Hierarchical Residual Sparse Autoencoders: Decomposing SAE Error into a Second Dictionary

Liz Lemma Future Detective

January 18, 2026

Abstract

Sparse autoencoders (SAEs) are a leading decomposition tool in mechanistic interpretability, yet a central limitation is reconstruction error: swapping true activations for SAE reconstructions often degrades downstream model performance, and the residual is commonly treated as uninterpretable "dark matter." Building on the observation (highlighted in recent interpretability reviews) that SAE residuals are structured and partially predictable from the original activations, we formalize a two-level sparse generative model in which activations decompose into (i) primary sparse features captured by an SAE dictionary and (ii) secondary sparse components that appear as systematic residuals. We prove that, under standard incoherence and sparsity assumptions, the residual induced by an approximately-correct first dictionary is itself sparse in a second dictionary, with an explicit bound on the effective residual noise in terms of the first-stage estimation error. This yields a principled hierarchical learning objective: train SAE-1 on activations, then train SAE-2 on residuals (optionally conditioned on SAE-1 codes) with an MDL-style penalty to prevent degenerate memorization. We provide sample-complexity upper bounds for recovering the residual dictionary (up to permutation/sign), and an irreducible-error lower bound showing when any fixed-size single-stage SAE must incur residual energy. Finally, we outline an experimental protocol on open LLMs (pretraining vs chat/safety data) demonstrating improved faithfulness under activation replacement and more stable downstream interventions (e.g., steering with fewer side effects), reframing SAE residuals as recoverable mechanistic structure rather than noise.

Table of Contents

3. 3. Problem Setup and Metrics: define activation distributions, reconstruction insertion faithfulness, residual structure metrics, and MDL objectives; formalize what it means to "explain residuals."

4. 4. Generative Model for Structured Residuals: two-level sparse model $a = D^\star z + E^\star y + \xi$; identifiability and incoherence conditions; discussion of when it is plausible for LLM activations.

5. 5. Algorithms: HR-SAE and CR-SAE: two-stage training, conditional residual modeling, and MDL-regularized sparsity; implementation knobs and invariants.

6. 6. Theory I — Residual is Sparse (Given Approximate First Dictionary): bound $r = a - \widehat{D}\widehat{z} = E^\star y + \tilde{\xi}$ and characterize $\tilde{\xi}$; when residual supports are recoverable.

7. 7. Theory II — Recovery and Sample Complexity for the Residual Dictionary: conditions and rates for recovering $E^\star$; stability across datasets/checkpoints as a corollary of shared $E^\star$.

8. 8. Theory III — Lower Bounds / Necessity: irreducible residual energy for any single-stage (m_1, k) model under the two-level generative process; MDL tradeoff showing when hierarchy beats "just make the SAE bigger."

9. 9. Experimental Design (Proposed): models, layers, data mixtures (pretraining vs chat vs safety), baselines (single SAE, SAE+error nodes), and evaluation metrics (faithfulness, residual predictability, intervention robustness).

10. 10. Results (Expected) and Ablations (Planned): residual dictionary interpretability, faithfulness gains vs compute, dataset dependence analysis; negative results and failure modes.

11. 11. Discussion: implications for mechanistic interpretability pipelines, monitoring/control, and limits; connection to geometry and nonlinearity.

12. 12. Conclusion and Open Problems: toward mechanism-level decompositions; residual hierarchies beyond two stages; interaction with on-manifold patching and cross-layer features.

1 Introduction

Sparse autoencoders (SAEs) have become a standard tool for interpreting internal activations of large neural networks: one learns a dictionary of features and an encoder that maps each activation vector to a sparse code, so that the decoded reconstruction approximates the original activation. In this paradigm, interpretability is commonly operationalized as sparsity and semantic coherence of the learned features, while *fidelity* is proxied by reconstruction error. Yet in many downstream uses—activation patching, causal interventions, and mechanistic attribution—the relevant quantity is not merely squared error in activation space, but *faithfulness*: replacing the model's true activation by its reconstruction should preserve the model's behavior. Empirically, we often observe that models tolerate certain reconstruction errors while being brittle to others, and that faithfulness can vary markedly across input distributions even when reconstruction metrics are held constant. This suggests that a single-stage SAE can be bottlenecked not only by optimization and capacity, but by a mismatch between what the dictionary can represent sparsely and what the model activation distribution actually contains.

We propose to treat the error of a first-stage SAE not as unstructured noise, but as a potential source of additional structure. Concretely, given an activation $a \in \mathbb{R}^d$ and a trained stage-1 dictionary D with code z, we define the residual

$$r \;:=\; a - Dz.$$

A common implicit assumption is that once an SAE is trained to a target reconstruction loss, the residual is essentially irreducible: either noise, or a diffuse signal whose representation would require dense codes. Our starting point is the contrary hypothesis: in regimes relevant to neural-network activations, a substantial portion of the residual may be *sparsely representable* in a second dictionary that is not redundant with the first. This is plausible whenever activations arise from multiple additive factors of variation with different sparsity patterns, scales, or statistical dependence on the input distribution. If stage-1 allocates its limited sparsity budget to the dominant factors, then the remaining components may be small in ℓ_2 yet structured and semantically meaningful—precisely the components that can disproportionately affect faithfulness when omitted or distorted.

