Do Sparse Autoencoders Reveal Faithful Concepts? A Neuron-Level Cross-Validation with QA-Based Probes

Anonymous ACL submission

Abstract

Sparse Autoencoders (SAEs) have recently been adopted to interpret large language models by decomposing hidden activations into sparse "concept features." These features are often assumed to provide faithful semantic explanations of individual neurons. However, despite their growing use, it remains unclear whether SAE-discovered concepts genuinely reflect neuron semantics or instead arise from reconstruction-induced artifacts. We propose a framework for cross-validating SAE-based neuron interpretations using an independent semantic probe. Specifically, we leverage question–answering embeddings (QA-Emb), in which each feature corresponds to an explicit yes/no semantic query posed to a language model, as an external reference signal. For each neuron, we derive parallel interpretations from SAE features and QA probes, and quantify their semantic agreement using embedding-based similarity and sign-consistency criteria. Across experiments on large-scale news data and multiple auxiliary domains, we find that while many neurons exhibit consistent interpretations, a substantial fraction shows partial or strong disagreement. Further analysis reveals recurring failure modes in which SAE features appear coherent but still lack support from independent QA evidence. Our results demonstrate that independent semantic cross-validation is essential for assessing the trustworthiness of reconstruction-based neuron interpretations in large language models.

1 Introduction

Understanding the internal representations of large language models (LLMs) has become a central challenge in natural language processing, both for scientific insight and for practical concerns such as robustness, controllability, and safety. A growing body of work aims to interpret individual neurons or latent dimensions in LLMs by associating them with human-interpretable concepts, often through analysis of activation patterns, top-ranked tokens, or synthetic stimuli (Kowalska and Kwaśnicka, 2025; Lin et al., 2025). Recently, Sparse Autoencoders (SAEs) have gained prominence as a promising approach to neuron-level interpretability. By learning a sparse, often overcomplete decomposition of hidden activations, SAEs aim to extract a set of "concept features" that can be interpreted independently and composed to explain neuron behavior (Shu et al., 2025). Qualitative analyses of SAE features frequently reveal coherent token sets or natural-language descriptions, leading to the widespread assumption that these features correspond to meaningful semantic units (Cunningham et al., 2023). However, interpretability does not necessarily imply faithfulness. SAE features are learned under reconstruction and sparsity objectives that do not explicitly enforce alignment with neuron semantics (Heap et al., 2025). As a result, it remains unclear whether the concepts identified by SAEs genuinely reflect the functional role of a neuron, or whether they capture abstractions introduced by the autoencoding process itself (Cho et al., 2025). Crucially, most existing evaluations of SAE interpretability are self-referential, relying on the same features or reconstruction statistics used during training, with limited independent validation (Lin and Bloom, 2023; Paulo et al., 2025).

In this work, we ask: *Can we trust the semantic "concept features" discovered by Sparse Autoencoders as faithful interpretations of individual neurons in large language models?* To address this question, we propose a cross-validation framework that evaluates SAE-based neuron interpretations against an independent semantic probe. Specifically, we leverage question–answering embeddings (QA-Emb), a question-driven semantic representation in which each dimension corresponds to the model's response to an explicit yes/no semantic question. While QA-Emb was originally introduced in a different context, we repurpose it

here purely as an external, human-interpretable semantic probe that is independent of SAE training objectives. A detailed description of QA-Emb is provided in Appendix A. For each neuron, we derive two parallel interpretations: one from SAE-discovered concept features, and one from QA-based semantic probes. We then quantify their agreement using embedding similarity measures and sign-consistency criteria, and categorize neurons into agreement, partial agreement, and disagreement regimes. This setup allows us to move beyond anecdotal interpretability examples and systematically analyze when SAE interpretations align with an independent semantic signal and when they fail. Our experiments reveal that while SAE and QA interpretations often agree, a substantial fraction of neurons exhibit systematic disagreement, even when SAE features appear highly coherent in isolation. Through detailed case studies and controlled faithfulness checks, we identify recurring failure modes in which SAE features assign misleading semantic labels that are not supported by QA evidence. We further analyze how agreement patterns vary across layers and datasets, highlighting the role of inductive biases introduced by reconstruction-based interpretability. Overall, our work makes three contributions:

- We propose a cross-modal validation framework that evaluates neuron interpretations by contrasting reconstruction-based explanations with independent semantic probes.
- We conduct a systematic study of agreement and disagreement between reconstruction-based and probe-based neuron interpretation methods in large language models.
- We identify systematic failure modes where semantically plausible interpretations diverge across methods, and show that faithfulness-based analysis can distinguish interpretations that are more explainable of neuron behavior.

2 Related Work

Neuron and Feature Interpretability in Large Language Models Interpreting individual neurons and latent features in neural networks has long been a central goal in representation learning and model analysis. Early work focused on visualizing neuron activations through maximally activating inputs, top-ranked tokens, or manually curated examples, aiming to associate neurons with human-interpretable concepts (Feldhus and Kopf, 2025; Choi et al., 2024). In the context of large language models, neuron-level interpretability has been explored through activation analysis, probing classifiers, and dataset-based inspection methods. These approaches often reveal that individual neurons respond to recurring lexical, syntactic, or semantic patterns, but also highlight the prevalence of polysemantic neurons whose activations cannot be easily explained by a single concept (Bricken et al., 2023; Bills et al., 2023). While these methods provide valuable qualitative insights, they typically rely on heuristic criteria or manual inspection, making it difficult to systematically evaluate the faithfulness of the resulting interpretations (Bills et al., 2023; Choi et al., 2024).

Sparse Autoencoders for Concept Discovery Sparse Autoencoders (SAEs) have recently emerged as a prominent approach for interpreting high-dimensional representations in neural networks. A central motivation for SAE-based interpretability is the widely observed polysemanticity of individual neurons, particularly in large language models, where a single neuron may respond to multiple unrelated concepts (Bricken et al., 2023; Bills et al., 2023). SAEs aim to address this issue by decomposing neuron activations into a set of sparse, often overcomplete latent features that are hypothesized to be more nearly monosemantic and thus easier to interpret in isolation (Elhage et al., 2022; Kim et al., 2020). Applied to large language models, SAEs have been used to extract latent "concept features" from hidden layers, with empirical analyses showing that individual features often correspond to coherent sets of tokens or concise natural-language descriptions (Bricken et al., 2023). This has led to the growing practice of treating SAE features as semantically meaningful units and using them as the basis for neuron- and feature-level interpretation. However, SAE-based representations are learned solely through a reconstruction-driven objective under sparsity constraints, without direct supervision or guarantees of semantic faithfulness (Le et al., 2012; Makhzani and Frey, 2014; Song et al., 2025). As a result, while SAE features are often assumed to correspond to approximately monosemantic concepts, the extent to which they faithfully capture the functional semantics of individual neurons remains an open question. Existing evaluations primarily focus on reconstruction quality or internal feature coherence, and rarely

incorporate independent semantic validation.

