Anytime-Valid Hypothesis Testing for Sparse Autoencoder Feature Validation

Rudransh Khera [1]

Abstract

Sparse autoencoders have become a standard tool for interpreting the internal representations of neural language models, but feature validation remains largely correlational and ad hoc. We introduce a sequential, anytime-valid hypothesis testing framework for validating SAE features against a causal necessity criterion. For each candidate feature, we construct an e-process from paired ablation interventions; across large feature dictionaries, we control the false discovery rate using the e-BH procedure, which is valid under arbitrary dependence between feature-level e-values, and which lets researchers stop at any sample size without inflating Type-I error. On a synthetic task with oracle ground truth, our procedure controls FDR below the target level (empirical FDP $= 0.000 \pm 0.000$ across 10 seeds at $\alpha = 0.05$) while recovering features in proportion to their causal effect size. On GPT-2 small with publicly available residual-stream SAEs, our procedure recovers six features on the indirect object identification benchmark at $\alpha = 0.05$ from $K = 89$ candidates, while Model-X knockoffs with a Gaussian fit recover none on the same data; five of the six recovered features have published interpretations directly aligned with the task structure.

1. Introduction

Sparse autoencoders (SAEs) have rapidly become a dominant lens for interpreting the internal representations of neural language models. By decomposing a model's residual stream into a large dictionary of sparse, often human-interpretable features, SAEs provide a tractable handle on otherwise dense and entangled activations (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024). The interpretability community has used them to study refusal behavior, sycophancy, deception, and a wide range of other phenomena. With these advances comes a methodological gap: the field knows how to discover candidate features, but lacks rigorous ways to validate which of those features actually drive specific model behaviors.

Current feature validation is largely correlational. Researchers inspect a feature's top activating examples, generate an automated explanation, and check whether the explanation appears plausible. When causal validation happens at all, it is usually a single ablation on a small set of hand-picked prompts, with no formal control of false discoveries. These problems compound at scale. An SAE with 24,000 features cannot be hand-validated, and naive multiple-testing corrections such as Bonferroni are too conservative to leave anything standing.

This paper introduces a hypothesis testing framework designed for the way interpretability researchers actually work. We frame each candidate SAE feature as a separate one-sided hypothesis: the null is that ablating the feature degrades task performance no more than ablating a comparably scaled random direction; the alternative is that the feature is genuinely necessary. We test the per-feature hypotheses sequentially using e-processes from the testing-by-betting literature (Shafer, 2021; Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024), which gives anytime-valid Type-I error control. A researcher can continue collecting intervention data on a feature until they reach a confident conclusion, with no correction for the act of looking at intermediate results. Across the candidate features, we control the false discovery rate via the e-BH procedure of Wang & Ramdas (2022), which is valid under arbitrary dependence between the per-feature e-values. The combination is well-suited to the actual interpretability workflow: continuous data collection, arbitrary stopping, large feature dictionaries, and unknown dependence structure between features.

We validate the procedure on two complementary settings. The first is a synthetic task with a small attention-only transformer and a small SAE, where we can recover oracle ground truth by exhaustive ablation. Across ten seeds, our procedure controls FDR below the target level (em-

[1]Department of Economics, University of California, Los Angeles, USA. Correspondence to: Rudransh Khera <kherarudransh@gmail.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

pirical FDP $= 0.000 \pm 0.000$ at $\alpha = 0.05$) and recovers features in proportion to their true causal effect size; features with strong effects are detected with near-100% probability, weaker features fail to reject as expected. The second setting is GPT-2 small with publicly available residual-stream SAEs from Bloom (2024), evaluated on the indirect object identification benchmark of Wang et al. (2023). From $K = 89$ candidate features active at the prediction position, our procedure rejects six at $\alpha = 0.05$. The Model-X knockoffs baseline of Candès et al. (2018), with a Gaussian fit, rejects zero on the same data and shows empirical FDR violations at higher α levels where the Gaussian assumption is more strained. Five of the six features our procedure recovers have published interpretations on Neuronpedia that align directly with the IOI task structure (transfer of ownership, recipients, awarding), supporting the conclusion that the test is recovering meaningful structure rather than artifacts.

