SUPPLEMENTARY MATERIAL FOR: 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

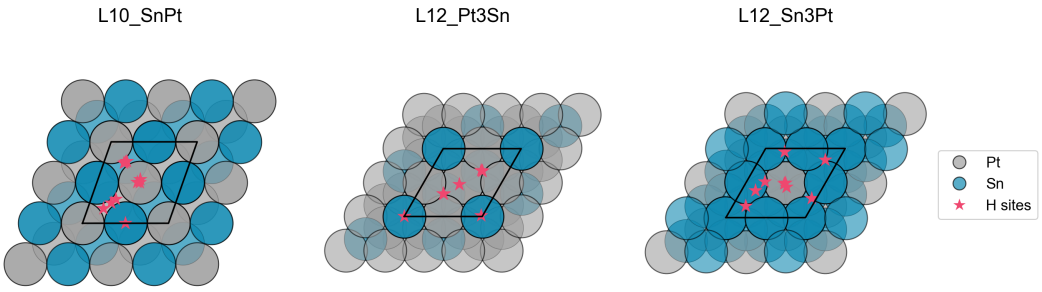
Contents

[*]Corresponding author
[†]Equal contribution

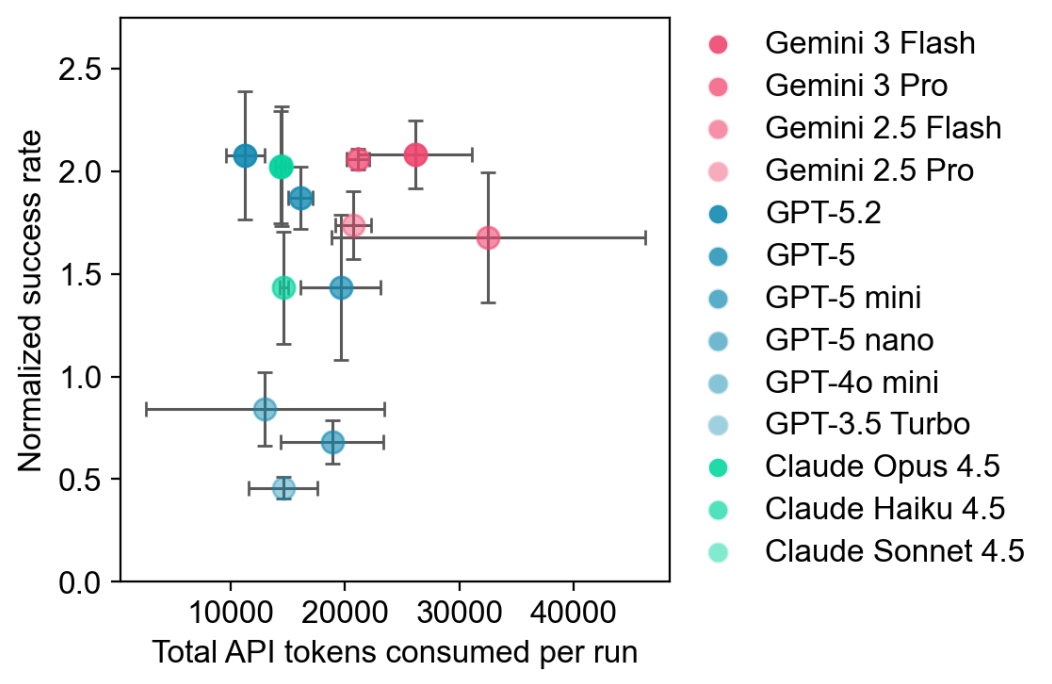
Supplementary Table 1: Language models evaluated in this study. All 13 models were used for single-shot screening; models marked with † were additionally benchmarked in the iterative CatAgent workflow.

Provider	Model	Mode
Google	Gemini 3 Flash†	Single-shot + CatAgent
	Gemini 3 Pro	Single-shot
	Gemini 2.5 Flash	Single-shot
	Gemini 2.5 Pro	Single-shot
OpenAI	GPT-5.2†	Single-shot + CatAgent
	GPT-5	Single-shot
	GPT-5 mini	Single-shot
	GPT-5 nano	Single-shot
	GPT-4o mini†	Single-shot + CatAgent
	GPT-3.5 Turbo†	Single-shot + CatAgent
Anthropic	Claude Opus 4.5	Single-shot
	Claude Sonnet 4.5	Single-shot
	Claude Haiku 4.5	Single-shot