Question-Based and Probe-Based Interpretability An alternative line of work studies interpretability through explicit probes, in which predefined semantic properties are tested against model representations. Probing classifiers and diagnostic tasks have been widely used to assess whether specific linguistic or semantic information is encoded in model activations (Alain and Bengio, 2018; Belinkov, 2021; Conneau et al., 2018). More recently, question-based approaches have been proposed to construct interpretable feature spaces by asking language models explicit yes/no questions and using their responses as semantic features (Benara et al., 2024). These methods offer transparent and human-readable representations, as each feature corresponds directly to a semantic query. While question-based probes provide explicit semantic grounding, they are not designed to reconstruct hidden representations or explain individual neurons in isolation. Instead, they serve as an external semantic signal with inductive biases distinct from reconstruction-based methods.

Positioning of the Present Work Prior work has largely studied reconstruction-based interpretability methods, such as sparse autoencoders, and probe-based semantic representations in isolation. In particular, SAE features are often implicitly treated as faithful, approximately monosemantic decompositions of polysemantic neurons, despite being optimized without explicit semantic supervision. In contrast, our work focuses on systematically cross-validating this assumption by comparing SAE-based neuron interpretations with independent, question-based semantic probes. By analyzing when these two signals agree or diverge, we aim to assess the reliability of SAE-based interpretations and to identify recurring failure modes in neuron-level semantic explanations.

3 Method

We propose a neuron interpretation framework that cross-validates activation-grounded semantic features with independent question-based semantic probes. At a high level, our method derives two complementary interpretations for each neuron—one based on sparse autoencoder (SAE) features and one based on QA-based semantic probes—and evaluates their semantic alignment to assess interpretability and faithfulness. Figure 1 provides an overview of the full pipeline. Given a collection of input text spans, neuron activations are processed along two parallel pathways. The upper pathway extracts SAE features that are directly grounded in model activations, while the lower pathway derives QA-based semantic hypotheses by correlating neuron activations with predefined semantic questions. These two interpretations are then aligned in a shared semantic space, cross-filtered to identify consistently supported concepts, and optionally synthesized into a refined neuron-level interpretation using a constrained LLM.

3.1 SAE-Based Neuron Interpretation

Let $h \in \mathbb{R}^d$ denote the hidden activation vector at a fixed layer of a language model. A sparse autoencoder (SAE) maps h to a sparse latent representation $z \in \mathbb{R}^{d_{\text{sae}}}$, from which h can be approximately reconstructed. For a given neuron n, we identify the SAE latent features whose decoder weights have the strongest influence on that neuron. These latent features are treated as candidate semantic explanations for the neuron's behavior. Each SAE feature is associated with a set of top-activating tokens or text spans, and may additionally include a short natural-language description derived from these activations. Aggregating the most influential SAE features yields an activation-grounded, data-driven semantic interpretation for each neuron.

3.2 QA-Based Neuron Interpretation

In parallel, we derive an independent semantic interpretation using QA-based semantic probes. Intuitively, QA-based interpretation treats neuron activity as being explained by explicit semantic questions, such as whether an input expresses a topic, attribute, or relation (Benara et al., 2024). More details about this method can be found in Appendix B. Each input span is associated with a QA embedding, where each dimension corresponds to a predefined binary semantic question. These questions are answered by a language model, producing a sparse vector of yes/no responses that serves as an interpretable semantic representation (see Figure 1, top branch). For a fixed neuron, we measure the statistical association between the neuron's activation values and each QA dimension across inputs. The QA-based interpretation of a neuron is defined by the subset of questions whose responses exhibit the strongest positive or negative association with that neuron's activations. These questions serve as explicit semantic hypotheses about the neuron's

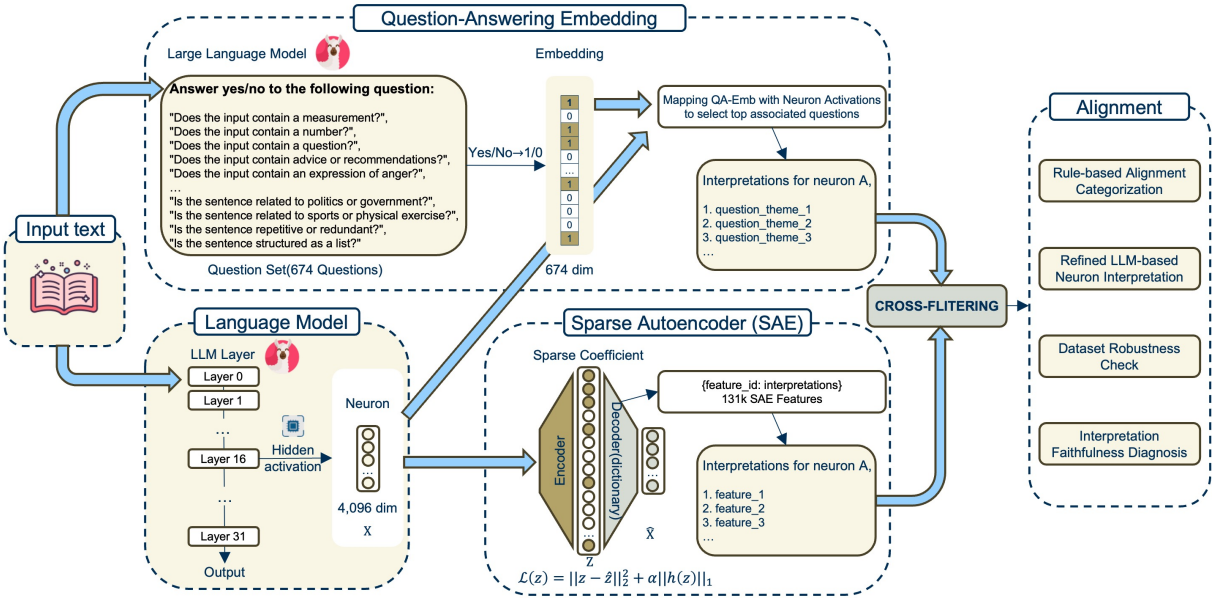


Figure 1: Overview of the proposed SAE–QA cross-validation framework. For each neuron, semantic interpretations are derived through two independent pathways: sparse autoencoder (SAE) features grounded in model activations, and QA-based semantic probes that provide external semantic hypotheses. Their semantic alignment is evaluated to assess agreement, disagreement, and faithfulness of neuron-level interpretations. Optionally, cross-filtered evidence from both pathways can be synthesized into a refined interpretation using a constrained language model.

functional role. Unlike SAE features, QA probes are predefined and independent of neuron activations, providing an external semantic signal for interpretation.

