CatAgent: Multi-Agent Orchestration for Electrocatalyst Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

The discovery of efficient hydrogen evolution reaction catalysts is hindered by the combinatorial scale of candidate material space, which limits conventional screening approaches. Here we present CatAgent, an autonomous multi-agent workflow driven by large language models that achieves a 2.27-fold increase in discovery rates over the random baseline. The framework integrates hypothesis generation, surrogate-model evaluation, and iterative critic feedback to navigate chemical space toward optimal hydrogen adsorption strength. We benchmark 13 state-of-the-art language models in 'single-shot mode' and 'iterative CatAgent mode' across alloy compositions spanning $L1_0$ and $L1_2$ crystallographic phases. Critic-enabled iterations improve performance for most model architectures, though self-refinement effectiveness varies across model families. Top-performing models such as Gemini 3 Flash and GPT-5.2 concentrate proposals near zero theoretical overpotential, identifying successful candidates twice as efficiently as random sampling. In contrast, outdated models such as GPT-3.5 Turbo fail to formulate consistent hypotheses and exhibit unreliable search behavior. This work provides the first demonstration that state-of-the-art language models can autonomously reshape high-throughput catalyst screening into reasoning-guided exploration.

1 Introduction

Scalable renewable energy storage demands efficient electrocatalysts for hydrogen production, yet state-of-the-art systems rely on scarce noble metals such as platinum and iridium [1]. This limitation drives an intensive search for earth-abundant alternatives, with bimetallic alloys emerging as promising candidates due to their tunable electronic properties. For the hydrogen evolution reaction (HER), the hydrogen adsorption energy (ΔE_H) serves as a foundational descriptor for catalytic activity, following a well-established volcano relationship where optimal performance correlates with near-thermoneutral binding [2, 3]. Specifically, targeting a free energy of adsorption near zero requires an electronic adsorption energy of approximately -0.27 eV, once zero-point energy and entropy corrections are accounted for [4, 5].

However, navigating the vast chemical space of electrocatalysts presents a significant combinatorial challenge [6]. Conventional high-throughput screening is computationally intensive, often requiring exhaustive sampling of myriad adsorption sites, surface terminations, and adsorption configurations [7]. Such enumeration typically relies on density functional theory (DFT), which remains prohibitively expensive for large-scale exploration. To overcome these bottlenecks, researchers have increasingly turned to active learning [8] and machine learning interatomic potentials [9], which can accelerate discovery by an order of magnitude or more compared to manual exploration [10]. These frameworks allow for rapid navigation through thousands of intermetallic configurations by prioritizing promising candidates for high-fidelity evaluation.

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.

Recently, the integration of large language models (LLMs) [11, 12] and transformer architectures [13] into materials science has introduced new paradigms for catalyst design [14–16]. Models such as CatBERTa and CatGPT have demonstrated that transformer-based representations can substitute expensive 3D descriptors in early-stage screening and enable inverse design of catalyst structures [17, 18]. Beyond these surrogate models, LLM-driven agentic workflows have emerged for fully autonomous experimentation. The A-Lab achieved a 63% success rate in synthesizing computationally predicted inorganic materials through closed-loop integration of ML-based synthesis planning, robotic experimentation, and active learning [19]. Similarly, multi-agent frameworks such as DREAMS have shown that LLM-based planners can orchestrate complex DFT simulation workflows with expert-level accuracy [20], while agentic systems for metal-organic framework discovery have demonstrated the capacity to autonomously navigate vast combinatorial spaces through iterative hypothesis generation and validation [21]. Frameworks such as Adsorb-Agent have further illustrated how LLM-driven reasoning can strategically identify stable adsorption configurations without exhaustive geometric enumeration by combining proposal engines with surrogate models such as graph neural networks [22]. These advancements demonstrate that LLM-based reasoning can drive autonomous navigation of chemical space, motivating the development of agentic workflows for catalyst discovery.

In this study, we investigate LLM-driven candidate selection for the screening of bimetallic HER catalysts within the $L1_0$ and $L1_2$ composition space. We compare the efficacy of a single-shot LLM approach against an iterative agentic workflow, CatAgent, with and without an integrated critic module for refined reasoning. Using a pretrained graph neural network potential (uma-s-1p1)[23] as surrogate model for rapid property evaluation, we assess how different LLM-orchestration strategies navigate bimetallic surface chemistry toward optimal hydrogen adsorption strength. The main contributions of this work are:

- A comprehensive benchmark of 13 state-of-the-art LLMs across three major providers (Google Gemini, OpenAI GPT, and Anthropic Claude), establishing quantitative metrics for chemical reasoning over 1,998 bimetallic compositions.
- CatAgent, a multi-agent framework integrating hypothesis generation, surrogate-model simulation, and critic feedback, achieving discovery rates up to 2.27-fold above random baseline.
- Systematic analysis revealing architecture-dependent search behaviors: top-performing models (Gemini 3 Flash, GPT-5.2) concentrate proposals near optimal adsorption energy, while outdated models (GPT-3.5 Turbo) exhibit dispersed exploration.

