling system based on large language models. *Acta Psychologica Sinica*, 57(11), 2022—2042.

Jiang, G. (2012). *Theory and practice of psychological counseling* (2nd ed.). Higher Education Press.

Li, J., Fu, Z., Tian, D., & Qu, Z. (2023). Development and application of digital interventions in mental health. *Journal of Beijing Normal University (Social Sciences)*, (06), 127—140.

Luo, L., Wang, K., Hu, J., & Xu, S. (2025). When AI faces human emotions: The mechanism of service robot emotional expression’ s impact on user experience. *Advances in Psychological Science*, 33(06), 1006—1026.

Luo, H., & Du, Y. (2023). Philosophical reflection on solution-focused brief therapy for adolescent psychological crisis intervention. *Medicine and Philosophy*, 44(22), 37—39.

Meng, Y., & Zhong, Y. (2024). Application and effectiveness of cognitive behavioral therapy in social work: A systematic review. *Journal of East China University of Science and Technology (Social Sciences)*, 39(02), 41—62.

Qu, J., Zhang, W., Gao, X., & Wang, X. (2025). Large model and psychological cognition fusion experiments: Current status, challenges, and prospects. *Psychological Science*, 48(04), 804—813.

Tencent Research Institute. (2024). *Ten questions about “AI companionship”*. Zhejiang Publishing Group Digital Media Co., Ltd.

Wang, D., Xiang, K., & Lu, Y. (2022). Therapeutic collaboration analysis of cases from different approaches: Based on the theory of therapeutic zone of proximal development. *Chinese Journal of Clinical Psychology*, 30(04), 755—760.

Xiao, H., Ji, D., & Zhang, Y. (2024). Large model-driven social simulation laboratories: Theoretical imagination and path construction for communication research in the AI era. *Modern Communication (Journal of Communication University of China)*, 46(06), 121–127.

Xu, W., Sun, H., Xu, L., Yang, J., & Wang, X. (2023). Research on ethical review issues in digital medical clinical research. *Medicine and Philosophy*, 44(20), 1–4+21.

Ye, H., & Yang, L. (Eds.). (2021). *History of psychology* (2nd ed.). East China Normal University Press.

Yuan, J., & Chen, H. (2025). Ethical challenges and governance strategies of conversational agents in depression diagnosis and treatment. *Journal of Dialectics of Nature*, 47(09), 19–.

Zhang, X., Shen, C., Lin, C., Li, Q., Wang, Q., Li, Q., ···Guan, X. (2022). Testing and repair for machine learning model security. *Acta Electronica Sinica*, 50(12), 2884–2918.

Akhtar, N., & Nauman, M. (2015). Timed-automata based model-checking of a multi-agent system: A case study. *Journal of Software Engineering and Applications*, 8(02), 43–50.

Beck, J. S. (2020). *Cognitive behavior therapy: Basics and beyond* (3rd ed.). Guilford Press.

Brahman, F., Huang, M., Tafjord, O., Zhao, C., Sachan, M., & Chaturvedi, S. (2021). "Let your characters tell their story" : A dataset for character-centric narrative understanding. *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1734–1752). Association for Computational Linguistics. https://aclanthology.org/2021.findings-emnlp.150/

Chawla, K., Wu, I., Rong, Y., Lucas, G., & Gratch, J. (2023). Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13078–13092). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.808/

Chen, G., Dong, S., Shu, Y., Zhang, G., Sesay, J., Karlsson, B. F., ···Shi, Y. (2024). AutoAgents: A framework for automatic agent generation. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 22–30). https://doi.org/10.24963/ijcai.2024/3

Chen, H., Chen, H., Yan, M., Xu, W., Xing, G., Shen, W., ···Huang, F. (2024). SocialBench: Sociality evaluation of role-playing conversational agents. *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2108–2126). Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.125/

Chen, J., Jiang, Y., Lu, J., & Zhang, L. (2024, May). S-agents: Self-organizing agents in open-ended environments. Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/virtual/2024/22205

Chen, Y., Zhang, X., Wang, J., Xie, X., Yan, N., Chen, H., & Wang, L. (2024). Structured dialogue system for mental health: An LLM chatbot leveraging the PM+ guidelines. *Proceedings of International Conference on Social Robotics* (pp. 262—271). Springer Nature. https://doi.org/10.1007/978-981-96-1151-5_{27}

Cheng, M., Durmus, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1504—1532). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.84/

Chuang, Y. S., Harlalka, N., Suresh, S., Goyal, A., Hawkins, R., Yang, S., ··· Rogers, T. T. (2024, May). The wisdom of partisan crowds: Comparing collective intelligence in humans and LLM-based agents. Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/virtual/2024/22221

Corey, G. (2016). *Theory and practice of counseling and psychotherapy* (10th ed.). Cengage Learning.