3.3 Semantic Alignment Between SAE and QA Interpretations

To assess whether SAE-based and QA-based interpretations describe a common underlying semantic concept, we compare their semantic content in a shared embedding space. Specifically, we compute semantic similarity between textual representations derived from SAE features (e.g., feature descriptions or token sets) and QA question texts. In addition to semantic similarity, we consider the directionality of neuron activations. This allows us to distinguish cases where the semantic content aligns with activation behavior from cases where similar semantics correspond to opposite activation directions. Based on these criteria, neurons can exhibit varying degrees of agreement or disagreement between SAE- and QA-based interpretations.

3.4 Cross-Filtering and Interpretation

While SAE and QA interpretations provide complementary semantic evidence, each alone may contain spurious or weakly grounded concepts. We therefore apply a cross-filtering step that retains only semantic elements that are consistently supported by both methods. This step identifies a shared semantic subspace between activation-grounded SAE features and externally defined QA probes. To synthesize a unified neuron-level interpretation, we employ a constrained LLM-based aggregation step. The language model is provided with: (i) cross-filtered SAE token evidence directly grounded in neuron activations, (ii) cross-filtered QA questions serving as independent semantic confirmation, and (iii) SAE feature descriptions capturing higher-level abstractions. The LLM is explicitly constrained to produce a concise semantic summary that is strictly supported by the provided evidence and may not introduce new semantic content. In this framework, the LLM functions as a semantic aggregator rather than a free-form generator.

3.5 Evaluation Overview

To evaluate the quality of refined neuron interpretations, we introduce the **Qualitative Interpretation Score (QIS)**, a composite metric designed to assess semantic coherence, cross-method agreement, grounding in evidence, and semantic focus. QIS is used exclusively for evaluation and does not influence the construction of neuron interpretations. Full definitions and implementation details are provided in the experimental section and Appendix F.

4 Experiments

This section describes the experimental setup used to instantiate and evaluate the proposed SAE–QA cross-validation framework, including model selection, datasets, implementation details, and evaluation protocols. Code will be released upon acceptance.

4.1 Model and Layers

All experiments are conducted on `meta-llama/LLaMA-3.1-8B`, which has a hidden dimensionality of $d_{\text{in}} = 4096$ (AI@Meta, 2024). Unless otherwise specified, our primary analysis focuses on a middle layer (layer 16), which prior work has suggested to contain rich semantic information (Skean et al., 2024, 2025). To assess layer-wise variation and faithfulness, we additionally analyze an early layer (layer 4) and a late layer (layer 28) in selected experiments.

4.2 Datasets

Our primary dataset consists of 9,800 text segments sampled from CC-News, a large-scale news corpus (Mackenzie et al., 2020). To evaluate robustness across domains and languages, we additionally include auxiliary datasets covering financial text, scientific abstracts, programming-related text, and non-English summaries. Dataset details are provided in Appendix G. For each dataset, we extract fixed-length token snippets of 64 tokens and segment each snippet into overlapping spans of 16 tokens using a sliding window, resulting in 49 spans per snippet. Each span is treated as an independent input for neuron activation analysis.

4.3 Representations

Hidden State Extraction For each input span, we extract hidden activations from the selected layer(s) using SAELens (Bloom et al., 2024). This yields an activation matrix of size $[N, d_{\text{in}}]$, where N denotes the number of input spans.

Sparse Autoencoder Configuration We use pretrained Sparse Autoencoders (SAEs) from *Llama Scope* to analyze hidden activations of the selected model and layer (He et al., 2024). Each SAE maps residual stream activations to a higher-dimensional sparse latent space. For each neuron, we select the top-ranked SAE features based on the absolute decoder weight magnitude connecting latent features to that neuron.

QA-Based Semantic Probes Each input span is associated with a QA embedding derived from a fixed set of binary yes/no questions (Benara et al., 2024). For each neuron, we compute Pearson correlation coefficients between its activation values and each QA dimension across inputs. QA dimensions with the strongest positive and negative correlations are selected to form the neuron's QA-based interpretation.

4.4 Implementation Details

To compare SAE- and QA-based interpretations, we embed textual representations using `all-mpnet-base-v2`, a sentence embedding model (Reimers and Gurevych, 2019) and compute cosine similarity in the resulting semantic space. All similarity thresholds, correlation settings, and selection hyperparameters are fixed across experiments. Exact values are reported below.

4.5 Evaluation Protocol

Agreement Categories Neurons are assigned to one of four categories based on semantic similarity and sign consistency between SAE and QA interpretations: *agreement*, *partial agreement*, *weak agreement*, or *disagreement*. Neurons with cosine similarity ≥ 0.45 are labeled as *agreement*, those with similarity in $[0.30, 0.45)$ as *partial agreement*, and those with similarity < 0.30 as *weak agreement*. Cases with high semantic similarity but opposite activation signs are labeled as *disagreement*. These thresholds are selected empirically based on preliminary analysis, balancing semantic coherence and coverage across neurons. We observe that the overall agreement trends and qualitative conclusions are robust to moderate variations of these thresholds. Additional details on threshold selection are provided in Appendix C.

Qualitative Interpretation Score (QIS) We evaluate refined neuron interpretations using QIS, which measures: (i) internal semantic consistency, (ii) semantic overlap between SAE and QA evidence, (iii) grounding of the refined summary in the provided evidence, and (iv) semantic focus. QIS is computed as a fixed weighted combination of these components and is used solely for evaluation.

Visualization We visualize SAE–QA alignment using similarity matrices, token-level UMAP overlap plots, and neuron-wise agreement distributions.