2 Methodology

2.1 Dataset

The search space for catalyst discovery is defined by bimetallic alloys composed of 37 metals: Al, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, La, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, and Bi [24]. These metals are combined into two distinct composition families: the $L1_2$ structure with a stoichiometry of metal1:metal2 = 0.75:0.25, and the $L1_0$ structure with a stoichiometry of 0.50:0.50. The total candidate space consists of 1,998 unique entries, comprising 1,332 $L1_2$ and 666 $L1_0$ candidates. For each candidate, adsorption sites are enumerated according to symmetry, resulting in 9 unique sites for $L1_2$ and 10 for $L1_0$.

Adsorption energies are calculated for every enumerated site, and the representative value for a given composition is defined as the minimum adsorption energy across all sites. The adsorption energy ΔE_{H} is calculated relative to gas-phase $\mathrm{H_2}$ as follows:

$$\Delta E_{\mathrm{H}} = E_{\mathrm{slab+H}} - E_{\mathrm{slab}} - \frac{1}{2}E_{\mathrm{H_2}}.$$

The target adsorption energy is -0.27 eV, corresponding to a near-thermoneutral free energy of adsorption ($\Delta G_{\mathrm{H}} = \Delta E_{\mathrm{H}} + 0.24\,\mathrm{eV} \approx 0$) that maximizes HER activity according to Sabatier's principle [4, 8]. A successful candidate is defined as falling within the window of $-0.57\,\mathrm{eV} \leq \Delta E_H \leq 0.03\,\mathrm{eV}$. The baseline dataset distribution corresponds to the minimum H adsorption energies across the 1,998 candidates, utilizing Universal Models for Atoms (UMA, uma-s-1p1 checkpoint) as a surrogate evaluator [23]. Among machine learning interatomic potentials benchmarked for hydrogen

adsorption energy prediction, UMA-s achieves the lowest total MAE (0.168 eV) combined with high structural reliability (84.58% normal relaxation rate), making it well-suited for rapid screening [25, 26].

2.2 CatAgent workflow

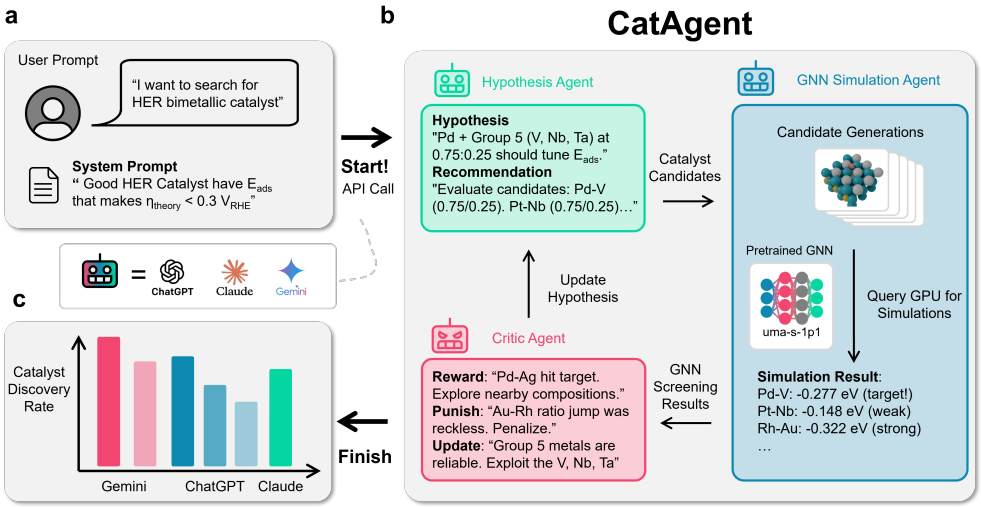

Figure 1: Schematic representation of the CatAgent multi-agent workflow architecture. (a) User and system prompts specify the screening objective and acceptance criterion. (b) Iterative loop of LLM-driven hypothesis formulation, candidate suggestion, UMA-based energy evaluation, and critic feedback for successive refinement. (c) Performance evaluation of screening outcomes.

Two primary screening approaches are employed: single-shot screening and the iterative CatAgent workflow. In the single-shot mode, the large language model (LLM) performs a combined role of hypothesis formulation and candidate generation to propose 200 candidates in a single non-iterative pass (Fig. 2a). This mode was evaluated using 13 models: Gemini 3 Flash, Gemini 3 Pro, Gemini 2.5 Flash, Gemini 2.5 Pro, GPT-5.2, GPT-5, GPT-5 mini, GPT-5 nano, GPT-4o mini, GPT-3.5 Turbo, Claude Opus 4.5, Claude Haiku 4.5, and Claude Sonnet 4.5.

The CatAgent workflow utilizes a multi-agent architecture comprising LLM-driven hypothesis, simulation, and critic agents. After the user specifies the screening objective and acceptance criterion (Fig. 1a), each run explores 200 candidates over 20 steps, proposing 10 candidates per step. We investigate two variants of this workflow: a critic-enabled and a critic-disabled configuration. In the critic-enabled variant, the critic agent evaluates stepwise outcomes to identify recurring success and failure patterns, and its feedback is used by the hypothesis agent to refine the search strategy in subsequent steps (Fig. 1b). After termination of the iterative loop, overall screening performance is evaluated using the metrics defined in Section 2.3 (Fig. 1c), enabling systematic comparison of screening outcomes across models and workflow configurations. The workflow was benchmarked using Gemini 3 Flash, GPT-5.2, GPT-4o mini, and GPT-3.5 Turbo; for all models and providers, default sampling hyperparameters were used.