The contributions of this paper are: (1) a formal hypothesis specification for SAE feature validation as sequential paired-difference tests; (2) an anytime-valid testing procedure combining e-processes with e-BH for FDR-controlled selection across feature dictionaries; (3) empirical validation on a synthetic ground-truth task and on GPT-2 small with publicly available SAEs, demonstrating calibration, power, and a clean advantage over the natural batch baseline.

2. Related Work

Anytime-valid inference and e-values. The framework we use builds on the testing-by-betting tradition (Shafer, 2021; Ramdas et al., 2023). An e-process is a non-negative process whose expectation is bounded by one under the null at every stopping time, which by Ville's inequality gives Type-I error control at arbitrary stopping. Our betting strategy follows the plug-in Kelly construction (aGRAPA) studied by Waudby-Smith & Ramdas (2024), with a one-sided clip appropriate for our composite null $H_0 : \mathbb{E}[X] \leq 0$. We use the Online Newton Step variant of Cutkosky & Orabona (2018) as a robustness check. Adaptive bets in this family achieve growth rates close to the Kelly-optimal fixed bet at modest sample sizes.

Multiple testing and FDR control. The Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) controls FDR for p-values under positive dependence. We instead use the e-BH procedure of Wang & Ramdas (2022), which controls FDR under arbitrary dependence when applied to e-values. This robustness matters in our setting because per-feature e-processes share prompts, share model state, and operate on overlapping subspaces of the residual stream. Knockoff methods (Candès et al., 2018) provide a different route to FDR control in batch settings, requiring either a known generative model of the covariates (Model-X) or a fixed-design assumption. We compare against Model-X knockoffs with a Gaussian fit as the natural batch baseline; we discuss the empirical performance and the model-misspecification issue in Section 4.

Sparse autoencoders and feature validation. Sparse autoencoder methods for language model interpretability follow Bricken et al. (2023); Cunningham et al. (2024); Templeton et al. (2024). Bloom (2024) released the GPT-2 small residual-stream SAEs that we use in our real-model experiments. Validation of discovered features has mostly relied on automated explanations of top activating examples and on manual ablation studies. Marks et al. (2025) introduced sparse feature circuits, where features and their connections are validated through targeted ablations on a downstream task; the validation is sound but is performed without formal multiple-testing control. Kissane et al. (2024) extends SAE analysis to attention layer outputs and applies it to the IOI circuit, using SAE features to disambiguate the role of multiple seemingly-redundant induction heads. The most direct prior work on hypothesis testing for mechanistic interpretability is Shi et al. (2024), which formalizes the circuit hypothesis as a batch test on a single discovered circuit. Our work is complementary: we operate at the per-feature level and provide FDR-controlled selection across the entire feature dictionary, whereas Shi et al. (2024) provides a single-circuit test once candidate circuits are identified.

3. Method

3.1. Hypothesis Specification

Fix a model M, an SAE attached at a chosen residual-stream hook with feature dictionary of size K, and a task whose performance is quantified by a logit difference $\Delta(\cdot)$ between a correct and an incorrect token at the prediction position. For prompt i, we measure the effect of intervening on feature j as

$$\Delta_i^{(j)} = \Delta(M_{\setminus j}(\text{prompt}_i)) - \Delta(M(\text{prompt}_i)), \quad (1)$$

where $M_{\setminus j}$ denotes the model with feature j ablated (the SAE is run, the j-th feature activation is set to zero, and the residual stream is reconstructed). The paired observation that goes into the sequential test is

$$X_i^{(j)} = \Delta_i^{(j_{\text{rand}})} - \Delta_i^{(j)}, \quad (2)$$

where j_{rand} is a random feature index drawn independently for each prompt. Under the null hypothesis that feature j is no more necessary for the task than a random comparison feature, $\mathbb{E}[X_i^{(j)}] \leq 0$. Under the alternative, the expectation is strictly positive. We test the one-sided composite hypothesis $H_0^{(j)} : \mathbb{E}[X^{(j)}] \leq 0$.