5 Results and Analysis

5.1 Overall SAE–QA Alignment

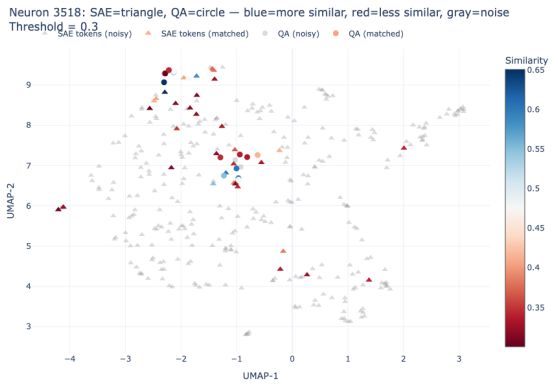
We begin by measuring semantic agreement between SAE-based and QA-based interpretations on the main dataset at the primary analysis layer (Layer 16). Each neuron is assigned to one of four agreement categories based on semantic similarity and sign consistency, as defined in Section 4.5. Table 1 summarizes the distribution of neurons across agreement categories in this primary setting. A

Agreement Category	Proportion of Neurons (%)
Agreement	29.3
Partial Agreement	60.8
Weak Agreement	9.9
Disagreement	< 0.1

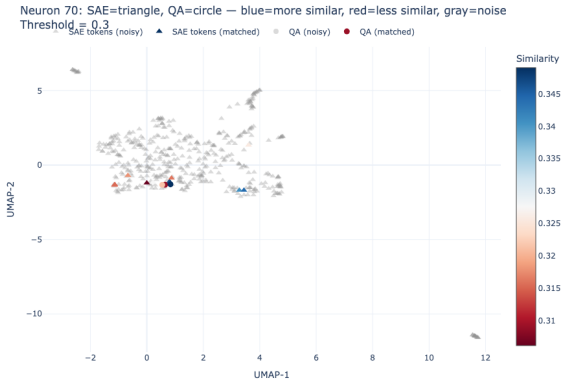
Table 1: Distribution of neurons across agreement categories for the main dataset at the middle layer.

substantial majority of neurons fall into the agreement or partial agreement regimes, indicating that SAE features often capture semantic information consistent with independent QA probes. A smaller but non-negligible fraction of neurons exhibit weak agreement, while explicit disagreement cases are exceedingly rare. Interestingly, in the few disagreement cases we observe, the associated SAE features often appear semantically coherent when viewed on their own, for example by inspecting their top activating tokens. However, these apparent interpretations are not supported by the neuron's activation patterns as measured by QA probes. This suggests that some SAE features can be intuitively interpretable at the surface level, which is patterns tied to the literal form of the input rather than a single underlying semantic function—into a single abstraction, yet fail to faithfully reflect the semantic factors that actually drive neuron behavior.

Qualitative Alignment Patterns To better understand what semantic agreement entails, we visualize the alignment between SAE-derived semantic representations and QA question texts using a cross-filtered UMAP projection computed per neuron. In each view, points below the similarity threshold are rendered in gray, while above-threshold points are highlighted to reveal whether SAE- and QA-derived semantics co-localize. Figure 2 shows two representative neurons: one with strong co-localization (agreement) and one with limited overlap (weak alignment). An interactive version of these visualizations (clickable points with linked SAE/QA matches) is provided in the supplementary material. Neurons in the agreement regime



(a) **Well-aligned neuron.** SAE-highlighted tokens and QA questions overlap strongly in the high-similarity region.



(b) **Poorly-aligned neuron.** Limited overlap between SAE and QA; highlighted regions are largely method-specific.

Figure 2: Cross-filtered UMAP views for two representative neurons on the main dataset (Layer 16). Points below the similarity threshold are gray; above-threshold points are highlighted, allowing visual inspection of whether SAE and QA derived semantics co-localize.

typically exhibit consistent semantic themes across activation-grounded SAE tokens, SAE feature descriptions, and QA questions. These cases suggest that reconstruction-based and question-based interpretability can converge on a shared semantic explanation when neuron behavior is well captured by both signals. To better understand why such convergence fails for other neurons, we examine cases of weak alignment and disagreement in more detail. We find that weakly aligned neurons often display method-specific semantic clusters: SAE features may form coherent and interpretable descriptions when examined in isolation, yet these descriptions are not supported by the neuron's activation patterns as measured by QA probes. This observation indicates that a coherent SAE feature

description does not necessarily imply behavioral faithfulness, and motivates a systematic analysis of failure modes enabled by our cross-validation framework.

Failure Modes of SAE-Based Interpretability A central contribution of our framework is its ability to expose systematic failure modes of SAE-based neuron interpretation. Figure 3 provides an annotated example that illustrates how such failures manifest in practice. The highlighted neuron (Neuron 70) exhibits a semantically rich and internally coherent SAE feature description, yet shows minimal alignment with QA-based evidence or activation behavior. Closer inspection reveals that the SAE interpretation aggregates heterogeneous surface-level regularities. These regularities include punctuation usage, numerical formatting patterns, and loosely related topical cues that frequently co-occur in the data but do not correspond to a unified semantic role. In contrast, QA probes consistently indicate a narrower functional role related to technological problem-solving and mathematical reasoning. This discrepancy suggests that reconstruction objectives can induce abstractions that appear meaningful but are not faithful to the neuron's causal contribution. An interactive version of this example, allowing inspection of individual points and their corresponding SAE and QA matches, is provided in the supplementary material. Additional failure cases and detailed token-level analyses are provided in the appendix I.

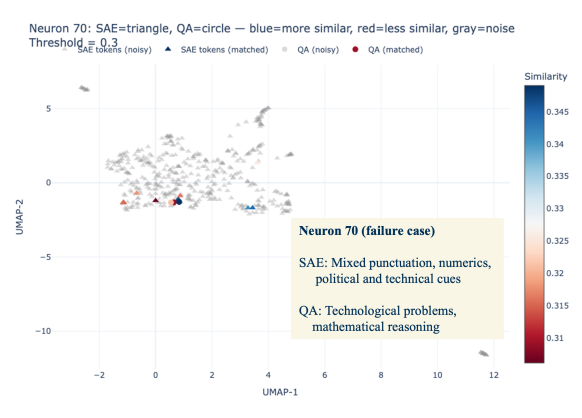


Figure 3: Annotated failure case illustrating a representative SAE–QA disagreement. The circled point corresponds to Neuron 70, previously shown in the interactive visualization (Figure 2(b)). Despite a semantically rich SAE feature description, the SAE-highlighted region shows little overlap with QA-derived semantics, which instead indicate a narrower functional role centered on technological and mathematical reasoning.

Constrained LLM-Based Neuron Interpretation To synthesize a unified neuron-level explanation from cross-validated evidence, we apply a constrained LLM-based semantic aggregation step. The LLM integrates cross-filtered SAE tokens, QA questions, and SAE feature descriptions into a concise interpretation supported by activation-grounded and independently validated evidence. Figure 4 shows the distribution of QIS values for neurons ranked by SAE–QA semantic similarity. Neurons in the top similarity group exhibit substantially higher mean QIS values (approximately ~ 0.52) compared to neurons in the bottom group (approximately ~ 0.32). This result indicates that neurons with stronger SAE–QA alignment tend to admit clearer and more coherent interpretations after refinement. Importantly, we also observe a non-trivial overlap in QIS distributions between the two groups. We find that LLM-aggregated interpretations are consistently more specific and faithful than interpretations derived from either SAE or QA alone. An illustrative example is provided in Appendix E. In agreement cases, aggregation reinforces shared semantics. In disagreement cases, aggregation resolves ambiguities by prioritizing activation-grounded and cross-validated evidence, often discarding misleading abstractions introduced by SAE features.