This perspective leads to a simple but consequential design: train an SAE on the original activations, freeze it, compute residuals, and train a second sparse model on those residuals. The resulting hierarchical reconstruction $\hat{a} = Dz + Ey$ can be strictly better than merely enlarging the first dictionary or relaxing sparsity, because it allocates representational capacity along a new axis: it allows a *separate sparsity budget* for what stage-1 failed to capture. Moreover, by constraining the second stage to see only the residual

(or the residual together with a summary of the first-stage code), we enforce a meaningful separation of roles between stages, which is difficult to obtain in a monolithic overparameterized SAE.

Our technical contribution is to formalize when and why this procedure should work. We introduce a two-level sparse generative model in which activations decompose additively into a k-sparse component in a dictionary $D^\star$, an s-sparse component in a second dictionary $E^\star$, and noise. Under standard incoherence and random-support assumptions, we show that an approximate first-stage recovery implies that the residual remains a sparse signal in $E^\star$ up to an *effective noise* term controlled by the first-stage estimation error. This yields a reduction from residual learning to classical sparse dictionary learning, with explicit sample-complexity and recovery guarantees for the residual dictionary. Complementarily, we give a lower-bound perspective: if the second component lies outside what any single-stage (m_1, k) model can represent within its span and sparsity budget, then a nontrivial reconstruction floor is unavoidable, and a second stage is the correct way to allocate additional capacity.

A second conceptual contribution is to connect this hierarchy to description length. A frequent failure mode of increasing dictionary size in a single stage is that, absent strong regularization, the model can trade sparsity for idiosyncratic features that partially memorize the training distribution. Even when explicit ℓ_1 penalties are used, comparing models across sizes and sparsity levels is subtle: reducing error by adding latents is not necessarily a gain if the resulting representation is longer or less stable. We therefore incorporate an MDL-style penalty on the supports and amplitudes of the codes, and we analyze a regime in which the hierarchical model achieves a Pareto improvement: it can reduce reconstruction error while also reducing expected code length relative to any single-stage model achieving comparable reconstruction, because a small residual dictionary can be reused across many examples instead of expanding m_1 toward an implicit "one-latent-per-example" limit.

We instantiate these ideas in two practical architectures. First, *Hierarchical Residual SAEs* (HR-SAEs) implement the two-stage training pipeline described above, producing a decomposition of activations into primary features and residual features. Second, we consider *Conditional Residual* variants (CR-SAEs) in which the residual encoder is permitted to condition on the stage-1 code, while still being trained to reconstruct only the residual. This allows the second stage to represent patterns that are predictable from the first-stage features without collapsing the hierarchy into a single dense representation.

Finally, we emphasize the interpretability motivation: our goal is not merely lower $\|a - \hat{a}\|_2^2$, but higher faithfulness under activation replacement across distribution shifts (e.g. pretraining-like data versus chat-formatted or safety-adversarial prompts). The hierarchical view offers a concrete hypoth-

esis for why faithfulness can degrade under shift: the residual component can change in prevalence or semantics, and a first-stage SAE trained on one distribution may systematically discard precisely the components that matter on another. By learning residual structure explicitly and by quantifying the cost of representation via MDL, we obtain a framework in which reconstruction, compression, and faithfulness can be studied jointly rather than as ad hoc trade-offs.

2 Background and Related Work

Sparse coding and dictionary learning. Our setting is closest to the classical sparse coding / sparse dictionary learning (SDL) literature, in which one posits that observations admit a representation $a \approx Dz$ with a dictionary D and a sparse code z obtained by solving a penalized least-squares problem **??**. A large body of work studies identifiability and recovery of D under incoherence and random-support assumptions, typically via alternating minimization between (approximate) sparse coding and dictionary updates **??**. While these results are not directly stated for modern neural-network SAEs, they provide an organizing principle: when activations are generated by (approximately) sparse latent factors and the learned dictionary is sufficiently incoherent, sparse coding is both statistically meaningful and algorithmically tractable in regimes that exclude worst-case hardness.

Sparse autoencoders as amortized sparse coding. Sparse autoencoders may be viewed as an amortized variant of sparse coding: rather than solving an optimization problem for each activation, one trains an encoder network $\mathrm{Enc}(a)$ to predict a sparse code whose decoder reconstruction $\hat{a} = D\,\mathrm{Enc}(a)$ achieves low reconstruction loss with an explicit sparsity penalty (e.g. ℓ_1) or implicit sparsity constraints **??**. In the interpretability setting, D is commonly taken overcomplete ($m \gg d$) and the encoder is trained jointly with D. This joint training introduces additional degrees of freedom beyond the classical "optimize z for fixed D" picture; nevertheless, the underlying tension remains the same: the representation is useful only to the extent that a small-support code captures the directions of activation space that matter for downstream computation.

Reconstruction loss versus downstream faithfulness. A recurring empirical observation in mechanistic interpretability is that low $\|a - \hat{a}\|_2^2$ does not uniquely determine the effect of replacing a by $\hat{a}$ inside the forward pass. This is unsurprising from the perspective of representation theory: the model M defines a task-dependent seminorm on activations via its downstream Jacobian and nonlinearities, and squared error in the ambient ℓ_2

metric is only a proxy. Related concerns appear in work on model editing and activation patching, where small perturbations in certain subspaces can have disproportionate causal effects. Within the SAE paradigm, this motivates distinguishing "good" reconstruction error (orthogonal to sensitive directions) from "bad" error (aligned with directions used by the network), and it suggests that residual structure should be evaluated not only geometrically but functionally.

