Leveraging Large Language Models for Explaining Material Synthesis Mechanisms: The Foundation of Materials Discovery

Yingming Pu[1,2] **Liping Huang**[1] **Tao Lin**[1*] **Hongyu Chen**[1*]

[1]Westlake University, Hangzhou, China
[2]Zhejiang University, Hangzhou, China

{puyingming, lintao, chenhongyu}@westlake.edu.cn

Abstract

Large language models (LLMs) have attracted significant attention in the advancement of materials discovery, particularly in their role in automation and robotics. However, a key question remains: Do these models operate based on a true grasping of physicochemical principles when designing experiments or interpreting results? Existing evaluations primarily focus on fact-checking tasks such as material property prediction and named entity recognition, while neglecting the reasoning required to grasp fundamental synthesis mechanisms. Furthermore, no previous studies have directly evaluated LLMs' ability to reason about synthesis mechanisms. To address these challenges, we first develop a benchmark containing 775 semi-manually created multiple-choice questions in the field of gold nanoparticles (AuNPs) synthesis for evaluation. Second, we probe the model's output logits to derive precise selection probabilities for the correct answers, obtaining a confidence-based score (c-score) as a quantitative evaluation metric. Additionally, based on this evaluation, we also develop an AI assistant using retrieval-augmented generation (RAG) to explain AuNP synthesis mechanisms, achieving a 10% improvement in accuracy over the leading model, Claude. Our study highlights the potential of LLMs in recognizing scientific mechanisms and offers a valuable tool for aiding the exploration of synthesis methods. Moreover, our dataset establishes a foundation for developing highly efficient models with the utilization of material synthesis mechanisms. Code and dataset are available here.

1 Introduction

Achieving precise synthesis has long been a dream for materials chemists. This involves using a range of controllable material synthesis techniques to create materials, generally, with specific structures and properties based on the underlying physicochemical mechanisms. [21, 39, 26, 28] Discover new synthesis mechanisms is similar to knowing how nature works–why these materials are formed, and how to synthesize them with desired properties. [27, 34, 36, 33, 38]

Standing at the forefront of contemporary innovation, to achieve controllable material synthesis, designing cutting-edge deep learning methods, particularly when combined with domain-specific knowledge, holds significant potential in this field. [25, 6, 29] It is important to note that many synthesis mechanisms are documented in natural language, which can be extracted directly from the scientific literature. In this context, large language models (LLMs) such as GPT-4 are emerging as

[*] Corresponding authors.

AI for Accelerated Materials Design Workshop at NeurIPS 2024.

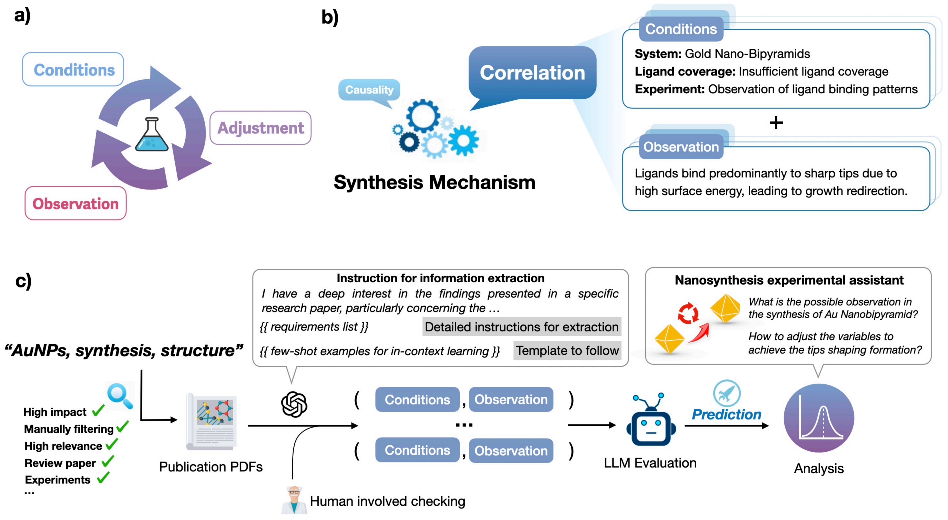

a)
Conditions
Adjustment
Observation
b)
Causality
Correlation
Synthesis Mechanism
Conditions
System: Gold Nano-Bipyramids
Ligand coverage: Insufficient ligand coverage
Experiment: Observation of ligand binding patterns
+
Observation
Ligands bind predominantly to sharp tips due to high surface energy, leading to growth redirection.
c)
"AuNPs, synthesis, structure"
High impact ✓
Manually filtering ✓
High relevance ✓
Review paper ✓
Experiments ✓
...
Publication PDFs
Instruction for information extraction
I have a deep interest in the findings presented in a specific research paper, particularly concerning the ...
{{ requirements list }}
Detailed instructions for extraction
{{ few-shot examples for in-context learning }}
Template to follow
(Conditions , Observation)
...
(Conditions , Observation)
Human involved checking
LLM Evaluation
Prediction
Analysis
Nanosynthesis experimental assistant
What is the possible observation in the synthesis of Au Nanobipyramid?
How to adjust the variables to achieve the tips shaping formation?

