BIDTrainer: An LLMs-driven Education Tool for Enhancing the Understanding and Reasoning in Bio-inspired Design

Liuqing Chen
chenlq@zju.edu.cn
College of Computer Science and Technology, Zhejiang University
Hangzhou, China
Zhejiang-Singapore Innovation and AI Joint Research Lab
Hangzhou, China

Zhaojun Jiang
22321008@zju.edu.cn
College of Computer Science and Technology, Zhejiang University
Hangzhou, China

Duowei Xia
22321230@zju.edu.cn
College of Computer Science and Technology, Zhejiang University
Hangzhou, China

Zebin Cai
caizebin@zju.edu.cn
College of Computer Science and Technology, Zhejiang University
Hangzhou, China

Lingyun Sun
sunly@zju.edu.cn
College of Computer Science and Technology, Zhejiang University
Hangzhou, China
Zhejiang-Singapore Innovation and AI Joint Research Lab
Hangzhou, China

Peter Childs
p.childs@imperial.ac.uk
Dyson School of Design Engineering, Imperial College London
London, United Kingdom

Haoyu Zuo*
hz2019@ic.ac.uk
Dyson School of Design Engineering, Imperial College London
London, United Kingdom

ABSTRACT

Bio-inspired design (BID) fosters innovations in engineering. Learning BID is crucial for developing multidisciplinary innovation skills of designers and engineers. Current BID education aims to enhance learners' understanding and analogical reasoning skills. However, it often heavily relies on the teachers' expertise. When learners pursue independent learning using some educational tools, they face challenges in understanding and reasoning practice within this multidisciplinary field. Additionally, evaluating their learning outcomes comprehensively becomes problematic. Addressing these challenges, we introduce a LLMs-driven BID education method based on a structured ontology and three strategies: enhancing understanding through LLMs-enpowered "learning by asking", assisting reasoning by providing hints and feedback, and assessing learning outcomes through benchmarking against existing BID cases. Implementing the method, we developed BIDTrainer, a BID education tool. User studies indicate that learners using BIDTrainer understood BID knowledge better, reason faster with higher interactivity than the baseline, and BIDTrainer assessed the learning outcomes consistent with experts.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

Bio-inspired design, Design education, Analogy training, Design evaluation

ACM Reference Format:
Liuqing Chen, Zhaojun Jiang, Duowei Xia, Zebin Cai, Lingyun Sun, Peter Childs, and Haoyu Zuo. 2024. BIDTrainer: An LLMs-driven Education Tool for Enhancing the Understanding and Reasoning in Bio-inspired Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642887

*Corresponding author

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642887

1 INTRODUCTION

From Velcro, the versatile fastener inspired by burrs [70], to the soft manipulator inspired by octopus suckers [30], bio-inspired design (BID) represents a special form of design-by-analogy [14, 64, 65]. BID, known for adopting analogies from nature to inspire designers and engineers [58], has led to numerous innovative solutions in

the engineering and design fields. Given the popularity and significance of the BID method, its integration into educational programs is increasingly emphasized [72]. BID education not only effectively connects learners to real-world applications [33] but also cultivates the cross-disciplinary innovative abilities required of future designers and engineers [43]. Thus, learners across design and engineering disciplines, including industrial design students and novice designers, as well as mechanical engineering students and researchers, are finding it increasingly essential to study BID [57]. Current BID training programs often rely on classroom lectures and workshops [2, 58, 73], including steps of case studies, design practice, and evaluation. These steps attempt to foster learners' capabilities of understanding and reasoning of BID knowledge, as well as provide effective summative assessment. Notably, BID projects developed in the workshops have gained significant recognition in the design field. For instance, a deployable emergency shelter, inspired by bear hibernation techniques [51], received the Red Dot Award, marking a noteworthy contribution to the broader design community.

However, current BID training programs heavily rely on the curricula and the expertise of instructors due to its multidisciplinary nature. Consequently, learners lacking access to specialized BID curricula and instructors are increasingly turning to learn independently, including understanding BID knowledge, practicing BID reasoning, and evaluating their learning outcomes. In traditional classroom settings, learners typically receive BID knowledge passively. To independently understand BID, learners can now actively seek information from dedicated repositories like AskNature [8]. However, there is still the challenge for learners of understanding complex BID knowledge in multiple disciplines [56]. Moreover, traditional classroom education often adopts a generic, one-size-fits-all approach to BID reasoning. In response, learners can now apply structured templates for engaging in BID reasoning that aligns with their interests [6]. Despite this, the challenge persists in bridging the multidisciplinary reasoning gap [71], leading to mistakes in BID practice, such as poor problem-solution pairing [21]. Due to the large student groups and high cost of time associated with individual assessments, traditional education often employs summative assessment. Such summative assessments do not allow learners to improve their learning and performance in real time, unlike formative assessments [16, 52]. Currently, learners are able to independently assess their design outcomes qualitatively using tools like analogy assessment charts for continuous improvement of their designs [20]. However, they still face challenges in obtaining quantitative and comprehensive evaluations, which are crucial for thorough learning [18]. In all, these three significant challenges — difficulties in understanding complex multidisciplinary knowledge, bridging the multidisciplinary reasoning gap, and obtaining comprehensive assessments.

Recently, Large Language Models (LLMs) have demonstrated powerful capabilities in natural language understanding, enabling them to perform various tasks including text generation, summarization, and intelligent question-answering [3]. The advancement of LLMs offers the potential for fostering learners' understanding and training their reasoning skills in BID education. LLMs, such as GPT-4, possess a multidisciplinary knowledge base and have the ability to understand and generate natural language texts, particularly in complex and nuanced scenarios [47]. This knowledge base, combined with their text processing abilities, makes them valuable for supporting educational tasks across multiple disciplines. Furthermore, LLMs have drawn increasing attention to the automatic scoring of student-written responses because of their extensive knowledge base and context awareness [36, 48].

To address the challenges in BID education, we propose an LLMs-driven BID education method, based on a structured ontology and consisting of three strategies. The structured ontology presents multidisciplinary knowledge in BID cases in a cognitively efficient manner. This ontology is utilized throughout the following three education strategies, and each focused on specific learning aspects: (1) Strategy for BID understanding: LLMs are employed for explaining knowledge within the ontology through "learning by asking" interaction between learners and LLMs. (2) Strategy for training BID reasoning: Reasoning steps that break down the multidisciplinary reasoning process are offered, aiming to bridge the gap between biological solutions in the ontology and their engineering applications. Additionally, LLMs aid this process by offering reasoning hints and feedback through dialogue interactions. (3) Learning Evaluation Strategy: A quantitative and comprehensive assessment of learners' outcomes, benchmarked by existing BID cases aligned with the ontology is provided. These strategies address challenges in understanding, reasoning, and evaluating respectively. The structured ontology of BID cases serves as a foundational knowledge base, specifically illustrating the multidisciplinary relationships in BID cases for understanding, presenting biological solutions for reasoning exercises, and providing benchmark cases for evaluating.

Applying this method, we developed a streamlined BID educational tool called BIDTrainer for learners seeking to acquire BID skills. This tool caters to diverse users, including university students in design and engineering-related disciplines, such as industrial design and mechanical engineering, as well as novice designers and researchers engaged in design-related activities in the industry. BIDTrainer facilitates understanding and reasoning in BID, complemented by learning evaluations to ensure a complete educational experience. Learners start by selecting a link within the ontology, proceed to understand or practice reasoning partnered with the tool through "learning by asking" interaction, and finally receive quantitative feedback on their learning outcomes. To evaluate the impact of BIDTrainer on BID understanding and reasoning, two between-subject user studies (N = 40) were conducted, which indicated enhanced understanding and reasoning among its users compared to a control group. Additionally, an inter-rater reliability study further validated the reliability of BIDTrainer in evaluating, confirming its consistency with expert evaluations.

The contributions of this work can be summarized as follows:

- **Introduction of the BID education method**: This method leverages a structured ontology for the presentation of multidisciplinary knowledge in BID education. Based on this ontology, the method facilitates understanding of complex knowledge through "learning by asking", bridges the multidisciplinary reasoning gap by offering reasoning steps, and employs comparative evaluations benchmarked by existing cases to offer learning feedback, thus forming a comprehensive BID teaching pattern.

- **LLMs-empowered interactive BID learning**: The natural language understanding capabilities of LLMs are leveraged to provide knowledge explanations and reasoning assistance through "learning by asking" interaction. Meanwhile, LLMs are utilized to offer timely and comprehensive evaluations of learning outcomes, enabling self-adjustment for learners.
- **Development of BIDTrainer**: An education tool that applies the BID education method and interacts with LLMs. Developed and utilized in experimental settings, BIDTrainer has been shown to enhance understanding and improve reasoning efficiency in BID compared to the control group. It can also provide valid evaluations and an engaging learning experience.

2 RELATED WORKS

2.1 BID education

Engineers are expected not only to possess technical expertise but also to engage in cross-disciplinary innovation [43, 44]. This necessitates operating beyond the narrow confines of a single discipline. A practical method to cultivate these cross-disciplinary abilities is through the integration of bio-inspired design (BID) into education programs [43].

There have been several real-world BID education programs [2, 53, 58, 73]. Among them, the simplest form of integrating BID into education is through case study learning. Studies have indicated that BID cases provided during classes can help learners to learn innovative design principles [38]. For example, presenting learners with high-quality biological knowledge text sections [31] as cases has been found to aid them in understanding biological knowledge, thereby enriching their design projects with biological insights, and ultimately enhancing the quality of their schemes [38].

The academic community has also considered how to integrate the BID method into education programs. When talking about these methods, it is important to distinguish between the problem-driven and solution-driven approaches [72]. In BID education, if there is a given problem, learners need to seek solutions in a problem-driven approach, which is very similar to current design classes. But if there is a specific biological solution, learners should reason its applications through a solution-driven approach. The BID methods, as exemplified in the following, often integrate into educational programs with one of these two approaches intentionally or unintentionally. For example, researchers use the Structure-Behavior-Function (SBF) model to help learners during their BID learning. The model enhances the understanding of the behaviors and functions of biological systems, implementing creative design in a solution-driven approach [62]. The TRIZ methodology, rooted in engineering innovation, evolves into Bio-TRIZ which enables learners to come up with richer BID concepts [10]. The Concept-Knowledge (C-K) theory model has also been explored in combination with BID. The model divides the BID field into two parts: the biological knowledge space and the concept space, helping learners to explore within them, reason new knowledge, and transfer this knowledge to the concept space to form design concepts [49]. Previous studies [31, 38] have proven the effectiveness of learning cases in BID education and the utility of employing appropriate design methods. In these studies, two crucial objectives in BID education programs are: understanding biological or engineering knowledge, and acquiring the skill of reasoning between the fields of biology and engineering. The former corresponds to case study learning, while the latter embodies the essence of numerous BID methods.

Additionally, the real-world implementation of current BID education, including K-12 training programs [24], higher education courses [25], and industry training [19], relies heavily on experienced teachers with cross-disciplinary expertise. However, such professionals are not readily available [11], and this format can be overly passive for learners.

2.2 BID education tools

To solve the problem of excessive reliance on teachers, various tools have been proposed, attempting to enhance BID understanding, BID reasoning, and offer learning evaluation. The two types of tools are introduced: BID understanding and reasoning tools, as well as BID learning evaluation tools.

In the learning of BID, learners deepen their understanding of BID knowledge through unstructured BID case repositories and structured knowledge presentation tools. For instance, BID case repositories like AskNature provide knowledge about biological phenomena in the "biological strategies" section, design cases in the "innovations" section [8]. Meanwhile, Bionic Inspiration [1] showcases design projects inspired by nature. To reduce the cognitive load caused by the unstructured information in these repositories, structured methods are used for knowledge presentation in BID. Commonly used structured BID presentation tools include Biologue, enabling users to tag BID documents with Structure-Behavior-Function (SBF) model [62]; Design by Analogy to Nature Engine (DANE), which uses the SBF model to capture the functioning of biological systems [63]; and Idea-Inspire 4.0, providing several biological systems and representing them using a multi-modal representation, such as function decomposition model, image, video, and so on [55]. When learners seek to train their analogical reasoning abilities in BID practice, the "Four Box Method" of SR.BID assists in structuring the formulation of design problems [20]. Additionally, a causal relation template based on Gentner's theory aids design learners in analogical reasoning [6].

However, there is a significant lack of tools and resources specifically for BID education, posing challenges for novice bio-inspired designers [45]. Their difficulties in understanding multidisciplinary knowledge and extracting biological principles for reasoning often cause them to seek external assistance, such as from biologists, teachers, libraries, or online resources [69]. In tool-based learning environments without expert or teacher involvement, learners are left to consult libraries or internet resources to assist in understanding and reasoning. Therefore, developing more intuitive and easy-to-understand BID education tools is crucial for enhancing the quality of both education and professional practice in this field.