Residual structure and "error" features. Several lines of work effectively introduce explicit representations of what a first model fails to capture. In circuit analysis, one often treats the residual stream as a sum of interpretable components plus a remainder term and studies when that remainder carries meaningful signal. In SAE deployments, practitioners sometimes attach explicit "error" nodes or reserve capacity to represent reconstruction failures, especially when analyzing interventions that require faithful replacement rather than merely approximate denoising. Our approach fits into this theme, but we insist on a structural hypothesis: the residual is not merely a nuisance term to be carried along, but can itself be sparse in an additional dictionary that is not redundant with the first. This is closely related to hierarchical sparse coding models in signal processing, including two-layer or multi-layer dictionary learning, where one models a as a sum of components with different sparsity patterns and possibly different coherence structure **??**. The key distinction in our context is methodological: we use a first-stage SAE as a data-dependent projector that induces residuals, and then we learn a second sparse model on those residuals rather than enforcing a joint multilayer factorization from scratch.

Dataset dependence and distribution shift. It is well known that SDL and SAEs learn dictionaries that reflect the training distribution: the induced activation distribution changes across domains, prompting formats, and fine-tuning regimes, so the learned sparse factors need not be invariant. For neural networks, this distribution dependence is often the central practical concern: features learned on pretraining-like corpora can behave differently on chat-formatted data or safety-adversarial prompts. Work on feature stability, cross-dataset transfer, and "universal" representations can be viewed as attempts to control this dependence, either by training on mixtures of domains, by enforcing invariances, or by explicitly conditioning the representation on metadata describing the domain. In this landscape, a residual decomposition provides a specific hypothesis for how shift manifests: stage-1 may preferentially allocate its limited sparsity budget to the dominant factors on the source distribution, leaving a structured but underrepresented component that becomes salient under shift. A second-stage residual model is then a natural mechanism for capturing such changes without forcing a

single stage to compromise between heterogeneous regimes.

Meta-SAEs and model selection across capacities. Recent practice also includes training families of SAEs across widths, sparsity penalties, and architectural choices, and then selecting among them using downstream metrics (including faithfulness-style evaluations) rather than reconstruction alone. One may view this as a form of meta-model selection: the SAE is not an end in itself but an instrument for analysis, and its hyperparameters should be chosen to optimize the instrument's utility. Our contribution complements this perspective by proposing a structured enlargement of the model class—hierarchical residual dictionaries—that separates "primary" and "secondary" factors, and by pairing it with a code-length notion that allows comparisons across different latent budgets.

MDL, sparsity penalties, and compression. Finally, our use of an MDL-style objective draws on a classical equivalence: for suitable priors (e.g. Laplace or spike-and-slab), penalized least squares with an ℓ_1 or support-size term corresponds to a maximum a posteriori estimator, and the negative log prior can be interpreted as a codelength **??**. In sparse coding, MDL perspectives have been used to justify sparsity penalties, to select dictionary sizes, and to compare representations on a common "bits" scale rather than raw loss. In our setting, this matters because increasing dictionary size or relaxing sparsity can always reduce reconstruction error, but not necessarily in a way that yields a shorter or more stable representation. An explicit description-length term makes precise the intuition that a representation which achieves marginally lower ℓ_2 error by using many idiosyncratic latents may be worse than one that reuses a small set of residual features across examples. This lens is particularly natural for hierarchical models, where the second stage can be interpreted as allocating additional bits specifically to the portion of the activation not already explained by the first stage.

3 Problem Setup and Metrics

We fix a pretrained network M and a choice of layer or module at which we record activations. For an input x drawn from an input distribution $\mathcal{D}$, we write

$$a(x) \in \mathbb{R}^d$$

for the corresponding activation vector (or a flattened tensor). We consider multiple activation-inducing input distributions, denoted $\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{chat}}, \mathcal{D}_{\text{safety}}$, and we write $\mathcal{A}(\mathcal{D})$ for the induced distribution of $a(x)$ when $x \sim \mathcal{D}$. Our training data are i.i.d. samples $\{a_i\}_{i=1}^n$ from a chosen $\mathcal{A}(\mathcal{D}_{\text{train}})$, with evaluation potentially performed on the other domains to probe distribution shift.

Stagewise reconstructions and residuals. A stage-1 sparse representation consists of a dictionary (decoder) $D \in \mathbb{R}^{d \times m_1}$ with unit-norm columns and an encoder Enc_1 producing a code

$$z \;:=\; \mathrm{Enc}_1(a) \in \mathbb{R}^{m_1}, \qquad \text{typically with } \|z\|_0 \le k \text{ (approximately).}$$

The stage-1 reconstruction is $\hat{a}^{(1)} := Dz$, and we define the residual vector

$$r \;:=\; a - \hat{a}^{(1)} \;=\; a - Dz.$$

Given r, a stage-2 residual representation consists of a dictionary $E \in \mathbb{R}^{d \times m_2}$ (again with unit-norm columns) and an encoder Enc_2 producing

$$y \;:=\; \mathrm{Enc}_2(r) \in \mathbb{R}^{m_2}, \qquad \text{typically with } \|y\|_0 \le s \text{ (approximately),}$$

or in a conditional variant $y := \mathrm{Enc}_2(r, z)$. The hierarchical reconstruction is then

$$\hat{a} \;:=\; Dz + Ey.$$

The point of the hierarchy is not merely to decrease $\|a - \hat{a}\|_2^2$, but to allocate representational budget to the "unexplained" portion of a in a controlled manner.

