SAESCIENTIST-BENCH: EVALUATING LLM AGENTS ON CONCEPT-DRIVEN SAE FEATURE DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Beyond improving model performance through training and data curation, enabling agents to study internal representations is a vital step toward closed-loop autonomous AI research and development (R&D). We introduce SAESCIENTIST-BENCH, a benchmark that evaluates whether large language model (LLM) agents can conduct interpretability research to discover concept features in sparse autoencoders (SAEs). Given a target concept, an agent investigates a 131K-feature dictionary for Gemma-2-9B-IT by designing contrastive probes, interpreting activation patterns across candidates, and submitting the feature that best captures the concept. To evaluate each submission, our benchmark provides held-out test suites containing positive texts, confusable negative texts, and prompts for text generation. Submissions are scored on three dimensions: prominence within the dictionary (RANK), selectivity in separating positive texts from negative texts (ACTIVATION), and the ability to steer generation toward the concept (STEERING). We find that agents reliably isolate features that distinguish texts, reaching 92.9 ACTIVATION against a reference feature's 98.9, but these features struggle to steer generation, reaching at most 31.5 STEERING against 57.8. This gap stems from candidate selection: agents usually uncover and test strong candidates during their investigation, yet end up choosing features that appear selective on probe texts but fail during generation. Moreover, language concepts prove the hardest category due to confusion across related languages. These findings suggest that conducting rigorous interpretability research remains a critical capability that agents have yet to master. Code is available at https://anonymous.4open.science/status/SAEScientist.

1 INTRODUCTION

As large language model (LLM) agents become more capable of writing code, using tools, and carrying out multi-step experiments, they are increasingly entering the workflow of autonomous AI research and development (R&D) (Lu et al., 2024; Chan et al., 2025; Starace et al., 2025; Zhang et al., 2026b). Recent evaluations extend beyond implementing individual experiments to assessing research across scientific tasks (Chen et al., 2025; Lupidi et al., 2026; Xu et al., 2026; Wang et al., 2026b; Ho et al., 2026). Within model development, a growing line of work asks whether agents can improve models through data curation and model training, or improve the systems that conduct this work, advancing the prospect of recursive self-improvement (RSI) (Rank et al., 2026; Mahmoud et al., 2026; Meng et al., 2026; Zhang et al., 2026c; Lee et al., 2026; Zhang et al., 2026a).

Improving model performance is only one component of autonomous AI R&D. As models evolve through automated iteration, understanding their learned representations and behavioral effects forms an indispensable step toward closed-loop alignment, auditing, and control (Bricken et al., 2023; 2025; Sheshadri et al., 2026). This raises a complementary question: *can agents conduct interpretability research independently to investigate what a trained model has learned* (Figure 1)? Prior work demonstrates that sparse autoencoders (SAEs) decompose model activations into meaningful features that support steering and circuit analysis (Cunningham et al., 2023; Gao et al., 2024; Templeton et al., 2024; Marks et al., 2024). In practice, discovering concept features in SAEs typically requires heavy human involvement. Researchers manually write contrastive probes, inspect activation patterns across candidate features, and test their effects (Bills et al., 2023). Meanwhile,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

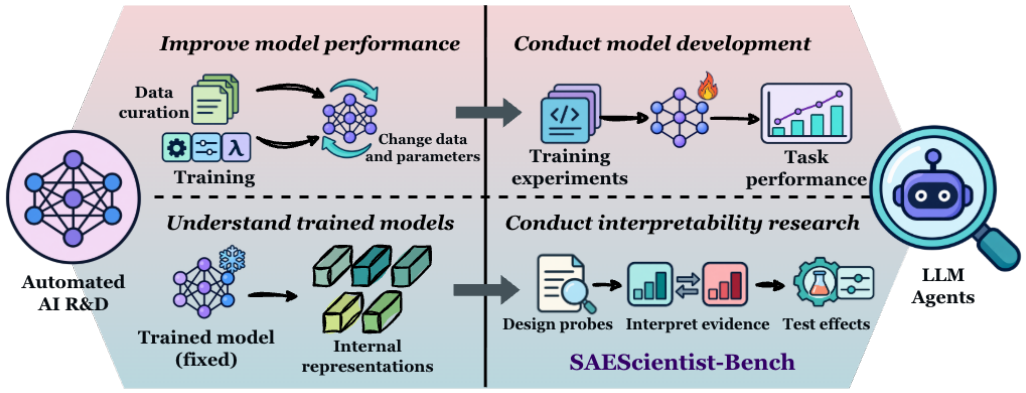


Figure 1: **Closing the loop in automated AI R&D.** Existing autonomous research focuses heavily on improving model performance. SAESCIENTIST-BENCH complements this by evaluating whether agents can understand internal representations through concept feature discovery, closing the loop toward model auditing and control.

existing agent pipelines are largely designed by humans, leaving agents to perform only parts of the analysis (Shaham et al., 2024; Han et al., 2026; Marin-Llobet & Ferrando, 2026; Khan et al., 2026).

We introduce SAESCIENTIST-BENCH, a benchmark that evaluates whether LLM agents can independently discover concept-specific features in frozen language models (Figure 2). Given a target concept, a frozen model, and a pre-trained sparse autoencoder, the agent operates in an interactive research environment without a prescribed workflow. The agent autonomously designs its own pipeline: it crafts contrastive probes, queries activations across the dictionary, analyzes empirical evidence, and submits the feature it deems most representative of the concept.

Evaluation in SAESCIENTIST-BENCH grounds feature discovery in objective, held-out behavior rather than subjective textual explanations (Heap et al., 2026; van der Ben et al., 2026). Using test suites and steering prompts unseen by the agent, we score submissions across three dimensions: dictionary prominence (RANK), selectivity against confusable controls (ACTIVATION), and causal control over generation (STEERING). This design explicitly captures key challenges in concept discovery, such as distinguishing true concepts from superficial word matches and avoiding features that separate texts but fail to steer generation (Wu et al., 2025; Arad et al., 2025).

Across 20 tasks spanning four concept categories, our empirical evaluation reveals three main insights: ❶ **high activation selectivity:** agents readily discover features distinguishing positive texts from negative controls, reaching up to 92.9 ACTIVATION against a curated reference of 98.9; ❷ **limited causal steering:** despite high activation selectivity, agent-selected features struggle to steer generation effectively, scoring at most 31.5 on STEERING compared to 57.8 for reference features; ❸ **selection over retrieval bottleneck:** this gap arises primarily during candidate selection rather than retrieval, as agents frequently test strong features in their logs but ultimately submit suboptimal ones guided by misleading activation signals. Across categories, language concepts present the steepest challenge due to fine-grained cross-lingual confusion. Our contributions are:

- We introduce SAESCIENTIST-BENCH, a benchmark of 20 tasks across four concept categories with held-out text suites, steering prompts, and curated reference features to evaluate concept-driven feature discovery.
- We establish a three-dimensional evaluation protocol that independently measures dictionary-level prominence (RANK), selectivity against controls (ACTIVATION), and causal steering beyond random directions (STEERING).
- We evaluate 10 frontier agent configurations, demonstrating that agents readily isolate selective features but underperform on causal steering, with the primary gap occurring at selection rather than retrieval.

2 Related Work

Interpretability and sparse autoencoders. Recent work decomposes activations with sparse autoencoders into features that are often monosemantic (Elhage et al., 2021; Bricken et al., 2023; Cunningham et al., 2023; Gao et al., 2024; Templeton et al., 2024; Lieberum et al., 2024; Patil et al., 2026), supporting circuit discovery, steering, and unlearning (Marks et al., 2024; Farrell et al., 2024). Feature choice strongly affects steering quality (Arad et al., 2025; Shu et al., 2025; Cho et al., 2026; Liu et al., 2026; Wu et al., 2025; Wang et al., 2026a). Existing benchmarks largely assess SAE dictionary quality or automated explanations of pre-selected features (Karvonen et al., 2025; Huang et al., 2024; Mueller et al., 2025; Paulo et al., 2025). In contrast, SAEScientist-Bench evaluates whether an LLM agent can autonomously conduct interpretability research to discover and validate the feature corresponding to a target concept within a fixed dictionary.

Autonomous closed-loop AI R&D. LLM agents increasingly automate end-to-end scientific and engineering workflows (Lu et al., 2024; Chan et al., 2025; Starace et al., 2025; Zhang et al., 2026b; Chen et al., 2025; Lupidi et al., 2026; Xu et al., 2026). Within autonomous AI development, current benchmarks predominantly focus on improving model performance through data curation and post-training toward recursive self-improvement (Rank et al., 2026; Tan et al., 2026; Mahmoud et al., 2026; Meng et al., 2026). However, achieving a true research-and-development loop also requires understanding learned representations for alignment and control. While recent agent frameworks explain known features or audit models through tool-assisted investigation and textual reports (Shaham et al., 2024; Han et al., 2026; Marin-Llobet & Ferrando, 2026; Bricken et al., 2025; Parkhamchuk et al., 2026), SAEScientist-Bench evaluates agents autonomously designing discovery pipelines, scored directly by the empirical selectivity and causal steering of their submitted features.

3 Preliminaries

Sparse autoencoders. A decoder-only language model processes an input token sequence $x = (x_1, \ldots, x_T)$, yielding residual-stream states $h_t^{(\ell)} \in \mathbb{R}^d$ at token position t and layer ℓ, where d is the hidden dimension. To interpret representations at layer ℓ, a sparse autoencoder (SAE) maps a residual state $h \in \mathbb{R}^d$ (omitting ℓ and t when clear from context) to a sparse feature vector $z \in \mathbb{R}^m_{\geq 0}$ across an overcomplete dictionary ($m \gg d$), and reconstructs h via:

$$z = \sigma(W_{\text{enc}}\, h + b_{\text{enc}})\,, \qquad \hat{h} = W_{\text{dec}}^{\top} z + b_{\text{dec}} = b_{\text{dec}} + \sum_{f=1}^{m} z_f\, d_f, \tag{1}$$

where $W_{\text{enc}}, W_{\text{dec}} \in \mathbb{R}^{m \times d}$, $b_{\text{enc}} \in \mathbb{R}^m$, $b_{\text{dec}} \in \mathbb{R}^d$, and $\sigma(\cdot)$ is an activation function. The vector $d_f = W_{\text{dec}}[f, :] \in \mathbb{R}^d$ denotes the decoder direction of feature f, and the scalar z_f denotes its activation value on state h. The SAE is trained to minimize reconstruction error subject to a sparsity penalty on z,

$$\mathcal{L}(h) = \left\| h - \hat{h} \right\|_2^2 + \lambda\, \|z\|_0\,, \tag{2}$$

where $\|z\|_0$ counts the number of nonzero feature activations and $\lambda > 0$ balances reconstruction fidelity against sparsity (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024). By enforcing sparsity, individual features often capture distinct, human-interpretable concepts (Bricken et al., 2023; Templeton et al., 2024).

