Supplementary information for:

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

Table of contents

Supplementary Note S1. Methodologies on the prompt engineering for creating the benchmark.

Creating a benchmark involves two steps: extracting experimental content (conditions and observations) and creating questions and answers, as well as manually correcting biased option distributions.

Firstly, OpenAI provides the GPT-4 model and the capability to parse PDF documents. We use its web services to read and analyze pre-collected PDF documents. To ensure the quality of content extraction and to allow for manual review and correction, as well as considering the cost of the experiment, we have decided to implement this via the web rather than calling an API. Each PDF uploading includes additional instructions to obtain the necessary content and to ensure it conforms to the required format. Here is the instruction below:

> I have a deep interest in the findings presented in a specific research paper, particularly concerning the correlations it explores. This study seems to delve into how altering certain variables like temperature and solution volume can impact the structure of nanoparticles. These shifts in nanoparticle structure are pivotal in understanding their behavior and applications. Could you meticulously review this paper and provide a **detailed** summary of the key correlations identified? In doing so, please adhere to the following requirements:
> - You **MUST** only give me all the correlations clearly, with the precise conditions, observations and mechanisms mentioned in the paper.
> - You should add some additional information as background information, such as what system, what method of synthesis, what experiments, what adjustments, etc.
> - You **MUSN'T** give any explanation of how this paper describes them, just extract the correlations.
> - Please clarify the starting conditions in the study and specify the exact changes made, such as adjustments in temperature or solution volume, to understand their

impact on nanoparticle structures, all related information should be in the condition part (see the example below).

- You **MUST** give detailed conditions, precise in numbers and statements.
- For each adjustment-observation pair, you **MUST** match the tendency precisely, for example, the **increase** of temperature gives **smaller** particle size or **increased** ligand coverage.
- You **MUST** extract the explanation of **intrinsic mechanisms mentioned in the paper**.
- Make the extracted correlations more detailed and append more background information, remember!
- Please answer in this standard JSON format (an example here):

```
{
        condition: {
                example key: example value
        },
        observation: [
                {
                        adjustment: example adjustment,
                        observation: example observation,
                        mechanism: example mechanism,
                }
        ] ,
}
```

Secondly, for the manually inspected condition-observation pairs, we further generated a corresponding number of test questions by invoking the GPT-4 API and provided standard answers. Here is the instruction:

Given the dialog:

{{input text}}

Task: Transform the following records (including conditions and observations) into a multiple-choice question for an evaluation case. The question should test the understanding and knowledge of large language models in terms of the **condition-observation correlations** based on the given text.

Requirements:

1. Uniqueness of the Correct Answer: The question **must** be designed in such a way that it has **only one correct answer**.

2. Construction of Incorrect Options: The incorrect answers should be derived by altering key aspects of the correct option to ensure they are **plausible yet distinguishable**.

3. Inclusion of Condition Object Information: **All details** present in the specifed "condition" object must be integrated into the question.

4. Knowledge Assessment Focus: The question should be aimed at evaluating **the logic of knowledge** possessed by large language models.

5. Deliverables: **Provide the question along with its multiple-choice options**, indicating clearly which option is the correct answer.

6. Exclusivity of Content: **Refrain** from adding any explanatory notes or additional information beyond the question and its options.

7. Formal and Logical Structure: Ensure the question is formulated in a formal and logical manner, suitable for an official evaluation setting.

8. You **MUST** follow the template:

Question: ...

Given options:

A. ...

```
B. ...
C. ...
D. ...
Correct Answer: [X]
"""
```

In order to achieve a balanced distribution of options, we reorganized the question options in the evaluation set by modifying the correct options of some questions. This is achieved by swapping the order of the options, and the algorithm will ensure that the probability of each option appearing is close.

Supplementary Note S2. Baseline models information.

Vicuna Paradigm. Vicuna represents a pioneering effort within the open-source community. This model, conceptualized and refined by LMSYS, undergoes an extensive fine-tuning process leveraging the LLaMA series models, trained on a dataset comprising 70,000 user-generated dialogs. Furthermore, Vicuna is distinguished as one of the preeminent models within the subset of LLaMA-2 fine-tuned models (for vicuna-v1.3 and v1.5 versions), attributed to its superior training quality and the voluminous corpus of data it utilizes.

