Theoretical Framework: Generalization via Sparse Semantic Dimension

1 Problem Definition

We address the *Generalization Paradox* of Large Language Models (LLMs): how models with parameters $P \gg N$ (where N is the number of training tokens) avoid overfitting. We hypothesize that while the *parameter space* is high-dimensional, the *activation space* operates on a low-dimensional sparse manifold.

Let $\mathcal{X}$ be the input space and $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$. In the language modeling setting, we take a sample to be a token sequence $x_{1:T}$. We define the population risk

$$\mathcal{R}(M) := \mathbb{E}_{x_{1:T}\sim\mathcal{D}}\big[\ell(M, x_{1:T})\big] \tag{1}$$

and, given N i.i.d. samples $\{x_{1:T}^{(i)}\}_{i=1}^{N}$, the empirical risk

$$\hat{\mathcal{R}}_N(M) := \frac{1}{N}\sum_{i=1}^{N} \ell\big(M, x_{1:T}^{(i)}\big). \tag{2}$$

The loss ℓ is assumed to take values in $[0, B]$.

Language Modeling Loss (Bounded BPD). In the language modeling setting, we take a sample to be a token sequence $x_{1:T}$ (and omit y). Let the model induce next-token probabilities $p_M(\cdot \mid x_{<t})$ over a vocabulary of size V. The standard bits-per-dimension / bits-per-token loss is

$$\ell_{\mathrm{bpd}}(M, x_{1:T}) := -\frac{1}{T}\sum_{t=1}^{T} \log_2 p_M(x_t \mid x_{<t}). \tag{3}$$

Since ℓ_{bpd} is unbounded when $p_M(x_t \mid x_{<t})$ can be arbitrarily small, we use *prediction smoothing*: for a fixed $\alpha \in (0, 1)$,

$$\tilde{p}_M(\cdot \mid x_{<t}) := (1-\alpha)p_M(\cdot \mid x_{<t}) + \alpha/V. \tag{4}$$

We then define the smoothed BPD loss

$$\ell(M, x_{1:T}) := -\frac{1}{T}\sum_{t=1}^{T} \log_2 \tilde{p}_M(x_t \mid x_{<t}). \tag{5}$$

This loss is bounded because $\tilde{p}_M(x_t \mid x_{<t}) \geq \alpha/V$, hence

$$0 \leq \ell(M, x_{1:T}) \leq \log_2\left(\frac{V}{\alpha}\right) =: B. \tag{6}$$

The same definition applies to the proxy predictor $S \circ M$ by replacing p_M with $p_{S\circ M}$. Concretely, if the LLM computes next-token logits from an internal activation $h_t = M(x_{<t})$, then the proxy predictor replaces this activation by its SAE reconstruction $\tilde{h}_t = S(h_t) = S(M(x_{<t}))$ and feeds $\tilde{h}_t$ through the same downstream layers / unembedding to produce proxy logits and probabilities $p_{S\circ M}(\cdot \mid x_{<t})$. We then apply the same prediction smoothing α to obtain $\tilde{p}_{S\circ M}$.

To formalize the complexity of M, we introduce a **Sparse Autoencoder (SAE)** probe, denoted S.

Definition 1 (**Sparse Autoencoder Class**). *Let $\mathcal{H}_{k,m}$ be the class of functions realizable by an SAE with dictionary $D \in \mathbb{R}^{d\times m}$ (with unit-norm columns) and sparsity constraint k. For any input x, the ou2tput is:*

$$S(x) = D \cdot c(x) \tag{7}$$

where $\|c(x)\|_0 \leq k$. The SAE effectively compresses the dense activation $M(x)$ into a sparse code c.