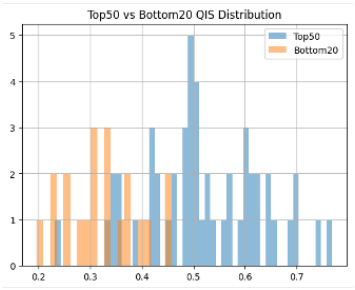


Figure 4: Distribution of Qualitative Interpretation Score (QIS) values for neurons grouped by SAE–QA similarity ranking. Higher QIS values indicate more coherent and well-grounded neuron interpretations.

5.2 Dataset Robustness Analysis

Finally, we assess whether the observed agreement patterns generalize beyond the primary dataset. We repeat the SAE–QA cross-validation procedure on several auxiliary datasets spanning different domains and languages. Across datasets, we observe qualitatively similar distributions of agreement and

disagreement, which means over 90% of neurons in each dataset show at least one aligned SAE–QA concept, indicating that the SAE–QA alignment is not dependent on dataset-specific lexical statistics, but instead captures higher-level neuron behaviors that generalize across domains. We further analyze the overlap of aligned neurons across datasets to assess whether the same neurons tend to express similar semantic concepts in different corpora. When comparing each auxiliary dataset against CCNews, we observe that approximately 4,000 neurons (out of the full layer population) exhibit overlapping aligned SAE–QA concepts across datasets. This indicates that the overlap is not driven by a single dominant dataset, but reflects a broadly shared set of neuron-level semantic units that persist across domains. Taken together, these results provides strong evidence that the proposed interpretation pipeline captures intrinsic neuron semantics rather than dataset-specific patterns.

6 Faithfulness-Oriented Diagnostics

To assess faithfulness more directly, we design a controlled diagnostic that evaluates whether competing interpretations better predict neuron activation behavior under targeted semantic interventions. Rather than assuming either SAE- or QA-based interpretations to be correct, we test each interpretation as a falsifiable hypothesis.

Diagnostic Setup For neurons exhibit failure modes, we prompt a language model to propose candidate semantic hypotheses corresponding to the SAE and QA interpretations, respectively. These hypotheses serve as competing explanations for the neuron's activation behavior. For each hypothesis pair, we construct controlled text subsets that isolate the corresponding semantic dimensions. In particular, we generate 2×2 contrastive text conditions that independently vary the two hypothesized factors (e.g., language form and religious content), yielding four conditions that differ along exactly one dimension at a time. Neuron activation is measured using a simple detector defined as the maximum activation over tokens in the input text. We evaluate each hypothesis by computing the area under the ROC curve (AUC), where activation scores are used to distinguish positive and negative contrast conditions. An AUC of 1 indicates that the hypothesis fully explains the neuron's behavior, an AUC of 0.5 indicates no predictive power, and an AUC of 0 indicates an incorrect explanation.

Layer-wise Faithfulness Outcomes Applying this diagnostic to neurons with SAE–QA disagreement reveals clear layer-dependent patterns. Among middle-layer neurons, a clear majority ($\sim$60%) are more faithfully explained by QA-based hypotheses under controlled diagnostics. In contrast, late-layer neurons more frequently favor SAE-based interpretations, with approximately two-thirds of neurons exhibiting higher AUC scores for SAE-based hypotheses. Early-layer neurons often yield ambiguous outcomes, with nearly half of neurons showing no clear predictive advantage for either interpretation. Representative neuron-level case studies illustrating each outcome are provided in Appendix J.

Implications While individual faithfulness diagnostics yield mixed outcomes, a consistent pattern emerges when results are grouped by layer. For neurons in the middle layer, failure cases identified via SAE–QA disagreement are more often resolved in favor of QA-based interpretations under controlled diagnostics. In contrast, early-layer neurons frequently yield ambiguous results, with neither method demonstrating clear predictive superiority. For late-layer neurons, SAE-based interpretations more often align with activation behavior under contrastive testing. These observations suggest that faithfulness is not a property of an interpretation method in isolation, but rather a function of its interaction with layer-specific representational characteristics.

7 Conclusion

Overall, our results demonstrate the importance of independent semantic cross-validation for neuron-level interpretability. By comparing reconstruction-based interpretations with external semantic probes, we show that apparent interpretability does not necessarily imply faithfulness, and that systematic failure modes can be revealed through cross-modal discrepancies. We further introduce a constrained LLM-based aggregation procedure and a quantitative metric, QIS, to assess the coherence, grounding, and focus of neuron interpretations. Taken together, our findings suggest that interpretability quality depends on more than raw semantic similarity, highlighting the need for principled diagnostic tools when evaluating neuron explanations in large language models.

Limitations

Overall, our results demonstrate the importance of independent semantic cross-validation for neuron-level interpretability. By comparing reconstruction-based interpretations with external semantic probes, we show that apparent interpretability does not necessarily imply faithfulness, and that systematic failure modes can be revealed through cross-modal discrepancies. We further introduce a constrained LLM-based semantic aggregation procedure and a quantitative metric, QIS, to evaluate the coherence, grounding, and focus of neuron interpretations. Together, these findings suggest that interpretability quality depends on more than raw semantic similarity, and that neuron explanations can be assessed in a principled and quantitative manner. Rather than identifying a universally superior method, our results highlight the need for diagnostic tools that account for interactions between interpretation methods and model representations, especially as interpretability claims are applied to larger and higher-stakes models.

References

AI@Meta. 2024. Llama 3 model card.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Vinamra Benara, Chandan Singh, John X. Morris, Richard Antonello, Ion Stoica, Alexander G. Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings by asking llms questions. *Preprint*, arXiv:2405.16714.

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

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. 2024. Saelens. https://github.com/decoderesearch/SAELens.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Seonglae Cho, Harryn Oh, Donghyun Lee, Luis Eduardo Rodrigues Vieira, Andrew Bermingham, and Ziad El Sayed. 2025. Faithfulsae: Towards capturing faithful features with sparse autoencoders without external dataset dependencies. *Preprint*, arXiv:2506.17673.

Dami Choi, Vincent Huang, Kevin Meng, Daniel D Johnson, Jacob Steinhardt, and Sarah Schwettmann. 2024. Scaling automatic neuron description. https://transluce.org/neuron-descriptions.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *Preprint*, arXiv:1805.01070.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy$_m odel/index.html$.