The simulation agent executes an evaluation pipeline using UMA: (i) construction of the bulk phase, (ii) bulk relaxation, (iii) construction of a (2×2) three-layer slab with 20 Å of vacuum, (iv) slab relaxation, (v) adsorption site enumeration, (vi) relaxation of the H+slab adsorbed structure with the bottom two layers fixed, and (vii) calculation of ΔE_{H} using the gas-phase H_2 energy. Geometry optimizations were performed using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm with a force criterion of $f_{\max} = 0.05$ eV/Å and a maximum of 100 steps. The codes, exact model snapshots, full prompts, and all raw outputs for the single-shot and CatAgent workflows are available at `https://osf.io/t8he3/overview?view_only=1255c0a1cb1646bd9616d634314cd5e9`.

2.3 Metrics and statistics

The performance of the screening process is evaluated using the success probability, defined as:

$$P_{\text{success}} = \frac{|\{i \,|\, \Delta E_{\text{H},i} \in [-0.57,\ 0.03]\,\text{eV}\}|}{200}$$

where the success window corresponds to the target adsorption energy of -0.27 eV $\pm$ 0.30 eV. The random-choice success probability is defined as the probability of selecting a successful candidate from the baseline distribution. Among the 1,998 candidates in the search space, 699 fall within the success window, yielding a random baseline of $699/1998 = 35.0\%$. The improvement factor is calculated as the ratio of the observed success probability to this random baseline; a value of 1.0 indicates performance equivalent to random selection, while values exceeding unity suggest an ability to identify compositions closer to the target adsorption energy.

Token usage is reported as the sum of input and output tokens. Cached tokens are excluded to ensure consistent comparison across providers where caching metadata may not be available. For $L1_0$ compositions, symmetric pairs are considered equivalent and deduplicated within and across steps. All statistical values derived from triplicate runs are reported as mean $\pm$ standard deviation.

3 Results and discussion

3.1 Intrinsic performance comparison of language models

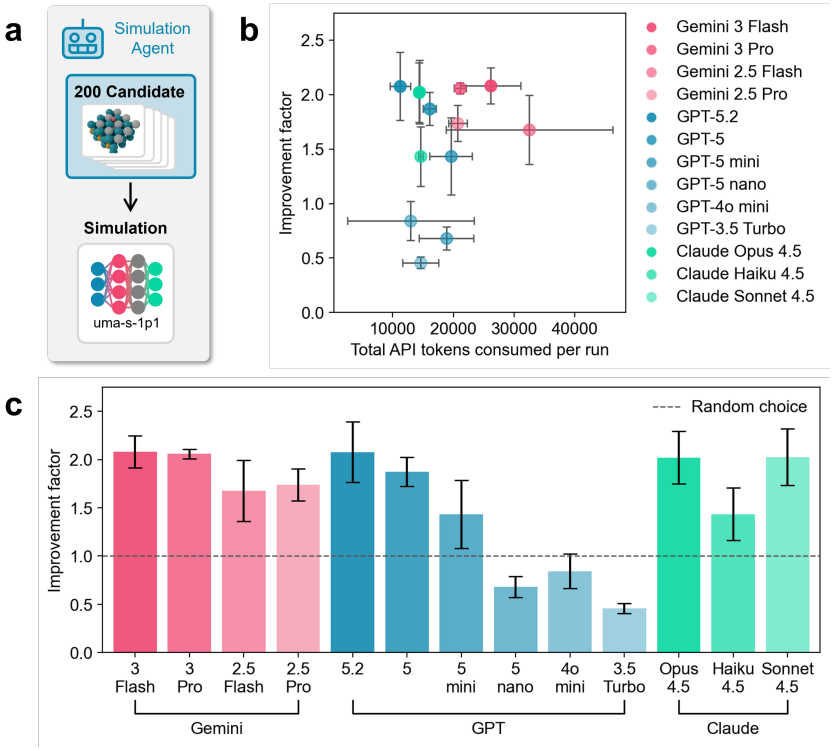

Figure 2: Single-shot screening performance analysis. (a) Workflow schematic. (b) Relationship between discovery efficiency and computational resource consumption across the 13 language models. (c) Improvement factors of the language models. Mean $\pm$ standard deviation from triplicate runs.

Single-shot benchmarking reveals substantial variation in chemical reasoning capabilities across the 13 language models evaluated (Fig. 2a). Resource consumption differs markedly among models, ranging from approximately 26,120 tokens per trial for Gemini 3 Flash to considerably lower values for more efficient architectures (Fig. 2b). Despite this variation in computational cost, performance differences are even more pronounced: Gemini 3 Flash achieves the highest mean improvement

factor of 2.08, demonstrating robust navigation of the chemical composition space in a single attempt, whereas GPT-3.5 Turbo exhibits an improvement factor of only 0.46, performing below the random baseline (Fig. 2c). This disparity across model architectures indicates that LLM training and design fundamentally determine single-shot predictive accuracy for catalyst discovery tasks.

Further analysis reveals that the number of unique candidates explored correlates strongly with the resulting improvement factor. While the experimental target is 200 unique catalyst candidates per trial, several models fail to meet this goal; GPT-4o mini averages only 116 candidates. Notably, resource efficiency does not track directly with overall performance: GPT-5.2 requires an average of only 78 tokens per successful candidate, representing a favorable balance between computational cost and discovery effectiveness. These metrics provide a comprehensive baseline for intrinsic model capabilities before introducing iterative agentic workflows.

