Explaining Headline A/B Test Outcomes with Sparse Autoencoder Features

Drejc Pesjak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
drejc.pesjak@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Jan Rupnik
Nordoon Inc.
Palo Alto, California, USA

Abstract

A/B testing is a standard method for optimizing news headlines, yet it reveals only *which* variant wins, not *why*. We propose representing headline pairs through Sparse Autoencoder (SAE) features extracted from a large language model. For each within-test pair we construct a *context* vector (shared topic) and a *diff* vector (wording differences), apply histogram-based binarization to suppress activation noise, and train gradient-boosted trees followed by RuleFit to extract interpretable feature patterns. On approximately 57,000 pairs from the Upworthy Research Archive, our best model achieves 61.8% pairwise accuracy and 0.76 AUC. The dominant predictive features are *diff* features with recognizable labels such as "clickbait-style framing," "curiosity gaps," "emotional intensity," and use of the demonstrative "this." We also test a *hierarchy hypothesis*—that topic context gates which diff features matter—but find no stable topic-gated regimes. The results demonstrate that SAE features provide a viable middle ground between opaque dense embeddings and shallow bag-of-words representations for interpretable text analysis.

Keywords

large language models, sparse autoencoders, A/B testing, headline optimization, interpretable machine learning, mechanistic interpretability

1 Introduction

Many decisions in digital media come down to a choice between two variants of the same content, where small differences in wording translate into large differences in reach. A/B testing settles such questions empirically: show both variants to random audiences and measure which one wins. The outcome, however, is merely a winner and a loser. An explanation of why would be more useful, as it carries over to headlines that were never tested.

Moving from winner identification to interpretable explanation requires a text representation that a statistical model can learn from. Early approaches—bag-of-words, TF-IDF, n-grams—are interpretable but miss context and meaning. Dense contextual embeddings from Transformer models [14] such as BERT [6] capture rich semantics but are opaque: no single dimension has a human-readable meaning.

Mechanistic interpretability research offers a promising resolution. Sparse Autoencoders (SAEs) [2, 5] trained on LLM activations decompose dense activation vectors into a much larger, sparse set of features—each corresponding to a recognizable concept such as "shocking scenarios" or "use of the word *this*." These features can be labeled automatically [1], and the GemmaScope-2 project [10] provides pre-trained SAEs for the Gemma model family with human-readable descriptions hosted on Neuronpedia.

In this work we use SAE features as the representation layer for analyzing A/B-tested news headlines from the Upworthy Research Archive [9]. We encode each headline through Gemma-3-4B-IT[1], decompose the activations with a 16,384-width SAE, and construct pairwise *context + diff* vectors that separate shared topic content from wording differences. After binarizing the features to suppress activation noise, we train XGBoost and RuleFit models to identify which SAE features predict the winner and to extract human-readable rules.

We also investigate a hierarchical regime hypothesis: the possibility that the topic or context of an article gates which "difference" features matter for clickability, and test this through unsupervised clustering and model-based recursive partitioning.

2 Related Work

Headline A/B testing and engagement. The Upworthy Research Archive [9] made a large corpus of real headline A/B tests publicly available for research. Gligorić et al. [8] analyzed the same archive with logistic regression on hand-crafted linguistic features (word count, sentiment, named entities, syntactic patterns), finding small but significant effects of specific wording choices on click-through rates. Robertson et al. [13] used randomized Upworthy trials to show that negatively worded headlines drive higher consumption. Clickbait detection work [3] has identified surface-level engagement cues such as forward references and sensational language. Our approach differs from all of these in using SAE features, which provide richer, model-derived interpretable dimensions rather than manually defined linguistic variables or opaque dense vectors.

Sparse autoencoders and interpretable text representations. Sparse autoencoders trained on neural network activations learn overcomplete dictionaries where individual features correspond to human-recognizable concepts [2, 5]. The GemmaScope project [10] provides pre-trained SAEs for the Gemma model family, including JumpReLU variants [12] that use a learnable threshold to enforce sparsity. Feature labels can be generated automatically by having an LLM describe each feature's activation pattern [1]. Most directly related to our work, Movva et al. [11] train SAEs on sentence embeddings and use L1-penalized logistic regression

[1] google/gemma-3-4b-it on Hugging Face.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.89

to select predictive features for hypothesis generation. Our work differs in three ways: (i) we use residual-stream SAEs rather than embedding-space SAEs, (ii) we introduce a context+diff decomposition for pairwise comparisons, and (iii) we use tree-based models and RuleFit for conjunctive rule extraction rather than linear feature selection.