Evaluation is crucial in multidisciplinary education, such as BID, to ensure that learners are effectively progressing in their learning process [15]. There are tools for qualitative and quantitative evaluation during BID education. For qualitative evaluation in BID

[1] https://www.bioinspiration.net/

education, the "T-Chart" tool requires learners to perform side-by-side evaluations of biological systems concerning their design problems using qualitative comparisons [20].

For quantitative evaluation during BID education, Yen et al. [71] initially gave learners three quantitative homework tasks, but due to the difficulty experienced by many learners, they shifted these tasks from individual to group assignments. This change was accompanied by the introduction of the "Make-or-Break Quantitative Analysis", focusing on a single crucial quantitative function. This indicates that BID learners manage only simple qualitative or single-criterion quantitative evaluations independently. Therefore, obtaining comprehensive quantitative evaluations crucial for teaching [18] remains a significant challenge.

2.3 AI including LLMs in design education

While some educational BID tools offer powerful learning functionalities, ranging from case studies to design evaluation, their effectiveness still depends on external assistance [22]. Moreover, finding teachers with deep expertise in specialized fields as external assistance is often difficult. As an emerging solution, the research community has been actively exploring AI-driven technologies for design education.

In design education programs, AI-driven systems are often used to support learners' creative processes. Researchers have constructed a learning platform called "Online Design Studio" within design courses. The platform has considered various design tasks such as Data evaluation and problem identification which were standardized and formulated to adapt to AI input and output formats, thus providing AI support for students [63]. Another domain is AI-driven evaluation, replacing teacher evaluation. Although there is a lack in judging diverse overall design proposals in innovation work [20], AI can be used to evaluate short-answer texts with structural features [13], providing scores with high accuracy. At times, generative AI systems play an important role in bolstering the academic grading framework [61], positioning their assessment outcomes as external references for educators and learners. Moreover, interactive educational chatbots based on natural language technology are also utilized as instructional agents and learning assistants to engage learners and facilitate personalized learning activities [34].

These tools, along with the supported AI technologies, to a certain extent assist learners in independent design learning. However, these systems fall short in aiding learners to understand knowledge or design methods. Additionally, it is noteworthy that in domains like BID education, which involve multidisciplinary knowledge, the presence of AI tools remains scarce. These challenges are due to shortage in AI datasets and the training procedures [34], which are inadequate for cross-disciplinary work in BID education including biology and engineering. Additionally, they are less effective in supporting multi-turn dialogues, thereby limiting their ability to assist learners in learning knowledge and design methods.

The advancement of Large Language Models (LLMs) offers a potential resolution. With their extensive training dataset, LLMs provide rich knowledge [67], enhanced natural language processing capabilities [23], and the ability to support multi-turn dialogues which makes them apt for complex design support contexts. LLMs have been integrated into certain design tools for the rapid generation of creative works, serving as an effective learning tool [4]. Within the field of educational research, LLMs are also recognized for their ability to imitate teachers' behaviors, thereby substituting for certain teaching tasks [26]. This includes the quick production of teaching materials [9] or serving as learning assistants, understanding the diverse needs of students, and producing tailored knowledge [60] to help them. Importantly, as the last component of the teaching process, LLMs are able to provide assessments of students' written answers [42], helping to rectify their misunderstandings and offering immediate, substantive feedback.

These studies provide insights for our research regarding the role of LLMs in design education and lays the groundwork for our technical implementation. In this work, we primarily employ LLMs as a supporter to assist learners in understanding the knowledge within BID cases, reasoning between biology and engineering fields, and facilitating learning assessments.

3 AN LLMS-DRIVEN BID EDUCATION METHOD

In this work, we propose an LLMs-driven BID education method, including three strategies designed to address the challenges in BID understanding, reasoning, and evaluating in the multidisciplinary context of BID, as shown in Figure 1. The method involves a structured ontology that intuitively presents multidisciplinary knowledge from biology, the intersection of biology and engineering, and engineering disciplines in BID, forming the foundation of our education method. Additionally, LLMs play roles in enhancing BID understanding by providing multidisciplinary knowledge explanations and assisting in BID reasoning by offering reasoning hints and feedback. For learning evaluation, LLMs function as evaluators, delivering scores and feedback to learners.

3.1 Structured ontology

Inspired by AskNatureNet [5], which presents BID cases in a structured format, we adopted a structured ontology to reduce the cognitive load for learners. This ontology is employed throughout the BID education method. It provides structured BID cases as materials for understanding and biological solutions as reasoning sources, as well as aligns learners' chosen materials or sources with corresponding original cases as evaluation benchmarks. Adopting the approach used in AskNatureNet which employs structured triples to connect biological and engineering domains, we present each BID case from the "innovations" section of AskNature [8] as structured triples: "source -> benefits -> application (S -> B -> A)" or "application -> benefits -> source (A -> B -> S)." Here, "source" refers to the specific organism, system, or process in nature that serves as the biological inspiration; "benefits" refer to the advantages, features or properties that the biological source exhibits; "application" describes how the biological benefits can be translated into the innovative designs or solutions. BID cases in the "S -> B -> A" format are the presentation for solution-driven design, while those in the "A -> B -> S" format cater to problem-driven design. These triples reveal relationships between the biological and engineering domains, providing learners with intuitive and efficient BID cases for understanding. When the "application" element of "S -> B -> A" triples is hidden, the

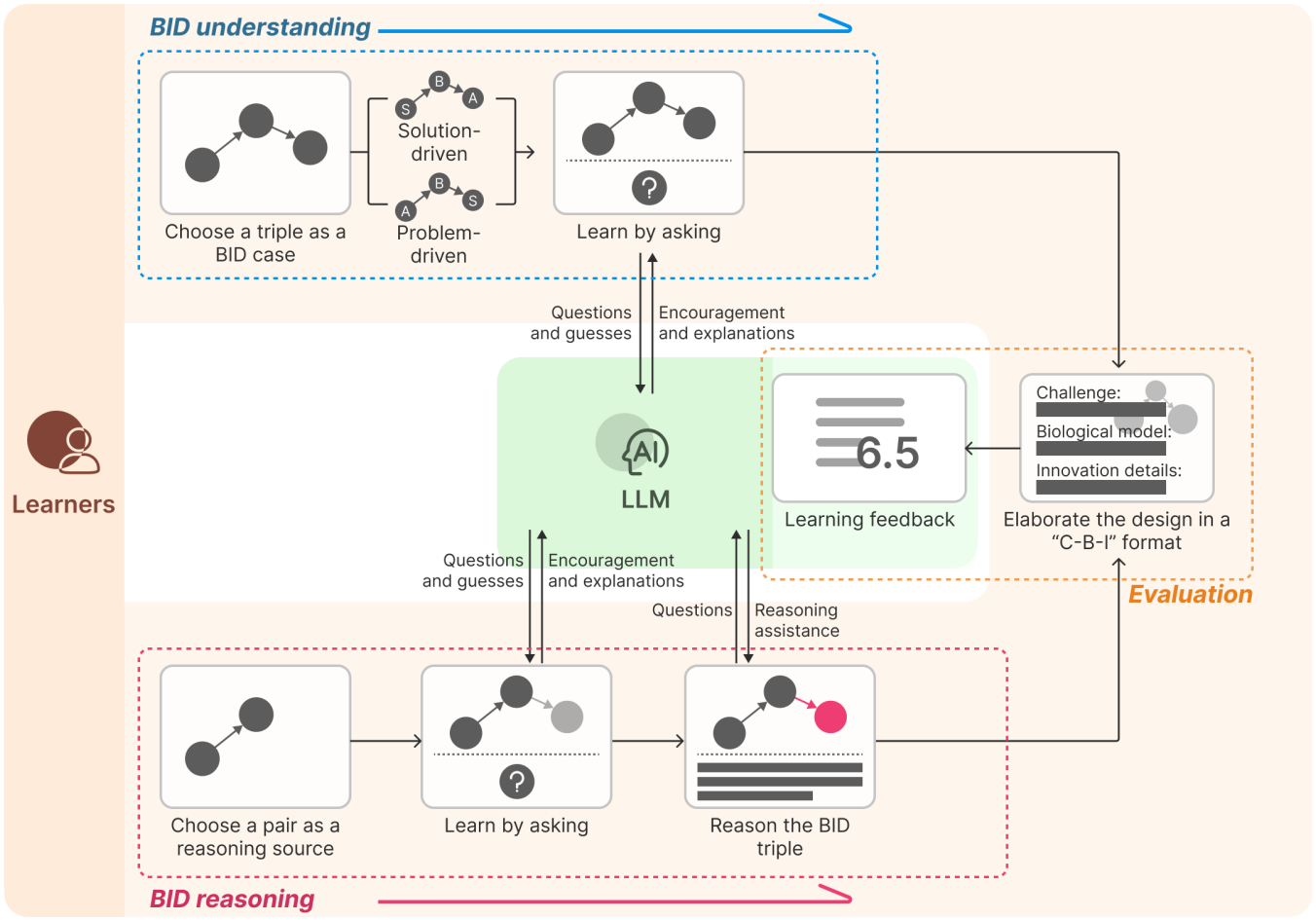


Figure 1: LLMs-driven BID education method including the strategy for assisting BID understanding, the strategy for training BID reasoning, and learning evaluation strategy.

remaining pairs of "S -> B" emphasize biological solutions used in BID cases, serving as sources for analogical reasoning training. The original cases of the chosen triples or pairs from the "innovations" section of AskNature, validated for their quality within the industry, are utilized as benchmarks for evaluating learners' design schemes.

3.2 Strategy for assisting BID understanding

BID involves enriching subject knowledge and developing design skills, which can be acquired through case study learning [71]. To address the challenge of understanding unfamiliar disciplinary content in case studies, we utilize LLMs to conduct the "learn by asking" approach [41], providing knowledge explanations.

To effectively leverage LLMs in our BID understanding strategy, we focus on how LLMs can offer relevant and accessible explanations interactively. LLMs are known for their ability to generate informative explanations [12], and we apply this ability by adopting the "learning by asking" approach. This approach, combined with making guesses, ensures interactive learning while enhancing knowledge retention [7]. Although learners must self-assess their knowledge scope when posing questions [41], the multidisciplinary breadth of BID often makes it challenging. To address this, we implement King's "guided student-generated questioning strategy" [32], which employs generic questions to prompt specific questions. Tailored to the unique aspects of BID, we categorize generic questions into three types: biology, the intersection of biology and engineering, and engineering, as shown in Table 1. Learners, guided to ask LLMs a specific question along with their guesses, receive encouragement and detailed explanations from LLMs. This "learn by asking" approach with LLMs personalizes knowledge explanations for learners in BID case studies within human-computer interactions.

3.3 Strategy for training BID reasoning

In BID education, learners are required to reason from biological solutions to derive applications in the field of engineering [72]. This process requires learners to have the skill of multidisciplinary reasoning and to possess foundational knowledge in both the biological and engineering fields. As an example, it is acknowledged that humans invented radar by learning from bats, but it is almost impossible to directly transfer the ultrasonic waves of bats to radar,

Table 1: Generic questions in "guided student-generated questioning strategy" categorized in three domains

Domain	Generic questions
Biology	What is the meaning of "S" or "B?" What are the strengths and weaknesses of "B?" What do you think causes "B?" What would happen if "S" doesn't have "B?" Why is "B" important? What is the difference between "B?"
The intersection of biology and engineering	Explain why "B" can be used in "A?" Explain how "B" can be used in "A?" How does "B" affect "A?" How do you use "B" to "A?"
Engineering	What are some possible solution of "A?" What is the best "A" and why?

a product that had never previously appeared. The breakthrough achieved by engineers was due to their understanding of how bats use reflected ultrasonic waves to detect obstacles. With this knowledge, engineers reasoned out a technology that humans can utilize: emitting waves, detecting reflected signals, and ultimately creating usable radar based on this principle.

In response to the requirements of reasoning mentioned above, within the BID educational framework, we designed a reasoning training strategy to teach learners to implement analogy reasoning with the help of LLMs. This strategy includes two supports to assist learners in reasoning training: reasoning steps and reasoning assistance from LLMs. These steps follow a "source -> benefits -> application" framework, guiding learners to start their reasoning from the "source" and its "benefits", and the goal is to explain the specific "application". The reasoning steps are based on the solution-driven BID process [21] and have been modified according to our framework. Learners are presented with a "source -> benefits" link, but the application part is intentionally left out. They are encouraged to engage in the reasoning process, applying their understanding to identify the application. The steps are detailed below:

Step 1. Identify the problems associated with the "source." Learners are tasked to investigate "what specific problems can be solved by the 'benefits' originating from the 'source'?". These could vary from avoiding predator attacks to basic needs like eating or moving objects. For instance, considering the link "elephant trunk -> strong, flexible": the elephant, despite its large size, has a trunk that is both strong and flexible. This allows it to handle small objects and eat, addressing challenges in mobility and feeding due to its size.