3.2 CatAgent performance comparison across language models

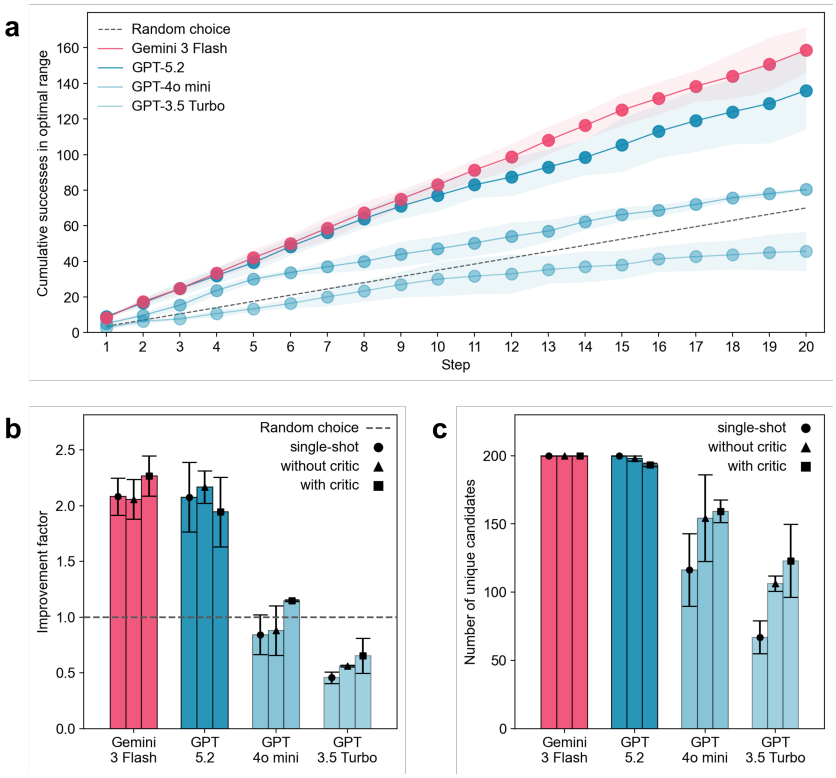

Figure 3: CatAgent workflow performance across model architectures. (a) Cumulative discovery of successful candidates during critic-guided iterative search. (b) Improvement factors and (c) unique candidate counts for single-shot, non-critic iterative, and critic-enabled iterative workflows across Gemini 3 Flash, GPT-5.2, GPT-4o mini, and GPT-3.5 Turbo. Mean $\pm$ standard deviation from triplicate runs.

The introduction of the CatAgent workflow allows for a direct comparison of discovery performance across different architectural designs: the single-shot baseline, the CatAgent workflow without critic, and the CatAgent workflow with critic. This analysis is focused on a subset of four models: Gemini 3 Flash, GPT-5.2, GPT-4o mini, and GPT-3.5 Turbo. As demonstrated in Fig. 3b, for Gemini 3 Flash, the addition of a critic module is consistent with an increase in the improvement factor of approximately 0.21 compared to the iterative workflow lacking a critic. Similarly, GPT-4o mini demonstrates a notable improvement in its improvement factor, rising from 0.88 to 1.15 when the critic module is integrated. These results suggest that iterative self-correction and the inclusion of an evaluation step generally enhance the model's ability to narrow down the chemical search space toward the desired adsorption energy target.

However, the efficacy of the critic module is not universal across all model architectures. Contrary to the general trends observed in other models, GPT-5.2 exhibits a lower improvement factor when a critic is employed compared to iterative trials conducted without one. In some instances, success probability can fall below the level of a random choice, particularly for models like GPT-3.5 Turbo. This performance deficit can be attributed to unfocused search behaviors or a failure to generate the required number of unique candidates. Additionally, the integration of a critic can influence the volume of output (Fig. 3c); with a critic, GPT-4o mini explores approximately 159 unique candidates per trial, which represents an increase over its single-shot baseline of 116.

