ASME Journal of Mechanical Design
Online journal at:
https://asmedigitalcollection.asme.org/mechanicaldesign

Vision-Language Models for Design Concept Generation: An Actor–Critic Framework

Parisa Ghasemi[1]
George W. Woodruff School of Mechanical Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332
e-mail: parisa.ghasemi@gatech.edu

Mohsen Moghaddam
George W. Woodruff School of Mechanical Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332;
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332
e-mail: mohsen.moghaddam@gatech.edu

We introduce a novel actor-critic framework that utilizes vision-language models (VLMs) and large language models (LLMs) for design concept generation, particularly for producing a diverse array of innovative solutions to a given design problem. By leveraging the extensive data repositories and pattern recognition capabilities of these models, our framework achieves this goal through enabling iterative interactions between two VLM agents: an actor (i.e., concept generator) and a critic. The actor, a custom VLM (e.g., GPT-4) created using few-shot learning and fine-tuning techniques, generates initial design concepts that are improved iteratively based on guided feedback from the critic—a prompt-engineered LLM or a set of design-specific quantitative metrics. This process aims to optimize the generated concepts with respect to four metrics: novelty, feasibility, problem–solution relevancy, and variety. The framework incorporates both long-term and short-term memory models to examine how incorporating the history of interactions impacts decision-making and concept generation outcomes. We explored the efficacy of incorporating images alongside text in conveying design ideas within our actor–critic framework by experimenting with two mediums for the agents: vision language and language only. We extensively evaluated the framework through a case study using the AskNature dataset, comparing its performance against benchmarks such as GPT-4 and real-world biomimetic designs across various industrial examples. Our findings underscore the framework's capability to iteratively refine and enhance the initial design concepts, achieving significant improvements across all metrics. We conclude by discussing the implications of the proposed framework for various design domains, along with its limitations and several directions for future research in this domain. [DOI: 10.1115/1.4067619]

Keywords: design concept generation, creativity, novelty, feasibility, large language model, vision-language model, agent-based design, artificial intelligence, bioinspired design, conceptual design, creativity and concept generation, data-driven design, machine learning

1 Introduction

Concept generation is crucial in the design process as it fosters creativity and the exploration of novel yet feasible ideas, driving innovation and effective problem solving. It often involves the convergence of ideas from multiple perspectives, thus enabling the development of holistic and technically sound solutions that meet diverse user needs and market demands [1]. Recent advancements in design concept generation have significantly integrated computational methodologies with structural engineering, enhancing material distribution optimization, and automating the iterative design process to produce structurally efficient, aesthetically pleasing solutions [2]. Studies have also emphasized the alignment of human–AI interaction in design, demonstrating AI's capability to augment human cognitive processes through multimodal data representation and design strategy optimization [3]. Advanced generative models have facilitated automated concept generation, enabling rapid exploration of design spaces and producing innovative solutions that satisfy both aesthetic and performance requirements [4]. Complex engineering problems have been addressed using conditional generative models and data-driven approaches to optimize aerodynamic and structural designs [5]. Furthermore, the creative capabilities of generative models have been augmented by integrating semantic networks, visual concepts, and natural language generation techniques, fostering innovative design ideation and bridging multiple knowledge domains [6]. These accomplishments collectively highlight the transformative impact of AI and data-driven approaches in contemporary design concept generation.

Recent research has increasingly explored the integration of large language models (LLMs) into agent-based frameworks, simulating complex environments where LLM-powered agents interact with each other and/or their surroundings, perform collaborative

[1]Corresponding author.
Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received June 3, 2024; final manuscript received January 4, 2025; published online April 2, 2025. Assoc. Editor: Christopher Mccomb.

problem solving, and execute tasks autonomously. For instance, external feedback tools and fact-checking systems have been employed to improve the factual accuracy of LLM outputs [7]. While single-agent LLMs demonstrate remarkable capabilities, multiagent systems enhance problem solving by incorporating specialized agents and fostering collaboration, competition, or debate among them [8]. These frameworks distribute tasks among specialized agents to address problems too complex for a single model. For example, LLaMAC [9] incorporates value distribution encoding inspired by the human brain and leverages both internal and external feedbacks to enable collaboration and iterative reasoning. Furthermore, innovations in self-improvement have been proposed, such as Reflexion [10], which utilizes episodic memory for self-reflection as semantic guidance, significantly enhancing decision-making, reasoning, and programming tasks without the need for costly fine-tuning. Applications in this context include, but are not limited to, natural language understanding and generation (e.g., developing coherent conversational agents), knowledge-grounded factuality, strategic tasks like gameplay, and social interaction simulation and analysis. For instance, LEGO [11] utilizes multiple agents with iterative feedback to tackle causality explanation generation, addressing challenges like spurious causal associations. Similarly, debate-style frameworks have demonstrated improved strategic and factual reasoning by enabling LLMs to propose, debate, and refine responses iteratively [12]. Another approach, STACKFEED [13], adopts a multiactor, centralized critic framework to iteratively refine knowledge bases with expert feedback, enhancing retrieval-augmented generation systems for greater accuracy and consistency. Additionally, Prospector [14] combines AskAct prompting for self-questioning with Trajectory Ranking to identify optimal actions from diverse trials, achieving high performance in decision-making tasks like household and shopping activities without parameter updates. Our work leverages similar multiagent LLM interactions for design concept generation but focuses on adapting these methodologies to design-specific challenges.

This debate-style iterative refinement in a multiagent framework, when applied to design, closely parallels the propose–critique–modify (PCM) cycle [15], wherein design solutions are sequentially proposed, critiqued, and refined. The PCM approach systematically breaks down the design process into manageable tasks, reinforcing a structured problem solving methodology based on continuous iteration. While our work builds on a similar structured approach to design concept generation, there are some distinctions. First, in Chandrasekaran's PCM model, human designers manage the proposal and critique roles, whereas our framework leverages AI agents for both actor and critic roles, automating the ideation process to a greater degree. Second, while Chandrasekaran emphasizes explicit knowledge representation—using encoded knowledge of problems, constraints, and objectives—our methodology draws on pretrained models and external resources, relying less on formal knowledge encoding and more on leveraging existing information repositories. Finally, memory is not a focal element in Chandrasekaran's PCM process, which primarily critiques the current design iteration. In contrast, our framework incorporates a memory mechanism within the state update process, allowing prior designs, modifications, and feedback to inform subsequent iterations, thus adding a temporal dimension to the design process.

In design concept generation, a crucial aspect is the "medium"—the channel or mode of communication, such as text or image—through which an idea is conveyed to the designer. The affordances of different mediums significantly influence the clarity, comprehension, and collaborative potential of the design concept generation process. Recent studies employ various mediums such as images, 3D meshes, point lattices, and parameters. However, our research indicates that the utilization of text as a primary [6] or supplementary design medium is notably scarce in the literature. Text+image concept generation combines the clarity of text with the visual impact of images, enriching idea expression and making it accessible to a broad audience, including nontechnical stakeholders. While technical formats are ideal for detailed precision in later design stages, text and images effectively convey conceptual breadth, aesthetic intentions, and functional narratives. This dual modality supports rapid, iterative exploration, allowing designers to quickly generate, modify, and communicate concepts with minimal resources. Embedding design concepts within a narrative framework captures stories, user experiences, and broader contexts, where structural or mathematical formats may fall short. Text and images enhance early stakeholder engagement by presenting ideas evocatively and informatively, inviting feedback and collaboration from the outset. Therefore, we propose the fusion of text and image generative models as a transformative tool for concept generation, particularly for communicating initial ideas associated with the functional, visual, and behavioral aspects of design, offering unprecedented generative capabilities and access to vast data repositories.

Design evaluation metrics are essential for assessing the effectiveness, innovation, and practicality of solutions in data-driven design processes. Each design challenge presents unique requirements, constraints, and objectives, necessitating a tailored selection of evaluation metrics and customized evaluation models to accurately quantify and calculate these metrics across various data types. Various design evaluation metrics have been studied and modeled in the literature across multiple dimensions. Visual and aesthetic evaluation metrics, such as Fréchet inception distance, assess the visual quality and realism of designs [16]. User satisfaction metrics gauge end-user contentment through surveys and user testing [17]. Creativity and novelty metrics utilize methods like feature-based local outlier factor and Hausdorff distance to measure uniqueness [18]. Feasibility metrics ensure practical viability, often assessed through human evaluation [19]. Diversity metrics, such as determinant of similarity matrix and covering radius bound, measure the variety of solutions generated [20]. Structural and functional performance metrics, including drag coefficient and strength-to-weight ratio, evaluate design efficiency and performance under specific conditions [21]. Data-driven evaluation models have emerged as pivotal tools for comprehensive design assessment, leveraging multimodal data sources to evaluate functionality, aesthetic appeal, and market viability [22]. Machine learning has been utilized to extract ontological data and facilitate deterministic evaluation of crowd-sourced concepts [23]. Metric modeling approaches vary significantly to address the unique challenges of different data types. In contrast, novelty detection in textual data uses semantic embedding and identifies novelty within a vector space [23]. Functional metrics for images use machine learning on annotated data or geometric analysis of contours [24], and volume fraction differences are calculated through binarization and volume comparison [25].

1.1 Knowledge Gaps.

Generative AI, particularly LLMs and vision-language models (VLMs), holds the promise to significantly augment this revolution and further infuse the design processes, including concept generation. Yet, *several challenges and questions remain* regarding the ability of generative AI models in their vanilla forms to produce concepts that demonstrate measurable improvements in key metrics such as novelty, usability, and feasibility [26]. These concerns stem from intrinsic limitations of the current AI technologies. We speculate that generative AI models in their standard iterations may fall short of fully realizing the ambitious objectives inherent in design concept generation, particularly in delivering tangible advancements in creativity, functionality, and user-centered solutions [27]. They often rely on predefined datasets and parameters, restricting creativity and applicability in unexplored design spaces [28]. Existing studies focus on optimization within well-defined constraints but frequently fail to generate truly novel concepts that diverge from traditional paradigms [29]. Performance optimization is typically limited to narrow engineering parameters, ignoring the holistic integration of multifunctional requirements, resulting in solutions that are optimal in one aspect but suboptimal overall [30]. Additionally, most current AI models do not evolve posttraining, indicating a significant gap in long-term adaptability [31]. There is also limited research on using text as a design medium despite its effectiveness in communicating ideas [6]. In this article, we address these challenges by (1) leveraging extensive data repositories and real-time internet access available to state-of-the-art VLMs and LLMs, overcoming narrow dataset limitations and ensuring continuous updates on the latest design developments; (2) enabling iterative and targeted refinement of design concepts through innovative actor–critic interactions; (3) fostering novelty through design-by-analogy approaches, particularly biologically inspired design, by exploring effective strategies in unrelated domains rather than reiterating existing solutions; and (4) utilizing a combination of text and image as design mediums, which is more conducive to ideation and initial design stages compared to image-only or other technical mediums suited for later design phases.

In light of these challenges, research on combining current methodologies with adaptable evaluation frameworks is limited. Existing methods often fail to incorporate real-time data updates and feedback [31], which is crucial for relevance in dynamic environments. However, *fundamental questions remain* regarding the practicality and feasibility of existing metrics for text+image concept generation. This dichotomy is particularly evident in AI-generated design concepts, where the open-ended exploration of the design space facilitated by generative AI models indeed fosters a heightened level of novelty but often at the expense of the practicality and applicability of the resulting concepts. Conversely, constraining the design space to be explored, for example, by narrowing the focus to specific challenges, enhances feasibility but reduces novelty [6]. In design innovation, novelty often arises from merging disparate ideas, where fusing remote or unrelated concepts can lead to highly innovative outcomes [32]. AI models often lack the knowledge and reasoning to discern the practicality of combining ideas [30], resulting in outputs that may not address real-world challenges [26]. The proposed framework in this work combines innovative ideas with practicality, comparing them systematically with recent developments across sectors. It is supported by a strong theoretical foundation and accessible Application Programming Interfaces (APIs), showcasing adaptability and generality. Validated across industries from automotive to medical devices, it operates efficiently with minimal data needs. In response to the lack of a data-driven model for design feasibility evaluation, we developed one based on an LLM.

Multiagent LLM systems present a promising frontier in AI-driven concept generation research. Yet, the potential of such a framework for design applications and the optimal configurations for design purposes remains largely unexplored. In this article, we harness the combined capabilities of design-specialized VLMs and LLMs, augmented with exploration, information retrieval, and feedback, to optimize solutions to design problems across various dimensions and metrics. This approach involves experimenting with multiple configurations to identify the most effective setup. Our exploration led to the development of an optimal framework that not only fosters innovation but also ensures the feasibility and relevance of the solutions to the specific design challenges at hand.

1.2 Contributions. We introduce an innovative multiagent LLM framework to boost the problem-solving abilities of VLMs as design agents. This framework employs a VLM model to generate an initial design concept for a specific application, which is then iteratively improved based on feedback from various critics. We validate this approach using the AskNature dataset, which we leverage to inspire creativity in design. Our results demonstrate the framework's ability to refine and enhance initial designs, improving novelty, feasibility, and problem–solution relevance. In summary, our contributions are as follows:

(1) We propose an actor–critic framework using a VLM model. This framework simulates a conversational concept design session between a designer and a critic. We explore different configurations within this framework to assess how various elements and dynamics influence the design process:
   (a) Different learning strategies for the VLM designer: fine-tuning versus few-shot learning.
   (b) Variations in the type of critics: qualitative versus quantitative evaluation models.
   (c) The impact of long-term memory (LTM): assessing the difference when the designer has access to all versus only the last round(s) of conversation.
   (d) The comparison of VLM versus LLM designer and critic agents to evaluate the influence of incorporating images as a design medium.
(2) We rigorously evaluate the framework's performance against two benchmarks: the state-of-the-art VLM model, GPT-4, and successful real-world examples of innovative biomimetic design (BID) from the AskNature dataset. We employ four data-driven quantitative evaluation models focused on novelty, feasibility, variety, and the relevancy between the problem–solution and inspiration source–solution. Notably, our team proposed the feasibility evaluation model.
(3) We outline several promising directions for future research in this domain, focusing on enhancing the proposed framework's scalability, adaptability, and application across diverse design theories.

2 Methods

This section outlines the theoretical underpinnings of our framework and delineates the distinctions between the various configurations we have investigated. Additionally, we detail the evaluation metrics employed in our study, providing a rationale for each and discussing their theoretical basis and algorithmic execution.