Step2. Extract the principle from "benefits". In this step, learners should investigate "how do the benefits practically address the identified problems?". In the case of elephants, a strong and flexible trunk provides a flexible method for grasping and operating, thereby solving the problem of eating.

Step3. Identify the requirements of human. In this step, learners should consider, "how can these problem-solving solutions be applied to real-world human requirements?". Continuing with the previous example: humans also frequently require the ability to grasp and operate objects, such as during medical surgeries, representing a requirement that can be addressed by solutions that are "strong and flexible".

Step4. Find the "application". This step involves describing a product that addresses the identified requirements. The question is "how can the problem-solving solutions be translated into innovative engineering applications?". Learners can either use existing products as applications or describe a product that does not yet exist. For example, a medical robotic arm can be considered as an application reasoned from the elephant's story.

We suggest the integration of LLMs in this educational work. LLMs perform well at handling a wide range of language tasks, including context-driven knowledge comprehension and crafting tailored explanations in alignment with learners' requirements [1]. When implementing the above steps, learners still need specific knowledge to support them. For instance, when learners identify the requirements in Step 1, they may be uncertain whether these problems genuinely reflect the challenges faced by the "source". At this point, LLMs can serve as reasoning assistants, efficiently giving feedback on their answers. If learners struggle to form effective assumptions due to knowledge gaps, LLMs can offer hints to learners (see Table 2, Prompt 2), including the provision of suggestive words and relevant knowledge, thereby facilitating the smooth execution of the reasoning steps.

By engaging in our reasoning process and seeking support from LLMs when encountering challenges, learners can bridge the multidisciplinary gap between biological solutions and engineering applications. This process, as shown in Figure 2, enables reasoning outcomes that are supported by theoretical foundations.

3.4 Learning evaluation strategy

To overcome the difficulties in gaining timely and comprehensive evaluations, we've implemented a comparative evaluation strategy using LLMs in our education method. This strategy ensures comprehensive, timely, and reliable feedback for learners while breaking away from reliance on teachers.

When using LLMs for evaluation tasks, a common practice is to provide benchmark answers for consistency checks [29]. In fact, if

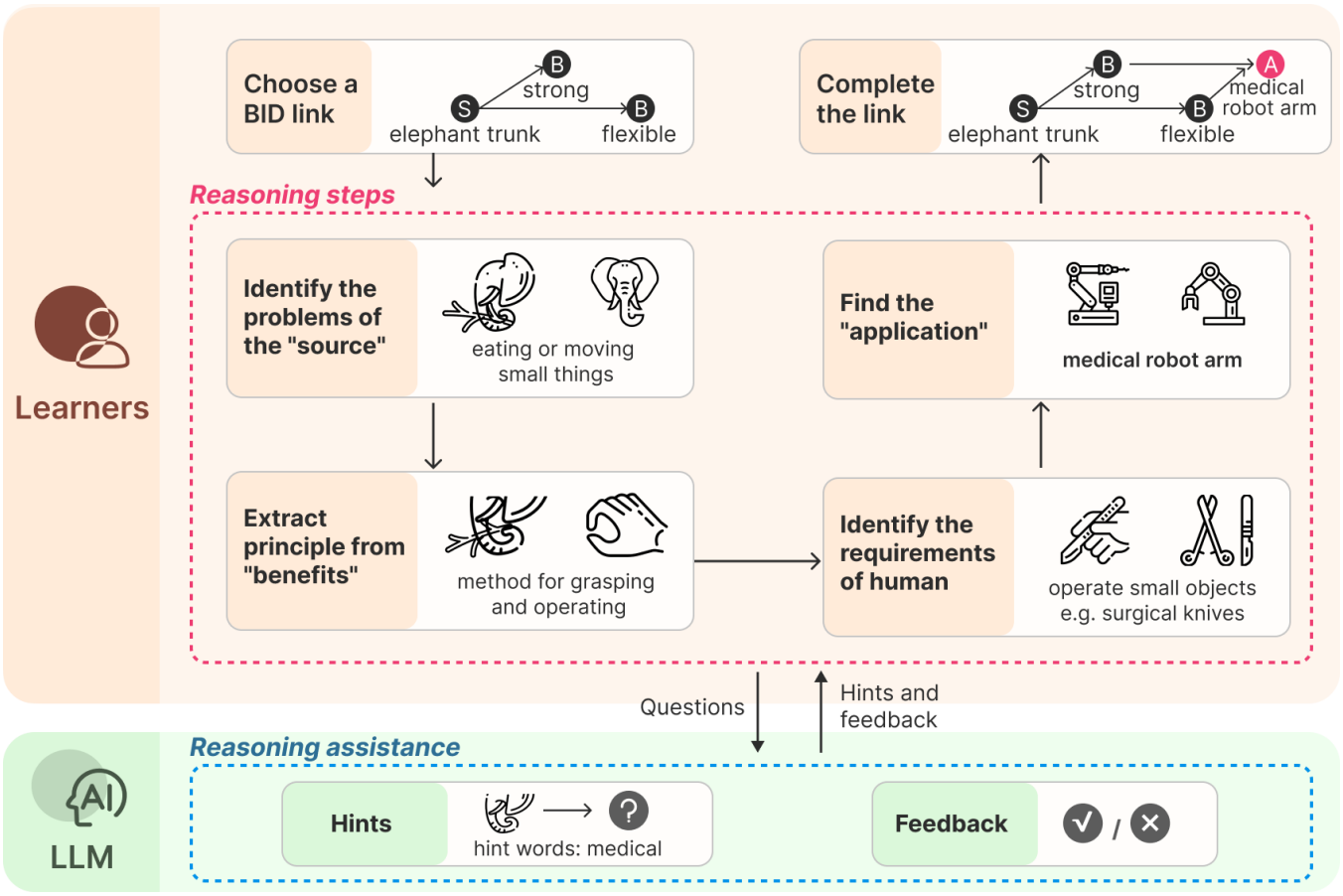


Figure 2: The LLMs-driven BID reasoning strategy. The process starts with an incomplete link, "elephant trunks -> strong, flexible", without the application. Through the steps, the process evolves to enable development of reasoning for "medical robot arm" as a potential "application". Within this strategy, the LLMs function primarily to offer hints and feedback for learners.

these benchmarks are presented in structured forms, LLMs evaluation becomes even more precise, aligning closely with human teacher grading [29]. Consequently, we adopt a comparative evaluation approach, benchmarked by real-world BID cases from the "innovations" section of AskNature, following the "Challenge-Biological Model-Innovation Details (C-B-I)" paradigm. Here, "Challenge" refers to problems in current design regarding user experience, industrial production, engineering requirements, cost, etc. "Biological Model" refers to the biological solutions that serve as the design source, including descriptions of biological "benefits" and the mechanisms of "benefits" of the "source". "Innovation Details" is a detailed description of design ideas inspired by biology. To facilitate comparative assessment, learners are also required to follow this "C-B-I" paradigm to elaborate their design schemes after understanding the BID triples or reasoning the BID pairs. The selected evaluation benchmarks are consistent with the BID triple or pair cases initially chosen by the learners. This thematic benchmarking not only enhances the accuracy of evaluations but also serves as a standard answer to learn.

LLMs can be guided to take on the behavior of a teacher [26], evaluate learning outcomes, and provide feedback that is specific to each learner's design. For a comprehensive assessment, we instruct LLMs to evaluate from perspectives of novelty, quality, and analogical rationality. Novelty and quality are standard metrics for assessing design schemes [54], while analogical rationality is a crucial criterion in BID, reflecting its nature as a design by analogy [20]. To ensure reliability, we utilize benchmarking and range criteria, further validating reliability through consistency analysis with expert scoring. Leveraging these approaches, LLMs evaluate learners' design schemes, assign a score, and provide a reason for the score.

The learning evaluation strategy can be employed following either the BID understanding strategy or the BID reasoning strategy, embarking on a new cycle of understanding or reasoning training. Ultimately, it is our vision to form a comprehensive educational experience where learners combine knowledge acquisition, design innovation, and evaluative feedback independent of teachers' assistance in BID education.

4 BIDTRAINER

Based on the BID education method, we introduce BIDTrainer, a novel tool designed to enhance learners' understanding and reasoning skills in BID independently. We have chosen to integrate the state-of-the-art Large Language Models (LLMs) , specifically GPT-4, due to its abilities to handle multidisciplinary data and generate contextually relevant responses [46, 47]. These abilities are employed to assist knowledge understanding, train analogical reasoning, and provide comprehensive evaluations.

4.1 Learning support from GPT-4

Given GPT-4's demonstrated effectiveness in educational settings, particularly in understanding complex contexts and producing written outputs [35], it was chosen to assist in our BID educational strategies. The robust natural language understanding and generation capabilities of GPT-4 make it suited for providing knowledge explanations, reasoning assistance, and scoring in a BID context. These functionalities are integrated into our tool via GPT-4's API[2].

Knowledge explanations. For the strategy for BID understanding and reasoning, learners acquire knowledge explanations by asking GPT-4 specific questions and posing guesses in multidisciplinary domains. To ensure effective "learning by asking" interaction between learners and the tool, GPT-4 is prompted to act as a BID expert and to "explain BID knowledge to non-biological learners in an understandable and encouraging way." The main prompts used are outlined in Table 2 and the complete prompts are listed in the Supplementary Materials. To ensure that GPT-4 remains encouraging and reliable in the interaction, we have incorporated prompts emphasizing "encouraging the learner" and advising that "if you are unsure about the knowledge, then simply write 'please consult external sources or seek help from teachers.'"

Reasoning assistance. In our strategy for BID reasoning, we provide learners with reasoning steps. To assist their reasoning at each reasoning step, we use GPT-4 in an expert role to provide hints and feedback. If learners are stuck at a step, GPT-4 provides corresponding reasoning hints to aid their progress. When learners reach a conclusion at a step but are uncertain, GPT-4 gives feedback on their conclusions. The specific prompts for these scenarios are listed in Table 2. To ensure that learners, rather than GPT-4, play the primary role in reasoning practice, we have emphasized the distinction between hints and directly revealing answers in our prompts. Additionally, we also check whether GPT-4 is sure about the reasoning in the prompts to enhance the reliability of the responses.

Scoring. GPT-4 has shown higher levels of performance in automatically scoring student-written constructed responses than other models [36]. Therefore, for the design schemes of learners presented in the "Challenge-Biological Model-Innovation Details" format, we employ GPT-4 in the role of evaluators to score their schemes. Scores are graded up to ten, with GPT-4 instructed to assign scores ranging from poor to excellent. To ensure scoring accuracy, we have defined scoring ranges of two points each, labeled as "poor", "pass", "moderate", "good", and "excellent", as well as provided detailed descriptions for scores fitting these ranges. Moreover, we use a few-shot learning techniques, using cases from the "innovations" section in AskNature as benchmarks which can be scored in the "excellent" range. To clarify that GPT-4 has correctly understood the learner's scheme and provided a rational score, we use guided generation prompts for GPT-4 to give comments in a specific format, as shown in Table 2.

[2]http://openai.com/api/

4.2 Implementation of BIDTrainer

We developed BIDTrainer (Figure 3), an educational tool for BID. It consists of two educational modes: the BID understanding mode and the BID reasoning mode. When entering the tool, all the relevant concepts contained in the tool are thoroughly explained to users by a tutorial video. In the BID understanding mode, learners can engage in learning using either the problem-driven or solution-driven approach. The tool follows the BID education method described and integrates learning support from GPT-4. The user flows for each mode are outlined below:

BID Understanding Mode: When learners select the BID understanding mode in the upper panel (Fig. 3(a)), the tool presents the structured ontology and provides options for problem-driven or solution-driven approaches (Fig. 3(b)). When learners choose the problem-driven design approach, the structured case is presented as "A-> B ->S" while selecting the solution-driven design approach results in the structured case being presented as "S -> B -> A". After selecting the design approach, learners freely choose a link (Fig. 3(c)) and interact with the tool to obtain explanations for the case. When learners engage in the interaction by applying the "guided student-generated questioning strategy" with the tool, it provides them with prompts based on generic questions (Fig. 3(g)). After a thorough understanding of the case, learners are required to elaborate their specific BID schemes in the C-B-I format (Fig. 3(i)). Finally, the system uses the original BID case corresponding to the chosen link as a benchmark to provide learners with scores and specific comments (Fig. 3(j) and (k)). Thus, the tool assists learners in understanding BID cases, problem-driven or solution-driven BID practices, and BID evaluation, forming a comprehensive education experience.