Nils Feldhus and Laura Kopf. 2025. Interpreting language models through concept descriptions: A survey. *Preprint*, arXiv:2510.01048.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *Preprint*, arXiv:2410.20526.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. 2025. Sparse autoencoders can interpret randomly initialized transformers. *Preprint*, arXiv:2501.17727.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*.

Edward Kim, Connor Onweller, Andrew O'Brien, and Kathleen McCoy. 2020. The interpretable dictionary in sparse coding. *Preprint*, arXiv:2011.11805.

Bianka Kowalska and Halina Kwaśnicka. 2025. Unboxing the black box: Mechanistic interpretability for algorithmic understanding of neural networks. *Preprint*, arXiv:2511.19265.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. 2012. Building high-level features using large scale unsupervised learning. *Preprint*, arXiv:1112.6209.

Johnny Lin and Joseph Bloom. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. *Software available from neuronpedia. org*.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, and 2 others. 2025. A survey on mechanistic interpretability for multi-modal foundation models. *Preprint*, arXiv:2502.17516.

Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. Cc-news-en: A large english news corpus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3077–3084, New York, NY, USA. Association for Computing Machinery.

Alireza Makhzani and Brendan Frey. 2014. k-sparse autoencoders. *Preprint*, arXiv:1312.5663.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. 2025. Automatically interpreting millions of features in large language models. *Preprint*, arXiv:2410.13928.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *Preprint*, arXiv:2503.05613.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. 2024. Does representation matter? exploring intermediate layers in large language models. *Preprint*, arXiv:2412.09563.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *Preprint*, arXiv:2502.02013.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. 2025. Position: Mechanistic interpretability should prioritize feature consistency in saes. *Preprint*, arXiv:2505.20254.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

A Background on Question–Answering Embeddings

Question–answering embeddings (QA-Emb) were introduced by Benara et al. (2024) as a framework for constructing explicitly interpretable semantic representations of language model behavior. Rather than learning latent dimensions whose meanings must be inferred post hoc, QA-Emb defines each representational dimension a priori via a human-readable semantic question.

At a high level, QA-Emb shifts the locus of interpretability from representation geometry to question design. Each dimension corresponds to a fixed yes/no query (e.g., whether an input concerns a person, an event, or an affective property), and the embedding value reflects the language model's response to that query. This design makes the semantic meaning of each coordinate transparent by construction, enabling direct inspection without auxiliary probing or interpretation models.

QA-Emb was originally proposed as a general-purpose semantic representation for analyzing language model behavior across tasks. Importantly, it does not assume that the resulting question set is exhaustive, complete, or cognitively grounded. Instead, its primary goal is transparency: each dimension encodes a clearly specified semantic hypothesis that can be individually examined, validated, or rejected.

Because QA-Emb does not rely on reconstruction objectives, sparsity constraints, or representation learning, it embodies inductive biases that differ fundamentally from feature-discovery methods such as Sparse Autoencoders. This conceptual separation makes QA-Emb well suited as an external semantic reference for diagnostic analysis, even when it is not treated as an authoritative or ground-truth semantic decomposition.

In this work, we adopt QA-Emb in this diagnostic spirit. We do not interpret QA-Emb as a model

of human semantics or neuron-level ground truth, but as a transparent, question-driven semantic lens that can be contrasted with reconstruction-based interpretations to expose differences in inductive bias.

B QA-Based Semantic Embeddings

We use question–answering embeddings (QA-Emb) as an independent semantic probe for neuron-level interpretation. Unlike reconstruction-based methods, QA-Emb provides a fixed, externally specified semantic feature space whose dimensions are defined by explicit yes/no questions, and are therefore independent of model training or representation learning objectives.

In our framework, QA-Emb is not used as a task representation or predictive embedding. Instead, it serves as a source of semantic hypotheses against which neuron behavior can be evaluated. For each input span, we obtain a binary-valued QA embedding by prompting a language model to answer a predefined set of semantic questions. Each dimension reflects the model's response to a specific semantic query.

To derive a QA-based interpretation for a neuron, we measure the statistical association between the neuron's activation values and each QA dimension across the dataset. Dimensions exhibiting strong positive or negative association are interpreted as candidate semantic factors influencing the neuron's behavior. This procedure yields a neuron-level semantic profile expressed entirely in question-defined coordinates.

Crucially, QA-Emb is treated as a non-authoritative reference rather than a ground-truth explanation. Agreement between QA-based and SAE-based interpretations suggests convergent evidence for a shared semantic factor, whereas disagreement highlights differences in inductive bias between question-driven and reconstruction-driven interpretability methods. Our analysis focuses on these agreement and divergence patterns rather than privileging either representation a priori.

C Threshold Selection for Semantic Agreement

The cosine similarity thresholds used to define agreement categories are chosen empirically. We evaluated a range of threshold values for high and moderate semantic similarity and inspected their effects on category distributions and qualitative interpretability.

We find that a threshold of 0.45 for strong agreement yields a stable set of neurons with high semantic coherence between SAE- and QA-based interpretations, while avoiding overly restrictive filtering. Lower thresholds increase coverage but admit semantically ambiguous cases, whereas higher thresholds reduce the number of interpretable neurons without substantially improving semantic consistency.

Importantly, we observe that the main agreement trends and qualitative failure modes discussed in the paper remain consistent across a reasonable range of threshold values, indicating that our conclusions are not sensitive to the exact choice of thresholds.

D LLM-Based Semantic Interpretation Prompt

Example Neuron Interpretation Prompt Below we show an example of the prompt used to generate a refined semantic interpretation for a single neuron. The example illustrates how cross-filtered tokens and QA questions are treated as primary evidence, while SAE feature descriptions serve only as secondary context.

```
You are an interpretability assistant
    for language models.
Your task is to summarize the semantic
    role of a neuron based on evidence
from two independent sources.

IMPORTANT:
- Cross-filtered TOKENS and QA QUESTIONS
     are the highest-trust evidence.
- SAE FEATURE DESCRIPTIONS are secondary
     and must not override shared
    evidence.

NEURON_ID: 3518

CROSS-FILTERED QA QUESTIONS:
1. an expression of anger
2. body of water
3. technical or specialized terminology
4. technological problem or bug

CROSS-FILTERED TOKENS:
Engineers, River, country, haven, island
    , islands, ire, land

SAE FEATURE DESCRIPTIONS:
1. specific locations and addresses
    related to events or activities
2. terms related to functional
    programming concepts and operations
3. references to military ranks and
    units
```

```
4. references to historical events or
    items related to mail delivery

Return ONLY a JSON object with a single
    field "summary".
```