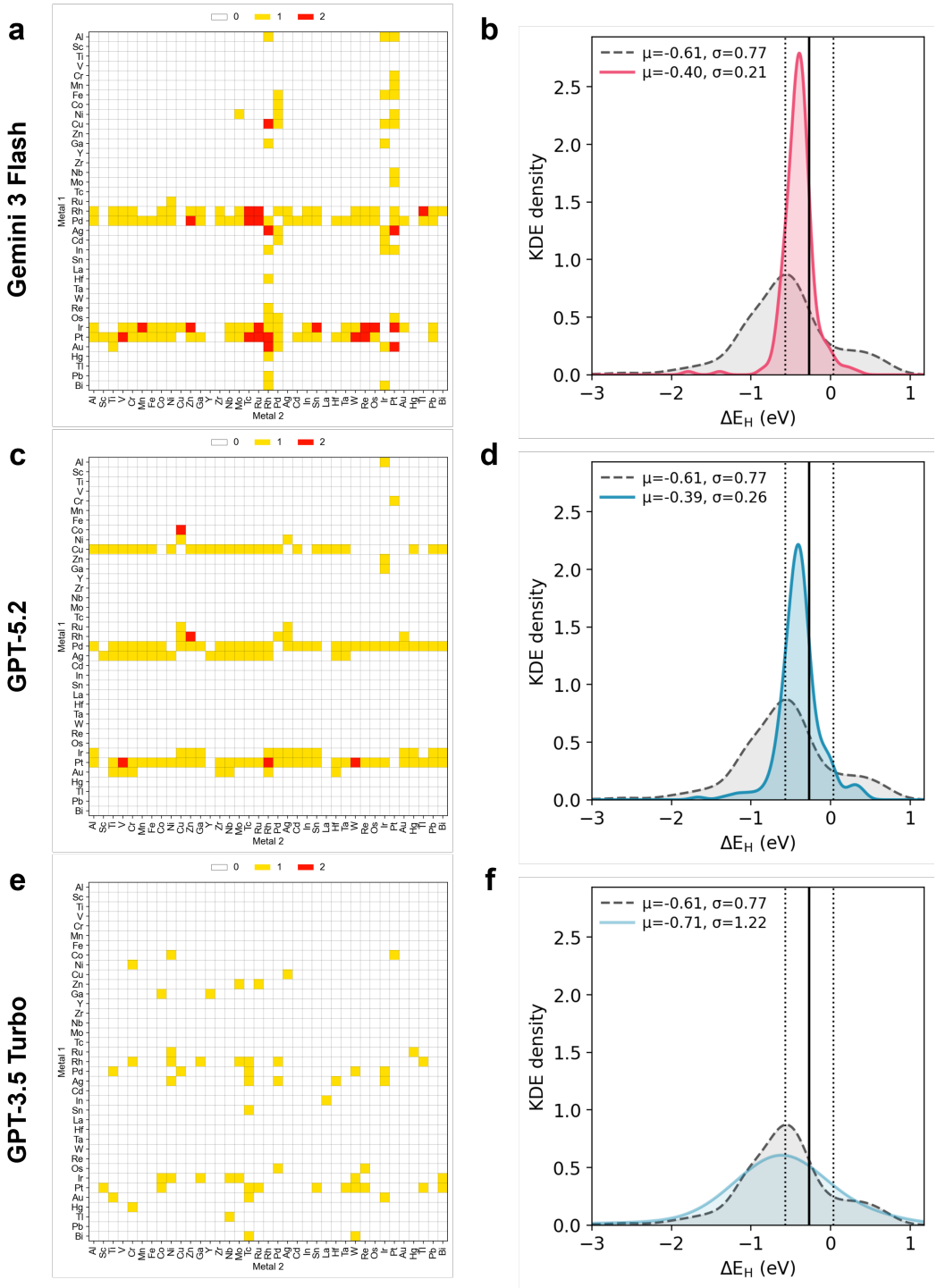
Token usage is reported as the sum of input and output tokens; cached tokens are excluded to ensure consistent comparison across providers. For $L1_0$ compositions, symmetric pairs are considered equivalent and deduplicated within and across steps.



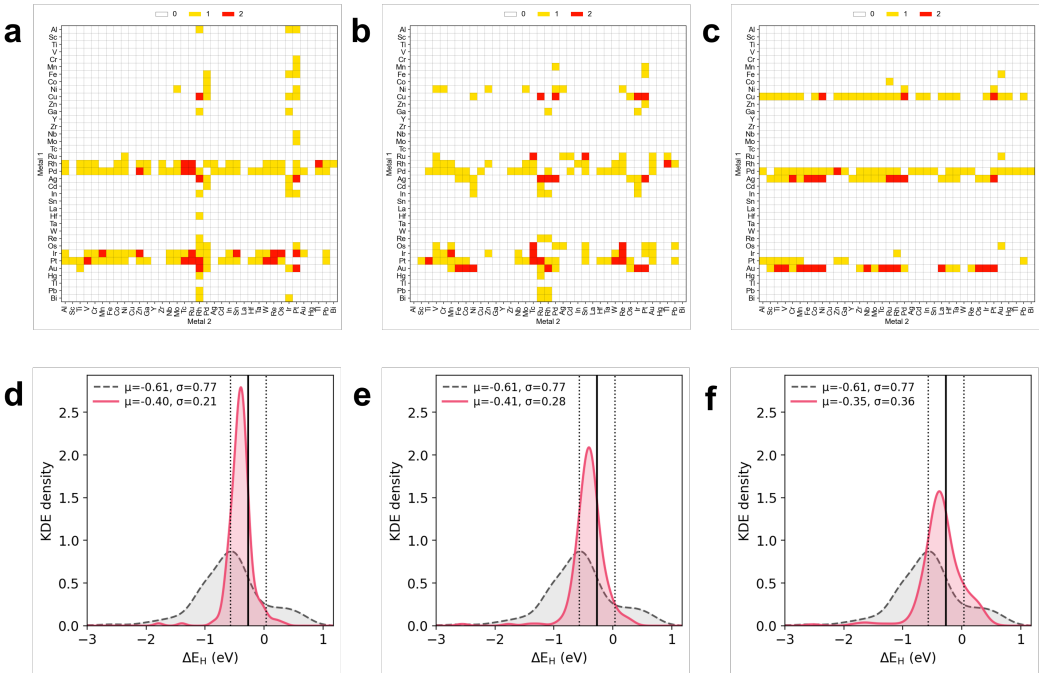
Supplementary Figure 1: Symmetry-unique adsorption sites enumerated for the $L1_2$ (9 sites) and $L1_0$ (10 sites) surface structures.