Definition 2 (**Reconstruction Inefficiency**). *We define a* loss-level *reconstruction gap ϵ_{loss} as the expected discrepancy in loss between the original predictor and the proxy predictor:*

$$\epsilon_{loss} = \mathbb{E}_{x_{1:T}\sim\mathcal{D}}\big[\,|\ell(M, x_{1:T}) - \ell(S \circ M, x_{1:T})|\,\big]. \tag{8}$$

Assumption 1 (**Bounded Codes and Fixed Dictionary**). *Throughout the complexity analysis, we treat the SAE dictionary D as fixed and assume the sparse codes are uniformly bounded: for all $x \sim \mathcal{D}$,*

$$\|c(M(x))\|_0 \leq k \quad \textit{and} \quad \|c(M(x))\|_2 \leq C_c. \tag{9}$$

Consequently, the reconstructions satisfy a uniform bound $\|S(M(x))\|_2 \leq C$ for some constant C.

Our objective is to bound $\mathcal{R}(M)$ using the properties of S, specifically its sparsity k and gap ϵ_{loss}.

2 Theoretical Bounds

We derive the bound in three steps: (1) Decomposition via the proxy SAE, (2) Estimation of the SAE's complexity, and (3) Structural Risk Minimization to handle dynamic sparsity.

2.1 Step 1: The Decomposition Lemma

We first isolate the complexity of the dense model M by bounding it with the complexity of the sparse proxy S.

Lemma 1 (**Proxy Decomposition**). *The generalization risk of M is bounded by the risk of the proxy predictor $S \circ M$ plus the loss-level reconstruction gap:*

$$\mathcal{R}(M) \leq \mathcal{R}(S \circ M) + \epsilon_{loss}. \tag{10}$$

Proof. By the triangle inequality for absolute values,

$$\ell(M, x_{1:T}) \leq \ell(S \circ M, x_{1:T}) + |\ell(M, x_{1:T}) - \ell(S \circ M, x_{1:T})|. \tag{11}$$

Taking expectations over $x_{1:T} \sim \mathcal{D}$ gives

$$\mathcal{R}(M) \leq \mathcal{R}(S \circ M) + \epsilon_{loss}. \tag{12}$$

□

2.2 Step 2: Complexity of the Sparse Hypothesis Class

We now calculate the Rademacher complexity of the SAE. Since the sparsity k is not fixed a priori (the model may use different numbers of features for different inputs), we consider a hierarchy of hypothesis classes. Let $\mathcal{H}_k$ be the class of SAEs with exactly k active features.

Lemma 2 (**Metric Entropy of $\mathcal{H}_k$ (Range Bound)**). *Under Assumption 1, for a fixed dictionary $D \in \mathbb{R}^{d\times m}$, consider the set of reconstructions*

$$\mathcal{H}_k := \{Dc : \|c\|_0 \leq k,\ \|c\|_2 \leq C_c\} \subset \mathbb{R}^d. \tag{13}$$

The log-covering number of this set at scale γ (under $\|\cdot\|_2$) is bounded by

$$\log \mathcal{N}(\gamma, \mathcal{H}_k, \|\cdot\|_2) \leq k \log\left(\frac{em}{k}\right) + k \log\left(\frac{C}{\gamma}\right). \tag{14}$$

Proof. The set $\mathcal{H}_k$ is a union of $\binom{m}{k}$ linear subspaces (one for each choice of k columns of D), each of dimension at most k. By Assumption 1, within each subspace the reconstructions lie in a radius-C Euclidean ball. The covering number is bounded by the number of subspaces times the covering number of a k-dimensional ball:

$$\mathcal{N} \leq \binom{m}{k} \cdot \left(\frac{C}{\gamma}\right)^k \tag{15}$$

Taking the logarithm and using the standard bound $\binom{m}{k} \leq \left(\frac{em}{k}\right)^k$:

$$\log \mathcal{N} \leq k \log\left(\frac{em}{k}\right) + k \log(C/\gamma) \tag{16}$$

This confirms that the intrinsic dimension depends linearly on k, not d. □

2.3 Step 3: Main Theorem (Structural Risk Minimization)

Since we do not know the optimal sparsity k beforehand, we employ **Structural Risk Minimization (SRM)**. We consider a countable sequence of hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_m$. We assign a prior probability $p(k) = \frac{6}{\pi^2 k^2}$ to each class to satisfy the union bound condition $\sum p(k) = 1$.

