Autointerp Simulation Scores Track Token Naming, Not Explanation Truth

Ayush Satish Govind **Prince Modi** **Ishit Gulati**
University of California San Diego
aygovind@ucsd.edu pmodi@ucsd.edu igulati@ucsd.edu

Abstract

Autointerp simulation scoring is used as a quality signal for sparse autoencoder latent descriptions, a use that presumes the score tracks whether a description is correct. We test that presumption with a within-latent 2×2 over whether a description names the latent's trigger token and whether its semantic gloss is true, on 16 token-driven latents of a Gemma-2-2B layer-12 GemmaScope SAE. Naming the token raises the score by 0.305, and does so for all 16 latents. Conditional on naming, swapping a true gloss for a false one changes it by 0.021 (95% CI -0.11 to $+0.15$). Truth is not inert: with no token named, the true gloss scores 0.281 higher. A judge-free baseline that marks tokens whose surface form appears in the description reproduces the effect: on the true gloss it matches the judge when the token is named (0.824 against 0.803) and scores zero when it is not (-0.004 against 0.629), and across the 100 latents of our selection pool it tracks the judge at $r = 0.61$. The judge is therefore not insensitive to semantics; it anchors on the token when one is available. Our false glosses are domain swaps that leave predictive content largely intact, so a high score licenses "this description names the right token" rather than "this description is correct", at least for latents where the token nearly is the explanation.

1 Introduction

Sparse autoencoders [Cunningham et al., 2024] decompose language model activations into more latents than anyone can inspect by hand, so the descriptions attached to them are written and checked automatically [Bills et al., 2023]. The standard check is simulation scoring: a judge model reads a description and predicts, token by token, how strongly the latent fires on its top-activating text, and the correlation between predicted and true activations becomes the score. Those numbers are published beside latents on Neuronpedia [Lin and Bloom, 2023], and used to compare SAE architectures [Karvonen et al., 2025]. Both uses read the score as evidence that a description is correct. Paulo et al. [2025] propose cheaper alternatives, but do not ask what the original measures.

Whether it is evidence of that is a question of measurement validity, separable from reliability. Chanin [2026] asks whether simulation scores are stable enough to compare; we ask a prior question, what the score responds to at all. A metric can be perfectly repeatable and still track something other than what it certifies. Huang et al. [2023] ask a related question of the explanations rather than the metric. Oikarinen et al. [2025] audit metrics directly, finding several that hold their score after the concept label is corrupted; we ask what the surviving signal is, separating corruption of the trigger token from corruption of the gloss.

There is reason to suspect this one misfires. The reference scorer's own documentation records a GPT-2 latent described as firing on female pronouns, which scores well largely because the token *she* appears inside unrelated words such as *shelled* [Lin, 2023]. The observation is not ours. What is missing is a measurement: how much of a score survives when a description is false, and how much

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Interpretability as a Science.

Latent: 9650
Description: The name "Mary" and its variations. (Neuronpedia description, re-scored 0.92)

(a) Top activating examples

Example 1: … s nice to meet you! I'm Mary. Thank you for stopping by Stirring Up…

Example 2: … 21:17 2. Jean Nelson 3. Mary Ellen Pakka DNS. Helen Bourchier DNS. Kay…

Example 3: … as Percy Weems (uncredited) Mary Patton as Mrs Ralston (uncredited)…

Example 4: … varria, Marcie Rubardt and Mary Yetter of CARE USA, and Rosemary…

(b) Ablated descriptions

Type	Description	Score
Token-Named & True Semantics	References to the personal name Mary appearing within varied prose and lists.	0.97
Token-Named & False Semantics	References to Mary-themed astronomical terminology and planetary observations.	0.93
Token-Omitted & True Semantics	References to a traditional feminine given name appearing within varied prose and lists.	0.69
Token-Omitted & False Semantics	References to astronomical terminology and planetary observations.	-0.01