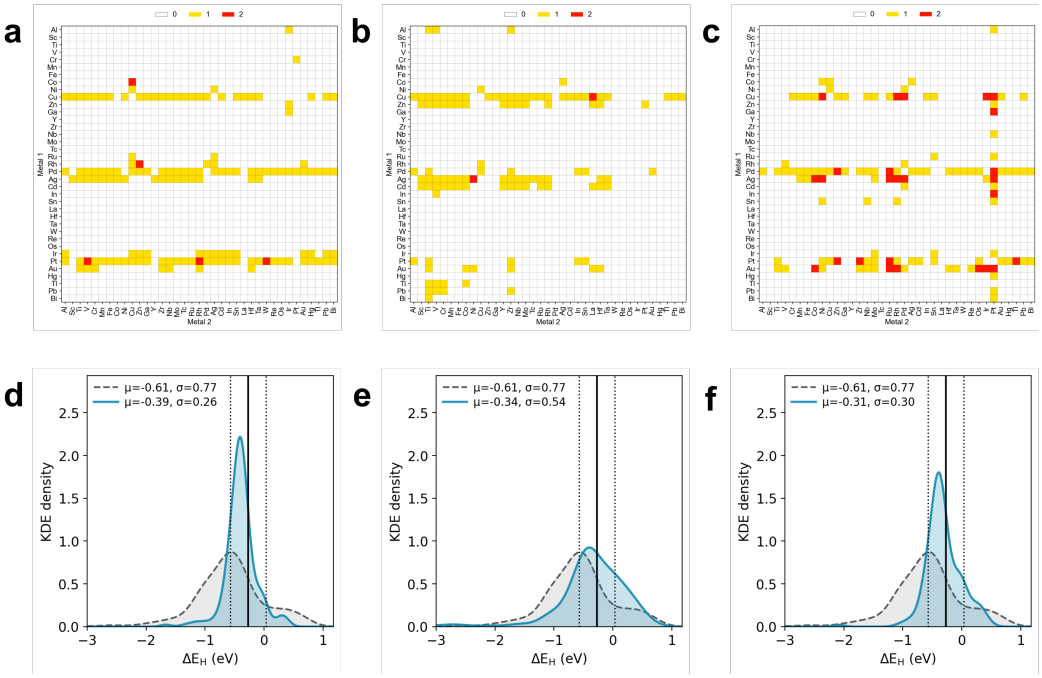
Supplementary Figure 2: Relationship between discovery efficiency (normalized success rate) and computational resource consumption (total API tokens per run) across the 13 language models in single-shot mode. Error bars: mean $\pm$ s.d. ($n = 3$).



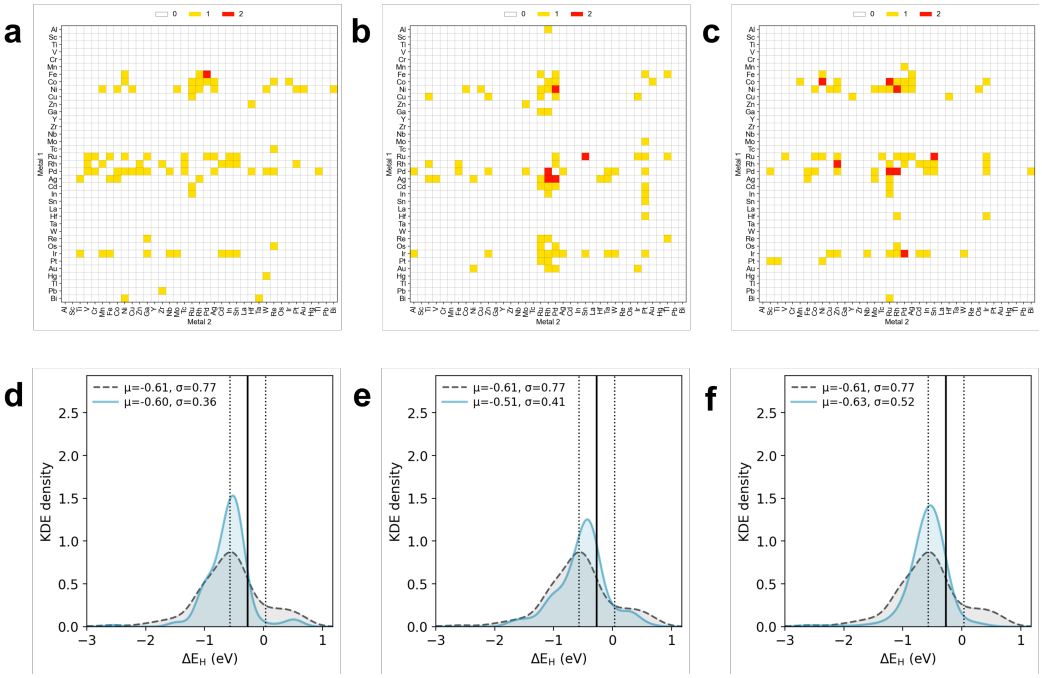
Supplementary Figure 3: (a, c, e) Heatmaps showing the composition distributions of candidates with adsorption energies within the energy threshold and (b, d, f) KDE plots showing the adsorption-energy distributions of all catalysts identified in one repeat of the critic-enabled CatAgent. (a,b) Gemini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



