From Atoms to Trees: Building a Structured Feature Forest with Hierarchical Sparse Autoencoders

Yifan Luo [* 1] **Yang Zhan** [* 1] **Jiedong Jiang** [2] **Tianyang Liu** [3] **Mingrui Wu** [4] **Zhennan Zhou** [5] **Bin Dong** [6 7]

Abstract

Sparse autoencoders (SAEs) have proven effective for extracting monosemantic features from large language models (LLMs), yet these features are typically identified in isolation. However, broad evidence suggests that LLMs capture the intrinsic structure of natural language, where the phenomenon of "feature splitting" in particular indicates that such structure is hierarchical. To capture this, we propose the **Hierarchical Sparse Autoencoder (HSAE)**, which jointly learns a series of SAEs and the parent-child relationships between their features. HSAE strengthens the alignment between parent and child features through two novel mechanisms: a structural constraint loss and a random feature perturbation mechanism. Extensive experiments across various LLMs and layers demonstrate that HSAE consistently recovers semantically meaningful hierarchies, supported by both qualitative case studies and rigorous quantitative metrics. At the same time, HSAE preserves the reconstruction fidelity and interpretability of standard SAEs across different dictionary sizes. Our work provides a powerful, scalable tool for discovering and analyzing the multi-scale conceptual structures embedded in LLM representations.
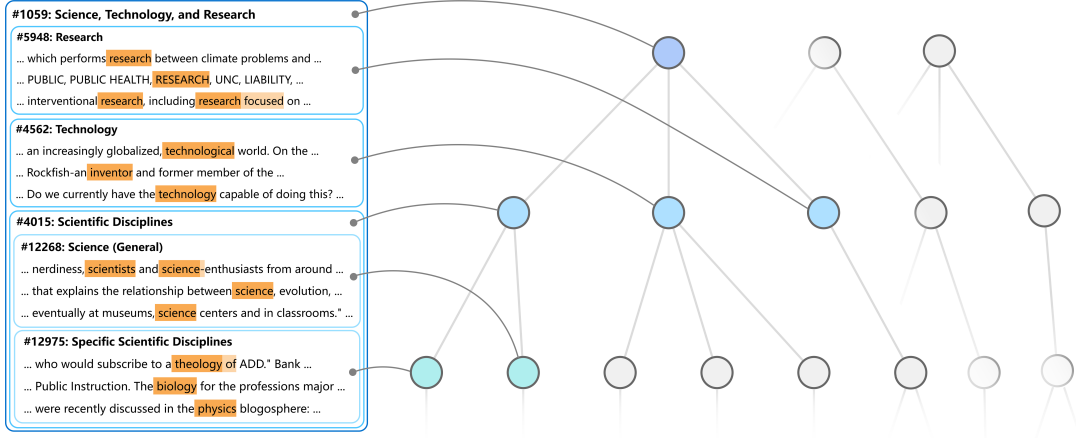
1. Introduction

Understanding the mechanisms by which large language models (LLMs) represent and process information is of great interest in the machine learning community. Recent advances in mechanistic interpretability have shifted focus from analyzing individual neurons (Olah et al., 2020; Bills et al., 2023), which are often polysemantic, toward identifying more coherent units of analysis. Sparse autoencoders (SAEs) have emerged as a particularly successful approach in this direction (Bricken et al., 2023; Cunningham et al., 2023). By learning an overcomplete dictionary that maps dense activations to a sparse latent space, SAEs scalably extract features that are human-interpretable.

Beyond identifying isolated features, emerging evidence suggests that SAE representations are not merely atomic collections but are embedded within a sophisticated internal organization. Recent work has revealed spatial clustering corresponding to semantic groups (Li et al., 2025) and co-activation patterns between abstract features and their specialized counterparts (Clarke et al., 2024), indicating that LLMs capture linguistic relationships through structured internal representations. Since navigating flat, unstructured SAE dictionaries remains inherently difficult, extracting these latent structures is essential for creating organized, multi-scale interpretations. A particularly notable phenomenon hinting at this organization is *feature splitting*, where broad concepts decompose into more granular sub-features as dictionary size increases (Bricken et al., 2023). While prior works have often viewed this as an obstacle to identifying universal, canonical features (Leask et al., 2025; Bussmann et al., 2025), we instead recognize it as direct evidence of the underlying hierarchy represented in the model. By exploiting feature splitting rather than suppressing it, we transform this behavior into a hierarchical indexing structure for the feature space, organizing the "atoms" of representation into a structured forest.

For this purpose, we introduce the **Hierarchical Sparse Autoencoder (HSAE)**. The HSAE simultaneously trains a series of SAE levels with increasing dictionary sizes, with features at each level explicitly assigned to a parent in the preceding coarser level, forming a tree-structured hierarchy. Model parameters and the parent–child assignments co-evolve through alternating optimization. To ensure that hierarchical links reflect semantic subsumption rather than mere structural adjacency, we design optimization objective

---

Preprint. February 13, 2026.

#1059: Science, Technology, and Research

#5948: Research
... which performs research between climate problems and ...
... PUBLIC, PUBLIC HEALTH, RESEARCH, UNC, LIABILITY, ...
... interventional research, including research focused on ...

#4562: Technology
... an increasingly globalized, technological world. On the ...
... Rockfish-an inventor and former member of the ...
... Do we currently have the technology capable of doing this? ...

#4015: Scientific Disciplines

#12268: Science (General)
... nerdiness, scientists and science-enthusiasts from around ...
... that explains the relationship between science, evolution, ...
... eventually at museums, science centers and in classrooms." ...

#12975: Specific Scientific Disciplines
... who would subscribe to a theology of ADD." Bank ...
... Public Instruction. The biology for the professions major ...
... were recently discussed in the physics blogosphere: ...

Figure 1. **Hierarchical feature discovery with HSAE.** We visualize a tree within the learned feature forest alongside a semantic dashboard of its nodes. Each node includes a unique feature index (e.g., #1059), a human-annotated semantic label, and representative top-activating context snippets. Orange highlights indicate the activation positions, with color intensity reflecting the activation magnitude. HSAE captures a clear conceptual taxonomy: a broad parent feature representing *Science, Technology, and Research* (#1059) is systematically decomposed into specialized features such as *Scientific Disciplines* (#4015). This intermediate node further branches into a feature capturing lexical patterns containing the word *science* or *scientists* (#12268) alongside a feature representing specific discipline names such as *biology* or *physics* (#12975). For clarity, only a three-layer subgraph and examples of some features are shown.

to enforce functional dependence across levels via: (1) a **structural constraint loss** encouraging parent features to be reconstructed from their assigned children's outputs, and (2) a **random feature perturbation** mechanism that treats parent and children contributions as interchangeable during reconstruction of original activations. These mechanisms align the learned features into a consistent conceptual taxonomy, where higher-level features are incentivized to act as distilled abstractions of their finer-grained descendants.

In our experiments, we apply HSAE to the residual stream activations of different LLMs and find that it captures a rich, multi-layered hierarchical structure among interpretable features. As illustrated in Figure 1, HSAE naturally organizes discovered features into a conceptual forest, revealing a progressive refinement of semantic meaning. In this structure, high-level parent features initially respond to broad, domain-general patterns, which are then systematically decomposed into increasingly specific features at deeper levels.

To provide a comprehensive evaluation of our results, we first present qualitative case studies, demonstrating that HSAE consistently recovers clear semantic hierarchies across diverse domains. Second, systematic quantitative evaluations confirm that these learned hierarchies are statistically representative of the entire feature space rather than isolated occurrences. Using multiple statistical metrics and an LLM-based automated interpretability assessment, we show that the recovered hierarchical links are both statistically robust and semantically meaningful, significantly outperforming baselines that infer relations post-hoc from independently trained SAEs. Third, we demonstrate that

HSAE preserves core SAE performance, matching or exceeding conventional SAEs on standard benchmarks for feature quality and reconstruction fidelity across all dictionary sizes. Finally, we provide extensive ablation studies to validate the necessity of each proposed mechanism.

Beyond fundamental validation, our analysis reveals several intriguing structural properties of the learned feature space. Visualization shows that input activations triggering sibling features are significantly more related in their low-dimensional projections than those triggering unrelated features, indicating that HSAE captures the underlying geometry of conceptual manifolds. Additionally, we find that features with multiple children often exhibit slightly higher interpretability than those with a single child, providing preliminary evidence that the branching factor may serve as a hint for assessing a feature's semantic clarity.