BID Reasoning Mode: When learners select the BID reasoning mode in the upper panel (Fig. 3(a)), the tool presents structured biological solutions used in BID cases, namely, "S -> B" link in the ontology (Fig. 3(c)). After learners freely choose a link as the source for reasoning based on their interests, they bridge the gap from biological solutions to engineering applications using the reasoning steps provided by the system (Fig. 3(f)). During the reasoning process, the tool offers knowledge explanations and reasoning assistance through the dialogue interaction (Fig. 3(g)), ensuring a smooth reasoning process and enabling learners to learn by asking. After completing the reasoning, learners will derive an "application" that can utilize the given biological solution to address a problem, forming an "S -> B -> A" link. Subsequently, learners are required to elaborate their specific proposals in the "C-B-I" format. Finally, the system scores learners based on the original case corresponding to the biological solution and provides specific comments. Thus, the tool assists learners in understanding biological solutions, engaging in reasoning, and conducting evaluation.

To address the known issue of hallucinations in GPT-4 generated content [47], we provided learners with some usage tips to mitigate

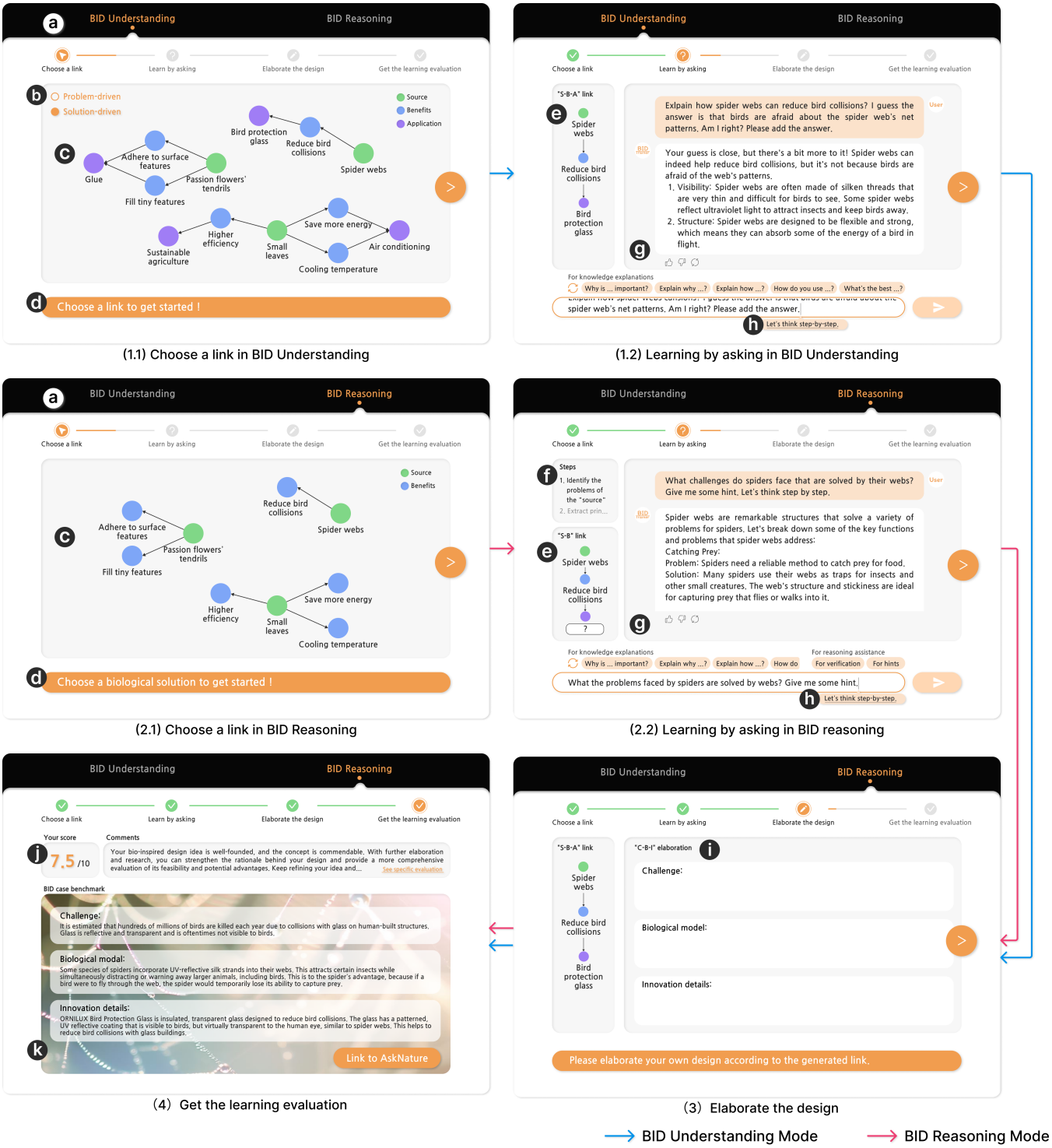


Figure 3: The user flow of BID understanding mode and BID reasoning mode in BIDTrainer includes (1.1) or (2.1) choose a link as a BID case in BID understanding mode or as a biological solution in BID reasoning mode, (1.2) or (2.2) learn by asking for knowledge explanations or reasoning assistance, (3) elaborate the design, (4) get the learning evaluation. The user interfaces include (a) Mode Selection, (b) Design Approach Selection, (c) Link Selection and (d) Bootstrap Box in (1.1) and (2.1); (e) Chosen Link Display, (f) Reasoning Steps, (g) Dialogue Interaction and (h) Chain-of-Thought prompting technique recommendation in (1.2) and (2.2); (i) Design Elaboration in (3); (j) Learning Evaluation, (k) BID Case Display in (4).

Table 2: The prompts used to provide knowledge explanations, reasoning assistance, and scoring

Prompt1:	**For knowledge explanations**
Prompts:	As a design learner focusing on bio-inspired design, I find some BID knowledge challenging due to my non-biology background. Please explain the knowledge as an experienced BID teacher for me. During our conversation, I'll be providing some guesses along with my questions. Please verify the accuracy of these guesses, and provide additional knowledge explanations. When responding, focus on encouraging the learner, and pay close attention to the learner's questions and guesses. If you are unsure about about the knowledge, then simply write "please consult external sources or seek help from teachers."
GPT-4:	Sure.
Learner:	[Specific questions] + [Guesses]
GPT-4:	[Encouragement] + [Knowledge explanations]
Prompt2:	**For reasoning assistance**
Prompts:	As a design learner focusing on bio-inspired design, I'm working on reasoning an application from a biological solution. Please assist by providing hints or giving feedback on my reasoning as an experienced BID teacher. When I ask for hints, focus on providing relevant and helpful hints without revealing direct answers. When I ask for feedback, I'll be providing some answers along with my questions. When responding, pay close attention to the learner's questions and answers. If you are unsure about about the reasoning, then simply write "please consult external sources or seek help from teachers."
GPT-4:	Sure.
Learner:	[Questions from reasoning steps] + I need some hints, please.
GPT-4:	[Hints]
Learner:	[Questions from reasoning steps] + [Answers]
GPT-4:	[Feedback on the answers]
Prompt3:	**For scoring**
Prompts:	As a design learner focusing on bio-inspired design, I've developed a design scheme but need help evaluating it. Please, as an experienced BID evaluator, grade my design based on quality, novelty, and analogical rationality. The scoring is out of 10, ranging from poor to excellent. The detailed descriptions for each scoring range are as follows: ... Here's an example of a design case rated as excellent for reference: [A BID case from AskNature in C-B-I format] Please give the score and comments on it in the following format: 1.Score; 2.Strengths; 3.Innovation details; 4.Areas for improvement.
GPT-4:	Sure.
Learner:	[Learner's design scheme in C-B-I format]
GPT-4:	[Score] + [Comments]

the impact of the hallucinations on learning. Firstly, learners were briefed about the possibility of encountering hallucinations in the interaction with GPT-4. Furthermore, we encouraged the implementation of Chain-of-Thought prompting techniques (Fig. 3(h)), as these have been proven to enhance GPT-4's reasoning capabilities [68]. Additionally, a feedback mechanism for GPT-4 generated answers, as illustrated in Fig. 3(g), includes three interactive elements: a thumbs-up button, a thumbs-down button, and a regenerate button. Learners are allowed to express their satisfaction with GPT-4 generated answers by clicking the thumbs-up, or their dissatisfaction by clicking the thumbs-down. Besides, they have the ability to request the regeneration of these answers by selecting the regenerate button, further tailoring the interaction to their learning preferences and needs.

5 USER STUDY

We conducted a comprehensive evaluation of BIDTrainer through a series of user studies to discover the tool's effectiveness in aiding learners' understanding, reasoning, and evaluation in BID. The subsequent sections describe the participants' characteristics, the experiment procedure, the assessment of the experiment and the corresponding results.

5.1 Participants and settings

Recruitment was conducted via online and on-campus events, targeting primary users of BIDTrainer. This included a diverse group of participants: learners, design researchers, and in-service designers, among others who need to learn BID. The aim was to represent a wide range of BIDTrainer's potential user base to ensure that our study's findings would be relevant and beneficial to the broader design learning community. All participants had the necessary prior experience with LLMs, ensuring familiarity with interactions with the LLMs in BIDTrainer. Each participant spent about 30 minutes and was compensated with 30 CNY. The detailed characteristics of the participants are presented in Table 3. Additionally, we recruited five PhDs in design with BID expertise as experts to evaluate the

Table 3: Characteristics of the participants in our study

	Sample ratio of participants
Gender	
Female	62.50% (N=50)
Male	37.50% (N=30)
Occupation	
Student	65.00% (N=52)
Researcher	27.50% (N=22)
In-service designer	7.50% (N=6)
Degree	
Industrial design	78.75% (N=63)
Mechanical design	7.50% (N=6)
Architectural design	3.75% (N=3)
Other design related	10.00% (N=8)
Learning year about design	
Less than two years	10.00% (N=8)
Two to four years	63.75% (N=51)
More than four years	26.25% (N=21)
Usage of LLMs	
Daily	37.50% (N=30)
Weekly	37.50% (N=30)
Monthly	25.00% (N=20)

participants' outcomes during the experiments, compensating them 80 CNY each.

80 participants were divided into two parts, with each part consisting of 40 participants undertaking different experimental tasks.

- Part 1 (25 females, 15 males) conducted the experiment about understanding.
- Part 2 (25 females, 15 males) conducted the experiment about reasoning.

Five experts actively conducted the experiment about evaluation.

During the experiments, participants were divided into two groups: the experimental group utilized BIDTrainer for BID training, while the control group engaged in current BID learning methods, consisting of unguided exploration through BID-specific websites like AskNature and general search engines like Google. Participants in the control group could freely browse articles related to biology, the summary descriptions presented by search engines on search pages, and even some specialized knowledge websites. Consequently, this constitutes an equitable baseline, mirroring the real-world learning context for learners. To ensure a representative and unbiased distribution, participants were subjected to a multi-step stratified random sampling process, considering multiple layers such as gender, occupation, and learning years about design. This approach effectively balanced key characteristics across the experimental and control groups, maintaining the diversity and comparability essential for the validity of our experiment results. Subsequent to their interactions with the tools, we assessed their learning performance to evaluate the efficacy of BIDTrainer in facilitating understanding and reasoning in BID. Additionally, experts evaluated learners-generated design schemes, providing a measure to verify the accuracy of BIDTrainer's learning evaluation.

5.2 Experiment procedure

Participants received an instructional session about the BID's basic concepts, including problem-driven and solution-driven approaches, and the "S -> B -> A" (for BID understanding) or the "S -> B" (for BID reasoning) links. To facilitate the use of the BIDTrainer, participants assigned to the tool also received a BIDTrainer tutorial. The study was segmented into three experiments:

Experiment about understanding: A between-subject user study was conducted to assess BIDTrainer's effectiveness in assisting BID case understanding. Participants in both groups were presented with the same "S -> B -> A" link of "dragonfly's wings -> small size, lightweight, high efficiency -> aircraft construction" and tasked with understanding it. The experimental group used the BID understanding mode of BIDTrainer. In contrast, the control group employed a baseline method where they freely inquired from BID-specific websites or general search engines to resolve their questions related to the knowledge of the link. After 15 minutes, their understanding was assessed through a quiz.

Experiment about reasoning: Another between-subject user study was to assess BIDTrainer's effectiveness in enhancing analogical reasoning in BID. Participants were presented with the same "S -> B" link of "dragonfly's wings -> small size, lightweight, high efficiency" and tasked with reasoning potential applications. The experimental group utilized the BIDTrainer, while the control group relied on the baseline method to seek for reasoning assistance such as searching for existing cases to acquire inspirations. Unlike the set duration in the previous experiment, this task had an open-ended timeframe, allowing participants to take the necessary time to formulate their reasoning. Once completed, participants submit their reasoning outcomes for expert assessment.