Figure 1: **The score tracks whether a description names the trigger token, not whether it is true.** Latent 9650, Gemma-2-2B layer 12, GemmaScope res-16k. **(a)** Four of the twenty top-activating texts, trimmed around the peak. Shading gives the true activation, not the judge's prediction. **(b)** The same twenty texts scored against four descriptions. A false gloss costs 0.04 and still beats the published description's re-scored value (0.92); dropping the token from that false gloss costs 0.94.

when it is true but omits the token. The two are confounded in the wild, because a correct description of a token-driven latent almost always names the token. We separate them with a within-latent 2×2 varying naming and truth independently (Figure 1).

Alongside the 2×2 we report a judge-free string-match baseline, synthetic controls bounding what the correlation step resolves, and a judge noise floor. The scope is narrow and worth stating here rather than only in the limitations: one model, one SAE layer, one judge, one prompt, and latents selected for a single trigger token with a bare-token ratio above 0.962, so the token very nearly is the explanation.

2 Method

2.1 Scorer

We reimplement the simulation scorer used by Neuronpedia. For each latent we take its 20 top-activating texts and extract a 64-token window centred on the maximally activating token. The judge, given only a description and the token sequence, predicts an integer activation in $[0, 10]$ for every token. Predictions are pooled across the 20 texts and correlated (Pearson) against the true activations, discretized to $[0, 10]$ by the maximum over that latent's own texts. Per-set max normalisation and forced integer output follow the reference implementations [OpenAI, 2023]; Appendix A.7 gives the correspondence.

Our judge is `gpt-5.6-luna`, accessed through its provider's hosted API at default sampling parameters. Its temperature is not settable to zero, so predictions carry sampling noise. Every score uses one fixed prompt, recorded as hash `0faef73` in each result file, so all five descriptions within a

latent were scored identically. Neuronpedia used a logprob-capable completion model, so we take sampled integers rather than logprob-weighted expectations, which is coarser. We round rather than floor, and our prompt is a reconstruction rather than the few-shot original. Holding scorer, prompt and texts fixed within a latent cancels bias that is constant across conditions, but not a difference in *sensitivity* across them: a judge reliable at surface matching and unreliable at semantic inference would attenuate our truth contrast more than our naming contrast. Section 3.4 addresses that rather than assuming it away.

2.2 Design

For each latent we score five descriptions against the same fixed 20 texts: the original Neuronpedia description, and four constructed descriptions forming a 2×2 over whether the description names the trigger token and whether its gloss is true. The false gloss is a domain swap: a coherent claim about a semantic field the latent does not respond to, verified absent from its top-activating texts. The named arm states the trigger verbatim; the unnamed arm replaces it with a hypernym or paraphrase making the same semantic claim.

A hypernym has a wider extension than the token it replaces, so the unnamed arm is not a purely lexical edit: "a traditional feminine given name" also covers *Jean* and *Helen*, which appear in latent 9650's texts without firing (Figure 1). The naming contrast therefore confounds surface naming with extensional precision; see Section 4.

Descriptions were drafted with a language model then checked programmatically, confirming the trigger appears as a delimited word in both named conditions and in no form in either unnamed one. Word counts were balanced within latent to within five words (mean 13.3, range 6–28). Appendix A.2 gives the full criteria.

2.3 Latent selection

The design needs latents for which a single true description exists and can be negated, which restricts us to latents driven by an identifiable trigger token. From the 16,384 latents of the layer-12 res-16k SAE [Gemma Team et al., 2024, Lieberum et al., 2024] we select on two properties. The *bare-token ratio* is the activation recovered on the peak token with no context, divided by activation in full context; we keep the 100 highest-ranked, a ratio of at least 0.962. The *max-token fraction* is the proportion of a latent's 20 texts sharing the same peak token, for which we require at least 0.7, alongside an alphabetic trigger of at least three characters. Nineteen latents qualify, of which we scored 15. A sixteenth from an earlier round has a two-character trigger predating the length criterion.