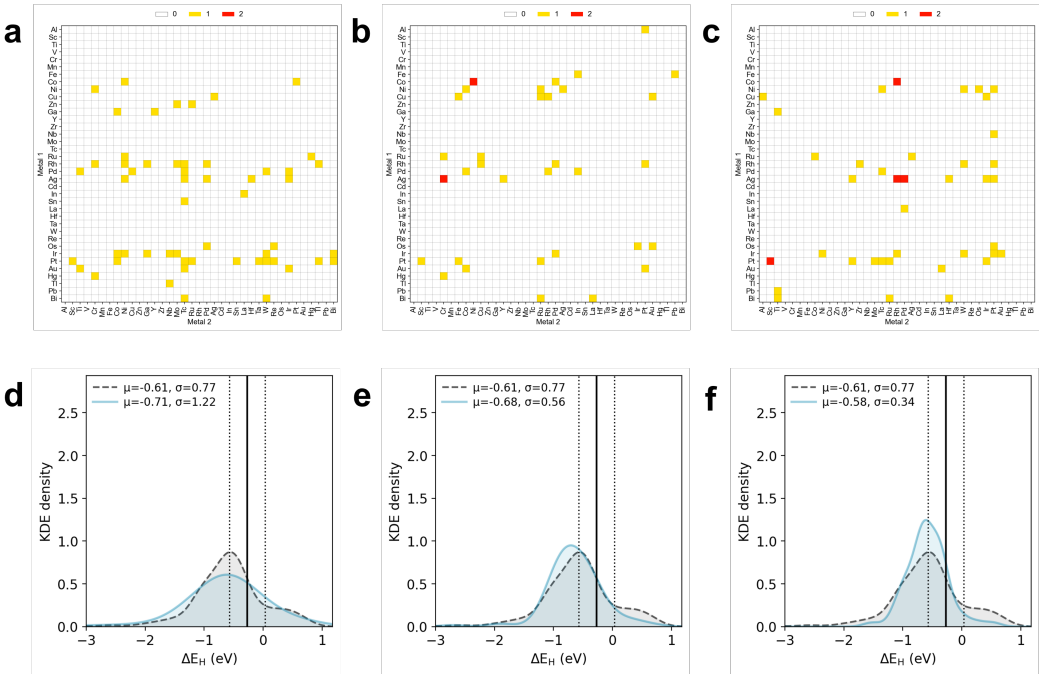
Supplementary Figure 4: Catalyst discovery patterns for Gemini 3 Flash using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e)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



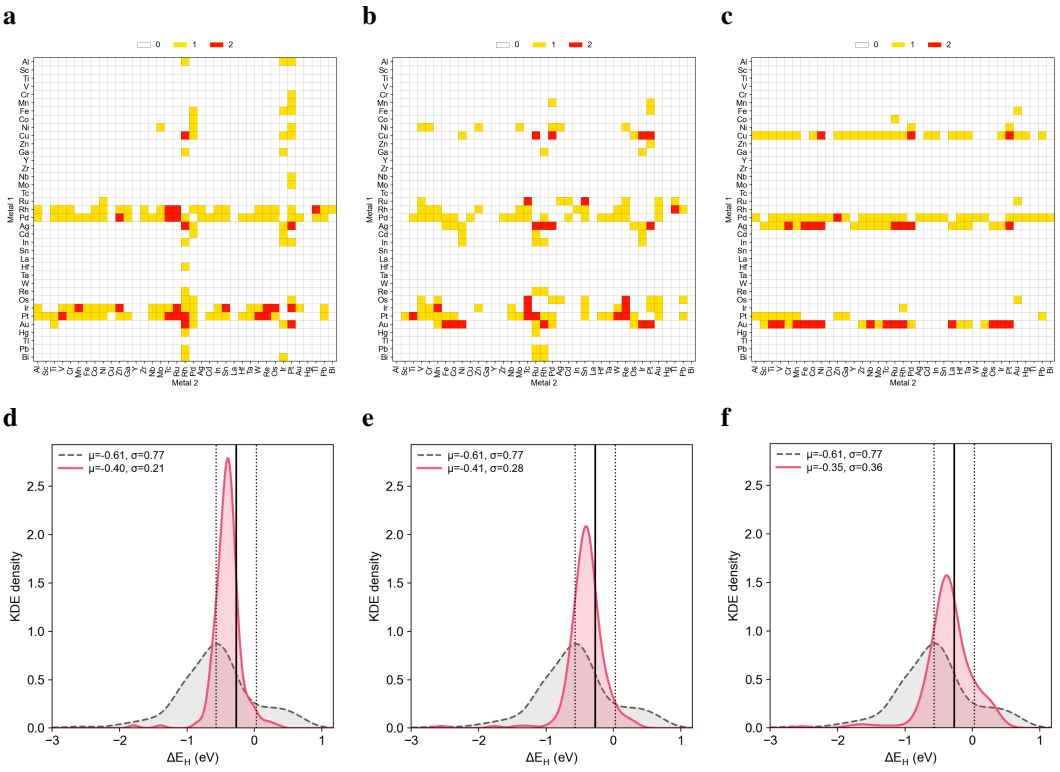
Supplementary Figure 5: Catalyst discovery patterns for GPT-5.2 using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