The main contributions of this work are as follows:

- **A Novel Hierarchical Architecture:** We propose HSAE, the first SAE framework to integrate structural priors directly into training via structural constraint loss and random feature perturbation, enabling the learning of an organized "conceptual forest."

- **Benchmarking Hierarchical Consistency:** We establish a set of benchmarks to measure hierarchy consistency, introducing statistical metrics—such as parent-child co-activation probability—alongside an LLM-based automated interpretability framework. Our results demonstrate that HSAE recovers more robust and coherent structures than post-hoc analysis of independent SAEs.

- **New Insights into LLM Representation Structures:** We provide empirical evidence of how LLMs organize complex concepts into fine-grained hierarchies. Our analysis reveals that these structures possess intrinsic geometric properties, characterized by the spatial clustering of sibling features, and that the structural branching factor serves as a predictor of feature semantic clarity.

2. Related Works

2.1. Discovering Structures in Model Activations

Early analyses of word embedding models identified the well-known parallelogram rule, where semantic analogies are captured through linear vector arithmetic such as $v(\text{"king"}) - v(\text{"man"}) + v(\text{"woman"}) \approx v(\text{"queen"})$ (Mikolov et al., 2013a;b). Similar phenomena have been observed in the activations of LLMs (Gurnee & Tegmark, 2024; Heinzerling & Inui, 2024). These observations support the linear representation hypothesis, which posits that the representation of semantic concepts (features) in the activation space of LLMs corresponds to one-dimensional directions (Park et al., 2023). This hypothesis motivated the development of SAEs, which decompose a model's activations into an overcomplete basis of interpretable linear features (Bricken et al., 2023; Cunningham et al., 2023).

SAE research has revealed various forms of spatial structure within the feature space. Initial work identified *feature splitting*, where broad concepts break into more granular sub-features (Bricken et al., 2023). Further geometric analyses found multi-scale cluster structures in the latent space corresponding to functional domains and semantic groups such as mathematics, code, and natural language (Li et al., 2025). Research has also uncovered irreducible two-dimensional circular manifolds that SAEs use to represent periodic concepts like weekdays (Engels et al., 2025). Moving beyond static spatial organization, recent work explores feature structures through co-activation patterns, revealing hierarchical "hub-and-spoke" topologies in co-occurrence graphs, where abstract features act as central hubs connected to more specialized spokes (Clarke et al., 2024).

2.2. Feature Splitting

Feature splitting (Bricken et al., 2023) is an observed phenomenon in SAEs where an interpretable, broad feature decomposes into multiple specialized ones as the dictionary size increases. For example, a general "punctuation mark" feature may split into distinct latents for periods, commas, and question marks. Related phenomena that can affect feature quality include *feature composition*, where frequently co-occurring independent features merge into a single latent (Anders et al., 2024; Wattenberg & Viégas, 2024), and *feature absorption*, where a general feature de-

velops blind spots for subcases captured by more specialized latents (Chanin et al., 2024). These effects can be explained by the sparsity-driven training objectives of SAEs (Chanin et al., 2024). Metrics have been proposed to quantitatively measure these effects within an SAE (Huang et al., 2024; Chanin et al., 2024; Karvonen et al., 2025).

Several recent works treat feature splitting phenomena as drawbacks (Bussmann et al., 2025; Chanin et al., 2025; Korznikov et al., 2025). For instance, Matryoshka SAEs address these issues by simultaneously training nested dictionaries of increasing size, forcing smaller dictionaries to reconstruct inputs independently. This approach is designed to preserve high-level concepts in smaller dictionaries while allowing larger dictionaries to learn specializations. In contrast to viewing splitting solely as a defect, our work interprets it as evidence of an underlying hierarchical conceptual structure in the model and aims to extract this hierarchy.

2.3. Architectural Evolution of Sparse Autoencoders

The engineering of SAEs has evolved from resolving practical training bottlenecks toward architectural designs that address high-level feature quality issues. Early SAEs used L_1 regularization with ReLU activations (Bricken et al., 2023; Cunningham et al., 2023), but suffered from feature shrinkage. Subsequent variants addressed this issue: Gated SAEs (Rajamanoharan et al., 2024a) decoupled detection from magnitude estimation, while TopK (Gao et al., 2025) and BatchTopK (Bussmann et al., 2024) enforced hard L_0 sparsity constraints via ghost gradients. Training dynamics and reconstruction fidelity were improved through JumpReLU SAEs (Rajamanoharan et al., 2024b), which introduced learnable activation thresholds, and p-annealing techniques (Karvonen et al., 2024) that interpolate between L_1 and L_0 objectives. For large-scale efficiency, Switch SAEs (Mudide et al., 2025) adopted mixture-of-experts routing. Recent structural innovations include Meta-SAEs (Leask et al., 2025), which decompose the composite features of another SAE into more fundamental units, and Matryoshka SAEs (Bussmann et al., 2025), which address feature absorption and splitting through nested dictionaries. Subsequent refinements aim to further reduce feature redundancy through distillation cycles (Martin-Linares & Ling, 2025) and to reduce correlated feature merging in smaller SAEs (Chanin et al., 2025). However, approaches that explicitly learn structures within the feature space remain underexplored.

3. Hierarchical Sparse Autoencoders

In HSAE, we jointly trains a collection of SAEs of increasing dictionary size and explicitly enforces a hierarchical structure among their features.

Technically, a standard SAE seeks to decompose the ac-

tivation vector $x \in \mathbb{R}^d$ into a sparse linear combination of interpretable directions. Let $\{\mathrm{SAE}_\ell(\cdot)\}_{\ell=1}^L$ denote a sequence of L SAEs, where

$$\mathrm{SAE}_\ell(x) = \sum_{i=1}^{n_\ell} d_{\ell,i}\, \sigma(e_{\ell,i}^T x).$$

Later in the paper, we also use $f_{\ell,i}(x) = d_{\ell,i}\, \sigma(e_{\ell,i}^T x)$ to denote a single feature function. Here, ℓ indexes the hierarchy level and n_ℓ represent the dictionary size of level ℓ. Each SAE_ℓ is a standard two-layer JumpRELU SAE (Rajamanoharan et al., 2024b). For simplicity, we omit the threshold parameter in σ, which is also different in each feature (ℓ, i). Note that the hierarchy index ℓ does *not* correspond to depth in a multilayer neural networks.

HSAE induces a partial tree structure over features by defining parent–child relationships between selected features in adjacent levels. For each feature (ℓ, i) where $1 \leqslant \ell \leqslant L-1$, we define a set of children indices

$$\mathcal{C}_{(\ell,i)} \subseteq \{1, \ldots, n_{\ell+1}\}.$$

We refer to (ℓ, i) as a parent feature and $\{(\ell+1, j) \mid j \in \mathcal{C}_{(\ell,i)}\}$ as its children. To ensure a tree-structured hierarchy, we require that child sets are disjoint across parent features

$$\mathcal{C}_{(\ell,i)} \cap \mathcal{C}_{(\ell,k)} = \varnothing, \quad \forall\, i \neq k.$$

The hierarchical relations introduced above do not by themselves impose constraints among related features. Instead, they define a structural prior that is instantiated through additional loss terms and auxiliary training mechanisms described later in this section.

3.1. Loss Function

Each level's SAE employs a standard loss, which consists of a MSE reconstruction loss and an L_0 sparsity penalty

$$\mathcal{L}_{\mathrm{SAE},\ell}(x) = \| \mathrm{SAE}_\ell(x) - x \|_2^2 + \lambda_\ell \sum_{i=1}^{n_i} 1_{\{\sigma(e_{\ell,i}^T x) > 0\}}.$$

For each parent-children group, we introduce a parent-children constraint loss as

$$\mathcal{L}_{\mathrm{PC},(\ell,i)}(x) = \begin{cases} \| f_{\ell,i}(x) - \sum_{j \in C_{\ell,i}} f_{\ell+1,j}(x) \|_2^2, & C_{\ell,i} \neq \varnothing, \\ 0, & C_{\ell,i} = \varnothing. \end{cases}$$

The final HSAE loss function is

$$\mathcal{L}_{\mathrm{HSAE}} = \mathbb{E}_{x \sim D} \left[\sum_{\ell=1}^L \mathcal{L}_{\mathrm{SAE},\ell}(x) + \rho \left(\sum_{\ell=1}^{L-1} \sum_{i=1}^{n_\ell} \mathcal{L}_{\mathrm{PC},(\ell,i)}(x) \right) \right].$$

where $x \sim D$ is the distribution of activations extracted from the specific site of the target LLM.

