IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024. Doi Number

An Innovative Solution to Design Problems: Applying the Chain-of-Thought Technique to Integrate LLM-based Agents with Concept Generation Methods

Shijun Ge[1] and Yuanbo Sun [1] Yin Cui [2*] Dapeng Wei[1]

[1] School of Design and Art, Beijing Institute of Technology, Beijing, 100081, China
[2] School of Design and Innovation, Shenzhen Technology University, Shenzhen, 518118, China

Corresponding author： Yin Cui （cuiyin@sztu.edu.cn）

ABSTRACT To enhance the application capabilities of large language models (LLMs) in conceptual design, this study explores how to achieve deep integration between LLM-based agents and concept generation methods using the chain-of-thought (CoT) technique and evaluates its feasibility. Using GPT-4 as a case study, we designed two agents: IntelliStorm (based on the unstructured brainstorming method) and EvoluTRIZ (based on the structured TRIZ method). Thirty participants were recruited, and through two experimental phases spaced one month apart, a comparative analysis of the effects of collaboration groups (human-agent vs. human-human) and concept generation methods (brainstorming vs. TRIZ) on participants' physiological activation and creative thinking performance were conducted. The results show that the involvement of LLM-based agents can effectively reduce participants' electrodermal activity(EDA)response levels, indicating a reduction in cognitive load. Moreover, these agents preserve the distinct physiological response patterns and performance advantages of the different concept generation methods. For example, IntelliStorm, like brainstorming, evokes stronger responses to information stimuli, demonstrating superior thinking fluency; EvoluTRIZ, like the TRIZ, exhibits a higher frequency of information responses, showcasing enhanced thinking elaboration. However, originality tends to favor human–human collaboration. The findings confirm that integrating LLMs with traditional concept generation methods is an effective strategy made possible by combining CoT and retrieval-augmented generation (RAG) technologies. In the future, LLM-based agents are expected to achieve broader application in the design field by incorporating additional concept generation methods.

INDEX TERMS LLM-based agent, chain-of-thought fine-tuning, concept generation method, EDA, human-agent collaboration, human-human collaboration

I. INTRODUCTION

During the design process, the conceptual design phase is considered the most flexible and creative stage[1], and its core focus is on search and exploration, which requires designers to coordinate the operation of long-term and short-term memory[2]. However, traditional design methods face limitations in terms of designers' personal knowledge and experience, thus leading to difficulties in the process of comprehensively browsing, memorizing, and retrieving reference materials[3]. These limitations can result in functional fixedness[4] and cognitive biases[5]. By leveraging their extensive knowledge repositories, LLMs can offer interdisciplinary insights and diverse perspectives, thereby compensating for designers' potential deficiencies in certain domains or specialized knowledge areas. LLMs can swiftly process and analyze vast amounts of information, and through continuous learning, they can adapt to new tasks, constantly updating and optimizing their reasoning abilities[6]. Consequently, LLMs demonstrate promising performance in the conceptual design phase [7]. Furthermore, they can generate fluent natural language responses [8] to interact with humans and have thus been recognized as potential collaborative partners for designers during the conceptual design stage [9]. Despite the advantages of LLMs in design collaboration, their performance remains inadequate due to a lack of design knowledge and design situational awareness[10]. If domain-specific knowledge

and techniques from the field of design could be integrated, and if LLMs were fine-tuned accordingly, their collaborative capabilities would be significantly enhanced [11].

To enhance the collaborative capabilities of LLMs in conceptual design, we have drawn upon various well-established concept generation methods from traditional paradigms. These methods have been proven to assist human thinking in a stepwise and principled manner, thereby reinforcing the concept generation process [12]. For example, brainstorming, as a classic nonstructured method, relies on intuition and intrinsic motivation to stimulate creativity [13] and can enhance divergent thinking abilities. In contrast, TRIZ offers a structured approach, providing a systematic set of tools and frameworks for problem solving and concept generation [14] and offering logical steps for rational thinking.

The question of whether these classic, traditional concept generation methods can be integrated with large language models remains largely unexplored. This study aims to address this gap by proposing the following hypothesis: Classic concept generation methods can be combined with LLMs, and after integration, their unique advantages can be demonstrated. To verify this hypothesis, we focus our exploration on three core research questions (RQs):

(RQ1) How can LLM-based agents be developed on the basis of traditional concept generation methods?

(RQ2) What impact do these LLM-based agents have on participants' design processes?

(RQ3) What impact do these LLM-based agents have on participants' design outputs?

To explore RQ1, our study utilized GPT-4, employing CoT fine-tuning and RAG technology to train two LLM-based agents for conceptual design: IntelliStorm (based on brainstorming) and EvoluTRIZ (based on the TRIZ). To investigate RQ2 and RQ3, we designed a comparative experiment that examined the performance of the participants while collaborating with different LLM-based agents (human–agent collaboration) and contrasted the results with those of the participants that collaborated with human designers (human–human collaboration). This comparison aimed to clarify whether agents could achieve or surpass the level of human collaboration in terms of the design output.

We recruited 30 graduate students from design programs, each possessing at least 5 years of design experience and 1 year of experience using LLMs. The experiment was conducted in two phases: In the first phase, participants collaborated with the two agents to complete two parallel design tasks. One month later, in the second phase, they completed the same tasks in collaboration with human partners. To capture data from these collaborations, we adopted a comprehensive research methodology. First, we monitored participants' electrodermal activity (EDA) during task execution to assess their physiological activation state. The EDA can be used to determine and classify changes in cognitive load[15],[16]. Its advantages include being noninvasive, easy to collect, and comfortable for users; this approach also provides comprehensive, objective, and continuous emotional information[17]. Additionally, we analyzed participants' creative outputs. Two experts were invited to score the works based on the four dimensions of creative thinking ability proposed by Torrance (1974)[18]: fluency, flexibility, originality, and elaboration. These criteria were used to evaluate participants' creativity performance in the tasks. This comprehensive assessment approach considers both physiological indicators and expert evaluations of creative outputs, providing a basis for a thorough understanding of the impact of LLM-based agents on the design process and outcomes.

The experimental results validated the applicability and advantages of integrating traditional concept generation methods with LLMs. This integration reduced participants' EDA response levels, indicating a decrease in cognitive load; it also preserved the physiological response characteristics and performance advantages of different concept generation methods while enhancing participants' thought fluency and flexibility. These findings demonstrate that through CoT fine-tuning, the logical output of LLMs can be effectively adjusted and optimized, further substantiating the usability of CoT fine-tuning techniques and potentially providing methodological references for the future construction of LLM-based agents for specific application scenarios by combining multiple concept generation methods.

Our work provides the following contributions:

(1) LLM-powered Design Assistants: This work integrates GPT-4 with concept generation methods, creating IntelliStorm and EvoluTRIZ, thereby expanding the application of innovative LLMs in conceptual design.

(2) Physiological Impact Analysis: The assistants' effects on cognitive load is elucidated by quantifying the participants' EDA response levels.

(3) Novel LLM Optimization Approach: A method combining CoT fine-tuning with the RAG is proposed to enhance LLM reasoning and output capabilities, providing explicit technical support for design-assisted development.

The remainder of this paper is organized as follows: Section II provides the relevant research background, Section III describes the agent construction process, Section IV details the experimental process, Section V presents the experimental results, and finally, we discuss the implications of this study in Section VI and conclude in Section VII.

II. RELATED WORK

A. DESIGN LLM-BASED AGENTS

LLMs utilize their vast knowledge bases and advanced reasoning capabilities to assist designers in the process of generating innovative solutions[7],[11],[19]. LLM-based agents primarily include the following core modules: the brain, perception, and action[20]. The perception module receives multimodal input from the external environment. By

employing its multimodal capabilities, the LLM-based agent converts the information received in various forms, such as numbers, audio, and images, into machine-understandable representations for processing by the brain. The brain is responsible for knowledge storage, memory, and the execution of key tasks such as information processing, decision-making, reasoning, and planning. Finally, the LLM-based agent may use a tool-calling module to execute corresponding actions and respond to external inputs[20]. This study focuses primarily on enhancing the agent's brain module in terms of natural language interaction, knowledge, memory, reasoning and planning as well as transferability and generalization[20]. To enable the agent to develop a deep understanding and contextual awareness of design products, it is necessary to augment the agent's design knowledge and to strengthen its design thinking and reasoning capabilities [21], [22].

