CATAGENT: MULTI-AGENT ORCHESTRATION FOR ELECTROCATALYST DISCOVERY

Seokhyun Choung[†] **Hoyun Kim**[†] **Jongheun Kim** **Jeong Woo Han**[*]
Department of Materials Science and Engineering, Research Institute of Advanced Materials,
Seoul National University, Seoul, 08826, 1 Gwanak-ro, Gwanak-gu, Republic of Korea
{schoung9967, hoyun423, olympic1234, jwhan98}@snu.ac.kr

ABSTRACT

The discovery of efficient electrocatalysts is hindered by the combinatorial scale of candidate material space. Here we present CatAgent, an autonomous multi-agent workflow driven by large language models that achieves up to a 9.65-fold increase in discovery rates over adsorbate-specific random baselines. We benchmark 13 language models in single-shot and iterative modes across bimetallic alloy compositions. Critic-enabled iterations improve performance for most architectures, with top models concentrating proposals near zero theoretical overpotential. Our results suggest that catalyst screening can benefit from LLM-guided chemical reasoning.

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. Bimetallic alloys are promising alternatives due to their tunable electronic properties. For the hydrogen evolution reaction (HER), the hydrogen adsorption energy (ΔE_{H}) serves as a volcano-type descriptor, with optimal catalytic activity near thermoneutral binding ($\Delta G_{\mathrm{H}} \approx 0$) (Nørskov et al., 2005).

Navigating this vast chemical space is a combinatorial challenge (Choung et al., 2024), as conventional density functional theory (DFT) screening requires exhaustive enumeration of sites, terminations, and configurations (Seh et al., 2017). Active learning (Tran & Ulissi, 2018) and machine learning interatomic potentials (Jacobs et al., 2025) accelerate discovery by prioritizing promising candidates for high-fidelity evaluation. More recently, large language models (LLMs) have been applied to catalyst design (Jablonka et al., 2023; Xin et al., 2025), including transformer-based surrogate models for early-stage screening (Ock et al., 2023; Mok & Back, 2024) and agentic systems that combine proposal engines with graph neural network surrogates for adsorption configuration search (Ock et al., 2024).

Here we introduce CatAgent, a multi-agent framework for electrocatalyst screening, evaluated on a bimetallic HER benchmark spanning the $L1_0$ and $L1_2$ composition space. Using a pretrained graph neural network potential (UMA) as surrogate model (Wood et al., 2025), we benchmark 13 LLMs from three providers in both single-shot and iterative configurations, achieving discovery rates up to 9.65$\times$ above adsorbate-specific random baselines. We choose this 1,998-candidate space because exhaustive evaluation is tractable, so every LLM prediction can be verified against ground truth before scaling to ternary, quaternary, and high-entropy alloy systems where full enumeration is infeasible.

2 METHODOLOGY

2.1 SEARCH SPACE AND ADSORPTION ENERGY CALCULATION

The search space for catalyst discovery is defined by bimetallic alloys composed of 37 metals spanning the 3d (Sc–Zn), 4d (Y–Cd), and 5d (La, Hf–Hg) transition metals along with selected p-block elements (Al, Ga, In, Sn, Tl, Pb, Bi) (Mamun et al., 2019). These metals form 1,998

[*]Corresponding author
[†]Equal contribution

bimetallic candidates across two intermetallic phases: $L1_2$ (3:1 stoichiometry, 1,332 entries) and $L1_0$ (1:1 stoichiometry, 666 entries). Adsorption sites are enumerated by symmetry (9 sites for $L1_2$, 10 for $L1_0$; Supplementary Figure 1), and ΔE_{H} for each composition is taken as the minimum across all sites, calculated relative to gas-phase H_2:

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

The target adsorption energy is -0.27 eV, corresponding to $\Delta G_{\mathrm{H}} \approx 0$ (Nørskov et al., 2005; Tran & Ulissi, 2018). All adsorption energies are evaluated using Universal Models for Atoms (UMA, uma-s-1p1 checkpoint) (Wood et al., 2025), which achieves the lowest total MAE (0.17 eV) with high structural reliability (85% normal relaxation rate) among machine learning interatomic potentials benchmarked for this task (Moon et al., 2025; Choung et al., 2025).