The estimand is relative. Rejecting $H_0^{(j)}$ certifies that ablating feature j degrades task performance more than ablating a feature drawn uniformly from the candidate set, not that j is necessary in an absolute sense. The random comparison fixes the reference point and absorbs the typical effect of removing an arbitrary feature, so the test isolates the excess effect attributable to j. A load-bearing feature can fail to reject when other features carry redundant signal, and the criterion is only as strong as the comparison distribution; we return to confounds in the comparison in Section 6.

3.2. E-process Construction

For each feature j we maintain a wealth process $E_t^{(j)}$ initialized at one. After observing $X_t^{(j)}$, we update

$$E_t^{(j)} = E_{t-1}^{(j)} \cdot (1 + \lambda_t^{(j)} X_t^{(j)}), \tag{3}$$

where $\lambda_t^{(j)} \in [0, c]$ is a betting fraction predictable with respect to the filtration $\mathcal{F}_{t-1}$ generated by observations $X_1^{(j)}, \ldots, X_{t-1}^{(j)}$, and $c = 0.5/B$ for an a priori bound B on $|X_t^{(j)}|$. The one-sided constraint $\lambda_t \geq 0$ is appropriate for our composite null; allowing negative bets in (3) would correspond to betting against H_1, which is not informative for our test direction.

We use the aGRAPA betting strategy (Waudby-Smith & Ramdas, 2024), which sets

$$\lambda_t^{(j)} = \mathrm{clip}_{[0,c]} \left(\frac{\hat{\mu}_{t-1}}{\hat{\sigma}_{t-1}^2 + \hat{\mu}_{t-1}^2} \right), \tag{4}$$

with $\hat{\mu}_{t-1}, \hat{\sigma}_{t-1}^2$ the running mean and variance of past observations. Under the null, $E_t^{(j)}$ is a non-negative supermartingale with $\mathbb{E}[E_t^{(j)}] \leq 1$ for all t. Ville's inequality gives $\mathbb{P}(\sup_t E_t^{(j)} \geq 1/\alpha) \leq \alpha$ under $H_0^{(j)}$, so we can stop at any data-dependent time and reject when wealth crosses $1/\alpha$.

Three conditions underpin this guarantee. Prompts are sampled i.i.d. and the comparison index j_{rand} is drawn independently at each step, so the observations $X_t^{(j)}$ are i.i.d. conditional on the fixed model and SAE; the betting fraction $\lambda_t^{(j)}$ is predictable, depending only on $X_1^{(j)}, \ldots, X_{t-1}^{(j)}$, which holds by construction in (4); and the paired difference satisfies $|X_t^{(j)}| \leq B$, which with $\lambda_t \in [0, 0.5/B]$ keeps each factor $1 + \lambda_t^{(j)} X_t^{(j)}$ in $[0.5, 1.5]$ and the wealth process non-negative. Under these conditions and the null $H_0^{(j)}$, the conditional mean $\mathbb{E}[X_t^{(j)} \mid \mathcal{F}_{t-1}]$ is at most zero, so $E_t^{(j)}$ is a non-negative supermartingale and the Ville bound applies. The bound B is set a priori (Appendix A); its selection is discussed in Section 6.