Interpretable machine learning. RuleFit [7] combines tree-derived conjunction rules with linear terms, producing models that are both predictive and human-readable. XGBoost [4] provides strong gradient-boosted baselines with built-in feature importance. Model-based recursive partitioning (MOB) [15] tests for parameter instability across subgroups, offering a principled way to discover heterogeneous treatment effects—we use it to test for topic-gated regime structure.

3 Methods

3.1 Dataset

We use the Upworthy Research Archive [9], which contains A/B test data from the news website Upworthy. Each "test" compared 2–5 headline variants for the same article, with real impression and click counts used to calculate click-through rate (CTR). We use the *confirmatory* split, comprising approximately 13,500 unique test IDs.

After preprocessing (HTML tag removal, deduplication, removal of single-headline tests), we sample within-test headline pairs. For each test with N headlines, we draw up to $\min\left(\binom{N}{2}, 10\right)$ random pairs, yielding approximately 57,289 pairs with balanced labels. The pair order is randomized, and the binary label is:

$$y = \mathbf{1}\left[\mathrm{CTR}_A > \mathrm{CTR}_B\right] \tag{1}$$

where CTR = clicks/impressions for each headline. We split into training (45,657 pairs) and test (11,632 pairs) sets using `GroupShuffleSplit` by `test_id`, ensuring that all pairs from the same test remain in the same split to prevent data leakage.

3.2 SAE Feature Extraction

Each headline is wrapped in a Gemma-3 chat template and passed through Gemma-3-4B-IT (Google's 4-billion parameter instruction-tuned model). We extract residual stream activations $\mathbf{h}_t \in \mathbb{R}^{2560}$ at layer 22 for each token position t, masking out instruction template tokens so that only headline content tokens remain.

The content-token activations are then encoded through a pre-trained JumpReLU SAE from the GemmaScope-2 project [10]:

$$\mathbf{z}_t = \mathrm{JumpReLU}_\theta\left(\mathbf{W}_{\mathrm{enc}}\,\mathbf{h}_t + \mathbf{b}_{\mathrm{enc}}\right) \tag{2}$$

where $\mathbf{W}_{\mathrm{enc}} \in \mathbb{R}^{d_{\mathrm{SAE}} \times d_{\mathrm{model}}}$, $d_{\mathrm{SAE}} = 16{,}384$, and the JumpReLU activation sets entries below a learned threshold θ to zero:

$$\mathrm{JumpReLU}_\theta(x) = x \cdot \mathbf{1}[x > \theta] \tag{3}$$

To obtain a single vector per headline, we max-pool across all content token positions:

$$\mathbf{v} = \max_{t \in \mathcal{T}_{\mathrm{content}}} \mathbf{z}_t \tag{4}$$

yielding a sparse vector $\mathbf{v} \in \mathbb{R}^{16384}_{\geq 0}$ where each dimension corresponds to a named SAE feature (e.g., "discovery and revelation," "shocking scenarios").

3.3 Pairwise Context + Diff Representation

Predicting CTR from a single headline's SAE vector would mostly capture which *topics* are inherently popular, not what makes one headline beat another for the same article. Since we have within-test pairs, we can focus on wording differences.

For a pair (A, B) with SAE vectors $\mathbf{v}_A$ and $\mathbf{v}_B$, we compute:

$$\mathbf{c} = \frac{\mathbf{v}_A + \mathbf{v}_B}{2} \qquad \text{(context: shared topic/content)} \tag{5}$$

$$\mathbf{d} = \mathbf{v}_A - \mathbf{v}_B \qquad \text{(diff: wording differences)} \tag{6}$$

and concatenate them into a single feature vector:

$$\mathbf{x} = [\,\mathbf{c} \parallel \mathbf{d}\,] \in \mathbb{R}^{32768} \tag{7}$$

The diff vector encodes *wording* (not just word) *differences* between the two variants in the SAE feature basis: it captures whatever changes in the model's representation when the phrasing is altered, from surface cues (e.g., demonstratives, punctuation, syntactic templates) to higher-level semantic or stylistic framing (e.g., emotion, moral framing, curiosity gaps). The context vector captures the shared topic, included to allow models to potentially learn context-dependent rules (e.g., "emotional language helps for political articles but not for lifestyle content").