3.3 Analysis of catalyst compositions discovered via CatAgent

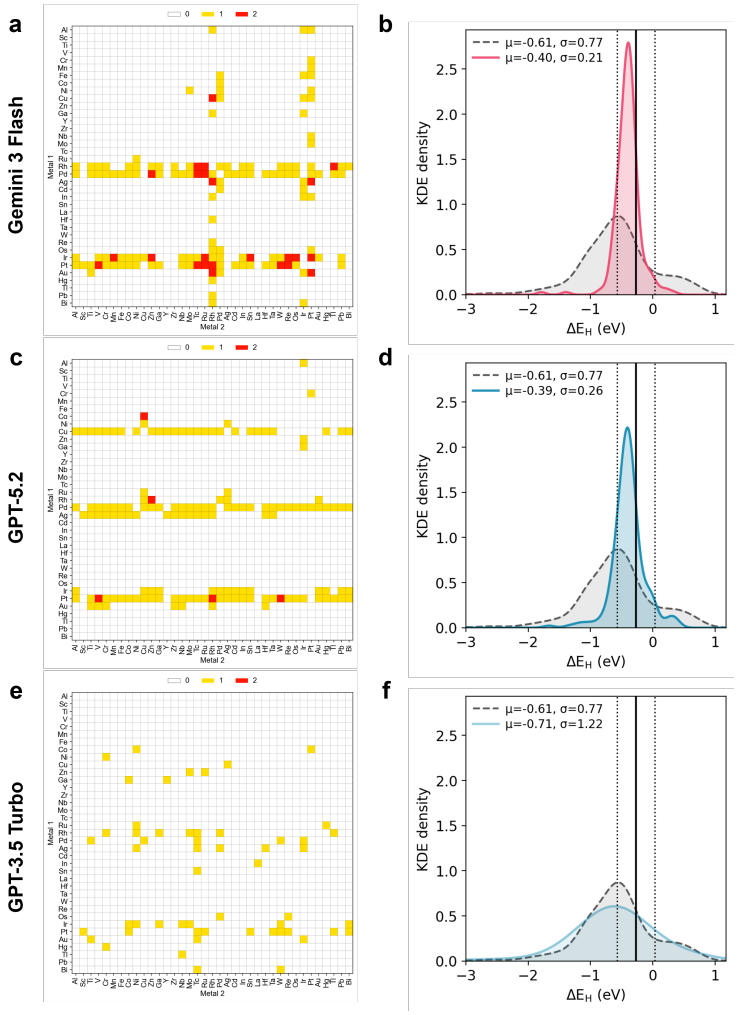

Figure 4: (a, c, e) Heatmaps showing the composition distributions of candidates with adsorption energies within the energy threshold and (b, d, f) KDE plots showing the adsorption-energy distributions of all catalysts identified in one repeat of the critic-enabled CatAgent. (a,b) Gemini 3 Flash; (c,d) GPT-5.2; (e,f) GPT-3.5 Turbo.

A qualitative analysis of the catalyst compositions discovered through the CatAgent workflows reveals different search behaviors, categorized as either focused or dispersed. The performance of the models is reflected in the adsorption-energy distribution mean and standard deviation relative to the broader dataset distribution. Models that exhibit focused search behaviors generate candidates with an energy distribution that has a smaller standard deviation and a mean shift toward the target, whereas dispersed search behaviors result in candidates that more closely resemble the original dataset's broad distribution.

Gemini 3 Flash exhibited a clear shift toward the target energy when using the CatAgent workflow with critic (Figs. 4a,b, and Fig. 5). The mean absolute deviation from the target decreased from 0.34 eV at baseline to 0.12 eV across three trials. This focusing effect is also evident in the narrowing of the energy distribution, with the standard deviation of adsorption energies reduced to an average of 36.7% of the baseline dataset value.

GPT-5.2 demonstrated the most precise alignment with the target energy among the models analyzed (Figs. 4c,d, and Fig. 6). In the CatAgent workflow with critic, GPT-5.2 achieved a mean absolute deviation of 0.08 eV from the target. The model shifted the mean candidate adsorption energy by 0.26 eV relative to the dataset baseline and reduced the standard deviation to 47.4% of the dataset value. This combination of a significant mean shift and a reduced standard deviation indicates that GPT-5.2 employs a focused search strategy that identifies compositions near the desired adsorption energy.

In contrast, GPT-3.5 Turbo exhibited dispersed search behavior with limited success in approaching the HER target (Figs. 4e,f, and Fig. 8). In the CatAgent workflow with critic, GPT-3.5 Turbo shifted the mean candidate adsorption energy by -0.05 eV relative to the dataset baseline across three trials, representing a move away from the target direction. Coupled with its reduced candidate generation count of 123, the resulting energy distribution remained largely unfocused, consistent with its lower normalized success rate. The contrasting energy distribution characteristics across these models highlight the varying capacity of language models to interpret and act upon chemical design targets within an agentic framework.

4 Conclusion

CatAgent, a multi-agent workflow driven by large language models, demonstrates substantial effectiveness in navigating complex chemical spaces for catalyst discovery. Through systematic evaluation of 13 language models in single-shot screening and four models within iterative multi-agent configurations, we establish that LLM-guided exploration can strategically identify promising bimetallic hydrogen evolution reaction catalysts from a search space of 1,998 compositions spanning $L1_0$ and $L1_2$ crystallographic phases.

The CatAgent framework integrates hypothesis generation, computational evaluation, and critic feedback mechanisms to achieve discovery rates up to 2.27-fold above random baseline selection. The integration of critic modules demonstrates architecture-dependent efficacy, with models such as GPT-4o mini exhibiting substantial performance improvements while GPT-5.2 showing decreased effectiveness, highlighting the heterogeneous nature of self-refinement capabilities across model families. Notably, performance remains robust across three major API providers—Google, OpenAI, and Anthropic—validating the generalizability of the approach beyond specific model implementations. Leading models like Gemini 3 Flash demonstrate focused search behaviors that concentrate candidate proposals near the target adsorption energy while substantially reducing distribution variance relative to the baseline dataset. Resource efficiency varies significantly across LLM architectures, with optimal configurations requiring as few as 78 tokens per successful candidate.

However, several opportunities for advancement remain. First, the current search space is relatively constrained. Scaling to broader compositional ranges, alternative crystal structures, and multi-component systems would better test framework capabilities. Second, whether domain-specific fine-tuning on materials science data could enhance chemical reasoning remains unexplored. Third, extending CatAgent to multi-objective optimization would enhance practical utility, as real-world catalyst design requires balancing activity, selectivity, stability, cost, and synthesizability simultaneously. This work establishes LLM-powered autonomous reasoning as a promising paradigm for accelerating electrocatalyst discovery in vast chemical spaces. Continued development in LLM reasoning capacity, prompting strategies, and experimental integration will be essential for realizing the full potential of AI-assisted materials discovery.