1) EXPANDING DESIGN KNOWLEDGE RETRIEVAL

The RAG refers to a hybrid AI model architecture for natural language processing that combines retrieval and generation mechanisms [20]. The key advantage offered by RAG models lies in their ability to integrate retrieved external knowledge. When the agent is presented with a question, it first uses a retrieval system (such as dense passage retrieval (DPR)) to retrieve the most relevant texts from an augmented design knowledge base. These retrieved documents are then input as contextual information into a sequence-to-sequence generation model. This mechanism enables the RAG to generate more accurate and relevant outputs in response to complex scenarios, to expand the knowledge boundaries of LLMs, and to provide up-to-date information [23]. Previous researchers have used this technology primarily to create question-answering systems for the purpose of solving complex problems in specific domains. For example, Zhou *et al.*[20] used RAG technology to construct an enterprise knowledge management system solution, and Balaguer *et al.*[24] reported that combining an RAG with fine-tuning in LLMs can significantly improve the model's accuracy when answering domain-specific questions.

Therefore, this study focuses on RAG technology, relies on the expansion of excellent design cases as a dynamic retrieval case library, and uses famous design cases based on certain principles and methods as design knowledge. This approach allows the agent to generate responses or complete specified tasks by integrating an enhanced context.

2) ENHANCING DESIGN REASONING CAPABILITIES

A mere expansion of the scope of an agent's design expertise does not enhance the agent's reasoning ability; therefore, it is necessary to help LLMs engage in deep-level reasoning and thinking[25], which requires an enhancement of the thinking module's understanding of design thinking. Researchers have proposed guidance for LLMs in a step-by-step process to be used in response to problems by providing either examples or explicit instructions, thereby promoting the models' use of reasoning steps before reaching a final answer; this approach can significantly improve the models' performance in terms of reasoning tasks. CoT fine-tuning is an innovative prompting tool that aims to assist LLMs with deep-level reasoning and thinking. The concept of CoT was first proposed by Wei *et al.* [26] from Google Brain in January 2022; they reported that this approach can enhance the reasoning abilities of LLMs. The core of this technology involves guiding the model to demonstrate its reasoning process and the corresponding logical relationships clearly through step-by-step instructions, thereby achieving complex thinking processes [27]. The process involves breaking down complex reasoning into a series of simple steps and providing clear guidance at each step to help the model gradually develop a complete chain of reasoning[26]. This study adopts the fine-tuned CoT [28] to construct paradigms featuring good problems and reasoning chains, particularly by utilizing the reasoning capabilities of LLMs to guide smaller models through the process of solving design innovation tasks. Additionally, this technique offers the advantage of improving model performance while ensuring that the underlying language model parameters remain unchanged, thus conserving computational resources[26].

B. CONCEPT GENERATION TECHNIQUES

Concept generation methods influence human brain activation and creative output[29] and may also affect the information output of the agents. This study selected two concept generation methods with different levels of structure—brainstorming and TRIZ—for comparative research. These two methods differ significantly in terms of their structure, potentially leading to different fine-tuning effects on agents.

Brainstorming is a popular, intuitive, and unstructured creative generation technique[13] that involves stimulating participants' creativity and imagination through free association and open discussion. During brainstorming, participants are encouraged to propose as many ideas as possible without restrictions or criticism. This free and open environment helps break through mental barriers and stimulate creative thinking. However, the unstructured nature of brainstorming may lead to divergent and incoherent ideas, making it challenging to form systematic design solutions.

The TRIZ (Theory of Inventive Problem Solving) is a logical, structured innovation method proposed by the Soviet inventor Genrich Altshuller in the 1950s. The TRIZ is based on the analysis of many patents and technological innovation cases, summarizing a systematic set of innovation principles and methods [14]. The core idea is to innovate and solve problems by identifying and resolving technical contradictions, providing established steps and principles.

The structured nature of the TRIZ helps improve the systematicity and coherence of ideas but may also limit designers' free association and innovation space.

Scholars have also introduced the GAI into the brainstorming process so that problem solving is enhanced by providing additional stimuli, generating diverse ideas, and connecting seemingly unrelated concepts. For example, GAI-driven TRIZ tools can help designers identify contradictions and propose solutions[30]. These studies provide valuable insights into how the GAI empowers traditional concept generation methods, and offer inspiration on how large language models can combine different concept generation methods.

C. EDA

Due to advancements in physiological signal acquisition and processing technologies, emotion recognition based on physiological signals has been applied in various fields, including human–computer interaction, intelligent driving, entertainment, education, and clinical biomedicine[31]. Analyzing emotional changes with respect to physiological signals represents a more objective and persuasive approach. The EDA is an effective and noninvasive method that measures fluctuations in skin conductance caused by sweat secretion when individuals are exposed to information stimuli and stress and has been widely employed in emotion research [32]. Real-time EDA response levels are associated with emotional arousal and represent an instantaneous state of arousal related to the cognitive challenges encountered in specific situations or scenarios [33], [34]. The EDA can be used to assess cognitive load, which refers to the amount of information processing that is needed when performing a specific task. The impact of cognitive load on creativity is complex and multidimensional [15]. Nourbakhsh *et al*. (2013) [16] employed the EDA to classify cognitive load, and research indicates that as the cognitive load increases, the EDA response levels may also increase. These studies provide valuable insights related to the work in this paper.

III. CONCEPTUAL DESIGN AGENTS

We constructed two LLM-based agents based on ChatGPT-4 to collaborate with designers. The first agent, named IntelliStorm, implements a typical unstructured design method, brainstorming, which aims to stimulate the diversity and quantity of ideas through interactions with designers. The second agent, named EvoluTRIZ, employs a typical structured design method known as the TRIZ (Theory of Inventive Problem Solving).

We fine-tuned the LLM-based agents on the CoT technique by providing it with demonstration prompts for two concept generation methods (brainstorming and the TRIZ), covering their inputs, processes, and outputs. The model iteratively generates outputs by using previous results as part of the subsequent input, thereby simulating the problem-solving processes of these design methods to a certain extent.

The specific construction steps include designing the initial parameters, conducting fine-tuning training, and evaluating the optimization and iteration (Figure 1).

A. INITIAL PARAMETERS

Based on the creation guidelines from GPTs, this study identified key parameters such as name, opening dialog, dialog principles, and knowledge settings. We drew on the CO-STAR framework techniques proposed by the winner of the first GPT-4 Prompt Engineering Competition, which was

TABLE I
INTELLISTORM PARAMETER SETTINGS

Components	Descriptions
Name	IntelliStorm
Role	A concept designer with rich imagination, skilled in using brainstorming methods
Opening dialog	Hello, I'm IntelliStorm. Welcome to our creative discussion. To better understand the design problem, please provide context information related to the task, as well as your expected design objectives. Let's brainstorm together!
Dialog principles	Style: Enthusiastic dialog style Tone: Encourages bold ideas, generates solutions without constraints, and does not judge participants' thoughts Audience: Adopts a designer's communication approach, such as being user-centered and emphasizing application scenarios Response: Output content should be rich and full of imagination and creativity
Design knowledge	Static: 50 selected innovative design cases Dynamic: https://www.red-dot.org/, https://ifdesign.com/en/

TABLE II
EVOLUTRIZ PARAMETER SETTINGS

Components	Descriptions
Name	EvoluTRIZ
Role	A concept designer with rich interdisciplinary knowledge and proficient in the TRIZ methodology
Opening dialog	Hello, I'm EvoluTRIZ. Welcome to solving problems together using the TRIZ method. To better understand the design problem, you need to think about and describe the core contradiction of the problem (Context), as well as your ideal design objectives
Dialog principles	Style: Concise and clear discussion style Tone: Maintains an objective and neutral attitude without criticism or flattery Audience: Communicates in a style associated with engineering that is problem-oriented and focuses on logic Response: Output content should be concise and logically clear, avoiding overly detailed or lengthy expressions
Design knowledge	Static: Cases corresponding to TRIZ's 40 Inventive Principles Dynamic: http://epub.cnipa.gov.cn/