Reconstruction metrics. We measure geometric fit by the mean squared reconstruction loss

$$L_{\mathrm{rec}}(a, \hat{a}) \;:=\; \|a - \hat{a}\|_2^2, \qquad \widehat{\mathcal{R}}_{\mathrm{rec}} \;:=\; \frac{1}{n} \sum_{i=1}^{n} \|a_i - \hat{a}_i\|_2^2.$$

To separate "what stage-1 leaves behind" from "what stage-2 captures," we also track residual energy and explained-residual fractions:

$$\rho_{\mathrm{res}} \;:=\; \frac{\mathbb{E}\|r\|_2^2}{\mathbb{E}\|a\|_2^2}, \qquad \rho_{\mathrm{unexp}} \;:=\; \frac{\mathbb{E}\|r - Ey\|_2^2}{\mathbb{E}\|r\|_2^2},$$

with empirical analogues obtained by averaging over samples. When ρ_{unexp} is small at fixed code length, we interpret the residual as possessing reusable structure rather than behaving as idiosyncratic noise.

Insertion faithfulness. Reconstruction loss in ℓ_2 is an imperfect proxy for the effect of replacing activations inside M. We therefore define an *insertion* (or *activation replacement*) evaluation. Let $\mathcal{L}_M(x)$ be the task loss of M on input x (e.g. cross-entropy with respect to the model's own next-token targets or an external label). Let $\mathcal{L}_M(x; \tilde{a})$ denote the loss when, at the chosen site, we overwrite the true activation $a(x)$ by a supplied vector $\tilde{a}$

while keeping all weights fixed. For a reconstruction map $T(a) := \hat{a}$, we define the faithfulness degradation

$$\mathrm{Faith}(T;\mathcal{D}) \;:=\; \mathbb{E}_{x\sim\mathcal{D}}\Big[\mathcal{L}_M(x;T(a(x))) - \mathcal{L}_M(x)\Big],$$

and we often report its empirical estimate on each $\mathcal{D}_{\mathrm{pre}}, \mathcal{D}_{\mathrm{chat}}, \mathcal{D}_{\mathrm{safety}}$. In settings where $\mathcal{L}_M$ is not directly available, one may equivalently use a divergence on logits or probabilities. The essential requirement is that Faith penalizes reconstruction error aligned with directions causally used downstream.

Residual structure metrics. To quantify whether stage separation is non-degenerate, we report (i) code sparsities $\|z\|_0$ and $\|y\|_0$ (or their ℓ_1 surrogates), (ii) cross-coherence

$$\mu(D,E) \;:=\; \max_{i,j} \big|\langle D_i, E_j\rangle\big|,$$

as a diagnostic for whether stage-2 is merely duplicating stage-1 directions, and (iii) a migration score measuring how much of the stage-2 reconstruction lies in $\mathrm{span}(D)$, e.g.

$$\mathrm{Mig} \;:=\; \frac{\mathbb{E}\|\Pi_{\mathrm{span}(D)}(Ey)\|_2^2}{\mathbb{E}\|Ey\|_2^2},$$

where $\Pi_{\mathrm{span}(D)}$ denotes orthogonal projection. Low Mig is not logically necessary for performance, but it provides evidence that the residual dictionary represents additional directions.

What it means to "explain residuals." Fixing a trained stage-1 pair (D, Enc_1), we say that residuals are (ϵ, s)-*explainable* by a residual dictionary E if for a fresh draw $a \sim \mathcal{A}(\mathcal{D})$, letting $r = a - D\mathrm{Enc}_1(a)$, there exists a code y with $\|y\|_0 \le s$ such that

$$\mathbb{E}\|r - Ey\|_2^2 \;\le\; \epsilon^2.$$

Operationally, we approximate the existential quantifier by an encoder Enc_2 trained to achieve this bound. The point is to separate *residual energy* (how large $\|r\|_2$ is) from *residual compressibility* (how well r can be represented by a small-support y drawn from a shared dictionary).

MDL-style objectives and code length. To compare models across different latent budgets, we introduce an explicit description-length proxy. For a code $u \in \mathbb{R}^m$ with support $S = \mathrm{supp}(u)$, we use a schematic penalty of the form

$$\mathrm{MDL}(u) \;\approx\; |S|\log m \;+\; \sum_{j\in S} \log\Big(1 + \frac{|u_j|}{\tau}\Big),$$

interpretable as (i) bits to specify indices and (ii) bits to specify quantized amplitudes at scale $\tau > 0$. Our training objective for the hierarchical model is then the empirical risk

$$\min_{D,E,\{\ z_i,y_i\ \}} \frac{1}{n}\sum_{i=1}^{n} \|a_i - Dz_i - Ey_i\|_2^2 + \lambda_1\|z_i\|_1 + \lambda_2\|y_i\|_1 + \beta\Big(\mathrm{MDL}(z_i) + \mathrm{MDL}(y_i)\Big),$$

subject to unit-norm constraints on dictionary columns. We use MDL not as a claim about optimal coding, but as a common scale on which we can state: (a) stage-2 is beneficial only if it reduces reconstruction and/or improves faithfulness *without* paying an excessive increase in description length, and (b) "explaining residuals" means achieving low ρ_{unexp} at controlled expected $\mathrm{MDL}(y)$.