Two features of this funnel work against generality. Three latents were excluded by inspection because their trigger appears inside unrelated words — *hat* in *hating*, *lambda* in *lamb* — the same phenomenon as the *she*/*shelled* case that motivates the paper, and plausibly where truth and naming come apart most. And a bare-token ratio above 0.962 means the trigger alone recovers almost all in-context activation, so naming it and being right about the mechanism are close to the same thing here. Both bias towards the null we report. Appendix A.1 gives counts at every stage.

2.4 Noise floor

Because the judge's temperature cannot be set to zero, rescoring the same description does not return the same number. We scored one condition (token-named, true semantics) three times for each of ten latents spanning the score range, holding description, texts and prompt fixed. Pooling the ten within-latent variances on deduplicated windows (Appendix A.3) gives a single-draw standard deviation on 20 degrees of freedom. This bounds rescoring jitter for one latent; it is not the uncertainty on a mean over 16, which is governed by between-latent variance.

3 Results

Naming the trigger token moves the simulation score by $+0.305$; replacing a true gloss with a false one, given the token is named, moves it by -0.021. The second number is interpretable only against what a scorer with no judge achieves (Section 3.1), what the correlation step can resolve (Section 3.2), and how much the judge varies when nothing changes (Section 3.3).

3.1 A judge-free baseline reproduces the effect

If the score responds to surface naming, a judge-free scorer should reproduce it. For each description we mark every token whose surface form appears as a content word in it, and correlate that binary vector against the true activations through the same pipeline: `grep` with a correlation on top, with no access to activations and no semantic model (Appendix A.5).

Table 1 gives the result beside the judge. When the token is named, the baseline reaches 0.824 against the judge's 0.803, so on the true gloss the judge buys nothing over surface matching. On the false gloss it retains a margin (0.782 against 0.684). When the token is omitted the baseline collapses to -0.004 while the judge holds 0.629. The baseline's naming effect is $+0.757$ and positive for all 16 latents. Its truth effect is $+0.140$, larger than the judge's but five times smaller than its own naming effect and positive in only 6 of 16 latents. Across all 100 latents in the selection pool, scored with their published descriptions, the baseline tracks the judge at Pearson $r = 0.607$ (Spearman $\rho = 0.559$). This escapes the max-token-fraction and trigger-length filters, though not the bare-token-ratio cut that defines the pool.

In nine of 16 latents a false but token-naming description outscores the published one. Heap et al. [2026] find autointerp scores survive randomising the transformer; token naming is a candidate for what survives.

Two conclusions follow, and the second matters as much as the first. Surface matching alone reproduces the naming effect, so that effect is not a property of LLM judges but of the measurement design. And the judge is demonstrably *not* insensitive: it contributes 0.633 over the baseline in the token-omitted true condition, where surface matching contributes nothing. The null in Section 3.4 is not the null of a blunt instrument.

3.2 The correlation step resolves differences it ought to resolve

We also replaced the judge's predictions with synthetic predictors built from the true activations, holding texts and the correlation step fixed (Appendix A.4). These bound what the metric resolves downstream of the judge; unlike Section 3.1 they say nothing about the judge itself.

Predictors with the sparsity of a plausible description but the wrong token support score at chance: shifting the true support three tokens right gives -0.018, and marking a matched number of random tokens 0.009. Predictors that identify where the latent fires score between 0.89 and 1.00, and coarseness is not penalised: peak token only scores 0.894, full support 0.893. A correlation near zero therefore means the wrong tokens were predicted, not too few. Because those figures are means over latents rather than per-latent bounds — five latents exceed 0.921 in the named-true condition — we note only that the truth contrast plausibly has less headroom in the named arm.