Each pair is weighted by the absolute CTR difference to emphasize pairs with clearer outcomes:

$$w = \left|\mathrm{CTR}_A - \mathrm{CTR}_B\right| \tag{8}$$

3.4 Feature Binarization

Raw SAE activations contain noise near the JumpReLU cutoff. The non-zero values exhibit characteristic distributional structure (Figure 1) that we exploit for denoising.

For *context* features (the per-pair mean activation) the distribution has two overlapping components. When only one headline of the pair fires the feature the mean is half the single-headline activation, producing a steeply decaying exponential just above the JumpReLU cutoff. When both headlines fire, the mean of two draws yields a bell-shaped body at higher values. For *diff* features the distribution has a central Gaussian at zero (both headlines activate the feature similarly), flanked by exponential-like tails shifted into the positive and negative directions (one headline activates the feature substantially more than the other).

We separate noise from signal with a histogram-based "walk" threshold. For context, we walk rightward from the smallest non-zero bin: when 4 consecutive bins are all higher than the current bin, the transition from the decaying component to the bell-shaped body has been reached, yielding threshold $\tau_{c,j}$ and binary encoding $\tilde{c}_j = \mathbf{1}[c_j > \tau_{c,j}]$. For diff features the same walk proceeds outward from zero in both directions, yielding symmetric thresholds $\pm\tau_{d,j}$ and trinary encoding:

$$\tilde{d}_j = \begin{cases} +1 & \text{if } d_j > +\tau_{d,j} \\ -1 & \text{if } d_j < -\tau_{d,j} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

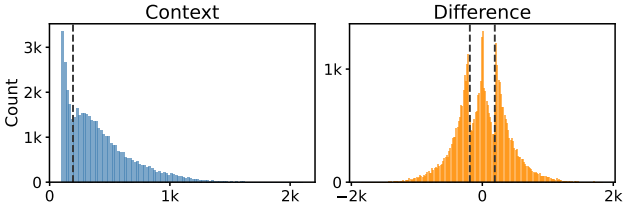


Figure 1: Context (left) and difference (right) activations for SAE feature 604 ("harmful unethical behavior"). Dashed lines mark walk thresholds.

3.5 Models

We evaluate the following models, all on the same group-based train/test split:

Logistic Regression. SGD-trained logistic regression with L2 regularization ($\alpha = 10^{-4}$) on the full 32,768-dimensional continuous feature vectors (we apply StandardScaler without centering to preserve sparsity), serving as a linear baseline.

XGBoost on binarized features. Gradient-boosted trees [4] with 100 estimators, maximum depth 6, learning rate 0.1, and column subsampling of 0.7, trained on the binarized/trinarized 32,768 int8 features. Sample weights are set to w (Equation 8). Only 467 out of 32,768 features receive non-zero importance from this model.

RuleFit on XGBoost-selected features. We extract the 467 features with non-zero XGBoost importance and train a RuleFit model [7] (25 estimators, tree size 4, maximum 500 rules). RuleFit generates both linear terms (each binarized feature with a coefficient) and conjunction rules (tree-derived feature combinations), producing a fully interpretable model.

Additional baselines. We also train a depth-8 decision tree (for visual interpretability), XGBoost with depth-1 stumps (to test whether additive features suffice), and per-cluster XGBoost models (to test the hierarchy hypothesis).

4 Results

4.1 Main Results

Experimental evaluation compares all models on the same group-based train/test split with sample weights $w = |\mathrm{CTR}_A - \mathrm{CTR}_B|$. Table 1 summarizes the results giving three commonly used evaluation metrics: classification accuracy (Acc), harmonic mean of precision and recall (F1) and area under the Receiver Operating Characteristic curve (AUC).

Table 1: Performance of all evaluated models on the held-out test set.

#	Method	Acc	F1	AUC
1	Logistic Regression	0.606	0.606	0.750
2	Decision Tree	0.556	0.551	0.630
3	XGBoost (continuous)	0.616	0.616	0.759
4	XGBoost per-cluster	0.571	0.564	0.673
5	**XGBoost (binarized)**	**0.618**	**0.619**	0.759
6	XGBoost stumps (d=1)	0.577	0.578	0.681
7	Decision Tree (binarized)	0.553	0.549	0.625
8	RuleFit (467 features)	0.588	0.556	**0.761**

Logistic regression is competitive (AUC 0.750), while XGBoost performs slightly better (0.759), indicating useful but limited nonlinear interactions. Depth-6 trees outperform stumps (0.759 vs. 0.681 AUC). Binarization slightly improves accuracy (61.6% → 61.8%) and concentrates importance on 467 of 32,768 features. RuleFit reaches comparable AUC (0.761) with 52 conjunction rules and 443 linear terms.