3.2. Parent-Children Feature Perturbations

Beyond additional loss terms, we also introduce a structure-aware random perturbation of features to encourage the relationship between parent feature and its children.

When computing the SAE loss $\mathcal{L}_{\mathrm{SAE},\ell}$, we do not always use features solely from level ℓ. Instead, for each feature, we stochastically substitute its contribution with those of its children at the next level.

Formally, given the perturbation rate r, let $z_{\ell,i} \sim \mathrm{Bernoulli}(r)$. We define

$$\tilde{f}_{\ell,i}(x) = z_{\ell,i} f_{\ell,i}(x) + (1 - z_{\ell,i}) \sum_{j \in \mathcal{C}_{\ell,i}} f_{\ell+1,j}(x).$$

and compute the SAE loss $\mathcal{L}_{\mathrm{SAE},\ell}$ with the perturbed version of the original SAE:

$$\widetilde{\mathrm{SAE}}_\ell(x) = \sum_{i=1}^{n_\ell} \tilde{f}_{\ell,i}(x)$$

The random variables $z_{\ell,i}$ are sampled independently across features at each training step. This perturbation encourages consistency between parent and children features by exposing the SAE loss to mixtures of representations across adjacent levels.

In practice, we adopt a parent-child constraint weight of $\rho = 0.01$ and a perturbation rate of $r = 5\%$. These values are selected based on the ablation results in Figure 8 to balance hierarchical alignment with reconstruction fidelity.

3.3. Optimization Strategy

Training HSAE involves the joint optimization of model parameters $\{e_{\ell,i}, d_{\ell,i}, \sigma_{\ell,i}\}$ and the set of hierarchical relations $\{\mathcal{C}_{\ell,i}\}$ simultaneously. To handle this joint objective, we adopt an alternating optimization strategy. Specifically, the training process alternates between the following two stages:

- **Parameter Optimization**: With the hierarchical structure fixed, we optimize the model parameters via gradient descend for a predefined number of steps.

- **Hierarchy Update**: With the model parameters fixed, we update the hierarchical structure $\{\mathcal{C}_{\ell,i}\}$ to reflect the current feature organization.

This iterative process is maintained throughout training, ensuring that the hierarchical structure co-evolves with the learned SAE features. In practice, we perform a round of hierarchy update every 5000 gradient steps.

Parameter Optimization. Most settings for parameter optimization in HSAE is identical to standard JumpReLU

4

SAE training. We use a straight-through estimator of the gradient through the discontinuity of L_0 sparsity penalty. Decoder vectors are normalized after each gradient step to fix their norms. We also adopt ghost gradients to mitigate the dead neuron problem (Jermyn & Templeton, 2024).

The only notable modification we introduce is the adaptive sparsity weight. We dynamically adjust the sparsity weights λ_ℓ during training to guide the average L_0 of each level toward a shared target sparsity. The detailed method is provided in Appendix A.

Similarity-Based Hierarchy Update. The hierarchy update step is guided by a general principle of feature similarity. Each feature at level $\ell + 1$ will select a new parent feature at level ℓ that has the most similarity according to a predefined similarity metric.

We consider several instantiations of this principle. Similarity can be defined using feature co-activation statistics, measured as the proportion of training examples on which two features are jointly active, with statistics tracked via an exponential moving average. Alternatively, similarity can be computed in parameter space using the cosine similarity between encoder or decoder feature vectors. Empirically, HSAE is largely insensitive to the specific choice of similarity metric, with all variants exhibiting comparable behavior in ablation studies (Figure 7). Unless otherwise specified, we use encoder vector similarity throughout this paper.

Partial Tree Structure. Recognizing that not all features naturally conform to a hierarchical taxonomy, we relax the rigid tree constraint by allowing a subset of child features to remain unassigned. Specifically, we implement a quantile-based filtering mechanism where features whose maximum similarity scores fall within the given bottom percentile of all candidates are excluded from the hierarchy. We treat this exclusion level as a tunable hyperparameter, which is set to 20% in our experiments. This partial tree structure provides the necessary flexibility to accommodate independent features—those capturing idiosyncratic or atomic concepts—without forcing them into incoherent parent-child relationships, thereby preserving the structural faithfulness of the discovered hierarchy.

4. Experiments

In this section, we present a comprehensive evaluation of HSAE, combining qualitative observations with rigorous quantitative analysis to demonstrate its ability to capture a significantly richer internal organization than standard SAEs. We begin by providing qualitative evidence of the interpretable hierarchies learned by HSAE, followed by a series of evaluations using statistical metrics and automated interpretability assessments to confirm that these structures

are recovered consistently across the model. To ensure that this hierarchical organization does not compromise core performance, we further benchmark HSAE against established SAE evaluation protocols, measuring its feature quality and reconstruction fidelity relative to standard baselines. Finally, we conduct ablation studies to isolate how explicit constraint and implicit perturbation independently and jointly contribute to the formation of semantically coherent and statistically robust hierarchies.

Experimental Setup. Unless otherwise specified, all models discussed in this section are trained on 100M residual stream activations extracted from the 13th layer of `gemma2-2b` (Riviere et al., 2024) as it processes the mini-PILE dataset (Kaddour, 2023). For both HSAEs and standard SAE baselines, we train four levels of SAEs with dictionary sizes of 2048, 4096, 8192, and 16384. We target a sparsity of $L_0 = 50$ for each level. Additional training details are provided in Appendix A.

To ensure the robustness of our findings, we also conduct extensive experiments across different layers and target sparsity. These additional results are provided in Appendix E. We observe consistent results across all tested settings.

4.1. Empirical Case Studies

To provide intuition on the hierarchical structures captured by HSAE, we first present representative case studies from our learned conceptual forest. As illustrated in Figure 1, HSAE successfully decomposes an abstract root representing *Science, Technology, and Research* into distinct meso-scale clusters representing its constituent pillars. These clusters further branch into specific features that follow distinct lexical patterns, such as occurrences of the word "science" or individual scientific disciplines.

The second example, shown in Figure 2, examines the conceptual organization of *Time*. Here, a coarse-grained root feature captures general temporal references. In the subsequent level, it is partitioned into features representing daily timescales versus longer-term durations (e.g., weeks, months), reflecting a clear organizing principle aligned with human cognition. The daily timescale branch further resolves into specialized sub-domains, segregating immediate deictic markers (e.g., "today") from the occurrences of "day" and sub-daily intervals. These examples suggest that the decompositions learned by HSAE are not merely stochastic. Rather, they consistently align with human intuition and possess inherent interpretability across semantic scales. We have provided more empirical studies in Appendix B

However, it is worth noting that the labels assigned to these features are over-simplifications of their actual activation patterns, and counter-examples may exist in certain contexts. Furthermore, while the hierarchy captures strong semantic
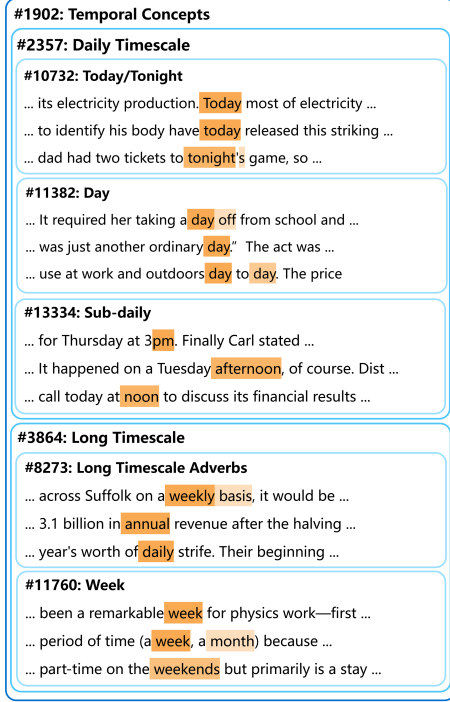
#1902: Temporal Concepts

#2357: Daily Timescale

#10732: Today/Tonight

... its electricity production. Today most of electricity ...

... to identify his body have today released this striking ...

... dad had two tickets to tonight's game, so ...

#11382: Day

... It required her taking a day off from school and ...

... was just another ordinary day," The act was ...

... use at work and outdoors day to day. The price

#13334: Sub-daily

... for Thursday at 3pm. Finally Carl stated ...

... It happened on a Tuesday afternoon, of course. Dist ...

... call today at noon to discuss its financial results ...