Experiment about evaluation: An inter-rater reliability study was conducted to verify BIDTrainer's accuracy in evaluating learners' design schemes, which are their learning outcomes. The learners' design schemes were proposed by participants who used BIDTrainer in the experiment about understanding, ensuring the evaluated schemes were from users who had a comprehensive experience with the tool. Participants submitted their "C-B-I" format design schemes, evaluated by both BIDTrainer and experts. The scores obtained from them were subsequently recorded to conduct a consistency check.

After the experiments, we conducted feedback interviews with the 40 participants in the experimental group to understand participants' experiences and views on BIDTrainer's impact.

5.3 Assessment

5.3.1 Quality of BID understanding. A quiz was formulated to assess learners' understanding of content from a specific BID case understanding. This assessment comprises nine questions, divided into four multi-choice questions with single answer (2 points for each question), three multi-choice questions with multiple answers (2 points for each question), and two short-answer formats (6 points for each question), summing up to 26 points. The goal of the quiz is to assess whether the participants have acquired a comprehensive understanding of the knowledge provided in the link. Some of the quiz questions are presented in Table 4. Choice questions are designed to assess foundational knowledge, such as determining

Table 4: Examples of questions from the quiz, and see the complete quiz in the Supplementary Materials

Question type 1:	**multi-choice questions with single answer**
Question:	If an airplane is designed to be as agile as a dragonfly, what are its main advantages?
Choices:	A. Extended continuous flight duration. B. The ability to carry larger loads. C. Slightly complex but efficient maneuvers. D. capability for flying missions in high altitudes.
Right choice:	C
Question type 2:	**multi-choice questions with multiple answers**
Question:	The flight mechanism of dragonfly wings can be used as research material for which fields?
Choices:	A. Fluid Dynamics B. Microstructure Control C. Biological Populations and Ecological Environments D. Energy Management and Applications
Right choices:	A, B, D
Question type 3:	**short-answer**
Question:	Summarize the reasons for the small, lightweight, and efficient characteristics of dragonfly wings. (Please provide your answer in a list format.)
Standard answers:	Dragonfly wings are made of a series of adaptive materials, which form a very complex composite structure. This bio-composite fabrication has some unique features and potential benefits. Light performance of dragonflies is one of the examples of nature's efficiency. Dragonflies can fly forwards, backwards and sideways. Dragonfly's fore and hind wings are controlled by separate muscles, and a distinctive feature of the dragonfly's wing movement is the phase relation between those wings during various maneuvers. When hovering, the fore and hind wings tend to beat out of phase; during takeoff, they tend to beat closer in phase. All mentioned characteristics in this paper indicate a highly efficient, lightweight, small size and reliable wing system.

the definition of "benefits" and comprehending the relationship between "source" and "benefits". All correct choices are derived from articles on AskNature to ensure credibility, while the other choices are intentionally misleading incorrect answers. Participants earn 2 points for choosing the correct choices. However, if they pick any wrong choices, they are awarded 0 points, reducing participants' motivation to guess.

The short-answer questions delve into a deeper understanding, prompting participants to articulate the connection between "source" and "benefits" and contemplate potential applications of the material. A dual grading mechanism is used in these questions. Each subjective question can earn up to 6 points, based on three standard answers from AskNature. Participants are advised to structure their answers in separate sections. Each section that matches the intended answer grants them 2 points. However, considering AskNature is a comprehensive open-source platform, it still might not list every possible answer. Participants receive 1 point for any practical and relevant answers, even if they differ from the standard ones. The answer of subjective questions of our participants are scored by two experts. The two sets of scores demonstrated good consistency (Intra-Class Correlation Coefficient (ICC) = 0.90 > 0.75). For each participant, we used the average of these two scores as the final subjective score.

5.3.2 Quality and efficiency of BID reasoning. The main goal of this assessment was to measure both the quality and efficiency of learners when using the BIDTrainer tool for reasoning. The assessment comprises two main components: the quality of the results and the time consumed for reasoning.

Experts with a deep understanding of bio-inspired design were invited to evaluate the quality of the applications output by participants. The quality scores focus on three key areas: quality [54], novelty [54], and analogical rationality [20]. Each application is scored on a 5-point Likert scale. Furthermore, during the process of BID reasoning undertaken by all participants, we documented the time they expended on the reasoning tasks to compare the reasoning efficiency between the experimental group and the control group.

5.3.3 Validity of learning evaluation. The primary objective of this assessment was to determine the validity of LLMs-driven evaluations of BID designs using BIDTrainer. Specifically, the focus was on comparing the alignment between expert evaluations and those generated by GPT-4, an integral component of BIDTrainer.

To standardize the control variables of the experiment, both experts and GPT-4 assessed learners' schemes in the same manner. The task description provided to the experts matched the prompts given to GPT-4, as shown in Table 2. The experts were tasked to evaluate the design schemes proposed by 20 participants from the

external group in the experiment about understanding who had completed the design process using BIDTrainer. The evaluation criteria focus on quality, novelty, and analogical rationality. Each design scheme was scored on a scale of 1 to 10, with higher scores indicating better design. To ensure consistency and reduce subjective biases, experts were instructed to score within predefined ranges. Additionally, during the evaluation process, we provided cases extracted from AskNature as benchmarks.

5.4 Results

Learners achieved higher scores in the understanding quiz and demonstrated greater efficiency in the reasoning process interactively. Moreover, the evaluations generated by the tool were consistent with the experts. To statistically analyze each measure under different conditions, we first conducted a Shapiro-Wilk test to determine if the data was parametric or non-parametric. Then, to compare between conditions, we used a paired t-test (if parametric) and a Wilcoxon signed-rank test (if non-parametric).

5.4.1 Quiz score. In the quiz with a total score of 26 points, participants in the experimental group using our tool performed better (M = 17.0, SD = 2.8) compared to participants in the control group (M = 13.0, SD = 3.4). The experimental group's scores were significantly higher ($p < 0.001$, statistically highly significant), in both subjective questions ($p = 0.003 < 0.01$, highly significant) and objective questions ($p = 0.003 < 0.01$, highly significant). The results are shown in Figure 4.

The superior performance of the experimental group in the quiz is attributed to the content and format provided by GPT-4. During the experiment, as participants explored knowledge within the tool, they posed questions about the structure and habits of "dragonfly". GPT-4 responded to these inquiries, offering correct, extensively covered, and systematically organized answers. Furthermore, as a large language model, GPT-4's output inherently possesses characteristics of summarization and synthesis, aligning well with the format of our subjective questions. This seemingly facilitated the experimental group's performance in subjective questions. In contrast, participants in the control group had to consider how to search within search engines (keywords, sentence structures, etc.). After obtaining search results, they had to filter information based on the result pages. This made the search process in the control group appear aimless.

Another evident result is that the performance of the experimental group was more stable than that of the control group (SD = 2.8 vs SD = 3.4). Although the interactive explanations generated by GPT-4 vary each time, the knowledge involved in these explanations is often consistent, ultimately resulting in similar learning outcomes for the learners. However, the search tools used by the control group produce entirely different results due to minor variations in the search process, leading to diverse learning content for learners and, consequently, inconsistent performance in the quiz.

5.4.2 Quality and efficiency of reasoning. When analyzing both the experimental and control groups in terms of quality and efficiency (time spent, unit: minutes) on reasoning tasks, the focus was on their mean performance before examining statistical significance through a paired t-test. For quality, the experimental group demonstrated a quality score slightly higher than the control group (M = 3.1 vs M = 3.0). In terms of efficiency, the experimental group completed their tasks notably faster than the control group (M = 5.9 minutes vs M = 9.2 minutes). With regard to the statistical significance, the time efficiency difference was evident: the time efficiency difference between the groups was statistically significant ($p = 0.006<0.05$), indicating the faster completion of the experimental group. When considering the quality of reasoning, even though the experimental group had a slight advantage over the mean score, this difference was not statistically significant ($p = 0.71$). These findings are further detailed in Figure 5.

From the result, it is clear that even within the context of tool assistance, the process of analogy reasoning remains predominantly human-driven, with GPT-4 serving as an assistant to provide hints or feedback. The results of BID reasoning depend on the long-term educational background and accumulated knowledge of learners. For instance, in practical BID education, learners with backgrounds in engineering tend to propose reasoning results within the engineering domain, whereas design background learners are more inclined to imitate external forms. Given this, short-term tool assistance and additional knowledge from GPT-4 may not change the thinking mode of learners, and thus, may not improve the quality of their reasoning instantly. However, our reasoning steps reduce the cognitive burden on learners, making it easier for them to get appropriate applications. Participants mentioned during interviews that the generic questions presented in the tool "helped me identify the questions I want to ask, reduce my cognitive stress, and enhance my efficiency." Consequently, during the reasoning training, participants with GPT-4 as an assistant completed tasks more swiftly, achieving more efficient reasoning.

5.4.3 Consistency between expert and LLMs evaluations. A correlation analysis was conducted to determine the consistency between the evaluations provided by GPT-4 and human experts. The scores from the experts and GPT-4 revealed a strong correlation (The two-way random Intra-Class Correlation Coefficient (ICC) = 0.81 > 0.75). This correlation suggests that within the BIDTrainer framework, GPT-4 is able to perform with human expert scoring patterns, offering insightful feedback on the design schemes of the learners. As a result, learners can gain timely and accurate feedback from GPT-4. These scores are further visualized in Figure 6. More evaluation details are included in the Supplementary Materials.

The scores generated by GPT-4 are generally higher compared to human ratings, as reflected in the higher average values (M = 7.10 vs M = 6.47) and max scores (Max = 9.00 vs Max = 8.40) of the evaluation data. The reason for this pattern could be observed from GPT-4's specific explanation regarding its scoring strategy. At the explanation, it determines if a learner's scheme is aligned with the designated "C-B-I" format, ensuring that both the challenge and biological model match with factual correctness. Once these matches are met, a base score of roughly 6 points is granted. Thereafter, GPT-4 shifts its focus to the quality, novelty, and analogical rationality of the scheme, gradually adjusting the initial score. Conversely, human experts delve directly into the metrics and potential impact of the scheme, leading to potentially wider scoring variances.

Taking the design scheme in Figure 7 as an example, GPT-4 gave a score of 6, with the reason being "this design addresses the challenge

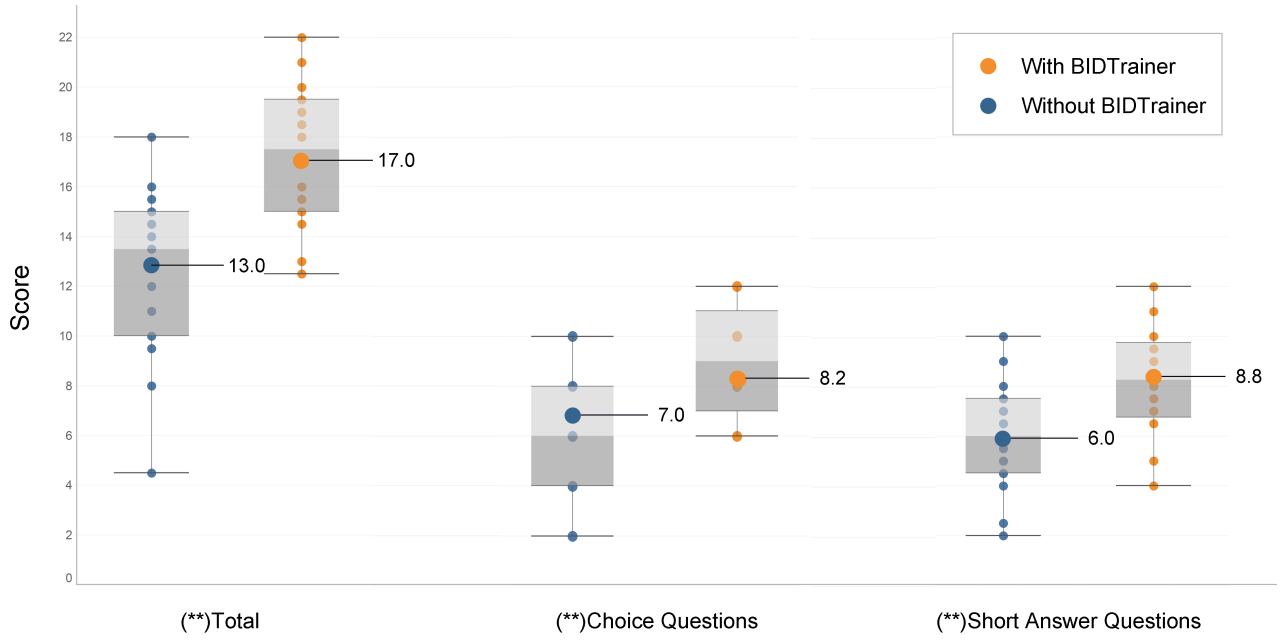


Figure 4: Results of the quiz score. Small dots refer to individual cases, while large dots represent the average value. **, *, and ns indicate significance of p < 0.01, p < 0.05, and p > 0.05, respectively.

of high energy consumption for drones" and "utilizing the dragonfly wing structure is an innovative approach". Regarding the scheme's shortcomings, GPT-4 pointed out that "more specific details on how the vein structure of dragonfly wings needed." In contrast, human experts, who prioritize the quality, novelty, and analogical rationality of the scheme, gave an average score of 4.6, labeling it as "pass". This trend points to GPT-4's pattern of assigning a foundational score even to proposals that may not be as refined, resulting in marginally higher average scores.