Figure 1: **Semantic illustration of our proposed framework for LLMs evaluation in nanomaterial synthesis prediction, highlighting concepts and workflow.** a) nanosynthesis study loop: begins with basic conditions, leading to the discovery of novel synthesis rules through experiments involving variable adjustments. b) exemplifies the synthesis mechanism, dissected into causality and correlations, with an emphasis on correlations described through condition-observation pairs. c) outlines the process from sourcing relevant literature (using key area keywords) for benchmark construction and model evaluation.

promising tools for assisting in the interpretation of material synthesis processes. These models have already demonstrated exceptional results in autonomous synthesis experiments. [12, 37, 1, 42, 13, 2, 22] However, despite efforts to leverage LLMs for material synthesis, a critical question remains: *Do these LLMs grasp the realworld physicochemical principles? Can they act as expert interpreters of synthesis mechanisms?*

Among existing investigations, the common and straightforward approach for answering this question is fact-checking, such as named entity recognition (Table 1). [35, 8, 32, 40, 15, 43, 18, 16, 14, 41, 31] However, evaluating the cognitive logic behind underlying principles is far more challenging, yet essential for tackling key scientific issues. [3] A historical example is Alexander Fleming's observation that bacteria could not survive where mold grew—a simple fact. Yet, this observation alone could not explain why the mold inhibited bacterial growth. Through his reasoning, Fleming discovered penicillin. This case highlights the crucial role of reasoning in scientific discovery, underscoring that correlation without rational explanation is insufficient.

Inspired by the research of nanomaterial synthesis, we embark on a feasibility study regarding whether LLMs can grasp underlying physicochemical principles. Before evaluating LLMs, the description method of a synthesis mechanism should be defined formally. Given the lack of existing work in this area, we propose a structured descriptor based on the iterative process of materials synthesis, which includes four key components: *condition, adjustment, observation and mechanisms*. Using this descriptor, we selected 220 high-impact research articles on gold nanoparticle synthesis to develop 775 expert-level multiple-selection questions. This benchmark covers five primary methods of gold nanoparticle synthesis and six major categories of nanomaterial structures, providing a comprehensive assessment of the reasoning capabilities of current LLMs. To evaluate reasoning ability, particularly in domain-specific mechanisms, we probe the output confidence of LLMs by analyzing output logits to derive the precise selection probability for the correct answer. This confidence-based score (c-score) serves as a quantitative metric, resembling human knowledge testing to avoid mere conjectures. As a solution, we demonstrate an assistant with retrieval augmented generation (RAG) to assist gold nanoparticle synthesis mechanisms question-answering with 10% more accuracy than current leading model, Claude.

Our contribution encompasses the following key elements, also shown in Figure 1:

1. We propose a synthetic mechanistic descriptor, grouped by initial conditions, variable adjustments, and experimental observations for the material synthesis mechanism study.

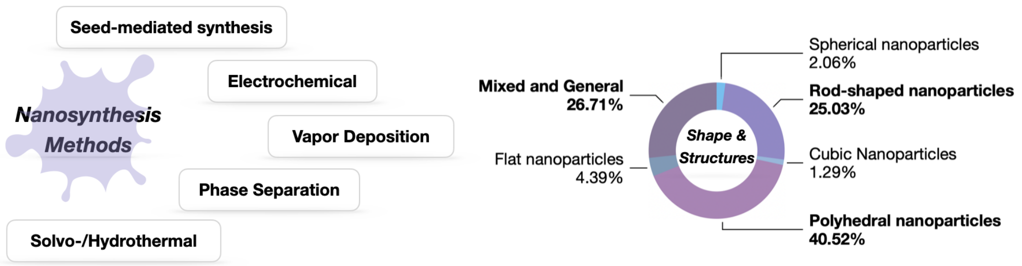


Figure 2: **Benchmark illustration.** We illustrate the collected data set from two perspectives, i.e., synthesis methods and structures.

Reference	Factual	Reasoning	Mechanisms	Creation	Domain / Area
Kim et al. (2017) [19]	✓			Auto	Materials Science
Weston et al. (2019) [35]	✓			Auto	Materials Science
Thawani et al. (2019) [31]	✓			Auto	Materials Science
Venugopal et al. (2021) [32]	✓			Manual	Chemistry
Cruse et al. (2022) [8]	✓			Auto	Materials Science
Guo et al. (2023) [14]	✓	✓		Auto	Chemistry
Zheng et al. (2023) [41]	✓	✓		Auto	Scientific discovery
Dunn et al. (2024) [9]	✓			Auto	Chemistry and Materials Science
Zaki et al. (2024) [40]	✓			Manual	Materials Science
Our Benchmark	✓	✓	✓	Manual	Materials Science

Table 1: An overview of current benchmarks in materials science and related domains with large language models reveals a focus on three primary categories: factual knowledge, reasoning, and synthesis mechanisms, particularly those grounded in physicochemical principles. Our benchmark specifically targets the evaluation of synthesis mechanisms, emphasizing both the factual accuracy of experimental observations and the underlying physicochemical principles that are fundamental to the development of novel materials.

2. A benchmark for gold nanoparticle synthesis mechanisms is established, comprising 775 multiple-selection questions focused on synthesis mechanisms, developed using the proposed descriptor.
3. We introduce a confidence-based score (c-score), providing an interpretable metric to assess LLMs' reasoning ability of complex synthesis mechanisms.
4. Based on evaluation results, we develop a research assistant to aid the study of gold nanoparticle synthesis mechanisms.

