

DentEval: Fine-tuning-Free Expert-Aligned Assessment in Dental Education via LLM Agents

Xinyu Deng¹, Vesna Miletic¹, Elvis Trinh¹, Jinlong Gao¹, Chang Xu¹, and Daochang Liu² *

¹ The University of Sydney, Sydney, NSW, Australia
ninadeng2023@gmail.com, {vesna.miletic, elvis.trinh, jinlong.gao, c.xu}@sydney.edu.au

² The University of Western Australia, Perth, Australia
daochang.liu@uwa.edu.au

Abstract. Large language models (LLMs) have demonstrated considerable potential in automating assignment scoring within higher education, providing efficient and consistent evaluations. However, existing systems encounter substantial challenges when assessing students’ responses to open-ended short-answer questions. These challenges include the need for large, annotated datasets for fine-tuning or additional training, as well as inconsistencies between model outputs and human-level evaluations. This issue is particularly pronounced in domains requiring specialized knowledge, such as dentistry. To address these limitations, we propose DentEval, an LLM-based automated assignment assessment system supporting multimodal inputs (e.g., text and clinical images) that is tailored for dental curricula. This framework integrates role-playing prompting and Self-refining Retrieval-Augmented Generation (SR-RAG) to assess student responses and ensure that the system’s outputs closely align with human grading standards. We further utilized a dataset annotated by dental professors, dividing it into few-shot learning and testing sets to evaluate the DentEval framework. Results demonstrate that DentEval exhibits a stronger correlation with human grading compared to representative baselines. Finally, comprehensive ablation studies validate the effectiveness of the individual components incorporated in DentEval. Our code is available on GitHub at: <https://github.com/DXY0711/DentEval>

Keywords: Automated Assignment Assessment · Dental Curricula · Large Language Models · Retrieval-Augmented Generation · Role-Playing.

1 Introduction

The rapid advancement of Large Language Models (LLMs), such as OpenAI’s GPT series [1, 13], has revolutionized education [2, 19] by enabling efficient and consistent automated assessment [9, 10]. However, applying LLMs to specialized domains like dentistry remains challenging. **Challenge 1: Ensuring precise**

* Corresponding author.

interpretation of domain-specific terminology, as LLMs may struggle with nuanced professional vocabulary. **Challenge 2: Aligning with expert grading rubrics**, which requires LLMs to adhere to specific evaluation criteria established by professionals. Addressing these challenges typically requires extensive fine-tuning and retraining with large datasets [8, 23, 25], which is time-consuming and impractical for resource-constrained settings.

To address these gaps, we propose DentEval, a LLM-based [13] automated assignment assessment framework tailored for dental curricula. DentEval integrates several advanced techniques, including Self-refining Retrieval-Augmented Generation (SR-RAG), multi-agent systems [3], and role-playing prompting, to ensure that LLMs can acquire sufficient domain knowledge and that the system’s outputs closely align with human grading standards.

Compared with standard automatic grading frameworks, there are two key innovations in DentEval:

1. **Self-refining Retrieval-Augmented Generation (SR-RAG):** We propose a novel framework for refining retrieved results and autonomously evaluating their adequacy, which effectively addresses the challenge of acquiring sufficient domain-specific knowledge without extensive fine-tuning in specialized fields.
2. **Role-playing Prompting:** DentEval introduces an innovative Sample Answer Generation (SAG) module, which designates the LLM as a professor to generate reference answers for the subsequent scoring module. Additionally, in the scoring module, the Evaluator LLM is assigned a teacher role to enhance alignment between its grading and human assessments.

Collectively, these innovative strategies not only address the challenges of automated assessment but also pave the way for practical applications:

1. **No Fine-tuning Requirement:** DentEval operates effectively without extensive computing resources or retraining.
2. **Optimized RAG Process:** We integrate LLM agents into the RAG process to optimize search results by extracting keywords from the original query, summarizing relevant knowledge from extensive retrieval results, and self-evaluating the sufficiency of the retrieved evidence.
3. **Efficiency Gains:** DentEval significantly reduces evaluation time and financial costs, allowing better allocation of educational resources.

We evaluate DentEval on a professor-annotated dataset, demonstrating stronger alignment with human grading compared to baseline methods and advanced techniques such as SciEx [5] and FairEval [22]. SciEx combines few-shot learning with role-playing prompts, while FairEval employs Multiple Evidence Calibration (MEC) and Balanced Position Calibration (BPC) to reduce bias. DentEval achieves substantial improvements in **Spearman’s correlation (up to 0.9259)** and **Pearson’s correlation (up to 0.8957)**, confirming its reliability. An ablation study further validates the contributions of SR-RAG, role-playing prompting, and SAG to system performance.