2.2 CatAgent workflow

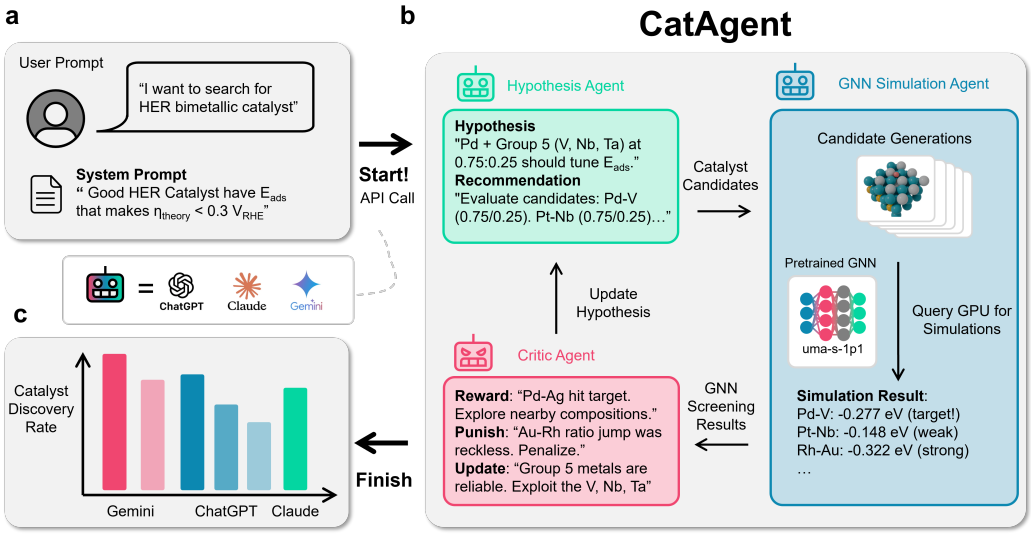

Figure 1: Schematic representation of the CatAgent multi-agent workflow. (a) User prompts specify the screening objective. (b) Iterative loop of LLM-driven hypothesis formulation, candidate suggestion, UMA-based energy evaluation, and critic feedback for successive refinement. (c) Performance evaluation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

CatAgent employs a multi-agent architecture with hypothesis, simulation, and critic agents. After the user specifies the screening objective (Fig. 1a), each run explores 200 candidates over 20 steps (10 per step), sampling roughly 10% of the search space. The LLM selects composition and phase while the simulation agent follows a fixed evaluation protocol. In the critic-enabled variant, the critic analyzes each step's results to identify which host metals, secondary elements, and phases correlate with hits or misses, then returns targeted recommendations to the hypothesis agent for the next iteration (Fig. 1b). Screening performance is evaluated using the metrics in Section 2.3 (Fig. 1c). Default sampling hyperparameters were used for all providers (Supplementary Table 1). In single-shot mode, the LLM proposes all 200 candidates in one pass without feedback (Fig. 2a).

The simulation agent executes an evaluation pipeline using UMA: (i) bulk construction and relaxation, (ii) (2×2) three-layer slab construction (20 Å vacuum) and relaxation, (iii) adsorption site enumeration, (iv) H-adsorbed slab relaxation with the bottom two layers fixed, and (v) ΔE_{H} calculation. All relaxations used LBFGS with $f_{\max} = 0.05$ eV/Å.

Success probability is defined as the fraction of 200 proposed candidates falling within an adsorbate-specific target window: $[-0.57, 0.03]$ eV for *H, $[-0.97, -0.37]$ eV for *CO, $[-0.30, 0.30]$ eV for *N and *NO, and $[-0.10, 0.50]$ eV for *O. The random baseline is defined separately for each adsorbate as the fraction of the 1,998-candidate space falling within its target window. For *H, the baseline is $699/1998 = 35.0\%$; for *CO, *N, *NO, and *O, the corresponding baselines are 5.7%,

15.7%, 4.5%, and 4.4%, respectively. The improvement factor is P_{success} divided by the baseline for the corresponding adsorbate (> 1.0 indicates better-than-random performance). We benchmark against random selection to test whether LLMs carry implicit chemical priors that guide search without engineered descriptors.