4.2 Top Predictive Features

Table 2 shows the ten most important linear terms from the RuleFit model. All are *diff* features—the SAE dimensions that capture how two headlines differ—confirming that *what distinguishes the headlines* matters far more than *what the article is about*. Context features appear only rarely among the top importances.

Table 2: Top 10 RuleFit linear terms by importance. All are diff features. A positive coefficient means that headline A activating this feature more than headline B predicts that A wins (higher CTR).

Rank	SAE Feature Label	Coef	Imp
1	"this" (demonstrative)	+0.14	0.076
2	clickbait-style titles	+0.11	0.070
3	remarkable, amazing qualities	+0.11	0.068
4	discovery and revelation	+0.10	0.061
5	subtitles	−0.12	0.061
6	past actions or events	+0.11	0.059
7	shocking or horrifying scenarios	+0.10	0.057
8	harmful unethical behavior	+0.08	0.056
9	math and trick questions	+0.09	0.056
10	stereotypes about gender roles	+0.09	0.055

Table 3: Examples of interpretable conjunction rules extracted by RuleFit. Positive and negative diff values favor headline A and B, respectively.

#	Rule	Coef
1	`clickbait titles` > 0.5	+0.07
2	`unexpected things` > 0.5	+0.06
3	`intense negative emotions` ≤ -0.5	−0.05
4	`recognizing or noticing` ≤ -0.5	−0.14
5	`joke, humor` > -0.5 and `things or anything` ≤ 0.5 and ctx: `asking questions` > 0.5	+0.03
6	`regulation, risk` > 0.5 and `past events` > -0.5	−0.04

The features tell a coherent story: headlines that use the demonstrative "this" (creating a curiosity gap), employ clickbait-style framing, evoke emotional intensity (shock, outrage), or promise revelation tend to get more clicks. Dry, subtitle-style text (feature 5, negative coefficient) hurts clickability.

4.3 Example Interpretable Rules

Table 3 shows selected RuleFit conjunction rules capturing interactions among binarized SAE features.

For instance, the fifth rule states that in the context of question-framed articles (context feature active), if the diff in humor is not strongly negative and the diff in "things or anything" vagueness is low, the more humorous variant tends to win. The sixth rule shows that regulatory/governance language combined with past-event framing *hurts* clickability.

Illustrative headline pair. Consider a real A/B test from the Upworthy archive where variant A is *"The Secret About Work Life In 20 Other Countries That Will Surprise You"* and variant B is *"In Japan, They Have A Word For Working Yourself To Death."* Both headlines cover the same topic (international work culture), so their context vectors are similar. However, variant A strongly activates the diff features for "clickbait-style titles" (+1) and "discovery and revelation" (+1) through its "The Secret…That Will Surprise You" framing, while variant B is specific and informative but does not trigger these engagement features. The model predicts A as the winner, consistent with the observed A/B test outcome.

4.4 Hierarchy Hypothesis

A central question was whether *context-dependent rules* govern clickability—for example, whether emotional language helps for political headlines but hurts for lifestyle content. We tested this through three approaches:

K-Means clustering on SVD-reduced context vectors (10 clusters) followed by per-cluster XGBoost models. The aggregated per-cluster predictions achieved 57.1% accuracy (vs. 61.8% global), a 4.7 percentage point degradation. The clusters fragmented the data without capturing distinct clickability regimes.

Model-Based Recursive Partitioning (MOB) [15] using context features as splitting variables and per-leaf Lasso regression on diff features. The MOB tree found statistically significant parameter instability (via the supLM test), producing 10 leaves with different regression coefficients. However, performance was severely overfit: train $R^2 = 0.39$ vs. test $R^2 = 0.04$, because each leaf had too few samples relative to the 16,384 diff features.

XGBoost tree dump analysis. We audited the split variables in the depth-6 binarized XGBoost model. Of 1,553 total splits, only 188 were on context features (12.1%), with the remaining 1,365 on diff features. Context splits appeared primarily near leaves, not at the root—suggesting the ensemble does not implement "context-first gating" in any strong sense.