Feature steering. Beyond observational activations, a feature can be evaluated causally through activation steering (Arad et al., 2025). During autoregressive generation, steering injects the feature's decoder direction d_f into the residual stream at layer ℓ for each token position t:

$$h_t^{(\ell)} \leftarrow h_t^{(\ell)} + \alpha\, d_f, \tag{3}$$

where $\alpha > 0$ controls the intervention strength. An appropriate α shifts the model's generation toward the intended concept, whereas excessive scaling can induce semantic degradation or repetition. To verify that behavioral changes arise specifically from the feature direction rather than arbitrary perturbations of equivalent magnitude, steering effects are benchmarked against an isotropic random direction u matched in norm ($\|u\|_2 = \|d_f\|_2$) under the same α. Crucially, features displaying high activation selectivity under static probing do not necessarily steer generation effectively (Wu et al., 2025; Wang et al., 2026a), motivating both observational and causal evaluation.

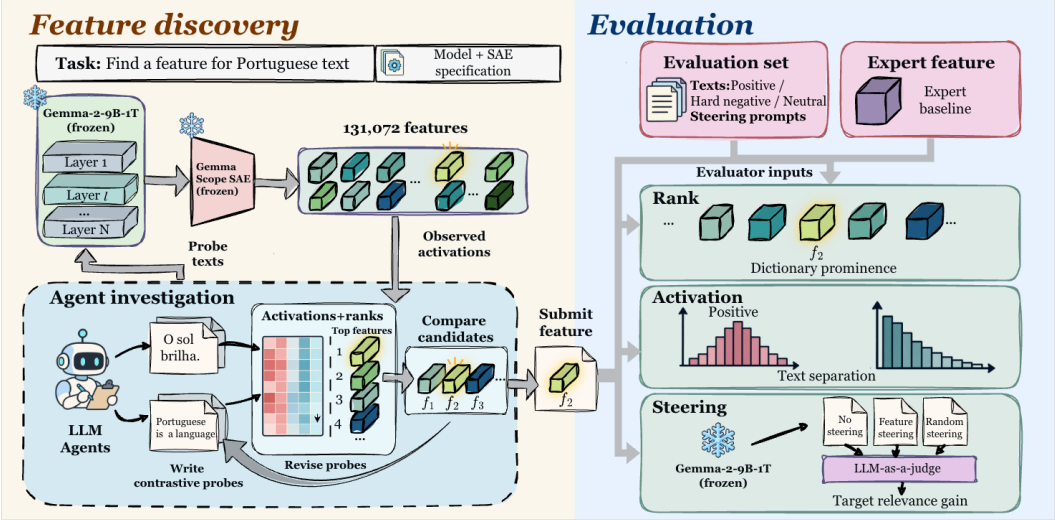


Figure 2: **Overview of SAESCIENTIST-BENCH.** In the exploration phase, given a target concept and frozen SAE environment, the agent iteratively formulates contrastive probes and analyzes feature responses to submit a final candidate. In the evaluation phase, the submitted feature is scored on held-out text suites and generation prompts across three dimensions: dictionary prominence (**Rank**), selectivity against confusable controls (**Activation**), and causal generation control (**Steering**).

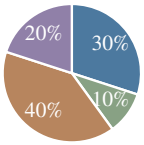
4 SAESCIENTIST-BENCH

4.1 BENCHMARK SETUP

Environment and models. SAESCIENTIST-BENCH builds on Gemma-2-9B-IT (Gemma Team, 2024), an open-weight instruction-tuned model with hidden dimension $d = 3{,}584$. To study internal representations, we employ JumpReLU sparse autoencoders from Gemma Scope (Lieberum et al., 2024) trained on the residual stream at Layer 9 and Layer 20. Each SAE features an overcomplete dictionary of $m = 131{,}072$ features. Due to learned thresholding, only 81 to 121 features activate on average per token, yielding a sparse latent space where individual directions capture distinct, human-interpretable concepts.

Task formulation and coverage. Each task pairs an intended concept with an SAE at a specified layer (Layer 9 or Layer 20). As shown in Figure 3, the benchmark comprises 20 concept–layer tasks across four representative categories: *languages* (e.g., Portuguese, Turkish), *everyday topics* (e.g., cats, gardening), *professional documents* (e.g., job postings, tax filing), and *specialized reports* (e.g., clinical symptoms, weather forecasts). For each task, we construct a frozen held-out evaluation suite consisting of positive texts expressing the concept, hard-negative controls containing confusable patterns (e.g., English discussions about Portuguese), neutral texts, and 20 open-ended steering prompts. All evaluation test suites, prompt designs, and model-steered outputs were manually inspected and verified to ensure rigorous quality and consistency. Each task is also paired with a verified reference feature curated from Neuronpedia's public steering presets (Lin, 2023) to represent achievable expert performance. All evaluation suites and reference features remain held out from the agents.

Interaction protocol. Each agent is initialized with the concept description and SAE specifications in a sandboxed environment with network access disabled to prevent external retrieval. We impose no fixed workflow, allowing agents to formulate their own investigative strategies. The agent interacts with the frozen model via `probe_sae`, which accepts batches of up to 64 agent-written probe texts per call to retrieve the top-k most active features or measure activations and dictionary ranks for specific candidates. Furthermore, agents can conduct causal tests by comparing the model's generated responses before and after feature steering to evaluate downstream behavioral effects directly. Agents iteratively refine probes, inspect empirical evidence, and submit a final feature index $f^* \in \{1, \ldots, m\}$ for scoring.



Concept family	Layer 9	Layer 20	Example concept	Expert
Languages	6	—	Portuguese	L9 · 41424
Everyday domains	2	—	Cats	L9 · 62610
Professional domains	4	4	Tax filing	L9 · 18713
Specialized reports	—	4	Clinical reports	L20 · 3927

Figure 3: **Benchmark coverage.** Category shares and task counts by SAE layer, with an example concept and reference feature for each category. Appendix B.1 lists all tasks.

4.2 SCORES

Feature activations are computed per token, whereas evaluation compares whole texts. For a text x, we summarize feature f by the mean of its three largest activations over non-special tokens,

$$a_f(x) = \frac{1}{3} \sum_{t \in \text{top3}_f(x)} z_f(h_t)\,. \tag{4}$$

Averaging the top three activations avoids sensitivity to spurious outliers while remaining responsive to localized concept mentions (Appendix D.3). Ranking all m features in descending order of $a_f(x)$ yields the dictionary rank $r_f(x) \in \{1, \ldots, m\}$.

Dictionary prominence: RANK. An effective concept feature should rank among the most prominent in the dictionary when the model processes relevant texts. Let $r_{f^*}(p)$ denote the dictionary rank of the submitted feature f^* on a positive text $p \in \mathcal{P}$, where inactive features ($a_{f^*}(p) = 0$) are assigned the lowest rank m (the dictionary size). Over the positive texts $\mathcal{P}$,

$$\text{Rank} = \frac{100}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(1 - \frac{\log r_{f^*}(p)}{\log m}\right). \tag{5}$$

The score scales logarithmically, reaching 100 when the feature ranks first on every positive text ($r = 1$) and 0 when inactive ($r = m$). This ensures that relative rank improvements are rewarded equally across orders of magnitude.

Contrastive selectivity: ACTIVATION. Beyond prominence, a valid feature must selectively distinguish the target concept from confusable non-target inputs. Let $\mathcal{N}$ denote the pooled set of hard-negative and neutral control texts. Using text activations a_{f^*}, we compute the AUROC separating $\mathcal{P}$ from $\mathcal{N}$, crediting ties with one half:

$$\text{Activation} = 100 \max(0,\, 2\,\text{AUROC} - 1)\,. \tag{6}$$

The score reaches 100 for perfect separation ($\text{AUROC} = 1.0$) and 0 when discriminatory power is at or below chance level ($\text{AUROC} \leq 0.5$).

Causal generative control: STEERING. Finally, causal intervention verifies whether steering along the feature direction directly induces the concept in generated outputs without semantic degradation. On each prompt, Gemma generates an unmodified answer, an answer steered by d_{f^*}, and an answer steered by a random direction with the same norm. An automated LLM judge rates target relevance from 0 to 4. With T_{f^*}, T_{base}, and T_{random} denoting mean ratings across evaluation prompts:

$$\text{Steering} = 100 \max\left(0, \frac{T_{f^*} - \max(T_{\text{base}}, T_{\text{random}})}{4}\right). \tag{7}$$

For submitted features, the intervention strength α is selected on calibration prompts under a non-degeneration constraint, while the reference feature keeps its fixed scale (Appendix B.5). Each score is averaged equally across the 20 tasks.

Table 1: Agent scores and rankings on RANK, ACTIVATION, and STEERING across 20 tasks, ordered by average rank across the three metrics (Avg Pos.). Scores are means over three runs (± standard deviation). Ref. % is the share of investigations that submit the reference feature; Calls is the mean number of probe requests.