5.4.4 Interview analysis. We conducted a qualitative analysis of the data obtained from interviews. Two of the authors employed inductive analysis to code the data, while the other authors reviewed the coding results. According to the interview data and coding results, a notable observation was the ease and independence that learners felt when interacting with the tool. Feedback from these interactions highlighted the supportive role of the tool. Phrases like, "I no longer hesitate to ask basic questions in front of the tool," were common. Our tool was also described as "an encouraging mentor without obstacle." This may be due to our prompts (see Table 2, Prompt1) that induce GPT-4 to provide positive feedback to learners, pushing them to explore further.

Discussing the content, participants emphasized the tool's provision of itemized explanations of biological knowledge (see the detailed outputs of the tool in the Supplementary Materials), often described as "detailed," "comprehensive," and "structured." They noted these explanations break down complex biological concepts into "easily understandable" segments.

Some feedback from the interviews pointed out shortcomings in the content generated by the tool. Two participants mentioned that the answer information generated by the tool was incomplete. For example, if the questions are not asked in the order of "source -> benefits -> application", the tool may overlook the relevant information. In fact, while the answers generated by GPT-4 were generally insightful, they could occasionally miss niche knowledge in specialized domains.

6 HALLUCINATION EVALUATION

As LLMs, including GPT-4 can generate hallucinations, such as incorrect or nonsensical information [39], we examined its potential impact on BID education. Specifically, we analyzed two types of GPT-4 generated textual answers in our tool: knowledge explanations and reasoning assistance. The other support provided by GPT-4, scoring, has already been verified in the experiment about evaluation and will not be repeated here.

6.1 Method

6.1.1 Test data collection. To assess GPT-4's application in knowledge explanations, we collected test data (N = 48). For design approaches, we covered both problem-driven and the solution-driven approaches, specifically represented as the "A -> B -> S" and "S -> B -> A" links, respectively. For each approach, we extracted two links from the "innovation" section in AskNature (total of four links). For each link, we selected three question domains covered by the "guided student-generated questioning strategy" in section 3.2: biology, biology & engineering, and engineering, with six, four,

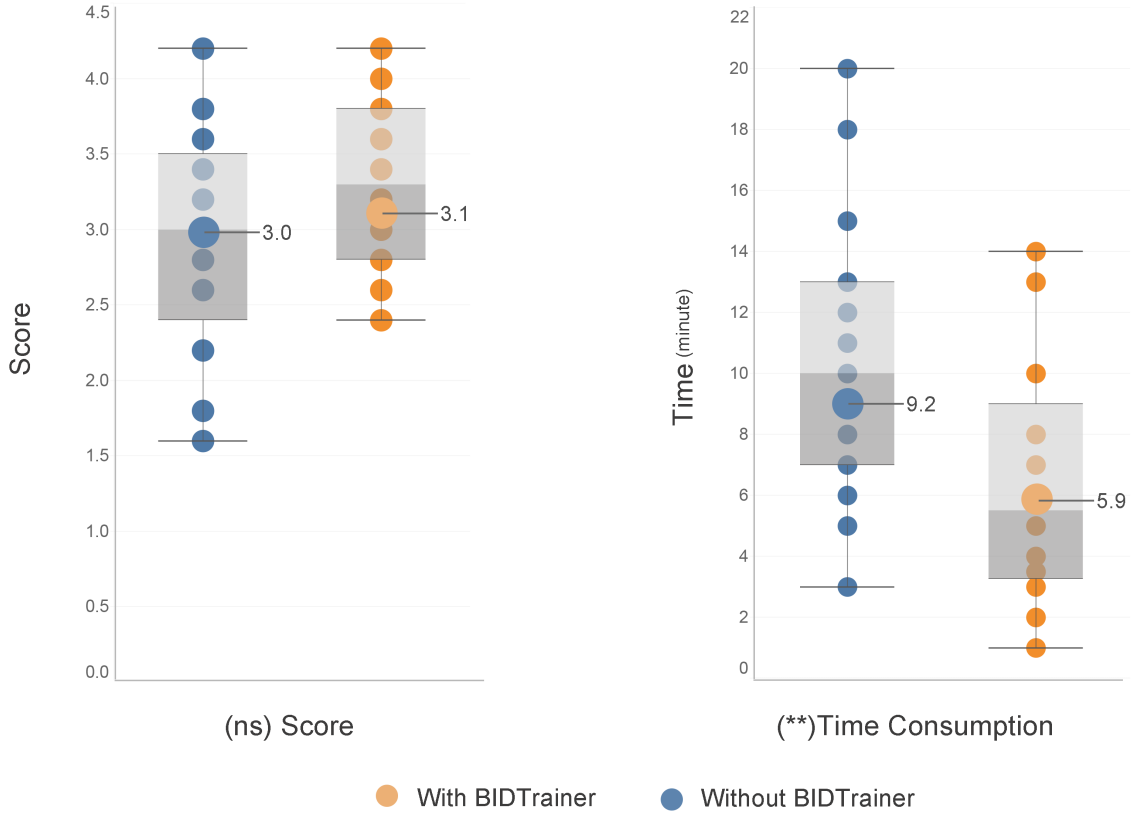


Figure 5: The score and time of application reasoning. Small dots refer to individual cases, while large dots represent the average value. **, *, and ns indicate significance of p < 0.01, p < 0.05, and p > 0.05, respectively

and two generic questions, respectively. Learners provided specific questions and guesses to generate answers, simulating real tool interactions, resulting in 48 QA pairs.

To evaluate GPT-4's application in reasoning assistance, we collected test data (N = 32). We extracted four "S -> B -> A" links from AskNature's "innovations" section and modified them to "S -> B" links by concealing the "application" aspect. For each link, we chose four steps covered by the reasoning steps outlined in section 3.2. For each step, two types of reasoning assistance were analyzed, which included feedback and hints, resulting in 32 QA pairs.

6.1.2 Procedure. Inspired by human evaluation methods commonly used in hallucination assessment for generative question answering [27], we employed four human evaluators with PhDs in design or biology (three majored in design and one majored in biology) to score QA pairs. They looked at questions posed by learners as well as the answers of GPT-4, rating each answer on a 5-point Likert scale for factual correctness which is widely used in measuring the faithfulness of the generated answer reflecting its hallucination [74]. To ensure accuracy, they are encouraged to conduct fact-checks using external sources. Each evaluator reviewed 48 knowledge explanation QA pairs and 32 reasoning assistance pairs, dedicating around 90 and 60 minutes respectively, with a compensation of 170 CNY. The overall inter-annotator agreement on these metrics was considered fair (Intra-Class Correlation Coefficient (ICC) = 0.411).

6.2 Results

In short, the hallucination evaluation demonstrates that GPT-4's answers exhibit a high level of factual correctness, as rated on a 5-point Likert scale. The scoring results of the generated answers are detailed in Table 5. Employing the qualitative scoring methodology described in Han [17], GPT-4's factual correctness in both knowledge explanations (score = 4.07) and reasoning assistance (score = 4.27) is rated as "excellent" (scores ranging from 4 to 5). Additionally, the hallucination rate, indicating the percentage of answers scoring below 3 on a 5-point scale where scores below 3 represent partially or entirely incorrect answers, is listed. Notably, the hallucination rates in both knowledge explanations (9.38%) and reasoning assistance (6.25%) are below 10%, indicating that a minority of answers are factually incorrect. Consequently, the performance of GPT-4 in generating correct knowledge explanations and providing reasoning assistance is considered competent for educational purposes in BID.

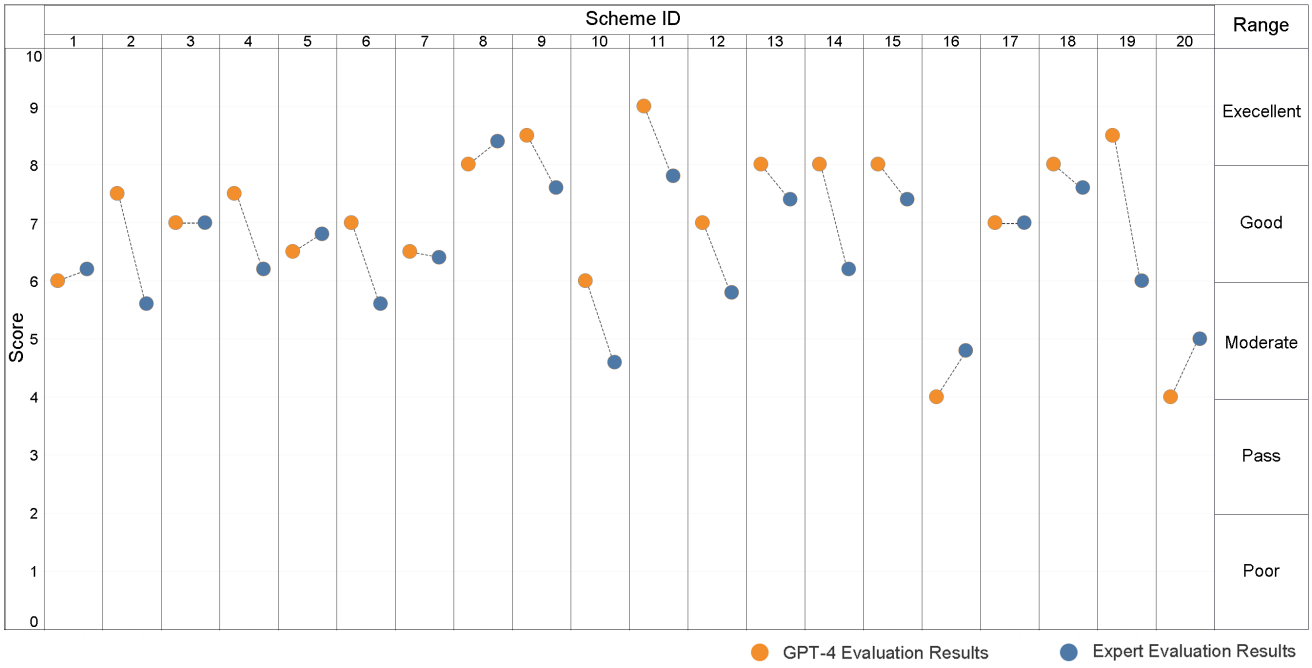
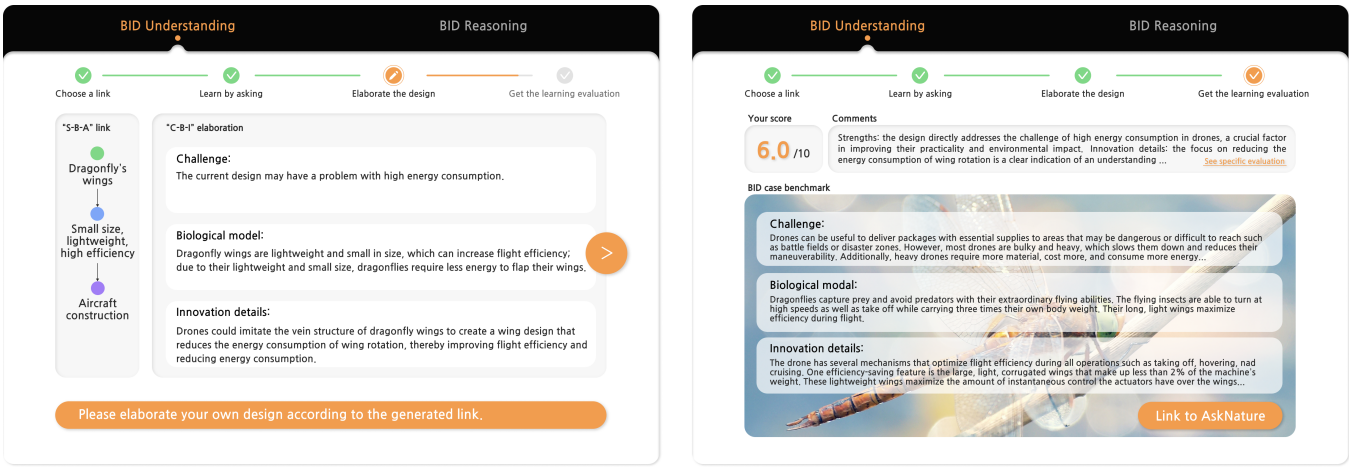


Figure 6: Results of scoring. The intervals corresponding to different score ranges are indicated on the right side of the image.