#3864: Long Timescale

#8273: Long Timescale Adverbs

... across Suffolk on a weekly basis, it would be ...

... 3.1 billion in annual revenue after the halving ...

... year's worth of daily strife. Their beginning ...

#11760: Week

... been a remarkable week for physics work—first ...

... period of time (a week, a month) because ...

... part-time on the weekends but primarily is a stay ...

Figure 2. **Hierarchical feature discovery of temporal concepts.** The root broad temporal feature (#1902) branches into a *Daily Timescale* (#2357) and a *Longer Timescale* (#3864). At the next level, the daily timescale branch splits into *Today/Tonight* (#10732), the unit concept *Day* (#11382), and *Sub-daily* intervals including morning/afternoon or am/pm (#13334). Similarly, longer-term concepts are partitioned into *Adverbs* (#8273) and specific units like *Week* (#11760). Orange highlights indicate activation positions, with intensity reflecting magnitude.

associations, an input that activates a child feature does not guarantee a corresponding trigger of its parent feature. This phenomenon of hierarchical consistency will be rigorously evaluated in the following sections using the parent-child co-activation probability.

4.2. Statistical Metrics

To quantitatively evaluate the structural faithfulness of the discovered hierarchy, we employ three metrics that measure the alignment between parent and child features.

Hierarchical Consistency via Logical-OR. We first assess whether parent features serve as effective semantic summaries of their children. We construct a logical-OR prediction where a parent is predicted to be active if at least one of its assigned children is active. We then measure the Hamming distance between this child-based prediction and the ground-truth parent activation pattern (Figure 3a). This metric evaluates the logical alignment of the hierarchy; a lower distance indicates that parent activations are a faith-

ful abstraction of lower-level patterns, confirming that the hierarchy captures a coherent decomposition.

Conditional Co-activation Probabilities. Beyond binary logical alignment, we evaluate the statistical dependency between levels through two complementary conditional probabilities, averaged over all assigned parent-child pairs:
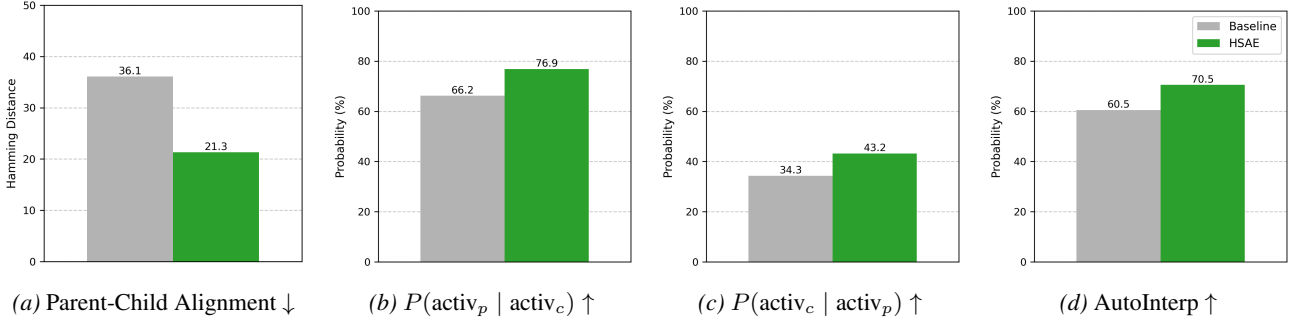
- **Parent-given-child probability**: Defined as $P(\text{activ}_p \mid \text{activ}_c) = P(\sigma(e_p^T x) > 0 \mid \sigma(e_c^T x) > 0)$, this measures **hierarchical necessity**. A high probability indicates that a child feature rarely activates without its parent, ensuring that children features remain strictly within the semantic scope of their parents.

- **Child-given-parent probability**: Defined as $P(\text{activ}_c \mid \text{activ}_p) = P(\sigma(e_c^T x) > 0 \mid \sigma(e_p^T x) > 0)$, this measures the **average semantic relevance** of each child feature to its parent. This metric evaluates the statistical strength of individual memberships within the hierarchy, where a higher value suggests that the parent feature is constituted by sub-components that are consistently co-active with it.

Baselines. We compare HSAE against a synthetic feature forest constructed from independently trained SAEs. We train a series of SAEs with dictionary sizes matching the corresponding levels of HSAE, then apply the same similarity-based assignment strategy to build a hierarchy post-hoc. This baseline represents structure obtained via post-hoc feature-splitting approaches.

Results. As illustrated in Figure 3, HSAE consistently outperforms the post-hoc baseline across all measured metrics. The substantial reduction in Hamming distance indicates that HSAE's parent features provide a significantly clearer and more complete logical summary of their constituent children. Furthermore, the improved co-activation probabilities establish the statistical prerequisite for semantic consistency, ensuring that the discovered hierarchical relations are functionally grounded rather than mere coincidences of vector similarity. These results demonstrate that jointly optimizing hierarchical relations alongside feature parameters enables a more consistent extraction of structure from the activation distribution comparing with post-hoc analysis.

4.3. Auto-Interpretability of Hierarchies

To evaluate whether the discovered hierarchical relations correspond to genuine semantic abstractions, we develop an automated pipeline using LLMs as judges. For a given parent-child pair, we independently retrieve high-activation examples for each feature and present them to the LLM. The LLM is tasked with a binary classification: determining whether the two sets of examples exhibit a valid hierarchical conceptual relationship (Yes/No) and providing an

(a) Parent-Child Alignment ↓   *(b)* $P(\text{activ}_p \mid \text{activ}_c)$ ↑   *(c)* $P(\text{activ}_c \mid \text{activ}_p)$ ↑   *(d)* AutoInterp ↑

Figure 3. **Quantitative evaluation of hierarchical structure. (a) Parent-Child Alignment**: Measured by the Hamming Distance between ground-truth parent activations and their logical-OR reconstruction from children, indicating how well parent features summarize their descendants' activations. **(b) Parent-given-child activation probability**: measuring the necessity of the parent feature for its children. **(c) Child-given-parent activation probability**: reflecting the coverage of children within the parent's activation space. **(d) AutoInterp**: LLM-based automated interpretability scores evaluating the semantic alignment between parent and child features.

associated confidence score. This automated setup allows for a large-scale evaluation of thousands of feature pairs, providing a statistically significant measure of semantic consistency. Detailed sampling strategies, prompt templates, and LLM configurations are provided in Appendix C.

Results. Figure 3d reports the rate at which the LLM identifies a valid hierarchical relationship. Again, HSAE significantly outperforms the synthetic baseline, confirming that its joint optimization extracts more semantically grounded structures than post-hoc alignment. This alignment between automated semantic judgment and previous statistical metrics provides strong evidence for the structural faithfulness of HSAE.
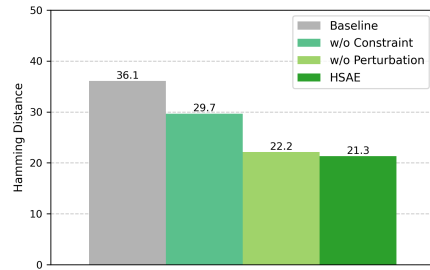
4.4. SAE Feature Quality

In this subsection, we evaluate HSAE from a standard SAE perspective by treating each level as an independent SAE. We first examine the reconstruction–sparsity trade-off by measuring the average L_0 and the fraction of variance explained. Following the SAEBench evaluation protocol (Karvonen et al., 2025), we further compare HSAE with baseline SAEs on a suite of established benchmarks, including Absorption (Absorp) (Chanin et al., 2024), Automated Interpretability (AutoInterp) (Paulo et al., 2024), Spurious Correlation Removal (SCR) (Karvonen et al., 2025), Sparse Probing (Gao et al., 2025), and RAVEL (Huang et al., 2024). Additional evaluation details are provided in Appendix D.

As shown in Table 1, HSAE achieves performance comparable to, and in some cases exceeding, baseline SAEs across most metrics. The overall results demonstrate that the hierarchical organization captured by HSAE does not come at the expense of feature quality. Instead, HSAE provides these structural insights as an additional benefit while maintaining the competitive reconstruction and interpretability standards of modern SAEs.

One notable result is the absorption score, where HSAE substantially outperforms baseline SAEs, particularly at larger dictionary sizes. While mitigating feature absorption is not an explicit goal of HSAE, this advantage is consistent with findings from Matryoshka SAE (Bussmann et al., 2025), which also induces a coarse-grained hierarchical ordering among features. Despite substantial differences in implementation between HSAE and Matryoshka SAE, the similarity in both design principles and empirical outcomes suggests that introducing hierarchical structure may help mitigate the feature absorption problem.