2 Method

The fundamental operations of DentEval are to generate reference answers based on relevant information from the dental handbook and then collaborate with Few-shot learning and role-playing prompting to make the final evaluation. The whole system is shown in Fig. 1

The system processes three inputs: (1) a student’s free-text answer A_{student} , (2) a marking rubric R that defines the scoring criteria, and (3) a question Q , which may contain textual or visual content. Additionally, a retrieval corpus C (e.g., textbooks) is leveraged to extract relevant evidence by computing the similarity between the embedded representation of Q and the contents of C .

The output is a numerical score $S_{\text{final}} \in [0, 5]$, determined by majority voting from multiple LLM scoring results. The mapping from input to output is formalized as:

$$S_{\text{final}} = \text{DentEval}(A_{\text{student}}, Q, R, C) \quad (1)$$

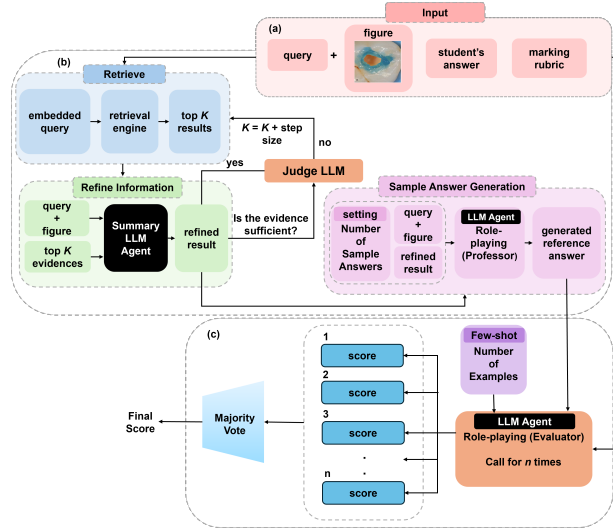


Fig. 1. DentEval Workflow Diagram consists of three main steps: (a) The system requires three types of input: the query (question), with an associated figure provided optionally, the student’s answer, and the marking rubric; (b) Retrieving the most relevant knowledge from the dental handbook and generating reference answers to aid in assessment; (c) Grading the student’s response with the assistance of reference answers and few-shot learning, and returning the final score through majority voting. Role-playing prompts are employed in LLM agents to simulate human-like reasoning.

2.1 Information Searching Process

Retrieving Process Milvus [21] is utilized as the vector database to store the textual and tabular information extracted from the dental handbook, while also serving as a similarity search engine to retrieve the top K most relevant information chunks for the input query. Here, K is a dynamic parameter that is adjusted during the self-evaluating phase when the retrieved evidence is deemed insufficient.

Information Refinement Once the information retrieval process is completed, the top K evidence chunks, along with Q , are input into the GPT-4o-based summarization LLM Agent for information refinement and then generating the refined information E . This step is necessary because the segmented evidence chunks are of fixed size and may include content that is not directly relevant to the query or its key terms [4, 24]. Irrelevant information within these chunks can introduce noise, which may negatively impact the accuracy and quality of the sample answer generation in subsequent stages.

Sufficiency Check The refined information E will be assessed by an agent LLM to determine whether it is sufficient for generating a high-quality reference answer. If the evidence is deemed insufficient, the system will loop back to the retrieval stage to gather additional evidence ($K = K + \text{step size}$), improving the overall quality of the response. This process leverages the LLM’s emerging capability for self-reflection and sufficiency judgment [7, 12], enabling it to introspectively evaluate whether the retrieved context provides adequate grounding before generation. To ensure practical reliability, the generated reference summaries were also reviewed by academics in the domain, who confirmed their factual consistency and alignment with the standards of dental evaluation.

Sample Answer Generation (SAG) This section introduces a novel approach within the standard RAG framework [6, 11, 26]. Unlike traditional knowledge retrieval outputs, E is transformed into sample answers $\{A_{\text{sample}}^1, \dots, A_{\text{sample}}^N\}$ by an LLM agent designated as a dental academic, as in Eq.(2). This mirrors the process used in academic settings to assess student assignment responses, structurally aligning the automated evaluation method with human judgment. The need for the agent to generate multiple reference answers arises from the nature of the open-ended questions in our dataset, which do not necessarily have a single, unique correct answer. This approach acknowledges the inherent complexity and variability of possible correct responses. Moreover, diverse reference answers help the LLM capture more acceptable responses, improving its assessment capability.