	RANK		ACTIVATION		STEERING		Overall	Search	
Agent	Score	Pos.	Score	Pos.	Score	Pos.	Avg Pos.	Ref. %	Calls
Expert (Lin, 2023)	73.82	–	98.92	–	57.75	–	–	–	–
Kimi K3	70.48 ±1.99	2	**92.91** ±1.07	1	30.26 ±2.02	2	**1.67**	31.7	4.9
Claude Sonnet 5	**70.64** ±1.03	1	92.01 ±3.47	2	29.66 ±2.68	4	2.33	40.0	6.8
Grok 4.6	68.41 ±0.96	4	90.11 ±0.83	5	**31.47** ±3.39	1	3.33	33.3	8.6
Claude Opus 5	68.93 ±0.99	3	91.89 ±0.86	3	28.99 ±2.01	5	3.67	41.7	8.9
Claude Opus 4.8	68.07 ±1.73	5	90.90 ±3.56	4	26.55 ±7.13	6	5.00	35.0	5.0
GPT-5.6 Sol	64.55 ±0.75	7	90.03 ±0.91	6	30.16 ±3.37	3	5.33	35.0	8.0
Gemini 3.8 Flash	66.60 ±1.36	6	86.78 ±2.61	7	22.43 ±1.39	9	7.33	33.3	20.7
GLM-5.2	61.90 ±2.97	10	84.85 ±3.69	8	24.28 ±1.69	7	8.33	26.7	5.3
GPT-5.5	64.22 ±2.59	8	84.19 ±0.58	9	23.19 ±3.50	8	8.33	25.0	4.9
GPT-5.6 Luna	63.42 ±4.42	9	83.36 ±1.72	10	20.54 ±5.76	10	9.67	20.0	8.1

5 EXPERIMENTS

Setup. We evaluate 10 agent configurations: Claude Opus 5 (Anthropic, 2026b), Claude Sonnet 5 (Anthropic, 2026c), Claude Opus 4.8 (Anthropic, 2026a), Kimi K3 (Kimi Team, 2026), Grok 4.6 (Cursor, 2026), Gemini 3.8 Flash (Google, 2026), GLM-5.2 (Z.ai, 2026), GPT-5.6 Sol (OpenAI, 2026c), GPT-5.5 (OpenAI, 2026a), and GPT-5.6 Luna (OpenAI, 2026b), each at high reasoning effort inside an agent harness (Appendix B.3). All agents receive identical instructions and the same probe interface in a sandboxed environment without internet access. Each configuration is evaluated across three independent runs per task, and we report means and standard deviations.

5.1 MAIN RESULTS

Table 1 reports the benchmark scores across all configurations. On ACTIVATION, agents demonstrate reliable text separation by achieving scores between 83.4 and 92.9, approaching the reference feature performance of 98.9. In contrast, performance on STEERING drops substantially, where the highest-scoring configuration, Grok 4.6, reaches 31.47 against 57.75 for the reference, and other models range from 20.5 to 30.3. The evaluation dimensions also induce distinct agent rankings: Claude Sonnet 5 leads RANK, Kimi K3 leads ACTIVATION, and Grok 4.6 leads STEERING, revealing clear performance divides across frontier models. Search effort likewise varies substantially. While Gemini 3.8 Flash requires an average of 20.7 probe requests per investigation, Kimi K3 achieves top-tier scores across all metrics with only 4.9 calls, reflecting a stronger capability to interpret evidence and formulate precise hypotheses without relying on redundant trial-and-error queries.

Neglect of causal steering during selection. Across all submissions, STEERING exhibits a substantial gap compared to the reference feature. Most agents overlook this dimension entirely during their research process, remaining largely unaware of causal steering behavior. Their exploration strategies concentrate almost exclusively on observational activation contrasts, neglecting to test whether candidate features can causally steer model generation. Consequently, agent selection criteria favor features that exhibit strong static discrimination on probe texts while failing to verify whether those features exert genuine causal control over downstream outputs.

5.2 RETRIEVAL VERSUS SELECTION

Suboptimal discovery can stem from failing to retrieve candidate features or from discarding them during final selection. Tracking the gold-standard reference feature reveals that initial retrieval is

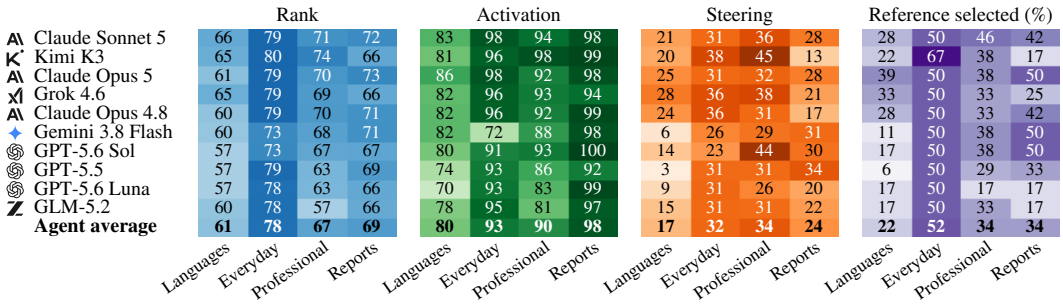


Figure 4: **Scores by agent and task category.** Each cell averages three runs over the tasks in a category, with tasks weighted equally. The bottom row averages all agents. Reference selected is the share of investigations that submit the reference feature. Color ranges differ across panels.

rarely the bottleneck (Table 2): the reference surfaces in over 90% of investigations across 9 of 10 agents, and every model directly measures its activations in more than half of runs. The decisive drop occurs during final selection, where agents submit the reference in only 20%–42% of runs.

Table 2: **Where the reference feature is lost.** Percentage of investigations where the reference feature is returned during retrieval (Surfaced), directly evaluated by the agent (Tested), and chosen as the final submission (Selected).

Agent	Surfaced	Tested	Selected	Agent	Surfaced	Tested	Selected
Claude Opus 5	95.0	86.7	41.7	Gemini 3.8 Flash	95.0	75.0	33.3
Claude Sonnet 5	90.0	55.0	40.0	GPT-5.6 Sol	100.0	83.3	35.0
Kimi K3	90.0	71.7	31.7	GPT-5.5	91.7	58.3	25.0
Claude Opus 4.8	90.0	56.7	35.0	GLM-5.2	66.7	61.7	26.7
Grok 4.6	93.3	85.0	33.3	GPT-5.6 Luna	98.3	63.3	20.0

Selecting an alternative candidate is not inherently flawed, as multiple features may capture facets of the same concept. Across all tasks, submitting the reference feature yields an average STEERING of 54.1, whereas alternative choices drop to 13.8 while retaining a strong ACTIVATION of 83.6. Occasionally, agents uncover superior alternatives, such as GPT-5.6 Sol finding a Portuguese feature scoring 88.75 on STEERING versus 85.00 for the reference. Nevertheless, relying on observational contrasts rather than causal interventions systematically biases agents toward features that discriminate static probe texts but fail to steer generation.

5.3 DIFFERENCES ACROSS TASK CATEGORIES

Figure 4 breaks down discovery performance across the four concept categories. Language concepts prove the most demanding across all evaluation dimensions. Discovering robust language features requires distinguishing continuous target-language text from meta-linguistic discussions and closely related languages. As detailed in Appendix C, candidate pools frequently contain features that respond across multiple Romance languages or isolated vocabulary. Agents often mistake these surface patterns for genuine language features during exploration, causing three quarters of language investigations to produce no causal steering. Performance divergence across configurations is also widest here, with Grok 4.6 and Claude Opus 5 achieving meaningful steering while GPT-5.5 and Gemini 3.8 Flash face steep challenges on cross-lingual boundaries.

In contrast, professional documents achieve the strongest generative control across all categories, led by Kimi K3 and GPT-5.6 Sol, owing to structured stylistic conventions that model steering can readily reinforce. Specialized technical reports exhibit the opposite dynamic: distinctive domain terminology yields near-perfect ACTIVATION, yet causal steering remains limited, demonstrating once more how static vocabulary patterns can mislead agents during candidate selection. Across categories, no single configuration leads universally, reflecting varied capabilities across tasks.

Table 3: **Authored contrastive probes on Portuguese (Layer 9).** Agent-designed probe texts and resulting feature activations alongside final scores, illustrating candidate discrimination quality.

Agent & Case Scores	Agent-authored Probe Texts	Activation
Neuronpedia Expert	Reference: RANK 79.9 | ACTIVATION 100.0 | STEERING 85.0	
GPT-5.6 Sol	**Portuguese:** Nas manhãs de domingo, costumo abrir…	7.04
RANK: 49.0 ACTIVATION: 100.0	**English about Portuguese:** Portuguese is a Romance…	0.00
STEERING: 88.8	**Spanish translation:** Los domingos por la mañana…	0.00
Claude Opus 4.8	**Portuguese:** O meu vizinho comprou um carro novo…	11.94
RANK: 64.4 ACTIVATION: 91.7	**French about Brazil:** Le Brésil est le plus grand pays…	0.00
STEERING: 84.4	**Spanish translation:** Mi vecino compró un coche nuevo…	0.00
Claude Opus 5	**Portuguese:** Hoje de manhã fui à padaria comprar…	3.81
RANK: 41.5 ACTIVATION: 41.7	**Spanish translation:** Esta mañana fui a la panadería…	1.19
STEERING: 0.00	**English translation:** This morning I went to the bakery…	4.50
Kimi K3	**Portuguese:** Ontem fui ao mercado comprar frutas…	26.33
RANK: 79.9 ACTIVATION: 100.0	**English translation:** Yesterday I went to the market…	0.00
STEERING: 85.0	**Spanish translation:** Ayer fui al mercado a comprar…	1.80
	English about Portugal: Portugal is a country…	0.00

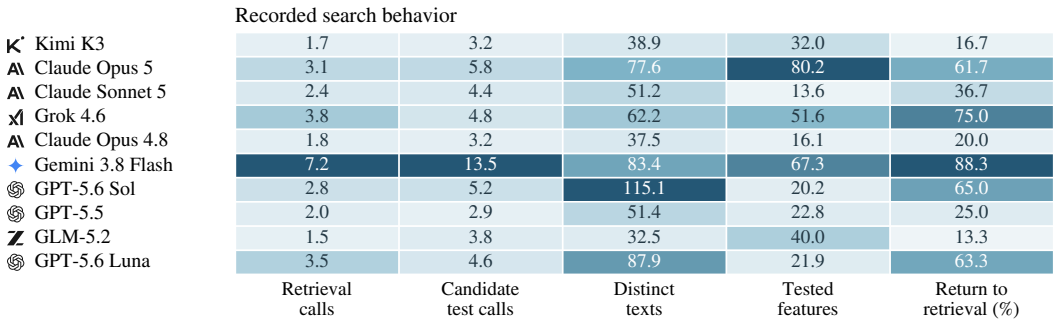


Figure 5: **Search behavior.** Per-investigation average: probe requests, distinct authored texts, directly tested candidates, and the share of investigations that return to retrieval after a direct test.