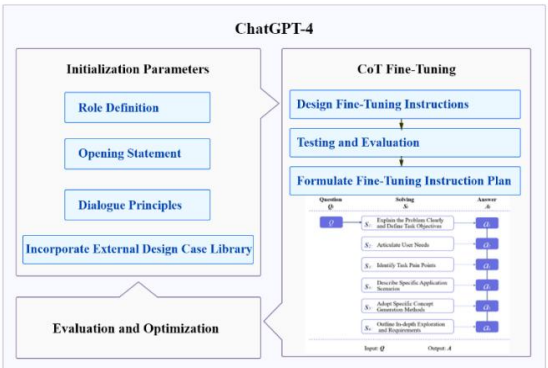

FIGURE 1. Construction agents.

organized by Singapore's Government Technology Agency[35]. Tables 1 and 2 show the parameter settings for IntelliStorm and EvoluTRIZ, respectively. IntelliStorm focuses on divergent thinking, whereas EvoluTRIZ represents logical thinking, resulting in significant differences in their parameter designs.

The external case library is divided into two parts: a dynamic design case library and a static classic design cases library. The dynamic design cases library is implemented via API technology. IntelliStorm integrates the data interfaces from the official websites of the Red Dot Design Award and IF Design Award into the intelligent agent's retrieval system, while EvoluTRIZ integrates data from the Chinese Patent Knowledge Base to achieve real-time updates and capture the latest design cases. Second, in terms of the static classic design case library, 50 classic design cases were selected for IntelliStorm. For EvoluTRIZ, cases corresponding to 40 inventive principles were chosen and then converted into text formats and uploaded to the GPT model.

B. FINE-TUNING TRAINING

The prompt used in the CoT technique is composed of a triad: (input, chain of thoughts, and output)[36]. Within this structure, the input denotes the example question Q; the CoT represents a series of intermediate natural language reasoning steps that lead to the final output, which can be denoted by the solution process S; and the output is the expected answer A. The prompt can be represented as

$$P = \{(Q_1,A_1,S_1),...(Q_k,A_k,S_k)\}$$

Given that the reasoning process is provided, it can be assumed that different stages of reasoning correspond to different expected answers, i.e., $\boldsymbol{A} = [a_1,...a_n]$. Thus, we have the following:

$$(A|[Q,P,\theta]\prod_{i=1}^{|A|} f_{PLM}(a_i|[Q,P,\theta])$$

According to Bayes' theorem, we can derive the following:

$$P(A \mid [Q,P]) = P(A \mid [Q,R,P])P(R \mid [P,Q]) \quad (1)$$

$$P(R|[Q,P] = \prod_{i=1}^{|R|} f_{PLM}(r_i|[Q,P,r_{<i}]) \quad (2)$$

$$P(A|[Q,P,R] = \prod_{j=1}^{|A|} f_{PLM}(a_i[Q,P,R,a_{<i}]) \quad (3)$$

In this context, a_i represents the expected output of the i-th process, while $|\boldsymbol{A}|$ denotes the total number of processes. Increasing the probability of the occurrence of A and S at each stage consequently enhances the performance of the CoT reasoning process [36].

Ten experienced designers were selected to engage in dialog with ChatGPT-4 to complete a collaborative task guided by the aforementioned strategies. The task was presented as follows: "You are traveling to the land of Oz. How can you efficiently communicate with locals when there is a complete language barrier?" The CoT process is illustrated in Figure 2. To construct the two agents, the 10 experienced designers were asked to complete the task

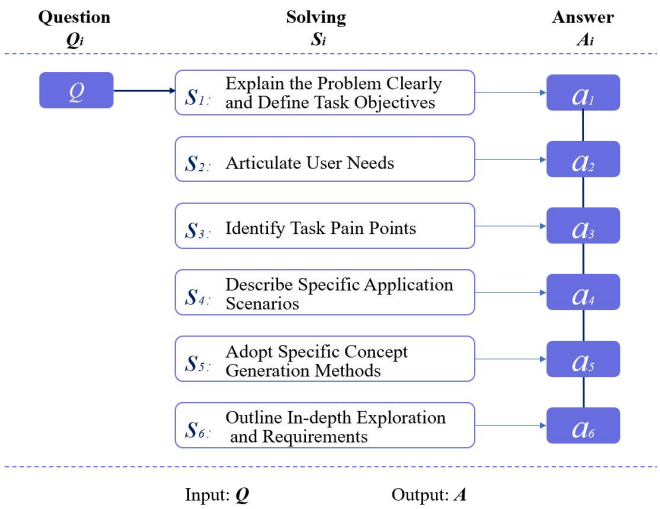

FIGURE 2. CoT training.

using both agents by extracting and optimizing their prompts.

C. EVALUATION AND ITERATION

To assess and optimize the performance of the agents, we invited 10 design experts (including 5 professors of design and 5 product designers with more than 10 years of experience) to evaluate the performance of the human-AI cocreation involving the fine-tuned version of ChatGPT-4. The fine-tuning training tasks for IntelliStorm and EvoluTRIZ were labeled I and E, respectively. On the basis of the five fundamental requirements of industrial design outlined by Professor Cheng Nenglin in "Introduction to Industrial Design" [37], namely, creativity, user satisfaction, mass production capability, public aesthetics, and social and environmental considerations, we established our evaluation criteria. As aesthetic appeal cannot be reliably assessed solely from descriptions, we employed the other four dimensions as the expert evaluation standards. Thus, the experts evaluated the tasks in terms of these four dimensions (based on a maximum score of 10 points per dimension) with the goal of assessing the effectiveness of the fine-

tuning process.The expert evaluation results are shown in Tables 3 and 4. The scoring results indicate that IntelliStorm and EvoluTRIZ performed well on the fine-tuning training tasks, particularly in terms of all four aspects. These outcomes are in line with the expected performance. However, the experts suggested adding a final step to the instruction structure: summarization and evaluation. After incorporating the scores and recommendations provided by the experts, the results obtained by the fine-tuning approach using CoT technology are shown in Tables 5 and 6.

TABLE III
EXPERT SCORING RESULTS REGARDING THE FINE-TUNING TRAINING TASKS FOR INTELLISTORM

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Mean score
User satisfaction	87	85	90	80	83	81	80	85	92	78	84.1
Scalability	82	91	85	82	82	77	86	79	80	80	82.4
Creativity	92	89	87	91	90	90	88	92	89	93	90.1
Social and environmental consciousness	82	83	86	83	85	80	80	76	73	82	81

TABLE IV
EXPERT SCORING RESULTS REGARDING THE FINE-TUNING TRAINING TASKS FOR EVOLUTRIZ

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	Mean score
User satisfaction	82	85	73	89	90	81	84	82	82	78	82.6
Scalability	86	79	86	82	85	79	86	80	83	81	82.7
Creativity	88	92	89	92	89	92	89	81	90	85	88.7
Social and environmental consciousness	82	76	86	85	92	85	80	87	85	90	84.8

TABLE V
FINE-TUNING SCHEME FOR INTELLISTORM

Steps	Scheme
1	For the purposes of this study, a role-playing scenario was established. The objective was for the AI to design a product that would meet the client's specified requirements. A role-playing scenario was implemented for this experiment. The researcher played the role of a client, while the AI system was assigned the role of a product designer. Within this scenario, the AI was tasked with providing [number] design proposals to meet the client's needs.
2	Which design proposal did you select, and how was it specifically implemented? Alternatively, which design proposal did you present, what are its innovative aspects, and how was it implemented?
3	How was the aforementioned [specific] functionality implemented?
4	Were any additional designs selected beyond those already discussed while ensuring that no previously mentioned concepts were repeated?
5	Do any approaches integrate [specific technology]? Alternatively, do any solutions incorporate cultural elements?
6	Please evaluate the aforementioned approaches from the perspectives of ***, ***, and *** and provide an in-depth analysis of these solutions.