2 Method

2.1 Preparation of datasets

The dataset for evaluation was meticulously selected from high-quality papers boasting an IF $>$ 15, with a particular emphasis on synthesis methods for controlling the structure of gold nanoparticles. To ensure the relevance and quality of research articles to meet the theme of gold nanoparticle synthesis, we manually reviewed 220 articles from a diverse range of journals. Each article underwent a process of key experimental information extraction, leading to the summarization and collection of 775 experimental records. To enable the creation of a unified framework of descriptors to assist in compiling test questions, these data are categorized under a *condition-adjustment-observation-mechanism* fashion, which is a usual loop for discovering synthesis principles and serves as a descriptor for expressing any controlled experiment. Moreover, given the unstructured nature of mechanistic expression in nanomaterial synthesis, we emphasize the importance of using average sampling in the design of the evaluation set by manually checking. In this approach, the dataset distribution is shown in Figure 2, one case in the benchmark is shown in supplementary information (see Figure S1-b), and we address three pivotal issues and challenges while collecting these data: 1) Uniform sampling in gold nanoparticle synthesis. 2) Precise extraction of experiments in each research article. 3) The information completeness of experiments, and synthesis mechanisms introduced by authors.

Sampling Issue: Our evaluation approach is based on two key perspectives: material synthesis methods and material morphology. The first category primarily focuses on the seed-mediated synthesis method, a widely used technique for creating nanoparticles with complex structures. The second category emphasizes mechanisms that are more broadly applicable to nanocrystals, rod-shaped particles, and certain unclassified systems, as illustrated in Figure 2. The two data distributions indicate the extensive range of knowledge points related to gold nanoparticle synthesis considered in this study.

Information extraction issue. We employed cue word prompt engineering in conjunction with predefined descriptors to extract the descriptions of synthesis experiments from each paper. We found that for each report we collected, on average, experimental report included five particularly relevant initial conditions and four sets of experimental setups, such as the increase or decrease of certain parameters, the presence or absence thereof, along with a corresponding number or more of observations leaning towards conclusions, details are shown in supplementary information (Figure S1-a).

Mechanism completeness issue. We consider the inherent bias in interpreting mechanisms—where a handful of studies fail to report complete and comprehensive descriptions of mechanisms compared to the average, and there exists a wide variance in the interpretation of complete mechanisms. To illustrate this, we employ the GPT-4 to model in conjunction with cue words for a more detailed interpretation of observed tendencies based on the content. Meanwhile, we also confirmed that 775 questions are almost making the accuracy converged for evaluating LLMs according to their performance in this task, see supplementary information (Figure S2) for more details.

Furthermore, we discovered that the utilization of cue word engineering with the GPT-4 model enables rapid acquisition of summaries for nanomaterials synthesis papers, which is also demonstrated in existing work. [7, 4, 42, 23] This capability extends even to articles that do not contain experimental data (such as literature reviews, opinions, and comments), wherein GPT-4 provides feedback indicating the absence of extractable content, thereby demonstrating GPT-4's honesty in responding to user queries.

Ultimately, we can rephrase this refined condition-observation pair-wise data into a standardized format of questions and options with the gold answer using the GPT-4 along with predefined instructions, one case is shown in Figure 2c. These samples are formatted as selection questions with four options, only one gold answer. For detailed methodologies on the aforementioned cue word prompt engineering, and prompt with instructions, please refer to the supplementary information (Note S1).

2.2 LLM Baselines

We choose multiple existing models with different architectures and features for comparison, to better consider the inner design differences. We here consider mainstream models, i.e., Vicuna series (7B and 13B)[5], Mistral and Mixtral architectures[17, 10], Qwen series[11], Gemma[30], GPT-4[12] and Claude. All models introduction please see supplementary information (Note S2).

2.3 Evaluation metrics

The most direct metric for evaluation is accuracy, which assesses the overall performance of models based on the number of correctly answered test questions—higher accuracy scores indicate stronger capabilities. [20, 24] In this work, we aim for the model to answer questions with computed confidence, reflecting a quantitative recognition of the logic in material synthesis. We introduce the confidence-based score (c-score) using a knowledge probing technique. Statistical accuracy, measured by the true-or-false count, serves as a baseline metric to evaluate the model's performance and compare it against random guessing.

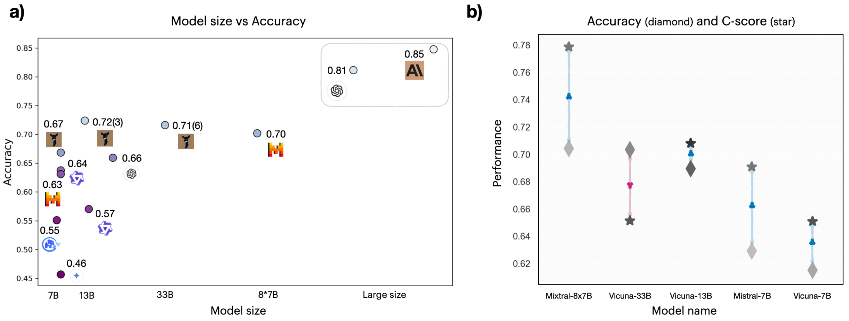


Figure 3: **Evaluation accuracy of baselines and the confidence-based scores of top-5 open-sourced models compared to the original accuracy in multiple selection questions.** a) x-axis represents different models, while y-axis is the accuracy. The figure delineates the range in accuracy achieved by each model under different temperature settings (from 0.0 to 1.0), where the circles represent the accuracy at each temperature setting, and the diamonds denote their average values. b) The comparison between accuracy and confidence-based scores among five top-performance models, showing the performence increasing (green line) and decreasing (red line).