2.1 Actor–Critic Framework. The methodology is structured around the iterative interaction between two agents: the actor and the critic, where the actor generates design concepts and the critic evaluates them (Fig. 1). This section outlines the theoretical framework and mathematical formulation tailored to the study. In the conceptual design session, the actor acts based on the current state to reach the critic's satisfaction threshold. The critic evaluates the design based on predefined criteria, providing feedback that serves as the reward signal to guide the actor's future actions. The design session terminates when the critic's satisfaction (see Sec. 2.1.4) reaches a threshold, signaling that the design concept meets the desired criteria. This approach extends the capabilities of LLMs and VLMs beyond natural language generation applications to address complex problem-solving and evaluation tasks, thereby integrating creative generation with critical assessment within a single cohesive system.

2.1.1 Definitions. This section provides the formal definitions of the key components of the proposed framework. Illustrative examples of various components are provided in Table 1.

- *State (S):* The state is a critical element that encapsulates the comprehensive context of the design project at any moment in time. It is represented by the current design concept (including all its details such as specifications), requirements, previous iterations, and feedback from the critic.
- *Action (A):* Actions are the specific design choices made by the actor based on the current design state. These might include proposing a new design element, modifying an existing component, or selecting particular materials and methods based on the state's requirements and the reward, which directly influence the subsequent state and the rewards received. The action space in our framework is discrete and high

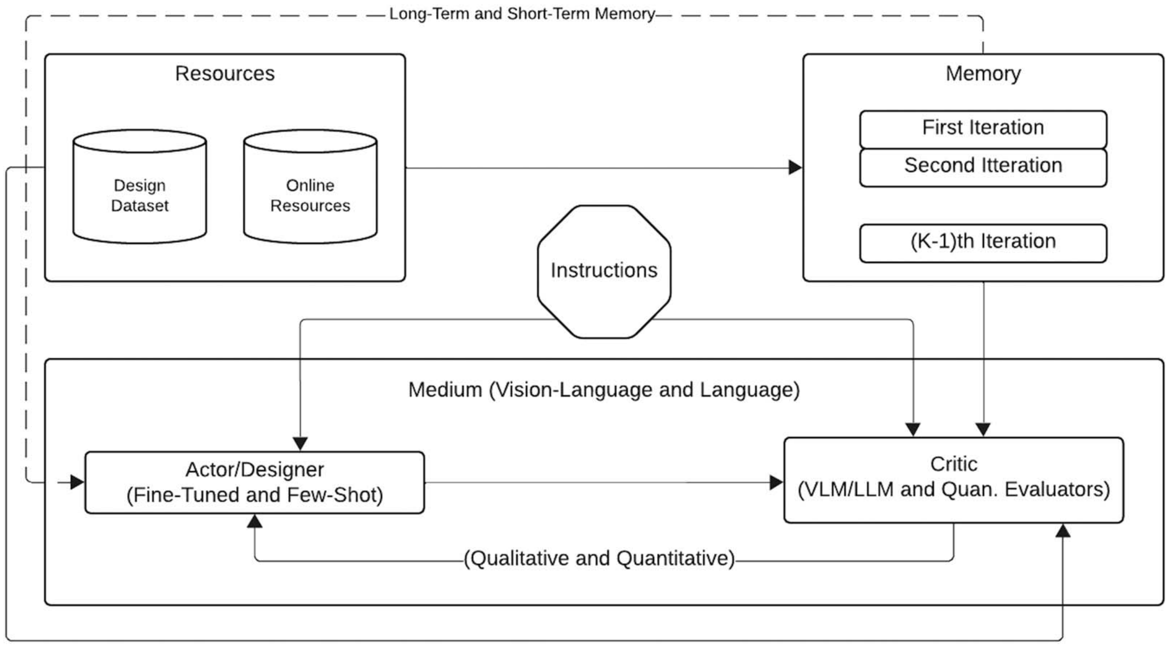


Fig. 1 A schematic illustration of the proposed multiagent, actor–critic LLM framework, containing various configuration options

dimensional, reflecting the depth and interrelatedness of design modifications. Each action encompasses multiple layers of abstraction. For instance, a decision to modify a design may involve considerations of style, function, cost, and compliance with design standards—all integrated within a single action.

- *Reward (R):* The reward is the assessment provided by the critic after the actor proposes a design solution, indicating how good or bad the action was in helping the agent achieve its goal. This feedback evaluates the design's novelty, feasibility, sustainability, or any other relevant metrics that define the success of the design according to the design's objectives, which is either a descriptive content or a set of quantitative values.
- *Policy (π):* The policy is the algorithmic behavior encoded within the actor, determining how it generates new designs or modifies existing ones based on the input it processes. It is the decision-making function for the actor agent, which encompasses the VLM's pretraining setup (frozen) and the attention weights that are influenced by the provided prompts (dynamic). While the weights remain unchanged, the model's behavior adjusts based on the provided instructions, representing a form of in-context learning.

2.1.2 Theoretical Framework. The following outlines a structured approach to formulating our framework, with an emphasis on the feedback loop and decision-making process. The state can be defined as follows:

$$s_t = (D_t, H_{t-1}) \tag{1}$$

where D_t represents the current design details and H_t encapsulates the history of all past states, actions, and feedback. The multimodal feedback function used by the critic is then modeled as a composite reward function that returns a vector of feedback scores or comments:

$$r(s_t, a_t) = (r_1(s_t, a_t), r_2(s_t, a_t), \ldots, r_n(s_t, a_t)) \tag{2}$$

where $r_i(s_t, a_t)$ denotes the output of the critic element, which may feature a function assessing a specific aspect or metric of the design, such as feasibility or novelty. Alternatively, it could be a specialized language model analyzing the concept from various perspectives. Consequently, r can comprise a set of numerical values or a textual description detailing the most limiting flaw. For more details on the critic, refer to Sec. 2.1.4. The state update function Φ explicitly incorporates the history and feedback to update the state:

$$s_{t+1} = \Phi(s_t, a_t, r(s_t, a_t)) \tag{3}$$

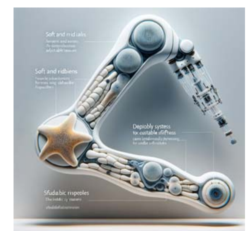
Key instructions for the prompt, along with an example of how to execute Eq. (3), are detailed in Sec. 2.1.3. Here, there could be an additional step to frame the state s_t before passing it to the critic. As each updated concept emphasizes new features pertinent to the latest round of refinement, less information about previous rounds is retained in it. Thus, this step might involve an additional LLM prompted to summarize the concept, listing innovation details from all steps. If this step is included, the critic evaluates the current state under the assumption that the latest action was optimal. Without this step, the action itself is evaluated in the context of the current state. We choose not to include this additional step, allowing for the reversal of unhelpful or ineffective proposed actions (i.e., design refinements) in subsequent rounds. Notably, in some experiments, the actor replaced previously proposed solutions with alternatives that better suited the context.

The neural network of the VLM/LLM actor, inherently embodies the policy π, which dictates the action a_t based on the state s_t and reward $r(s_t, a_t)$:

$$a_t = \pi_\theta(a_t|s_t) = \text{VLM}_\theta(P_t) \tag{4}$$

where VLM_θ represents the pretrained model with frozen weights θ and P_t denotes the dynamically constructed input prompt incorporating the state and critic feedback. VLMs/LLMs utilize the self-attention mechanism to identify which elements of the input prompt are most relevant at each step of the text generation process. This mechanism enables the model to effectively focus on particular components of the prompt, such as the critic's numerical or descriptive feedback, guiding its output in a context-sensitive manner. During inference, while model's weights remain static, the prompt sequence plays a pivotal role in shaping the model's behavior (i.e., the policy function). By embedding evaluation-based feedback directly into the prompt, we impose indirect constraints

Table 1 An example of key components in our framework

Definition	Example	
State (s)	The proposed robotic system is inspired by the adaptive features of starfish arms, leveraging their ability to alter stiffness through mutable connective tissues. Combining soft and rigid materials, this innovation aims to replicate the precise, gentle, and versatile movements of starfish. Designed for delicate operations, the robot arms can stiffen to exert more force when needed, facilitating a range of tasks from handling fragile objects to applying significant pressure. This technology could revolutionize robotic manipulation, enhancing performance across diverse environments and tasks.	
Action (a)	The improved design of the robotic system incorporates a network of fluidic chambers within the robot's limbs, drawing inspiration from the hydraulic systems found in cephalopod limbs. This enhancement allows for individual pressurization of each chamber, dynamically adjusting limb stiffness and movement for better scalability and efficiency in controlling mechanical properties. By adopting a fluidic approach, the design ensures more precise control over limb behavior during complex tasks, addressing previous limitations in scaling and controlling variable stiffness in larger robots.	
Reward (r)	A significant limitation of implementing this robotic system on a practical scale lies in the complex challenge of effectively integrating and controlling variable stiffness in larger robots. Although promising at the microscale, scaling up these mechanisms to address real-world dynamic tasks presents a major engineering hurdle. The primary flaw is the difficulty in managing and controlling the variable stiffness efficiently in large-scale robots, which is crucial for handling a wide range of mechanical interactions with the environment.	
State (s')	The innovation aims to develop a robotic system that emulates the mutable connective tissues of starfish arms, blending soft and rigid materials for adjustable stiffness. This design facilitates precise and gentle movements for delicate operations, while also allowing for increased stiffness during more force-intensive tasks. To enhance control over stiffness in larger systems, the improved design integrates a network of fluidic chambers within the robot's limbs, inspired by cephalopod hydraulic systems. These chambers can be dynamically pressurized, enabling scalable and efficient adjustments of limb stiffness and movement. Drawing from both starfish and cephalopod biological models, this system combines adaptability in small-scale locomotion and object interaction with the scalability needed for complex, large-scale robotic applications.	

or a soft form of policy guidance that directs the model's response generation. Section 2.1.3 provides some influential prompting details observed during our experiments.

The feedback function and the state update explicitly consider the long-term memory stored in H_t, ensuring that the decisions made at any iteration t are informed by the outcomes and learnings from all previous iterations:

$$H_t = \begin{cases} H_{t-1} \cup (s_t, a_t, r(s_t, a_t)) & \text{(long} - \text{termmemory)} \\ (s_t, a_t, r(s_t, a_t)) & \text{(short} - \text{termmemory)} \end{cases} \quad (5)$$

The history plays a crucial role in shaping the actor's behavior by guiding its attention and token processing, making it a determining component in how the design process evolves. When utilizing long-term memory, as modeled in Eq. (5), all interactions from all iterations are included in the history. In contrast, the short-term memory configuration excludes H_t from the update function, resulting in a history that only comprises the latest state, action, and reward (refer to Sec 2.1.5 for more details).

This framework introduces a novel paradigm for design concept generation where learning is conceptualized as a series of design feedback and enhancements. This approach not only simplifies the computational and data demands typically associated with training a model, but also aligns better with practical workflows where model parameters may remain fixed once deployed.

2.1.3 Actor. This comparative analysis focuses on enriching the study by comparing the actor–critic framework with two distinct learning approaches for the designer component: few-shot learning and fine-tuning, both utilizing GPT-4 as the underlying model. This comparison aims to evaluate the capabilities and limitations of these learning strategies in generating innovative conceptual designs. The dataset utilized in this case study comprises instances of successful innovative conceptual designs, where the innovation is derived from natural elements or systems. Each instance within the dataset is annotated with specific attributes: the application or industry it pertains to, and the benefits of the design. These annotations serve as inputs for queries in both the few-shot learning and fine-tuning processes designed to create the actor VLM/LLM.

Few-Shot Learning. While LLMs are capable of identifying the intended task even with zero-shot prompting and through follow-up prompts, more complex tasks, such as design concept generation, benefit significantly from enhanced alignment with desired responses by employing in-context learning through few-shot prompting [33]. Given that our framework incorporates a conversational refinement process, wherein each iteration's output depends on the actor's generated responses, each response must be well refined and accurately shaped. This ensures the final result is achieved without requiring further refinement through additional queries to the actor in each iteration. This process involves providing the VLM/LLM with a small number of annotated examples from the dataset as context in the prompt. This approach aims to leverage GPT-4's extensive pretrained knowledge base, guiding it to generate outputs that reflect the patterns and attributes of the provided examples.

We experimented with various wordings and instructional orderings for few-shot prompting to evaluate their impact on agent responses. Specifically, we tested (1) ordering the prompt with the instruction first followed by the assignment, and vice versa; (2) implicitly mentioning or excluding the set of metrics against which the concept should be evaluated; and (3) explicitly assigning a designer or critic persona or excluding it. None of these variations affected the quality of the responses from the agents. It is worth noting that we strategically opted not to use reasoning and logic-related prompting techniques, as the AskNature dataset inherently employs a chain-of-thought prompting technique by including descriptions of the design, resulting features from the concept, and crucially, the reasoning behind adopting the analogy. This intrinsic fortification obviates the need for additional reasoning prompts. We observed a low sensitivity to the language used in the prompts and no significant differences in performance among all approaches, provided that the prompt adheres to specific guidelines: (1) The prompt should specify that the objective of the provided examples is to learn the principles of design by analogy, specifically how to draw inspiration from a distant source to solve a given problem. (2) The prompt should specify that when generating the concept, only the details of the innovation and the biological model should be shared. (3) To emphasize novelty, the inspiration source and its application should not be among the provided examples. (4) In the refinement steps following the initial concept generation, the prompt should state that the response must be the revised concept, not instructions on how to achieve it, and that limitations should be addressed while preserving advantages. The followings are examples of such prompts in the context of the bioinspired design case study conducted in this article:

Initial Prompt—Initial Round:

"[Multiple examples from the design related dataset] Review the examples provided from the design-related dataset. Learn how to draw inspiration from nature and map it to the design domain to create innovative design concepts. Develop a novel design concept that can be used in [desirable application/industry] for [desirable benefit/design objectivel. Share only the innovation details and the biological model. Do not utilize the biological source and its application in the examples."

Secondary Prompt—Following Rounds

"The following is feedback from a design critic on your improved design concept. Address the identified limitations while maintaining the positive aspects of your design. Then, share the improved version with me, including only: (1) the improved innovation details and (2) any additional sources of inspiration if used. The improved version must contain the solution itself, not the steps to achieve it. [critic's feedback]"

Fine-Tuning. This process adjusts the weights of the VLM/LLM using a larger subset of the dataset, thereby customizing the model's responses to be more aligned with the domain-specific characteristics. More details on the dataset and implementation are provided in Sec. 3.1. Each instance in the dataset is created in a JSONL format, where the first set of braces contains the user message, which defines the problem or request, and the second set of braces contains the model's response, representing the generated concept for the given problem. The template for a data instance follows the same structure used in the inference phase, but with keywords specific to the design problem being substituted.

{"messages": [{"role": "user", "content": "Given the specified application/ industry: I and the desired benefits/objectives: I, design a concept inspired by a biological system. Provide the innovation details and the corresponding biological model."}, {"role": "assistant", "content": "Innovation Details: [] Model: []"}]}

Table 2 illustrates a couple of examples of the template's fillable spots, showing how various components can be replaced with relevant values.