References

[1] C Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020.

[2] Hyeonjung Jung, Seokhyun Choung, and Jeong Woo Han. Design principles of noble metal-free electrocatalysts for hydrogen production in alkaline media: combining theory and experiment. *Nanoscale Advances*, 3(24):6797–6826, 2021.

[3] Seokhyun Choung, Heejae Yang, Jinuk Moon, Wongyu Park, Hyeokjoon June, Chaesung Lim, and Jeong Woo Han. Theoretical tuning of local coordination environment of metal-nitrogen doped carbon catalysts for selective chlorine-evolution reaction. *Catalysis Today*, 425:114358, 2024.

[4] Jens Kehlet Nørskov, Thomas Bligaard, Ashildur Logadóttir, John R. Kitchin, Jingguang G. Chen, Svetlozar Pandelov, and Ulrich Stimming. Trends in the exchange current for hydrogen evolution. *Journal of The Electrochemical Society*, 152(2):J23–J26, 2005. doi: 10.1149/1.1856988.

[5] Xiangyun Xiao, Sungsu Kang, Seokhyun Choung, Jeong Woo Han, Jungwon Park, and Taekyung Yu. Synthesis of metal cation doped nanoparticles for single atom alloy catalysts using spontaneous cation exchange. *Journal of Materials Chemistry A*, 11(6):2857–2867, 2023.

[6] Seokhyun Choung, Wongyu Park, Jinuk Moon, and Jeong Woo Han. Rise of machine learning potentials in heterogeneous catalysis: Developments, applications, and prospects. *Chemical Engineering Journal*, 494:152757, 2024.

[7] Zhi Wei Seh, Jakob Kibsgaard, Colin F Dickens, IB Chorkendorff, Jens K Nørskov, and Thomas F Jaramillo. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science*, 355(6321): eaad4998, 2017.

[8] Kevin Tran and Zachary W. Ulissi. Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution. *Nature Catalysis*, 1(9):696–703, 2018. doi: 10.1038/s41929-018-0142-1.

[9] Ryan Jacobs, Dane Morgan, Siamak Attarian, Jun Meng, Chen Shen, Zhenghao Wu, Clare Yijia Xie, Julia H Yang, Nongnuch Artrith, Ben Blaiszik, et al. A practical guide to machine learning interatomic potentials–status and future. *Current Opinion in Solid State and Materials Science*, 35:101214, 2025.

[10] L. Kavalsky, V. I. Hegde, E. Muckley, M. S. Johnson, B. Meredig, and V. Viswanathan. By how much can closed-loop frameworks accelerate computational materials discovery? *Digital Discovery*, 2:1112–1125, 2023. doi: 10.1039/D2DD00133K.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne 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.

[15] Negin Orouji, Jeffrey A. Bennett, Richard B. Canty, Long Qi, Shijing Sun, Paulami Majumdar, Chong Liu, Núria López, Neil M. Schweitzer, John R. Kitchin, Hongliang Xin, and Milad Abolhasani. Autonomous catalysis research with human–AI–robot collaboration. *Nature Catalysis*, 8:1135–1145, 2025. doi: 10.1038/s41929-025-01430-6.

[16] Hongliang Xin, John R. Kitchin, and Heather J. Kulik. Towards agentic science for advancing scientific discovery. *Nature Machine Intelligence*, 7:1373–1375, 2025. doi: 10.1038/s42256-025-01110-x.

[17] Janghoon Ock, Chakradhar Guntuboina, and Amir Barati Farimani. Catalyst energy prediction with catberta: Unveiling feature exploration strategies through large language models. *ACS Catalysis*, 13(24): 16032–16044, 2023. doi: 10.1021/acscatal.3c04956.

[18] Dong Hyeon Mok and Seoin Back. Generative pretrained transformer for heterogeneous catalysts. *Journal of the American Chemical Society*, 146(49):33712–33722, 2024. doi: 10.1021/jacs.4c11504.

[19] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624:86–91, 2023. doi: 10.1038/s41586-023-06734-w.

[20] Ziqi Wang, Hongshuo Huang, Hancheng Zhao, Changwen Xu, Shang Zhu, Jan Janssen, and Venkatasubramanian Viswanathan. DREAMS: Density functional theory based research engine for agentic materials simulation. *arXiv preprint*, 2025.

[21] Theo Jaffrelot Inizan, Sherry Yang, Aaron Kaplan, Yen-hsu Lin, Jian Yin, Saber Mirzaei, Mona Abdelgaid, Ali H Alawadhi, KwangHwan Cho, Zhiling Zheng, et al. System of agentic ai for the discovery of metal-organic frameworks. *arXiv preprint arXiv:2504.14110*, 2025.

[22] Janghoon Ock, Tirtha Vinchurkar, Yayati Jadhav, and Amir Barati Farimani. Adsorb-agent: Autonomous identification of stable adsorption configurations via large language model agent. 2024. arXiv:2410.16658.

[23] Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

[24] Osman Mamun, Kirsten T. Winther, Jacob R. Boes, and Thomas Bligaard. High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Scientific Data*, 6:76, 2019. doi: 10.1038/s41597-019-0080-z.