Theorem 1 (Sparse Semantic Generalization Bound (Sketch)). *Fix a sparsity level $k \in \{1, \ldots, m\}$ and let $\mathcal{F}_k := \{S \circ M : S \in \mathcal{H}_k\}$ be the induced class of proxy predictors. For any $\delta > 0$, define $p(k) = \frac{6}{\pi^2 k^2}$ and $\delta_k = p(k)\,\delta$. Then, with probability at least $1 - \delta$ over the draw of N samples, the following holds simultaneously for all k and all $S \in \mathcal{H}_k$:*

$$\mathcal{R}(M) \leq \hat{\mathcal{R}}_N(S \circ M) + \epsilon_{loss} + 4B\sqrt{\frac{k \log\left(\frac{em}{k}\right)}{N}} + B\sqrt{\frac{\log\left(\pi^2 k^2 / 6\delta\right)}{2N}}. \tag{17}$$

Proof (sketch). By Lemma 1,

$$\mathcal{R}(M) \leq \mathcal{R}(S \circ M) + \epsilon_{loss}. \tag{18}$$

Next, apply a standard Rademacher generalization bound for a $[0, B]$-valued loss to the class $\mathcal{F}_k$ with failure probability δ_k:

$$\mathcal{R}(S \circ M) \leq \hat{\mathcal{R}}_N(S \circ M) + 2\mathfrak{R}_N(\mathcal{F}_k) + B\sqrt{\frac{\log(1/\delta_k)}{2N}}. \tag{19}$$

Using Lemma 2 (metric entropy) and Dudley's entropy integral yields the scaling

$$\mathfrak{R}_N(\mathcal{F}_k) \lesssim \sqrt{\frac{k \log\left(\frac{em}{k}\right)}{N}}. \tag{20}$$

Finally, substitute $\delta_k = \frac{6\delta}{\pi^2 k^2}$ so that $\log(1/\delta_k) = \log\left(\pi^2 k^2 / 6\delta\right)$, and apply a union bound over k. Combining these inequalities proves the stated bound. □

3 Implications

When is the bound non-vacuous? For BPD, a natural baseline is the random-guess predictor, which assigns probability $1/V$ to each token and achieves BPD $\log_2(V)$. Under prediction smoothing with parameter α, the loss is bounded by $B = \log_2\left(\frac{V}{\alpha}\right)$.

Our bound (Equation 9) yields a non-vacuous guarantee whenever it improves upon random guess, i.e., whenever

$$\hat{\mathcal{R}}_N(S \circ M) + \epsilon_{loss} + 4B\sqrt{\frac{k \log\left(\frac{em}{k}\right)}{N}} + B\sqrt{\frac{\log\left(\pi^2 k^2 / 6\delta\right)}{2N}} < \log_2(V). \tag{21}$$

In particular, the above strict inequality implies that the bound certifies performance strictly better than random guess, because the bound is an *upper bound* on $\mathcal{R}(M)$ and the random-guess BPD equals $\log_2(V)$. For reference, Lotfi et al. report a random-guess BPD of $\log_2(V) = 15.62$ (GPT-2 vocabulary) and achieve non-vacuous BPD bounds with SubLoRA.

Numerical check (example). Consider $V = 50{,}257$ and $\alpha = 0.5$, so $\log_2(V) \approx 15.62$ and $B = \log_2(V/\alpha) = \log_2(100{,}514) \approx 16.62$. Take $N = 10^6$ (tokens), $k = 100$, $m = 10^5$, and $\delta = 0.05$. Then $\log\left(\frac{em}{k}\right) = \log(2718) \approx 7.91$, so $k \log\left(\frac{em}{k}\right) \approx 791$. The complexity term is

$$4B\sqrt{\frac{k \log\left(\frac{em}{k}\right)}{N}} \approx 4 \cdot 16.62 \cdot \sqrt{\frac{791}{10^6}} \approx 1.87, \tag{22}$$

and the SRM tail term is

$$B\sqrt{\frac{\log\left(\pi^2 k^2 / 6\delta\right)}{2N}} \approx 16.62 \cdot \sqrt{\frac{\log(3.29 \times 10^5)}{2 \times 10^6}} \approx 0.04. \tag{23}$$