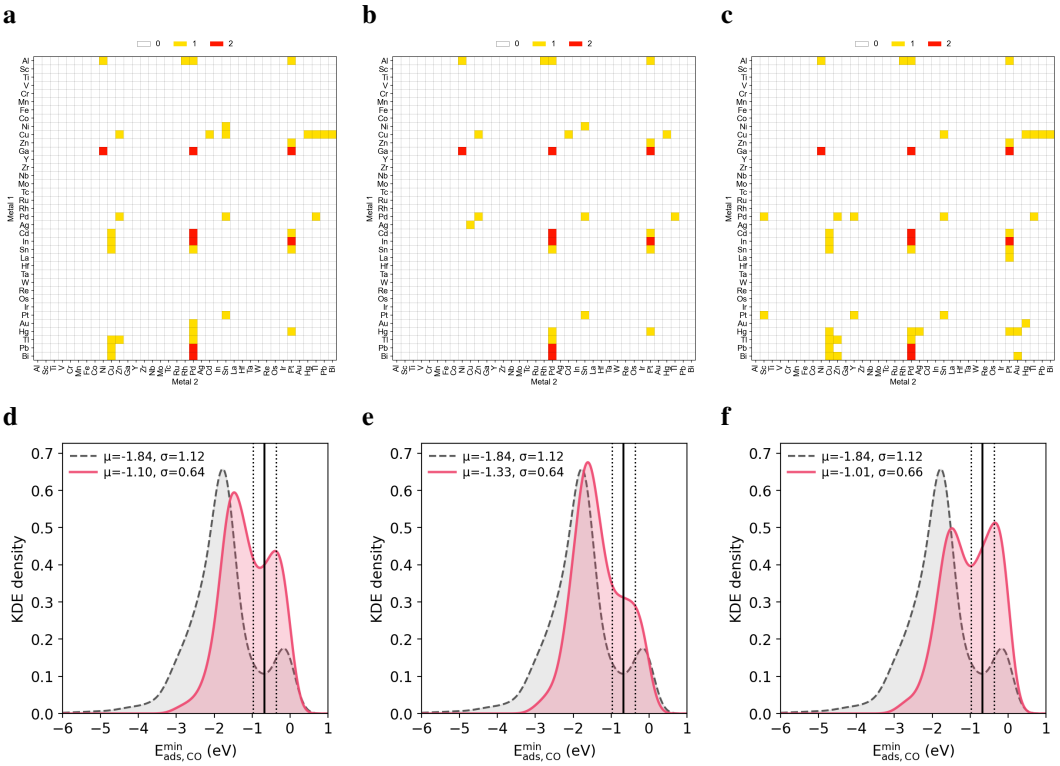
Supplementary Figure 6: Catalyst discovery patterns for GPT-4o mini using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



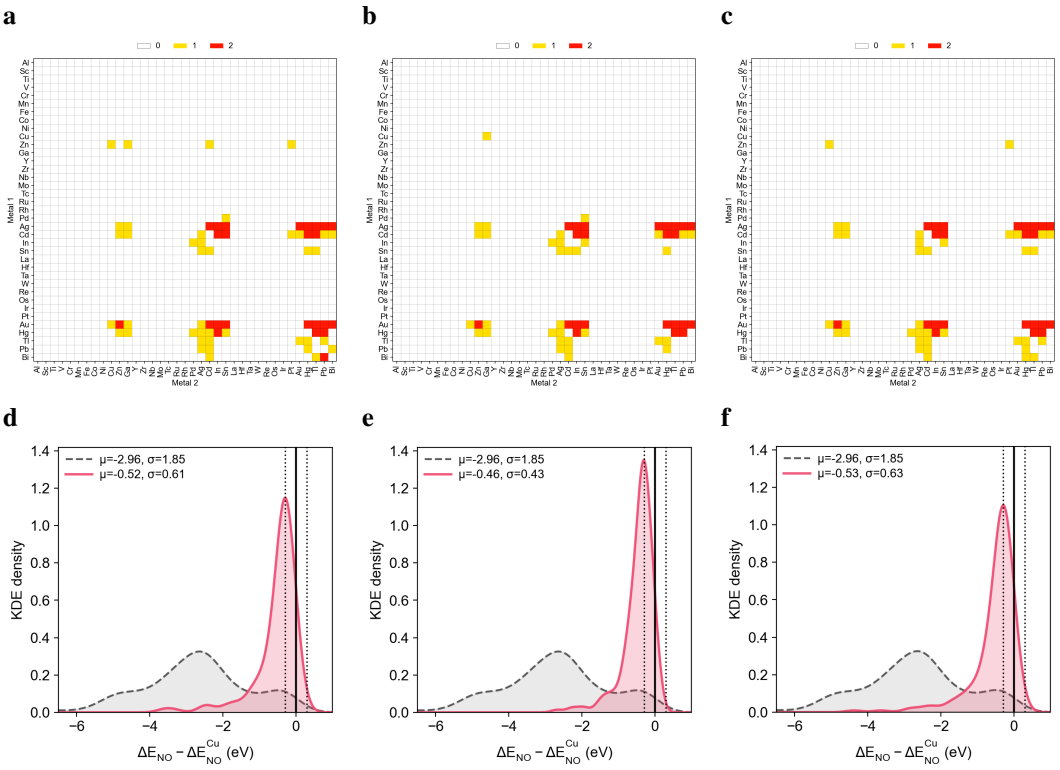
Supplementary Figure 7: Catalyst discovery patterns for GPT-3.5 Turbo using critic-enabled CatAgent across three independent trials. (a-c) composition heatmaps of successful candidates; (d-f) kernel density estimates of adsorption energy distributions for all explored candidates. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