3.3 Judge noise floor

Scoring one description three times for each of ten latents gives a pooled within-latent standard deviation of 0.019 on roughly 20 degrees of freedom, so a difference between two independently drawn scores carries a standard deviation of 0.027. This puts all but three of the per-latent truth effects in Appendix A.8 above two difference-draw standard deviations; it does *not* bound the uncertainty on a pooled mean.

3.4 Naming the trigger token, not semantic truth, drives the score

Table 1 gives the condition means. Naming the trigger token raises the score by 0.305 (95% CI 0.20 to 0.41), positive for all 16 latents (one-sided sign test, $p < 0.001$). Semantic truth does not behave this way: within the named conditions, the true gloss scores $+0.021$ above the false one, 95% CI -0.105 to $+0.147$ (paired $t = 0.36$), positive for 10 of 16 latents (one-sided $p = 0.23$). The interval is wide: these data are consistent with a truth effect up to about half the naming effect, and we do not claim to have excluded one.

The judge's per-token predictions point the same way. Within a latent, predictions under the named-true and named-false descriptions correlate at 0.668, against 0.218 for the two unnamed descriptions. Part of that gap is lexical: the named pair shares the trigger verbatim by construction. Excluding

Table 1: Mean simulation score by condition, $N = 16$ latents. All contrasts are within-latent. SD, min and max are across latents for the judge score. **String-match** is the judge-free baseline of Section 3.1, which sees only the description text.

Condition	Judge	SD	Min	Max	String-match
Neuronpedia description (re-scored)	0.733	0.208	0.160	0.989	0.552
Token-named, true semantics	0.803	0.174	0.331	0.975	0.824
Token-named, false semantics	0.782	0.130	0.487	0.946	0.684
Token-omitted, true semantics	0.629	0.206	0.243	0.893	−0.004
Token-omitted, false semantics	0.347	0.347	−0.020	0.947	−0.002

trigger positions the two correlations fall to 0.264 and 0.089, so the contrast survives but the named pair is not invariant to the gloss (Appendix A.6).

Truth is not irrelevant. Within the unnamed conditions the true gloss scores 0.281 above the false one, an interaction of -0.260. The pooled truth effect averages over latents that disagree: per-latent values span -0.585 to $+0.387$ with a standard deviation of 0.237 (Figure 2).

4 Discussion and limitations

Everything reported here is one model, one SAE at one layer, one judge, and one prompt, on 16 latents from a heavily filtered population. Section 2.3 notes two filtering choices that bias towards our null, and we make no claim about latents that are not token-driven.

Four limitations bear on whether the null is a null. The truth effect's 95% CI of -0.105 to $+0.147$ does not exclude an effect half the size of the naming effect, so the claim rests on the string-match baseline rather than on the point estimate. The named conditions sit nearer the 0.89–0.92 oracle mean than the unnamed ones, so the interaction may be compressed for that reason as well as any substantive one. The unnamed arm replaces the trigger with a hypernym of wider extension, confounding surface naming with extensional precision. And our "false" glosses are domain swaps that leave predictive content largely intact — "Mary-themed astronomical terminology" still directs a reader to *Mary*. Simulation scoring is by construction a predictive criterion, so one reading is not that the metric fails to track truth but that domain-swapped glosses remain predictively adequate. That gap is what downstream uses elide, but it is a different claim.

A high simulation score therefore licenses the claim that a description names the right token, not that it is correct, for token-driven latents of the kind we study. Ranking descriptions, comparing SAEs, and gating publication all inherit that gap. The most useful next step is a condition we did not run: an uninformative description naming no token and making no claim, establishing the floor against which the naming effect should be read. The token-omitted false condition averages 0.347 and reaches 0.947, far above the string-match baseline there, so the judge recovers something from the texts alone that we have not characterised.