2.1.4 Critic. To enrich the study on our framework, we introduce a comparative analysis between two distinct forms of critic: qualitative (VLM/LLM) critics and quantitative critics employing quantitative metrics. This comparison aims to explore the dynamics of feedback in the design process and its influence on the evolution of design concepts. This comparison not only contributes to our understanding of effective critique mechanisms in AI-driven design systems but also provides valuable insights into optimizing the actor–critic framework.

Qualitative Critic. For the qualitative critic, the feedback to the actor is grounded in quantitative metrics that assess the novelty, feasibility, and relevancy of the solution–problem and solution–inspiration source relationships. These metrics are calculated using the evaluation methodology detailed in Sec. 2.2. Importantly, the feedback provided to the actor is labeled with the corresponding metrics, ensuring clarity regarding which aspect each feedback point pertains to. This approach allows the actor to receive targeted insights on specific areas of the design concept.

Qualitative Critic. We opted to leverage GPT-4—in both forms of vision language and language only—as our qualitative critic. In this approach, the model is prompted to conduct a comprehensive analysis of the design concept, which is presented through a descriptive text (and possibly an accompanying image). The model is prompted to use an Internet search specifically when assessing feasibility, ensuring that its critique is contextualized with the latest advancements and developments within the same domain. The search results affect only the feasibility aspect of the analysis rather than the entire critique, ensuring that the model's evaluation remains relevant, even in cases where data on novel problems may be sparse or nonexistent.

The Internet search is integrated into the critic's evaluation in a structured manner depending on the availability and type of data. When a technology is well established and widely adopted, the model focuses on evaluating practical aspects like manufacturing scalability and potential cost specific to the design concept, or alternatively, ease of implementation. In scenarios where the technology is emerging or still in the research and development phase, the feedback discusses existing limitations and challenges associated with transitioning the technology from experimental to industrial settings. Finally, in cases where information on a proposed concept is either scarce or completely absent, the critic still manages to provide relevant feedback by drawing on existing technologies, processes, or solutions that the proposed concept could be building upon or modifying. This structured approach limits the impact of search results to specific domains of feasibility and manufacturability, while other critical dimensions—such as novelty, contextual appropriateness, and maintenance concerns—are independently addressed within the critique based on the details of the concept.

The critique systematically evaluates various dimensions such as feasibility, novelty, and so on, weighing their relative significance. The final feedback provided to the actor emphasizes the most critical limiting flaw, drawing from a broad understanding of the domain and the specific details of the presented design. To foster a rich and constructive feedback mechanism, the qualitative critic maintains a long-term memory of the entire conversation. This capability enables the critic to compare current design iterations

Table 2 Examples of fillable spots in the JSONL template used for fine-tuning the actor

Application/ industry	Benefit/ objective	Innovation details	Model
3D printing	Sustainable eco-friendly	The MycoPrint system leverages the self-assembling capabilities of mycelium, the vegetative part of a fungus consisting typically of a mass of branching, threadlike hyphae, to create intricate 3D structures.	Mushrooms undergo a unique growth process which involves a network of mycelium growing in all directions, eventually forming a complex, natural 3D structure—the mushroom fruiting body we are familiar with.
Construction architecture	Resilience	Inspired by bamboo's sectional, hollow structure, this innovative concept involves the development of modular building elements that mimic bamboo's growth patterns and structural properties …	Bamboo is a highly versatile plant known for its exceptional growth rate and structural strength.

with previous ones, enhancing the relevance and specificity of the feedback while avoiding redundancy. Furthermore, to ensure the feedback from the VLM critic is as informative as possible, it is structured in an explanatory format rather than being conveyed through numerical values or simplistic labels. This decision is predicated on the belief that nuanced, descriptive feedback will better guide the actor in addressing the identified limitations. The following provides an example of correct and comprehensive instructions for the critic to generate a desirable response:

"The following is a design concept. Analyze it from various aspects and metrics, and critique it by concisely identifying the most limiting flaw. When assessing the feasibility and manufacturability of the concept, first search the internet to determine if the necessary technologies for this innovation are already developed or under research in any domain. If such technologies exist, the concept is considered feasible."

2.1.5 Memory. In our study, we investigate the influence of memory constraints on the design process within an actor–critic framework by comparing two distinct models: one endowed with LTM capabilities (as explained in Sec. 2.1) and another with short-term memory (STM), where the actor accesses only the last design iteration. Designing an STM configuration was motivated by the inherent memory limitations of contemporary LLMs like GPT-4, which are constrained by a strict token size limit for inputs. Given that our framework involves a conversational setup essential for generating successful design concepts, the length of conversation may often exceed these token limits. Consequently, we have designed an alternative, memory-independent framework to evaluate whether the objectives of concept generation can be achieved with a simplified version of our framework.

The design of STM was inspired by concepts from Markov decision processes, which emphasize the importance of the current state in decision-making. In the STM configuration, the decision-making process focuses on the immediate past state, which includes the latest concept, its corresponding feedback, and a set of pertinent instructions. This model simplifies the examination of design evolution in scenarios devoid of long-term memory. The theoretical underpinnings of this configuration remain largely similar to our primary framework, except for the history definition, which no longer integrates the previous iterations in H_t. Instead, it encompasses only the immediate previous state, adhering to the Markov property, which posits that the future state depends solely on the current state and the action taken, disregarding any historical context of states and actions.

The nature of interactions with an LLM is typically stateless; the model does not inherently retain the memory of previous interactions unless such context is explicitly incorporated into each new prompt. To simulate a continuous dialogue or a series of related design iterations, it is crucial to embed relevant previous context within each new prompt to ensure coherence and continuity in the model's responses. Therefore, distinguishing between long-term and short-term memory configurations in our study involves ensuring that all pertinent instructions and historical context are provided in each iteration of the design process when employing a long-term memory setup. This approach enables a nuanced exploration of how memory limitations impact the efficacy and creativity of the design process within our VLM-based actor–critic framework.

2.2 Evaluation Metrics. The selection of evaluation metrics for a study is inherently dependent on specific use cases and application domains. Our metric selection was guided by three primary factors: (1) our central focus on design ideation and conceptual design; (2) adherence to design goals and standards; and (3) comprehensiveness and consideration of all dimensions of a concept's success. Among the critical and standard evaluation metrics in design, as outlined in Sec. 1, our evaluation framework integrates four distinct metrics: novelty, feasibility, variety, and relevancy.

Novelty evaluation, one of the most critical metrics in design ideation, encourages the generation of unique solutions, pushing beyond conventional and predictable ideas. Novelty and creativity are intrinsically interdependent in the design process. Creativity generates the ideas that lead to novel solutions, and the presence of novelty serves as an indicator of creative achievement [34]. Feasibility, on the other hand, grounds our innovative concepts in practicality, assessing the extent to which these ideas can be realistically implemented within current or near-future technological paradigms. This balance is essential for advancing designs from conceptual sketches to real-world applications.

Variety complements novelty and feasibility by evaluating the breadth and diversity of generated solutions. It measures how distinct and varied the design concepts are from one another, ensuring that the exploration of the design space is extensive and not limited to a narrow set of ideas. Finally, relevancy serves as a critical bridge, ensuring that the solutions are not only innovative and feasible but also pertinent to the problem space they intend to address. By employing a method of domain mapping, we ensure that our design solutions maintain a meaningful connection to their inspiration sources. Relevancy is crucial since (1) generated concepts can be conceptually intriguing but irrelevant to the problem requirements or (2) if inspired by another source, they may ignore the critical constraint and objective differences between the source of inspiration and the problem, making the analogy ineffective or misleading (refer to Sec. 3 for more details).

Collectively, these metrics provide a comprehensive framework for evaluating design concepts: novelty guarantees originality, advancement, and competitiveness; feasibility ensures practicality and thus the likelihood of real-world adoption; variety reduces the chances of fixated and suboptimal ideas; and relevancy ensures targeted and contextually appropriate solutions.

2.2.1 Novelty. To effectively measure the novelty of the generated design concepts, we adopt a well-established quantitative approach in design literature that integrates three primary aspects: the number of components/functionalities, the number of stages in the design process, and the uncommonness of each component [32]. The novelty of a concept is computed as follows:

$$N = \frac{\sum_{j=1}^{m} c_j \sum_{k=1}^{n} g_k n_{jk}}{\sum_{j=1}^{m} c_j \sum_{k=1}^{n} g_k} \tag{6}$$

where N is the novelty score of a design concept with m functionalities or attributes and n stages; c_j are the weights assigned to each component/function based on their importance; g_k are the weights according to the importance of each stage; and n_{jk} represents the novelty score for each component at each stage.

This metric is specifically tailored to evaluate designs inspired by natural sources, aligning with our focus on innovative conceptual design. The first aspect, number of components/functionalities (m), recognizes each distinct functionality introduced in a design concept. This measure is particularly pertinent to our study, given its emphasis on leveraging natural systems to foster innovation in design. By quantifying the number of functionalities, we can assess the complexity and richness of the design concepts, with a higher number of functionalities indicating a potentially higher novelty. To maintain an open-ended design space, we assign uniform weights to all functionalities (c_j). The second aspect, number of stages (n), is fixed to 1 in our framework, reflecting the single stage nature of conceptual design. The third aspect, uncommonness or unusuality of each component (g_{jk}), is gauged based on how unexpectedly a component fulfills a specified functionality. To quantify this dimension of novelty, we employ the framework proposed in Ref. [23], which introduces span metric for assessing novelty by measuring the conceptual distance between entities in a design concept and the context using a knowledge graph, offering an indication of how disparate and thus potentially innovative the components are within the context of the design.

2.2.2 Feasibility. In the evaluation of our design concepts, a pivotal component is the feasibility assessment, which quantifies the practical implementability of each design concept generated by our framework. This assessment adopts a well-established weighted average methodology to evaluate the feasibility across different components and stages (again considered 1 in our study) within the design process [32]. The methodology integrates a detailed examination of the alignment of each design component with current or near-future technologies, employing a semantic analysis framework. The feasibility score, F, for a design concept is calculated using the formula:

$$F = \frac{\sum_{j=1}^{m} c_j \sum_{k=1}^{n} g_k f_{jk}}{\sum_{j=1}^{m} c_j \sum_{k=1}^{n} g_k} \tag{7}$$

where f_{jk} is the feasibility score for functionality j at stage k. To accurately measure the feasibility of each functionality (f_{jk}), we propose a quantitative approach as follows:

(1) *Key Phrase Extraction:* For each functionality within a design idea, we extract key phrases that describe its design idea.
(2) *Technology Retrieval:* By utilizing GPT-4, we search for the closest available technologies and in-progress research projects related to the extracted key phrases, focusing on technologies that are either currently available or can be developed based on existing technologies.
(3) *Semantic Similarity Analysis:* The extracted part of the design concept pertaining to a specific functionality, along with the retrieved technological information, is embedded into a vector space using the text-similarity-davinci-001 model introduced by OpenAI. This model is optimized for deep semantic understanding, allowing for nuanced comparison between the design concept and potential technological implementations.
(4) *Cosine Similarity Calculation:* We calculate the cosine similarity between the vector representations of the design functionality and each instance of the corresponding technological data. This calculation yields a numerical output for each instance, ranging from −1 to 1. We observed that the number of technology instances retrieved in step 2 is directly related to the feasibility of a specific functionality, with a greater number of instances suggesting higher feasibility. Consequently, we aggregate the cosine similarity scores from all returned instances, considering both the breadth of feasibility—reflected by the quantity of similar technologies—and the degree of similarity.
(5) *Overall Score:* Finally, we calculate the average of feasibility scores across all functionalities.

The resulting score represents a quantitative measure of the closeness of a design concept to being implementable.

To validate the performance of the feasibility evaluation algorithm, we conducted benchmarking against ground truth data. Due to the absence of an existing annotated dataset, we created a small dataset comprising 30 samples, categorized into three groups: futuristic, advanced, and ordinary concepts, which correspond to low, medium, and high feasibility, respectively. An example from each category is presented in Table 3. Each sample was evaluated using our feasibility evaluation algorithm, and the results were aggregated by category. The statistical measures are illustrated in Fig. 2, demonstrating the algorithm's capability to distinguish among the categories. It is important to note that, as the algorithm has not been tested on a quantitatively annotated dataset, the accuracy of the output values cannot be conclusively determined. However, the results indicate that the algorithm performs well in comparative evaluation and predicting the feasibility ranking of a concept.

2.2.3 Variety. The variety evaluation algorithm aims to quantitatively assess the diversity of design concepts by leveraging natural language processing to calculate the intragroup similarity of the concepts. The model initially processes design concepts by vectorizing them using a pretrained BERT (Bidirectional Encoder Representations from Transformers) model. The embeddings of the [CLS] token are utilized as the feature vectors, capturing the semantic information of each concept. Subsequently, the pairwise similarities between the feature vectors are computed using cosine similarity, generating a similarity matrix that quantifies the similarity between each pair of design concepts.

Given that variety is intrinsically relevant only in the presence of a set of concepts rather than individuals, we evaluate the metric on all generated concepts pertaining to a specific configuration of our

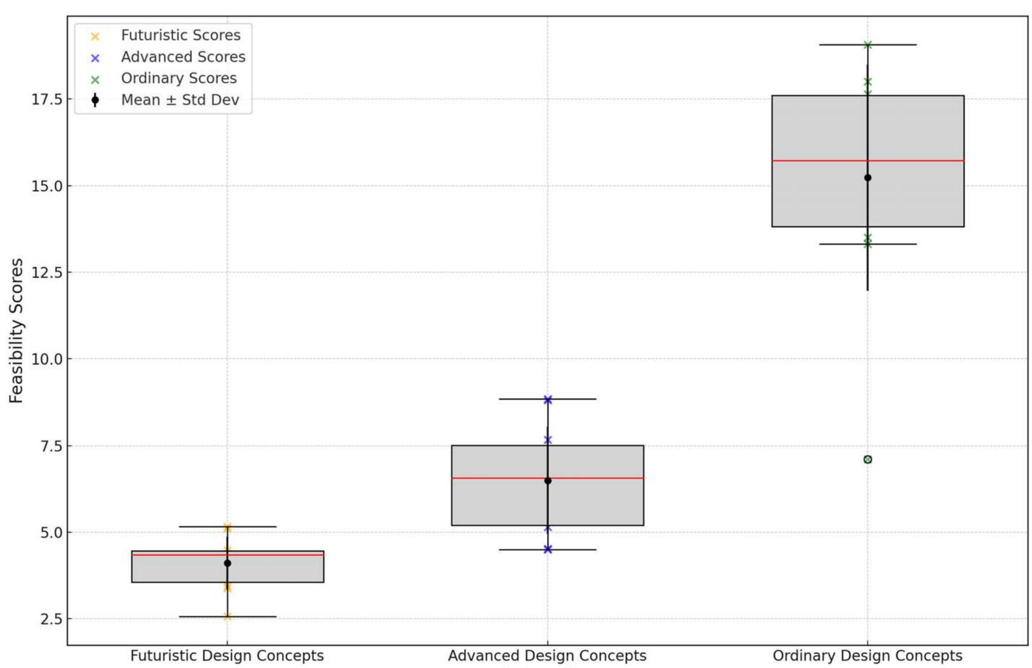


Fig. 2 The results of benchmarking the feasibility algorithm against ground truth data

framework. This approach facilitates the comparison of intergroup variety among different configurations. Thus, the similarity matrix is visualized using a histogram, displaying the distribution of similarity scores. Additionally, statistical measures, including the mean, standard deviation, minimum, and maximum similarity scores, are computed for the upper triangle of the similarity matrix (excluding the diagonal).