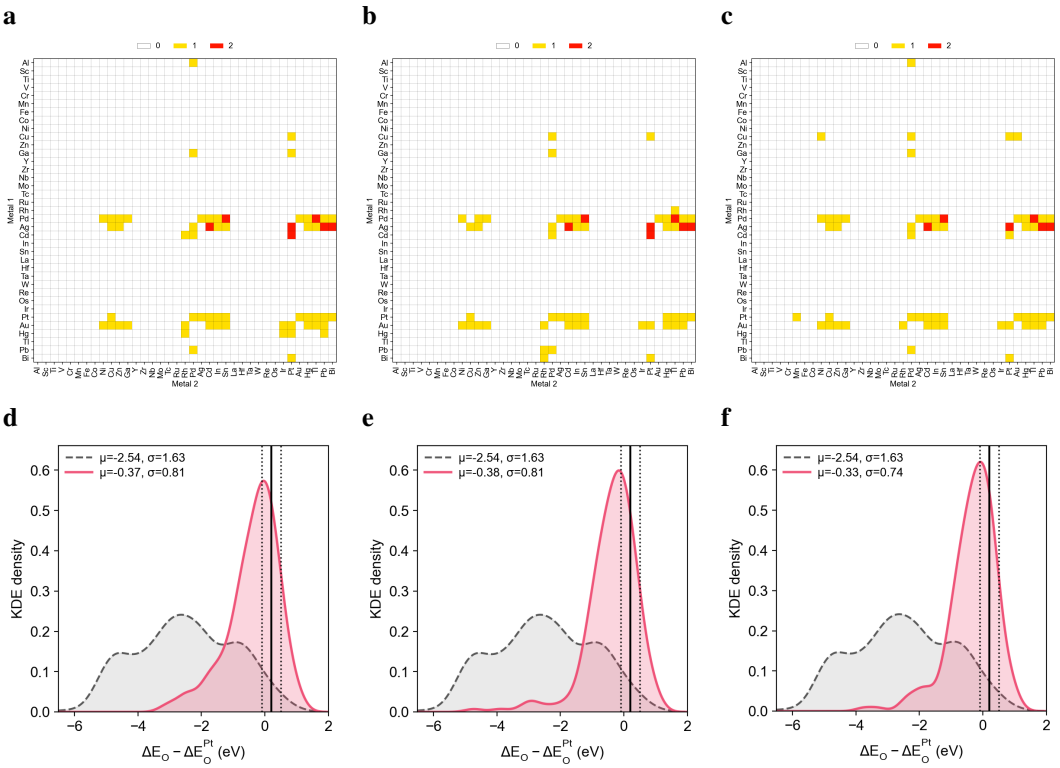
Supplementary Figure 8: CatAgent screening results for *H adsorption across three independent trials. (a–c) Composition heatmaps of successful candidates; (d–f) kernel density estimates of adsorption energy distributions. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



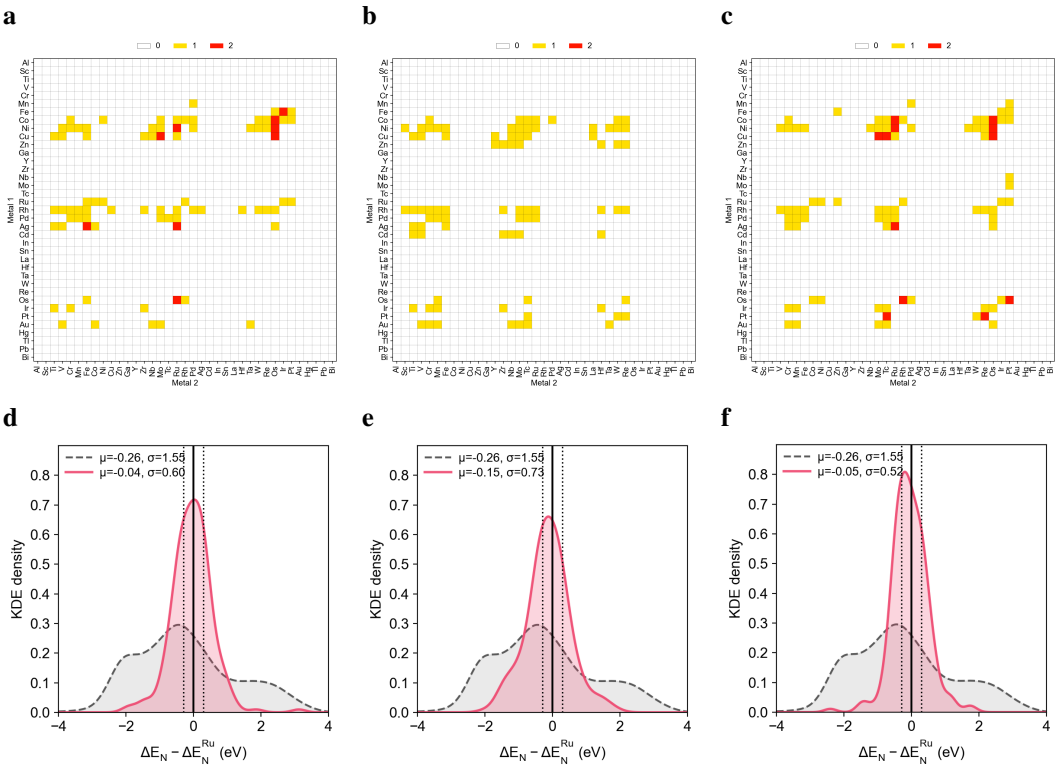
Supplementary Figure 9: CatAgent screening results for *CO adsorption across three independent trials. (a–c) Composition heatmaps of successful candidates; (d–f) kernel density estimates of adsorption energy distributions. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