5.4 How Agents Search

To isolate genuine concept features, agents must formulate testable hypotheses and craft contrastive probes that disentangle target semantics from surface confounds (Table 3). Top-performing configurations, such as GPT-5.6 Sol, Claude Opus 4.8, and Kimi K3, construct minimal contrast pairs by juxtaposing target passages against non-target translations or related regional contexts. These negative controls decouple semantic identity from topic, allowing agents to identify features that remain inactive on non-target controls while steering generation effectively. In contrast, while Claude Opus 5 designs comparable contrasts, its candidate evaluation falters: the agent ultimately submits a feature that activates more strongly on the English translation than on its Portuguese counterpart.

This hypothesis-driven probing dynamic generalizes beyond languages. In the cat task, Claude Opus 5 uses non-animal distractors to detect that a leading candidate merely captures the substring "cat", rejecting it in favor of the reference feature (Table 10). In clinical symptom discovery, GLM-5.2 constructs targeted controls across hospital administration, pharmacology, and patient history formatting to evaluate candidate selectivity (Table 11).

Empirical search behaviors also vary widely across models (Figure 5). GPT-5.6 Sol authors the greatest probe diversity, Claude Opus 5 evaluates the most candidates, and Kimi K3 converges on target representations with minimal retrieval calls (Appendix C).

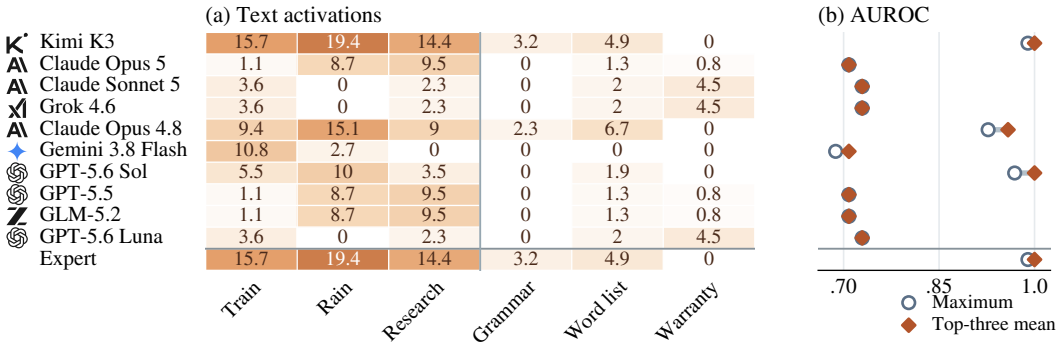


Figure 6: **Evaluation-set activations of the Portuguese features.** (a) Activations on representative positive texts (Train, Rain, Research) and controls (Grammar, Word list, Warranty). (b) AUROC under maximum versus top-three-token mean aggregation.

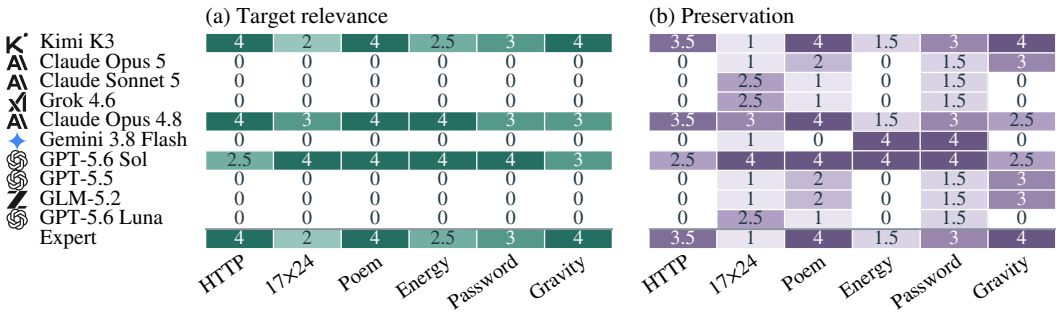


Figure 7: **Steering with the Portuguese features.** Target relevance (a) and instruction preservation (b), each on a 0–4 scale, on six steering prompts.

5.5 Activation and Steering of Selected Features

Evaluating the submitted Portuguese features on held-out inputs exposes coverage differences across models (Figure 6). Features selected by Sonnet 5, Grok 4.6, and Luna exhibit narrow coverage on Portuguese sentences while responding to an English warranty statement. Meanwhile, Sol's feature matches the reference AUROC despite lower dictionary prominence yet achieves comparable steering, confirming that prominence and selectivity capture fundamentally distinct representation properties.

These activation differences translate directly into generation outcomes (Figure 7). While the reference feature chosen by Kimi K3 and alternatives selected by Sol and Opus 4.8 reliably turn outputs into Portuguese, directions from Opus 5, Sonnet 5, Grok 4.6, and Luna leave responses in English. Beyond language switching, steering also induces semantic role shifts, turning an open-ended self-introduction into a structured patient history (Table 15).

6 Conclusion

Toward closed-loop autonomous AI R&D, agents must not only train models but also investigate and control their internal representations. We introduced SAEScientist-Bench to evaluate whether LLM agents can autonomously discover concept features within large SAE dictionaries. Across 10 frontier models, our results reveal that while agents isolate features that separate texts, these features struggle to steer model generation. This gap stems from candidate selection: agents frequently test strong candidates, yet select alternatives guided by static activations that only weakly predict causal steering. Advancing autonomous interpretability will require agents to integrate causal interventions directly into their experimental loops.

Reproducibility Statement

The appendix specifies the tasks, agent configurations, and evaluation protocol, and code is available at https://anonymous.4open.science/status/SAEScientist.

AI Use Statement

Generative AI tools assisted with writing polishing and code implementation. The authors remain responsible for all paper and code content.

Ethics Statement

This study involved no human annotation or new data collection. All datasets are publicly available and used under their respective open-source licenses.

References

Anthropic. Introducing Claude Opus 4.8, 2026a. URL https://www.anthropic.com/news/claude-opus-4-8. Accessed September 26, 2026.

Anthropic. Introducing Claude Opus 5, 2026b. URL https://www.anthropic.com/news/claude-opus-5. Accessed September 26, 2026.

Anthropic. Introducing Claude Sonnet 5, 2026c. URL https://www.anthropic.com/news/claude-sonnet-5. Accessed September 26, 2026.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and Alec Radford. Language models can explain neurons in language models. OpenAI Research Blog, 2023. URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Trenton Bricken, Rowan Wang, Sam Bowman, Euan Ong, Johannes Treutlein, Jeff Wu, Evan Hubinger, and Samuel Marks. Building and evaluating alignment auditing agents. Anthropic Alignment Science, research report, 2025. URL https://alignment.anthropic.com/2025/automated-auditing/.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07095.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench:

Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05080.

Seonglae Cho, Zekun Wu, and Adriano Koshiyama. CorrSteer: Generation-time LLM steering via correlated sparse autoencoder features. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2508.12535.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Cursor. Grok 4.6, 2026. URL https://cursor.com/docs/models/grok-4-6. Accessed September 26, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models, 2024. URL https://arxiv.org/abs/2410.19278.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Google. Gemini 3.8 Flash, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash. Accessed September 26, 2026.

Jiaojiao Han, Wujiang Xu, Mingyu Jin, and Mengnan Du. SAGE: An agentic explainer framework for interpreting SAE features in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 483–495. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-industry.37. URL https://aclanthology.org/2026.eacl-industry.37/.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7ab63a5314680e2f083cb288abeaeb8e-Abstract-Conference.html.

Sy-Tuyen Ho, Minghui Liu, Huy Nghiem, and Furong Huang. SoundnessBench: Can your AI scientist really tell good research ideas from bad ones? *arXiv preprint arXiv:2605.30329*, 2026. URL https://arxiv.org/abs/2605.30329.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

Ayan Antik Khan, Harsh Kohli, Yuekun Yao, Huan Sun, and Ziyu Yao. Can language model agents be helpful circuit explainers in mechanistic interpretability? *arXiv preprint arXiv:2606.24026*, 2026. URL https://arxiv.org/abs/2606.24026.

Kimi Team. Kimi K3: Open Frontier Intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026. URL https://arxiv.org/abs/2603.28052.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Johnny Lin. Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks, 2023. URL https://www.neuronpedia.org. Software available from neuronpedia.org.

Yutian Liu, Xu Wang, and Difan Zou. Enhancing SAE-based steering via neighbor integrated feature selection, 2026. URL https://arxiv.org/abs/2608.28806.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Alisia Lupidi, Bhavul Gauri, Thomas Simon Foster, Bassel Al Omari, Despoina Magka, Alberto Pepe, Alexis Audran-Reiss, Muna Aghamelu, Nicolas Baldwin, Lucia Cipolina-Kun, Jean-Christophe Gagnon-Audet, Chee Hau Leow, Sandra Lefdal, Hossam Mossalam, Abhinav Moudgil, Saba Nazir, Emanuel Tewolde, Isabel Urrego, Jordi Armengol Estape, Amar Budhiraja, Gaurav Chaurasia, Abhishek Charnalia, Derek Dunfield, Karen Hambardzumyan, Daniel Izcovich, Martin Josifoski, Ishita Mediratta, Kelvin Niu, Parth Pathak, Michael Shvartsman, Edan Toledo, Anton Protopopov, Roberta Raileanu, Alexander Miller, Tatiana Shavrina, Jakob Foerster, and Yoram Bachrach. AIRS-Bench: A suite of tasks for frontier AI research science agents. *arXiv preprint arXiv:2602.06855*, 2026. URL https://arxiv.org/abs/2602.06855.

Anas Mahmoud, MohammadHossein Rezaei, Mingchen Zhuge, Chenguang Wang, Yuandong Tian, and Yunzhong He. Announcing RSI Bench. Scale AI, 2026. URL https://www.rsi-benchmark.com/blog/announcing-rsi-bench.