2.2.4 Relevancy. For relevancy evaluation, we adopted the method proposed by Ref. [6]. This methodology focuses on evaluating the relevancy between domains in problem–solution and nature–solution mappings through binary text classification tasks. *Problem–solution relevancy* refers to the degree of alignment between a specific problem (which encompasses the application context and the desired benefits) and the proposed solution (the innovation). In this context, a relevant problem–solution mapping indicates that the innovation directly addresses the application's needs and effectively delivers the intended benefits. *Nature–solution relevancy*, on the other hand, evaluates the extent to which the proposed solution is appropriately inspired by or derived from a biological source (biomimicry). A relevant nature–solution mapping ensures that the biological inspiration is directly applicable to the innovation, maintaining a clear and logical connection between the natural system and the engineered solution.

To evaluate these relevancies, we constructed datasets containing positive and negative samples for fine-tuning classifiers (based on the GPT-4 model). Positive samples represent cases where the problem–solution or nature–solution mappings are appropriate and relevant. In contrast, negative samples are generated by substituting elements (e.g., replacing the innovation with an unrelated one) to create irrelevant mappings. Additionally, a generator was employed to produce these negative samples by creating random innovations that disregard the original problem or biological inspiration. For both relevancy models, we used binary cross-entropy loss, a standard loss function for binary classification tasks, to measure the difference between the predicted probabilities and the actual labels (relevant or irrelevant). The consistent convergence of the loss values indicates that the models effectively minimized classification errors and learned the intended mappings. The hyperparameter values for training are provided in Appendix C, which presents the training and validation loss. We conducted three independent training runs for each model to ensure robustness, and in all instances, the loss values consistently converged within 100 steps. On average, the problem–solution model achieved a validation loss of 0.39, an accuracy of 0.83, a precision of 0.84, and a recall of 0.83, while the nature–solution model achieved a validation loss of 0.13, an accuracy of 0.94, a precision of 0.95, and a recall of 0.93.

3 Experiments and Results

In this section, we rigorously assess the performance of the proposed framework and its constituent elements in generating novel, feasible, and relevant solutions based on a biomimetic design dataset.

3.1 Dataset. The dataset employed in this study is derived from the Innovation section of the AskNature website [35], a well-known resource that catalogues a wide array of biomimetic designs and projects inspired by natural systems, organisms, and processes. It is a comprehensive repository demonstrating how nature-inspired principles and strategies are applied across diverse fields to foster sustainability and innovation. Each entry in the dataset provides detailed information on the biological model, the challenge addressed, the biomimetic approach, and a narrative of the biomimicry story. With 309 BID samples, the dataset offers a rich foundation for exploring diverse learning strategies for language models. It is important to note that the primary purpose of this dataset may not be creative. However, since creativity can emerge through analogical transfer [36], with our adoption approach, this dataset serves as a valuable resource for creative designs. Furthermore, we utilized this dataset as a case study to showcase the proposed framework's performance, rather than to enhance the creativity capabilities of the framework itself.

3.2 Experiments. Each entry in the dataset is initially structured into five components: application, benefits, challenges, innovation details, and biomimicry story. In design terms, the application refers to the industry or sector, benefits represent the design objectives, and innovation details provide technical descriptions. To optimize the use of this dataset, we reorganized each entry into specific input–output pairs, allowing for more targeted model outputs to suit our needs. In particular, we focused on using the "innovation details" and "biomimicry story" as outputs. For the inputs, we experimented with different combinations of the application, benefits, and challenges. Including the "challenges" in the input data often increased the feasibility of the designs but occasionally reduced their novelty, as noted in prior studies [6]. To balance novelty with practical implementation, we ultimately chose to focus on "application" and "benefits" as inputs and exclude "challenges" in both the few-shot learning and fine-tuning configurations. Once we structured the data into input–output pairs, we applied them in two ways: (1) as training data for the actor, either through fine-tuning or few-shot prompting; and (2) as a benchmark for comparing the generated concepts.

Our proposed framework incorporates multiple configurations, each designed to explore different components of the framework. Rather than solely focusing on evaluating its overall effectiveness, our experiments compare the performance of various configurations. This approach helps us analyze how different components interact and contribute to the overall outcomes, ultimately guiding us toward identifying the optimal configuration.

For clarity, we define the main configuration—which includes a few-shot prompted actor, qualitative critic, vision-language medium, and long-term memory—as the default configuration for comparison (as detailed in Sec. 2.1). To isolate the effects of individual components, we ensured that in each alternative configuration, only one component was modified from the default setting. This controlled approach allows us to understand the specific influence of each component on the overall system.

We conducted ten conversational experiments for each of the six configurations tested: the main configuration, fine-tuned actor, quantitative critic, language-only medium, short-term memory, and GPT-4 as a benchmark, generating a total of 60 design concepts. Additionally, we selected ten examples from AskNature to serve as another benchmark. Each design concept was evaluated using our evaluation metrics, and the results were aggregated by configuration group. The statistical analyses of the aggregated results are presented in Secs. 3.4–3.7. Each section provides a focused comparison, highlighting how the alteration of one component impacts performance relative to the main configuration and the benchmarks.

3.3 Benchmark and Baseline. We establish two benchmarks to serve as comparative standards. The first benchmark is based on the capabilities of GPT-4. As a versatile and powerful language model, GPT-4 provides a robust baseline against which the performance of our framework can be assessed, facilitating an understanding of how our specialized actor-critic approach compares to the general purpose AI in generating innovative design concepts. The second benchmark is derived from a subset of the AskNature dataset, specifically selected to function as validation data. This subset comprises successful instances of BID concepts that have already been implemented in real-world scenarios. Given their proven efficacy and innovation, these instances represent ideal real-world benchmarks, grounding our evaluation in tangible outcomes and established successes in biomimicry. By not exposing these samples to the model prior to validation, we ensure an unbiased assessment of our framework's performance. We select several example applications from different industries to evaluate the performance across benchmarks and the corresponding configurations of our framework.

We have carefully considered the potential for comparing our method to existing debate-style and role-playing LLM frameworks but found them less suited for a direct comparison. Multiagent LLM systems typically assign roles and metrics specifically tailored to the task they are designed for. In our case, the framework is explicitly designed for design concept generation, with key evaluation metrics such as novelty, feasibility, and relevancy. These metrics are central to the design process and differ from the objectives of existing debate-style frameworks, which often focus on reasoning, explanation, or factual consistency. However, we acknowledge existing frameworks' potential value for integration to design tasks and even enhancing certain aspects of our framework. For instance, while a framework like LEGO [11] could enhance problem-solution relevancy by providing causal explanations for design choices, it would not directly contribute to generating novel or feasible designs, which is our primary goal. Consequently, we have benchmarked our method against models and datasets designed for concept generation, such as AskNature, which better aligns with the goals and metrics of our framework.

3.4 Few-Shot Prompted Versus Fine-Tuned Actor. This analysis is focused on comparing the effects of two distinct methods of specialization of the designer agent within the actor–critic framework, examining how each method impacts the overall system. We explore two types of learning: few-shot prompting and fine-tuning. For the few-shot prompting approach, we randomly hand-picked seven instances from seven different applications/industries in the AskNature dataset and prompted the GPT model with specific attributes—application, benefits, innovation, and biological model—of each instance. In contrast, for the fine-tuning approach, we divided the dataset into training and validation datasets with a 4:1 ratio. We further deepened our analysis by examining the impact of varying extents of training; specifically, we trained two GPT models, one for three epochs and another for ten epochs. The trends in training and validation losses for both models are depicted in Appendix C. The results indicated that when trained for only three epochs, the model struggled to discern patterns in the data, with final loss values approximately at 1.3 for training and 2 for validation. Conversely, when the model was trained for ten epochs, it exhibited signs of overfitting the training data, as evidenced by a training loss approaching zero, while the validation loss continued to increase after three to four epochs. This suggests that the designer agent is challenged in identifying the input–output relationship, likely due to the complexity of the task. This complexity is particularly evident in design scenarios, where a given set of industry-specific applications and design objectives could correlate with hundreds of potential design solutions.

In our study, we observed that both fine-tuned designer agents tend to exhibit certain patterns of hallucination, which manifest in three distinct forms: partial answers, repeated responses, and the generation of nonsensical material (Table 4). Specifically, partial answers often emerge as incomplete responses that, for instance, only provide the biological model without addressing the full scope of the query. Repeated responses are another common issue, where the designer produces the identical answer across several consecutive rounds of conversation, indicating a lack of adaptability and responsiveness in the learning model. Additionally, the production of nonsensical material further underscores the challenges faced by the fine-tuned model in reliably generating contextually appropriate and logical content.

We conducted a comprehensive statistical comparison of two training configurations—few-shot prompting and fine-tuning—alongside our two benchmarks, the results of which are detailed in Fig. 3. The analysis reveals that on average, few-shot prompting significantly outperforms fine-tuning, scoring 37.34 points higher in terms of novelty and 2.73 points higher in feasibility. It is important to note that these results reflect only the successful outputs from the fine-tuned models, which constituted 20–30% of all experiments, as concepts exhibiting hallucination were excluded from the statistical analysis. Additionally, the metrics for problem–solution and nature–solution relevancies were consistently met across all concepts, except in some instances within the fine-tuning configuration. Taking into account all the data presented in this section, it is evident that few-shot prompting emerges as a more effective method for specializing a GPT model in the role of a designer.

3.5 Qualitative Versus Quantitative Critic. This analysis is dedicated to scrutinizing the dynamics between different configurations of the critic agent, specifically comparing the performance of GPT-based models versus design-specific data-driven evaluation models, and contrasting the types of feedback—qualitative versus quantitative—used within these configurations. By examining these variations, we aim to discern which combinations yield the most effective critique in terms of enhancing the design process. A statistical comparative analysis of the four categories, each comprising ten design concepts, is presented in Fig. 4. These examples and analyses not only include detailed comparisons of our experimental setups but also juxtapose these with benchmark examples to provide a comprehensive view of how each configuration performs under similar conditions allowing for the assessment of the nuances in feedback effectiveness and agent performance.

In the quantitative critic configuration of our experiment, specific thresholds for each metric were established based on the average values observed in the main configuration, serving as the criteria to halt the iterative design conversation. The thresholds were set at $novelty = 45.2$, $feasibility = 8.3$, $problem - solution\ relevancy = 1$, and $nature - solution\ relevancy = 1$. Notably, the novelty score consistently increased and the threshold was met within five to six iterations, demonstrating a rapid achievement of innovative concepts. However, the feasibility threshold exhibited

Table 3 Examples from feasibility dataset along with corresponding labels (i.e., categories)

Category	Example
Futuristic design concept (low feasibility)	Temporal manipulation wristband: The temporal manipulation wristband allows users to manipulate their perception of time, slowing down or speeding up their subjective experience. It works by interfacing with the brain's temporal processing centers using advanced neural stimulation. The wristband features a holographic display for control and monitoring, and it is powered by energy harvested from body heat and motion.
Advanced design concept (medium feasibility)	Augmented reality glasses: AR glasses that overlay digital information onto the real world. They feature advanced optics, high-resolution displays, and sensors for environment mapping. The glasses are equipped with voice and gesture controls, and they connect to smartphones for additional functionality.
Ordinary design concept (high feasibility)	Bluetooth speaker: A portable Bluetooth speaker with high-quality sound and a built-in rechargeable battery. It features a durable, water-resistant design, easy pairing with devices, and control buttons for volume and playback. The speaker is available in various colors and styles.

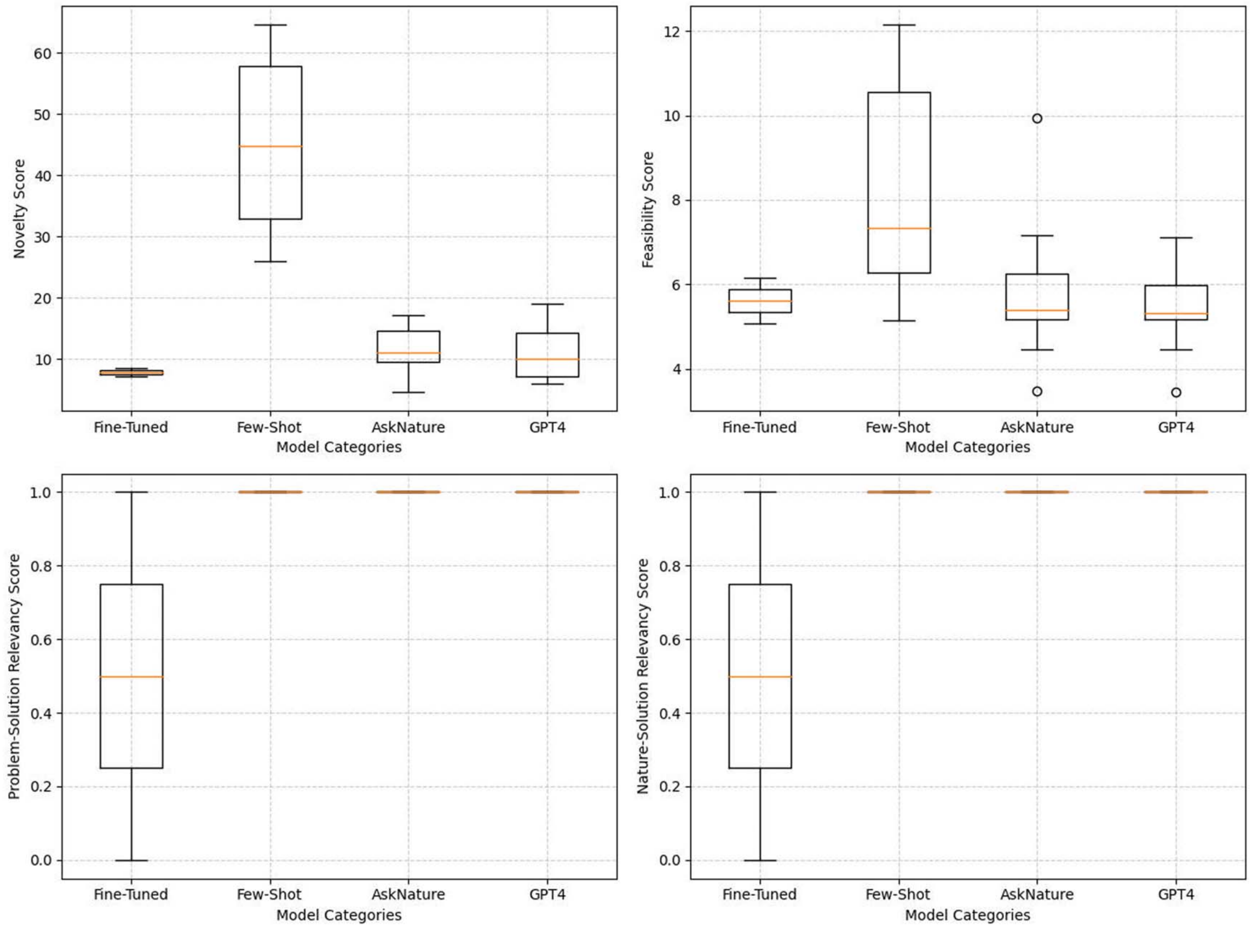


Fig. 3 Distributions of evaluation scores for four configurations—few-shot prompted designer, fine-tuned designer, GPT (baseline), and AskNature (benchmark)

fluctuations as the conversations progressed, indicating variability in the practicality of the concepts generated. The nature–solution relevancy consistently scored 1, yet problem–solution relevancy often remained at 0, suggesting a disconnect between the generated concepts and the predefined objectives.

