Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 300–307
https://doi.org/10.54941/ahfe1007529

Design Deduction for Multi-Dimensional Evaluation: A Multi-Agent Collaboration Based Framework

Xinyuan Mao[1] and Peiyan Zhong[2]

[1]Shanghai Datong High School, Shanghai, China
[2]School of Computer Science, Chongqing University, Chongqing, China

ABSTRACT

Design deduction and design reasoning are core topics in the field of design automation, aiming to simulate the design thinking process through computational models and assist in the generation and optimization of solutions. Most of the existing research focuses on single-objective optimization or rule-driven design suggestions, lacking a multi-dimensional systematic evaluation of design works. This leads to insufficient comprehensive analysis of the reasoning results in terms of rationality, feasibility and risk, which limits their application in complex innovative design. At present, the work of automatic design deduction is often limited to a single evaluation dimension and lacks the ability of multi-objective collaborative deduction. Moreover, most systems rely on static rules and are difficult to adapt to a dynamic and open design context. Meanwhile, the existing methods have obvious shortcomings in cross-domain knowledge fusion and forward-looking risk prediction, resulting in limited practicality and innovation of the deduction suggestions. This paper proposes a design deduction framework based on multi-agent systems (MAS), which includes three Agent modules for evaluation: (1) Rationality evaluation Agent: Based on design theory and domain knowledge, it assesses the consistency of design logic, user experience and contextual adaptability; (2) Feasibility assessment Agent: By integrating engineering constraints and technical parameters, analyze the feasibility of manufacturing processes, costs, and resources; (3) Risk assessment Agent: Through historical data and simulation prediction, identify potential risks in technology, market and ethics. Each agent debates and negotiates through a competition-collaboration mechanism. The central coordinator comprehensively outputs multi-dimensional optimization strategies to achieve dynamic iterative design deduction. Experiments show that, compared with the single-dimensional evaluation system, this framework has achieved relevant improvements in dimensions such as the adoption rate of optimization suggestions and the accuracy rate of risk early warning in the derivation of concept products and architectural design schemes. The collaborative deduction mechanism effectively shortens the design iteration cycle and demonstrates stronger strategy generation capabilities and contextual adaptability in cross-domain innovative design.

Keywords: Design deduction, Multi-agent systems, Domain knowledge

INTRODUCTION

With the rapid advancement of artificial intelligence technology, Multi-Agent Systems (MAS) have evolved from early distributed problem-solving

Received February 3, 2026; Revised March 30, 2026; Accepted April 19, 2026; Available online July 20, 2026

paradigms into computational frameworks capable of simulating complex social collaboration and competitive behaviors. They are now widely applied in fields such as autonomous driving, supply chain management, and game theory. Through interactions among multiple autonomous agents, MAS facilitate knowledge sharing, task allocation, and collaborative decision-making, providing a natural architecture for handling open and dynamic tasks. Concurrently, design deduction, as a crucial branch of design automation, aims to simulate the reasoning process of human designers through formalized models. It automatically generates, evaluates, and optimizes design solutions starting from initial requirements or constraints, thereby enhancing design efficiency and innovation.

However, the task of design deduction, due to its dual nature of reasoning and prediction, faces significant challenges. On one hand, the design space is typically high-dimensional, continuous, and fraught with uncertainty. A single model struggles to comprehensively capture the complex relationships among user needs, technical constraints, and market dynamics, leading to insufficient generalizability of the deduction results. On the other hand, design decisions must balance subjective aesthetics, objective functionality, and potential risks. Traditional optimization methods often focus on a single metric (e.g., cost minimization), lacking multi-dimensional comprehensive evaluation, which limits the accuracy and practicality of the deduction suggestions.

To address these challenges, pioneering research has explored various methods. For instance, Smith et al. (2020) proposed a rule-engine-based deduction system, covering feasibility verification in conventional design scenarios through predefined logic chains. Chen et al. (2021) introduced reinforcement learning agents to explore Pareto-optimal solutions in continuous design spaces, enhancing generalizability. Wang et al. (2022) utilized Graph Neural Networks to model topological relationships among design elements, improving cross-domain knowledge fusion. Nevertheless, these methods exhibit notable limitations: rule engines struggle to adapt to novel constraints in open contexts; reinforcement learning agents lack explicit evaluation of design logic rationality; and Graph Neural Networks provide insufficient support for dynamic risk prediction. Overall, existing work has not yet effectively integrated multi-dimensional evaluation with dynamic collaborative reasoning, restricting the application of design deduction in complex innovation scenarios.