LLM Output Here is the output for refined interpretation for this neuron.

```
"summary": "This neuron is associated
    with geographical locations,
particularly bodies of water and land-
    related terms, often appearing in
technical or specialized contexts such
     as engineering or problem-solving
discussions."
```

E Example of Aggregated Interpretation Refinement

To illustrate how LLM-based aggregation improves specificity and faithfulness of neuron interpretations, we present a representative example from our analysis. We consider Neuron 3518(l28), which exhibits partial semantic alignment between SAE-based and QA-based interpretations. After cross-filtering, the primary QA evidence associated with this neuron includes the following semantic questions:

- an expression of anger
- body of water
- technical or specialized terminology
- technological problem or bug

The corresponding cross-filtered SAE evidence consists of the following top activating tokens:

- Engineers
- River
- country
- haven

When considered independently, neither interpretation provides a fully satisfactory explanation of the neuron's functional role. The SAE tokens mix geographical and occupational cues without indicating a clear semantic focus, while the QA questions span both topical (e.g., bodies of water) and functional (e.g., technological problems) dimensions. We apply LLM-based aggregation to jointly summarize the shared and activation-grounded evidence. The resulting refined interpretation is:

> *"This neuron is related to geographical locations, particularly bodies of water, and technical or specialized terminology, possibly referring to engineers or experts working on or near rivers, islands, or countries, potentially addressing technological problems or bugs."*

This aggregated interpretation is more specific than either the SAE-based or QA-based interpretation alone. It retains activation-grounded semantic elements (e.g., geographical references such as *river* and *country*) while prioritizing functional cues supported by QA probes (e.g., technical problem-solving). At the same time, the aggregation discards misleading or weakly supported abstractions that arise when SAE features or QA probes are considered in isolation. This example illustrates how LLM-based aggregation leverages cross-validated evidence to produce neuron interpretations that are both semantically coherent and more faithful to observed activation behavior.

F Qualitative Interpretation Score (QIS)

While semantic similarity between SAE-based and QA-based interpretations provides a coarse measure of cross-modal convergence, it does not capture the quality of the *refined neuron-level interpretation* produced after cross-filtering and LLM-based semantic aggregation. In particular, simple similarity scores do not distinguish between focused, coherent interpretations and vague or fragmented ones.

To address this limitation, we introduce the **Qualitative Interpretation Score (QIS)**, a composite metric designed to quantify how well a refined neuron summary captures the shared semantic structure supported by both SAE and QA evidence. Importantly, QIS does not assume that SAE and QA interpretations must match exactly. Instead, it evaluates whether the refined summary accurately represents the *shared semantic subspace* between the two methods while maintaining internal consistency and semantic focus.

Overview Given a neuron with: (i) a set of SAE-derived semantic descriptions, (ii) a set of QA-derived semantic questions, and (iii) a refined neuron summary produced via constrained LLM-based aggregation, QIS combines four complementary components:

- **Intra-set coherence**: whether semantic evidence within each method (SAE or QA) is internally consistent.
- **Cross-set overlap**: the degree of semantic overlap between SAE and QA evidence.
- **Summary grounding**: how well the refined summary aligns with evidence from both SAE and QA.
- **Semantic focus**: a dispersion-based penalty that captures whether the supporting evidence forms a tight semantic cluster around the summary.

Together, these components reward neuron interpretations that are coherent, cross-validated, and semantically focused, while penalizing interpretations that mix unrelated concepts or rely on overly generic abstractions.

Formal Definition Let S denote the set of SAE semantic descriptions for a neuron, Q denote the set of QA semantic questions, and r denote the refined neuron summary. All elements are embedded into a shared semantic space.

We define the four components as follows:

(a) Intra-set coherence. For each set $X \in \{S, Q\}$, intra-set coherence is computed as the average pairwise cosine similarity within X.

(b) Cross-set overlap. Cross-set overlap is defined as the average cosine similarity between elements of S and Q.

(c) Summary grounding. Summary grounding is defined as the average cosine similarity between r and elements in $S \cup Q$, measuring whether the summary lies near the semantic center of the supporting evidence.

(d) Semantic focus. Semantic focus is defined as a dispersion-based penalty measuring the average distance between evidence elements and the summary r. Higher dispersion corresponds to lower focus.

We define the final quality score as a weighted sum of grounding, overlap, intra-set consistency (SAE and QA), and a dispersion penalty:

$$\text{QIS} = w_g G + w_o O + w_s I_S + w_q I_Q - w_d D,$$

with all weights fixed across experiments.

Weighting Rationale The weights are assigned as: Grounding 0.4, Overlap 0.3, SAE coherence 0.15, QA coherence 0.15, Dispersion 0.1. We assign higher weight to summary grounding and cross-set overlap, as these directly reflect whether the refined interpretation captures the shared semantic structure supported by both methods. Intra-set coherence terms capture internal consistency within each method and are assigned lower weight. The dispersion term serves as a mild penalty to discourage overly diffuse or generic interpretations.

G Dataset Details

Table 2 summarizes all datasets used in our experiments, including their domains, languages, and the text fields from which token snippets are extracted.

For each dataset, we use the primary free-text field as the source for snippet extraction (e.g., `article`, `sentence`, `content`, `func_code_string`, or `text`), depending on the dataset schema. All datasets are processed using the same tokenization and segmentation procedure: fixed-length snippets of 64 tokens are extracted and further segmented into overlapping contiguous spans of 16 tokens with a sliding window, resulting in 49 spans per snippet. No dataset-specific preprocessing or filtering is applied unless otherwise stated.

H Interactive UMAP Visualizations

We include interactive cross-filtered UMAP visualizations as an offline HTML artifact. To view, open `interactive/umap_viewer.html` in a modern browser (Chrome/Firefox). Each point is clickable and reveals the corresponding neuron, along with the top-matching SAE tokens/feature descriptions and QA questions. No external network access is required.

I Details on SAE–QA Disagreement Cases

We present additional neuron-level disagreement cases to illustrate recurring failure modes of SAE-based interpretations. In these cases, SAE features yield semantically rich but internally diffuse descriptions, while QA-based probes reveal a more constrained functional role that is not captured by the SAE interpretation. Such cases highlight the limits of reconstruction-driven feature discovery when neurons respond to heterogeneous surface patterns rather than a coherent semantic concept. For example, neuron 70 exhibits a representative

Dataset	Domain	Lang.	Text Field Used	#Segments	Usage
CC-News (Mackenzie et al., 2020)	News	EN	article	9,800	Main
takala/financial_phrasebank (Malo et al., 2014)	Finance	EN	sentence	4,312	Robustness
MedRAG/pubmed (Xiong et al., 2024)	Scientific	EN	content	4,312	Robustness
code-search-net/code_search_net (Husain et al., 2019)	Programming	EN	func_code_string	4,312	Robustness
hugcyp/LCSTS (Hu et al., 2015)	Non-English Summaries	ZH	text	4,312	Robustness

Table 2: Dataset details and text fields used for snippet extraction.

disagreement pattern in which SAE-derived features appear to capture diverse high-level semantic domains, while QA-based evidence indicates a narrower functional sensitivity.