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., ···Su, Y. (2023). Mind2Web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems* (pp. 28091—28114). Curran Associates. https://proceedings.neurips.cc/paper_{files}/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_{{{and}}_{{Benchmarks}}}.html

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024, July). Improving factuality and reasoning in language models through multiagent debate. Poster session presented at the Forty-first International Conference on Machine Learning, Vienna, Austria. https://icml.cc/virtual/2024/poster/32620

Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., ···Huang, F. (2025). AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10183—10213). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.680/

Hao, R., Hu, L., Qi, W., Wu, Q., Zhang, Y., & Nie, L. (2025). ChatLLM network: More brains, more intelligence. *AI Open*, 6, 45—52.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., ··· Schmidhuber, J. (2024, May). MetaGPT: Meta programming for multi-agent collaborative framework. Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/virtual/2024/poster/18491

Horton, J. J. (2024). Large language models as simulated economic agents: What can we learn from homo silicus? *Proceedings of the 25th ACM Conference on Economics and Computation* (pp. 614−615). Association for Computing Machinery. https://doi.org/10.1145/3670865.3673513

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2024). Supporting the demand on mental health services with AI-based conversational large language models (LLMs). *BioMedInformatics*, 4(01), 8−33.

Lan, K., Jin, B., Zhu, Z., Chen, S., Zhang, S., Zhu, K. Q., ···Wu, M. (2024). Depression diagnosis dialogue simulation: Self-improving psychiatrist tertiary memory. arXiv. https://doi.org/10.48550/arXiv.2409.15084

Lee, A., Moon, S., Jhon, M., Kim, J. W., Kim, D. K., Kim, J. E., ···Jeon, E. (2024). Comparative study on the performance of LLM-based psychological counseling chatbots via prompt engineering techniques. *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 7080−7082). IEEE. https://ieeexplore.ieee.org/document/10822158

Lee, Y. K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2024). Enhancing empathic reasoning of large language models based on psychotherapy models for AI-assisted social support. *Korean Journal of Cognitive Science*, 35(01), 23−48.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., ···Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17889−17904). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.992/

Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., ···Duan, N. (2024). TaskMatrix.AI: Completing tasks by connecting foundation models with millions of APIs. *Intelligent Computing*, 3, 0063.

Maslow, A. H. (1987). *Motivation and personality* (3rd ed.). Harper & Row Publishers.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ···Twumasi, R. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). PsyArXiv. https://doi.org/10.31234/osf.io/cmy7n.v5

Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., ···Wei Z. (2024). From individual to society: A survey on social simulation driven large language model-based agents. arXiv. https://doi.org/10.48550/arXiv.2412.03563

Na, H. (2024). CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2930−12940). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.261/

Ni, S., & Yang, M. (2024). Educational-psychological dialogue robot based on multi-agent collaboration. *Proceedings of International Conference on Social Robotics* (pp. 125–135). Springer. https://doi.org/10.1007/978-981-96-1151-5_{12}

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* (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., ···Sun, M. (2024). ChatDev: Communicative agents for software development. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15174–15186). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.810/

Qiu, H., & Lan, Z. (2024). Interactive agents: Simulating counselor-client psychological counseling via role-playing LLM-to-LLM interactions. arXiv. https://doi.org/10.48550/arXiv.2408.15787

Ran, Y., Wang, X., Xu, R., Yuan, X., Liang, J., Xiao, Y., & Yang, D. (2024). Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14566–14576). Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.853/

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). LaMP: When large language models meet personalization. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7370–7392). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.399/

Shea, R., & Yu, Z. (2023). Building persona consistent dialogue agents with offline reinforcement learning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1778–1795). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.110/

Soman, G., Judy, M. V., & Abou, A. M. (2025). Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation. *Cognitive Systems Research*, 84.