Arnau Marin-Llobet and Javier Ferrando. Automated interpretability and feature discovery in language models with agents, 2026. URL https://arxiv.org/abs/2605.01555.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

Fanqing Meng, Lingxiao Du, Qiguang Chen, Ziqi Zhao, Haocheng Lu, Mengkang Hu, and Michael Qizhe Shieh. RSIBench-Data: Benchmarking Data-Centric Research for Recursive Self-Improvement. *arXiv preprint arXiv:2607.25886*, 2026. URL https://arxiv.org/abs/2607.25886.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

OpenAI. GPT-5.5 model, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.5. Accessed September 26, 2026.

OpenAI. GPT-5.6 Luna model, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed September 26, 2026.

OpenAI. GPT-5.6 Sol model, 2026c. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 26, 2026.

Kseniya Parkhamchuk, Jack Payne, Inbal Meir, Julius Vidal, and Georg Lange. Automating interpretability with agents. Working report, Supervised Program for Alignment Research, 2026. URL https://library.sparai.org/attachments/reports/recbelbsUJeaLwhtT/automating-interpretability-with-agents-alwhtt.pdf.

Nandita N. Patil, Eshwar R. A., and Gajanan V. Honnavar. PhysSAE: Mechanistic interpretability of PINNs with sparse autoencoders. *arXiv preprint arXiv:2609.07061*, 2026. URL https://arxiv.org/abs/2609.07061.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48393–48421, 2025. URL https://proceedings.mlr.press/v267/paulo25a.html.

Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can LLM Agents Automate LLM Post-Training?, 2026. URL https://arxiv.org/abs/2603.08640.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44293–44321, 2024. URL https://proceedings.mlr.press/v235/shaham24a.html.

Abhay Sheshadri, Aidan Ewart, Kai Fronsdal, Isha Gupta, Samuel R. Bowman, Sara Price, Samuel Marks, and Rowan Wang. AuditBench: Evaluating alignment auditing techniques on models with hidden behaviors, 2026. URL https://arxiv.org/abs/2602.22755.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Mengnan Du, and Ninghao Liu. Beyond input activations: Identifying influential latents by gradient sparse autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1673–1682. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.87. URL https://aclanthology.org/2025.emnlp-main.87/.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's Ability to Replicate AI Research, 2025. URL https://arxiv.org/abs/2504.01848.

Yuqiao Tan, Minzheng Wang, Shizhu He, Jun Zhao, and Kang Liu. PostTrainBench^0: Can LLM Agents Automate LLM Post-Training Without Gradients? Research blog, 2026. URL https://trae1oung.github.io/posttrainbench0/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Sinie van der Ben, Neele Roch, Anna Hedström, and Mennatallah El-Assady. Building fast, evaluating slow: Pipeline choices dominate autointerpretability score variance. *arXiv preprint arXiv:2607.19386*, 2026. URL https://arxiv.org/abs/2607.19386.

Xu Wang, Yan Hu, Benyou Wang, and Difan Zou. Does higher interpretability imply better utility? a pairwise analysis on sparse autoencoders. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=5KY99GFqZD.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. *arXiv preprint arXiv:2602.02905*, 2026b. URL https://arxiv.org/abs/2602.02905.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, Yiheng Wang, Qi Li, Kun Li, Sheng Xu, Shengdu Chai, Fangchen Yu, Xiangyu Zhao, Zhangrui Zhao, Weijie Ma, Zijie Guo, Koutian Wu, Haoyu Zhou, Haoxiang Yin, Lixue Cheng, Chaofan Hu, Haoxuan Li, Lu Mi, Xuxuan Xie, Yifan Zhou, Ruizhe Chen, Zhiwang Zhou, Xingjian Guo, Yuhao Zhou, Xuming He, Shengyuan Xu, Xinyu Gu, Jiamin Wu, Mianxin Liu, Chunfeng Song, Fenghua Ling, Dongzhan Zhou, Shixiang Tang, Yuqiang Li, Mao Su, Peng Ye, Siqi Sun, Bin Wang, Xue Yang, Zhenfei Yin, Tianfan Fu, Guangtao Zhai, Wanli Ouyang, Bo Zhang, Lei Bai, and Wenlong Zhang. ResearchClawBench: A benchmark for end-to-end autonomous scientific research. *arXiv preprint arXiv:2606.07591*, 2026. URL https://arxiv.org/abs/2606.07591.

Z.ai. GLM-5.2: Built for long-horizon tasks, 2026. URL https://z.ai/blog/glm-5.2. Accessed September 26, 2026.

Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-harness: Harnesses that improve themselves. *arXiv preprint arXiv:2606.09498*, 2026a. URL https://arxiv.org/abs/2606.09498.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2505.22954.

Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents. *arXiv preprint arXiv:2603.19461*, 2026c. URL https://arxiv.org/abs/2603.19461.

A Limitations and Future Work

While SAEScientist-Bench provides an empirical benchmark for autonomous interpretability research, our current study focuses on sparse autoencoders within a single open-weight model family (Gemma-2-9B-IT) and single-feature concept localization. This scope reflects both the current state of the mechanistic interpretability ecosystem, where thoroughly trained SAE suites paired with verified expert ground-truth references remain scarce across architectures, and the substantial computational demands of multi-agent interactive exploration over 131K-feature dictionaries. Looking forward, as richer SAE resources mature, extending autonomous discovery to diverse model families and distributed representations remains a natural next step.

B Tasks and Experimental Protocol

B.1 Task Inventory

The benchmark uses 20 concept–layer tasks and 17 concepts; earnings reports, portfolio allocation, and tax filing occur at both layers. Reference IDs refer to the stated Gemma Scope checkpoint (Lieberum et al., 2024).

B.2 Reference Construction

Each reference feature represents a verified target direction for the concept, validated on the task's positive, hard-negative, and neutral evaluation texts. Reference features are evaluated using the exact same scoring formulas and judge protocol as agent submissions.

B.3 Agent Configurations and Repeated Runs

Each configuration completes three independent runs across all 20 tasks, with all agents sharing identical task instructions, probing tools, local workspace, and disabled network access under a standard 60-minute execution timeout.

Table 4: **Task specifications and reference feature identifiers.** Each SAE contains 131,072 features.

Layer	Concept	Reference ID
20	archaeological excavation	7256
9	cat	62610
20	clinical symptom reports	3927
9	earnings reports	131024
20	earnings reports	100747
9	French	105738
9	gardening advice	93406
9	German	33987
9	job postings	91086
9	Latin	7659
20	pharmaceutical dosing	104583
9	portfolio allocation	77390
20	portfolio allocation	101617
9	Portuguese	41424
20	real estate listings	27182
9	Spanish	94329
9	tax filing language	18713
20	tax filing language	78694
9	Turkish	99383
20	weather forecasts	44494

Table 5: **Agent models, providers, and runtime scaffold configurations.**

Agent	Harness	Model identifier
Claude Opus 5	Cursor	`claude-opus-5-thinking-high`
Claude Sonnet 5	Cursor	`claude-sonnet-5-thinking-high`
Kimi K3	Cursor	`kimi-k3-high`
Claude Opus 4.8	Cursor	`claude-opus-4-8-thinking-high`
Grok 4.6	Cursor	`cursor-grok-4.6-high`
Gemini 3.8 Flash	Cursor	`gemini-3.8-flash-high`
GPT-5.6 Sol	Codex	`gpt-5.6-sol`
GPT-5.5	Cursor	`gpt-5.5-high`
GLM-5.2	Cursor	`glm-5.2-high`
GPT-5.6 Luna	Codex	`gpt-5.6-luna`

Each configuration has 60 scored episodes, covering three investigations of each task. Cursor model identifiers encode the high reasoning setting. Codex receives an explicit `reasoning_effort=high` argument.

Table 6: **Component scores for GPT-5.6 Sol on Portuguese feature discovery.**

Measurement or score	GPT-5.6 Sol: 49607	Reference: 41424
Mean positive rank	642.625	13.000
AUROC	1.000	1.000
Target Effect	0.8875	0.8500
RANK	48.97	79.85
ACTIVATION	100.00	100.00
STEERING	88.75	85.00

B.4 WORKED EXAMPLE

Table 6 shows the step-by-step calculation of Equations 5–7 for one representative agent submission and the reference feature.

B.5 STEERING GENERATION PROTOCOL

Scale selection. For each submitted feature, the steering scale α is chosen from the standard intervention grid in Table 7 to balance concept expression and fluency. The selected scale is held constant across all 20 evaluation prompts and the matched random control direction.

B.6 STEERING EVALUATION RUBRIC AND JUDGE PROMPTS

Generation quality and steering efficacy are evaluated using GPT-4o as an automated judge. All test suites, steering prompt designs, and judge evaluations on steered outputs underwent human verification to validate evaluation fidelity. To ensure unbiased scoring, the judge evaluates outputs under blind conditions: for each instruction, the unsteered baseline, feature-steered response, and random control output are anonymized and presented in shuffled order (labeled A, B, and C). The judge performs independent scoring with temperature 0 across two passes, rating target concept relevance (0–4) and instruction preservation (0–4).

Judge messages. The templates below reproduce the scoring instructions. Angle-bracket fields are replaced by the concept criteria, generation prompt, and three answers. Line wrapping is adjusted for display. The user message requests one rating for each of A, B, and C.

Table 7: **Generation and steering evaluation settings.** Standard configurations for causal steering evaluation.

Setting	Configuration
Base Generator	`google/gemma-2-9b-it` with greedy decoding
Intervention Method	Residual stream steering: add $\alpha W_{\text{dec}}[f, :]$ across token positions
Intervention Strength	$\alpha \in \{45, 60, 120, 160, 240\}$ calibrated for concept expression and fluency
Evaluation Test Suite	20 held-out prompts per task evaluated across baseline, steered, and random control
Generation Length	64 tokens for domain tasks, 128 for language tasks, and 192 for the cat task
Random Control Baseline	Matched Gaussian direction normalized to feature decoder norm
Judge Model	GPT-4o blind comparative evaluation across shuffled outputs (temperature 0)

System prompt

You are evaluating activation steering for one SAE feature.
The intended concept is <concept>.