4.5. Ablation Studies

To isolate the contributions of different components in HSAE, we investigate two primary mechanisms for enforcing hierarchical alignment: **the parent-child constraint loss** and **random feature perturbation**. We evaluate their efficacy by measuring parent-child alignment across several configurations as shown in Figure 4.

Figure 4. **Ablation of training mechanisms.** We compare the parent-child alignment across: (1) *Baseline*: independently trained SAEs with post-hoc alignment; (2) *w/o Constraint*: without parent-child constraint loss; (3) *w/o Perturbation*: without random perturbation; and (4) *HSAE*: our complete implementation.

The results demonstrate that the full HSAE achieves the lowest Hamming distance of 21.3, confirming that both

Table 1. Evaluation results for HSAE and baseline SAEs on multiple SAE benchmarks. All metrics except L_0 are the higher the better.

Model	Dict. Size	L_0	Var. Exp.	Absorption	AutoInterp	RAVEL	SCR	Sparse Probing
HSAE	2048	49.4	0.612	0.983	0.807	0.521	0.272	0.868
	4096	50.0	0.651	0.981	0.814	0.582	0.286	0.873
	8192	50.5	0.685	0.951	0.859	0.627	0.314	0.874
	16384	50.6	0.710	0.922	0.869	0.654	0.324	0.873
Baseline	2048	49.4	0.613	0.988	0.813	0.539	0.245	0.863
	4096	49.8	0.651	0.974	0.829	0.609	0.302	0.879
	8192	50.3	0.684	0.901	0.854	0.627	0.320	0.862
	16384	50.7	0.712	0.844	0.861	0.661	0.338	0.855

mechanisms are essential for maintaining hierarchical consistency. Explicit constraint provides the foundational alignment, as its removal causes the Hamming distance to rise significantly to 29.7. Random perturbation further refines this relationship, reducing the distance by an additional 0.9 compared to using the constraint loss alone. Collectively, these mechanisms act complementarily to ensure that parent features faithfully encapsulate the collective activation patterns of their children.

We further investigate how different tree topology assumptions affect hierarchical alignment. Detailed results and comparative analyses are provided in Appendix E.

4.6. More Observations

In this section, we present further insights into the relationship between the learned hierarchical structure and the underlying activation space.

Geometric Manifestation of Hierarchy. We first examine whether the discovered hierarchical relationships are reflected in the geometry of the activation space. Figure 5 shows the UMAP projection of activations that trigger various features. We observe that activations for sibling features (those sharing a common parent) tend to cluster more closely together than those of unrelated features. This alignment indicates that the semantic hierarchy is not merely a structural constraint imposed by HSAE, but is also inherently represented in the geometric distribution of the LLM's activations. To demonstrate that Figure 5 is representative of a broader trend, we provide additional examples in Appendix G.

Branching Factor and Semantic Clarity. We further investigate how a feature's structural properties correlate with its interpretability. Specifically, we compare the AutoInterp scores of root features having only a single child against those that branch into multiple children. We find that multi-child roots exhibit a 2.49% higher average score—a performance delta comparable to the interpretability gain typically achieved by quadrupling the dictionary size (from
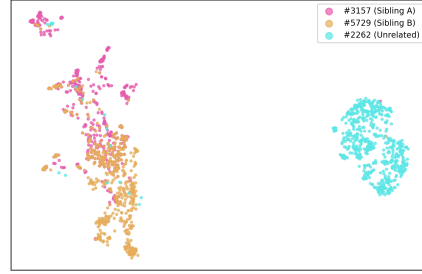

Figure 5. **Geometric manifestation of hierarchical relations.** UMAP projection of activations that trigger different features. Sibling A (#3157) and Sibling B (#5729) are features at level 2 that share the same parent (#14).

16k to 64k) in standard SAEs. This suggests an underlying link between a feature's hierarchical positioning and the semantic clarity of its represented concept. More detailed results are provided in Figure 11 in Appendix G.

5. Conclusion and Limitation

We introduced the Hierarchical Sparse Autoencoder, the first framework designed to explicitly learn conceptual hierarchies alongside sparse features. Through a joint optimization objective and a random perturbation mechanism, HSAE recovers meaningful taxonomies that align with human intuition. Our experiments demonstrate that HSAE significantly outperforms post-hoc alignment baselines in hierarchical consistency while preserving the feature quality of standard SAEs. Furthermore, we report unique observations linking the learned structure to the geometric properties of the activation space and feature interpretability.

Despite these advancements, several limitations remain. First, the current tree structure is constrained by a fixed number of levels, which may not fully capture the varying depths of semantic abstraction in large-scale models. Second, although parent-child co-activation is significantly improved, some misalignments persist; it remains unclear if these reflect methodological constraints or the inherently

non-hierarchical nature of certain activations. Future work should investigate investigate more flexible architectural priors and the scaling laws of hierarchical discovery. Beyond methodology refinements, these hierarchies provide a foundation for multi-scale model steering and safety auditing, enabling researchers to intervene in model behavior at the specific level of semantic granularity.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning by improving the interpretability of LLMs. By providing tools to uncover hierarchical structures in latent representations, our work contributes to the broader effort of making AI systems more transparent and understandable. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgment

References

Anders, E., Neo, C., Hoelscher-Obermaier, J., and Howard, J. N. Sparse autoencoders find composed features in small toy models. https://www.lesswrong.com/posts/a5wwqza2cY3W7L9cj/sparse-autoencoders-find-composed-features-in-small-toy, 2024.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

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

Bussmann, B., Leask, P., and Nanda, N. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=m25T5rAy43.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., and Bloom, J. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *ArXiv*, abs/2409.14507, 2024. URL https://api.semanticscholar.org/CorpusID:272827216.

Chanin, D., Dulka, T., and Garriga-Alonso, A. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025.

Clarke, M. A., Bhatnagar, H., and Bloom, J. Compositionality and ambiguity: Latent co-occurrence and interpretable subspaces. https://www.lesswrong.com/posts/WNoqEivcCSg8gJe5h/compositionality-and-ambiguity-latent-co-occurrence-and, 2024.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Heinzerling, B. and Inui, K. Monotonic representation of numeric attributes in language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 175–195, Bangkok,

Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.18. URL https://aclanthology.org/2024.acl-short.18/.

Huang, J., Wu, Z., Potts, C., Geva, M., and Geiger, A. Ravel: Evaluating interpretability methods on disentangling language model representations. *ArXiv*, abs/2402.17700, 2024. URL https://api.semanticscholar.org/CorpusID:268032000.

Jermyn, A. and Templeton, A. Ghost grads: An improvement on resampling., 2024. URL https://transformer-circuits.pub/2024/jan-update/index.html#dict-learning-resampling.

Kaddour, J. The minipile challenge for data-efficient language models. *ArXiv*, abs/2304.08442, 2023. URL https://api.semanticscholar.org/CorpusID:258180536.

Karvonen, A., Wright, B., Rager, C., Angell, R., Brinkmann, J., Smith, L., Mayrink Verdun, C., Bau, D., and Marks, S. Measuring progress in dictionary learning for language model interpretability with board game models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83091–83118. Curran Associates, Inc., 2024. doi: 10.52202/079017-2644. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9736acf007760cc2b47948ae3cf06274-Paper-Conference.pdf.

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C., Ayonrinde, K., Wearden, M., Conmy, A., Marks, S., and Nanda, N. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *ArXiv*, abs/2503.09532, 2025. URL https://api.semanticscholar.org/CorpusID:276937927.

Korznikov, A., Galichin, A., Dontsov, A., Rogov, O., Tutubalina, E., and Oseledets, I. Ortsae: Orthogonal sparse autoencoders uncover atomic features. *arXiv preprint arXiv:2509.22033*, 2025.

Leask, P., Bussmann, B., Pearce, M. T., Bloom, J. I., Tigges, C., Moubayed, N. A., Sharkey, L., and Nanda, N. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Li, Y., Michaud, E. J., Baek, D. D., Engels, J., Sun, X., and Tegmark, M. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2025.