[25] Jinuk Moon, Uchan Jeon, Seokhyun Choung, and Jeong Woo Han. Catbench framework for benchmarking machine learning interatomic potentials in adsorption energy predictions for heterogeneous catalysis. *Cell Reports Physical Science*, 6(12), 2025.

[26] Seokhyun Choung, Miyeon Kim, Jinuk Moon, and Jeong Woo Han. From atomic motif to realistic single atom catalysts through machine learning interatomic potentials. *ACS Energy Letters*, 10(12):6288–6296, 2025.

[27] LangChain AI. Langchain: License (mit license). GitHub repository, . URL `https://github.com/langchain-ai/langchain/blob/master/LICENSE`. License: MIT. Accessed 2026-01-31.

[28] LangChain AI. Langgraph: License (mit license). GitHub repository, . URL `https://github.com/langchain-ai/langgraph/blob/main/LICENSE`. License: MIT. Accessed 2026-01-31.

[29] FAIR-Chem. Fair-chem: License (mit license). GitHub repository. URL `https://github.com/FAIR-Chem/fairchem/blob/main/LICENSE.md`. License: MIT (software). Accessed 2026-01-31.

[30] Meta AI and FAIR Chemistry. Uma model weights: License (fair chemistry license v1). Hugging Face model repository (facebook/UMA). URL `https://huggingface.co/facebook/UMA/blob/main/LICENSE`. License: FAIR Chemistry License v1 (model weights). Accessed 2026-01-31.

[31] OpenAI. Openai services agreement. OpenAI Policies. URL `https://openai.com/policies/services-agreement/`. Terms governing use of OpenAI services for businesses/developers (incl. API). Accessed 2026-01-31.

[32] Google. Gemini api additional terms of service. Google AI for Developers. URL `https://ai.google.dev/gemini-api/terms`. Additional terms for Gemini API usage. Accessed 2026-01-31.

[33] Anthropic. Anthropic commercial terms of service. Anthropic Legal. URL `https://www.anthropic.com/legal/commercial-terms`. Terms governing use of Anthropic API keys and related commercial offerings. Accessed 2026-01-31.

A Appendix

A.1 List of Assets

- **LangChain** [27] [MIT License]
- **LangGraph** [28] [MIT License]
- **FAIR-Chem /** `fairchem-core` [29] [MIT License]
- **UMA pretrained potential** (`uma-s-1p1`) [30] [FAIR Chemistry License v1]
- **OpenAI API + GPT-family model endpoints** [31] [Proprietary (API Terms of Use)]
- **Google Gemini API endpoints** [32] [Proprietary (API Terms of Service)]
- **Anthropic Claude API endpoints** [33] [Proprietary (API Terms of Service)]

A.2 Compositional distribution and energetic profiles of catalysts identified via CatAgent with critic

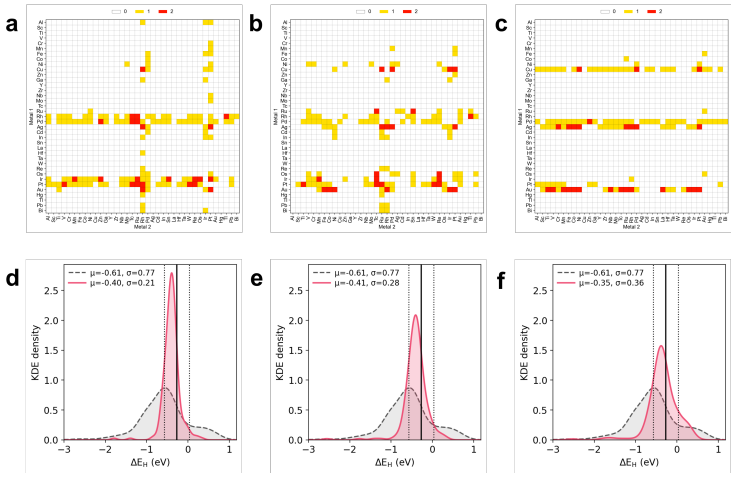

Figure 5: Catalyst discovery patterns for Gemini 3 Flash using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.

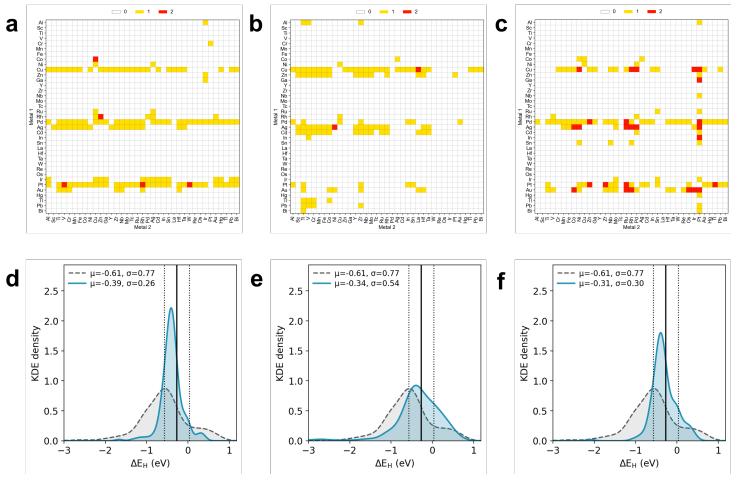

Figure 6: Catalyst discovery patterns for GPT-5.2 using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.