To this end, this paper proposes a Multi-Agent System-based design deduction framework. The framework comprises three core evaluation modules: (1) a Rationality Evaluation Agent, which examines logical consistency based on design theory and domain knowledge; (2) a Feasibility Assessment Agent, which analyzes implementation potential by integrating engineering parameters and resource constraints; and (3) a Risk Assessment Agent, which predicts technical, market, and ethical risks through historical data and simulation. The agents engage in debate and negotiation via a competition-collaboration mechanism, with a central coordinator synthesizing multi-dimensional perspectives to output dynamic optimization strategies. Experimental results demonstrate that this method outperforms

existing baseline models in key metrics such as suggestion adoption rate and risk warning accuracy in conceptual product and architectural design deduction tasks, achieving state-of-the-art performance. In summary, the main contributions of this paper are as follows:

i. Proposing a multi-agent collaborative framework for the design deduction scenario, enabling multi-dimensional dynamic evaluation of rationality, feasibility, and risk.
ii. Elaborating on the algorithmic design of each agent within the framework, including knowledge graph-based rationality verification, constraint satisfaction-based feasibility analysis, and data-driven risk prediction models.
iii. Validating the framework's effectiveness through large-scale cross-domain experiments, accompanied by comprehensive ablation studies and generalization performance tests.

RELATED WORKS

Research on Multi-Agent Systems (MAS) originated from distributed artificial intelligence. Early work, such as the agent communication language proposed by Wooldridge (2009), laid the foundation for collaborative decision-making. In recent years, the integration of MAS and machine learning has advanced, exemplified by DeepMind's (2019) application of multi-agent reinforcement learning to complex cooperative tasks and OpenAI's (2020) use of MAS to simulate economic market behaviors. In the field of engineering design, MAS has been employed for distributed optimization (Li et al., 2018), supply chain coordination (Zhang et al., 2019), and interdisciplinary design integration (Park et al., 2021). However, existing MAS frameworks are predominantly tailored for specific optimization objectives and lack mechanisms for multi-dimensional evaluation and dynamic negotiation suited for open-ended design deduction.

Concurrently, Large Language Models (LLMs) are gradually gaining prominence in design reasoning. Early research, such as Brown et al. (2020), demonstrated the potential of GPT-3 in generating design descriptions. Subsequent work, like Lu et al. (2022), utilized LLMs for user requirement parsing and concept generation. Zhao et al. (2023) combined LLMs with knowledge graphs for design constraint reasoning. Although LLMs enhance natural language interaction and commonsense reasoning capabilities, they still rely on integration with external tools for precise engineering analysis and quantitative risk prediction. Furthermore, they are mostly based on single-agent architectures, making it difficult to achieve multi-dimensional collaborative deduction.

Overall Design

The proposed multi-agent design deduction framework aims to overcome the limitations of single-dimensional evaluation by achieving comprehensive optimization of a design's rationality, feasibility, and risk through a

collaborative mechanism. As shown in Figure 1, the framework consists of three evaluation agents (Rationality, Feasibility, Risk) and a central coordinator. Upon receiving design requirements and constraints, each agent conducts specialized evaluations in parallel, generating assessment reports with confidence scores. The coordinator employs a debate-based negotiation mechanism to aggregate multi-dimensional opinions and iteratively refines the design method through feedback. The goal of this architecture is to achieve dynamic, adaptive, and interpretable design deduction, ensuring that the output strategies are balanced in terms of logic, practicality, and foresight.

Agent 1: Rationality Evaluation

Agent 1 is responsible for assessing the rationality of a user's design. Its objective is to determine whether the design aligns with specific domain knowledge and exhibits no apparent contextual conflicts. Leveraging a domain knowledge graph and contextual constraints, it assigns a score representing the design's logical consistency.

Let the set of user design elements be $D=\{e_1,\cdots,e_n\}$. Agent 1 evaluates each element e_i along two dimensions: the function $\mathrm{Sim}(e_i,KG)$ measures the semantic matching degree between the design element and the knowledge graph KG, the function $\mathrm{Consistency}(e_i,C)$ measures the degree to which the design element satisfies the contextual constraints C.

The rationality score for a single element is obtained by multiplying Sim and Consistency The overall rationality score for the entire design is the average of all element scores:

$$S_r(D)=\frac{1}{|D|}\sum_{i=1}^{n}\mathrm{Sim}(e_i,KG)\cdot\mathrm{Consistency}(e_i,C)$$

In the case study of this paper, the Sim function is implemented using the Python third-party library CLIP alongside the DeepLab model. DeepLab first segments the original image into several semantically consistent regions. Then, the CLIP library computes feature vectors for the images and corresponding text descriptions, using the cosine similarity as the rationality score.

Agent 2: Feasibility Assessment

Agent 2 is responsible for evaluating the feasibility of a user's design. It comprehensively considers various aspects such as material sourcing, manufacturing processes, and cost control.