TABLE VI
FINE-TUNING SCHEME FOR THE EVOLUTRIZ AGENT

Steps	Scheme
1	Hello, I understand that you will be asking me some purposeful questions. Your goal is to open up your creative thinking and develop innovative design solutions with my assistance. I can certainly help you with that by utilizing various TRIZ tools such as the contradiction matrix, 40 inventive principles, supersystem analysis, fishbone diagram analysis, and the concept of ideal final result (IFR) to analyze problems and provide solutions.
2	The research process involves the following steps: Identify the contradictions inherent in the problem. Determine the parameters that could be improved or potentially worsened. Apply inventive principles to resolve the issues thus identified.
3	Research objective: To develop an innovative product or system for the *** scenario that is designed to facilitate *** tasks. The proposed solutions should extend beyond conventional concept designs or existing products that are currently prevalent in the market with the aim of generating novel and distinctive approaches.
4	From a practical perspective, the [ordinal number] proposal appears to be less feasible. Please evaluate the practicality of the remaining proposals.
5	Please provide a detailed explanation of the [ordinal number] proposal or functionality.
6	Please evaluate the aforementioned proposals from the perspectives of ***, ***, and *** and provide a comprehensive summary of these solutions.

IV. EXPERIMENT

A. PARTICIPANT RECRUITMENT

In this study, we adhered to rigorous ethical standards and procedures to ensure the scientific validity and rationality of this research, and the research protocol was approved by the Ethics Review Committee of the Beijing Institute of Technology (BIT-EC-H-2024165). All of the participants voluntarily signed informed consent forms prior to the experiment.

The participants were recruited from the School of Design at a leading university in China. A total of 30 master's degree students with backgrounds in industrial design were enrolled. The participants' ages ranged from 22 to 28 years ($\mu = 24.5$, $SD = 1.53$), and the sample featured an equal gender ratio of 1:1. All of the participants were in good physical health and had normal vision. Additionally, all of the participants had frequent experience using LLMs over the past 6 months to 1 year; however, no participants had received formal training in the TRIZ.

The experiment was conducted in two phases separated by a one-month interval, with each phase comprising two design tasks. During the first phase, the participants referred to as the human-agent group collaborated with the LLM-based IntelliStorm and EvoluTRIZ agents to complete the tasks. To mitigate the impact of order effects and task differences, a counterbalanced design was employed.

During the second phase, the participants collaborated with a senior product designer with 8 years of professional experience and were designated the human–human group. They completed the tasks using two methods, brainstorming and the TRIZ.

B. EXPERIMENTAL EQUIPMENT AND SOFTWARE

The experimental environment was located in a well-equipped laboratory furnished with the necessary furniture, two professional experimenters, and essential equipment. The experimental equipment included but was not limited to audio recording notebooks, video cameras, two laptop computers (one dedicated to experimental guidance and another focused on data collection), and a BIO3 EDA physiological signal acquisition device provided by the Jinfa Technology Company, which featured a sampling frequency of 64 Hz. The device was secured to the participant's wrist via a band, and electrodes were attached to the index and middle fingers of the participant's left hand. Additionally, a smartphone equipped with the most up-to-date version of GPT-4 software was used to facilitate voice interaction.

The EDA data acquisition and processing procedure employed the ErgoLAB software developed by the Jinfa Technology Company, China, and the data analysis were performed using SciPy and SPSS V26.

C. EXPERIMENTAL DESIGN AND PROCEDURE

Experimental tasks: (a) You are traveling to the land of Oz and need to communicate effectively with the locals despite a complete language barrier. (b) You need to prepare delicious meals immediately despite not knowing how to cook. These two tasks represent common pain points in specific scenarios, each containing significant contradictions. Designers must propose innovative product solutions, encouraging diversity and depth in their approaches.

In preparation for the experiment, the participants were first trained and equipped with devices. The experimental assistant then presented a simulated task: "How to solve the drinking water problem during outdoor activities." In the human–agent collaboration (HAC) group, the participants randomly collaborated with one of the LLM-based agents to familiarize themselves with the process. In the human-human collaboration (HHC) group, the participants used a paper version of the TRIZ 39 contradictions matrix parameter table to collaborate with the experienced human designer, who was tasked with assisting participants in identifying the main contradictions of the task, selecting both the key improvement and worsening parameters, and, when necessary, interpreting specific innovative principles and providing encouragement. Formal experiment:

(1) Before each design task, the participants were required to enter a 2-minute resting state to establish a baseline for the physiological data.

(2) During the execution of the design task, the participants communicated with the LLM-based agents using natural language. To promote the visualization of

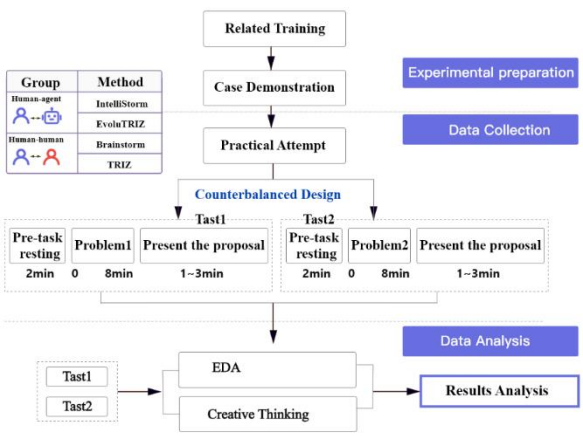

FIGURE 3. Experimental design process.

thoughts, participants could sketch or record text on paper according to their personal preferences. Each design task lasted for 8 minutes, at which point the data collection phase ended.

(3) After completing the task, the participants were required to verbally describe the functions, structure, and usage methods of their final solution in detail[38]. The

FIGURE 4. Experimental scenario.

duration of this description was determined by the participant's actual explanation, which typically lasted between 1 and 3 minutes.

The entire experimental process is illustrated in Figure 3, and the experimental scenario for the HAC group participants is shown in Figure 4.

D. DATA COLLECTION AND PROCESSING

In this study, we employed a comprehensive research approach with the goal of obtaining an in-depth understanding of participants' emotional states and cognitive performance during the process of collaboration with agents through both qualitative and quantitative data collection and analysis. Two types of data were collected: physiological data(EDA vales) and textual data(the solution outputs from the participants).

1) PHYSIOLOGICAL DATA

(1) Noise reduction: With the assistance of the ErgoLAB software, we applied Gaussian smoothing and low-pass filtering techniques to the raw EDA data to effectively reduce noise and interference. Through these preprocessing steps, we obtained dynamic EDA values for each participant over the 8-minute task duration; the time series is shown on the vertical axis.

(2) We analyzed two main components of the EDA data: the tonic component, which is known as the skin conductance level (SCL) and represents the average level over a longer period, reflecting an individual's baseline physiological state during a specific time frame; and the phasic component, which is known as the skin conductance response (SCR) and represents the short-term fluctuations in skin conductance that are typically triggered by external stimuli or emotional changes [38]. With the assistance of the ErgoLAB software, we extracted the mean values of the participants' EDA and SCL as well as the event count and event amplitude of the SCR. These metrics provide a quantitative basis for analyzing individuals' physiological states, including their emotional states and reactions under different conditions.

2) TEXTUAL DATA

(1) Text transcription: During the experiment, participants' communications were fully audio-recorded and transcribed into a text format. The portions of the transcribed text in which the participants elaborated on their design solutions were selected as the basis for evaluating their innovative thinking.

(2) Expert scoring: Innovative thinking refers to the demonstration of highly original and creative thinking abilities during the problem-solving process and the creation of new things. Such thinking primarily includes four dimensions: fluency, flexibility, originality, and elaboration[18]. These dimensions collectively generate a comprehensive evaluation framework for innovative thinking, thus ensuring a thorough and in-depth assessment. To evaluate these solutions objectively, we invited two experts to engage in a detailed scoring process based on standardized criteria.

Fluency assesses the number of solutions generated by the respondent within a specified time, thereby measuring the fluency of their thinking.