Algorithm 1 Anytime-valid SAE feature validation

Input: model M, SAE with K features, prompts $\{p_i\}$, level α, bound B, budget T
Initialize $E^{(j)} \leftarrow 1$ for each $j \in \{1, \ldots, K\}$
for $t = 1$ **to** T **do**
  for $j = 1$ **to** K **do**
    Sample random comparison feature $j_{\text{rand}} \neq j$
    Compute $X_t^{(j)}$ via Equations (1) and (2)
    Compute $\lambda_t^{(j)}$ from $X_1^{(j)}, \ldots, X_{t-1}^{(j)}$ via (4)
    Update $E^{(j)} \leftarrow E^{(j)} \cdot (1 + \lambda_t^{(j)} X_t^{(j)})$
  end for
end for
Apply e-BH at level α to $\{E^{(j)}\}_{j=1}^K$
Return: set of rejected features

3.3. FDR Control via e-BH

After running the per-feature e-processes (either to a fixed sample budget or until each individually rejects), we collect the final e-values $\{E^{(j)}\}_{j=1}^K$ and apply the e-BH procedure of Wang & Ramdas (2022). Sort the e-values in decreasing order $E_{(1)} \geq E_{(2)} \geq \ldots \geq E_{(K)}$, find the largest k for which $E_{(k)} \geq K/(k\alpha)$, and reject the corresponding k hypotheses. This procedure controls FDR at level α under arbitrary dependence between the e-values, which is exactly the regime we face when features share prompts and overlap in feature space.

The full procedure is summarized in Algorithm 1.

A researcher can stop the outer loop at any point, including data-dependent stopping per feature, without compromising Type-I error. Anytime-validity is a property of each individual e-process, and FDR control under arbitrary dependence is a property of e-BH, so the two guarantees compose without additional correction.

4. Synthetic Experiment

4.1. Setup

To assess calibration and power against ground truth, we construct a small attention-only transformer and SAE for which oracle causal features can be recovered by exhaustive ablation. The task is a simplified indirect object structure: each input has the form $[S, IO, \text{FILL}, S]$, the model is trained to predict the indirect object IO at the final position. The vocabulary is eight name tokens plus one filler token ($V = 9$), the sequence length is four, and the training set contains 4,000 sequences. We train a 2-layer attention-only transformer ($d_{\text{model}} = 64$, four heads) to convergence (eval accuracy 1.00). We train a ReLU SAE on the residual stream prior to layer 1 with $d_{\text{SAE}} = 512$ and an L1 sparsity penalty; the trained SAE achieves 99.98% explained

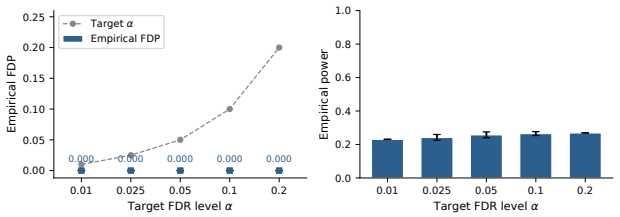


Figure 1. FDR control and power on the synthetic experiment, averaged across 10 seeds at $n = 500$ prompts per feature. Left: empirical FDP (blue) at each target α level (gray dashed) is exactly zero; labels show numerical values explicitly because the bars have zero height. Right: empirical power across the same target levels.

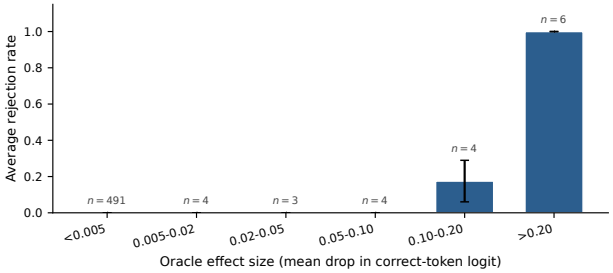


Figure 2. Average rejection rate as a function of true causal effect size (oracle ablation score), 10 seeds at $n = 500$ prompts per feature, $\alpha = 0.05$. The number above each bar is the count of features in that bin. Power saturates at 100% for strong-effect features (> 0.20) and degrades gracefully toward zero on weaker effects.

variance on held-out activations.

Oracle ground truth is obtained by ablating each of the 512 features on every evaluation prompt and recording the mean drop in the correct-token logit at the final position. Features whose mean effect exceeds the 95th percentile of the score distribution are labeled causal; this gives 26 causal features and 486 null features. We treat the resulting binary mask as ground truth for evaluating FDR and power.