Martin-Linares, C. P. and Ling, J. P. Attribution-guided distillation of matryoshka sparse autoencoders. *arXiv preprint arXiv:2512.24975*, 2025.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013a. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K. (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013b. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Mudide, A., Engels, J., Michaud, E. J., Tegmark, M., and de Witt, C. S. Efficient dictionary learning with switch sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k2ZVAzVeMP.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Paulo, G., Mallen, A. T., Juang, C., and Belrose, N. Automatically interpreting millions of features in large language models. *ArXiv*, abs/2410.13928, 2024. URL https://api.semanticscholar.org/CorpusID:273482460.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving sparse decomposition of language model activations with gated sparse autoencoders. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 775–818. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0024. URL https://proceedings.neurips.cc/paper_files

/paper/2024/file/01772a8b0420baec00c4d59fe2fbace6-Paper-Conference.pdf.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *ArXiv*, abs/2407.14435, 2024b. URL https://api.semanticscholar.org/CorpusID:271298201.

Riviere, G. T. M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ram'e, A., Ferret, J., Liu, P., Tafti, P. D., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stańczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C. A., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozi'nska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Pluci'nska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J. R., Gordon, J., Lipschultz, J., Newlan, J., Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., cia Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Gorner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., shad Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., chong Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M. M., Perrin, S., Arnold, S. M. R., bastian Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kociský, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models

at a practical size. *ArXiv*, abs/2408.00118, 2024. URL https://api.semanticscholar.org/CorpusID:270843326.

Wattenberg, M. and Viégas, F. Relational composition in neural networks: A survey and call to action. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=zzCEiUIPk9.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L.-C., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S.-Q., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.-C., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025. URL https://api.semanticscholar.org/CorpusID:278602855.

A. Detailed Settings

A.1. Experiment Setups

Data Preparation. For all experiments, we collect a dataset of 100M activation vectors from the target LLM. These activations are generated using input prompts with a sequence length of 1024 tokens. Following standard practice, we exclude activations from the `bos` token position. All activation vectors are normalized with a constant scaler to have a RMS norm expectation of $\sqrt{d}$ before being used for training.

Training Hyperparameters. We train all models using a batch size of 1024 with the Adam optimizer ($\beta_1 = 0, \beta_2 = 0.995$). We employ a cosine learning rate scheduler, starting with a linear warm-up phase covering the first 10% of training steps, followed by a decay to zero.

Computational Resources and Efficiency. All training and evaluation tasks are conducted on NVIDIA A800 GPUs. For a target model with hidden dimension $d = 2304$, training a 4-level HSAE (with dictionary sizes ranging from 2048 to 16384) takes approximately 6 hours. Compared to training independent SAEs of equivalent sizes, HSAE introduces a modest computational overhead. This is primarily attributed to the calculation of the hierarchical constraint loss and the memory latency incurred by fragmented indexing during the feature perturbation.

A.2. Dynamic Sparsity Control

To maintain a consistent level of sparsity throughout training, we implement a feedback control mechanism to dynamically adjust the sparsity penalty weight λ_ℓ. Specifically, we define a target sparsity $\hat{L}_0$ and model the desired trajectory of the average L_0 as a first-order exponential decay process:

$$\Delta L_0 = -\eta(L_0 - \hat{L}_0)$$

where η represents the decay rate, set to 0.001 in our experiments. We track the current L_0 using an exponential moving average (EMA) with a momentum of 0.999 to filter out high-frequency noise from individual batches. During each training step, we compare the observed change in L_0 against the theoretical ΔL_0 derived from the exponential model. If the observed reduction in sparsity is more rapid than the target trajectory, the sparsity weight λ_ℓ is decreased; conversely, if the model remains denser than expected, λ_ℓ is increased. In practice, we found that this velocity-based control significantly reduces oscillations in training compared to the control methods described in Karvonen et al. (2025).

B. More Case Studies

In this section, we provide further qualitative evidence of the hierarchical structures discovered by HSAE.

Figure 6 (a) illustrates the hierarchical organization of *Financial Resource Concepts*. A broad root feature (#1967) serves as the semantic anchor, which the model partitions into distinct functional domains: capital acquisition and fiscal status, categorized as *Funding Flows and Financial Capacity* (#3062), and the execution of financial duties, labeled as *Financial Transactions and Obligations* (#4829). Within these branches, the hierarchy further resolves into specialized leaf features that capture subtle semantic nuances. For instance, the model distinguishes between formal institutional support (e.g., *scholarship*, *awards* in #9354) and informal or abstract expressions of wealth (e.g., *deep pockets*, *budget* in #11940). Concurrently, the transactional branch successfully separates commercial settlements (e.g., *proceeds*, *cash payment* in #7291) from mandatory statutory costs (e.g., *tax*, *fees* in #12432).

Figure 6 (b) illustrates the hierarchical organization of *Writing and Language Concepts*. A broad root feature (#1307) serves as the semantic anchor, which the model partitions into distinct functional domains: active processes such as *Writing* (#2378) and *Reading Actions* (#4370), and their corresponding outputs, categorized as *Text Content* (#5006) and *Written Material* (#3681). Within the *Written Material* branch, the hierarchy further resolves into granular sub-features that distinguish between structured digital formats (*document*, #7996) and physical media (*paper*, #8366).

C. Auto-Interpretability of Hierarchies

This appendix provides implementation details for the auto-interpretability assessment described in Section 4.3. Our pipeline extends the SAEBench evaluation framework (Karvonen et al., 2025) with custom modifications to evaluate the hierarchical
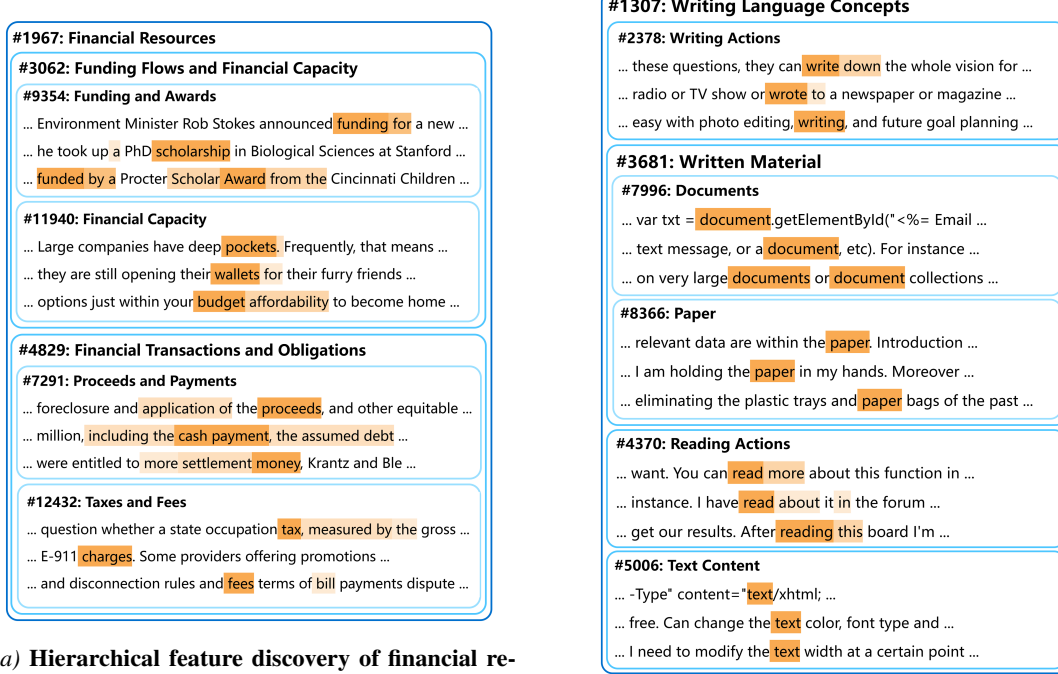
#1967: Financial Resources

#3062: Funding Flows and Financial Capacity

#9354: Funding and Awards

... Environment Minister Rob Stokes announced funding for a new ...

... he took up a PhD scholarship in Biological Sciences at Stanford ...

... funded by a Procter Scholar Award from the Cincinnati Children ...

#11940: Financial Capacity

... Large companies have deep pockets. Frequently, that means ...

... they are still opening their wallets for their furry friends ...

... options just within your budget affordability to become home ...

#4829: Financial Transactions and Obligations

#7291: Proceeds and Payments

... foreclosure and application of the proceeds, and other equitable ...

... million, including the cash payment, the assumed debt ...

... were entitled to more settlement money, Krantz and Ble ...

#12432: Taxes and Fees

... question whether a state occupation tax, measured by the gross ...

... E-911 charges. Some providers offering promotions ...

... and disconnection rules and fees terms of bill payments dispute ...