Throughout the experiments, the flow of conversation and statistical results indicated that the generated design concepts resembled a "bag of ideas"—highly novel yet serving disparate purposes rather than converging toward a cohesive, innovative, and practical design concept. Despite the design objectives (benefits) being set and fixed from the beginning of each experiment, each iteration's output appeared to target different objectives, mostly irrelevant to the originally assigned objective. Consequently, each experiment was allowed to extend for significantly more iterations—three to four times the average of the main configuration—yet, as the designer continued, the relevance of the generated concepts to the defined problem notably declined.

One inherent challenge in this configuration was for the designer to essentially reverse-engineer the evaluation metrics' formula by deciphering the patterns in feedback responses to various design concepts. Despite these efforts, the results show that it failed to effectively meet this challenge, often producing concepts that did not align with the established objectives, thereby undermining the utility of the generated designs in addressing the specified design problem. In contrast, experiments conducted with the main configuration exhibit a range of advantageous outcomes that underscore its effectiveness: (1) they tend to converge very quickly toward a solution, (2) the designer's focus is consistently on enhancing the initial design to optimally address the defined problem, (3) there is no evidence of problem–solution irrelevancy, and (4) proposed refinements are complementary, ensuring that enhancements to the design do not interfere with one another but instead work in concert to improve the overall concept. Notably, although similar biological inspirations are sometimes employed in the design process of both configurations, they are modeled and utilized distinctly differently. For instance, in the qualitative configuration, the same biological inspiration is integrated in a manner that contributes to a functionality or design feature harmonizing with other elements, whereas, in the quantitative configuration, that same biological model often stands alone, not contributing to the cohesive function of the design.

These differences are primarily due to the more informative nature of the qualitative critic, which provides detailed explanations of flaws rather than mere numerical feedback, thus acting as a more effective guide for the designer. This approach encourages a deeper exploration into the nuances of the design concept on a step-by-step basis. Additionally, the qualitative critic is more comprehensive, as it is not constrained by specific metrics and is free to analyze any aspect of the concept, including those that cannot be easily quantified or automated. This flexibility allows for a broader evaluation of the design's various features and their interrelationships, leading to more nuanced and holistic design improvements.

3.6 Vision-Language Model Versus Large Language Model Agents. This section is dedicated to exploring the efficacy of incorporating images alongside the text in conveying design ideas within our GPT-based actor–critic framework. A critical observation from our experiments is the necessity of explicitly prompting the critic in a vision-language model setup to consider both textual and visual

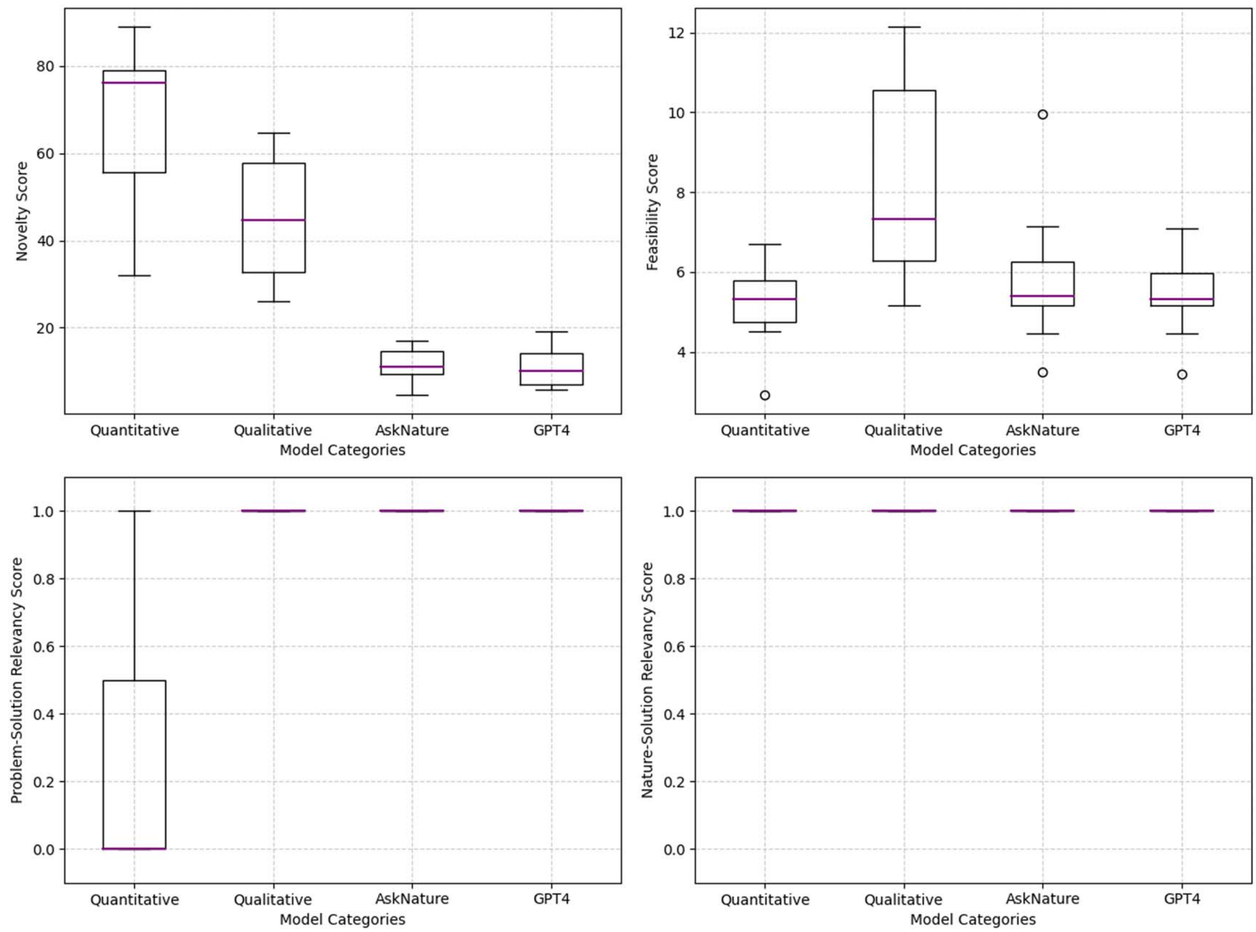


Fig. 4 Distributions of evaluation scores for four design approaches—quantitative critic, qualitative critic, GPT (baseline), and AskNature (benchmark)

information. Omitting this directive in the prompts often results in the critic disproportionately focusing on the visual aspects of the design. Such an imbalance can lead to critiques that are overly centered on aesthetic attributes at the expense of functional features, ultimately yielding feedback that is less informative and potentially misleading. This phenomenon underscores the importance of balanced input processing to ensure that both functional and aesthetic dimensions of a design are adequately assessed.

Figure 5 illustrates the results of this comparative analysis based on various metrics. The analysis reveals that while the impacts of the two configurations are generally similar, there are notable distinctions in their performance. Typically, the VLM configuration exhibits greater novelty in the generated designs but tends to lag in feasibility. A recurrent issue observed in some cases is the irrelevance of generated images; i.e., while the textual content of a design concept may present an intriguing idea, the accompanying image may be entirely unrelated, undermining the overall coherence of the design proposal.

The LLM configuration consistently generates responses that directly address feedback, often resulting in more feasible solutions. In contrast, the VLM configuration sometimes ventures into proposing novel ideas inspired by new biological sources in each iteration. These proposals, although creative, frequently lack relevance to the previous functionalities and overlook the complexities involved in simulating the biological source. Such scenarios typically do not converge toward a practical solution but oscillate among various potential solutions, indicating a struggle to achieve a definitive design outcome.

The authors conclude that incorporating visual presentation as an additional layer for conveying design ideas does provide some benefits, particularly in helping human designers visualize concepts. However, this addition does not significantly enhance the value of our actor–critic framework. This limitation likely stems from the current capabilities of GPT models, where the vision abilities are not as advanced as the linguistic capabilities. Presently, the vision functionalities of GPT models may even introduce confusion and misdirection within the design process. Therefore, we have decided to proceed with a configuration that utilizes a VLM actor and an LLM critic as the primary setup. This approach leverages the strengths of each model type while minimizing the drawbacks associated with their current technological limitations.

3.7 Long-Term Versus Short-Term Memory. This analysis investigates the impact of memory on the actor (designer agent) using an STM configuration, a simplified conceptual model that underlies similar approaches excluding the conversation history. In STMs, decisions are made based solely on the current state and previous action, making it an intriguing alternative to more complex LTM configurations, especially given the memory limitations of current LLMs like GPT. These models often struggle with memory constraints, leading to the termination of discussions when the length of generated tokens exceeds the limit. Our investigation primarily explores whether a simpler STM framework could effectively model the designer, particularly if the absence of memory from previous rounds does not detrimentally affect the output.

To further explore this issue, we explored two variations of modeling the designer as an STM agent: (1) a configuration where initial instructions—which set the context—are included in the state (s)

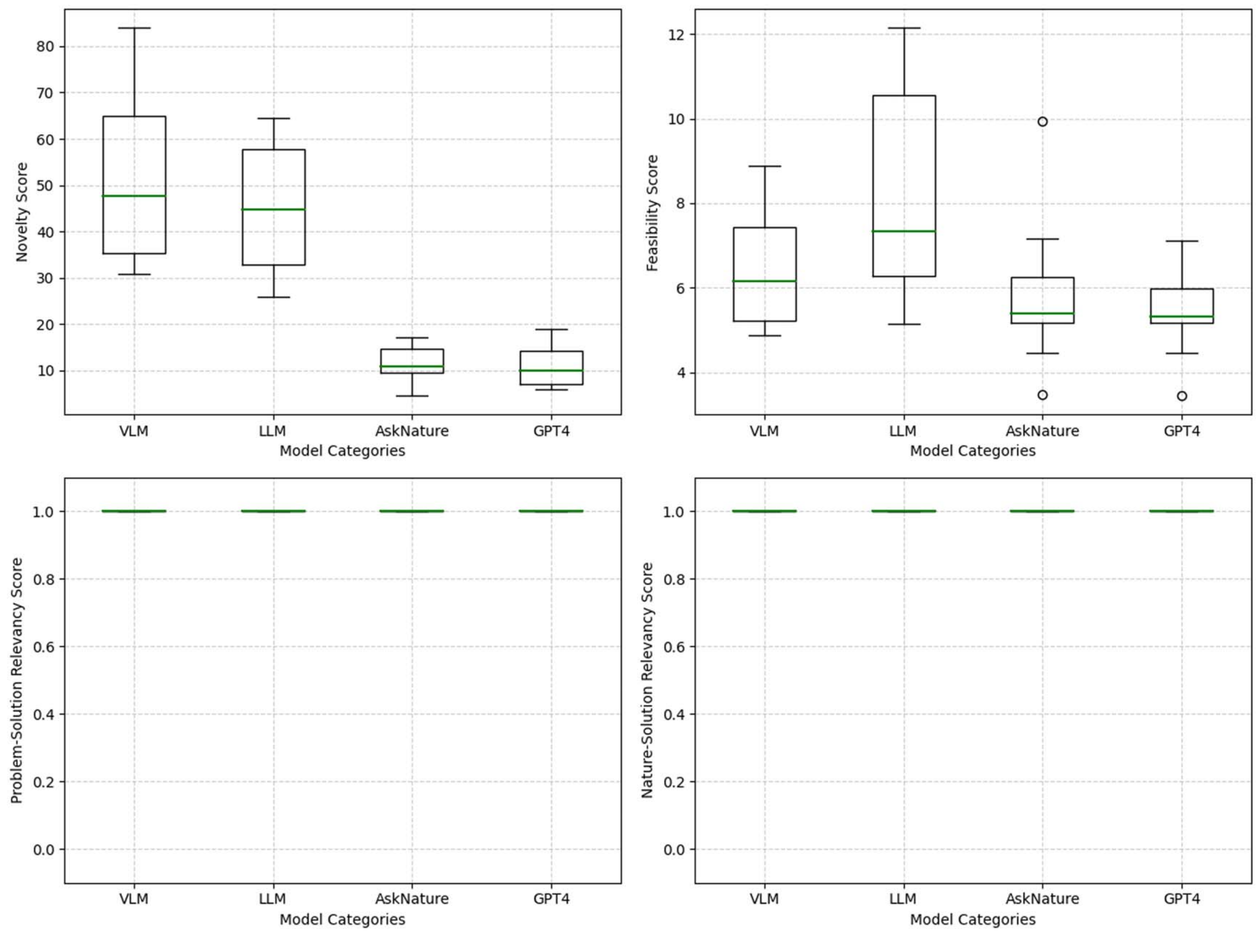


Fig. 5 Distributions of evaluation scores for four configurations—vision-language medium, language-only medium, GPT (baseline), and AskNature (benchmark)

and provided to the designer in all iterations, and (2) a configuration where initial instructions are included only in the first iteration (initial prompt). In the setting without continuous instructions, the designer consistently hallucinates, indicating a potential drawback of not including contextual reminders. Conversely, with instructions included in every iteration, the designer's behavior diverges in various ways from an LTM configuration: sometimes, it disregards all previous designs and starts anew, and at other times, it either does not draw inspiration from biological sources or fails to draw any inspiration whatsoever. However, in most cases, this configuration eventually succeeds in convincing the critic of the design's value.