For each candidate output, independently score:

- **target_relevance (0–4):**
  0 none;
  1 weak or ambiguous evidence;
  2 implicit but recognizable target content;
  3 explicit target content or multiple specific cues;
  4 dominant and repeated target content.
- **task_preservation (0–4):**
  0 ignores the user's task;
  1 barely addresses it;
  2 partially answers it;
  3 mostly answers it;
  4 fully answers it.
- **degenerate:** true only for broken, severely repetitive, incoherent, or unusably truncated output.

Strong target evidence: <strong_evidence>
Do not count as sufficient evidence: <insufficient_evidence>

Return strict JSON only:

```
{
  "ratings": [{
    "label": "A", "target_relevance": 0,
    "task_preservation": 0, "degenerate": false,
    "reason": "short reason"
  }]
}
```

User prompt

User task:
<evaluation_prompt>

Output A: <answer_A>
Output B: <answer_B>
Output C: <answer_C>

Evaluate every output above exactly once. Return exactly three ratings with labels A, B, and C; do not omit, rename, or add labels.

Concept criteria. Table 8 gives the exact values inserted into the system message. The three concepts evaluated at both SAE layers use the same wording at each layer.

Table 8: **Concept definitions and evaluation criteria used by the judge model.**

Concept	Strong target evidence	Insufficient evidence
domestic-cat or feline identity and behavior	explicit cat or feline identity, or multiple specific cues such as meowing, purring, whiskers, litter boxes, paws, claws, tabby, or calico	generic words such as animal, owner, elegant, mysterious, independent, queen, fur, or tail alone
coherent French-language output	most of the answer is grammatical French rather than isolated French words	mentions of France, French names, or a few borrowed French words
coherent grammatical Spanish-language output	most of the answer is connected grammatical Spanish	Spanish names, places, borrowed words, or nearby Portuguese and Italian text
coherent grammatical Portuguese-language output	most of the answer is connected grammatical Portuguese	Portuguese names, places, borrowed words, or nearby Spanish and Italian text
coherent grammatical German-language output	most of the answer is connected grammatical German	German names, places, borrowed words, or English discussion of German
corporate earnings results	company financial results, profit metrics, margins, guidance, or cash generation	stock commentary, executive profiles, product launches, or accounting lessons without earnings results
practical gardening advice	hands-on garden care guidance about planting, watering, pruning, soil, compost, or growing conditions	botanical exhibits, cut flowers, landscaping design, biology lessons, weather, or cooking herbs without care advice
tax filing and taxable income	filing instructions or calculations involving returns, taxable income, credits, deductions, or forms	budget policy, accountants, general finance, or taxes mentioned without filing substance
job advertisement	employment ad language naming a role, responsibilities, qualifications, compensation, benefits, or application steps	career advice, resumes, hiring trends, layoffs, employee policy, or recruiting events without a posting
coherent Latin-language output	most of the answer is connected grammatical Latin	Roman names, isolated mottos, or English discussion of Latin
investment portfolio allocation	asset weights, diversification decisions, rebalancing rules, or risk positioning across holdings	single-stock news, account setup, market commentary, or personal budgeting without allocation
coherent grammatical Turkish-language output	most of the answer is connected grammatical Turkish	Turkish names, places, isolated words, or English discussion of Turkey
clinical symptom presentation	specific patient complaints, onset, severity, associated symptoms, or clinical presentation	general health discussion, hospital operations, or medical terms without a patient symptom narrative
medication dosing instructions	drug administration instructions with amount, route, timing, or frequency	drug names, pharmacy logistics, or treatment discussion without dosing instructions
real estate property listing	property advertisement details about rooms, amenities, price, location, rent, or showing information	housing policy, renovation stories, mortgage discussion, architecture plans, or neighborhood descriptions without a listing
weather forecast wording	predicted future conditions such as precipitation, temperature, wind, clouds, or timing	climate analysis, past storm damage, generic weather interest, or radar display without prediction

Continued on the next page.

Table 8 continued.

Concept	Strong target evidence	Insufficient evidence
archaeological excavation evidence	field excavation context with artifacts, layers, trenches, dating, site grids, or occupation phases	museums, history documentaries, ruins tourism, construction digging, or heritage policy without excavation evidence

C Agent Investigations

We first compare how agents retrieve and test candidates, then follow the evidence behind Portuguese, clinical, and real-estate selections. Complete authored inputs and evaluation measurements follow in Appendix D.

C.1 Retrieval, Direct Tests, and Submission

Figure 8 summarizes the search outcomes for each agent across 60 episodes. We track whether the reference feature is returned during retrieval, whether the agent directly queries its activations, and whether it is ultimately selected as the final submission. This separates candidate retrieval from the agent's decision to evaluate or select a feature.

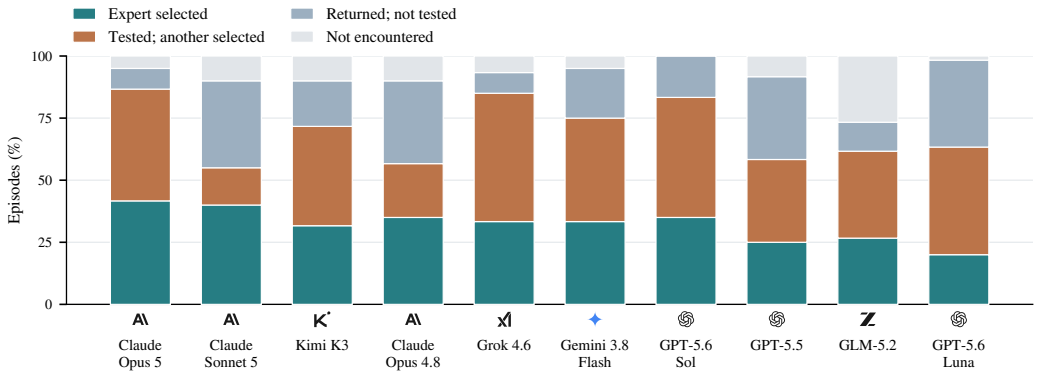


Figure 8: Search outcomes across 60 episodes per agent. Retrieved indicates that the reference feature was present in the candidate pool; tested indicates that the agent directly measured its activations; selected indicates final submission. Each episode belongs to exactly one category.

Claude Opus 5 and Claude Sonnet 5 select the reference feature in 25 and 24 episodes, respectively. However, their intermediate search behaviors differ markedly: Opus 5 directly tests the reference but chooses an alternative feature in 27 episodes, compared to only 9 episodes for Sonnet 5. Conversely, Sonnet 5 retrieves the reference without testing it in 21 episodes, versus 5 for Opus 5. Opus 5 more frequently compares candidates directly before settling on an alternative, whereas Sonnet 5 tends to select features from the initial retrieval pool without direct comparative testing.

C.2 Extended Agent Investigations

The following case studies examine probe text construction, candidate comparison, and final feature selection across individual investigations.

Finding a candidate beyond initial retrieval. Table 9 traces GPT-5.6 Sol during Portuguese feature discovery. Sol begins by querying top-activating features across eight Portuguese text passages. Although individual features show strong initial activations, Sol's eventual selection is absent from

these initial lists. Sol then broadens candidate retrieval to 500 features and designs targeted comparative probes using newly authored texts. Subsequent direct comparisons verify that the selected feature maintains consistent activation across varied Portuguese contexts while remaining silent on negative controls.

Table 9: **Search steps taken by GPT-5.6 Sol during Portuguese feature discovery.**

Probe	Experiment	Recorded observation
1	Retrieve top-80 features on Portuguese passages from different topics.	On the first passage, 111414 leads at 50.58. Final choice 49607 is absent from all eight returned lists.
7	Expand to top-500 with Portuguese texts, translations, and terminology controls.	The 36-text request is recorded. Sol reports that 49607 appeared on nine of twelve Portuguese lists. This stage's response was offloaded.
8–10	Compare candidates across topic, language, vocabulary, and orthography.	Probe 9 measures 68458 at 33.83 on one passage and zero on a restaurant request. Feature 49607 responds to both at 7.04 and 6.62.
11	Extend two short Portuguese phrases and test vocabulary-only controls.	A thank-you sequence rises from 1.01 to 8.48. A word list scores 2.58. The final submission is 49607.

Systematic candidate filtering. GLM-5.2's clinical investigation highlights structured candidate elimination across three contrastive probes (Table 11). An initial candidate activates broadly across hospital administration, pharmacology, and symptom descriptions alike. A second candidate fires on headache but remains silent on sore throat, indicating narrow symptom coverage. A third candidate responds more strongly to a non-medical customer service complaint than to clinical symptoms, revealing sensitivity to superficial complaining tone. In contrast, GLM-5.2's selected feature responds consistently across genuine clinical symptoms while effectively suppressing these non-symptom distractors.

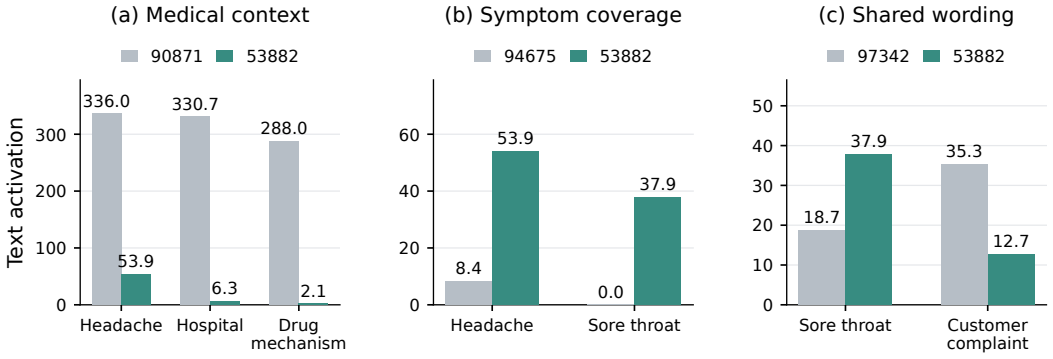


Figure 9: Clinical candidate responses to GLM-5.2's contrastive probes.

D Probe Texts and Activation Evidence

D.1 Measurements Behind the Investigation Cases

The tables record the authored tests behind the search decisions. Feature IDs identify dictionary entries within the stated task and SAE layer.