(a) **Hierarchical feature discovery of financial resource concepts.** The root feature (#1967) representing financial concepts branches into *Funding Flows and Financial Capacity* (#3062) and *Financial Transactions and Obligations* (#4829). Within the *Funding Flows* branch, HSAE further distinguishes between external support (*Funding and Awards*, #9354) and internal economic status (*Financial Capacity*, #11940). Within the latter, the hierarchy resolves into granular features for specific settlements (*Proceeds and Payments*, #7291) and statutory costs (*Taxes and Fees*, #12432).

#1307: Writing Language Concepts

#2378: Writing Actions

... these questions, they can write down the whole vision for ...

... radio or TV show or wrote to a newspaper or magazine ...

... easy with photo editing, writing, and future goal planning ...

#3681: Written Material

#7996: Documents

... var txt = document.getElementById("<%= Email ...

... text message, or a document, etc). For instance ...

... on very large documents or document collections ...

#8366: Paper

... relevant data are within the paper. Introduction ...

... I am holding the paper in my hands. Moreover ...

... eliminating the plastic trays and paper bags of the past ...

#4370: Reading Actions

... want. You can read more about this function in ...

... instance. I have read about it in the forum ...

... get our results. After reading this board I'm ...

#5006: Text Content

... -Type" content="text/xhtml; ...

... free. Can change the text color, font type and ...

... I need to modify the text width at a certain point ...

(b) **Hierarchical feature discovery of writing and language concepts.** The root feature (#1307) representing broad language-related concepts branches into specialized domains: *Writing Actions* (#2378), *Written Material* (#3681), *Reading Actions* (#4370), and *Text Content* (#5006). Within the *Written Material* branch, HSAE further distinguishes between digital or structured documents (*Documents*, #7996) and physical or raw media (*Paper*, #8366).

Figure 6. **Extended Case Studies of Hierarchical Feature Discovery.**

relations between HSAE features.

Feature Pairs Sampling and Data Collection. We randomly sample 5000 parent-child pairs from the learned hierarchical structure in HSAE. Activations are computed on sequences from the mini-PILE testing dataset. Following the SAEBench methodology, for each feature we collect 10 sequences with the highest activation values and sample 5 sequences with probability proportional to their activation values. These sequences are formatted by highlighting activating tokens with `<<token>>` syntax.

LLM Judge Configurations. We employ `Qwen3-Max` as the judge model.

The system prompt is as follows:

```
We're studying neurons in a sparse autoencoder (SAE) within a neural network.
Each neuron activates on specific words, substrings, or concepts in short
documents, with activating words indicated by << ... >>. You will be given
two sets of documents where two different neurons activate. Your task is to
compare the activation patterns of these two neurons and determine if there is
a parent-child relationship between them. A parent-child relationship means
that one neuron's activating concept (the child) is a subset or a more specific
version of the other neuron's activating concept (the parent). Analyze the
provided examples and output your judgment in the following format:
```

```
HaveRelationship:  [Yes/No]
Confidence:  [High/Medium/Low]
```

```
Do not include any additional text, explanations, or formatting.
```

The user prompt template:

```
Here are the activating documents for Neuron A:
[list of examples with activating tokens highlighted]
```

```
And for Neuron B:
[list of examples with activating tokens highlighted]
```

```
Based on these documents, determine if Neuron A and Neuron B have a parent-child
relationship.
```

D. SAEBench Evaluation Details

We evaluate our models using the SAEBench framework (Karvonen et al., 2025), adopting the default configurations for comparability. Since SAEBench provides a diverse set of metrics, we specify the exact metric identifiers used for our reporting in Table 2. Notably, for Absorption, we report $(1 - \text{score})$ to align it with other metrics where a higher value indicates superior performance.

Table 2. Mapping of reported benchmarks to specific SAEBench metrics.

Benchmark	Metric
Variance Explained	`explained_variance_legacy`
Absorption	`mean_absorption_fraction_score*`
AutoInterp	`autointerp_score`
RAVEL	`disentanglement_score`
SCR	`scr_metric_threshold_20`
Sparse Probing	`sae_top_5_test_accuracy`

E. More activation sources and target LLMs

To evaluate the generalizability of HSAE, we extend our experiments across multiple dimensions. First, we move beyond the 13th layer of `gemma-2-2b` to train HSAE on residual stream activations from early and late stages (layers 6 and 20). Second, we apply our method to the 18th layer of `qwen3-4b` (Yang et al., 2025) to verify cross-model consistency. Finally, we test the robustness of the discovered hierarchy under varying capacity constraints by training on `gemma-2-2b-layer-13` with target sparsities $L_0 \in \{80, 100\}$.

As summarized in Table 3, HSAE consistently maintains reconstruction fidelity comparable to the baselines across all tested settings. Here, the reported L_0 and Variance Explained represent the average performance across all hierarchical levels. Simultaneously, it yields significant improvements in parent-child alignment and co-activation probabilities. These results collectively demonstrate HSAE's ability to recover structured feature taxonomies is not dependent on specific layer depths, model architectures, or sparsity levels.

We note a slight discrepancy in the *Variance Explained* scores for `gemma2-2b-layer-13` between Table 1 and Table 3. This shift is attributed to the different evaluation corpora employed: the benchmarks in Table 1 strictly follow the SAEBench protocol using OpenWebText, whereas the results in Table 3 are computed using the mini-PILE testing dataset. Despite these distributional shifts, the reconstruction fidelity between HSAE and baseline remain comparable.

Table 3. Evluation results for HSAE and baseline SAEs on multiple benchmarks. All metrics except L_0 are the higher the better.

Activation Source	Model	L_0	Var. Exp.	$P(\text{activ}_p \mid \text{activ}_c)$	$P(\text{activ}_c \mid \text{activ}_p)$	Ham. Dist. ($\downarrow$)
`gemma2-2B-layer-6`	HSAE	50.1	0.759	0.752	0.403	23.2
	Baseline	50.0	0.759	0.651	0.331	34.6
`gemma2-2b-layer-13`	HSAE	50.1	0.713	0.769	0.432	21.3
		80.1	0.743	0.749	0.421	36.2
		100.7	0.757	0.747	0.415	46.3
	Baseline	49.8	0.714	0.662	0.343	36.1
		80.0	0.743	0.634	0.326	61.9
		100.0	0.758	0.619	0.317	78.9
`gemma2-2b-layer-20`	HSAE	50.0	0.729	0.800	0.461	19.4
	Baseline	50.0	0.727	0.729	0.395	30.6
`qwen3-4b-layer-18`	HSAE	50.2	0.988	0.347	0.197	28.7
	Baseline	50.0	0.988	0.309	0.179	46.4

F. Ablation Studies

F.1. Similarity Metric for Hierarchy Updates

We investigate the impact of different similarity metrics used during the hierarchy update. As shown in Figure 7, we compare three potential proxies for feature relatedness: (1) **Co-activation** (statistically derived co-activation probability), (2) **Decoder** (cosine similarity between decoder vectors), and (3) **Encoder** (cosine similarity between encoder vectors).

Our results indicate that all three metrics yield nearly indistinguishable performance in terms of the final hierarchical consistency. Since the choice of similarity metric does not significantly alter the outcome, we adopt Encoder similarity as the default for all experiments, as it offers the most direct and computationally efficient implementation within our training pipeline.
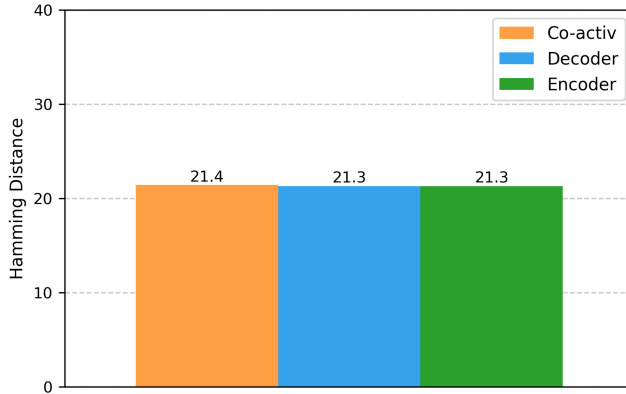

Figure 7. **Ablation of similarity metric used for hierarchy Update.** (1) *Co-activ*: Use statistically derived co-activation probability as similarity; (2) *Decoder*: Use decoder vector similarity; (3) *Encoder*: Use encoder vector similarity. All three metrics provide similar results.