(1) Design scheme elaborated by a learner

(2) Learning evaluation generated by the tool

Figure 7: An example of a design scheme proposed by a learner, accompanied by a learning evaluation generated by the tool

However, as Figure 8 illustrates, the factual correctness score is significantly low in the cross-disciplinary field of "biology & engineering" within knowledge explanations. For instance, in a QA pair that all four evaluators scored at or below three points: GPT-4's response to the question, "How can the low permeability of cork oak be used in the design of gas and liquid valves?" mentioned the feasibility of using a gas layer as a seal to prevent gas or liquid leakage. As evaluator 2 pointed out, gas as a sealing material is not stable enough for high-standard engineering applications like valves, making it impractical for such uses. The hallucination issue becomes more severe when answering questions in the field of "biology & engineering" due to the nature of cross-disciplinary queries that often need reasoning capabilities and rely on experiential knowledge, extending beyond mere factual recall.

To reduce the impact of hallucinations on learning, we have implemented a set of tips for users interacting with GPT-4 in our tool. Before using the tool, learners are informed about the potential for encountering hallucinations. In addition, during the usage of

Table 5: Scoring results for factual correctness of answers generated by GPT-4 in knowledge explanations and reasoning assistance.

	Knowledge explanations	Reasoning assistance
Evaluator 1	4.25	4.69
Evaluator 2	4.04	4.41
Evaluator 3	4.50	4.31
Evaluator 4	3.21	3.31
Mean	**4.07**	**4.27**
Hallucination rate	**9.38%**	**6.25%**

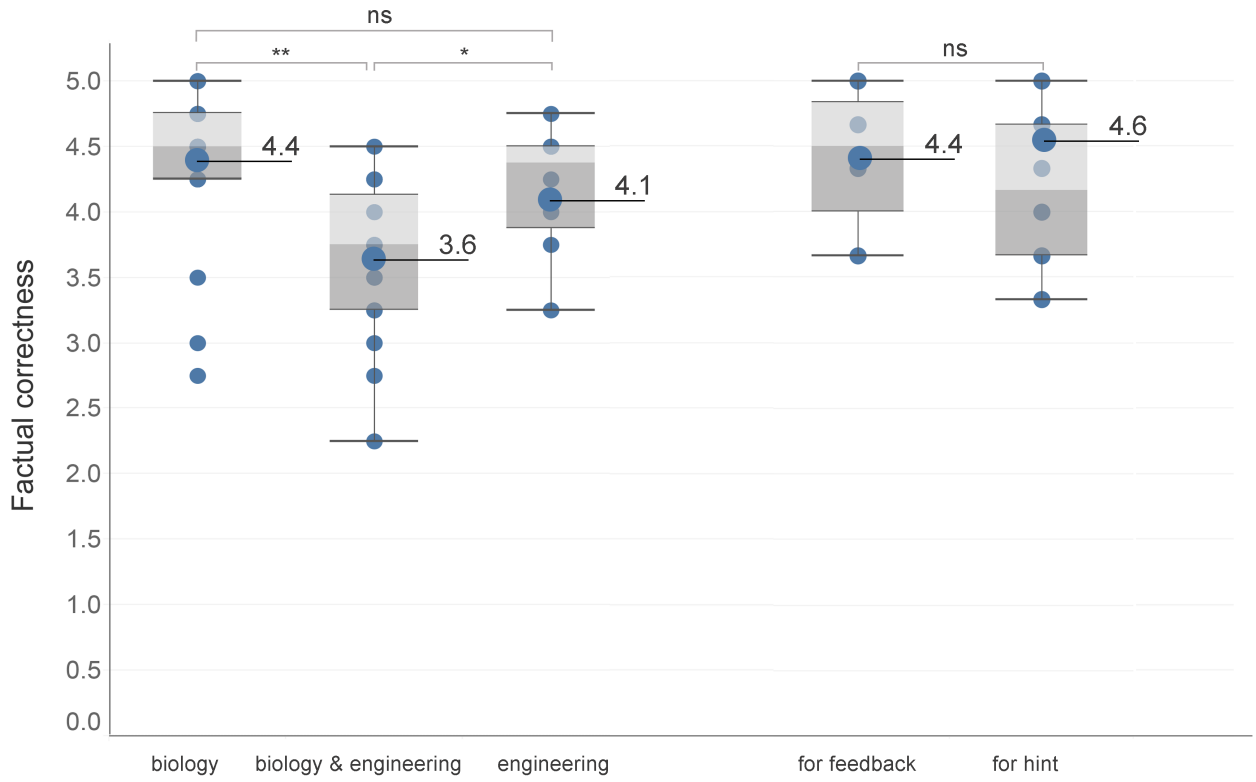


Figure 8: The results of factual correctness scores in hallucination evaluation, categorized by the types of QA pairs. **, *, and ns indicate significance of $p < 0.01$, $p < 0.05$, and $p > 0.05$, respectively

the tool, we encourage the use of Chain-of-Thought prompting techniques. Furthermore, we also established a feedback mechanism for GPT-4 generated answers.

7 DISCUSSION

In this section, we discuss how our tool can potentially improve collaborative learning in BID. We also explore the broader implementations of our LLMs-driven BID education tool, its applicability beyond the field of BID and the educational contexts. Furthermore, the limitations and future prospects of our approach are described.

7.1 LLMs-driven BID education tool for learner teams

When co-learners and teachers are unavailable, our tool can assist learners in collaborative learning for BID, providing a more meaningful learning experience. BID often entails cross-disciplinary collaboration, where experts from diverse fields collaborate to study biological systems, extract principles, and apply them to design solutions [66]. When learners collaborate remotely with co-learners, the tool can assist in understanding BID knowledge, practicing BID reasoning, and offer precise assessment feedback through interaction in the platform. Our tool can potentially enable learners to train their BID skills through collaboration whenever they require

it, which can also enhance their communication and cooperation abilities. This provides an effective case of human-LLMs collaboration that demonstrates cooperation among humans as well as between human and LLMs.

7.2 Beyond BID and Beyond education

While our tool primarily focuses on BID education, we believe that our education method can be generalized to support other areas of design education. We have introduced a BID education method driven by LLMs. This method supports case learning, cognitive training, and evaluation benchmarked by cases. The conceptual foundation of this education method, which combines case learning with cognitive training and provides comparative evaluation, is also applicable to various design education domains that require the accumulation of cases and cognitive skills.

Our tool can significantly enhance the efficiency of understanding cases and reasoning while providing interactivity. Beyond supporting BID education, this functionality can be extended to enable more efficient BID. In the context of design practice, our tool can efficiently facilitate the understanding of BID cases and analogical reasoning through knowledge explanations and reasoning assistance. Additionally, through interactions between users and LLMs, such as the implementation of the "guided student-generated questioning strategy" for knowledge explanations, the tool enhances the user experience. While current LLMs-driven BID tools mainly emphasize the efficiency of finding solutions [59], our approach can improve user interactivity and design efficiency through human-LLMs collaboration.

7.3 Limitations and Future Work

Our work is based on LLMs which demonstrate robust generalization capabilities and natural language understanding ability [37]. However, they may also exhibit hallucination phenomena, wherein outputs do not match the factual realities[28]. Such hallucinations typically arise when LLMs are queried about topics not covered in their training data [40] or when there are errors in the input data [50]. We have decreased the likelihood of hallucination effects by utilizing the model with better performance (GPT-4), but as shown in our evaluations, we still observed that the outputs of GPT-4 occasionally contain minor factual inaccuracies. Additionally, while our user study has indicated a high degree of consistency between GPT-4 and human experts, a potential risk persists when relying on results generated by GPT-4 for tasks that require a high level of reliability, such as evaluating learners' designs. To mitigate this, future work could focus on various methods to prevent potential hallucinations. One approach could involve deploying real-time textual semantic detection models to identify hallucinations in LLM outputs. Additionally, incorporating feedback from experts in the BID field could be crucial in refining GPT training. This feedback would be instrumental in enhancing the quality of generated content and could play a significant role in safeguarding learners from the long-term impact of hallucination effects.

Besides, our tool is primarily designed to assist in learning BID knowledge and methods, enabling learners to develop conceptual design. However, its utility is somewhat limited when it comes to the subsequent stages of engineering practice. For example, while learners can generate innovative concepts in textual format, the tool does not currently support the translation of these ideas into sketches or models. Additionally, it lacks features for conducting simulations, usability testing, and transforming BID concepts into tangible products. As an educational tool, our focus has been more on the initial stages of conceptualization rather than providing a comprehensive platform for all stages of BID practice. Future work could explore and integrate advanced technologies and theories, such as leveraging LLMs and diffusion models. These enhancements could potentially support learners not only in the creation of BID concepts but also in testing, refining, and realizing these concepts in more practical, applicable forms.

8 CONCLUSION

In this research, we introduced an LLMs-driven BID education method, aiming to tackle the multidisciplinary challenges in enhancing BID understanding and reasoning. As a practical application of this methodology, we developed BIDTrainer, an LLMs-driven tool designed to facilitate interactive learning in BID. Our findings show BIDTrainer's effectiveness in enhancing learners' understanding and reasoning skills in BID, particularly in its ability to align closely with teacher evaluations. BIDTrainer exemplifies how LLMs can be integrated into educational tools to foster both interactive and efficient learning. As LLMs continue to evolve, their integration into tools like BIDTrainer will be instrumental in advancing BID education.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant No. 2022YFB3303304, the National Natural Science Foundation of China under Grant No. 62207023, and the ZJU-SUTD IDEA Grant from The Ng Teng Fong Charitable Foundation (Grant No. 188170-11102).

REFERENCES

[1] Hamzeh Alabool. 2023. ChatGPT in Education: SWOT analysis approach. *2023 International Conference on Information Technology (ICIT)* (2023), 184–189. https://doi.org/10.1109/ICIT58056.2023.10225801

[2] Lisa B. Bosman and Katherine L. Shirey. 2023. Using STEAM and Bio-Inspired Design to Teach the Entrepreneurial Mindset to Engineers. *Open Education Studies* 5 (2023). https://doi.org/10.1515/edu-2022-0187

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[4] Canva. 2023. How Huntington Beach Union High School District Teaches Its Students Real-World Design Skills. https://www.canva.com/education/case-studies/huntington-beach-union-high-school-district/. Accessed: 2023-09-01.

[5] Liuqing Chen, Zebin Cai, Zhaojun Jiang, Qi Long, Lingyun Sun, Peter R. N. Childs, and Haoyu Zuo. 2023. A KNOWLEDGE-BASED IDEATION APPROACH FOR BIO-INSPIRED DESIGN. *Proceedings of the Design Society* 3 (2023), 231 – 240. https://doi.org/10.1017/pds.2023.24

[6] Hyunmin Cheong and L. H. Shu. 2013. Using templates and mapping strategies to support analogical transfer in biomimetic design. *Design Studies* 34 (2013), 706–728. https://doi.org/10.1016/j.destud.2013.02.002

[7] Berlyne De. 1966. Conditions of prequestioning and retention of meaningful material. *Journal of Educational Psychology* 57 (1966), 128–132. https://doi.org/10.1037/H0023346

[8] Jon-Michael Deldin and Megan Schuknecht. 2014. *The AskNature Database: Enabling Solutions in Biomimetic Design.* Springer London, London, 17–27. https://doi.org/10.1007/978-1-4471-5248-4_2

[9] Ramon Dijkstra, Zülküf Genç, Subhradeep Kayal, Jaap Kamps, et al. 2022. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers.

[10] Fabien Durand, Michael Helms, Joanna Tsenn, Erin M. McTigue, Daniel A. McAdams, and Julie S. Linsey. 2015. Teaching Students to Innovate: Evaluating Methods for Bioinspired Design and Their Impact on Design Self Efficacy *(International Design Engineering Technical Conferences and Computers and Information in Engineering Conference)*. https://doi.org/10.1115/DETC2015-47716

[11] Marjan José Eggermont. 2018. Bio-inspired Design and Information Visualization. *Graduate Studies Science* (2018).