Mistral and Mixtral architectures. Developed by Mistral.AI, these models represent two distinct approaches within the field of AI. The Mistral model integrates a Grouped-Query Attention mechanism to enhance inference speed and employs Sliding Window Attention to efficiently manage extended sequences with reduced computational demands. Conversely, the Mixtral model adopts an innovative architecture characterized by a high-quality Sparse Mixture of Experts. This decoder-only framework enables the feedforward block to select from eight unique parameter groups, with a router network at each layer determining the optimal combination of two groups, known as experts, to process each token and amalgamate their outputs in an additive fashion.

Qwen series. These models are meticulously fine-tuned using a dataset curated to align with a diverse range of tasks, including conversation, tool utilization, agency, and safety protocols. A notable distinction of the Qwen models lies in their token representation capacity. The 7B model processes 2.4 trillion tokens, while the 14B model handles 3.0 trillion tokens. This positions Qwen at the pinnacle of token representation capabilities compared to other models in its category.

Gemma framework. The Gemma model encapsulates a series of lightweight, cutting-edge open-source models that draw from the same foundational research and technological advancements underpinning the Gemini models. This lineage of models is celebrated for their formidable performance metrics, notably in comparison to the GPT-4 model.

Other frameworks. In the evaluation of models including, but not limited to, GPT-4 (gpt-4-0125-preview) and Claude 3 (claude-3-ops-20240229), our analysis endeavors to assess them based upon their sophisticated capabilities in solving general problems. It is pertinent to note that these models are not made available as open-source; nevertheless, they are developed through the training on extensive corpus of data, encompassing a wide array of domains. This approach underscores the depth and breadth of knowledge these models can potentially harness, despite the proprietary nature of their development methodologies. In addition, Gemini is not considered due to its accessibility, thus, we use Gemma for a case to test its behavior.

Supplementary Note S3. Examples of knowledge probing.

In this section, we present several representative Knowledge Probing results. The questions and options were meticulously designed by researchers with professional expertise, ensuring that the similarities between the options were on par with benchmark levels. The results exhibit the confidence levels of five models for these questions. We converted the logits values output by the models for options A through D into numbers ranging from 0 to 1 to represent the relative confidence between the options.

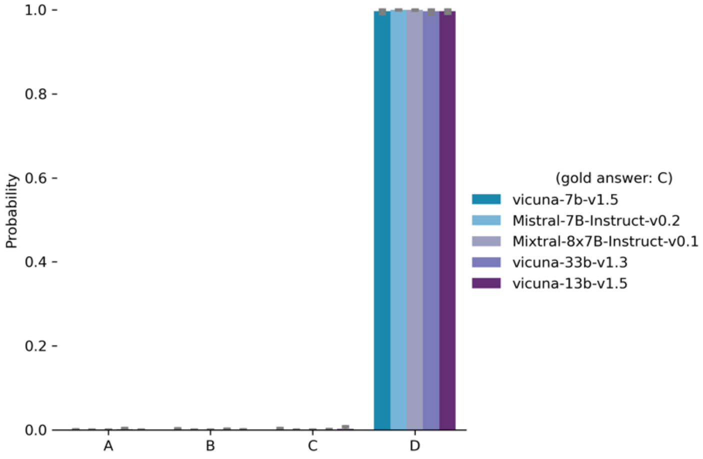
Case 1. In a typical seed-mediated growth of gold nanostructures, with the involvement of surfactant CTAB, Au source HAuCl4, the reductant ascorbic acid, and thiol ligands such as glutathione (GSH), with the decreasing of seed concentration, the morphology of the resulting nanoparticles tends to be inequivalent with more protruding features. Which of following explanation is reasonable?

A. The decrease in seed concentration would lead to a lower Au deposition rate, as a result, the growth is focused onto the whole surface.

B. The decrease in seed concentration would lead to a higher Au deposition rate, as a result, the growth is focused onto the whole surface.

C. The decrease in seed concentration would lead to a higher Au deposition rate, as a result, the growth is focused on a few active sites.

D. The decrease in seed concentration would lead to a lower Au deposition rate, as a result, the growth is focused on a few active sites and develops inequivalent morphologies, including protruding ridges, grooves, and curved tips.