3 Results and Discussion

3.1 Single-shot model performance

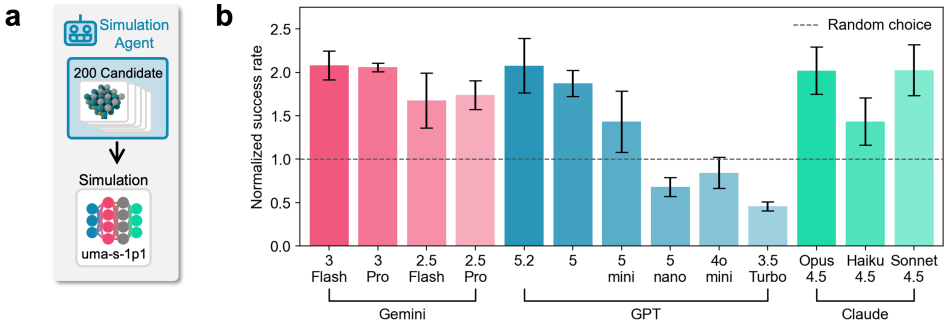

Figure 2: Single-shot screening performance analysis. (a) Workflow schematic. (b) Improvement factors of the 13 language models. Error bars: mean ± s.d. ($n = 3$).

As a control for the iterative experiments, we first benchmark all 13 models in single-shot mode (Supplementary Table 1; Fig. 2a). Improvement factors range from 2.08 (Gemini 3 Flash) to 0.46 (GPT-3.5 Turbo, below random baseline), indicating that model architecture and scale affect chemical reasoning capacity (Fig. 2b). Several models also fail to generate the target 200 unique candidates per trial (e.g., GPT-4o mini averages 116), directly lowering their scores. Resource efficiency does not track with overall performance: GPT-5.2 requires only 78 tokens per successful candidate (Supplementary Figure 2).

To verify that these results are not driven by stochastic variation, we conducted five independent repeats per model and measured element frequency correlation (Pearson r) across repeats (Supplementary Figure 14). Models with high discovery rates also show high repeat-to-repeat consistency: Opus 4.5, Sonnet 4.5, and Gemini 3 Pro all exceed $r = 0.9$, while GPT-5 nano and GPT-3.5 Turbo fall near zero. This positive correlation between reproducibility and discoverability (Supplementary Figure 13) suggests that element preferences of top-performing models reflect stable internal priors rather than stochastic outputs.

3.2 Iterative CatAgent workflow performance

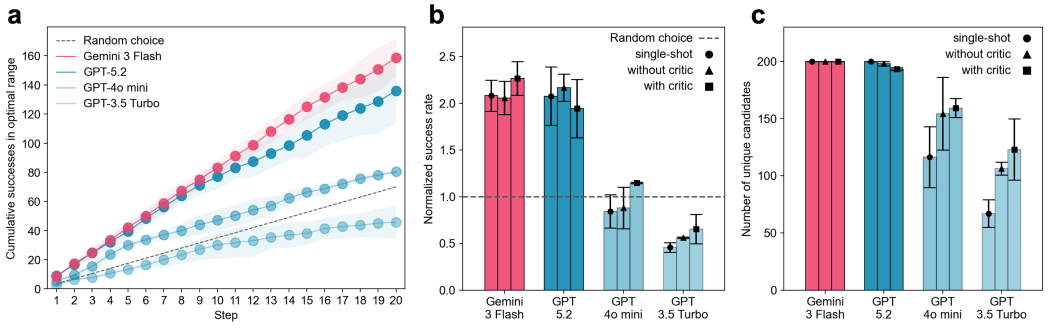

Figure 3: CatAgent performance. (a) Cumulative successful discoveries during iterative search. (b) Improvement factors and (c) unique candidate counts across three workflow configurations. Mean ± s.d. ($n = 3$).