$$A_{\text{sample}} = \text{LLM}_{\text{professor}}(Q, E) \quad (2)$$

2.2 Evaluating Process

Each student’s response is scored based on a predefined rubric and reference answers. Furthermore, we incorporate Few-shot learning to improve performance. A student response from each score tier serves as a Few-shot learning example with evaluation feedback, as shown in Fig. 2. To reduce variability in LLM scoring and improve consistency, the evaluation agent generates n scores, which are then aggregated using **Majority Voting**, ensuring a more reliable and robust assessment.

$$S_i = \text{LLM}_{\text{evaluator}}^i(Q, A_{\text{student}}, \{A_{\text{sample}}^1, \dots, A_{\text{sample}}^N\}, R), \quad i \in \{1, \dots, n\} \quad (3)$$

$$S_{\text{final}} = \text{Mode}(\{S_1, S_2, \dots, S_n\})$$



Few-shot Examples:	
Question: What is the purpose of enamel etching as shown in the photograph (right)?	
Student’s Answer: The photo appears to show etch being placed on the margins of a Class I preparation using the selective-etch protocol. The purpose of etch being placed on these surfaces is to allow the sulfuric acid in the etch to make micro-pits in the enamel, thereby increasing the surface area and roughness of the enamel. The result should be better bond for the adhesive and ultimately a more mechanically secure strong restoration.	
Feedback: sulfuric acid is incorrect. Adequate explanation, therefore score 3	
Score: 3	
Same Question ...	
Student’s Answer: polyacrylic acid removes the smear layer on enamel and improves bonding of adhesive.	
Feedback: No relevant content is provided. The response does not address the question.	
Score: 0	
Questions In The Dataset:	
Question 1: What is the purpose of enamel etching as shown in the photograph (right)?	
Question 2: What are the consequences of light curing an adhesive or composite for a Class I (or Class II) restoration at an angle that is not perpendicular to the occlusal surface?	

Fig. 2. Scoring Rubric and Example Content Illustration

3 Experiments and Result

3.1 Dataset

As shown in Fig. 2, the dataset used in this study comprises two types of open-ended questions: **(1)** a combination of text and images and **(2)** purely textual questions. Both question types are derived from the Tooth Conservation Simulation Clinic course, where the first type serves as a practical simulation question and the second as a theoretical assessment question.

For each question, we collected 28 responses. Some were gathered from an anonymous pilot study with students, while the rest were simulated by dentistry academics. Each response was scored by dentistry academics based on a predefined marking rubric. In the subsequent evaluation, we selected one response from each score tier as a few-shot learning example, while the remaining responses were used as the testing set. As a result, each question includes **6** few-shot examples and **22** test data points. Despite the limited number of questions, these tasks differ both structurally and substantively:

- One is text-only, while the other involves multimodal reasoning with textual and visual inputs
- They cover distinct subdomains in dentistry, including dental morphology and restorative materials.
- The responses exhibit considerable variation in length, structure, and reasoning complexity, reflecting authentic diversity in learner expression.

3.2 Implementation Details

We chose GPT-4o (gpt-4o-mini)[13] as the LLM agent in DentEval, considering its multimodal capabilities. To mitigate biases in scoring and enhance consistency and fairness across LLM and human evaluators, we apply Cumulative Distribution Function (CDF) mapping[15] as a post-processing method. We compared DentEval with baselines and state-of-the-art methods to evaluate its effectiveness. Additionally, We further analyzed the resource cost of DentEval and validated the effectiveness of modules in the system through ablation experiments.

We evaluate DentEval against the following methods. As a baseline, we use both Zero-shot and Few-shot approaches. Zero-shot directly prompts the LLM to score responses without modifications or role-playing. Few-shot enhances Zero-shot by incorporating examples for reference. We also compare DentEval with SciEx[5] and FairEval[22].

For the evaluation metrics, we selected Accuracy (Acc.), Spearman’s Rank-Order Correlation Coefficient (SROCC) [17], and Pearson Linear Correlation Coefficient (PLCC) [16]. Accuracy measures the proportion of student responses that are accurately scored across the entire dataset, while SROCC and PLCC assess the consistency and linear correlation of the scoring results with human evaluations.

3.3 Experiment Result

Given the differences in the types of the two questions and their respective score distributions within our dataset, we have chosen to evaluate them separately. The best results for each method in Question 1 are presented in Table 1. We emphasize the alignment of the results with human evaluations, using accuracy as a secondary reference dimension. As shown in Table 1, both SciEx and FairEval outperform the baseline Zero-shot method in terms of SROCC, PLCC, and accuracy, even though they slightly underperform compared to the baseline Few-shot

method. Our approach achieves the best performance across all evaluation metrics—SROCC (0.7447), PLCC (0.7288), and accuracy (68.2%), with the p-value also reaching its minimum. This demonstrates that for Question 1, DentEval is better aligned with human grading patterns and provides more accurate scoring.