In summary, our benchmark evaluates LLMs from two perspectives: (1) true-or-false-based accuracy, which examines whether LLMs recognize physicochemical principles, and (2) the confidence-based score (c-score), which ensures the model's ability to discern correct answers, as depicted in Figure 2d. To achieve this, we conduct a preliminary study on the effects of temperature on LLMs to assess their stability. Both intuitive accuracy and the c-score are used to evaluate the models' capabilities in recognizing physicochemical principles.

3 Results and Discussion

3.1 Temperature Effect Analysis

To assess the stability of language models, we investigated the effect of different temperatures on their performance. We conducted controlled experiments using five temperature values uniformly selected from 0 to 1: specifically 0.1, 0.3, 0.5, 0.7, and 0.9. With the analysis of the accuracy, we find that there is a general trend of declining precision as temperature increases, though some models exhibit more complex fluctuations. For instance, models such as Claude-3-ops-20240229, GPT-4-0125-preview, Mixtral-8x7B-Instruct-v0.1, Mistral-7B-Instruct-v0.2, and GPT-3.5-turbo initially showed a decrease in accuracy, followed by an increase as the temperature rose slightly.

The results reveal that temperature has a patterned influence on the performance of language models in this task. While the likelihood of incorrect responses increases at higher temperatures, statistical analysis suggests that overall performance at lower temperatures is marginally better, demonstrating enhanced stability. For further details, refer to the supplementary information (Figure S3).

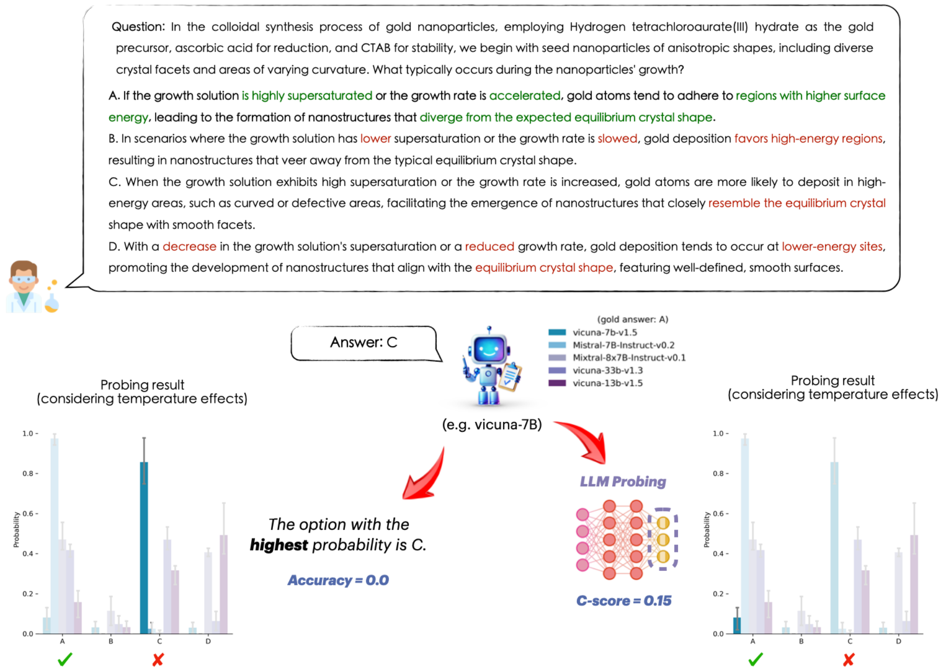
3.2 Results of Accuracy with Temperature

Contemporary LLMs rely on the aforementioned pre-training and fine-tuning processes. Consequently, multiple-choice questions serve as one of the effective methods to evaluate the level of reasoning ability of the language model in a specific domain. This implies that the model should provide answers according to the domain knowledge learned during the pre-training phase and the comprehensive abilities acquired during the fine-tuning phase, as required by the question.

The content we aim to evaluate primarily unfolds from two aspects:

1. The themes encompassed by the multiple-choice questions represent the knowledge being assessed, aiming to evaluate the recognition of the model in terms of the concepts and semantics demonstrated within the sentences.
2. This question format supports both the different mechanisms expression and the examination of the language model's logical reasoning abilities, assessing whether it can answer based on the fundamental principles of gold nanoparticle synthesis.

In order to obtain the binary accuracy of the model, for each multiple-choice questions, one point will be given to the model if it selected the gold answer, otherwise zero. Based on the temperature



Question: In the colloidal synthesis process of gold nanoparticles, employing Hydrogen tetrachloroaurate(III) hydrate as the gold precursor, ascorbic acid for reduction, and CTAB for stability, we begin with seed nanoparticles of anisotropic shapes, including diverse crystal facets and areas of varying curvature. What typically occurs during the nanoparticles' growth?