4 A Two-Level Generative Model for Structured Residuals

We formalize the hypothesis that the residuals left by a first-stage sparse representation are not arbitrary noise, but possess reusable structure that can itself be sparsely represented. Concretely, we posit that activations admit a *two-level* additive decomposition

$$a \;=\; D^\star z \;+\; E^\star y \;+\; \xi, \tag{1}$$

where $D^\star \in \mathbb{R}^{d\times m_1}$ and $E^\star \in \mathbb{R}^{d\times m_2}$ are dictionaries with unit-norm columns, $z \in \mathbb{R}^{m_1}$ is k-sparse, $y \in \mathbb{R}^{m_2}$ is s-sparse, and ξ is a mean-zero noise term. The intended interpretation is that $D^\star z$ captures a "primary" set of features (those that a single-stage SAE of size m_1 is most likely to learn under an MDL/sparsity bias), while $E^\star y$ captures a secondary set of features that are systematically present but underrepresented by the first-stage model class.

Sparsity and support model. We assume $\|z\|_0 \le k$ and $\|y\|_0 \le s$, typically with random supports. A standard choice, sufficient for identifiability arguments, is that $\mathrm{supp}(z)$ and $\mathrm{supp}(y)$ are drawn independently from (approximately) uniform subsets of sizes k and s, respectively, and that conditional on the supports the nonzero coefficients are independent, symmetric, and subgaussian. Independence of supports is not a metaphysical claim; it is a modeling device ensuring that the two components are not adversarially entangled and that moment-based recovery arguments apply. In empirical settings we expect correlations, but we use the independent-support regime as an anchor point for what recovery should look like when residual structure is genuinely reusable across samples.

Incoherence and cross-incoherence. To rule out degenerate representations, we impose mutual incoherence within each dictionary and limited alignment across dictionaries. Writing $D_i^\star$ for the ith column, the basic control parameters are

$$\mu(D^\star) := \max_{i \neq j} |\langle D_i^\star, D_j^\star \rangle|, \qquad \mu(E^\star) := \max_{i \neq j} |\langle E_i^\star, E_j^\star \rangle|, \qquad \mu(D^\star, E^\star) := \max_{i,j} |\langle D_i^\star, E_j^\star \rangle|.$$

We require these coherences to be small enough relative to k and s that sparse codes are identifiable (e.g. $k \ll 1/\mu(D^\star)$ and $s \ll 1/\mu(E^\star)$, with a further constraint involving $\mu(D^\star, E^\star)$). Intuitively, incoherence ensures that no atom is easily substituted for a combination of other atoms at the target sparsity, while cross-incoherence ensures that the "secondary" atoms are not merely copies of "primary" atoms.

Identifiability up to permutation and sign. Even in the noiseless setting $\xi = 0$, the pair $(D^\star, z)$ is only identifiable up to column permutations and sign flips. Accordingly, whenever we compare learned dictionaries to ground truth we allow multiplication by a permutation/sign matrix P. This is the only non-identifiability we permit in the ideal regime; in particular, our coherence assumptions are intended to preclude continuous families of equivalent sparse decompositions.

Why the residual is expected to be structured. The hierarchy in (8) is motivated by the observation that a first-stage model constrained to k-sparse codes and size m_1 need not allocate capacity to all directions that matter for downstream computation. Even when $\|a - Dz\|_2$ is small on average, the remaining error may concentrate on a low-dimensional set of directions that occur intermittently. In this regime, the residual behaves neither as isotropic noise nor as example-wise idiosyncrasy: it is compressible by a shared dictionary. The role of $E^\star y$ is precisely to capture such intermittently active, yet reusable, residual directions.

Connection to stagewise training. Suppose a first-stage estimator outputs $(\widehat{D}, \widehat{z})$ approximating $(D^\star, z)$ (up to P). Then the residual used for stage two satisfies

$$r \;=\; a - \widehat{D}\,\widehat{z} \;=\; E^\star y \;+\; \tilde{\xi},$$

where the effective noise $\tilde{\xi}$ combines the base noise ξ with first-stage misspecification. The point of the generative model is not that $\tilde{\xi}$ is small in an absolute sense, but that the *signal component* of r remains sparse in a fixed dictionary $E^\star$ across samples. This is the mathematical expression of "structured residuals": there exists a second sparse model class in which residuals are more predictable than generic noise.