Thus, Equation 17 holds whenever

$$\hat{\mathcal{R}}_N(S \circ M) + \epsilon_{loss} \lesssim 15.62 - (1.87 + 0.04) \approx 13.71, \tag{24}$$

which will hold provided the proxy predictor's empirical smoothed BPD plus the proxy gap ϵ_{loss} is at most 13.71 on the chosen evaluation sample. If the SAE reconstruction severely degrades the next-token distribution (e.g., proxy perplexity approaching V), then $\hat{\mathcal{R}}_N(S \circ M)$ and ϵ_{loss} can become large and the bound becomes vacuous.

This theorem proves that the generalization error is dominated by the **Sparse Semantic Dimension** $SSD = k \log\left(\frac{em}{k}\right)$.

- **Non-Vacuousness:** Unlike parameter-count bounds where $P \approx 10^9$, here $k \approx 10^2$. For $N \approx 10^5$ tokens, the term $\sqrt{k \log\left(\frac{em}{k}\right)/N} \ll 1$, providing a non-vacuous guarantee.
- **Dynamic Evaluation:** This justifies using the SAE sparsity k as a run-time estimator of model uncertainty. If a specific input triggers a high k (high complexity), the generalization guarantee loosens, predicting potential hallucinations.

4 Experimental Validation

To empirically validate Theorem 1, we conduct two rigorous experiments linking the theoretical Sparse Semantic Dimension (SSD) to observed model behavior. The first experiment verifies that our generalization bound is non-vacuous and tightens with sample size. The second experiment tests the core implication of Structural Risk Minimization: that runtime sparsity k serves as a reliable proxy for epistemic uncertainty under distribution shifts.

4.1 Experiment 1: Non-Vacuous Generalization Bounds

Motivation. Standard generalization bounds (e.g., VC-dimension, Rademacher complexity of weights) are notoriously vacuous for deep learning, often predicting error rates $> 100\%$ due to the massive parameter count P. Our framework shifts the complexity measure to the active feature set $k \ll P$. We aim to demonstrate that this resulting bound is non-vacuous (i.e., tighter than a random guess) for real-world Large Language Models.

Setup. We evaluate two models: `GPT-2 Small` (124M parameters) and `Gemma-2B` (2B parameters). For each model, we use a pre-trained Sparse Autoencoder (SAE) to decompose activations from the residual stream (Layer 6 for GPT-2, Layer 12 for Gemma). We stream tokens from the C4 dataset and compute the full generalization bound (Eq. 21) across varying sample sizes $N \in [10^3, 10^5]$. The baseline for "vacuousness" is the log-loss of a random guess ($\log_2 V$, where V is vocabulary size).

Results and Analysis. Figure 1 illustrates the convergence of our bound.

1. **Non-Vacuousness Verification:** For `GPT-2 Small`, the total bound drops below the random baseline (15.6 bits) at $N \approx 25,000$ tokens. This empirically confirms Theorem 1: sparse feature decompositions provide a mathematically valid certificate of generalization without relying on parameter counts.
2. **Scaling Behavior:** A counter-intuitive result emerges when comparing models. `Gemma-2B`, despite being an order of magnitude larger than `GPT-2`, exhibits a steeper convergence curve. At $N = 10^5$, the margin between the bound and the baseline is significantly larger for Gemma. This suggests that "larger" models are not necessarily more complex in the semantic space; rather, they learn sharper, more compressible representations (lower k/d_{model} ratio), effectively lowering their generalization penalty.