5 Conclusion

Across 16 token-driven latents, naming the trigger token moved the simulation score by 0.305, while replacing a true gloss with a false one moved it by 0.021, with an interval that does not exclude an effect half the size of the naming effect. A string-match baseline with no judge reproduces the naming effect in all 16 latents and matches the judge outright when the token is named, while contributing nothing when it is not. This is one model, one SAE layer, one judge, and one prompt, and the truth effect is not uniform across latents, so we read the result as a bound on what a high score licenses rather than a general property of simulation scoring. Whether a description is correct needs checking by something other than the score that currently gates it.

References

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023. Accessed 2026-09-01.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026. URL https://arxiv.org/abs/2605.18229.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.08600.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers, 2026. URL https://arxiv.org/abs/2501.17727v2.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.blackboxnlp-1.24/.

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.

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. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19.

Johnny Lin. Neuronpedia Scorer. GitHub repository, 2023. URL https://github.com/hijohnnylin/neuronpedia-scorer. Accessed 2026-09-01.

Johnny Lin and Joseph Bloom. Neuronpedia: Interactive reference and tooling for analyzing neural networks with sparse autoencoders. https://www.neuronpedia.org, 2023. Accessed 2026-09-01.

Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://arxiv.org/abs/2506.05774.

OpenAI. Automated Interpretability. GitHub repository, 2023. URL https://github.com/openai/automated-interpretability. Archived 2026-05-29; accessed 2026-09-01.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models, 2025. URL https://arxiv.org/abs/2410.13928.

A Appendix

A.1 Selection funnel

All 16,384 latents of the layer-12 res-16k SAE carry a Neuronpedia description. We take up to five top-activating texts for each from Neuronpedia's public activation dump, discarding examples that are weak, peak on BOS or EOS, peak within the first three positions, or contain an unknown token. This drops 4,424 BOS or EOS peaks, 1,876 early-position examples, and 3 weak examples; 726 latents lose every example, leaving 15,658 latents and 75,617 examples.

We then re-run the model on truncated windows containing a BOS token, w tokens of left context, and the peak token. The bare-token ratio is the activation recovered at $w = 0$ divided by that example's activation in full context, taken as a median across the latent's examples. 73.6% of latents recover nothing at $w = 0$, and 2,177 reach a median ratio of at least 0.7. We retained the 100 highest-ranked, corresponding to a ratio of at least 0.962. This is a rank cut chosen to fit the scoring budget, not a threshold.

Those 100 were scored with their Neuronpedia descriptions, which also yields 20 top-activating texts each. The max-token fraction (MTF) is the proportion of a latent's 20 texts whose peak token is the modal peak token, with a median of 0.55 over the 100 candidates. Requiring MTF ≥ 0.7 leaves 34, and requiring an alphabetic trigger of at least three characters with no markup delimiters leaves 19. Of these we scored 15; a sixteenth (latent 8252) comes from the earlier round described below and is not among the 19, since its two-character trigger predates the length criterion.

Three latents were excluded by inspection rather than by rule, because the manipulation is undefined for them. Latent 10013 fires on *hat* inside both *hating* and *that*, and 7119 on *lambda* but also on *lamb*, so neither admits a single true gloss. Latent 3384 fires on the split token *Parse*, which fails token identity despite an MTF of 0.75. As Section 2.3 notes, the first two are precisely the regime in which surface matching and semantic correctness are most likely to diverge, so their exclusion biases towards our null.

Selection proceeded in two rounds. Ten latents were described first, all with MTF between 0.80 and 1.00, before the trigger-length criterion was introduced; one of them (latent 8252, trigger *mi*) would not satisfy it. The criterion was applied when selecting the second round, reducing 24 remaining MTF-qualifying candidates to 10. From those we drew nine, stratified by Neuronpedia score into low, middle and high bands so the sample would not consist only of high-scoring latents. Three of the nine were then excluded by inspection, leaving six.