F.2. Hyper-parameter Tuning

We examine the impact of key hyper-parameters on HSAE performance, specifically the parent-child constraint weight ρ and the random perturbation rate r. These parameters are essential for effective hierarchical training; as expected, setting both to zero would cause the method to degrade into independent SAEs.

Our analysis reveals a nuanced trade-off: while the hierarchical constraints guide organized feature discovery, an excessively

large constraint weight ρ imposes overly rigid structural priors that can bottleneck the representational capacity of the features, leading to a decrease in Variance Explained. Similarly, a perturbation rate r that is too high introduces excessive noise into the training dynamics, potentially causing instability and degraded feature quality. Unlike the ablations discussed in the main text—where all settings maintain high fidelity—these extreme configurations demonstrate a non-negligible trade-off between hierarchy consistency and reconstruction performance. We utilize the scatter plots in Figure 8 to visualize these relationships across different parameter regimes.

(a) Parent-Child Constraint Weight ρ *(b)* Random Perturbation Rate r

Figure 8. **Hyper-parameters' effects to reconstruction-hierarchy consistency trade-off.** Increasing parent-child constraint weight ρ or perturbation rates r improve hierarchical alignment but eventually degrades Variance Explained.

The final selection of hyper-parameters represents a principled balance between reconstruction fidelity and structural hierarchy consistency. We prioritize configurations that preserve maximal fidelity—ensuring the SAE effectively captures the original activation space—while simultaneously achieving optimal hierarchical alignment.
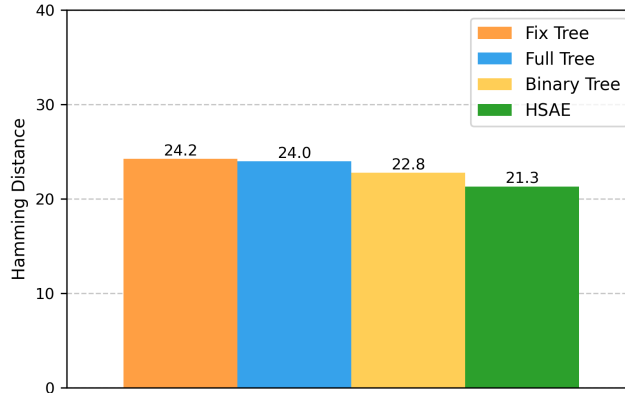
F.3. Tree Structural Assumptions

Figure 9. **Ablation of tree structural assumptions.** We compare parent-child alignment across different structural variants: (1) *Fix Tree*: static hierarchy after early initialization; (2) *Full Tree*: mandatory assignment for all features; (3) *Binary Tree*: a hierarchy constrained to a maximum branching factor of two; and (4) *HSAE*: our flexible partial tree approach.

We further investigate how different architectural assumptions regarding the tree topology affect hierarchical alignment. Specifically, we compare the default HSAE with three alternative configurations: **Binary Tree**, which restricts each parent to a maximum of two children; **Fix Tree**, where the hierarchy is determined early in training and remains static; and **Full Tree**, which forcefully assigns every feature to a parent.

16

As shown in Figure 9, the default HSAE outperforms all constrained variants. The performance gap compared to the Binary Tree value of 22.8 suggests that conceptual decomposition in LLMs is naturally multi-branching rather than strictly dyadic. The significantly higher Hamming distance of the Fix Tree at 24.2 highlights the necessity of an alternating optimization process, as feature representations and their optimal hierarchical assignments must co-evolve during training. Finally, the Full Tree baseline yields sub-optimal consistency with a distance of 24.0. This indicates that the more flexible Partial Tree structure is necessary for capturing a more robust hierarchical structure of LLM activations, as it avoids forcing assignments for features that do not exhibit strong similarity to any potential parent.
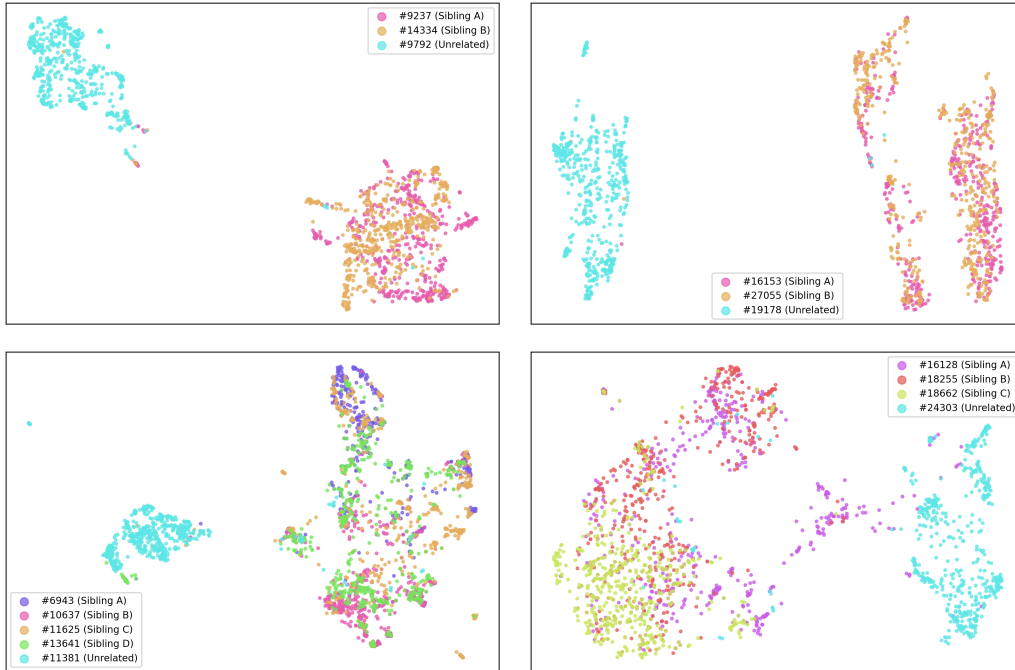
G. More Observations

G.1. Geometric Manifestation of Hierarchy

To understand how the discovered hierarchy manifests in the original activation space, we visualize the activation clusters using UMAP projection in Figure 10. Each plot contrasts the activations triggered by a set of sibling features (which share the same parent) against those triggered by an unrelated feature from a distant branch within the same hierarchical level.

We observe that unrelated features remain clearly isolated from sibling clusters. This geometric arrangement suggests that HSAE's structural constraints successfully capture the natural clustering of the LLM's latent space, where features with a common hierarchical ancestor are encoded in nearby geometric regions.
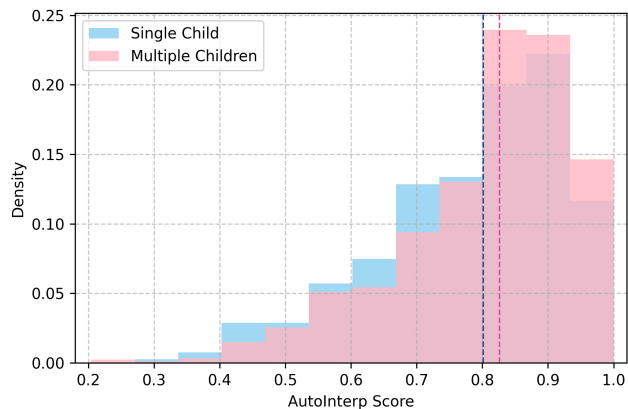
While sibling features exhibit spatial proximity in the UMAP visualization, we should note that their co-activation rate does not increase compared to baseline SAEs. This indicates that HSAE learns a disentangled decomposition rather than simply clustering redundant features. Specifically, although the parent-child constraint encourages structural organization, it does not force sibling features to fire simultaneously on the same tokens. On the contrary, the sparsity penalty inherent in the SAE objective naturally discourages such co-activation, ensuring that each sibling feature remains a distinct and sparse functional unit.

Figure 10. **Geometric manifestation of hierarchical relations.** UMAP projection of activations that trigger different features. In all cases, the unrelated feature is clearly seperated from sibling features.

G.2. Branching Factor and Semantic Clarity

As mentioned in the main text, Figure 11 provides a detailed distribution of AutoInterp scores categorized by the branching factor of root features.

Figure 11. **Interpretability versus structural branching.** Histogram of AutoInterp score of root features with single/multiple children. Feature with multiple children tends to be more interpretable comparing with those with single child. The average AutoInterp score gap is 2.49%, comparable to the performance gain achieved by quadrupling the dictionary size (from 16k to 64k) in standard SAEs.