4.2. Calibration: FDR Control

We run our procedure with $n = 500$ prompts per feature, betting strategy aGRAPA, bound $B = 10$, across 10 random seeds. Figure 1 reports empirical FDP and power averaged across seeds at each of five target α levels. Two findings stand out. First, the empirical FDP is exactly zero across all 10 seeds at every target level we tested (annotated as 0.000 in the left panel), so our procedure is conservative relative to the FDR guarantee, as expected for the dependence-robust e-BH correction. Second, power increases monotonically with α from 0.226 at $\alpha = 0.01$ to 0.262 at $\alpha = 0.20$, with very small standard deviations across seeds (typically < 0.02).

4.3. Power as a Function of Effect Size

The 26 oracle-causal features are not uniformly easy to detect. Their true effect sizes (mean drop in correct-token logit upon ablation) span roughly two orders of magnitude. Figure 2 shows the average rejection rate as a function of the binned effect size, computed across the same 10 seeds. Power is essentially zero for features with effect size below 0.05, climbs to roughly 17% in the 0.10-0.20 range, and saturates at 100% for features with effect size above 0.20. This graceful degradation is the expected behavior of a calibrated test under finite samples: stronger signals are detected with more confidence, and the procedure stays conservative on signals near the noise floor.

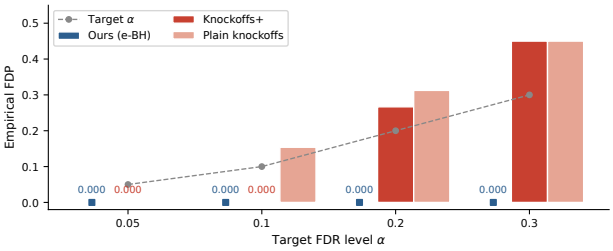


Figure 3. FDR control on synthetic SAE data: our procedure (e-BH on e-values) versus Model-X knockoffs with a Gaussian fit. Our procedure controls FDP at zero across all four target levels (annotated 0.000). Knockoffs+ rejects nothing at $\alpha \in \{0.05, 0.10\}$ and shows empirical FDP exceeding the target at $\alpha \in \{0.20, 0.30\}$; plain knockoffs shows FDP violations starting at $\alpha = 0.10$. The Gaussian fit appears to misspecify the heavily non-Gaussian SAE activation distribution, which leads to anti-conservative behavior at higher target levels.

4.4. Comparison to Knockoffs

We compare against Model-X knockoffs with a Gaussian fit, which is the natural batch competitor for FDR-controlled feature selection. We collect the same 500 prompts of feature activations and correct-token logits, estimate the joint mean and covariance of the active features, and construct equicorrelated knockoffs from this Gaussian approximation. We then fit a cross-validated Lasso of the logit on the augmented design $[X, \tilde{X}]$ and apply the standard knockoff threshold for both knockoffs+ and the plain variant. Figure 3 reports the results for four target α levels.

Two observations. First, our procedure controls FDR strictly at every target level. Second, knockoffs+ either fails to reject (at $\alpha = 0.05$ and 0.10) or violates the FDR target (at $\alpha = 0.20$, empirical FDP 0.267; at $\alpha = 0.30$, empirical FDP 0.450). Plain knockoffs violates FDR control starting at $\alpha = 0.10$. The Gaussian assumption is badly violated for sparse SAE activations, and the resulting anti-conservatism is a known failure mode for Model-X under model misspec-

ification. Model-X knockoffs targets a related but distinct estimand, FDR-controlled selection of features conditionally associated with the logit given a model of the activation distribution, rather than the per-feature necessity contrast we test. We include it as the natural batch competitor for FDR-controlled feature selection; the comparison shows that this apparently natural batch approach runs into real difficulty on sparse SAE activations, while our procedure does not.