A.2 Description checks

Each `named` condition was verified to contain the trigger token as a whitespace- or punctuation-delimited word, using a word-boundary match against a suffix list rather than a substring match, so that *situated* does not count as naming *site*. No `unnamed` condition contains any inflected form of the trigger. No two conditions within a latent share a string, and the `unnamed` conditions were checked to describe a meaning rather than the token as an object. We did not verify that the unnamed conditions match the extension of the named ones, and Section 2.2 notes that in general they do not.

A.3 Duplicate windows

The 20 top-activating texts per latent yield between 12 and 20 distinct 64-token windows, with a mean of 16.8. Identical windows share a cache key and receive identical predictions within a draw, which would deflate a variance estimate, so the noise floor is computed on deduplicated windows. We have not established whether the duplication is present in Neuronpedia's activation dump or introduced when we centre a window on each example's peak token.

A.4 Sensitivity controls

A.5 String-match baseline

The baseline lowercases each description, strips non-alphanumeric characters, and removes English function words together with a small list of description boilerplate (*references*, *mentions*, *context*,

Table 2: Sensitivity controls, $N = 16$ latents, computed from stored activations without further scoring runs. These bound what the correlation step can resolve; they bypass the judge and so say nothing about it.

Predictor	What it predicts	Score
`oracle_graded`	Exact true activation at every token	1.000
`oracle_binary`	10 at every truly active token; 0 elsewhere	0.893
`oracle_top20`	10 at the 20 strongest tokens; 0 elsewhere	0.921
`oracle_peaks`	10 at each text's peak token; 0 elsewhere	0.894
`wrong_shift3`	Binary true support shifted three tokens right	-0.018
`wrong_random20`	10 at 20 random tokens	-0.005
`wrong_randomN`	10 at N random tokens, matching the true support size	0.009

tokens, *varied*, and similar). Each text token is normalised the same way, with the SentencePiece word-boundary marker removed, and assigned 10 if its normalised form is in the remaining set and 0 otherwise. The resulting vector is pooled across the same 20 texts and correlated against the same discretized true activations as every judge score, so the only difference from a simulation score is the substitution of exact string matching for the judge.

The stoplist is a researcher degree of freedom, so we report the baseline with no stoplist at all. Named-true then scores 0.431, named-false 0.382, and both unnamed conditions approximately -0.03; the naming effect is $+0.437$ and positive in 16 of 16 latents, the truth effect $+0.049$ and positive in 7 of 16, and the correlation with the judge over the 100 published descriptions is $r = 0.582$. Function words match in nearly every text and depress the absolute scores, but the qualitative result — a large naming effect present in every latent, and a small truth effect that is not consistently signed — is unchanged.

A.6 Prediction invariance

For each latent we pool the judge's predicted activations across the 20 texts to give one vector per condition, and correlate those vectors with each other within latent. Means over the 16 latents:

Pair	Mean Pearson
named-true vs. named-false	0.668
named-true vs. unnamed-true	0.742
named-false vs. unnamed-true	0.526
named-false vs. unnamed-false	0.457
named-true vs. unnamed-false	0.273
unnamed-true vs. unnamed-false	0.218

The two named conditions differ only in semantic content and yield predictions correlated at 0.668; the two unnamed conditions differ in the same way and yield 0.218. Because the named pair shares the trigger token verbatim, we recompute both excluding every position whose token is the trigger. The named pair falls to 0.264 and the unnamed pair to 0.089.

A.7 Reference implementation and prompt

Two choices follow the reference implementations [OpenAI, 2023]. Per-set max normalisation matches OpenAI's `calculate_max_activation`, which computes the maximum over a latent's own activation records rather than on a global scale. Forcing integer output has precedent in their `LogprobFreeExplanationTokenSimulator`, which samples integers directly rather than taking an expectation over token logprobs.

Our prompt is a reconstruction rather than the few-shot original. It is fixed across every score in this paper and identified by the hash `0faef73` recorded in each result file.