These results indicate that while non-additive interactions exist (depth-6 substantially outperforms depth-1 stumps), they are predominantly *among diff features*—not between context and diff. The hierarchy hypothesis, as tested, is not supported.

5 Discussion and Conclusions

Our experimental analysis has shown that on our data the best model achieves 61.8% accuracy and 0.76 AUC. This is a modest but real signal, given that many A/B test pairs have very small CTR differences, which makes the ground-truth label itself noisy. The SAE features come from a general-purpose LLM that was never fine-tuned for headline effectiveness, yet they capture recognizable engagement patterns. Across all models the diff features dominate, so what distinguishes two headlines matters far more than what the article is about. The top features (clickbait framing, curiosity gaps, emotional intensity, use of "this") align with known headline optimization principles and provide independent empirical validation. Depth-6 trees outperform depth-1 stumps by 7.8 AUC points, suggesting that multi-way feature interactions carry signal, yet our attempts to find stable topic-gated regimes were unsuccessful. The interactions are among the diff features themselves rather than between context and diff, an informative negative result suggesting that hierarchical structure, if present, does not follow the topic dimension we assumed.

We conclude that Sparse Autoencoder features extracted from a large language model provide an interpretable representation for analyzing headline A/B test outcomes. The context + diff decomposition, combined with histogram-based binarization, produces human-readable feature importances and conjunctive rules at predictive performance comparable to continuous representations, placing SAE features between opaque dense embeddings and shallow bag-of-words representations.

Limitations and future work. The multi-stage pipeline compounds approximation errors, and the SAE feature labels are generated automatically, so they may not perfectly describe what each feature detects. All reported associations are predictive rather than causal: without controlled experiments we cannot claim that activating a particular SAE feature causes more clicks. Hard partitioning methods (k-means, MOB) may also miss soft or overlapping regime structure. Future work could explore larger SAE dictionaries (65k or 262k width), models fine-tuned for headline tasks, causal validation through controlled generation, and soft mixture-of-experts approaches to regime discovery. Future work should compare SAE features against lexical and dense representations and isolate the contributions of context and binarization through controlled ablations.

Acknowledgements

This research was financially supported by the Slovenian Research and Innovation Agency (ARIS) through the Young Researcher Grant (PR-13856), the research program Knowledge Technologies (P2-0103) and the Gravity project AI for Science (GC-0001) and European HE project DataPact (GA 101189771).

The authors acknowledge the use of Claude to support paper organization and language and grammar refinement. All AI-generated suggestions were carefully reviewed and revised by the authors, who take full responsibility for the content of this paper.

References

[1] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Joshua Clymer, et al. 2023. Language models can explain neurons in language models. *OpenAI blog*.

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Henighan, Tom Conerly, et al. 2023. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*. Anthropic.

[3] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: detecting and preventing clickbaits in online news media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 9–16.

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 4171–4186.

[7] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2, 3, 916–954.

[8] Kristina Gligorić, Ashton Anderson, and Robert West. 2023. Linguistic effects on news headline success: evidence from thousands of online field experiments (registered report). *PLOS ONE*, 18, 3, e0257091. doi:10.1371/journal.pone.0257091.

[9] J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quéré, and Charles Ebersole. 2021. The Upworthy Research Archive: a time series of 32,488 experiments in U.S. media. *Journal of Quantitative Description: Digital Media*, 1. Dataset: https://osf.io/jd64p/. doi:10.51685/jqd.2021.014.

[10] Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. 2025. Gemma scope 2: technical paper. https://huggingface.co/google/gemma-scope-2.

[11] Rajiv Movva, Kenny Peng, Nikhil Garg, Jon M. Kleinberg, and Emma Pierson. 2025. Sparse autoencoders for hypothesis generation. In *Forty-second International Conference on Machine Learning (ICML)* (Proceedings of Machine Learning Research). Vol. 267. PMLR. https://proceedings.mlr.press/v267/movva25a.html.

[12] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024. Jumping ahead: improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*.

[13] Claire E. Robertson, Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J. Van Bavel, and Stefan Feuerriegel. 2023. Negativity drives online news consumption. *Nature Human Behaviour*, 7, 5, 812–822. doi:10.1038/s41562-023-01538-4.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 5998–6008.

[15] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17, 2, 492–514.