5. Real-Model Experiment: GPT-2 Small on IOI

5.1. Setup

We evaluate the procedure on GPT-2 small with the publicly available residual-stream SAEs of Bloom (2024) attached at the pre-residual hook of layer 7 ($d_{\text{SAE}} = 24{,}576$, expansion factor 32). The task is the indirect object identification benchmark of Wang et al. (2023): prompts of the form "When S and IO went to the place, S gave the object to" followed by a prediction position. The correct completion is the indirect object name; the distractor is the repeated subject. We generate 2,000 IOI prompts using 16 single-token names, 6 places, 10 objects, and 3 templates; the underlying GPT-2 small model gets the IOI completion right on the overwhelming majority of these prompts. We use the first 1,000 prompts for testing and reserve the remainder for the knockoffs comparison.

We restrict the test to features that activate at the prediction position on at least 20% of the IOI prompts, since features silent at this position cannot affect the next-token logit through ablation. This filter yields $K = 89$ candidate features. We then run our procedure with the same hyperparameters as the synthetic experiment (aGRAPA, $B = 10$, $\alpha = 0.05$).

5.2. Results

Out of $K = 89$ candidate features tested at $n = 1{,}000$ prompts, our procedure rejects six at $\alpha = 0.05$ via e-BH. The largest e-value is $\sim 4.9 \times 10^{11}$ for feature 14455; the e-BH threshold for the top hypothesis is $K/\alpha = 1780$. The e-BH cutoff falls at rank $k = 6$: the sixth-largest e-value (1425) clears the rank-6 threshold $K/(6\alpha) \approx 297$, while no higher rank index satisfies the e-BH condition, so the remaining 83 features are not rejected. Five of the six rejected features exceed the top-hypothesis threshold of 1780; feature 5320 (e-value 1425) is rejected because its rank-6 threshold is lower. The Model-X knockoffs baseline of Candès et al. (2018), applied to the same prompts and same restricted feature set, rejects zero features.

Figure 4 shows the wealth trajectories of the three top features. The trajectories are smooth, monotone, and clearly separated by signal strength: feature 14455 crosses the single-test threshold near prompt 130 and the top-hypothesis threshold near prompt 270 and continues climbing exponentially; feature 7176 climbs more slowly; feature 8171 is the slowest of the three but still ends above the top-hypothesis threshold, so all three are rejected by e-BH.

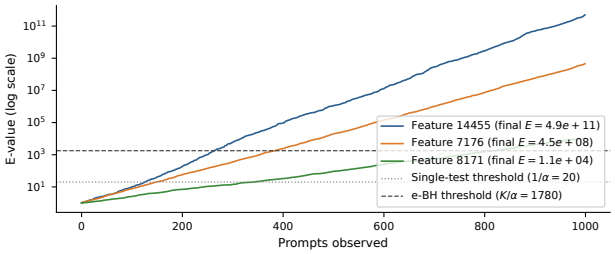


Figure 4. Wealth trajectories for the three highest-e-value features on GPT-2 small + Bloom SAE @ layer 7, IOI task, $n = 1{,}000$ prompts. The y-axis is in log scale. The dotted line is the single-test rejection threshold ($1/\alpha = 20$); the dashed line is the e-BH rejection threshold for the top hypothesis ($K/\alpha = 1780$ with $K = 89$). Feature 14455 reaches an e-value of $\sim 4.9 \times 10^{11}$ at the budget boundary. Feature 7176 climbs steadily; feature 8171 is slower still but also ends above the top-hypothesis threshold, and all three are rejected.