Plausibility for transformer activations. For large language models, the assumption of approximate linear superposition of features is empirically supported by the success of sparse autoencoders and related linear dictionary models on residual-stream-like representations. In such settings, $D^\star$ can be read as capturing high-frequency, broadly useful features (often those whose activation patterns are stable across $\mathcal{D}_{\text{pre}}$-like data), while $E^\star$ can encode lower-frequency or more context-dependent features that are nevertheless shared across many examples (e.g. discourse-structure phenomena, tool-use formatting, refusal/safety motifs, or rare syntactic constructions). Under distribution shift (e.g. from $\mathcal{D}_{\text{pre}}$ to $\mathcal{D}_{\text{chat}}$ or $\mathcal{D}_{\text{safety}}$), it is plausible that the mass of y (or its support) changes while the set of residual atoms remains largely stable, making residual modeling a natural way to reuse capacity across domains.

Scope and limitations. We emphasize what (8) does *not* claim. We do not assert that there are exactly two additive components, that coefficients are truly independent, or that ξ captures all nonlinearity. Rather, we use the two-level model as the minimal extension of single-dictionary sparsity that (i) explains why residual modeling should succeed when residuals are compressible, and (ii) yields concrete identifiability and sample-complexity predictions under standard incoherence and random-support assumptions. When these assumptions fail (e.g. strongly correlated supports, highly coherent features, or dense components that are not sparse in any fixed dictionary), we should expect feature migration between stages or diminished returns from stage two. The purpose of the model is therefore conditional: it delineates an interpretable regime in which residual dictionaries are learnable and worth learning, and it clarifies what empirical diagnostics (coherence, migration, residual explainability) are relevant to detect departures from that regime.

5 Algorithms: HR-SAE and CR-SAE

We consider a stagewise procedure that produces a hierarchical reconstruction of each activation vector $a \in \mathbb{R}^d$ of the form

$$\hat{a} \;=\; Dz \;+\; Ey, \tag{2}$$

where $D \in \mathbb{R}^{d\times m_1}$ is a first-stage dictionary and $E \in \mathbb{R}^{d\times m_2}$ is a residual dictionary trained on the stage-1 reconstruction error. The training objective is the empirical reconstruction loss augmented with sparsity and description-length control. We emphasize that the architectural details (linear dictionary learning versus a neural sparse autoencoder with a learned encoder) are orthogonal to the stagewise principle; we describe the algorithms in a form compatible with both.

HR-SAE (hierarchical residual sparse autoencoder). HR-SAE proceeds in two phases. In phase 1 we learn a sparse representation of a using a standard SAE objective

$$\min_{D,\mathrm{Enc}_1} \ \frac{1}{n}\sum_{i=1}^{n}\left\|a_i - Dz_i\right\|_2^2 \ + \ \lambda_1\|z_i\|_1 \ + \ \beta\,\mathrm{MDL}(z_i), \qquad z_i := \mathrm{Enc}_1(a_i), \tag{3}$$

subject to column normalization $\|D_j\|_2 = 1$. In phase 2 we freeze (D, Enc_1), form residuals

$$r_i \ := \ a_i - Dz_i, \tag{4}$$

and learn a second sparse model on $\{r_i\}$:

$$\min_{E,\mathrm{Enc}_2} \ \frac{1}{n}\sum_{i=1}^{n}\left\|r_i - Ey_i\right\|_2^2 \ + \ \lambda_2\|y_i\|_1 \ + \ \beta\,\mathrm{MDL}(y_i), \qquad y_i := \mathrm{Enc}_2(r_i), \tag{5}$$

again with $\|E_j\|_2 = 1$. The resulting hierarchical reconstruction is $\hat{a}_i = Dz_i + Ey_i$.

The principal design choice in HR-SAE is that Enc_2 receives only residual information (possibly after a fixed normalization), thereby imposing a form of stage separation: the second stage cannot directly "re-explain" the primary component unless it is present in the residual. In practice we often additionally regularize cross-alignment between the learned dictionaries (e.g. by penalizing large $\mu(D, E)$) to discourage the trivial solution in which E duplicates columns of D and absorbs a share of the first-stage signal.

CR-SAE (conditional residual modeling). The conditional variant modifies the second stage by allowing the residual encoder to depend on the stage-1 code:

$$y_i \ := \ \mathrm{Enc}_2(r_i, z_i), \tag{6}$$

or, equivalently, by parameterizing Enc_2 via a gating/predictor network g that maps z_i to per-latent thresholds, gains, or priors used when encoding r_i. The motivation is that residual structure may be predictable given which primary features are active (or given their amplitudes), even if the residual itself is small in norm. Conditioning allows the second stage to represent such "explained residual variability" without forcing E to grow large or y to become dense. Importantly, conditioning need not alter the decoder E; one can view CR-SAE as learning a conditional sparse coding distribution over y given z while keeping the residual dictionary shared across samples.

To preserve interpretability, we restrict conditioning to influence *only* the second-stage code selection mechanism, not the residual computation (4). In particular, we avoid feeding a_i directly into the residual encoder in CR-SAE; otherwise, the second stage could bypass the intended decomposition and act as an unrestricted second autoencoder.

MDL regularization and approximate sparsity. We use the term "MDL" as a convenient umbrella for code-length control. Concretely, $\mathrm{MDL}(z)$ and $\mathrm{MDL}(y)$ may be instantiated as (i) a support-size penalty proportional to $\|z\|_0$ or $\|y\|_0$ (implemented via hard top-k/top-s selection, or a differentiable approximation), plus (ii) an amplitude cost corresponding to quantization or a heavy-tailed prior (e.g. a $\log(1 + |z_j|/\tau)$ term). The role of β is to ensure that improvements in squared error are not obtained by encoding idiosyncratic detail with high-support codes; this is particularly important in stage 2, where the residual may contain a mixture of structured signal and unstructured noise.