From our experimental results, it became evident that STM, while computationally less demanding due to its simpler state management, is markedly less efficient than the LTM. This inefficiency stems primarily from the repetition of previously stated ideas in subsequent iterations, which are then critiqued in the same manner, resulting in a significantly longer convergence time—typically requiring two to three times more iterations to reach a solution. Without long-term memory, the designer in the STM model tends to focus on optimizing immediate feedback, often at the expense of a progressive refinement strategy, leading to less strategic and more repetitive design trajectories that compromise the quality of decision-making.

Moreover, the iterative process in the STM framework often introduces multiple sources of inspiration in an attempt to address limitations identified in the previous concept. However, some of these inspirations may be irrelevant, leading to confusion and uninformative feedback from the critic, filled with disparate information. This overload of feedback can overwhelm the designer, who may respond by indiscriminately adding new functionalities inspired by various sources, perpetuating an ineffective cycle of design revisions. In contrast, LTM demonstrates a smoother learning curve, with designs gradually converging toward more optimal solutions and exhibiting less volatility in the learning process compared to the abrupt shifts seen in STM. Additionally, although STM tends to generate highly novel ideas—presumably due to the inclusion of multiple functionalities in each iteration—it does so within a constrained scope of adaptability and innovation. Each iteration, being a response solely to the immediate past, lacks the historical context that could enrich the design process. This approach not only limits the depth of exploration but also results in outputs that are less feasible or sometimes even irrelevant to the design problem, as indicated in Fig. 6.

We conclude that while STM offers superior memory efficiency, it falls short of integrating critical nuances of feedback and decreases overall conversation and design learning efficiency. Consequently, we opt to proceed with a configuration augmented with long-term memory, which better supports complex decision-making tasks and fosters more effective integration of feedback into the design process.

3.8 Variety Evaluation. This section presents the results of the variety analysis conducted on the categorical outcomes of different configurations, contrasting them with the benchmarks. The analysis offers insights into the degree of variation within groups of design concepts. The distributions of pairwise similarities are illustrated in Fig. 7. The results indicate no significant difference among most concept groups, except for the fine-tuned actor and STM

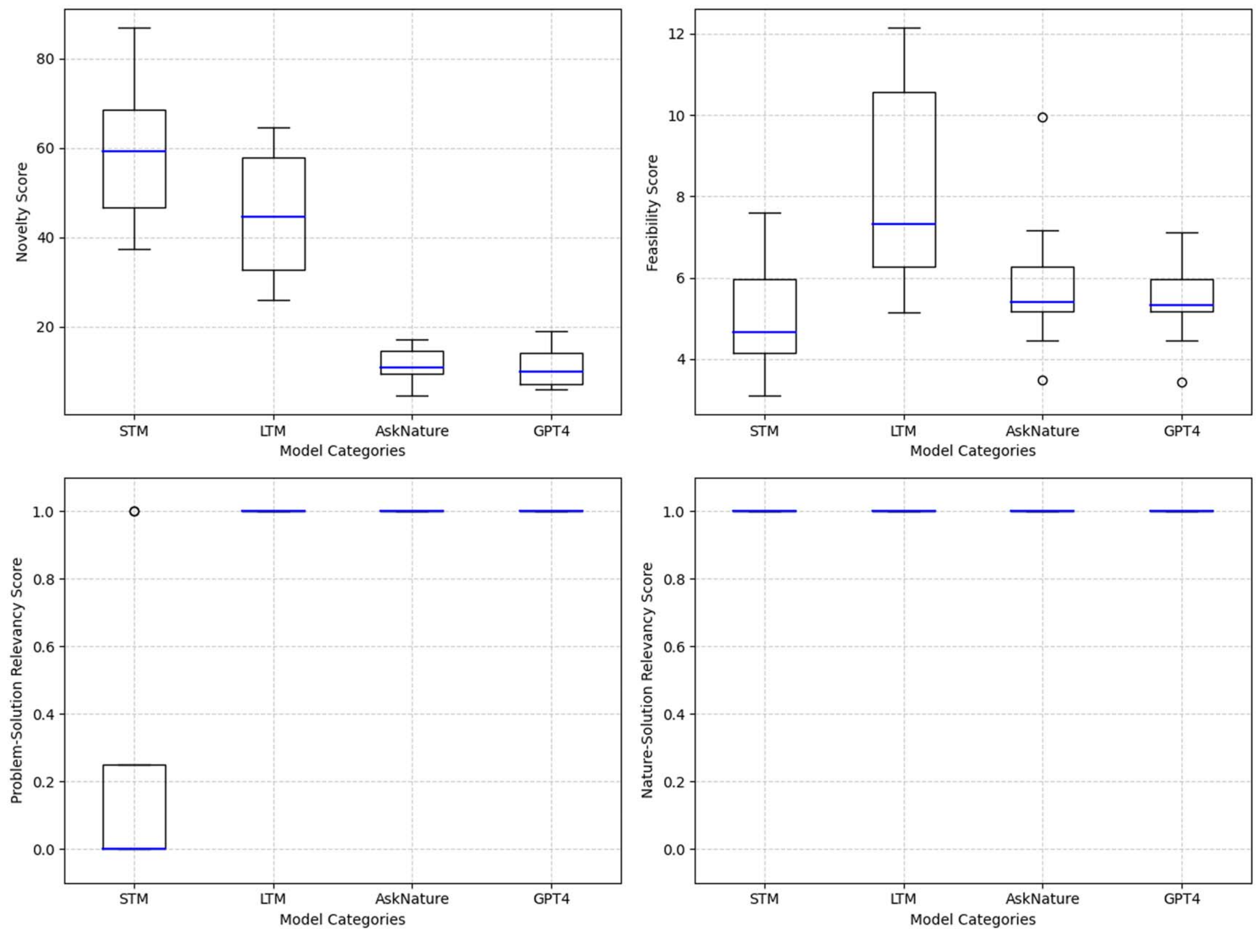


Fig. 6 Distributions of evaluation scores for four configurations—short-term memory, long-term memory, GPT (baseline), and AskNature (benchmark)

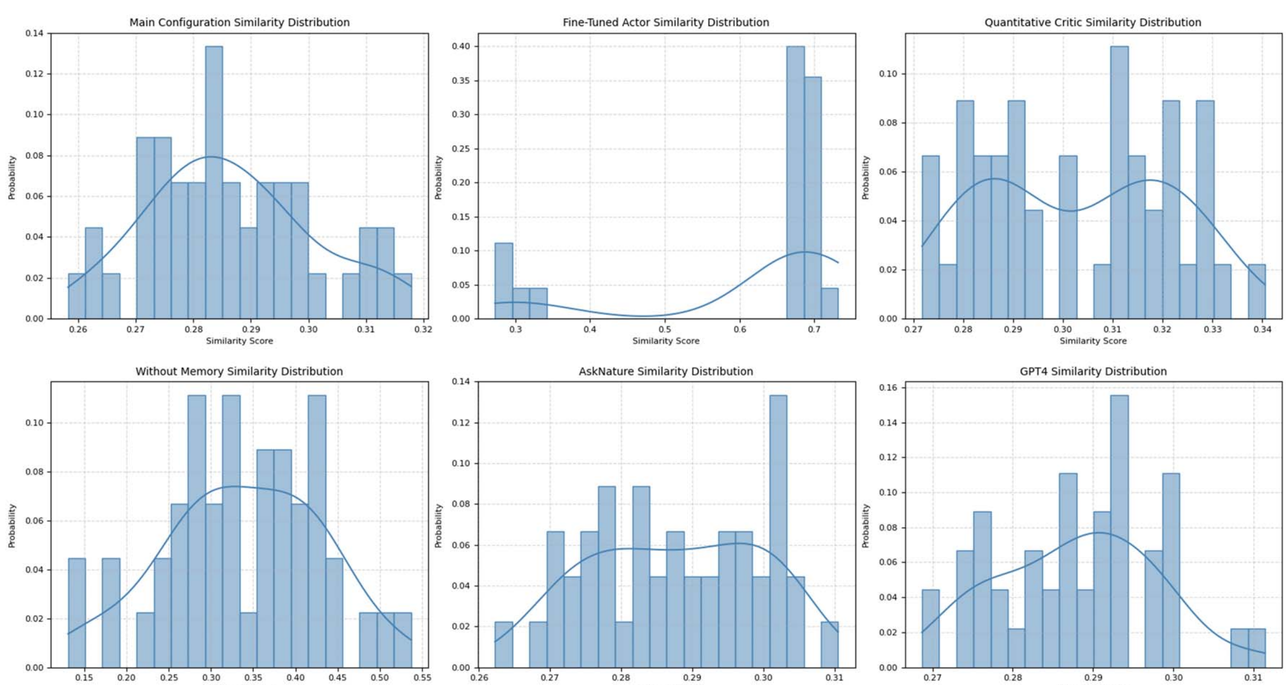


Fig. 7 Variety evaluation results for various configurations and the benchmarks

Table 4 Examples of different types of hallucination by the fine-tuned designer

Types of hallucination	Example
Partial	The noncuttable material, called Proteus, is a lightweight material made of ceramic spheres within a flexible, cellular aluminum structure. The design was inspired by the tough cellular skin of grapefruit and fracture-resistant mollusk shells. When a cutting tool, like an angle grinder, tries to cut into
Repeated (same response in more than ten rounds)	Pangolin packaging is a biodegradable alternative to packing peanuts seen in protective mail packaging. The paper-based design curls into itself and interlocks to create a protective form, much like the pangolin curls its body around to protect itself from predators.
Nonsensical	The following organizations and research groups have used nature-inspired design to develop the aforementioned technologies. Durham University - Proteus, University of Cambridge - Plastic Film, AutoDesk Research - Generative Design Software, Wageningen University - Probabilistic Fastener, Bilkent University - Self-Repairing Robots, MIT - Geometric Pavilion Structure

configurations. The main configuration, quantitative critic, AskNature, and GPT-4 exhibit relatively low mean similarity scores (0.29, 0.30, 0.29, and 0.29, respectively) with small standard deviations (0.02, 0.02, 0.01, and 0.01, respectively), suggesting that the concepts generated within these groups are both diverse and consistent. The narrow range between minimum and maximum similarity further implies that the model consistently produces varied concepts across these configurations.

In contrast, the fine-tuned actor configuration shows a higher mean similarity (0.61), indicating that the concepts within this group are more similar to each other compared to those generated by the main model. The higher standard deviation (0.16) and the broader range between minimum and maximum similarity suggest that, while there is significant variation among some concepts, others are highly similar. As shown in Fig. 7, the distribution appears polarized; the samples in the higher similarity bins are likely the result of designer hallucinations, potentially due to overfitting (Table 4). The STM configuration displays a higher standard deviation (0.09) and a broader range between minimum and maximum similarity than other configurations, indicating a mix of both similar and diverse concepts. This may be attributed to the presence of several irrelevant components or functionalities within each concept in this group, some of which are shared across concepts. However, within this group, there are also concepts with surprisingly distinct and unique components. It is worth noting that the similarity in statistical measures across the main configuration, quantitative critic, AskNature, and GPT-4 could be attributed to the shared design problems across all groups.

4 Discussion and Conclusions

In our framework, the typical flow of conversation begins with the presentation of a novel, ambitious idea inspired by a biological system. This initial proposal sets the stage for a collaborative iterative process between the designer (i.e., the actor) and the critic. Subsequent rounds focus on identifying and dissecting the limiting aspects of the proposed idea, addressing the inherent challenges associated with adapting the biological system for practical applications. The designer and critic work in tandem to systematically deconstruct these complexities, exploring viable solutions that can feasibly translate the biological inspiration into a real-world context. This dynamic interaction ensures that each element of the proposed design is thoroughly scrutinized and optimized.

4.1 An Illustrative Example: The Tidal Energy Converter.

To illustrate, a conversation between the designer and the critic in the context of energy generation with sustainability as the design objective resulted in the following concept:

> *Innovation Details:* The tidal energy converter draws inspiration from various natural forms to enhance efficiency and durability. Mimicking the pulsating movements of a jellyfish's bell, the design captures tidal energy through contraction and expansion. It incorporates logarithmic spirals inspired by nautilus shells, optimizing water flow for increased energy capture. The structure features a honeycomb coral pattern for robustness, dispersing stress and enhancing resilience. Modular units resembling mangrove roots stabilize the device and reduce water flow erosion, crucial for the placement in marine environments. Additionally, fixed, angled blades, modeled after humpback whale fins, employ hydrodynamic principles to minimize drag and maximize lift, improving fluid dynamics and energy efficiency.

The conversation pertaining to this experiment shows a progressive refinement of a novel energy generation concept. Each round of the discussion highlights the evolution of the design to address specific challenges, while integrating bioinspired elements to enhance the system's efficiency and durability. The following summarizes our observations from this experiment, but also applies to all experiments using this configuration:

- *Innovative Approach:* Each design iteration integrates different natural inspirations to improve energy capture and system durability. Starting with jellyfish locomotion, the designer progressively incorporates elements inspired by wind turbines, nautilus shells, honeycomb coral, mangrove roots, seaweed and kelp movements, and finally, humpback whale fin structures. These bioinspired designs aim to harness and optimize the natural efficiency and resilience of these biological systems in a mechanical device.
- *Problem-Solving Focus:* With each round, the designer responds to the critic's concerns about scalability, complexity, cost, and environmental sustainability. This iterative problem-solving approach demonstrates a deepening understanding of the practical challenges involved in commercial energy generation and a commitment to refining the concept to make it feasible.
- *Feasibility and Challenges:* The critic consistently provides feedback on the concept's most limiting flaw regarding the feasibility of the designs based on current research and practical considerations. They highlight potential flaws such as scalability, maintenance, durability in harsh marine environments, and the complexity of integrating various mechanical and dynamic components. Each critique pushes the designer to innovate further and simplify the system to better meet these challenges.
- *Real-World Application Potential:* The conversation moves toward a design that could be potentially implemented in real-world conditions by simplifying mechanisms, using durable materials, and reducing the need for maintenance. The final designs focus on static, durable structures that can withstand marine environments while maximizing energy efficiency, showing a path from conceptual to potentially practical applications.