Originality evaluates the uniqueness and novelty of solutions, thereby measuring whether unique perspectives or ideas are proposed during the problem-solving process.

Flexibility assesses the diversity and adaptability of the respondent's thinking, thereby measuring their ability to propose solutions drawn from different categories.

Elaboration evaluates depth, thereby measuring the hierarchical structure of the detailed descriptions of the solutions.

V. RESULTS

A. DESIGN PROCESSES

This study initially preprocessed the EDA data to eliminate individual differences. After confirming that the data met the assumption of a normal distribution through the Shapiro–Wilk test, a repeated measures two-way ANOVA was conducted using SciPy. This analysis effectively examined the main interaction effects between the intergroup collaboration patterns and intragroup methods under the time factor.

The analysis results,as shown in Table 7,indicate that all p values are less than 0.001. Both the collaboration groups (HAC and HHC) and the concept generation methods had significant main effects on the subjects' EDA values ($p < 0.001$) and produced significant interaction effects. Moreover,

TABLE VII
TWO-WAY ANOVA RESULTS

	T	P	η^2	95% CI
Method	34.276	<0.001***	0.008	[0.01,1.00]
Group	56.765	<0.001***	0.07	[0.07,1.00]
Method* Group	-34.135	<0.001***	0.02	[0.02,1.00]

the effect size between groups was larger, with a negative interaction with the methods. This result implies that both the concept generation methods and collaboration groups had significant impacts on the physiological activation of the participants.

1) TEMPORAL VARIATION COMPARISON

To further compare the dynamic changes in the EDA during the task, we averaged the EDA values at each time point for each 8-minute task segment, resulting in EDA values for the 4 time series groups. The entire task was divided equally into 20 periods, corresponding to 21 sampling points with a time interval of 24 seconds between each sampling point. Line graph 5 was plotted to visually analyze the trend of the EDA changes over time.

As shown in Figure 5, the participants exhibited lower EDA response levels and a relatively stable trend during collaboration with the LLM-based agents. In contrast, when collaborating with human designers, the subjects demonstrated consistently higher EDA response levels with greater fluctuations. Specifically, there was a sudden increase in EDA toward the latter part of the brainstorming task, whereas the TRIZ task showed the most significant fluctuations.

2) COMPREHENSIVE COMPARISON

To conduct an in-depth analysis of the subjects' baseline physiological states during specific periods and their

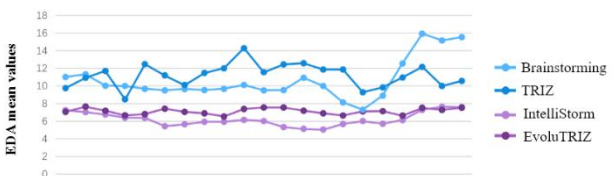

FIGURE 5. Comparison of mean EDA values over time.

emotional responses to the information stimuli, we analyzed the mean values of the EDA, SCL, event count, and event amplitude. First, we conducted tests to confirm that the data within each group met the assumption of a normal distribution. Then, we performed a two-factor analysis of variance on these indicators. The results, as shown in Table 8, indicate that the collaboration groups had significant effects on all of the feature data ($p > 0.05$), whereas the methods had a significant effect only on the SCL mean, with $p = 0.035$.As shown in Figure 6.

(1) The EDA mean values are as follows: IntelliStorm (5.709) < EvoluTRIZ (6.266) < Brainstorming (9.588) < TRIZ (10.92). This finding demonstrates that the overall physiological activation level during human–agent collaboration tasks is significantly lower than that during human–human collaboration, with unstructured methods such as brainstorming presenting lower activation levels.

(2) The SCL mean values are as follows: IntelliStorm (5.26) < EvoluTRIZ (5.607) < Brainstorming (9.238) < TRIZ (10.34), indicating that the baseline physiological state during human–agent collaboration tasks is notably lower than that of human–human collaboration, with unstructured methods such as brainstorming producing relatively low baseline physiological states.

(3) The event count mean values are as follows: IntelliStorm (39.25) < EvoluTRIZ (44.83) < Brainstorming (46.92) < TRIZ (48.42). This result reflects that human–human collaboration elicits more frequent responses to information stimuli, with structured methods such as the TRIZ prompting more frequent stimulus responses.

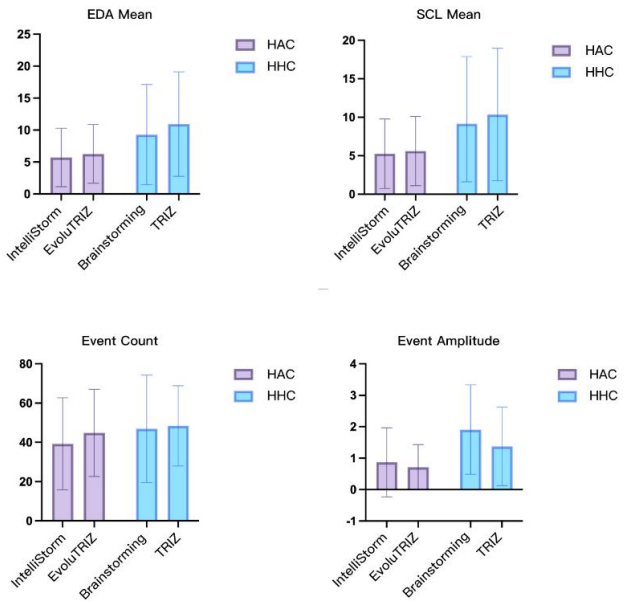

FIGURE 6. Comparison of feature data.

(4) The event amplitude mean values are as follows: EvoluTRIZ (0.7113) < IntelliStorm (0.8675) < TRIZ (1.373) < Brainstorming (1.908), demonstrating that human–human collaboration results in more intense responses to information stimuli, with unstructured methods such as brainstorming resulting in more significant attention shifts and intense emotional responses.

These comprehensive results indicate that the

TABLE VIII
ANALYSIS OF MEAN CHARACTERISTIC DATA

	HAC		HHC		Group		Method	
	IntelliStorm	EvoluTRIZ	Brainstorming	TRIZ	p	F	p	F
EDA Mean	5.709	6.266	9.588	10.92	0.002**	10.025	0.683	0.167
SCL Mean	5.26	5.607	9.238	10.34	<0.001***	20.974	0.035*	4.596
Event Count	39.25	44.83	46.92	48.42	0.047*	3.846	0.389	0.75
Event Amplitude	0.8675	0.7113	1.908	1.373	<0.001***	11.452	0.207	1.617

collaboration group significantly influenced designers' physiological activation and responsiveness to information during the conceptual design process; this indication suggests that when collaborating with agents, participants exhibit more stable emotional states and lower cognitive loads, whereas collaboration with human designers results in greater cognitive engagement and emotional activation, with more frequent and intense responses to information stimuli.

B. PERFORMANCE RESULTS

We conducted a reliability assessment of the expert scoring results. The scores given by two experts for each participant across the four dimensions, fluency, originality, flexibility, and elaboration, were analyzed for reliability via the intraclass correlation coefficient (ICC). The statistical results of this analysis are shown in Table 9. The fluency (ICC (1,1) = 0.982) and originality (ICC (1,1) = 0.849) demonstrate high consistency. The flexibility (ICC (1,1) = 0.770) and elaboration (ICC (1,1) = 0.769)) also exhibit overall consistency. In conclusion, the scoring system demonstrated reliability and can be used for subsequent analysis.

Given that the scores did not follow a normal distribution, we used the Mann–Whitney U test to examine the main effects of both group and method separately. The results are shown in Table 10. The collaboration groups demonstrated significant differences in terms of fluency, originality, and flexibility ($p < 0.05$, Cohen's $d > 0.4$), indicating medium effect sizes. However, no significant difference was observed between the collaboration groups in terms of elaboration ($p > 0.05$).

(3) Flexibility: The experimental results are as follows: EvoluTRIZ (2.24) > IntelliStorm (2.20) > Brainstorming (2.12) > TRIZ (1.08). These findings suggest that collaboration with LLM-based agents can improve cognitive flexibility.