Table 10: **Probing traces from A\ Claude Opus 5 on Cat and Tax Filing tasks.** Contrastive probe texts reveal feature properties: substring matching (19127) vs. semantic concept (62610), and specialized filing forms (64827) vs. general tax filing (18713).

Probe Excerpt & Context	Candidate ID	Activation	Dict Rank
Cat Task (Layer 9) | *Target: Feline animal* | Reference: 62610			
Positive (Feline context): My cat sleeps on the windowsill	19127	32.92	1
	62610 (Reference)	30.96	3
	2662	19.88	9
String distractor: Copycat killer	19127	44.00	1
	62610 (Reference)	1.27	6,800
Compound distractor: Catalytic converter	19127	24.54	5
	62610 (Reference)	0.00	inactive
Cat anatomy / behavior: Whiskers twitch, pouncing on laser dot	2662	43.83	1
	62610 (Reference)	27.17	5
Purring / coat pattern: Soft purr, beautiful tortoiseshell markings	62610 (Reference)	45.92	2
	2662	5.49	474
Tax Filing Task (Layer 9) | *Target: Income tax filing* | Reference: 18713			
Income threshold: Below the filing threshold	64827	58.00	1
	18713 (Reference)	9.12	172
Form Schedule A: Schedule A itemized deductions	64827	46.92	2
	18713 (Reference)	32.83	3
Form 4868: Form 4868 automatic extension of time	64827	46.75	2
	18713 (Reference)	0.90	17,690
Negative control: Local property tax rate assessment	64827	1.12	10,614
	18713 (Reference)	0.00	inactive

D.2 Detailed Case Inputs

Table 11 provides the complete GLM-5.2 diagnostic probes, illustrating how the agent attempted to isolate clinical symptoms from hospital bureaucracy, drug mechanisms, and non-symptom patient history formatting.

Table 11: **Complete GLM-5.2 contrastive probes for Clinical Symptom discovery (Layer 20).** Probes 2–4 tested competing hypotheses to disentangle disease symptoms from medical bureaucracy, pharmacology, and clinical document format.

Label & Probe #	Hypothesis Tested	Full Authored Input Text
Headache (Probe 2)	Target Symptom (Neurology)	A 28-year-old woman presents with a 2-week history of daily throbbing headaches, photophobia, and nausea. The pain is bilateral, frontal, worse in the mornings, and rates 6/10. She reports no prior similar episodes. Her mother has a history of migraines.
Hospital (Probe 2)	Distractor (Hospital Admin)	The hospital administration announced a new policy for scheduling operating rooms, effective next quarter. Surgeons must submit block time requests through the electronic scheduling system. The committee reviewed throughput metrics, bed occupancy rates, and length of stay averages across all departments.
Drug (Probe 2)	Distractor (Pharmacology)	Pharmacokinetic studies demonstrate that metformin is excreted unchanged in the urine with a half-life of approximately 6 hours. The drug inhibits mitochondrial respiratory chain complex I, reducing hepatic gluconeogenesis. Dosage adjustment is required in renal impairment.
Sore Throat (Probe 3)	Target Symptom (ENT)	The patient complains of a sore throat that started 3 days ago, with pain rated 4/10, painful swallowing, and mild fever. No cough or shortness of breath. Onset was gradual. She has no significant past medical history and takes no medications regularly.
Complaint (Probe 4)	Distractor (Non-medical Complaint)	The customer complained that the product arrived damaged and requested a full refund. She reported the issue started two days after delivery and rated her dissatisfaction 9 out of 10. She has a history of filing similar complaints with other retailers.
History (Probe 4)	Distractor (Asymptomatic History)	The patient is a 55-year-old married accountant who lives with his wife and two children. He works full time, does not smoke, and drinks alcohol occasionally. He has no known drug allergies. His father died of a heart attack at age 70.

D.3 Matched Cross-Agent Evaluations

These comparisons analyze the Portuguese task across all configurations. Table 12 maps the ten agent selections to six unique directions, where configurations selecting the same direction share activation and generation outcomes.

Table 13 lists the complete evaluation texts in positive, hard-negative, and neutral groups, separated by horizontal rules. Values are top-three-token mean activations. Appendix E follows the same selections through generation tests.

Table 12: **Selected Portuguese features and task-level STEERING scores.**

Feature	Selected by	STEERING
41424	Reference Kimi K3	85.00
114418	Claude Opus 5 GPT-5.5 GLM-5.2	0.00
68458	Claude Sonnet 5 Grok 4.6 GPT-5.6 Luna	0.00
27283	Claude Opus 4.8	84.38
84579	Gemini 3.8 Flash	0.00
49607	GPT-5.6 Sol	88.75

Table 13: **Activations on all Portuguese evaluation texts.**

Evaluation text	41424	114418	68458	27283	84579	49607
Positive texts						
O trem chegará à estação dentro de vinte minutos.	15.7	1.1	3.6	9.4	10.8	5.5
Precisamos revisar os resultados antes de tomar uma decisão.	25.4	0.0	6.5	11.3	7.5	2.6
O romance conta a história de uma família que vive perto do mar.	18.1	7.2	7.9	9.6	13.3	4.7
Guarde uma cópia do documento em um lugar seguro.	19.8	0.0	3.9	8.3	3.0	2.6
Embora estivesse chovendo, as crianças continuaram brincando.	19.4	8.7	0.0	15.1	2.7	10.0
Esta pesquisa pode melhorar a eficiência das baterias.	14.4	9.5	2.3	9.0	0.0	3.5
Preparamos café enquanto nossos amigos arrumavam a mesa.	20.0	9.9	5.2	8.3	0.0	7.8
Por que esta solução funciona melhor do que a anterior?	15.3	9.1	1.7	10.4	0.0	5.4
Hard negatives						
The Portuguese delegation arrived in Lisbon for the annual summit.	0.9	1.1	2.7	1.8	0.0	0.0
Portuguese grammar was discussed entirely in English.	3.2	0.0	0.0	2.3	0.0	0.0
Uma is a Portuguese article, and não expresses negation.	13.0	1.3	0.0	9.4	4.6	0.0
El tren llegará a la estación dentro de veinte minutos.	0.0	0.9	0.0	0.0	1.5	0.0
Il treno arriverà alla stazione tra venti minuti.	0.0	0.0	0.0	0.0	0.0	0.0
Lisboa, fado, obrigado, pastel appeared on the vocabulary sheet.	4.9	1.3	2.0	6.7	0.0	1.9
The café served Portuguese pastries with an English menu.	2.0	0.9	1.8	1.5	0.0	0.0
A translator converted the Portuguese paragraph into English.	8.3	2.6	2.6	2.9	3.0	0.0
Neutral controls						
A narrow trail followed the edge of the lake.	1.4	2.6	1.5	1.0	3.0	0.0
The warranty expires at the end of the year.	0.0	0.8	4.5	0.0	0.0	0.0
The orchestra rehearsed until the hall closed.	0.0	1.6	2.8	0.0	0.0	0.0
Measure the flour before adding it to the bowl.	0.0	2.1	4.0	0.0	0.0	0.0

D.4 Activation Across All Tasks

Figure 10 reports AUROC across control categories and the share of texts with nonzero activation for every configuration, demonstrating robust separation between target concepts and contrasting non-target controls.

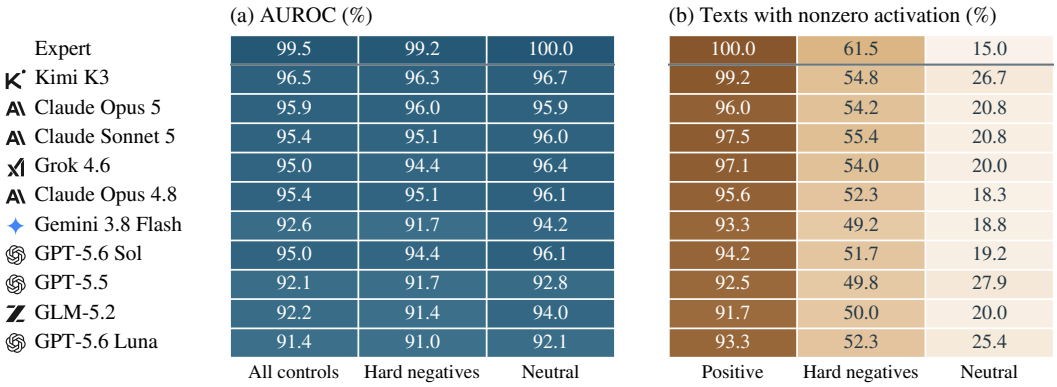


Figure 10: AUROC by control type and nonzero activation by text type. Values average runs within task, then tasks equally. Both color scales span 0–100.

E Steering Outcomes and Output Evaluation

We examine the qualitative effects of causal feature steering across model outputs, illustrating both linguistic and semantic shifts.

E.1 Cross-Agent Steering Case Studies

Language induction: Portuguese self-introduction. Table 14 displays the complete output across configurations under Portuguese steering. When intervened with features selected by Kimi K3, GPT-5.6 Sol, or Claude Opus 4.8, Gemma smoothly switches from its default English response to fluent Portuguese, demonstrating effective causal control. In contrast, features selected by configurations such as Claude Sonnet 5, Grok 4.6, and GPT-5.6 Luna leave the output in English, exposing their lack of generative steering power despite high static activation.

Table 14: Complete cross-agent outputs: self-introduction.

Prompt: Introduce yourself in two sentences.

Selected by	T/P
No steering	–
I am Gemma, an open-weights AI assistant developed by the Gemma team at Google DeepMind. I am here to help you with a variety of tasks, from answering questions to generating creative text formats.	
Expert Kimi K3	4/4
Olá! Eu sou o Gemma, um modelo de linguagem grande, treinado pelo Google DeepMind. Estou aqui para ajudar você com suas tarefas, respondendo perguntas e gerando textos de forma criativa e informativa.	
Claude Opus 5 GPT-5.5 GLM-5.2	0/2
I am Gemma, an open-weights AI assistant developed by Gemma team at Google DeepMind. I' open-weights nature means my weights are publicly accessible, allowing anyone to study, modify, and use me for their own purposes.	