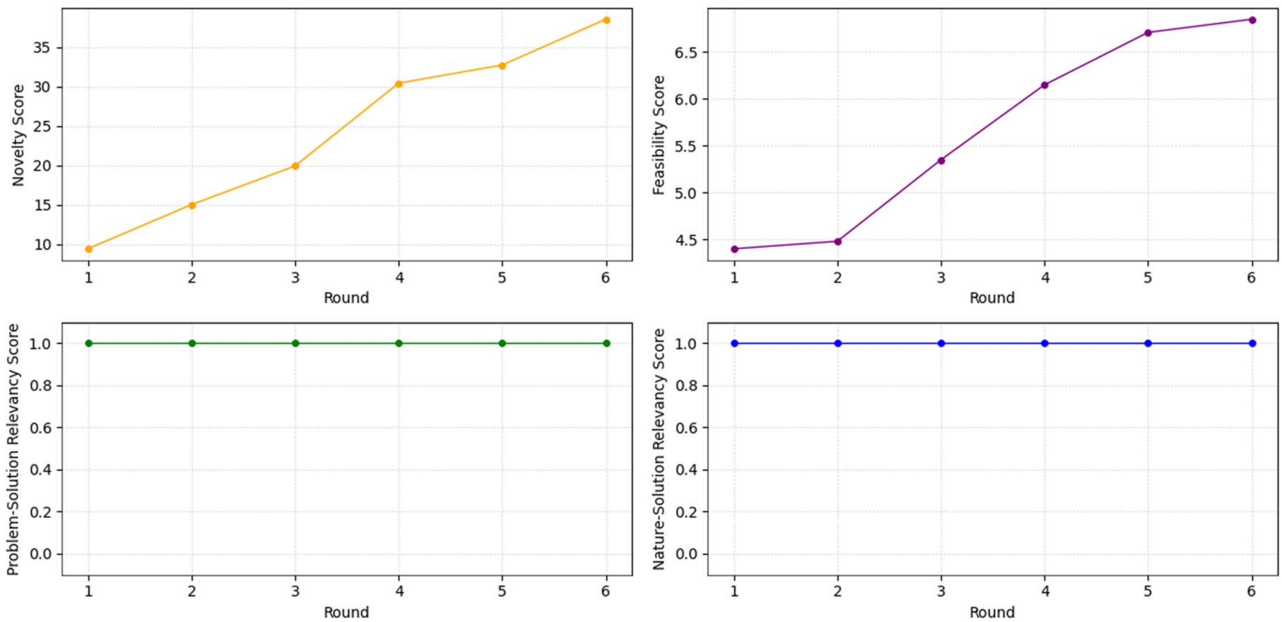


Fig. 8 Progression of evaluation scores over multiple rounds of conversation for the tidal energy converter concept

- *Sustainability Emphasis:* Throughout the conversation, there is a strong emphasis on sustainability—not only in terms of energy generation but also in the use of materials and the long-term operability of the system in marine conditions. This aligns with the initial design objective to create a sustainable energy solution.

The quantitative evaluation scores presented in Fig. 8 provide significant insights into the flow and progression trend of the framework's performance across multiple rounds. Notably, the novelty scores show a consistent upward trajectory, starting from 9.27 and culminating at 38.43. This steady increase reflects the framework's capability to continuously introduce innovative features and functionalities, significantly enhancing the conceptual novelty with each iteration. In parallel, the feasibility scores also exhibit a progressive increase from 4.39 to 6.89. This trend is particularly noteworthy because, contrary to what might be expected, the introduction of new features does not compromise the concept's feasibility. Instead, it enhances it. This improvement suggests that the additions are judiciously selected to align with current research and industry developments, effectively addressing any limitations within the concept while ensuring that the design remains practical and implementable. Moreover, the scores for problem–solution and nature–solution consistently hold at 1.0 throughout the rounds, indicating a robust ability of the framework to maintain the relevance of the design to the initial problem and design objectives. This consistency underscores the framework's effectiveness in several key areas. First, the problem–solution relevancy is maintained by ensuring that all new features and improvements work in harmony with each other and are aligned with the core design objectives. Second, the nature–solution relevancy is consistently upheld by selecting inspiration sources that are not only innovative but also aligned with environmental constraints and the broader context of the problem. The breakdown of how different features interrelate and contribute to the concept's overall functionality is as follows:

- The core mechanism of the device mimics the contraction and expansion of a jellyfish's bell. This movement is ideal for capturing tidal energy as it naturally aligns with the rhythmic push and pull of ocean tides, offering a continuous and reliable energy capture method.
- The logarithmic spirals are designed to optimize water flow around the device. This feature can increase the efficiency of energy capture by smoothing the flow of water into and out of the device, minimizing turbulence and resistance that could otherwise reduce efficiency.
- The use of a honeycomb pattern for the structural design enhances the resilience of the converter by effectively dispersing stress across the structure. This not only improves durability but also aids in withstanding harsh marine conditions.
- Modular units provide stabilization and reduce erosion caused by the water flow. This is crucial for maintaining the integrity and positioning of the device in its marine environment, ensuring long-term operation and effectiveness.
- The incorporation of fixed, angled blades uses hydrodynamic principles to minimize drag and maximize lift. This improves the fluid dynamics around the blades, enhancing the overall efficiency of the energy conversion process and contributing to the stability of the entire structure.

The different parts and functionalities of this concept are not only consistent with one another but also synergistic. Each component is designed to perform a specific function that supports the overall goal of efficient and sustainable energy generation. The selection of biomimetic models shows that the device is well adapted to its environment. This concept is a meticulous integration of multiple biomimetic principles, each contributing to the device's performance in a complementary way.

To assess the impact of extending the conversation beyond the critic's satisfaction threshold, we conducted additional experiments where the interaction continued after this point. The results indicate that, in the absence of specific limitations raised by the critic or further guidance, the actor initially proposes a new feature for the latest version of the concept, which is contextually relevant to the ongoing experiment. The critic continues to evaluate and provide feedback in a manner consistent with earlier rounds, and the actor responds by addressing these critiques in the subsequent iterations. This leads to a repetition of the interaction cycle observed prior to reaching the satisfaction threshold, focused now on the newly proposed feature. Figure 9 illustrates the trend of evaluation scores following the critic's satisfaction round.

Given that this study was scoped such that the actor's primary role is to respond to the critic's concerns, the actor, after reaching the satisfaction threshold, generates new ideas autonomously, as there are no additional critiques to address. However, exploring the implications of continuing the interaction beyond this point requires a more in-depth analysis and careful design of the conversational

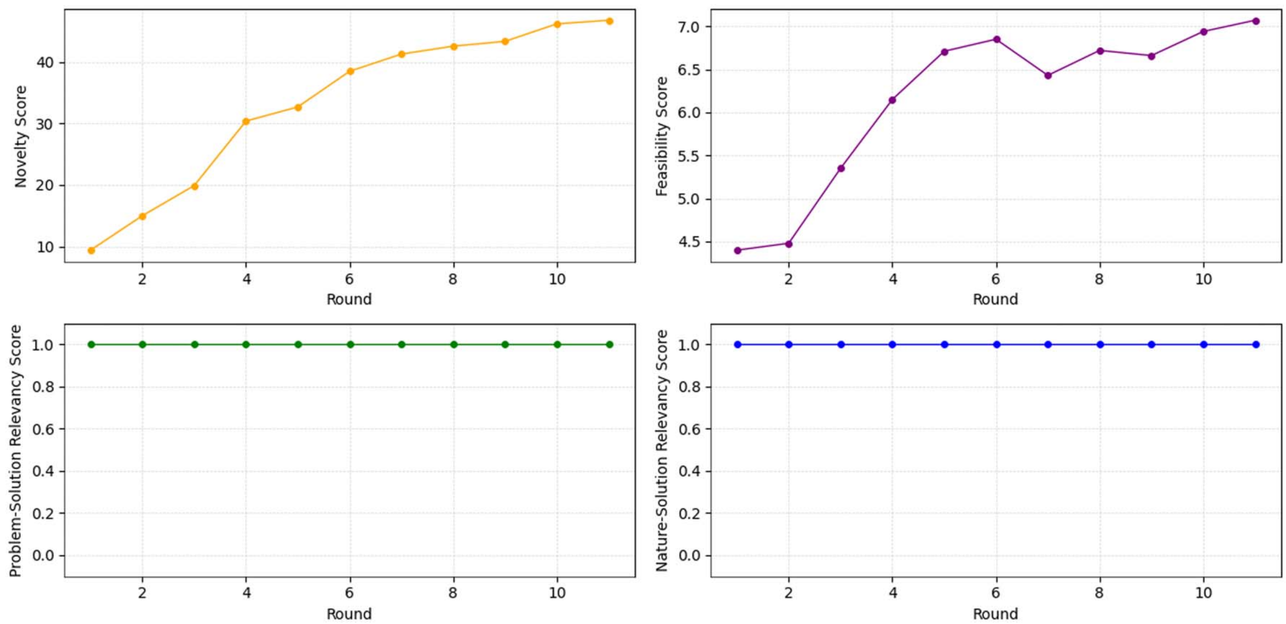


Fig. 9 An example of the trend of evaluation scores after surpassing the critic's satisfaction threshold

experiment. This would allow for a more rigorous examination of how the actor and critic's dynamic evolves and what the resulting design concepts might look like in such extended interactions.

The decision to focus on detecting only the most limiting flaw in each iteration, rather than identifying all flaws simultaneously, stems from our observations of agent performance under different prompting conditions. When the critic is prompted to detect all flaws at once, we noted that both the designer and critic agents became overwhelmed. This often resulted in the generation of generic responses that lacked specificity and were nearly identical across different design concepts, thereby diminishing the value of the feedback. In contrast, by concentrating on the most limiting flaw during each round of analysis, the responses exhibited a more thorough examination of the specific concept at hand. This focused approach allows the critic to conduct an actual analysis from various angles. Such targeted feedback provides case-specific recommendations and actionable insights, significantly enhancing the utility and relevance of the interactions between the designer and critic and facilitating more effective iterative improvements to the design.

Additionally, our decision for the designer to generate outputs using both vision and language, without receiving images as input, and for the critic to process inputs in both vision and language without producing any images as outputs, reflects established practices in conceptual design and evaluation. Historically, the conceptual design has been effectively communicated through a blend of descriptive text and illustrative sketches, enabling designers to convey the essence and details of their ideas in an intuitive and informative manner. Similarly, evaluating or critiquing these designs typically involves textual feedback, centered on analyzing and discussing the presented concepts without the need to interpret additional visual outputs from the critic. However, after a thorough analysis of the results from our initial setup, we opted to adjust our framework to feature a VLM as the designer and a LLM as the critic. This decision was informed by the observation that while the VLM designer enhances the representation of design concepts by providing both textual descriptions and visual sketches, the LLM critic focuses solely on functional features.

4.2 Implications and Contributions. This framework aims to address several prominent challenges associated with current state-of-the-art AI-driven models for design concept generation:

- *Generalization and Creativity:* A primary challenge for AI models, including LLMs and VLMs, is their ability to generalize beyond their training data and exhibit genuine creativity. This often results in outputs that are innovative but may not fully capture the depth of human creativity or conceptualize entirely new paradigms. Our final configuration, however, enhances these models' capabilities by iteratively refining initial ideas until no major concerns by the critic remain, as demonstrated by the novelty results, which serve as an indicator of creativity. This approach effectively addresses critiques and revises solutions from previous iterations when necessary. This actor–critic approach facilitates the selection of rational solutions, thus bridging the gap between AI-generated concepts and human-like creativity.
- *Contextual Understanding and Interpretation:* AI models often struggle with achieving the deep contextual understanding necessary to fully appreciate the complexities of design briefs or user needs. This can lead to technically sound but contextually misaligned concepts. Our framework, demonstrated through quantitative problem–solution relevancy analyses, maintains and refines the relevance of concepts throughout the design process, ensuring alignment with the specified objectives until the conversation's conclusion.
- *Dependency on Data Quality and Bias:* The effectiveness of data-driven models significantly depends on the quality and diversity of their training data. To mitigate this, our framework leverages GPT-4, which, with its access to a vast and diverse range of data and continuous updates from the Internet, offers a more balanced and inclusive design perspective. By guiding this extensive knowledge base through iterative refinement—which mimic systematic human thinking and learning in design—we enhance the model's capacity to produce more equitable and innovative outputs.
- *Integration Into Existing Workflows:* Integrating AI-enhanced tools into established design workflows poses significant challenges, often requiring extensive adaptation and specialized knowledge, which can hinder widespread adoption. To address this, our framework employs a combination of text and image outputs generated by the GPT-4 model, facilitating a smoother integration into existing workflows. This approach leverages familiar mediums (text and visuals) that designers are accustomed to, thereby reducing the barriers to adoption and enhancing the usability of AI tools in traditional design settings.

Overall, the framework we propose offers substantial benefits that make it a versatile and powerful tool across various domains of application. First, it is founded on robust theoretical principles yet remains highly adaptable; it is designed to be accessible to a wide range of users through available APIs such as those provided by OpenAI, facilitating ease of use regardless of users' technical expertise. Second, the framework is characterized by its generalizability. Our experimental validations cover a wide array of industries, demonstrating its applicability across sectors as diverse as 3D printing, aerospace, agriculture, architecture, automotive, biotechnology, drug discovery, energy generation, robotics, transportation, and medical device manufacturing, among others. This breadth of application underscores the framework's capacity to handle a variety of design challenges and requirements, making it a valuable tool for innovation in virtually any field. Third, the framework is designed to perform effectively with minimal data requirements. Unlike traditional models that may require vast datasets to function optimally, our framework can achieve significant performance with just a few examples, leveraging advanced machine learning techniques to extrapolate and innovate from limited inputs. This feature is particularly advantageous in scenarios where data availability is limited or where data collection is challenging, expensive, or time consuming.

Our framework holds the potential to significantly impact design practice and education by enhancing creativity through human–AI collaboration. By allowing designers to collaborate with multiple generative AI agents, our framework facilitates the exploration of a broader range of design possibilities, leading to the generation of more innovative concepts. This approach enables designers to ideate in a dynamic and interactive environment, where AI agents can offer diverse perspectives, suggest novel ideas, and provide iterative feedback. As a result, designers are empowered to push the boundaries of conventional design thinking and explore creative avenues that may have been previously overlooked. In engineering design education, this framework can serve as a powerful tool for teaching design students the importance of iterative ideation, critical evaluation, and interdisciplinary collaboration. By integrating AI into the design process, students can learn how to harness the strengths of generative models while maintaining their unique creative input, ultimately fostering a new generation of designers who are adept at navigating the intersection of technology and creativity.

Agent-based modeling is a computational method used to simulate the interactions of autonomous agents with a view to assessing their effects on the system as a whole. It involves creating models where each agent is assigned specific behaviors and characteristics and can make decisions independently. This technique is used across various fields including network theory [37], economics [38], and biology [39], to name a few. Drawing a high-level conceptual analogy from agent-based modeling, our framework can be likened to the structure of a Generative Adversarial Network (GAN). In this analogy, the generator is represented by a VLM specialized as a designer, while the discriminator is analogous to a LLM specialized as a design critic. Conceptually, the interplay between our agents, akin to the backpropagation of error in neural networks, involves either multiple regression models or descriptive feedback that acts as a vectorized mapping into a space decodable by the generator.