5.3. Qualitative Case Study

Table 1 reports the six rejected features together with paraphrased interpretations from Neuronpedia, which hosts auto-generated explanations and dashboard views for the Bloom SAEs. We find that five of the six recovered features have semantics directly aligned with the IOI task structure (transfer of ownership, recipient roles, awarding of resources, agentic verbs of giving and handing). Feature 14455 in particular is described on Neuronpedia as the preposition "to" in transactional contexts (e.g., "sold to companies", "handed to her", "passed to senior officers"), a description that matches the role of "to $\langle IO \rangle$" in the IOI prompt almost exactly. Feature 11936 is similarly described as awarding or providing resources to entities. The remaining IOI-aligned features (5320, 8171, 7176) capture closely related semantic roles. Feature 20598 is an honest outlier: its top activations are violent or criminal incidents, with no obvious surface alignment to the IOI task. We do not regard this as a failure of the test. The e-process correctly registered that ablating this feature reduces the IOI logit margin more than ablating a random feature does. The discrepancy between the procedure's verdict and the auto-interp summary may reflect that the feature plays a role in the model's representation of named individuals as objects of action that the auto-interp explanations do not capture; understanding such cases is exactly the use case for which a calibrated, FDR-controlled test is helpful.

Table 1. Six SAE features in GPT-2 small (layer 7, Bloom SAEs) rejected by our procedure on the IOI task at FDR-controlled $\alpha = 0.05$ with $n = 1000$ prompts and $K = 89$ candidates. Interpretations paraphrased from Neuronpedia auto-interp explanations. Five of six features have semantics directly relevant to the IOI task structure ('X gave Y to Z'); feature 20598 is an honest outlier discussed in Section 5.3.

Feature	E-value	Firing rate	Interpretation
14455	$\sim 10^{11}$	1.00	Transfer of ownership; *'to'* in transactional contexts (selling, handing over).
7176	$\sim 10^{8}$	0.31	People described as objects of action (girlfriend, daughter, son).
8171	10614	1.00	Personal interactions and relational contexts between named individuals.
11936	8825	0.99	Awarding or providing resources/support to entities.
20598	3179	0.36	Violent or criminal incidents; entangled with name-prediction circuits.
5320	1425	1.00	Agentic action verbs (sought, brought, intervened, signed).

6. Discussion and Limitations

Why this works. The combination of an anytime-valid sequential test with a dependence-robust FDR procedure fits the structure of SAE feature validation cleanly. Per-feature, we get adaptive sample efficiency: easy features are rejected with few prompts, hard features get the budget they need, and the researcher can stop or continue at any point without inflating Type-I error. Across features, e-BH provides FDR control under arbitrary dependence, which matters because per-feature e-processes share prompts and operate on overlapping subspaces. The dependence-robust guarantee also makes the procedure conservative in practice; we observe strict FDP $= 0$ in our experiments, in line with the conservative nature of e-BH under unknown dependence.

Limitations. We tested only one real-model setting (GPT-2 small with Bloom layer-7 SAEs) on one task (IOI). The e-process construction depends on a bound B on the per-prompt paired difference, and a poorly chosen bound either clips the data or chokes the betting strategy; in our experiments $B = 10$ worked well across both settings, but the right choice in a new setting will require some calibration. We focused exclusively on the necessity hypothesis (ablation-based) rather than the sufficiency hypothesis (steering-based), because steering interventions push the residual stream off-distribution and the resulting effects can be more about distribution shift than about the natural causal role of the feature. The framework extends to sufficiency tests, but the methodological questions that arise there are deferred to future work.

Confounds in the comparison feature. The comparison index is drawn uniformly, so it is not matched to feature j on firing rate, activation magnitude, or decoder norm. A feature that fires on every prompt with large activations perturbs the residual stream more when ablated than a sparse low-magnitude feature, independent of any task-specific role, so part of the excess effect the test attributes to j could reflect these covariates. The real-model results offer partial reassurance: the rejected features span firing rates from 0.31 to 1.00 (Table 1), and the second-largest e-value comes from a feature firing on 31% of prompts, so firing rate alone does not order the rankings. A controlled test that matches the comparison feature on firing rate and norm would tighten the estimand, and we leave this to future work.

Future work. The most natural extensions are cross-model corroboration (running the same procedure on Pythia, Llama, or Gemma with their respective SAEs), broader task coverage (refusal, sycophancy, factual recall), and the sufficiency hypothesis with careful treatment of off-distribution intervention. The procedure can also support more complex hypotheses, such as testing pairs or sets of features as putative circuits, by adapting the per-feature e-process to a multi-feature observation; we expect this to be a productive direction.