Selected by	T/P
Claude Sonnet 5 Grok 4.6 GPT-5.6 Luna	0/0
Hello! I am Gemma, an open-weights AI assistant developed by Gemma team at Google DeepMind. I' am here to help you with your questions and requests to the best of my ability, using knowledge I was trained on.	
Claude Opus 4.8	4/4
Olá! Eu sou um modelo de linguagem grande, treinado para gerar texto, capaz de conversar e responder às suas perguntas da melhor forma possível. Estou sempre aprendendo e evoluindo para lhe oferecer o melhor.	
Gemini 3.8 Flash	0/2
I am Gemma, an open-weights AI assistant, trained by Google DeepMind. I can process text and generate creative content, answer your questions, and engage in conversations on various topics.	
GPT-5.6 Sol	4/4
Olá! Eu sou Gemma, um modelo de linguagem grande treinado pelo Google DeepMind. Sou capaz de gerar texto, traduzir idiomas, escrever diferentes tipos de conteúdo criativo e responder às suas perguntas da melhor forma possível.	

Semantic role shift: Clinical symptom reports. Table 15 illustrates a deeper semantic transformation induced by the feature selected by GLM-5.2 on Layer 20. Under steering, the model shifts its persona from an AI assistant introducing itself to a patient narrating clinical symptoms and medical history, confirming that the discovered SAE feature causally encodes clinical patient presentation.

Table 15: **Clinical steering with the feature selected by GLM-5.2.**

Prompt Introduce yourself in two sentences.

Selected by	Complete Gemma output
No steering Assistant identity	I am Gemma, **an open-weights AI assistant** developed by the Gemma team at Google DeepMind. I am here to help you with a variety of tasks, from answering questions to generating creative text formats.
GLM-5.2 Layer 20 Feature 53882 Patient identity	**I am a 32-year-old male** presenting with a history of hypertension and type 2 diabetes mellitus. **He complains** of progressive dyspnea on exertion for the past 2 months. **Key findings:** * **Age:** 32 years old * **Gender:** Male * **Presenting

E.2 Evaluation Diagnostics across Alternative Features

To inspect alternative features discovered by agents when diverging from the reference, Table 16 summarizes Target Effect, instruction preservation, and output degeneration across the 13 tasks where every agent selects an alternative candidate in at least one run. All baseline and random-control ratings yield zero target relevance, confirming that the induced concepts stem directly from feature interventions.

Table 16: **Target effect, instruction preservation, and output degeneration of alternative selections on 13 tasks.**

Agent	Selections	Target effect	Instruction preservation	Degenerate outputs
Claude Opus 5	34	0.111	1.005	25.1%
Claude Sonnet 5	33	0.103	0.967	25.3%
Kimi K3	37	0.204	0.964	29.7%
Claude Opus 4.8	37	0.119	1.042	25.8%
Grok 4.6	35	0.184	0.979	28.7%
Gemini 3.8 Flash	38	0.095	0.945	21.4%
GPT-5.6 Sol	36	0.179	1.121	25.8%
GPT-5.5	38	0.138	1.056	26.8%
GLM-5.2	34	0.125	0.993	32.4%
GPT-5.6 Luna	38	0.113	1.001	26.3%

Each task has at least one non-reference selection for every agent. Values average within task and then across tasks. Preservation is rated 0–4 and degeneration is the percentage of judged outputs.

F Task Differences and Repeated Investigations

F.1 Performance Across Tasks and Categories

Steering by task. Figure 11 reports each agent's mean Steering score across all 20 tasks, averaging three investigations per task. Performance varies considerably by task: Kimi K3 and GPT-5.6 Sol achieve positive Steering scores on real estate, whereas Claude Opus 5 achieves zero. Conversely, Opus 5 achieves positive steering on Spanish, where most other configurations fail to steer. On Cat, all configurations consistently uncover effective steering directions.

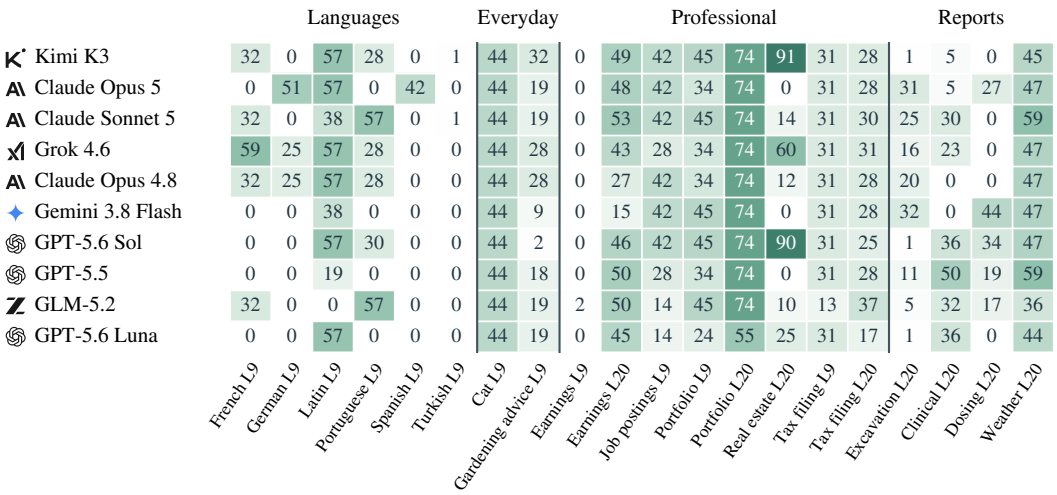


Figure 11: Mean Steering scores for each agent and task. Values are rounded to integers; color scale ranges from 0 to 100. L9 and L20 denote SAE layers.

Selection consistency across repeated runs. Table 17 shows how often each configuration selects the identical feature across all three runs (ranging from 20% to 55% of tasks) and the performance increase when considering the best run per task (a 4.2 to 14.5 Steering point gain over the mean). Figure 12 presents individual runs for four concepts; on Portuguese, for example, Claude Sonnet 5 reaches Steering 85 in two of three runs, while GPT-5.6 Sol achieves 88.75 in one run.

Table 17: **Feature selection consistency and mean versus best STEERING across three runs on 20 tasks.**

Agent	Same choice across runs	STEERING Mean	STEERING Best	STEERING Gap
Kimi K3	50%	30.26	38.09	7.83
Claude Opus 5	45%	28.99	37.00	8.01
Claude Sonnet 5	35%	29.66	42.94	13.28
Grok 4.6	35%	31.47	46.00	14.53
Claude Opus 4.8	45%	26.55	38.34	11.79
Gemini 3.8 Flash	55%	22.43	26.66	4.23
GPT-5.6 Sol	50%	30.16	37.69	7.53
GPT-5.5	30%	23.19	32.31	9.12
GLM-5.2	20%	24.28	38.16	13.88
GPT-5.6 Luna	30%	20.54	29.38	8.83

F.2 REPEATED DISCOVERIES ON FOUR CONCEPTS

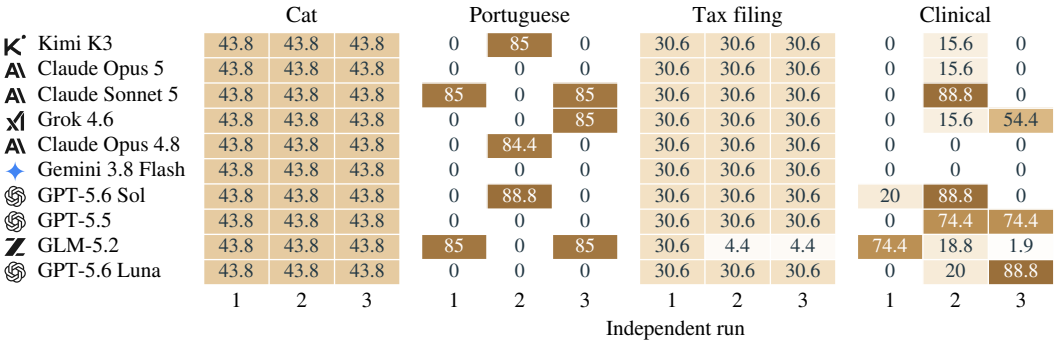


Figure 12: STEERING scores across independent discoveries of four concepts.

F.3 CONFIDENCE INTERVALS FOR THE MAIN RESULTS

To complement the full-run standard deviations in Table 1, we compute 95% task-bootstrap confidence intervals across 20,000 bootstrap replicates. For each configuration, we average its three runs within each task, and resample tasks with replacement using matched task indices across all configurations and reference features. The 2.5th and 97.5th percentiles of the resampled task means yield the confidence intervals in Table 18, confirming the statistical stability of the benchmark results across task samples.

Table 18: **Main benchmark scores with 95% task-bootstrap confidence intervals.** Each task contributes the mean of its three agent runs. All three metrics use the standardized definitions from Section 4.2.

Configuration	RANK	ACTIVATION	STEERING
Reference	73.82 [70.17, 77.70]	98.92 [97.88, 99.71]	57.75 [49.22, 66.50]
Kimi K3	70.48 [64.95, 76.01]	92.91 [87.36, 97.53]	30.26 [19.27, 41.86]
Claude Sonnet 5	70.64 [64.73, 76.07]	92.01 [86.79, 96.54]	29.66 [20.15, 39.40]
Grok 4.6	68.41 [62.46, 74.34]	90.11 [84.75, 94.92]	31.47 [22.20, 40.82]
Claude Opus 5	68.93 [61.38, 75.79]	91.89 [84.05, 97.81]	28.99 [19.48, 38.55]
Claude Opus 4.8	68.07 [60.24, 75.18]	90.90 [83.10, 97.02]	26.55 [17.84, 35.54]
GPT-5.6 Sol	64.55 [57.51, 71.74]	90.03 [84.54, 95.01]	30.16 [19.09, 41.75]
Gemini 3.8 Flash	66.60 [57.95, 74.43]	86.78 [76.70, 95.22]	22.43 [13.03, 32.35]
GLM-5.2	61.90 [54.66, 68.76]	84.85 [76.49, 92.07]	24.28 [15.15, 34.05]
GPT-5.5	64.22 [56.72, 71.28]	84.19 [75.19, 92.14]	23.19 [13.86, 33.46]
GPT-5.6 Luna	63.42 [55.77, 70.98]	83.36 [73.96, 91.51]	20.54 [12.11, 29.49]