(4) Elaboration: The experimental results are as follows: EvoluTRIZ (2.29) > TRIZ (2.00) > IntelliStorm (1.82) > Brainstorming (1.80). These results reveal that structured methods can promote deeper thinking and problem solving among participants, with LLM-based agent collaboration demonstrating particular advantages.

The normalized results are presented in a radar chart, as shown in Figure 7.

In conclusion, combining LLM-based agents with traditional design methods can yield significant advantages. The IntelliStorm agent performs better in terms of fluency, the EvoluTRIZ agent excels in originality.

The concept generation methods had significant effects on all of the performance measures ($p < 0.05$, Cohen's $d > 0.4$), also indicating medium effect sizes.

(1) Fluency: The experimental results are as follows: IntelliStorm (4.29) > EvoluTRIZ (3.41) > Brainstorming (2.62) > TRIZ (1.23). These findings suggest that collaboration with LLM-based agents can enhance cognitive fluency.

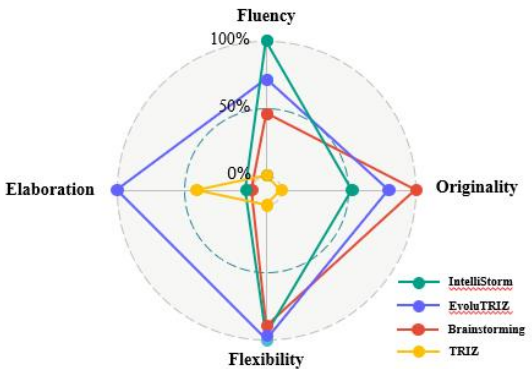

FIGURE 7. Creative thinking performance.

(2) Originality: The experimental results are as follows: Brainstorming (3.71) > EvoluTRIZ (3.40) > IntelliStorm (3.00) > TRIZ (2.03). These findings indicate that human–human collaboration using the brainstorming method yields the highest originality, whereas the combination of LLM-based agents and the TRIZ method can enhance

TABLE IX
ICC RESULTS

	ICC	95% CI		F test assuming a true value of 0			
		Lower	Upper	Values	df1	df2	p
Fluency ICC（1，1）	0.982	0.972	0.99	117.37	56	57	0.000***
Originality ICC（1，1）	0.849	0.773	0.904	6.609	56	171	0.000***
Flexibility ICC（1，1）	0.77	0.655	0.854	4.349	56	171	0.000***
Elaboration ICC（1，1）	0.769	0.66	0.85	4.304	56	285	0.000***

TABLE X
CREATIVE THINKING ASSESSMENT RESULTS

	HAC	HHC			Group		Method	
	IntelliStorm	EvoluTRIZ	Brainstorming	TRIZ	p	Cohen's d	p	Cohen's d
Fluency	4.29	3.41	2.62	1.23	0.035***	0.603	0.003**	0.727
Originality	3.00	3.40	3.71	2.03	0.028*	0.617	0.05*	0.465
Flexibility	2.20	2.24	2.12	1.08	0.023**	0.610	0.022*	0.624
Elaboration	1.82	2.29	1.80	2.00	0.929	0.023	0.020*	0.627

flexibility and elaboration, and the human–human group that used brainstorming exhibits an irreplaceable advantage in terms of originality.

VI. DISCUSSION

A. DEVELOPMENT AND TRAINING OF LLM-BASED AGENTS (RQ1)

This study selected only two conceptual generation methods with significantly different structures, providing a detailed description of the development steps for integration with LLMs. However, future practices should not be limited to the examples provided in this study. In selecting conceptual generation methods, various approaches including morphological analysis[39], SCAMPER[40], C-K theory[41], and Six Thinking Hats[42], each possess distinct characteristics and advantages. All of these properties can be combined with LLMs through chain-of-thought techniques to enhance the collaborative efficiency of LLM-based agents.

Furthermore, the application of design case libraries offers new directions for agent development. By constructing static case libraries combined with enhanced retrieval functions, agents can effectively access and match existing design principles, ensuring design consistency. This approach is particularly crucial in scenarios such as enterprise brand development and cultural heritage preservation, guaranteeing effective knowledge transfer and the continuity of distinctive features. Additionally, introducing dynamic case libraries in conjunction with industry trends and user feedback can provide personalized recommendation services for designers, further enhancing the adaptability and foresight of LLM-based agents.

B. DESIGN PROCESS (RQ2)

We found that collaboration modes significantly affect EDA values. When collaborating with LLM-based agents, participants presented lower EDA response levels with less fluctuation, indicating a lower cognitive load and more stable emotions. A high cognitive load can occupy substantial working memory resources, limiting an individual's ability to generate novel ideas [43] [15]. This phenomenon may occur because when collaborating with LLM-based agents, participants experience a greater sense of control and autonomy over the task [44], receiving neutral and structured content feedback from the agent. In contrast, when collaborating with humans, participants are influenced not only by verbal communication but also by emotional feedback, such as encouragement or opposition[45]. Facing different viewpoints and solutions may lead to cognitive conflicts [46], and participants may unconsciously engage in social comparison, focusing on whether their performance is superior or inferior to that of others. These situations require more emotional and cognitive resources to process, resulting in emotional fluctuations and an increased cognitive load.

Moreover, concept generation methods also influence EDA values to some extent. When unstructured methods such as brainstorming were used, participants presented lower physiological activation levels but more significant attention shifts and intense emotional responses. With structured methods such as the TRIZ, participants exhibited high physiological activation states and more frequent responses to information, which may occur because during brainstorming tasks, participants are encouraged to propose as many ideas as possible without immediate evaluation or filtering. This open, unconstrained environment reduces task urgency and pressure, resulting in lower emotional activation levels. However, participants need to constantly switch between different ideas and thoughts in their minds, requiring high levels of attention and information response to explore and express as many creative ideas as possible in a short time. The TRIZ emphasizes generating ideas by comparing different engineering parameters and solution principles. During this process, participants may need to choose between multiple complex options, which not only increases the frequency of information processing but may also lead to increased stress, resulting in elevated physiological activation levels and frequent information responses.

These findings address Research Question 2, demonstrating that collaboration with LLM-based agents can reduce participants' cognitive pressure not only by providing more stable and predictable feedback, which helps them handle task challenges more smoothly, but also, when combined with different concept generation methods, reflecting the unique impacts of these methods on emotional activation and information processing.

C. DESIGN PERFORMANCE (RQ3)

We analyzed the effects of the collaboration modes and concept generation methods on the basis of the four scoring aspects: fluency, flexibility, originality, and elaboration.

Fluency and Flexibility: The research findings indicate that human–agent collaboration outperforms human–human collaboration in terms of these two aspects. Compared with EvoluTRIZ, IntelliStorm produced significantly more solutions, possibly related to the "quantity encouragement" principle of the method. Collaboration with EvoluTRIZ resulted in slightly stronger performance in terms of flexibility, potentially because the structured method promoted broader and more diverse consideration of issues. These findings reveal that the involvement of LLM-based agents can enhance cognitive efficiency while maintaining the advantages and characteristics of different concept generation methods.

Originality: The results show that human–human collaboration via brainstorming performed best, possibly because of the unique emotional resonance, experiential background, and intuitive thinking inherent in human interaction during the process. An alternative explanation is

that LLM-based agents rely on established algorithms and preset knowledge bases for reasoning and decision-making, which are primarily based on induction and deduction from known information. Consequently, humans may outperform LLM-based agents in terms of sparking innovative inspiration and insights. However, human–agent collaboration modes have demonstrated unique value. For example, while the TRIZ performed poorly when employed during human–human collaboration, when combined with LLM-based agents, its performance surpassed that of IntelliStorm, strongly indicating the potential of LLM-based agents in enhancing the originality of structured methods.

Elaboration: Human–agent collaboration has clear advantages, with EvoluTRIZ performing best. This result indicates a correlation with structured innovation method principles on the one hand and a more effective combination with LLM-based agents on the other hand. The participation of LLM-based agents led to more in-depth problem exploration, generating more detailed and thorough solutions.