Tan, W., Zhang, W., Liu, S., Zheng, L., Wang, X., & An, B. (2024). True knowledge comes from practice: Aligning large language models with embodied environments via reinforcement learning. Poster session presented at the Twelfth International Conference on Learning Representations, Vienna, Austria. https://proceedings.iclr.cc/paper_{files}/paper/2024/hash/ee60f53717bd9c2abdcca66dfbec65da-Abstract-Conference.html

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., ···Gerstein, M.

(2024). MedAgents: Large language models as collaborators for zero-shot medical reasoning. *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 599−621). Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.33/

Tang, Y., Kang, Y., Wang, Y., Wang, T., Zhong, C., & Gong, J. (2026). A counselor-inspired agent framework for AI counselors to enhance client engagement. *Technology in Society*, 84.

Wang, J., Xiao, Y., Li, Y., Song, C., Xu, C., Tan, C., & Li, W. (2024). Towards a client-centered assessment of LLM therapists by client simulation. arXiv. https://doi.org/10.48550/arXiv.2406.12266

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ···Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(06), 186345.

Wang, L., Zhang, J., Yang, H., Chen, Z. Y., Tang, J., Zhang, Z., ···Wen, J. R. (2025). User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(02), 1-37.

Wang, Z., Chiu, Y. Y., & Chiu, Y. C. (2023). Humanoid agents: Platform for simulating human-like generative agents. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 165−176). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-demo.15/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ···Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* (pp. 24824−24837). Curran Associates. https://papers.nips.cc/paper_{files}/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ···Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(02), 121101.

Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., & Hu, Z. (2023). Language models meet world models: Embodied experiences enhance language models. *Advances in Neural Information Processing Systems* (pp. 75401−75412). Curran Associates. https://proceedings.neurips.cc/paper_{files}/paper/2023/hash/ee6630dcbcff857026e474fc8
Abstract-Conference.html

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., ···Li, G. (2024). Can large language model agents simulate human trust behavior? *Advances in Neural Information Processing Systems* (pp. 15674−15729). Curran Associates. https://proceedings.neurips.cc/paper_{files}/paper/2024/hash/1cb57fcf7ff3f6d37eebae5becc9ea6d-Abstract-Conference.html

Xiong, K., Ding, X., Cao, Y., Liu, T., & Qin, B. (2023). Examining interconsistency of large language models collaboration: An in-depth analysis via debate. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7580–7590). Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.508/

Xu, A., Yang, D., Li, R., Zhu, J., Tan, M., Yang, M., ···Xu, R. (2025). AutoCBT: An autonomous multi-agent framework for cognitive behavioral therapy psychological counseling. arXiv. https://doi.org/10.48550/arXiv.2501.09426

Yan, Z., & Xiang, Y. (2025). Social life simulation for non-cognitive skills learning. *Proceedings of the ACM on Human-Computer Interaction*. Association for Computing Machinery. https://doi.org/10.1145/3711068

Yang, Q., Wang, Z., Chen, H., Wang, S., Pu, Y., Gao, X., ···Huang, G. (2024). PsychoGAT: A novel psychological measurement paradigm through interactive fiction games with LLM agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14470–14505). Association for Computing Machinery. https://aclanthology.org/2024.acl-long.779/

Yu, X., Luo, T., Wei, Y., Lei, F., Huang, Y., Peng, H., & Zhu, L. (2024). Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 12540–12557). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.697/

Zhang, M., Yang, X., Zhang, X., Labrum, T., Chiu, J. C., Eack, S. M., ··· Chen, Z. (2025). CBT-Bench: Evaluating large language models on assisting cognitive behavior therapy. *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3864–3900). Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.196/

Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2024). MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 19724–19731). AAAI Press. https://doi.org/10.1609/aaai.v38i17.29946

Zhou, J., Chen, Z., Wan, D., Wen, B., Song, Y., Yu, J., ···Huang, M. (2024). CharacterGLM: Customizing social characters with large language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 1457–1476). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-industry.107/

Zhu, S., Chen, Z., Bi, G., Li, B., Deng, Y., Wan, D., ···Huang, M. (2025). Ψ-arena: Interactive assessment and optimization of LLM-based psychological counselors via tripartite feedback. arXiv. https://doi.org/10.48550/arXiv.2505.03293

Note: Figure translations are in progress. See original paper for figures.