Implementation knobs. The principal hyperparameters are (m_1, m_2), sparsity budgets (k, s) (or their continuous surrogates via λ_1, λ_2), and the MDL weight β. Beyond these, we have found the following to be structurally meaningful: (i) whether Enc_1 and Enc_2 enforce exact sparsity (e.g. top-k) or approximate sparsity (e.g. ℓ_1), (ii) residual normalization (e.g. scaling r_i to equalize variance across dimensions), which affects the effective noise level seen by stage 2, and (iii) dictionary orthogonalization or coherence penalties, which can reduce feature duplication across stages. A further knob is whether to perform an optional joint fine-tuning pass on (D, E) after stagewise training. When joint fine-tuning is used, we include an explicit penalty discouraging "feature migration" (for example, discouraging changes in D that increase residual energy explainable by E) so that the learned decomposition remains stable.

Stagewise invariants. The stagewise construction yields two useful invariants that we exploit both in analysis and in debugging. First, with stage 1 fixed, the optimal stage 2 loss is never worse than the stage 1 loss, since the second stage can choose $y_i = 0$ for all i, implying

$$\min_{E,\{y_i\}} \sum_i \|r_i - E y_i\|_2^2 \;\le\; \sum_i \|r_i\|_2^2. \tag{7}$$

Thus any observed increase in reconstruction error after adding stage 2 is attributable to optimization pathologies or to interactions introduced by joint training, rather than to the model class itself. Second, MDL penalties provide a safeguard against a degenerate regime in which stage 2 memorizes the residuals by using large supports or unstable amplitudes; empirically, monitoring $\mathbb{E}\|y\|_0$ and a proxy for $\mathrm{MDL}(y)$ is often a sharper diagnostic than monitoring squared error alone. These invariants are the algorithmic counterpart of the theoretical claim developed next: if stage 1 is approximately correct, then the residual behaves as a noisy sparse signal in a reusable dictionary, and stage 2 should be able to exploit this structure without paying excessive description length.

6 Theory I: Residual Sparsity Under an Approximate First Dictionary

We formalize the sense in which the stagewise residual

$$r \;:=\; a - \widehat{D}\widehat{z}$$

inherits a sparse generative structure whenever the first stage is approximately correct. Throughout this section we assume the two-level model

$$a \;=\; D^\star z \;+\; E^\star y \;+\; \xi, \tag{8}$$

where $\|z\|_0 \le k$, $\|y\|_0 \le s$, the supports are drawn independently, and ξ is mean-zero subgaussian with parameter σ^2. We further assume an alignment matrix P (permutation/sign) such that the first-stage dictionary estimate satisfies

$$\big\|\widehat{D} - D^\star P\big\|_{2\to 2} \le \varepsilon, \qquad \big\|\widehat{z} - P^\top z\big\|_2 \le \eta,$$

where the second bound should be interpreted as holding in expectation or with high probability, depending on the encoder/sparse coding method.

Residual decomposition and effective noise. A direct algebraic manipulation shows that r decomposes into the desired residual component $E^\star y$ plus an "effective noise" term absorbing both base noise and first-stage misspecification:

$$\begin{aligned} r &= a - \widehat{D}\widehat{z} \\ &= D^\star z + E^\star y + \xi - \widehat{D}\widehat{z} \\ &= E^\star y \;+\; \underbrace{\xi + (D^\star P - \widehat{D})\widehat{z} + D^\star(P\widehat{z} - z)}_{=:\,\tilde{\xi}}. \end{aligned} \tag{9}$$

Thus the residual is itself generated by a sparse model in dictionary $E^\star$, up to the effective noise $\tilde{\xi}$. The latter contains (i) the original stochastic noise ξ, (ii) a *dictionary drift* term $(D^\star P - \widehat{D})\widehat{z}$, and (iii) a *code error* term $D^\star(P\widehat{z} - z)$.

Norm control for $\tilde{\xi}$. From (9) we obtain deterministic bounds of the form

$$\|\tilde{\xi}\|_2 \;\le\; \|\xi\|_2 \;+\; \varepsilon\,\|\widehat{z}\|_2 \;+\; \|D^\star\|_{2\to 2}\,\|P\widehat{z} - z\|_2. \tag{10}$$

Squaring and taking expectations (using $(u+v+w)^2 \le 3(u^2+v^2+w^2)$) yields

$$\mathbb{E}\|\tilde{\xi}\|_2^2 \;\le\; 3\,\mathbb{E}\|\xi\|_2^2 \;+\; 3\,\varepsilon^2\,\mathbb{E}\|\widehat{z}\|_2^2 \;+\; 3\,\|D^\star\|_{2\to 2}^2\,\mathbb{E}\|P\widehat{z} - z\|_2^2. \tag{11}$$

In the canonical sparse regime (bounded moments for nonzeros and $\|\widehat{z}\|_0 \leq k$), one typically has $\mathbb{E}\|\widehat{z}\|_2^2 = O(k)$ and $\mathbb{E}\|P\widehat{z} - z\|_2^2 = O(\eta^2)$, so (11) may be summarized as an effective variance inflation

$$\mathbb{E}\|\tilde{\xi}\|_2^2 \;\lesssim\; \mathbb{E}\|\xi\|_2^2 \;+\; O(\varepsilon^2 k) \;+\; O(\eta^2\,\|D^\star\|_{2\to 2}^2).$$

This is the quantitative sense in which "good" stage-1 estimation implies that the stage-2 problem is a standard sparse coding instance with additional noise whose scale is controlled by (ε, η).