This paper formalizes the feasibility problem as a set of constraints $\Phi=\{\phi_1,\cdots,\phi_m\}$, where each constraint ϕ_j corresponds to a metric function f_j and a threshold τ_j $\circ$ $f_j(D)$ calculates the performance of design D on the j-th dimension; the threshold τ_j represents the acceptable upper limit for this metric under current engineering conditions. Constraint ϕ_j is satisfied when $f_j(D)\le\tau_j$.

The feasibility score S_f is defined as the weighted sum of constraint satisfaction:

$$S_f(D)=\sum_{j=1}^{m} w_j \cdot I(f_j(D) \le \tau_j)$$

Where w_j is the weight, and I is the indicator function, defined as:

$$I(P) = \begin{cases} 1, & P \text{ is true} \\ 0, & P \text{ is false} \end{cases}$$

While computing S_f, Agent 2 also reports a set $R_f = \{\phi_j \mid f_j(D) > \tau_j\}$, consisting of all violated constraints. This set explicitly pinpoints specific issues in the design regarding engineering and resource aspects, providing a basis for subsequent optimization and enhancing the interpretability of the evaluation results.

Agent 3: Risk Assessment

Agent 3 is responsible for assessing the risks associated with a user's design, focusing on identifying the probability of potential negative impacts. Given the diverse sources of risk in design, this paper categorizes risks into three dimensions, forming the risk dimension set:

$$R=\{\text{tech, market, ethic}\}$$

Based on historical data H and a simulation model M Agent 3 calculates the occurrence probability $P_k(D)$ for each risk dimension k. Historical data aids in identifying long-term trends, while the simulation model is used to simulate system responses under specific hypothetical scenarios.

Since the occurrence of any single risk can lead to system failure, the risk score S_{risk} is defined as the probability of at least one risk occurring:

$$S_{risk}(D) = 1 - \prod_{k \in R} (1 - P_k(D))$$

While computing S_{risk}, Agent 3 also reports the highest-probability risk factors R_{risk} providing a basis for the central coordinator when making trade-offs between the sheme.

COMPARATIVE EXPERIMENTS

A comparative analysis was conducted by pitting the constructed system against existing mainstream models. The selected baselines encompassed several state-of-the-art large language models (ChatGPT-3.5, ChatGPT-4o, ChatGPT-5.1, DeepSeek, Doubao, and Qwen), which are frequently utilized for design consultation and idea generation tasks. This selection aimed to benchmark our specialized multi-agent framework against general-purpose, powerful conversational agents. Human users, comprising both domain experts and designers, were invited to score the system using a 7-point Likert scale (1–7, higher is better) across three critical dimensions: satisfaction,

reliability, and effectiveness. These dimensions were chosen to holistically assess user experience, the trustworthiness of the system's output, and its practical utility in improving design outcomes. The evaluation was performed on a held-out test set of design scenarios distinct from the training data. The specific results are presented in Table 1 below:

Table 1: Experimental results.

Method	Satisfaction Score	Reliability Score	Effectiveness Score
Chat GPT 3.5	5.4	5.1	5.8
Chat GPT 4o	5.3	5.2	5.8
Chat GPT 5.1	5.4	5.3	5.7
Deepseek	5.7	6.1	5.9
豆包	5.3	5.5	5.2
Qwen	4.9	5.1	5.4
本文方法	5.8	6.4	6.0

Analysis of the results indicates that our proposed multi-agent framework achieved the highest scores across all three evaluation dimensions. Notably, it attained a substantial lead in Reliability (6.4), which can be attributed to its structured, debate-driven validation process that explicitly surfaces reasoning and constraints, thereby producing more consistent and justifiable outputs compared to the more monolithic and sometimes opaque reasoning of single LLM agents. The Effectiveness score (6.0) also demonstrates a meaningful improvement, suggesting that the multi-dimensional optimization strategies generated by the collaborative agents are perceived as more actionable and valuable for refining designs. While the lead in Satisfaction (5.8) is more modest, it signifies that the added complexity of the multi-agent process did not negatively impact usability and was appreciated for the comprehensive feedback provided. The strong performance of models like DeepSeek highlights the advanced capabilities of modern LLMs, yet our framework's specialized architecture for design deduction provides a measurable, consistent edge, particularly in producing reliable and effective design guidance. This comparative study empirically validates the core thesis that a dedicated multi-agent collaborative approach surpasses single-model, general-purpose assistants in complex design evaluation tasks.

CONCLUSION

This paper proposes a Multi-Agent System-based design deduction framework. Through collaborative debate among rationality, feasibility, and risk assessment agents, it achieves multi-dimensional dynamic optimization of design solutions. Experiments demonstrate that the framework outperforms existing methods in terms of suggestion adoption rate, risk warning accuracy, and iteration efficiency, while also exhibiting robust cross-domain generalization capabilities. Future work will explore agent self-evolution mechanisms and the integration of real-time data streams to further enhance adaptability to open innovation environments.