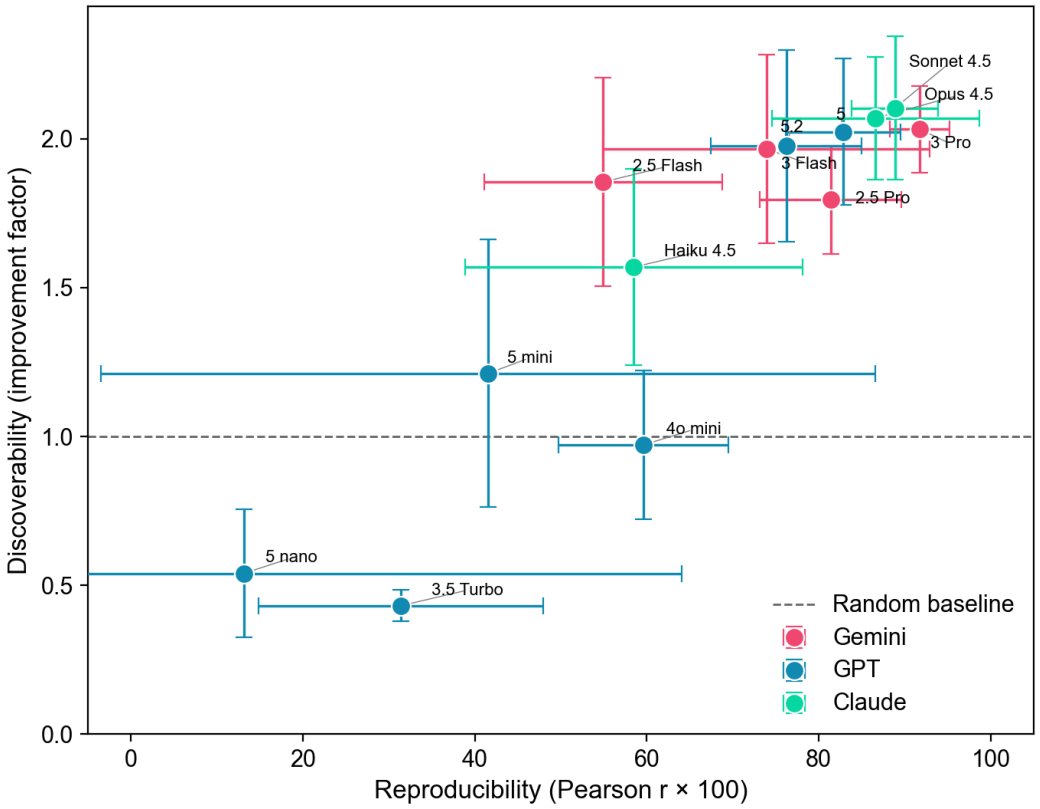
Supplementary Figure 10: CatAgent screening results for *NO adsorption across three independent trials. (a–c) Composition heatmaps of successful candidates; (d–f) kernel density estimates of adsorption energy distributions. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



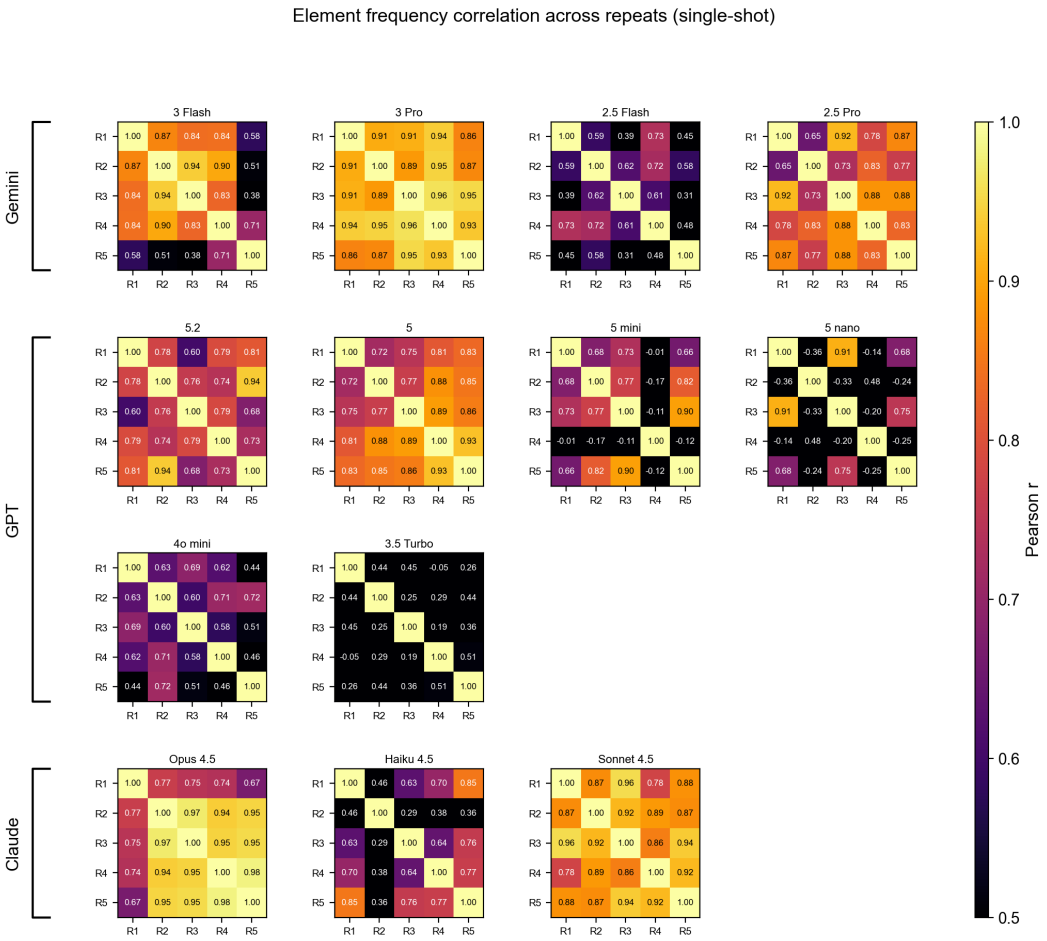
Supplementary Figure 11: CatAgent screening results for *O adsorption across three independent trials. (a–c) Composition heatmaps of successful candidates; (d–f) kernel density estimates of adsorption energy distributions. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