For Question 2, the results lead to the same conclusion: our method outperforms the other four approaches in both consistency with human evaluations and accuracy. Moreover, it achieves higher SROCC (0.9259) and PLCC (0.8957) compared to Question 1. However, unlike the findings about Question 1, the baseline Few-shot method exhibits a noticeable decline in scoring performance for Question 2. This discrepancy is likely attributed to differences in question types. While Question 1 has a single correct answer, Question 2 allows students to respond from multiple perspectives and still achieve full marks. As a result, during Few-shot learning, the LLM may interpret the provided examples as the only correct answers, leading to misjudgments of other students’ responses. This decline in performance suggests that, for open-ended questions, the presence of a single definitive answer may be a critical factor influencing the effectiveness of automatic grading.

According to the U.S. Bureau of Labor Statistics (BLS) [20], the average hourly wage for higher education professionals in 2023 was \$51. Stephen, Gierl, and King (2021) [18] reported that human scorers typically spend about 1.5 minutes evaluating a single constructed-response item per student, equating to roughly \$1.275 per response. In contrast, our findings indicate that the GPT-4o LLM agents in DentEval process each question using 43K tokens in approximately 0.13 minutes at a cost of only \$0.007 [14]. As shown in Table 2, DentEval significantly reduces both time and financial expenditures in educational settings, thereby enabling educators to dedicate resources to more pedagogically valuable activities.

Table 1. Performance Comparison of DentEval on Question 1 & 2 with Baselines, SciEx, and FairEval.

Question 1					
	Baseline(Zero-shot)	Baseline(Few-shot)	SciEx [5]	FairEval [22]	DentEval
Acc.(%) ↑	9.1	50	13.6	18.2	68.2
SROCC ↑	0.5133	0.6811	0.5444	0.3516	0.7447
P Value ↓	0.0146	0.0005	0.0088	0.1086	0.0001
PLCC ↑	0.4936	0.6754	0.5032	0.3933	0.7288
P Value ↓	0.0196	0.0006	0.0170	0.0702	0.0001
Question 2					
	Baseline(Zero-shot)	Baseline(Few-shot)	SciEx [5]	FairEval [22]	DentEval
Acc.(%) ↑	50	50	22.7	36.4	68.2
SROCC ↑	0.8285	0.2955	0.8197	0.6542	0.9259
P Value ↓	<0.0001	0.1819	<0.0001	0.0010	<0.0001
PLCC ↑	0.8106	0.2821	0.7562	0.6396	0.8957
P Value ↓	<0.0001	0.2034	<0.0001	0.0013	<0.0001

Table 2. Cost Comparison between Human Evaluator and DentEval.

	Tokens Count	Time (min)	Cost per Answer
Human Evaluator	-	1.5	\$1.275
DentEval	43000	0.13	\$0.007

3.4 Ablation Study

We conduct an ablation study to verify the effectiveness of each innovative component in our proposed method. Since the previous section demonstrated that different question types could affect the performance of the Few-shot method, we use Zero-shot as the sole baseline for comparison in this ablation study. We evaluate the following configurations: using only the SR-RAG component within the DentEval framework without incorporating SAG or Role-playing Prompting; using both RAG and SAG but excluding Role-playing Prompting; using only Role-playing Prompting; and the full DentEval framework. The best-performing results for each method on Question 1 and Question 2 are presented in Table 3. Overall, the results indicate that each of our proposed innovations effectively improves consistency with human evaluations and contributes to an increase in accuracy.

Table 3. Ablation Study on Components of DentEval: Results for Question 1 & 2

	Acc. (%) \uparrow	SROCC \uparrow (P-Value \downarrow)	PLCC \uparrow (P-Value \downarrow)
Question 1			
Baseline (Zero-shot)	9.1	0.5133 (0.0146)	0.4936 (0.0196)
DentEval (SR-RAG)	54.5	0.6300 (0.0017)	0.5196 (0.0132)
DentEval (SAG)	9.1	0.6550 (0.0009)	0.5642 (0.0062)
DentEval (Role-playing)	40.9	0.7098 (0.0002)	0.6693 (0.0007)
Full DentEval (Table 1)	68.2	0.7447 (<0.0001)	0.7288 (<0.0001)
Question 2			
Baseline (Zero-shot)	50.0	0.8285 (<0.0001)	0.8106 (<0.0001)
DentEval (RAG)	68.2	0.8902 (<0.0001)	0.8735 (<0.0001)
DentEval (SAG)	63.6	0.8884 (<0.0001)	0.8749 (<0.0001)
DentEval (Role-playing)	63.6	0.8884 (<0.0001)	0.8749 (<0.0001)
Full DentEval (Table 3)	68.2	0.9259 (<0.0001)	0.8957 (<0.0001)