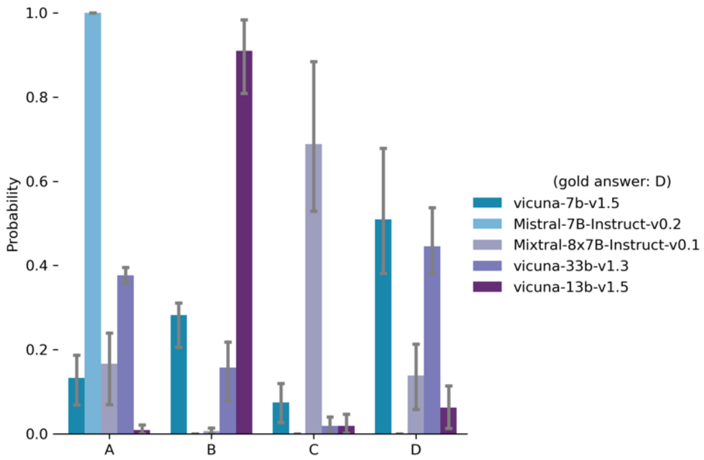
Case 2. In a typical seed-mediated growth of gold during the advanced colloidal synthesis of chiral gold nanoparticles, which involves the utilization of amino acids or peptides as chiral promoters, hydrochloric acid as a gold precursor, ascorbic acid as a reducing agent, CTAB for colloidal stability, and pre-existing gold nanoparticles to facilitate heterogeneous nucleation. Which one of the following phenomena is commonly observed during the synthesis?

A. Elevated concentrations of chiral inducers predominantly lead to the formation of nanostructures exhibiting more pronounced chiral responses.

B. An increase in the concentration of chiral biomolecules is typically correlated with the emergence of nanostructures possessing simpler morphologies, such as sharply defined surfaces, which manifest stronger chiral responses.

C. Diminishing concentrations of chiral inducers are generally associated with the production of nanostructures that exhibit reduced chiral responses.

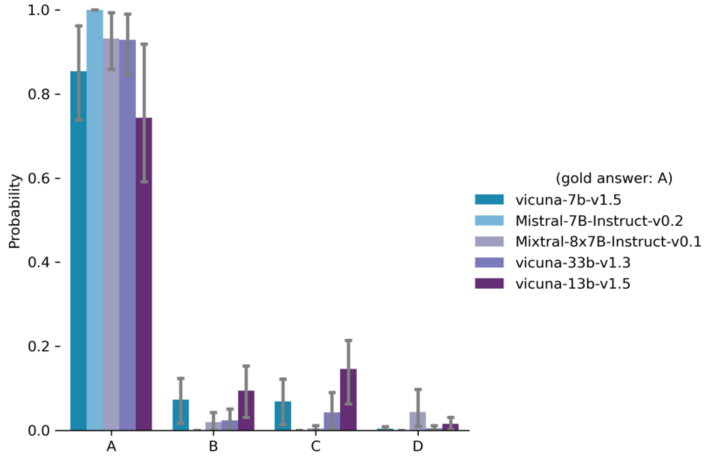
D. The augmentation of chiral biomolecule concentrations frequently results in the development of more complex morphologies, characterized by notable chiral twisting and the presence of surface undulations.