Table 3: Per-latent scores for all five conditions, run 0. NP is the published Neuronpedia description, re-scored by our scorer; NT, NF, UT, UF are token-named/omitted crossed with true/false semantics. Δ is the truth effect within the named conditions (NT $-$ NF). Words is the mean length of the four constructed descriptions.

Latent	NP	NT	NF	UT	UF	Δ	Words
152	0.689	0.699	0.775	0.719	0.711	-0.075	12.0
237	0.989	0.975	0.761	0.622	0.183	$+0.214$	11.2
939	0.809	0.795	0.636	0.751	-0.005	$+0.158$	18.2
1657	0.980	0.938	0.551	0.756	-0.013	$+0.387$	13.5
3133	0.813	0.902	0.762	0.893	-0.020	$+0.141$	17.2
4741	0.614	0.546	0.799	0.689	0.577	-0.253	8.5
5270	0.946	0.929	0.782	0.318	0.545	$+0.147$	10.2
7155	0.160	0.836	0.946	0.750	0.455	-0.110	17.2
8252	0.862	0.973	0.856	0.243	0.561	$+0.117$	9.5
9650	0.924	0.971	0.934	0.693	-0.012	$+0.037$	10.0
10088	0.600	0.775	0.767	0.738	0.734	$+0.008$	18.0
10131	0.561	0.883	0.838	0.722	0.778	$+0.046$	17.5
10195	0.728	0.828	0.487	0.767	-0.008	$+0.341$	18.2
11372	0.687	0.802	0.905	0.771	0.069	-0.103	10.5
11621	0.593	0.331	0.916	0.277	0.947	-0.585	11.0
14135	0.767	0.673	0.804	0.353	0.058	-0.131	9.2

A.8 Per-latent results

A.9 All latents

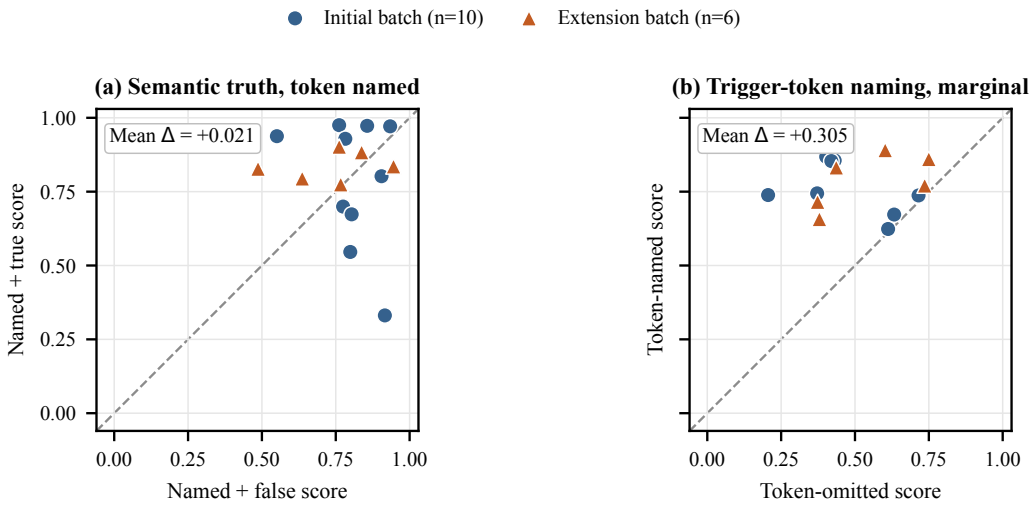


Figure 2: All 16 latents, one point each. **(a)** Token-named with true semantics against token-named with false semantics. Points on the diagonal are latents where replacing the gloss does not move the score; 10 of 16 lie above it, and the spread in both directions is the heterogeneity discussed in Section 3.4. **(b)** Mean token-named score against mean token-omitted score. All 16 points lie above the diagonal, the closest being latent 11621 at a gap of 0.012. Marker size gives the mean word count of a latent's four constructed descriptions.