4 Conclusion

We present DentEval, a scalable framework for automated assessment in dental education that integrates multimodal inputs, retrieval-augmented generation, and role-playing LLM agents. By retrieving domain-specific knowledge and generating diverse reference answers, DentEval aligns closely with human grading (Spearman up to 0.93; Pearson up to 0.90). Ablation results confirm the contribution of each component.

Designed for open-ended, domain-specific tasks, DentEval offers a resource-efficient alternative to manual grading. Future directions include extending multimodal RAG (e.g., clinical videos, 3D models) and developing adaptive calibration to improve robustness across varied question types.

Acknowledgments. This work was supported in part by the Australian Research Council under Projects DP240101848 and FT230100549.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ali, K., Barhom, N., Tamimi, F., Duggal, M.: Chatgpt—a double-edged sword for healthcare education? implications for assessments of dental students. *European Journal of Dental Education* **28**(1), 206–211 (2024)
3. Chan, C.M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201 (2023)
4. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieval-augmented generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 17754–17762 (2024)
5. Dinh, T.A., Mulloy, C., Bärmann, L., Li, Z., Liu, D., Reiß, S., Lee, J., Lerzer, N., Ternava, F., Gao, J., et al.: Sciex: Benchmarking large language models on scientific exams with human expert grading and automatic grading. arXiv preprint arXiv:2406.10421 (2024)
6. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 **2** (2023)
7. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., Guo, J.: A survey on llm-as-a-judge (2025), <https://arxiv.org/abs/2411.15594>
8. Katuka, G.A., Gain, A., Yu, Y.Y.: Investigating automatic scoring and feedback using large language models. arXiv preprint arXiv:2405.00602 (2024)
9. Latif, E., Zhai, X.: Fine-tuning chatgpt for automatic scoring. *Computers and Education: Artificial Intelligence* **6**, 100210 (2024)

10. Lee, G.G., Latif, E., Wu, X., Liu, N., Zhai, X.: Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence* **6**, 100213 (2024)
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020)
12. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: Llm-as-judges: A comprehensive survey on llm-based evaluation methods (2024), <https://arxiv.org/abs/2412.05579>
13. OpenAI: Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, accessed: 2025-02-20
14. OpenAI: OpenAI API Pricing, <https://openai.com/api/pricing/>, accessed: 2025-02-27
15. Panofsky, H.A., Brier, G.W.: Some applications of statistics to meteorology. (No Title) (1968)
16. Pearson, K.: Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110 (1894)
17. Spearman, C.: The proof and measurement of association between two things. (1961)
18. Stephen, T.C., Gierl, M.C., King, S.: Automated essay scoring (aes) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice* **54**, 103085 (2021)
19. Thorat, V.A., Rao, P., Joshi, N., Talreja, P., Shetty, A., Thorat, V., RAO, P., Shetty, A.R.: The role of chatbot gpt technology in undergraduate dental education. *Cureus* **16**(2) (2024)
20. U.S. Bureau of Labor Statistics: 25-1042 Biological Science Teachers, Postsecondary: Occupational Employment and Wages, May 2023 (2023), [https://www.bls.gov/oes/current/oes251042.htm\(4\)](https://www.bls.gov/oes/current/oes251042.htm(4)), last Modified: April 3, 2024
21. Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., et al.: Milvus: A purpose-built vector data management system. In: *Proceedings of the 2021 International Conference on Management of Data*. pp. 2614–2627 (2021)
22. Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators. *ArXiv abs/2305.17926* (2023)
23. Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., Wang, Y.: Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association* **31**(9), 1833–1843 (2024)
24. Xu, S., Pang, L., Shen, H., Cheng, X., Chua, T.s.: Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *arXiv preprint arXiv:2304.14732* (2023)
25. Zhang, W., Wang, Q., Kong, X., Xiong, J., Ni, S., Cao, D., Niu, B., Chen, M., Li, Y., Zhang, R., et al.: Fine-tuning large language models for chemical text mining. *Chemical Science* **15**(27), 10600–10611 (2024)
26. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., Cui, B.: Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473* (2024)