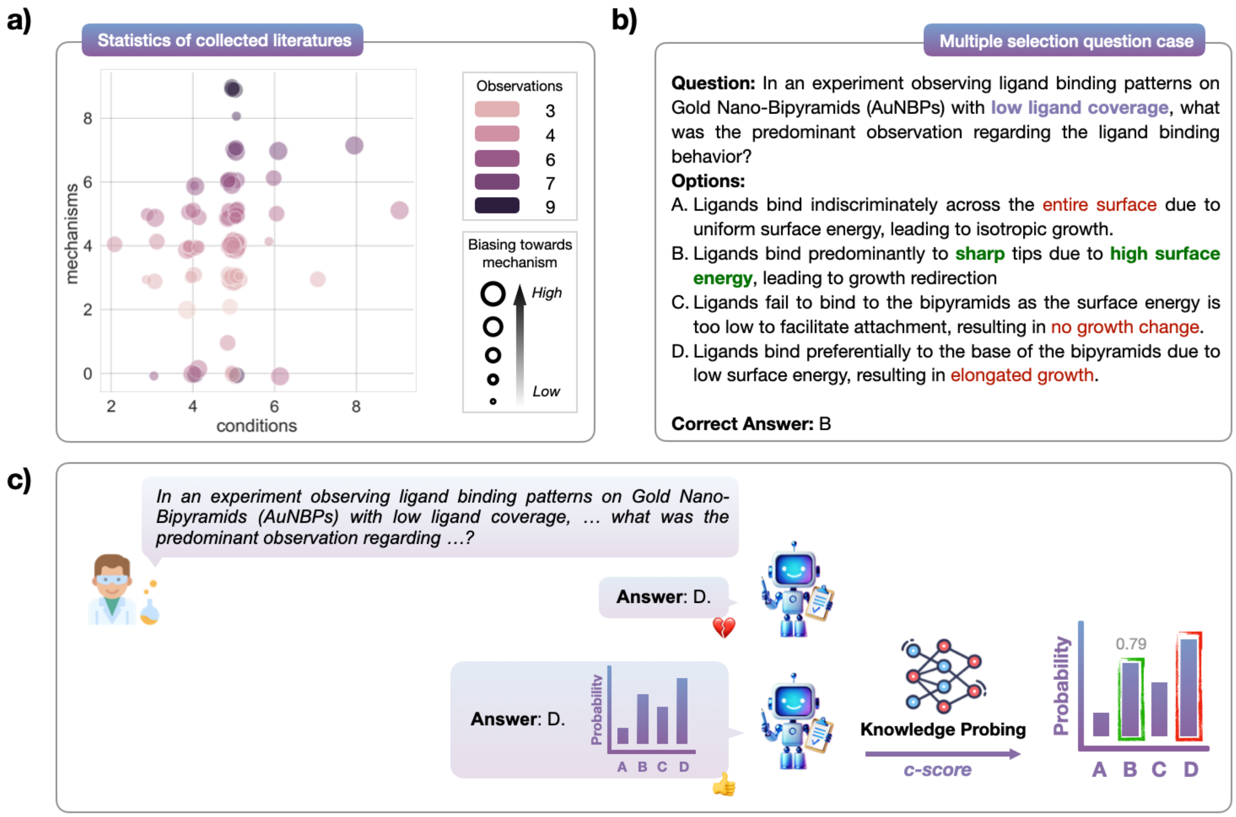
Case 3. During the seed-mediated colloidal synthesis of gold nanoparticles, employing hydrochloric acid as the gold precursor, ascorbic acid as the reduction agent, and CTAB for stabilization, with the initial seed nanoparticles characterized by anisotropy in shape (varied crystal facets and differing curvature sites), which phenomenon is prevalently observed throughout the overgrowth phase?

A. In conditions of elevated growth rates or supersaturation of the colloidal solution, the preferential attachment of gold atoms occurs at sites of high surface energy, including areas with marked curvature, defects, considerable strain, or minimal passivation by capping agents, leading to nanostructures that deviate from the equilibrium crystal shape.

B. Under conditions of reduced growth rates or lower supersaturation of the colloidal solution, gold atoms tend to deposit preferentially at high-energy sites, resulting in nanostructures that diverge significantly from the equilibrium crystal morphology.
C. With higher growth rates or solution supersaturation, gold atoms predominantly attach to high-energy sites, including areas with marked curvature, defects, considerable strain, or minimal passivation by capping agents, facilitating the formation of nanostructures that conform to equilibrium crystal shapes, characterized by well-defined facets.
D. In scenarios of diminished growth rates or supersaturation, the deposition of gold atoms favors sites of lower surface energy, such as flat crystal facets, promoting the development of nanostructures that align with the equilibrium crystal configuration, exhibiting well-faceted surfaces.



Supplementary Figure S1. The collected data set for evaluation.



Question: In an experiment observing ligand binding patterns on Gold Nano-Bipyramids (AuNBPs) with low ligand coverage, what was the predominant observation regarding the ligand binding behavior?

Options:

A. Ligands bind indiscriminately across the entire surface due to uniform surface energy, leading to isotropic growth.

B. Ligands bind predominantly to **sharp** tips due to **high surface energy**, leading to growth redirection

C. Ligands fail to bind to the bipyramids as the surface energy is too low to facilitate attachment, resulting in no growth change.