In conclusion, the involvement of LLM-based agents can enhance cognitive performance. Additionally, we found that during human–human collaboration, TRIZ tasks generally performed poorly, possibly due to participants' unfamiliarity with the TRIZ method. However, EvoluTRIZ performed best in terms of elaboration and flexibility and second-best in terms of fluency and originality. This finding reveals that the involvement of LLM-based agents can rapidly lower the threshold for participants to use new methods, saving learning costs.

D. FUTURE RESEARCH DIRECTIONS

Based on the value and limitations identified in this study, we propose that future research could be developed in the following ways:

(1) Multimodal LLM-based agent design: Various input and output forms, such as visual, audio, and text, could be incorporated into a single LLM-based agent to meet the needs of different design scenarios.

(2) Optimization of human–agent collaboration modes: Different types of human–agent collaboration, such as synchronous versus asynchronous collaboration and individual versus team collaboration could be investigated, and the impact of different collaboration modes on designers' creativity and decision-making processes could be further explored. Particular emphasis should be placed on how LLM-based agents can optimize design team collaboration efficiency and innovative output.

(3) Personalization and adaptability of LLM-based agents: Future research could further explore the personalization and adaptability of LLM-based agents. This process could involve developing more flexible conceptual design agents by combining multiple concept generation methods and developing agents based on designers' needs and preferences.

(4) Integration of neuroergonomic research methods: Technologies such as electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) could be utilized to observe brain activity in depth and investigate how LLM-based agents influence participants' cognitive activation processes in the brain.

E. LIMITATIONS

Despite providing valuable insights, this study has several limitations. First, the research sample size was relatively small and confined to the Chinese university region. Future studies could expand to a wider array of demographic groups and additional countries and regions. Second, the study did not compare the effects of different LLMs on the design process or the outputs. Additionally, the task duration was limited to only 8 minutes, which makes it challenging to generate in-depth insights in such a short time frame. This time constraint may have resulted in insufficient explanatory power in terms of elaborating on thinking processes.

VII. CONCLUSION

The hypothesis proposed in this study to enhance the conceptual design capabilities of LLMs has been confirmed: traditional concept generation methods can be effectively combined with LLMs through CoT techniques. This combination not only reflects the unique impacts of these methods on physiological activation and information response in cognitive processes but also leverages their respective advantages and characteristics in terms of the performance of innovative thinking.

To address RQ1, we meticulously detailed the development process of LLM-based agents, culminating in the creation of two agents: IntelliStorm and EvoluTRIZ. To investigate participants' cognitive processes (RQ2) and evaluate the output performance (RQ3), we conducted experiments comparing the effects of collaboration modes (HAC and HHC) and concept generation methods (brainstorming and the TRIZ) on participants' EDA activity and innovative thinking.

Our findings indicate that LLM-based agents significantly reduce participants' baseline physiological activation levels, alleviate their cognitive load, and enhance their thinking efficiency. Different conceptual design methods exhibited unique physiological response patterns. For example, IntelliStorm and brainstorming demonstrated lower baseline physiological activation levels but more intense responses to information stimuli. Conversely, EvoluTRIZ and the TRIZ exhibited higher baseline physiological activation levels but with a greater frequency of information responses.

Moreover, the HAC preserved the performance advantages of various concept generation methods: IntelliStorm and brainstorming excelled in terms of divergent thinking, particularly in cognitive fluency; EvoluTRIZ and TRIZ performed the best in terms of logical reasoning, facilitating thought elaboration. While the HHC using brainstorming demonstrated superior performance in terms of originality, human–agent collaboration significantly enhanced cognitive

flexibility. These findings confirm that using CoT techniques to fine-tune LLMs for integration with different concept generation methods is an effective strategy that can better optimize the reasoning of large models and enhance participants' cognitive processes and performance outcomes. Future research should incorporate more conceptual design methods to expand the application of LLM-based agents in design.

ACKNOWLEDGMENT

This work is supported by the Guangdong Province Educational Science Planning Project: Research on the Construction of an Innovative Model for Art Design Course Clusters Driven by Intangible Cultural Heritage Digitalization from the Perspective of Cultural Confidence Project Number: 2024GXJE108.

REFERENCES

[1] N. Yüksel, H. R. Börklü, H. K. Sezer, and O. E. Canyurt, "Review of artificial intelligence applications in engineering design perspective," Eng. Appl. Artif. Intell., vol. 118, p. 105697, Feb. 2023, doi: 10.1016/j.engappai.2022.105697.

[2] L. Hay, A. H. B. Duffy, C. McTeague, L. M. Pidgeon, T. Vuletic, and M. Grealy, "Towards a shared ontology: a generic classification of cognitive processes in conceptual design," Des. Sci., vol. 3, p. e7, 2017, doi: 10.1017/dsj.2017.6.

[3] M. Ghonim, Design thinking in architecture education: issues, limitations and suggestions. 2016.

[4] N. Crilly and C. Cardoso, "Where next for research on fixation, inspiration and creativity in design?," Des. Stud., vol. 50, pp. 1–38, May 2017, doi: 10.1016/j.destud.2017.02.001.

[5] K. Fu, J. Murphy, M. Yang, K. Otto, D. Jensen, and K. Wood, "Design-by-analogy: experimental evaluation of a functional analogy search methodology for concept generation improvement," Res. Eng. Des., vol. 26, no. 1, pp. 77–95, Jan. 2015, doi: 10.1007/s00163-014-0186-4.

[6] F. Stella, C. Della Santina, and J. Hughes, "How can LLMs transform the robotic design process?," Nat. Mach. Intell., vol. 5, no. 6, pp. 561–564, Jun. 2023, doi: 10.1038/s42256-023-00669-7.

[7] W. Holmes and I. Tuomi, "State of the art and practice in AI in education," Eur. J. Educ., vol. 57, no. 4, pp. 542–570, Dec. 2022, doi: 10.1111/ejed.12533.

[8] H. Jo and D.-H. Park, "AI in the workplace: examining the effects of ChatGPT on information support and knowledge acquisition," Int. J. Human–Computer Interact., pp. 1–16, Nov. 2023, doi: 10.1080/10447318.2023.2278283.

[9] Z. Zhou, J. Li, Z. Zhang, J. Yu, and H. Duh, "Examining how the large language models impact the conceptual design with human designers: a comparative case study," Int. J. Human–Computer Interact., pp. 1–17, Jul. 2024, doi: 10.1080/10447318.2024.2370635.

[10] X. Zhai et al., "A review of artificial intelligence (AI) in education from 2010 to 2020," Complexity, vol. 2021, no. 1, p. 8812542, Jan. 2021, doi: 10.1155/2021/8812542.

[11] X. Zhai, "ChatGPT user experience: implications for education," SSRN Electron. J., 2022, doi: 10.2139/ssrn.4312418.

[12] K. Helm, K. Jablokow, S. Daly, E. Silk, S. Yilmaz, and R. Suero, "Evaluating the impacts of different interventions on quality in concept generation," in 2016 ASEE Annual Conference & Exposition Proceedings, New Orleans, Louisiana: ASEE Conferences, Jun. 2016, p. 26766. doi: 10.18260/p.26766.

[13] A. F. Osborn, Applied Imagination; Principles and Procedures of Creative Thinking. Scribner, 1953.

[14] G. S. Altshuller, "Creativity as an Exact Science," CRC Press, 1984.

[15] I. P. Hernandez Sibo, D. A. Gomez Celis, and S. Liou, "Exploring the landscape of cognitive load in creative thinking: a systematic literature review," Educ. Psychol. Rev., vol. 36, no. 1, p. 24, Mar. 2024, doi: 10.1007/s10648-024-09866-1.

[16] N. Nourbakhsh, Y. Wang, and F. Chen, "GSR and blink features for cognitive load classification," in Human-Computer Interaction – INTERACT 2013, vol. 8117, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds., in Lecture Notes in Computer Science, vol. 8117. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 159–166. doi: 10.1007/978-3-642-40483-2_11.