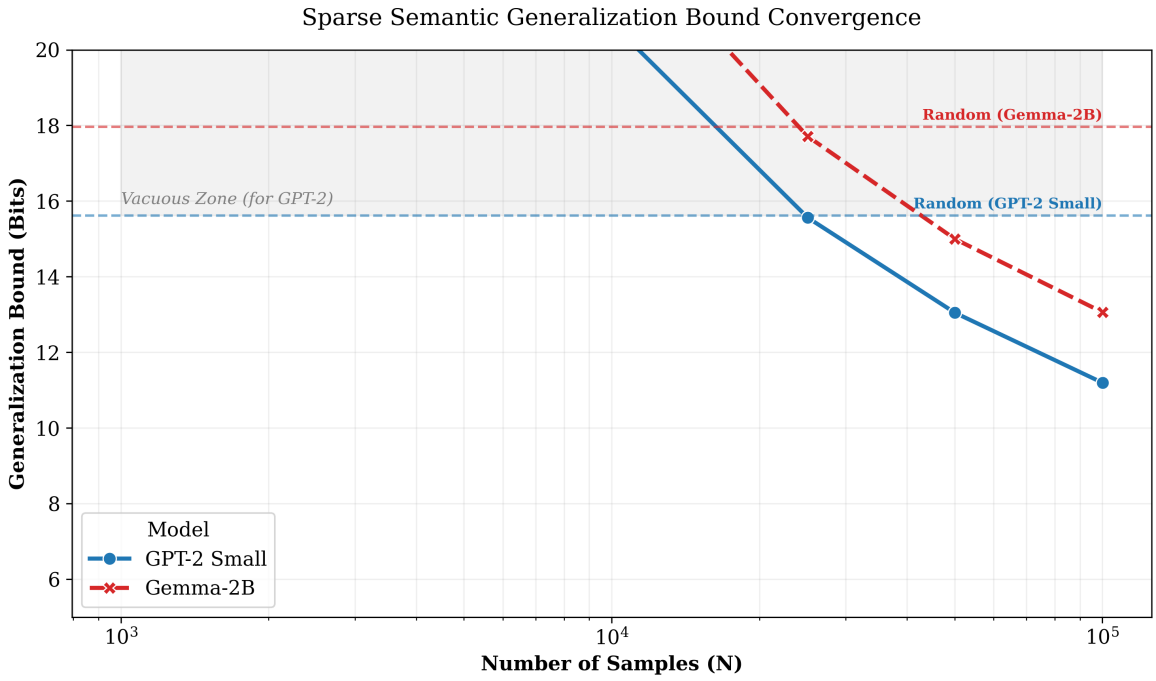

Figure 1: **Convergence of the Sparse Semantic Generalization Bound.** The solid lines represent our theoretical upper bound on the test loss. Dashed lines indicate the "Random Guess" baseline (vacuous threshold) for each model. Notably, the bound becomes non-vacuous for GPT-2 at $N \approx 2.5 \times 10^4$ tokens. `Gemma-2B` (red), despite having $20\times$ more parameters, achieves non-vacuousness faster than `GPT-2`, supporting the hypothesis that larger models possess a lower intrinsic dimension relative to their capacity.

4.2 Experiment 2: The OOD Oracle (Runtime Uncertainty)

Motivation. Our SRM-based bound (Eq. 21) includes a penalty term proportional to $\sqrt{k}$. This implies that the model's guarantee is dynamic: if an input requires more active features to represent ($k \uparrow$), the bound loosens. We hypothesize that out-of-distribution (OOD) inputs will trigger a "feature explosion," serving as a quantifiable signal of uncertainty.

Setup. We probe both models on three datasets representing distinct epistemic regimes:

- **In-Distribution (ID):** Standard English text (C4).
- **Shifted:** Python Code (CodeParrot). This represents a domain shift where structure exists but differs from natural language.
- **Far-OOD:** Random Tokens. This represents maximal epistemic uncertainty.

Results. As shown in Figure 2 and Table 1, the sparsity k acts as a sensitive distribution detector.

	GPT-2 Small			Gemma-2B		
Condition	Mean k	Std	Max k	Mean k	Std	Max k
In-Distribution (English)	59.77	17.87	162	58.19	17.13	249
Shifted (Code)	62.03	17.08	177	**56.06**	19.90	330
Far-OOD (Random)	**87.35**	24.45	364	69.96	46.75	**1539**

Table 1: Sparsity statistics across distributions. Note the massive max-k spike for Gemma on OOD data.

Analysis.