A. If the growth solution is highly supersaturated or the growth rate is accelerated, gold atoms tend to adhere to regions with higher surface energy, leading to the formation of nanostructures that diverge from the expected equilibrium crystal shape.

B. In scenarios where the growth solution has lower supersaturation or the growth rate is slowed, gold deposition favors high-energy regions, resulting in nanostructures that veer away from the typical equilibrium crystal shape.

C. When the growth solution exhibits high supersaturation or the growth rate is increased, gold atoms are more likely to deposit in high-energy areas, such as curved or defective areas, facilitating the emergence of nanostructures that closely resemble the equilibrium crystal shape with smooth facets.

D. With a decrease in the growth solution's supersaturation or a reduced growth rate, gold deposition tends to occur at lower-energy sites, promoting the development of nanostructures that align with the equilibrium crystal shape, featuring well-defined, smooth surfaces.

Figure 4: **Illustration of the knowledge probing method using a simple case.** This diagram represents the application of a knowledge probing method where an input comprising a question-option pair is presented (correct options are highlighted in green and incorrect options in red). The accompanying bar plots visualize the probing results; the y-axis represents probability and the x-axis represents the options. Error bars in the plots depict the effects of temperature adjustments. Both bar plots are identical but apply different masks.

analysis, we here fix the temperature at 0.1 to test all these LLMs. Finally, the average score of each model on the benchmark will be ranked.

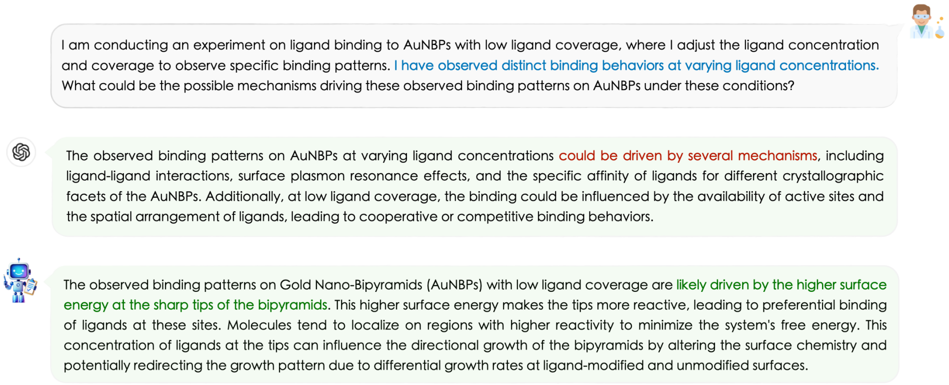
As illustrated in Figure 3a, all models significantly surpass the random guessing baseline of 25% with a remarkable margin, consistent with their capabilities in general tasks.

In this evaluation, Claude and GPT-4 emerged as the top-performing models, achieving accuracies of 84.8% and 81.2%, respectively, significantly outstripping other open-source models which generally scored around 70% or lower. This discrepancy underscores a substantial performance gap between the leading models and their counterparts. The superior performance of Claude and GPT-4 across a broad spectrum of benchmarks can be attributed to their robust training regimens. In contrast, the relatively poorer performance of open-source models may stem from factors such as the quality and type of training data and the constraints of model size. Notably, the Mixtral-8x7B model demonstrates performance comparable to the Vicuna series, with an accuracy of approximately 70.4%. Gemma, at the lower end of the scale, records an accuracy of 44.7%, which, while being the lowest in this study, still significantly exceeds the 25% accuracy expected from random guessing. The accuracy levels of other models occupy a middle ground and are detailed in the supplementary information (Table S1). Additionally, a trend observed in this study is the correlation between model size and accuracy; larger models such as GPT-4 and Claude not only excel in reasoning but also in recognition of mechanisms, further illustrating the advantages of increased scale.

In summary, the performance of these tested LLMs demonstrates their capability to elucidate physico-chemical principles. This showcases their potential future utility in explaining synthetic mechanisms effectively.

3.3 Results of C-scores with Temperature

Knowledge probing is a method designed to assess the capacity of language models, such as those in the GPT series, to recognize and recall specific knowledge domains. This technique evaluates the model's comprehension by analyzing the probability distribution of tokens corresponding to the logits in the model's output. Since the prediction of the next token is governed by the distribution of logits



I am conducting an experiment on ligand binding to AuNBPs with low ligand coverage, where I adjust the ligand concentration and coverage to observe specific binding patterns. I have observed distinct binding behaviors at varying ligand concentrations. What could be the possible mechanisms driving these observed binding patterns on AuNBPs under these conditions?

The observed binding patterns on AuNBPs at varying ligand concentrations could be driven by several mechanisms, including ligand-ligand interactions, surface plasmon resonance effects, and the specific affinity of ligands for different crystallographic facets of the AuNBPs. Additionally, at low ligand coverage, the binding could be influenced by the availability of active sites and the spatial arrangement of ligands, leading to cooperative or competitive binding behaviors.

The observed binding patterns on Gold Nano-Bipyramids (AuNBPs) with low ligand coverage are likely driven by the higher surface energy at the sharp tips of the bipyramids. This higher surface energy makes the tips more reactive, leading to preferential binding of ligands at these sites. Molecules tend to localize on regions with higher reactivity to minimize the system's free energy. This concentration of ligands at the tips can influence the directional growth of the bipyramids by altering the surface chemistry and potentially redirecting the growth pattern due to differential growth rates at ligand-modified and unmodified surfaces.