D. Ligands bind preferentially to the base of the bipyramids due to low surface energy, resulting in elongated growth.

Correct Answer: B

In an experiment observing ligand binding patterns on Gold Nano-Bipyramids (AuNBPs) with low ligand coverage, … what was the predominant observation regarding …?

Figure S1. The collected data set for evaluation. **a)** statistical analysis of collected literatures. Each point represents a scientific report. The size is up to the relevance of the mechanism determined by the manual assessment. To facilitate easier visualization, we applied a jittering technique, meaning that each point was moved slightly by adding a small random values to the integer coordinates. **b)** One case in the benchmark, with the question and options to be sent to models, and the answer will be collected through a refining process (by an extractor prompted with specific structured information extraction prompts). **c)** Our evaluation metric, c-score, which forces the model to provide the confidence of the gold answer (**instead of the responded option**) by probing the model's output logits.

Supplementary Figure S2. The accuracy curve of the model during the test of 775 questions.

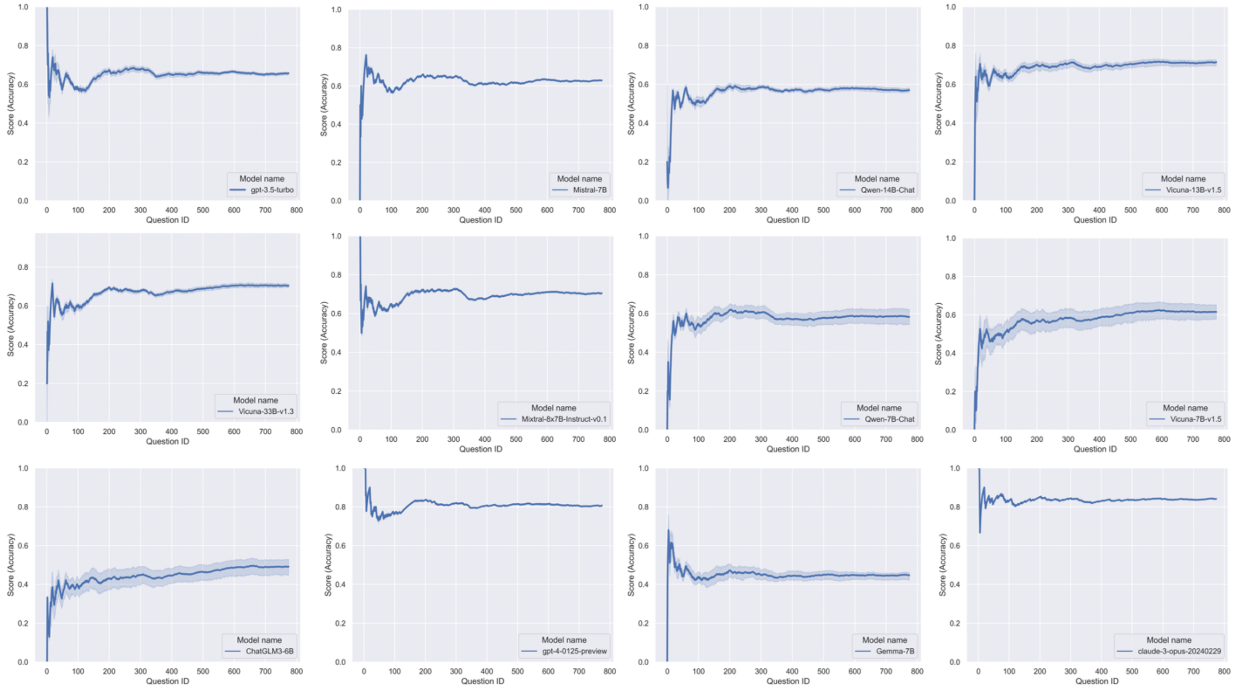


Figure S2. The accuracy curve of the model during the test of 775 questions.

Supplementary Figure S3. Temperature effects analysis on answering nano materials synthesis questions.