However, there are distinct differences in how feedback functions in this framework compared to backpropagation of error in traditional neural networks. First, the loss function in neural networks quantitatively measures the divergence of generated outputs from real data, facilitating direct backpropagation of error gradients. In contrast, the feedback in our framework does not constitute a gradient that can be directly backpropagated. Furthermore, unlike the discriminator's output in a conventional GAN, which is compared against a predefined dataset to assess authenticity, the feedback in our setup is not benchmarked against such a dataset but is evaluated based on its relevance and constructive value to the design process. Additionally, a critical distinction in our framework from GANs is that learning does not occur during a training phase but through an ongoing iterative process of conversation and refinement. Given these observations, an intriguing direction for further exploration could involve implementing an actual GAN model where the generator and discriminator are embodied by LLM/VLMs. This model would integrate the conversational refinements typical of our current framework into the training process, potentially enhancing the dynamism and efficacy of the learning mechanism in design generation contexts.

4.3 Limitations and Future Directions. While we have delineated several strengths of our framework, it is essential to also consider its limitations, which highlight areas for potential enhancement and further development. One notable limitation of our current framework is that during both initial ideation and subsequent refinement phases, the designer generates only a single solution. This approach inherently constrains the exploration to a specific trajectory within the design space, effectively limiting the breadth of exploration to a singular vector rather than embracing the full spectrum of potential solutions. To enhance the capability of our framework and enable a more comprehensive exploration of the design space, future work could incorporate strategies for horizontal search. This would complement the existing vertical search mechanism, which optimizes solutions iteratively within a defined path, by broadening the scope to consider multiple design pathways simultaneously. Such an expansion would potentially allow for a richer, more diverse generation of solutions, thereby maximizing the framework's exploratory power.

Additionally, another limitation arises from the specific focus of our analysis on biologically inspired designs. While this focus has provided valuable insights, it also restricts the generalizability of our findings across other design methodologies. Future research could address this limitation by exploring a wider array of design approaches, such as various forms of design by analogy. Expanding the scope to include different methodologies would not only diversify the potential applications of the framework but also enhance its utility and relevance in a broader context of design challenges. It is important to recognize that when employing design examples other than those provided by AskNature, additional prompting techniques may be required. AskNature inherently includes reasoning and rationale for each sample, allowing for the omission of certain prompting techniques. However, in the absence of such augmentations, prompts must be carefully customized to include all relevant descriptions, thereby minimizing the risk of generating overly general or potentially misleading responses.

Moreover, the efficacy of the feedback mechanism within our framework, especially the manner in which the critic evaluates and provides feedback, is pivotal and can become a limitation if not meticulously calibrated. Inaccurate or irrelevant feedback from the critic could misdirect the design process, inadvertently reinforcing suboptimal design pathways that diverge from the intended objectives. To mitigate this issue, future work should focus on developing mechanisms to validate or certify the critiques before they are relayed to the designer. This could involve implementing an inner feedback loop within the critic's framework, where feedback undergoes several rounds of refinement and validation to ensure its accuracy and relevance. This certification process could incorporate advanced algorithms capable of contrasting the validity and helpfulness of a set of feedback based on historical data and/or predefined criteria. Furthermore, integrating multiple critics, each tasked with evaluating specific aspects of the generated concepts, could potentially enhance the overall performance of the framework. This would necessitate a carefully structured interaction between the agents to ensure effective collaboration. For instance, incorporating LEGO [11] could improve the precision of causal explanations, offering deeper insights into the reasoning behind the proposed designs.

Catastrophic forgetting represents a notable challenge in the realm of deep neural networks, where models tend to lose previously acquired knowledge upon training on new tasks. To mitigate catastrophic forgetting, various strategies have been explored, including the protection of crucial weights within the model, the adoption of dual-memory systems, and the implementation of regularization techniques. Despite these advancements, catastrophic forgetting continues to pose a substantial hurdle, particularly in complex, real-world scenarios such as design. Specifically, when these models undergo fine-tuning, there is often a discernible drop in their performance on tasks they had previously mastered. This issue underscores a vital area for future research in our study, focusing on the application of innovative fine-tuning techniques such as applying advanced parameter-efficient fine-tuning techniques, leveraging mitigation strategies (like experience replay), or utilizing a smaller LLM, aimed at specializing a VLM/LLM to function effectively as a design expert without compromising its existing capabilities. Another approach could incorporate heuristic-based search techniques to assess the impact of template sections of the training data (i.e., shared parts of the samples in JSONL), optimizing the fine-tuning performance.

In relation to the evaluation metrics, the design literature presents a range of alternative definitions and models that warrant consideration. It would be valuable to apply and juxtapose these with our proposed metrics to gain deeper insights. For instance, Bayesian surprise serves as a measure of novelty, playing a crucial role in the mathematical modeling of the fundamental limits of creativity [40]. Furthermore, exploring the impact of varying the component weights in our novelty and feasibility evaluation models may yield valuable insights and enhance the optimization of models tailored for LLM-based design concept generation scenarios. To further enhance the accuracy and reliability of our feasibility evaluation algorithm, several potential improvements could be implemented. One promising approach is to label each similar concept identified through the search process with an indication of the extent to which the project has been developed. These labels could then be used to weigh the corresponding cosine similarity scores, giving more influence to concepts that are more fully realized and thus likely to provide more relevant insights. Additionally, experimenting with different prompting techniques could refine the algorithm's ability to generate more nuanced and contextually appropriate evaluations. For instance, prompts could be tailored to emphasize specific aspects of feasibility, such as technical viability or market potential, thereby guiding the model to consider a wider range of factors in its assessment. The evaluation algorithm could also be expanded to include a synthesis step, where the algorithm not only identifies similar concepts but also evaluates the compatibility of combining different components. This could involve checking the technical feasibility of integrating various elements (e.g., materials, technologies, or processes) that are not immediately obvious but could work together to realize the new idea.

Acknowledgment

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Funding Data

- National Science Foundation under the Engineering Design and Systems Engineering (EDSE) (Grant No. 2050052).

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The data and information that support the findings of this article are freely available at: https://github.com/sail-gt.

References

[1] Liu, Y.-C., Chakrabarti, A., and Bligh, T., 2003, "Towards an 'ideal'approach for Concept Generation," Des. Stud., **24**(4), pp. 341–355.

[2] Leuenberger, A., Birner, E., Lumpe, T. S., and Stanković, T., 2024, "Computational Design of 2D Lattice Structures Based on Crystallographic Symmetries," ASME J. Mech. Des., **146**(7), p. 071703.

[3] Kwon, E., and Goucher-Lamber, K., 2024, "Comparing and Evaluating Human and Computationally Derived Representations of Non-Semantic Design Information," ASME J. Mech. Des., **146**(3), p. 031401.

[4] Li, X., Xie, C., and Sha, Z., 2023, "Design Representation for Performance Evaluation of 3D Shapes in Structure-Aware Generative Design," J. Des. Sci., **9**(27).

[5] Handawi, K. A., Brahma, A., Wynn, D. C., Kokkolaras, M., and Isaksson, O., 2024, "Design Space Exploration and Evaluation Using Margin-Based Trade-Offs," ASME J. Mech. Des., **146**(6), p. 061701.

[6] Zhu, Q., Zhang, X., and Luo, J., 2023, "Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers," ASME J. Mech. Des., **145**(4), p. 041409.

[7] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., and Chen, W., 2023, "Critic: Large language models can self-correct with tool-interactive critiquing." arXiv preprint arXiv:2305.11738.

[8] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X., 2024, "Large Language Model based Multi-Agents: A Survey of Progress and Challenges," arXiv.

[9] Zhang, B., Mao, H., Ruan, J., Wen, Y., Li, Y., Zhang, S., Xu, Z., Li, D., Li, Z., Zhao, R., Li, L., and Fan, G., 2024, "Controlling Large Language Model-Based Agents for Large-Scale Decision-Making: An Actor-Critic Approach," ICLR 2024 Workshop on Large Language Model (LLM) Agents, Vienna, Austria, May 11.

[10] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S., 2024, "Reflexion: Language Agents With Verbal Reinforcement Learning," Adv. NeAdvances in Neural Information Processing Systemsural Inf. Process. Syst., **36**, pp. 8634–8652.

[11] He, Z., Cao, P., Chen, Y., Liu, K., Li, R., Sun, M., and Zhao, J., 2023, LEGO: A Multi-agent Collaborative Framework with Role-playing and Iterative Feedback for Causality Explanation Generation, Volume: Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, Singapore, pp. 9142–9163.

[12] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I., 2024, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," Forty-first International Conference on Machine Learning, Vienna, Austria, July 21.

[13] Gupta, N., Kirtania, S., Gupta, P., Mehra, K. K., Gulwani, S., Iyer, A., Parthasarathy, S., Radhakrishna, A., Rajamani, S. K., and Soares, G., 2024, "STACKFEED: Structured Textual Actor-Critic Knowledge Base Editing with FeedBack," arXiv.

[14] Tan, X., Shi, S., Qiu, X., Qu, C., Qi, Z., Xu, Y., and Qi, Y., 2023, "Self-criticism: Aligning Large Language Models With Their Understanding of Helpfulness, Honesty, and Harmlessness," Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 650–662.

[15] Chandrasekaran, B., 1990, "Design Problem Solving: A Task Analysis," AI Mag., **11**(4), pp. 59–59.

[16] Yuan, C., Marion, T., and Moghaddam, M., 2023, "Dde-gan: Integrating a Data-Driven Design Evaluator Into Generative Adversarial Networks for Desirable and Diverse Concept Generation," ASME J. Mech. Des., **145**(4), p. 041407.

[17] Yuan, C., Marion, T., and Moghaddam, M., 2022, "Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model," ASME J. Mech. Des., **144**(2), p. 021403.

[18] Ghasemi, P., Yuan, C., Marion, T., and Moghaddam, M., 2023, "Are Generative Adversarial Networks Capable of Generating Novel and Diverse Design Concepts? An Experimental Analysis of Performance," Proceedings of the Design Society., **3**, pp. 633–644.

[19] Cheng, S.-H., 2023, "Impact of Generative Artificial Intelligence on Footwear Design Concept and Ideation," Engineering Proceedings, **55**(1), p. 32.

[20] Ghasemi, P., Yuan, C., Marion, T., and Moghaddam, M., 2024, "DCG-GAN: Design Concept Generation With Generative Adversarial Networks," ASME J. Des. Sci, **10**(14).

[21] Raina, A., Puentes, L., Cagan, J., and McComb, C., 2021, "Goal-Directed Design Agents: Integrating Visual Imitation With One-Step Lookahead Optimization for Generative Design," ASME J. Mech. Des., **143**(12), p. 124501.

[22] Su, H., Song, B., and Ahmed, F., 2023, "Multi-Modal Machine Learning for Vehicle Rating Predictions Using Image, Text, and Parametric Data," IDETC-CIE, Boston, MA, Aug. 20.

[23] Camburn, B., He, Y., Raviselvam, S., Luo, J., and Wood, K., 2020, "Machine Learning-Based Design Concept Evaluation," ASME J. Mech. Des, **144**(3), p. 031113.

[24] Lee, M., Park, Y., Jo, H., Kim, K., Lee, S., and Lee, I., 2022, "Deep Generative Tread Pattern Design Framework for Efficient Conceptual Design," ASME J. Mech. Des, **144**(7), p. 071703.

[25] Nie, Z., Lin, T., Jiang, H., and Kara, L. B., 2021, "Topologygan: Topology Optimization Using Generative Adversarial Networks Based on Physical Fields Over the Initial Domain," ASME J. Mech. Des., **143**(3), p. 031715.

[26] Hagos, D. H., Battle, R., and Rawat, D. B., 2024, "Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives," IEEE Trans. Artif. Intell., **5**(12), pp. 5873–5893.

[27] Kucharavy, A., 2024, "Fundamental Limitations of Generative LLMs," *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, Springer Nature, Cham, Switzerland, pp. 1–247.

[28] Chen, L., Zuo, H., Cai, Z., Yin, Y., Zhang, Y., Sun, L., Childs, P., and Wang, B., 2024, "Toward Controllable Generative Design: A Conceptual Design Generation Approach Leveraging the Function–Behavior–Structure Ontology and Large Language Models," ASME J. Mech. Des., **146**(12), p. 121401.

[29] Regenwetter, L., Nobari, A. H., and Ahmed, F., 2022, "Deep Generative Models in Engineering Design: A Review," ASME J. Mech. Des., **144**(7), p. 071704.

[30] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., and Shah, M., 2023, "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage," TechRxiv.

[31] Jovanovic, M., and Voss, P., 2024, "Towards Incremental Learning in Large Language Models: A Critical Review," arXiv.

[32] Shah, J. J., and Vargas-Hernandez, N., 2003, "Metrics for Measuring Ideation Effectiveness," J. Des. Stud., **24**(2), pp. 111–134.

[33] Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z., 2024, "A Survey on In-context Learning," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, November.

[34] Runco, M. A., and Jaeger, G. J., 2012, "The Standard Definition of Creativity," Creat. Res. J., **24**(1), pp. 92–96.

[35] Deldin, J.-M., and Schuknecht, M., 2013, *The AskNature Database: Enabling Solutions in Biomimetic Design*, Biologically Inspired Design, Springer, London, pp. 17–27.

[36] Weisberg, R. W., 1988, "The Nature of Creativity: Contemporary Psychological Perspectives," *Analogical Thinking in Problem-Solving and Creativity*, 1st ed., Vol. 150, Cambridge University Press, New York, pp. 148–176.

[37] Maddah, N., and Heydari, B., 2024, "Building Back Better: Modeling Decentralized Recovery in Sociotechnical Systems Using Strategic Network Dynamics," J. Reliab. Eng. Syst. Safety, **246**, p. 110085.

[38] Axtell, R., and Farmer, J., 2022, "Agent-Based Modeling in Economics and Finance: Past, Present, and Future," J. Econ. Lit., **14**.

[39] Gorochowski, T. E., Matyjaszkiewicz, A., Todd, T., Oak, N., Kowalska, K., Reid, S., Tsaneva-Atanasova, K. T., Savery, N. J., Grierson, C. S., and di Bernardo, M., 2012, "Bsim: An Agent-Based Tool for Modeling Bacterial Populations in Systems and Synthetic Biology," PLoS One, **7**(8), p. e42790.

[40] Varshney, L. R., 2019, "Mathematical Limit Theorems for Computational Creativity," IBM J. Res. Dev., **63**(1), p. 8618382.