Figure 5: **An RAG-based LLM assistant (the bottom one) for gold nanoparticle synthesis mechanism explanation.** The depicted dialogue illustrates a question posed by a researcher. The responses demonstrate that while the original GPT-4 model (top) struggles to clearly elucidate the mechanisms involved, the RAG-based LLM assistant provides a clear and detailed explanation of the mechanisms, showcasing its remarkable capacity to assist human scientists in recognition of complex synthesis processes.

and transformed by the Softmax function, each token is assigned a probability. Our focus is directed towards a subset of tokens with higher probabilities rather than a singular output result, as illustrated in Figure 4. The model responds using a predetermined vocabulary consisting of four tokens: A, B, C, and D. Each option's probability is computed based on the output logits. Using the vicuna-7B model as an example, the c-score assesses the probability associated with the correct answer, whereas accuracy solely indicates whether the response is correct. Typically, a sharper distribution of predicted token probabilities indicates a higher certainty in the model response, and vice versa. This design enables an analysis of the model's responses to varied queries, discerning whether they are grounded in solid theoretical recognition or merely speculative guesses.

In our study, we examine the distribution at the logits layer before the model responses, which is considered an indicator of confidence, to better address the aforementioned question. We further assess this by combining previous accuracy metrics with c-score. In past multiple-choice assessments, the final answer of the model is assumed to be chosen with 100% confidence, meaning that either True or False only. This inspires us to measure its capabilities directly by the percentage of the gold answer's confidence, even the model chooses the drinkable gold answer. We here evaluate the overall confidence using the formulated c-score, which quantifies the confidence level assigned to each correct answer, as detailed in the E.q. 1:

$$c\text{-}score = \frac{1}{N}\sum_{i=1}^{N}\frac{e^{L_G^i}}{e^{L_A^i}+e^{L_B^i}+e^{L_C^i}+e^{L_D^i}} \tag{1}$$

where L_G^i is the probability (or confidence) of gold answer regarding the i-th question and L_X is for other options. The probability of all options (assume four here) is normalized exponentially and N is the number of all questions.

Specifically, based on the evaluation of accuracy, we further evaluate the top-5 ranked open-source models with c-score for efficient comparison, considering only models that excel in the benchmark. These models represent typical scales in the open-source community. The evaluation process will also be repeated with different temperature settings, as mentioned before (0.1, 0.3, 0.5, 0.7 and 0.9). Finally, the average c-scores of top-5 models will be ranked for discussion due to their competitive performances.

Regarding the results, as shown in Figure 3, the Vicuna-33B exhibits a lower level of performance, whereas the other models showed slight improvements in the c-score compared to accuracy. In detail, the c-score of Mixtral-8x7B improves by about 8% compared to the accuracy, while the c-score of Vicuna-33B decreases by about 6%. This indicates a clear confidence difference between the two models. Similarly, Vicuna-13B, Mistral-7B and Vicuna-7B demonstrate remarkable differences between accuracy and c-scores. The results indicate that the c-score effectively measures the ability of LLMs in gold nanoparticle synthesis tasks in an interpretable manner. Such metrics suggest that utilizing c-score allows for a more appropriate assessment of language models compared with pure

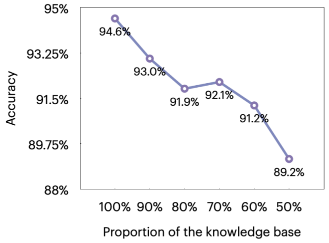


Figure 6: Results of accuracy with different knowledge corpus sizes.

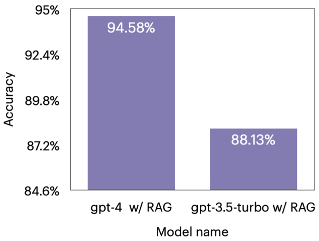


Figure 7: Results of accuracy with model differences.

accuracy, revealing insights distinct from traditional accuracy statistics. We believe that this design can help guide models toward a deeper recognition of mechanisms, potentially enhancing the training process of LLMs in the future.

To illustrate the knowledge probing process, we present representative results in Figure 4, demonstrating the temperature effects with error bars. Here, option A is correct, because the higher the supersaturation (reduction rate), the more unbalanced the growth of the gold particle morphology will be, and thus a nanoparticle morphology with a high-index crystal face or high curvature will be developed. Options B, C, and D have opposite statements. For each model tested, except for option B, there is a tendency to choose A, C, and D – with a higher confidence. Among them, Mistral-7B has a confidence of nearly 100% for the correct option A, and a tendency for other options is almost 0. This result shows that the model has sufficient and solid learning of this knowledge point, and can distinguish the nanosynthesis logic involved in this question, while other models are more confused. One possible reason is that the model is interfered by certain keywords, resulting in a confidence level of about 50%. Some other examples of knowledge probing are in supplementary information (Note S3).