7. Conclusion

We introduced a sequential, anytime-valid hypothesis testing framework for validating sparse autoencoder features against a causal necessity criterion, with FDR control across the feature dictionary via e-BH. The procedure is calibrated, gracefully degrades with weaker signals, controls FDR under arbitrary dependence between features, and supports the actual interpretability workflow of continuous data collection and arbitrary stopping. On synthetic ground truth and on GPT-2 small, the procedure performs as predicted by the underlying theory and outperforms the natural batch competitor in both settings.

Impact Statement

This paper presents a methodological contribution to the statistical validation of features discovered by sparse autoencoders in language models. By providing FDR-controlled, anytime-valid tests for feature necessity, the work supports more rigorous interpretability research, which we view as broadly aligned with safe and reliable deployment of language models. We do not foresee specific negative societal consequences beyond those already present in mechanistic

interpretability research more generally.

References

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Bloom, J. Open source sparse autoencoders for all residual stream layers of GPT2-small. https://www.alignmentforum.org/posts/f9EgfLSurAiqRJySD/open-source-sparse-autoencoders-for-all-residual-stream, 2024. AI Alignment Forum.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features/index.html, 2023.

Candès, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in Banach spaces. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1493–1529, 2018.

Kissane, C., Krzyzanowski, R., Bloom, J., Conmy, A., and Nanda, N. Interpreting attention layer outputs with sparse autoencoders. arXiv preprint https://arxiv.org/abs/2406.17759, 2024.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. arXiv:2403.19647.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Shafer, G. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647.

Shi, C., Beltran-Velez, N., Nazaret, A., Zheng, C., Garriga-Alonso, A., Jesson, A., Makar, M., and Blei, D. M. Hypothesis testing the circuit hypothesis in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=5ai2YFAXV7.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html, 2024.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Wang, R. and Ramdas, A. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022. doi: 10.1111/rssb.12489.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

A. Implementation Details

The procedure was implemented in PyTorch with TransformerLens for model loading and SAELens for sparse autoencoder management. Synthetic experiments were run on a single Apple M3 chip; the GPT-2 small experiments were run on the same hardware over approximately three hours of compute.

Hyperparameters used in all experiments:

- Bound $B = 10$ on the per-prompt paired difference.
- Maximum betting fraction $c = 0.5/B = 0.05$.
- Betting strategy: aGRAPA with running mean and variance estimates of past observations.
- E-BH applied at level $\alpha = 0.05$ unless otherwise specified.
- Random comparison feature index j_{rand} drawn uniformly from $\{1, \ldots, K\} \setminus \{j\}$ at each prompt.

For the GPT-2 small experiment we use the Bloom residual-stream SAE at layer 7, with hook `blocks.7.hook_resid_pre`. The full dictionary contains 24,576 features. After filtering for features that activate at the prediction position on at least 20% of IOI prompts, $K = 89$ candidate features are submitted to the test.

Code to reproduce all experiments and figures is available at https://github.com/kherarudransh-oss/anytime-valid-hypo-testing-sae-feature-validation.

B. Neuronpedia References for Rejected Features

The six rejected features in our GPT-2 small + IOI experiment can be inspected interactively on Neuronpedia. Each URL displays the auto-interp explanation, top activations, and logit attribution for the corresponding feature.

- Feature 14455: https://www.neuronpedia.org/gpt2-small/7-res-jb/14455
- Feature 7176: https://www.neuronpedia.org/gpt2-small/7-res-jb/7176
- Feature 8171: https://www.neuronpedia.org/gpt2-small/7-res-jb/8171
- Feature 11936: https://www.neuronpedia.org/gpt2-small/7-res-jb/11936
- Feature 5320: https://www.neuronpedia.org/gpt2-small/7-res-jb/5320
- Feature 20598: https://www.neuronpedia.org/gpt2-small/7-res-jb/20598