SAE-Based Interpretation. The top SAE features associated with this neuron are characterized by highly heterogeneous descriptions, including references to political and military events, sports statistics, mathematical expressions, programming syntax, browser functionality, historical events, and customer service language. Correspondingly, the top-activating tokens span a wide range of surface forms, such as punctuation marks, numerical sequences, dates, programming-related symbols (e.g., parentheses, operators, code fragments), and common function words.

While individual SAE feature descriptions are semantically plausible in isolation, their aggregate lacks a coherent unifying concept. Instead, the features collectively emphasize structural properties of text, including heavy punctuation, numerical formatting, code-like syntax, and structured or templated expressions. This results in an interpretation that is semantically rich but diffuse, making it difficult to ascribe a clear functional role to the neuron based on SAE evidence alone.

QA-Based Interpretation. In contrast, the QA-based probes for Neuron 70 show strong positive associations with questions related to technological concepts, technological problems or bugs, and mathematical reasoning. These questions consistently point to technical or formal contexts in which structured symbols, numerical values, and code-like patterns are prevalent.

Disagreement Analysis. The disagreement arises because SAE-based interpretations surface a broad set of domain-level labels (e.g., politics, sports, military history) that are not supported by the QA-based evidence. The QA probes suggest that the neuron's activation is driven less by semantic content in these domains and more by the presence of structurally dense, technical, or formally expressed text. In this case, SAE features appear to conflate multiple surface-correlated patterns into semantically labeled features, obscuring the neuron's underlying functional sensitivity.

Failure Mode Characterization. This example illustrates a recurring failure mode in which SAE features assign semantic labels to neurons that are primarily sensitive to formatting, symbolic structure, or syntactic regularities. Although such features reconstruct activations effectively, they do not faithfully capture the neuron's semantic role. QA-based probes provide an external constraint that exposes this mismatch by highlighting the neuron's consistent association with technical and formal contexts rather than the diverse semantic categories suggested by SAE features.

J Representative Neuron Case Studies

In this appendix, we present representative neuron-level case studies to illustrate the qualitative behavior underlying the layer-wise faithfulness trends reported in the main text. Rather than relying solely on semantic plausibility, each case study evaluates faithfulness by testing whether a proposed interpretation can *predict neuron activation under controlled semantic interventions*. Neurons are selected to exemplify QA-favored, SAE-favored, and ambiguous diagnostic outcomes.

J.1 QA-Favored Middle-Layer Neuron

We first examine a representative neuron from a middle layer (layer 16) for which QA-based hypotheses provide a more faithful explanation than SAE-based interpretations. This neuron is associated with QA concepts related to technical or bug-oriented content. Controlled contrasts demonstrate that QA-derived semantic conditions reliably predict activation behavior, while alternative explanations fail to generalize.

```
=== Neuron 70 @ Layer 16 ===

TECH vs CONTROL:
  AUC         : 0.8007
  Cliff delta : 0.6014
  t-test pval : 4.02e-22
  means       : 0.5235 vs 0.4020
```

```
MATH vs CONTROL:
  AUC         : 0.6661
  Cliff delta : 0.3322
  t-test pval : 3.05e-14
  means       : 0.4680 vs 0.4020

TECH vs MATH:
  AUC         : 0.7204
  Cliff delta : 0.4408
  t-test pval : 1.30e-08

Conclusion:
  Neuron activation is more controllable
       by TECH/BUG-style text,
  supporting the QA-based hypothesis.
```

The consistently high AUC values and large effect sizes indicate that the QA-based interpretation captures a functionally relevant trigger for this neuron. In contrast, SAE-based interpretations for this neuron do not yield comparable predictive power under matched controls, suggesting lower faithfulness despite their semantic richness.

J.2 SAE-Favored Late-Layer Neuron

We next analyze a representative late-layer neuron (layer 28) for which SAE-based interpretations provide a more faithful explanation. Initial QA-based probes suggest sensitivity to religious content; however, controlled conditional tests reveal that this apparent effect does not generalize once language form is accounted for. In contrast, SAE-derived hypotheses related to language form remain predictive across conditions.

```
=== Neuron 1216 @ Layer 28 ===

Language main effect (Non-religious):
  A1B0 vs A0B0 AUC : 0.848

Religion main effect (Non-foreign):
  A0B1 vs A0B0 AUC : 0.889

Religion within FOREIGN language:
  A1B1 vs A1B0 AUC : 0.472

Language within RELIGION:
  A1B1 vs A0B1 AUC : 0.630

Conclusion:
  Apparent religion effects disappear
      under language control,
  while language-form effects persist.
```

These results falsify the QA hypothesis that religion is the primary semantic driver for this neuron. Although religion-related content initially appears predictive, its effect does not generalize across language conditions. In contrast, the SAE-based interpretation—sensitivity to language form or encoding—remains stable under targeted interventions, indicating higher faithfulness.

J.3 Ambiguous Early-Layer Neuron

Finally, we consider an early-layer neuron (layer 4) that exhibits weak and inconsistent predictive power for both SAE- and QA-based interpretations. For this neuron, neither hypothesis achieves robust discrimination, and observed effects are sensitive to experimental conditioning.

```
=== Neuron 162 @ Layer 4 ===

Encoding effect (CY vs EN, Non-financial
    ):
  AUC : 0.540

Finance effect (FIN vs Non-FIN, English)
    :
  AUC : 0.465

Conclusion:
  Mixed or weak effects; neither
      hypothesis generalizes reliably.
```

Both QA-based semantic probes and SAE-based feature descriptions yield only marginal improvements over chance and fail to generalize across subconditions. This suggests that the neuron may encode low-level, entangled, or highly context-dependent features that are not well captured by either semantic abstraction.

Together, these case studies demonstrate that faithfulness depends on whether an interpretation can predict neuron behavior under controlled interventions, not merely on semantic plausibility. QA-based and SAE-based methods each succeed in different regimes, while some neurons admit no clean semantic explanation. These findings motivate a comparative, falsification-oriented approach to neuron interpretability.