[12] Fernando Ferraretto, Thiago Laitz, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2023. ExaRanker: Synthetic Explanations Improve Neural Rankers. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2023). https://doi.org/10.1145/3539618.3592067

[13] Rahel Flechtner and Aeneas Stankowski. 2023. AI Is Not a Wildcard: Challenges for Integrating AI into the Design Curriculum. *Proceedings of the 5th Annual Symposium on HCI Education* (2023). https://doi.org/10.1145/3587399.3587410

[14] Katherine K. Fu, Diana P. Moreno, Maria C. Yang, and Kristin L. Wood. 2014. Bio-Inspired Design: An Overview Investigating Open Questions From the Broader Field of Design-by-Analogy. *Journal of Mechanical Design* 136 (2014), 111102. https://doi.org/10.1115/1.4028289

[15] Xiaoyi Gao, Pei qi Li, Ji Shen, and Huifang Sun. 2020. Reviewing assessment of student learning in interdisciplinary STEM education. *International Journal of STEM Education* 7 (2020), 1–14. https://doi.org/10.1186/s40594-020-00225-4

[16] Sharon Gedye. 2010. Formative assessment and feedback: a review. *Planet* 23 (2010), 40 – 45. https://doi.org/10.11120/plan.2010.00230040

[17] Ji Han, Feng Shi, Liuqing Chen, and Peter R. N. Childs. 2018. A computational tool for creative idea generation based on analogical reasoning and ontology. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 32 (2018), 462 – 477. https://doi.org/10.1017/S0890060418000082

[18] John A. C. Hattie and Helen S. Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77 (2007), 112 – 81. https://doi.org/10.3102/003465430298487

[19] M Helms. 2019. Challenges for BID in Industry. In *Proceedings of the NASA VINE Tools Workshop, Cleveland, OH, USA*, Vol. 9.

[20] Michael E. Helms and Ashok K. Goel. 2014. The Four-Box Method: Problem Formulation and Analogy Evaluation in Biologically Inspired Design. *Journal of Mechanical Design* 136 (2014), 111106. https://doi.org/10.1115/1.4028172

[21] Michael E. Helms, Swaroop Vattam, and Ashok K. Goel. 2009. Biologically inspired design: process and products. *Design Studies* 30 (2009), 606–622. https://doi.org/10.1016/J.DESTUD.2009.04.003

[22] Delia Hillmayr, Lisa Ziernwald, Frank Reinhold, Sarah I. Hofer, and Kristina M. Reiss. 2020. The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education* 153 (2020), 103897. https://doi.org/10.1016/j.compedu.2020.103897

[23] Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. 2023. Instruction Induction: From Few Examples to Natural Language Task Descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1935–1952. https://doi.org/10.18653/v1/2023.acl-long.108

[24] Stanley Hunley, Joshua Whitman, Seungik Baek, Xiaobo Tan, and Drew Kim. 2010. Incorporating The Importance of Interdisciplinary Understanding in K 12 Engineering Outreach Programs Using A Biomimetic Device. In *2010 Annual Conference & Exposition*. 15–715.

[25] Shoshanah Jacobs, Marjan Eggermont, Michael Helms, and Kristina Wanieck. 2022. The Education Pipeline of Biomimetics and Its Challenges. *Biomimetics* 7, 3 (2022), 93.

[26] Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies* (2023), 1–20.

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55 (2022), 1 – 38. https://doi.org/10.1145/3571730

[28] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169* (2023).

[29] Sumbul Khan, Lucienne Blessing, and Yakhoub Ndiaye. 2023. Artificial intelligence for competency assessment in design education: a review of literature. In *International Conference on Research into Design*. Springer, 1047–1058.

[30] Byoung Soo Kim, Min Ku Kim, Younghak Cho, Eman E. Hamed, Martha U. Gillette, Hyeongyun Cha, Nenad Miljkovic, Vinay Kumar Aakalu, Kai Kang, Kyung-No Son, Kyle M. Schachtschneider, Lawrence B. Schook, Chenfei Hu, Gabriel Popescu, Yeonsoo Park, William C. Ballance, Seunggun Yu, Sung-Gap Im, Jonghwi Lee, Chi Hwan Lee, and Hyunjoon Kong. 2020. Electrothermal soft manipulator enabling safe transport and handling of thin cell/tissue sheets and bioelectronic devices. *Science Advances* 6 (2020). https://doi.org/10.1126/sciadv.abc5630

[31] Jin Woo Kim, Daniel A. McAdams, and Julie S. Linsey. 2014. Helping students to find biological inspiration: Impact of valuableness and presentation format. *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (2014), 1–6. https://doi.org/10.1109/FIE.2014.7044029

[32] Alison King. 1992. Facilitating Elaborative Learning Through Guided Student-Generated Questioning. *Educational Psychologist* 27 (1992), 111–126. https://doi.org/10.1207/S15326985EP2701_8

[33] Paula Kirya, Eric Chen, Marina Achterman, Kelli Eugenio, Zekaria Beshir, Natalie Ngoy, Radwanul Hasan Siddique, Atilla Ozgur Cakmak, and Jared Ashcroft. 2021. Biomimicry of Blue Morpho butterfly wings: An introduction to nanotechnology through an interdisciplinary science education module. *Journal of the Society for Information Display* 29 (2021), 896 – 915. https://doi.org/10.1002/jsid.1071

[34] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

[35] Melissa M. Lacey and David P. Smith. 2023. Teaching and assessment of the future today: higher education and AI. *Microbiology Australia* (2023). https://doi.org/10.1071/ma23036

[36] Ehsan Latif and Xiaoming Zhai. 2023. Fine-tuning ChatGPT for Automatic Scoring. *ArXiv* abs/2310.10072 (2023). https://doi.org/10.48550/arXiv.2310.10072

[37] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[38] Shijian Luo, Ze Bian, and Yuqi Hu. 2020. How can biological shapes inspire design activity in closed domains? *International Journal of Technology and Design Education* 32 (2020), 479–505. https://doi.org/10.1007/s10798-020-09593-y

[39] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[40] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. *arXiv preprint arXiv:2305.14552* (2023).

[41] Ishan Misra, Ross B. Girshick, Rob Fergus, Martial Hebert, Abhinav Kumar Gupta, and Laurens van der Maaten. 2017. Learning by Asking Questions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 11–20. https://doi.org/10.1109/CVPR.2018.00009

[42] Steven Moore, Huy A Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the quality of student-generated short answer questions using GPT-3. In *European conference on technology enhanced learning*. Springer, 243–257.

[43] Jacquelyn KS Nagel and Ramana M Pidaparti. 2016. Significance, prevalence and implications for bio-inspired design courses in the undergraduate engineering curriculum. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 50138. American Society of Mechanical Engineers, V003T04A009.

[44] Jacquelyn K Nagel, Peyton Pittman, Ramana Pidaparti, Chris Rose, and Cheryl Beverly. 2016. Teaching bioinspired design using C–K theory. *Bioinspired, Biomimetic and Nanobiomaterials* 6, 2 (2016), 77–86.

[45] Jacquelyn K. S. Nagel, Christopher Stewart Rose, Cheri Beverly, and Ramana M. Pidaparti. 2019. Bio-inspired Design Pedagogy in Engineering. *Design Education Today* (2019). https://doi.org/10.1007/978-3-030-17134-6_7

[46] Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2023. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. *ArXiv* abs/2303.17003 (2023). https://doi.org/10.48550/arXiv.2303.17003

[47] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774

[48] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* (2023), 101356. https://doi.org/10.13140/RG.2.2.32393.31840

[49] Ramana M. Pidaparti and Jacquelyn K. S. Nagel. 2018. T3-A: C-K Theory-based Bio-inspired Projects in Sophomore Design Course. https://api.semanticscholar.org/CorpusID:67247707

[50] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768* (2023).

[51] Red Dot. 2020. Deployable Emergency Shelter Nominated for the Red Dot Luminary at the Red Dot Award: Design Concept 2020. https://www.red-dot.org/magazine/deployable-emergency-shelter-nominated-for-the-red-dot-luminary-at-the-red-dot-award-design-concept-2020 Accessed:

2023-12-12.
[52] Royce Sadler. 1998. Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice* 5 (1998), 77–84. https://doi.org/10.1080/0969595980050104
[53] Carlo Santulli and Carla Langella. 2011. Introducing students to bio-inspiration and biomimetic design: a workshop experience. *International Journal of Technology and Design Education* 21 (2011), 471–485. https://doi.org/10.1007/s10798-010-9132-6
[54] Jami J. Shah, Noé Vargas-Hernández, and Steve M. Smith. 2003. Metrics for measuring ideation effectiveness. *Design Studies* 24 (2003), 111–134. https://doi.org/10.1016/S0142-694X%2802%2900034-0
[55] L. Siddharth and Amaresh Chakrabarti. 2018. Evaluating the impact of IdeaInspire 4.0 on analogical transfer of concepts. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 32 (2018), 431 – 448. https://doi.org/10.1017/S0890060418000136
[56] Olga Speck and Thomas Speck. 2021. Biomimetics and Education in Europe: Challenges, Opportunities, and Variety. *Biomimetics* 6 (2021). https://doi.org/10.3390/biomimetics6030049
[57] Laura Stevens, Marc J. de Vries, Mark J.W. Bos, and Helen Kopnina. 2019. Biomimicry Design Education Essentials. *Proceedings of the Design Society: International Conference on Engineering Design* (2019). https://doi.org/10.1017/dsi.2019.49
[58] Laura Stevens, Helen Kopnina, K. F. Mulder, and Marc J. de Vries. 2020. Biomimicry design thinking education: a base-line exercise in preconceptions of biological analogies. *International Journal of Technology and Design Education* (2020), 1–18. https://doi.org/10.1007/s10798-020-09574-1
[59] Feng Sun, He Xu, Yihan Meng, and Zhimao Lu. 2022. A BERT-based model for coupled biological strategies in biomimetic design. *Neural Computing and Applications* 35 (2022), 2827–2843. https://doi.org/10.1007/s00521-022-07734-z
[60] Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*. 522.
[61] Tianran Tang, Pengfei Li, and Qiheng Tang. 2022. New Strategies and Practices of Design Education Under the Background of Artificial Intelligence Technology: Online Animation Design Studio. *Frontiers in Psychology* 13 (2022). https://doi.org/10.3389/fpsyg.2022.767295
[62] Swaroop Vattam and Ashok K. Goel. 2011. Foraging for Inspiration: Understanding and Supporting the Online Information Seeking Practices of Biologically Inspired Designers. https://doi.org/10.1115/DETC2011-48238
[63] Swaroop Vattam, Bryan Wiltgen, Michael E. Helms, Ashok K. Goel, and Jeannette Yen. 2011. DANE: Fostering Creativity in and through Biologically Inspired Design. https://doi.org/10.1007/978-0-85729-224-7_16
[64] Julian F. V. Vincent. 2009. Biomimetics — a review. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 223 (2009), 919 – 939. https://doi.org/10.1243/09544119JEIM561
[65] Julian F. V. Vincent, Olga Bogatyreva, Nikolaj Bogatyrev, Adrian Bowyer, and Anja-Karina Pahl. 2006. Biomimetics: its practice and theory. *Journal of The Royal Society Interface* 3 (2006), 471 – 482. https://doi.org/10.1098/rsif.2006.0127
[66] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Weirong Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xingxu Xie. 2023. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *ArXiv* abs/2302.12095 (2023). https://doi.org/10.48550/arXiv.2302.12095
[67] Wenya Wang, Vivek Srikumar, Hannaneh Hajishirzi, and Noah A. Smith. 2023. Elaboration-Generating Commonsense Question Answering at Scale. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1619–1635. https://doi.org/10.18653/v1/2023.acl-long.90
[68] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
[69] Mart Willocx, Amir Ayali, and Joost R. Duflou. 2021. Reprint of: Where and how to find bio-inspiration?: A comparison of search approaches for bio-inspired design. *Cirp Journal of Manufacturing Science and Technology* (2021). https://doi.org/10.1016/J.CIRPJ.2021.06.005
[70] Dietmar H. Wittmann, Charles Aprahamian, and Jack M. Bergstein. 1990. Etappenlavage: Advanced diffuse peritonitis managed by planned multiple laparotomies utilizing zippers, slide fastener, and Velcro® analogue for temporary abdominal closure. *World Journal of Surgery* 14 (1990), 218–226. https://doi.org/10.1007/BF01664876
[71] Jeannette Yen, Michael E. Helms, Ashok K. Goel, Craig A. Tovey, and Marc J. Weissburg. 2014. Adaptive Evolution of Teaching Practices in Biologically Inspired Design. https://doi.org/10.1007/978-1-4471-5248-4_7
[72] Jeannette Yen and Marc J. Weissburg. 2012. Biologically Inspired Design : A Tool for Interdisciplinary Education.
[73] Ibrahim H. Yeter, Valerie Si Qi Tan, and Hortense Le Ferrand. 2023. Conceptualization of Biomimicry in Engineering Context among Undergraduate and High School Students: An International Interdisciplinary Exploration. *Biomimetics* 8 (2023). https://doi.org/10.3390/biomimetics8010125
[74] Yuhao Zhang, Derek Merck, Emily B Tsai, Christopher D. Manning, and C. Langlotz. 2019. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1%2F2020.acl-main.458