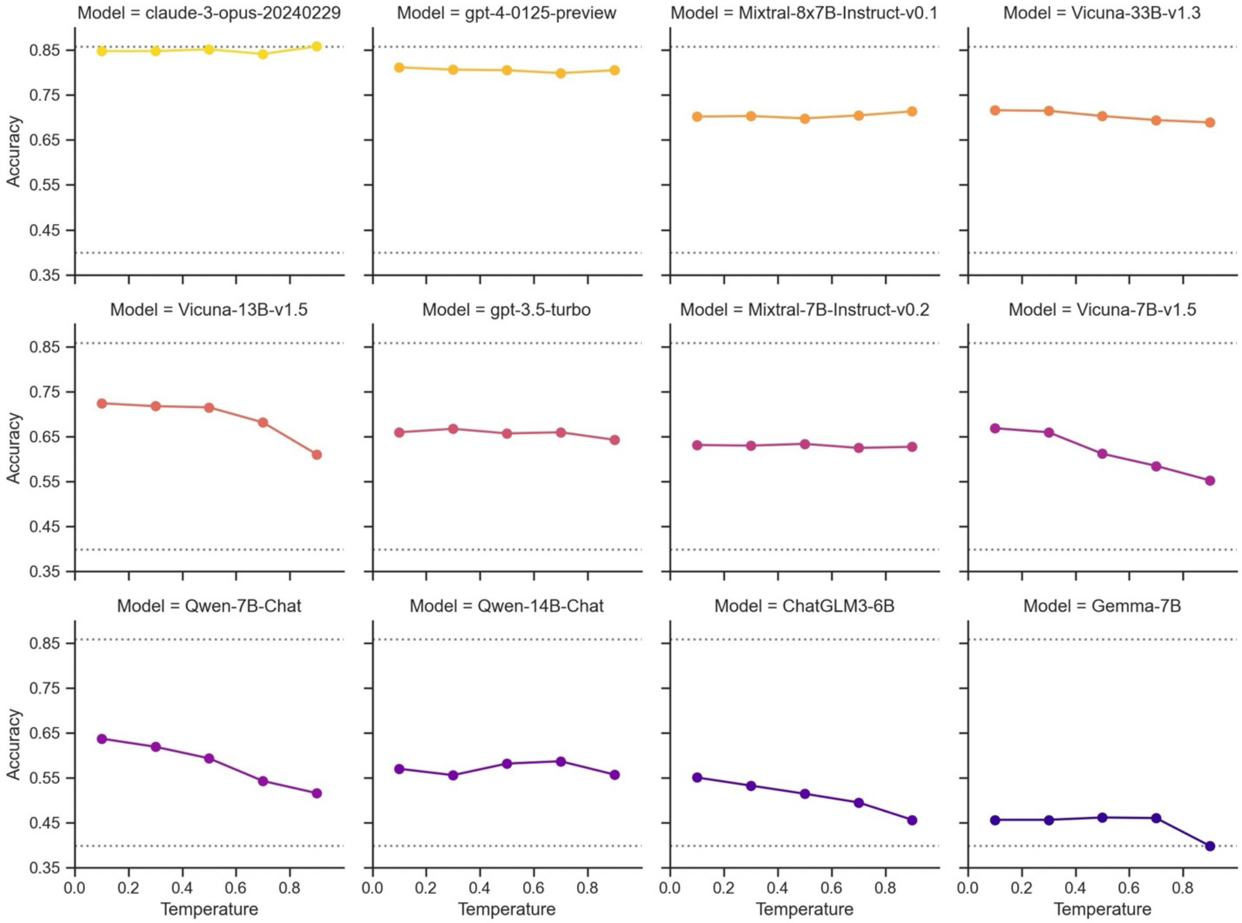


Figure S3. Temperature effects analysis on answering nano materials synthesis questions.

Model name	**Accuracy**
claude-3-opus-20240229	0.8477
gpt-4-0125-preview	0.8116
Mixtral-8x7B-Instruct-v0.1	0.7019
Vicuna-33B-v1.3	0.7161
vicuna-13b-v1.5	0.7239
gpt-3.5-turbo	0.6594
vicuna-7b-v1.5	0.6684
Mistral-7B	0.6310
Qwen-7B-Chat	0.6374
Qwen-14B-Chat	0.5703
chatglm3-6b	0.5510
Gemma-7B	0.4568

Table S1. Accuracy of all the tested models. Mixtral-8x7B and Vicuna-33B are identified by their temperatures as they both in a high-similar accuracy.