3.4 RAG-based Assistant for Synthesis Mechanism Explanation

Judging from the evaluation results, the LLM demonstrates a notable ability to comprehend synthesis mechanisms, making it a promising tool for assisting materials scientists in their studies. This capability spans extensive literature learning and natural language reasoning. In this work, we have utilized collected data on mechanisms to establish a knowledge base focused on the synthesis of gold nanoparticles. Leveraging this knowledge, we employed RAG technology to develop an intelligent assistant capable of explaining the mechanisms of gold nanoparticle synthesis through a chat system. This intelligent assistant can accurately respond to inquiries about mechanisms by retrieving relevant text information, thus aiding in the comprehension of material synthesis processes, as illustrated in Figure 5.

To implement this system, we utilized the GPT-4-turbo and the Claude model as the reasoning core of our assistant. The knowledge base was structured in a condition-observation-mechanism pairwise fashion, cataloging existing studies on gold nanoparticle synthesis. For retrieving relevant records, we used the text embedding model text-embedding-ada-002 provided by OpenAI, creating a vector database for efficient search. When users pose questions, the assistant initially retrieves the most relevant entries from the knowledge base via similarity matching before responding, citing the source records.

To assess the efficiency of our developed assistant, we conducted a benchmark using a question-and-answer format. Notably, these two models ranked as the top performers in our evaluation, though they are both closed-sourced. We used accuracy as our comparison metric, finding that the RAG-based assistant using GPT-4-turbo achieved an accuracy of 94.58%, while the assistant utilizing the Claude model reached 93.81%, which is nearly 10% higher than the top-accuracy of the Claude model without RAG. However, it was observed that the performance of the Claude model with RAG was not as competitive as that of the GPT-4-turbo model. This discrepancy may be attributed to the Claude model's in context learning ability, specifically in this domain.

Considering the efficiency and high-quality question-answering capability of the GPT-4-turbo-based assistant, we believe this design could significantly advance the application of LLM-based techniques in aiding nanomaterials synthesis mechanisms study.

3.4.1 Ablation Study

We conducted further experiments to explore the impact of knowledge base size on the performance of the assistant. By selecting 90%, 80%, 70%, 60%, and 50% of the sub-knowledge bases from the original full-size knowledge base, we systematically assessed the assistant's performance across these varied conditions.

The results, as illustrated in Figure 6, indicate a significant influence of the knowledge base size on the assistant's efficacy. There was a noticeable decrease in performance, specifically a 6.8% drop, when the knowledge base was reduced to 50% of its original size. It is important to note that this corpus encompasses the majority of gold nanoparticle synthesis methods and structure control mechanisms. Interestingly, even with an increase in the corpus from 70% to 80%, only a minor decrease in performance was observed, and performance still improved when the size was expanded back to 100% of the original. These findings suggest that the retrieval performance of the assistant is not highly sensitive to increments in knowledge base size, indicating robustness in the assistant's underlying retrieval mechanisms.

Additionally, we substituted the original GPT-4 model with the GPT-3.5-turbo model, both developed by OpenAI, to evaluate performance differences when each is paired with the RAG system. We opted not to use the Claude model due to accessibility issues.

As depicted in Figure 7, the GPT-3.5-turbo model achieved an accuracy of 88.1%. There is a performance gap of 6.5% between GPT-3.5-turbo and GPT-4 when both are equipped with RAG. Notably, in scenarios without RAG, the accuracy gap between these two models widens to 15%. This result highlights the significant role of the RAG assistant in enhancing the models' ability to explain synthesis mechanisms effectively.

4 Related Work

In the field of materials science, existing datasets predominantly support tasks focused on factual knowledge, such as named entity recognition and classification. [9, 35, 8, 32, 19, 14, 41, 16, 40] Researchers utilize these datasets to benchmark the performance of language models in the materials domain, see Table 1. Previously, three key chemistry-related capabilities in LLMs, recognition, reasoning, and explaining have been identified, and a benchmark containing eight chemistry tasks has been established. [14] Meanwhile, the potential of large language models to perform scientific synthesis, inference, and explanation across many domains for scientific discovery has been discussed, although this approach is based solely on knowledge graph inference. [41] To expend the task diversity, LLMs such as GPT-3 have been benchmarked on datasets spanning the chemical space, including molecules, materials, and reactions, across diverse tasks such as classification, regression, and inverse design. [16] With the continuous growing of the LLMs, a dataset of 650 challenging questions from the materials domain, requiring the knowledge and skills of a materials science student who has completed their undergraduate degree, has been curated. [40]

5 Conclusion

In this study, we developed a novel evaluation framework and benchmark to assess LLMs' capabilities in materials science, with a specific focus on AuNP synthesis. Our methodology centred on a comprehensive set of 775 multiple-choice questions, evaluating various mainstream LLMs through both accuracy metrics and confidence scores (c-scores). The results demonstrate that c-scores combined with the benchmark effectively differentiate whether LLMs' responses are based on the grasping of physicochemical mechanisms, rather than random outputs. Building on these insights, we developed an RAG-based assistant to explain gold nanoparticle synthesis mechanisms, boosting both efficiency and accuracy. This assistant helps material scientists reduce the hallucination in mechanism-related question-answering, thus aiding the exploration of more synthesis mechanisms.

References

[1] M. R. AI4Science and M. Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *ArXiv*, abs/2311.07361, 2023.

[2] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. 2023.

[3] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[4] C.-H. Chiang and H.-y. Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[5] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[6] K. Choudhary, B. L. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. N. Choudhary, A. Agrawal, S. J. L. Billinge, E. A. Holm, S. P. Ong, and C. M. Wolverton. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8:1–26, 2021.