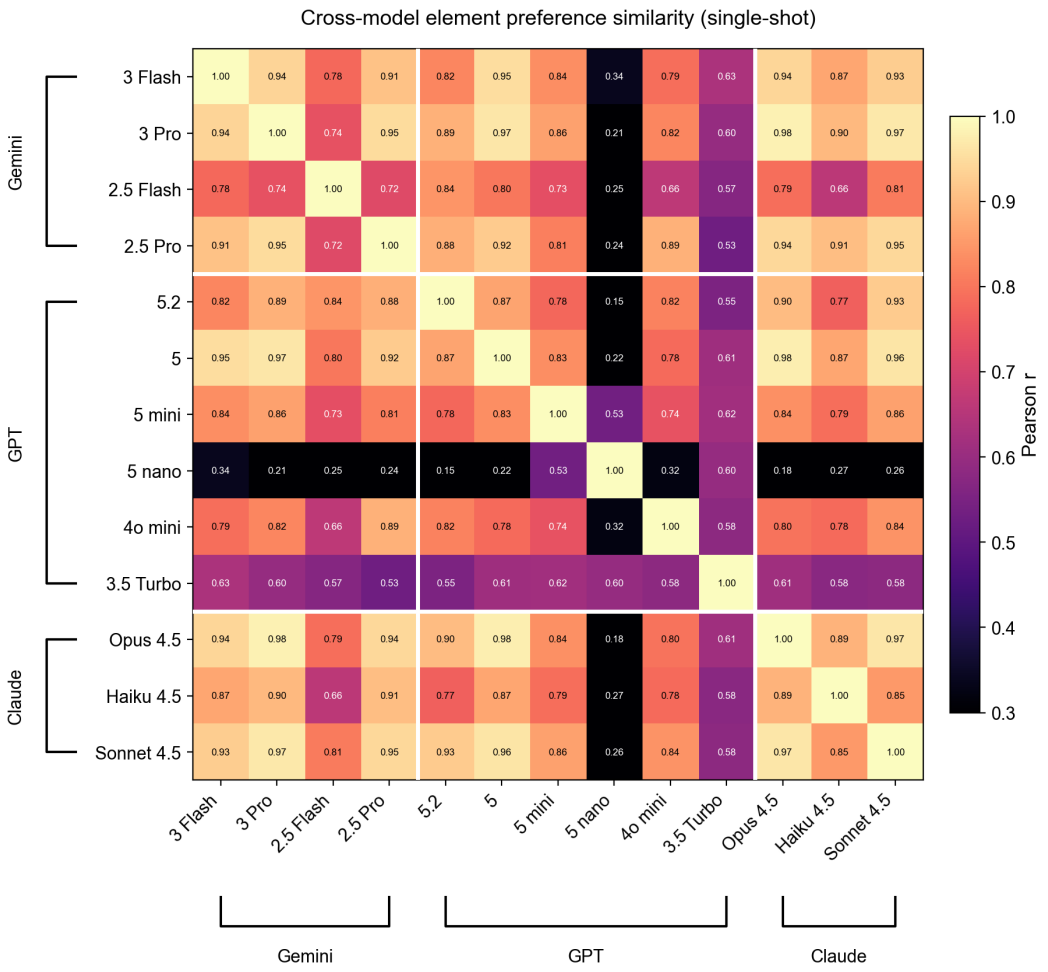
Supplementary Figure 12: CatAgent screening results for *N adsorption across three independent trials. (a–c) Composition heatmaps of successful candidates; (d–f) kernel density estimates of adsorption energy distributions. (a,d) Trial 1, (b,e) Trial 2, (c,f) Trial 3.



Supplementary Figure 13: Reproducibility versus discoverability for all 13 language models in single-shot mode ($n = 5$ independent repeats). The x-axis shows the mean pairwise Pearson correlation of element selection frequencies across repeats; the y-axis shows the improvement factor over random baseline. Error bars indicate ± 1 s.d. across repeat pairs (x) and repeats (y). The dashed line marks the random baseline (improvement factor = 1.0). Models in the upper-right quadrant are both reproducible and effective.



Supplementary Figure 14: Element frequency correlation across five independent repeats for each of the 13 language models in single-shot mode. Each 5×5 matrix shows pairwise Pearson r between element selection frequency vectors of two repeats. Models are grouped by provider (Gemini, GPT, Claude). High-performing models (e.g., Opus 4.5, Sonnet 4.5, Gemini 3 Pro) show consistently high correlations ($r > 0.9$), while low performers (e.g., GPT-5 nano, GPT-3.5 Turbo) show near-zero or negative correlations.



Supplementary Figure 15: Cross-model element preference similarity in single-shot mode. Each cell shows the Pearson r between two models' aggregated element selection frequency vectors (summed across five repeats and normalized). White lines separate provider groups. Most high-performing models show strong agreement ($r > 0.8$) regardless of provider, while GPT-5 nano diverges from the consensus ($r < 0.35$ against most models).