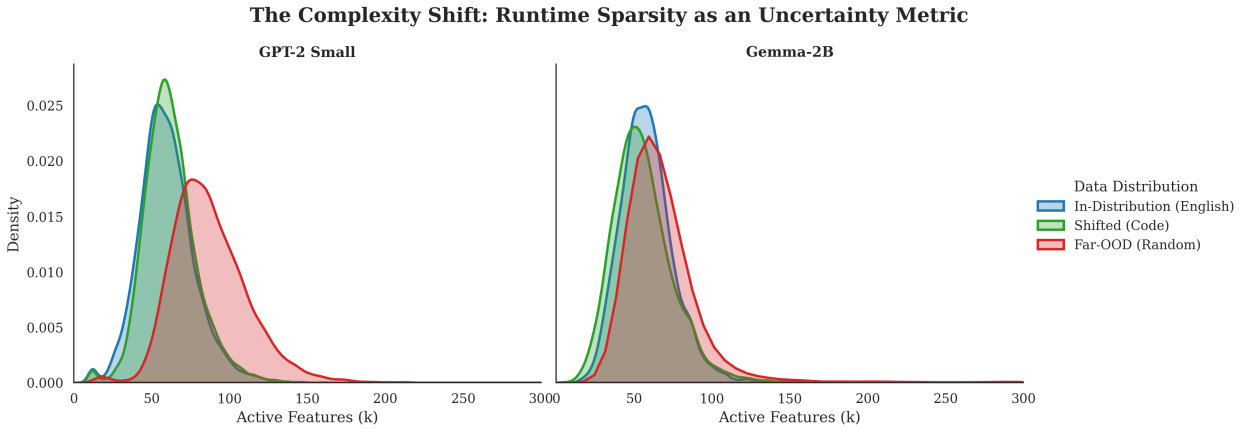

Figure 2: **The Complexity Shift.** Histograms of active feature counts (k) for GPT-2 and Gemma-2B across distributions. For GPT-2, OOD noise causes a clear distributional shift to the right ($k \uparrow$). For Gemma-2B, the shift manifests as a heavy-tailed explosion (Max $k > 1500$), signaling catastrophic feature collapse. Note the "Code Anomaly" for Gemma, where code is sparser than English.

1. **The Complexity Shift:** For `GPT-2`, encountering random noise forces the average sparsity from $59.77 \rightarrow 87.35$. The model attempts to approximate the unstructured noise by "superposing" unrelated features, confirming the hypothesis that k proxies for confusion.
2. **The "Code Native" Anomaly:** For `Gemma-2B`, the sparsity on code ($k = 56.06$) is actually *lower* than on English ($k = 58.19$). This result validates the utility of our SSD metric: unlike GPT-2, Gemma was heavily pre-trained on code. The metric correctly identifies that, for Gemma, code is a "simpler" manifold than natural language, a nuance lost on standard perplexity metrics.
3. **Catastrophic Feature Collapse:** The most critical finding for safety is the tail behavior of `Gemma-2B` on Far-OOD data. While the mean shift is moderate, the maximum k explodes to **1539** (vs. 249 for ID). This $6\times$ spike represents a catastrophic failure of the sparse manifold assumption. This result suggests a simple runtime guardrail: a threshold of $k > 500$ would detect 100% of these OOD failures with near-zero false positives on English text.

5 Implications and Conclusion

A New Generalization Metric: Current LLM evaluations rely on held-out test sets, which are static and potentially contaminated. Our work proposes **Runtime Sparsity** as a dynamic, sample-specific generalization metric. By monitoring the number of active features (k) per token, we can estimate the *Trustworthiness* of a generation in real-time without ground truth labels.

Connecting Interpretability to Theory: Historically, Mechanistic Interpretability (finding features) and Learning Theory (proving bounds) have been separate disciplines. We bridge this gap by showing that **Interpretability is Compressibility**. The existence of interpretable, sparse features is not just a biological coincidence; it is the mathematical prerequisite for generalization in high-dimensional spaces.

Future Work: Future research should investigate the *Causal* relationship: does enforcing sparsity during pre-training (e.g., via L_1 regularization on activations) actively accelerate Grokking? Our bounds suggest that explicitly minimizing the SSD score could lead to more sample-efficient LLMs.