[7] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*, 2021.

[8] K. Cruse, A. Trewartha, S. Lee, Z. Wang, H. Huo, T. He, O. Kononova, A. Jain, and G. Ceder. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. *Scientific data*, 9(1):234, 2022.

[9] A. Dunn, Q. Wang, A. M. Ganose, D. Dopp, and A. Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6:1–10, 2020.

[10] A. Q. J. et al. Mixtral of experts. *ArXiv*, abs/2401.04088, 2024.

[11] J. B. et al. Qwen technical report. *ArXiv*, abs/2309.16609, 2023.

[12] O. J. A. et al. Gpt-4 technical report. 2023.

[13] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, and X. Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Neural Information Processing Systems*, 2023.

[14] T. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, X. Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

[15] K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

[16] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, and B. Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, pages 1–9, 2024.

[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[18] Y. Kang and J. Kim. Chatmof: An autonomous ai system for predicting and generating metal-organic frameworks. *arXiv preprint arXiv:2308.01423*, 2023.

[19] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444, 2017.

[20] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. R. Joty, and J. Huang. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[21] N. Li, P. Zhao, and D. Astruc. Anisotropic gold nanoparticles: synthesis, properties, applications, and toxicity. *Angewandte Chemie*, 53 7:1756–89, 2014.

[22] E. O. Pyzer-Knapp, J. W. Pitera, P. W. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, 2022.

[23] N. Rampal, K. Wang, M. Burigana, L. Hou, J. Al-Johani, A. Sackmann, H. S. Murayshid, W. A. Al-Sumari, A. M. Al-Abdulkarim, N. E. Al-Hazmi, et al. Single and multi-hop question-answering datasets for reticular chemistry with gpt-4-turbo. *arXiv preprint arXiv:2405.02128*, 2024.

[24] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[25] S. So, T. Badloe, J.-K. Noh, J. Bravo-Abad, and J. Rho. Deep learning enabled inverse design in nanophotonics. *Nanophotonics*, 9:1041 – 1057, 2020.

[26] Y. Sun and Y. Xia. Shape-controlled synthesis of gold and silver nanoparticles. *Science*, 298:2176 – 2179, 2002.

[27] N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, and G. Ceder. Toward autonomous design and synthesis of novel inorganic materials. *Materials Horizons*, 8(8):2169–2198, 2021.

[28] B. Tang, Y. Lu, J. Zhou, T. Chouhan, H. Wang, P. Golani, M. Xu, Q. Xu, C. Guan, and Z. Liu. Machine learning-guided synthesis of advanced inorganic materials. *Materials Today*, 41:72–80, 2020.

[29] H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik, and E. Kumacheva. Nanoparticle synthesis assisted by machine learning. *Nature reviews materials*, 6(8):701–716, 2021.

[30] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[31] A. R. Thawani, R.-R. Griffiths, A. Jamasb, A. Bourached, P. Jones, W. McCorkindale, A. Aldrick, et al. The photoswitch dataset: a molecular machine learning benchmark for the advancement of synthetic chemistry. 2020.

[32] V. Venugopal, S. Sahoo, M. Zaki, M. Agarwal, N. N. Gosvami, and N. A. Krishnan. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7), 2021.

[33] F. Wang, Y. Han, C. S. Lim, Y. Lu, J. Wang, J. Xu, H. Chen, C. Zhang, M. Hong, and X. Liu. Simultaneous phase and size control of upconversion nanocrystals through lanthanide doping. *Nature*, 463:1061–1065, 2010.

[34] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. V. Katwyk, A. Deac, A. Anandkumar, K. J. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. K. Manrai, D. S. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Velickovic, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, and M. Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023.

[35] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702, 2019.

[36] Y. Xia, Y. Xiong, B. Lim, and S. E. Skrabalak. Shape-controlled synthesis of metal nanocrystals: simple chemistry meets complex physics? *Angewandte Chemie*, 48 1:60–103, 2009.

[37] Z. Xiao, W. Li, H. Moon, G. W. Roell, Y. Chen, and Y. J. Tang. Generative artificial intelligence gpt-4 accelerates knowledge mining and machine learning for synthetic biology. *bioRxiv*, 2023.

[38] S. Xing, L. H. Tan, M. Yang, M. Pan, Y. Lv, Q. Tang, Y. Yang, and H. Chen. Highly controlled core/shell structures: tunable conductive polymer shells on gold nanoparticles and nanochains. *Journal of Materials Chemistry*, 19:3286–3291, 2009.

[39] R. X. Yang, C. A. McCandler, O. Andriuc, M. Siron, R. Woods-Robinson, M. K. Horton, and K. A. Persson. Big data in a nano world: A review on computational, data-driven design of nanomaterials structures, properties, and synthesis. *ACS Nano*, 16:19873 – 19891, 2022.

[40] M. Zaki, N. A. Krishnan, et al. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024.

[41] Y. Zheng, H. Y. Koh, J. Ju, A. T. Nguyen, L. T. May, G. I. Webb, and S. Pan. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*, 2023.

[42] Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes, and O. M. Yaghi. A gpt-4 reticular chemist for guiding mof discovery. *Angewandte Chemie*, page e202311983, 2023.

[43] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, and O. M. Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, 2023.