REFERENCES

Du, H., Thudumu, S., Vasa, R., & Mouzakis, K. (2024). *A Survey on Context-Aware Multi-Agent Systems: Techniques, Challenges and Future Directions*. ArXiv.org. https://arxiv.org/abs/2402.01968

Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). *From LLM Reasoning to Autonomous AI Agents: A Comprehensive Review*. ArXiv.org. https://arxiv.org/abs/2504.19678

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., & He, C. (2024, February 5). *LLM Multi-Agent Systems: Challenges and Open Problems*. ArXiv.org. https://arxiv.org/abs/2402.03578

He, P., Lin, Y., Dong, S., Xu, H., Xing, Y., & Liu, H. (2025). Red-Teaming LLM Multi-Agent Systems via Communication Attacks. *Findings of the Association for Computational Linguistics: ACL 2022*, 6726–6747. https://doi.org/10.18653/v1/2025.findings-acl.349

Jiang, B., Xie, Y., Wang, X., Su, W. J., Taylor, C. J., & Mallick, T. (2024). *Multi-Modal and Multi-Agent Systems Meet Rationality: A Survey*. Openreview.net. https://openreview.net/forum?id=9Rtm2gAVjo

Jiang, F., Peng, Y., Dong, L., Wang, K., Yang, K., Pan, C., Dusit Niyato, & Dobre, O. A. (2024). Large Language Model Enhanced Multi-Agent Systems for 6G Communications. *IEEE Wireless Communications*, *31*(6), 48–55. https://doi.org/10.1109/mwc.016.2300600

Ke, Z., Jiao, F., Ming, Y., Nguyen, X.-P., Xu, A., Long, D. X., Li, M., Qin, C., Wang, P., Savarese, S., Xiong, C., & Joty, S. (2025). *A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems*. ArXiv.org. https://arxiv.org/abs/2504.09037

Li, A., Xie, Y., Li, S., Tsung, F., Ding, B., & Li, Y. (2024). *Agent-Oriented Planning in Multi-Agent Systems*. ArXiv.org. https://arxiv.org/abs/2410.02189

Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, *1*(1). https://doi.org/10.1007/s44336-024-00009-2

Maldonado, D., Cruz, E., Jackeline Abad Torres, Cruz, P. J., & Gamboa, S. (2024). Multi-agent Systems: A survey about its components, framework and workflow. *IEEE Access*, *12*, 1–1. https://doi.org/10.1109/access.2024.3409051

Wang, B., Yue, X., & Sun, H. (2023). *Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate*. ArXiv.org. https://arxiv.org/abs/2305.13160

Wang, J. (2025). *A Tutorial on LLM Reasoning: Relevant Methods behind ChatGPT o1*. ArXiv.org. https://arxiv.org/abs/2502.10867

Wang, Q., Wang, Z., Su, Y., Tong, H., & Song, Y. (2024). *Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?* ArXiv.org. https://arxiv.org/abs/2402.18272

Wu, J., Zhu, J., & Liu, Y. (2025). *Agentic Reasoning: Reasoning LLMs with Tools for the Deep Research*. https://i-newcar.com/uploads/allimg/20250221/2-250221164PCR.pdf

Xie, T., Gao, Z., Ren, Q., Luo, H., Hong, Y., Dai, B., Zhou, J., Qiu, K., Wu, Z., & Luo, C. (2025). *Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning*. ArXiv.org. https://arxiv.org/abs/2502.14768

Yang, L., Yu, Z., Cui, B., & Wang, M. (2025). *ReasonFlux: Hierarchical LLM Reasoning via Scaling Thought Templates*. ArXiv.org. https://arxiv.org/abs/2502.06772

Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y., Liu, Z., Zhou, B., Peng, H., Liu, Z., & Sun, M. (2024). *Advancing LLM Reasoning Generalists with Preference Trees*. ArXiv.org. https://arxiv.org/abs/2404.02078

Yusuf Izmirlioglu, Pham, L., Son, T. C., & Enrico Pontelli. (2024). A Survey of Multi-Agent Systems for Smartgrids. *Energies*, *17*(15), 3620–3620. https://doi.org/10.3390/en17153620

Zhang, G., Yue, Y., Li, Z., Yun, S., Wan, G., Wang, K., Cheng, D., Yu, J. X., & Chen, T. (2024). *Cut the Crap: An Economical Communication Pipeline for LLM-based Multi-Agent Systems*. ArXiv.org. https://arxiv.org/abs/2410.02506

Zhang, J., Wang, X., Ren, W., Jiang, L., Wang, D., & Liu, K. (2025). RATT: A Thought Structure for Coherent and Correct LLM Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(25), 26733–26741. https://doi.org/10.1609/aaai.v39i25.34876