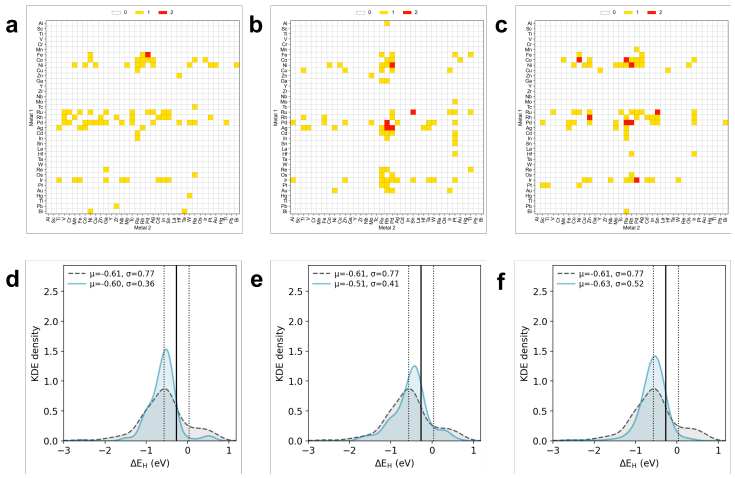

Figure 7: Catalyst discovery patterns for GPT-4o mini using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.

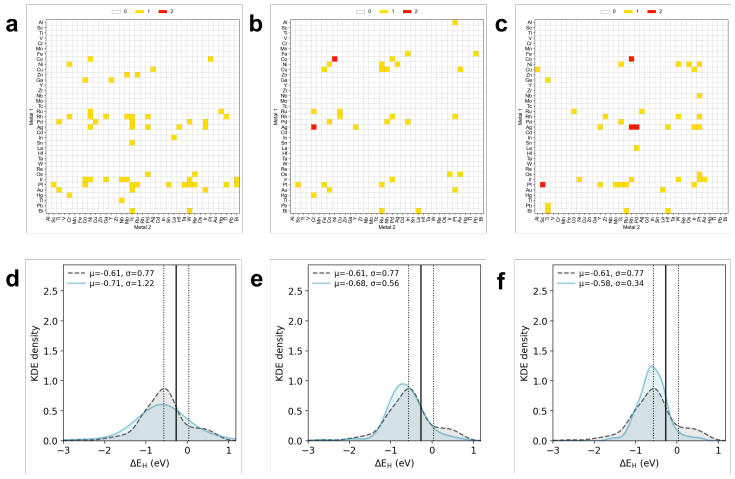

Figure 8: Catalyst discovery patterns for GPT-3.5 Turbo using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction correctly frame the core contribution — CatAgent (multi-agent iterative workflow) and a 13-model single-shot benchmark over bimetallic space — consistent with the Introduction's bullet-point contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the final paragraph of the Conclusion, the paper explicitly acknowledges limitations and open gaps—most notably that the search space is relatively constrained (limited compositions and phases), that domain-specific fine-tuning is not explored, and that the current framework does not yet address multi-objective optimization, framing these as clear directions for advancement.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is an empirical workflow/benchmark study.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Methodology section points to an anonymous artifact that contains the exact model snapshot identifiers, prompts, codes, and run commands needed to reproduce the reported figures. Detailed information can be found at `https://osf.io/t8he3/overview?view_only=1255c0a1cb1646bd9616d634314cd5e9`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The codes, full prompts, model snapshots, and raw data are available at `https://osf.io/t8he3/overview?view_only=1255c0a1cb1646bd9616d634314cd5e9`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the key experimental settings in Methodology, including the candidate budget (200), iterative schedule (20 steps × 10), UMA checkpoint (uma-s-1p1). The codes, experimental settings, full prompts, model snapshots, and raw data are available at `https://osf.io/t8he3/overview?view_only=1255c0a1cb1646bd9616d634314cd5e9`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Uncertainty reporting is defined explicitly: all statistics from triplicate runs are reported as mean ± standard deviation, and the main quantitative figures (Fig. 2 and Fig. 3) state this in their captions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Detailed information can be found at `https://osf.io/t8he3/overview?view_only=1255c0a1cb1646bd9616d634314cd5e9`.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work is a computational study of LLM-guided catalyst screening and does not involve human subjects, crowdsourcing, or personally identifiable data, and it uses standard scientific simulations without deceptive, surveillance, or discriminatory applications. In line with the NeurIPS Code of Ethics, the project's foreseeable impacts are predominantly beneficial (accelerating materials discovery).

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contribution of this work is methodological and empirical—benchmarking LLM-assisted workflows for electrocatalyst discovery.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release any high-risk generative model weights or scraped user data; it releases only workflow code, configuration files, and precomputed catalyst-screening outputs. Access to external LLMs and UMA is mediated through their providers' APIs or gated checkpoints under existing terms.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper (A.1 List of Assets) credits all non-original assets that are essential to the study.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets introduced by the paper (the CatAgent workflow code, the released raw experimental outputs, and the baseline dataset) are documented in `https://osf.io/t8he3/overview?view_only=1255c0a1cb1646bd9616d634314cd5e9`.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The study does not involve crowdsourcing or human-subject experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects or study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.