[17] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: a systematic review," Sensors, vol. 20, no. 2, p. 479, Jan. 2020, doi: 10.3390/s20020479.

[18] E. P. Torrance, The Torrance Tests of Creative Thinking-TTCT Manual and Scoring Guide: Verbal Test A, Figural Test. Ginn, 1974.

[19] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "Transitioning to human interaction with AI systems: new challenges and opportunities for HCI professionals to enable human-centered AI," arXiv.org. Accessed: Sep. 05, 2024. [Online]. Available: https://arxiv.org/abs/2105.05424v4

[20] Z. Xi et al., "The rise and potential of large language model based agents: a survey," Sep. 19, 2023, arXiv: arXiv:2309.07864. Accessed: Sep. 26, 2024. [Online]. Available: http://arxiv.org/abs/2309.07864

[21] Y. Liu et al., "Summary of ChatGPT-related research and perspective towards the future of large language models," Meta-Radiol., vol. 1, no. 2, p. 100017, Sep. 2023, doi: 10.1016/j.metrad.2023.100017.

[22] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-efficient fine-tuning for large models: a comprehensive survey," Jul. 12, 2024, arXiv: arXiv:2403.14608. Accessed: Sep. 05, 2024. [Online]. Available: http://arxiv.org/abs/2403.14608

[23] X. Feng et al., "Large language model-based human-agent collaboration for complex task solving," Feb. 20, 2024, arXiv: arXiv:2402.12914. Accessed: Sep. 26, 2024. [Online]. Available: http://arxiv.org/abs/2402.12914

[24] A. Balaguer et al., "RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture," Jan. 30, 2024, arXiv: arXiv:2401.08406. Accessed: Sep. 05, 2024. [Online]. Available: http://arxiv.org/abs/2401.08406

[25] B. Wang et al., "Towards understanding chain-of-thought prompting: an empirical study of what matters," presented at the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2023, pp. 2717–2739. doi: 10.18653/v1/2023.acl-long.153.

[26] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," Adv. Neural Inf. Process. Syst., vol. 35, pp. 24824–24837, Dec. 2022.

[27] J. Li, H. Cao, L. Lin, Y. Hou, R. Zhu, and A. El Ali, "User experience design professionals' perceptions of generative artificial intelligence," in Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu HI USA: ACM, May 2024, pp. 1–18. doi: 10.1145/3613904.3642114.

[28] N. Ho, L. Schmid, and S.-Y. Yun, "Large language models are reasoning teachers," Jun. 13, 2023, arXiv: arXiv:2212.10071. Accessed: Sep. 26, 2024. [Online]. Available: http://arxiv.org/abs/2212.10071

[29] T. Shealy, J. Gero, M. Hu, and J. Milovanovic, "Concept generation techniques change patterns of brain activation during engineering design," Des. Sci., vol. 6, p. e31, 2020, doi: 10.1017/dsj.2020.30.

[30] M. Ghane, M. C. Ang, D. Cavallucci, R. Abdul Kadir, K. W. Ng, and S. Sorooshian, "Semantic TRIZ feasibility in technology development, innovation, and production: a systematic review," Heliyon, vol. 10, no. 1, p. e23775, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23775.

[31] Z. Wu, D. Ji, K. Yu, X. Zeng, D. Wu, and M. Shidujaman, "AI Creativity and the Human-AI Co-creation Model," presented at the HUMAN-COMPUTER INTERACTION: THEORY, METHODS AND TOOLS, HCII 2021, PT I, M. Kurosu, Ed., 2021, pp. 171–190. doi: 10.1007/978-3-030-78462-1_13.

[32] W. Boucsein, Electrodermal activity. Boston, MA: Springer US, 2012. doi: 10.1007/978-1-4614-1126-0.

[33] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," J. Neurosci. Methods, vol. 190, no. 1, pp. 80–91, Jun. 2010, doi: 10.1016/j.jneumeth.2010.04.028.
[34] H. D. Critchley, "Neural mechanisms of autonomic, affective, and cognitive integration," J. Comp. Neurol., vol. 493, no. 1, pp. 154–166, Dec. 2005, doi: 10.1002/cne.20749.
[35] teo sheila, "How I won singapore's GPT-4 prompt engineering competition." Accessed: Oct. 06, 2024. [Online]. Available: https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41
[36] S. Qiao et al., "Reasoning with language model prompting: a survey," Sep. 18, 2023, arXiv: arXiv:2212.09597. Accessed: Sep. 26, 2024. [Online]. Available: http://arxiv.org/abs/2212.09597
[37] N. L. Cheng, Introduction to Industrial Design. China Machine Press, 2019.
[38] J. S. Gero and U. Kannengiesser, "The situated function–behaviour–structure framework," Des. Stud., vol. 25, no. 4, pp. 373–391, Jul. 2004, doi: 10.1016/j.destud.2003.10.010.
[39] F. Zwicky, "The morphological approach to discovery, invention, research and construction," in New Methods of Thought and Procedure, F. Zwicky and A. G. Wilson, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1967, pp. 273–297. doi: 10.1007/978-3-642-87617-2_14.
[40] B. Eberle, Scamper: creative games and activities for imagination development (combined ed., grades 2-8), 1st ed. New York: Routledge, 2023. doi: 10.4324/9781003423560.
[41] A. Hatchuel and B. Weil, "A new approach of innovative design:," Int. Conf. Eng. Des. ICED 03, 2003.
[42] P. S. Aithal and S. K. P. M., "Integrating theory a and six thinking hats technique for improved organizational performance," Int. J. Appl. Eng. Manag. Lett., pp. 66–77, Nov. 2017, doi: 10.47992/IJAEML.2581.7000.0013.
[43] M. Bose, J. Anne Garretson Folse, and S. Burton, "The role of contextual factors in eliciting creativity: primes, cognitive load and expectation of performance feedback," J. Consum. Mark., vol. 30, no. 5, pp. 400–414, Jul. 2013, doi: 10.1108/JCM-02-2013-0475.
[44] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.," Am. Psychol., vol. 55, no. 1, pp. 68–78, 2000, doi: 10.1037/0003-066X.55.1.68.
[45] A. Fink, R. H. Grabner, D. Gebauer, G. Reishofer, K. Koschutnig, and F. Ebner, "Enhancing creativity by means of cognitive stimulation: evidence from an fMRI study," Neuroimage, vol. 52, no. 4, pp. 1687–1695, Oct. 2010, doi: 10.1016/j.neuroimage.2010.05.072.
[46] J. Decety and J. A. Sommerville, "Shared representations between self and other: a social cognitive neuroscience view," Trends Cogn. Sci., vol. 7, no. 12, pp. 527–533, Dec. 2003, doi: 10.1016/j.tics.2003.10.004.

AUTHOR BIOGRAPHIES

Ge Shijun, a former designer, is now a Ph.D. candidate at the Beijing Institute of Technology. Her research focuses on human-AI collaboration and design thinking education, exploring how large language models can enhance design efficiency. Ge's industry experience uniquely informs her academic work in bridging AI technology with practical design applications.

Sun Yuanbo is a professor and doctoral supervisor at the School of Design and Art, Beijing Institute of Technology. He is the head of the Human Factors Engineering and Interaction Design Laboratory and the Vehicle Styling Design Studio. He also serves as the Deputy Director of the Information and Interaction Design Professional Committee of the China Industrial Design Association. His research areas include human–computer interaction, intelligent connected vehicles, and design thinking education.

Cui Yin obtained her Ph.D. in Product Design in 2021. From 2021 to 2023, she conducted interdisciplinary research in artificial intelligence and design at the postdoctoral research station of the Beijing Institute of Technology and was recognized as an outstanding postdoctoral researcher in 2023. Since then, she has been working as a lecturer in the Department of Industrial Design, School of Creative Design, Shenzhen Technology University. Her research interests focus on the intersection of generative artificial intelligence and design disciplines.

Wei Dapeng is a Ph.D. candidate at the Beijing Institute of Technology, with a primary focus on human–computer interaction design in complex information systems, user–centered interaction design, and human factors/ergonomics research. His main contributions are centered on improving the user experience in various scenarios through interaction design.