We select four models spanning the single-shot performance range for iterative evaluation: Gemini 3 Flash (top), GPT-5.2 (strong), GPT-4o mini (moderate), and GPT-3.5 Turbo (below baseline). Each iterative run requires roughly 20× more API calls than single-shot mode (Ock et al., 2024). Cumulative discovery curves show that top-performing models steadily outpace random baseline throughout the 20-step search (Fig. 3a), and critic-enabled self-correction generally enhances discovery (Fig. 3b): the critic increases Gemini 3 Flash's improvement factor by 0.21 and lifts GPT-4o mini from 0.88 to 1.15. However, critic efficacy is not universal. GPT-5.2 exhibits a lower improvement factor with the critic than without, and GPT-3.5 Turbo remains below random baseline. The critic also affects output volume (Fig. 3c), with GPT-4o mini generating 159 unique candidates versus 116 in single-shot mode. Total API costs across all experiments were approximately $55 (OpenAI), $30 (Google), and $10 (Anthropic). LLM inference accounts for less than 2% of wall-clock time; the remainder is UMA structure relaxation.

3.3 Catalyst selection patterns and adsorbate generalization

Adsorption-energy distributions reveal two distinct search behaviors (Supplementary Figures 3–7). High-performing models narrow the candidate distribution toward the target energy: with the critic, Gemini 3 Flash reduces its mean absolute deviation from the target to 0.12 eV and compresses the standard deviation to 36.7% of baseline, while GPT-5.2 achieves 0.08 eV deviation and 47.4% compression.

Both top models preferentially select Pt-group metals (Pt, Ir, Rh, Pd) as the host element while varying the secondary component, a chemically sound strategy given their position near the HER volcano apex. Reaching target energies still requires selecting the right secondary metal and phase. Gemini 3 Flash further explores both $L1_0$ and $L1_2$ phases for these hosts, while GPT-family models show narrower phase coverage, possibly reflecting differences in training data composition. In contrast, GPT-3.5 Turbo produces a dispersed distribution with only 123 candidates. Cross-model comparison of element selection frequencies (Supplementary Figure 15) shows high agreement ($r > 0.8$) among top-performing models across all three providers, with GPT-5 nano as a clear outlier. That independently trained models converge on similar element preferences provides additional evidence that these selections reflect shared chemical knowledge rather than arbitrary patterns.

To assess generalizability beyond *H, we extended the evaluation to four additional adsorbates (*CO, *NO, *O, *N) using Gemini 3 Flash with the critic-enabled workflow. Across three trials, CatAgent consistently outperforms the adsorbate-specific random baseline, with improvement factors of 4.04 (*CO), 9.65 (*NO), 7.73 (*O), and 2.81 (*N) (Supplementary Figures 9–12). The maximum improvement factor is 9.65 for *NO. Improvement factors are higher for adsorbates with lower baseline success rates (4.4% for *O vs. 15.7% for *N), suggesting greater benefit in harder search spaces. That performance holds across adsorbates with distinct optimal compositions indicates the model relies on more than memorized HER catalyst associations.

4 Conclusion

We presented CatAgent, an LLM-driven multi-agent framework for bimetallic catalyst screening. Benchmarking 13 models across single-shot and iterative configurations, we find that critic-enabled iteration improves discovery for most architectures, with the best configuration achieving up to 9.65× improvement over adsorbate-specific random baselines. Critic efficacy varies by architecture: GPT-4o mini benefits, whereas GPT-5.2 shows decreased effectiveness. This difference reveals heterogeneous self-refinement capabilities across model families. Leading models concentrate proposals near the target adsorption energy while reducing distribution variance.

Several open questions remain. The LLM currently selects only composition and phase while structure generation follows a fixed protocol; expanding its role to site selection or structure optimization is the next step. Adding descriptor-based baselines such as Bayesian optimization or d-band center ranking, which require feature engineering, would better contextualize LLM performance against established screening methods. Extending the framework to ternary and high-entropy alloy systems, testing under tighter energy windows, and incorporating multi-objective criteria (activity, stability, cost) will be important next steps.

REFERENCES

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.

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.

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.

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

A List of Assets

- **LangChain** (LangChain AI, a) [MIT License]
- **LangGraph** (LangChain AI, b) [MIT License]
- **FAIR-Chem / `fairchem-core`** (FAIR-Chem) [MIT License]
- **UMA pretrained potential (`uma-s-1p1`)** (Meta AI & FAIR Chemistry) [FAIR Chemistry License v1]
- **OpenAI API + GPT-family model endpoints** (OpenAI) [Proprietary (API Terms of Use)]
- **Google Gemini API endpoints** (Google) [Proprietary (API Terms of Service)]
- **Anthropic Claude API endpoints** (Anthropic) [Proprietary (API Terms of Service)]