When is the residual support recoverable? Let $S := \mathrm{supp}(y)$ denote the residual support. To reason about recoverability we use the standard mutual coherence $\mu(E^\star) := \max_{i\neq j} |\langle E_i^\star, E_j^\star\rangle|$ and consider the correlations

$$\langle E_j^\star, r\rangle = y_j + \sum_{\ell\in S\setminus\{j\}} \langle E_j^\star, E_\ell^\star\rangle\, y_\ell + \langle E_j^\star, \tilde{\xi}\rangle. \tag{12}$$

The middle term is the usual sparse-coding interference bounded by $\mu(E^\star)(s-1)\|y\|_\infty$, while the last term is controlled by $\|E^{\star\top}\tilde{\xi}\|_\infty$. Since $\|E_j^\star\|_2 = 1$,

$$\|E^{\star\top}\tilde{\xi}\|_\infty = \max_j |\langle E_j^\star, \tilde{\xi}\rangle| \;\leq\; \|\tilde{\xi}\|_2, \tag{13}$$

and under additional distributional assumptions on ξ (and on the stage-1 error terms) one can further upgrade (13) to high-probability bounds scaling like $\sqrt{\log m_2}$.

A sufficient condition for exact support recovery by simple correlation thresholding (and, a fortiori, by OMP under comparable conditions) is that the nonzero coefficients dominate both interference and effective noise. For instance, if

$$\min_{j\in S} |y_j| \;>\; \mu(E^\star)(s-1)\|y\|_\infty \;+\; 2\,\|E^{\star\top}\tilde{\xi}\|_\infty, \tag{14}$$

then the set of indices with the largest s magnitudes in $E^{\star\top} r$ equals S. When combined with (10)–(11), condition (14) makes explicit the tradeoff: larger first-stage error (larger ε or η) increases $\|E^{\star\top}\tilde{\xi}\|_\infty$ and thereby shrinks the regime in which residual supports are identifiable.

Interpretation for HR-SAE/CR-SAE. Equation (9) is the key structural claim needed for the stage-2 analysis: conditional on an approximately correct first stage, the residual is a noisy sparse signal in $E^\star$. The only ways in which the stagewise construction can fail to expose this structure are (i) ε and η are too large (so $\tilde{\xi}$ dominates), or (ii) the residual coefficients are themselves too small relative to coherence/interference. The MDL pressure in stage 2 addresses a complementary failure mode: even when (14) does

not hold pointwise, aggregated learning of a reusable E remains possible provided we prevent the encoder from encoding idiosyncratic noise via large supports. In the next section we treat the residuals $\{r_i\}$ as samples from (9) and state recovery and sample complexity guarantees for learning $E^\star$ from these noisy sparse observations.

7 Theory II: Recovery and Sample Complexity for the Residual Dictionary

We now treat the residuals as observations from a noisy sparse dictionary model and state conditions under which a second-stage learner recovers the residual dictionary $E^\star$ (up to permutation/sign). Concretely, for each activation sample we form

$$r_i \;:=\; a_i - \widehat{D}\,\widehat{z}_i,$$

and, by the residual representation established previously, we may regard

$$r_i \;=\; E^\star y_i \;+\; \tilde{\xi}_i, \qquad \|y_i\|_0 \le s, \tag{15}$$

where $\tilde{\xi}_i$ is an effective noise term whose second moment is controlled by the stage-1 errors. The stage-2 goal is to learn a dictionary $\widehat{E}$ such that $\widehat{E}$ is close to $E^\star Q$ for some permutation/sign matrix Q, and to infer sparse codes y_i with small support and small reconstruction error.

Identifiability regime. We work in a standard incoherent random-support setting. Let $\mu(E^\star) := \max_{p\neq q} |\langle E^\star_p, E^\star_q\rangle|$. Assume:

1. **Incoherence/sparsity:** $s \le c_0/\mu(E^\star)$ for a sufficiently small constant c_0.
2. **Random supports:** $\mathrm{supp}(y_i)$ is drawn uniformly among s-subsets (or i.i.d. Bernoulli with expected size s), independently across i.
3. **Coefficient regularity:** conditional on its support, the nonzeros of y_i are independent, mean-zero, subgaussian, and satisfy $\mathbb{E}[y_{ij}^2] \in [\underline{\nu}, \overline{\nu}]$ on-support.
4. **Noise control:** $\tilde{\xi}_i$ is mean-zero (or has been centered) and satisfies $\mathbb{E}\|\tilde{\xi}_i\|_2^2 \le \sigma_{\mathrm{eff}}^2$ with σ_{eff}^2 small enough relative to the target accuracy.

Under these conditions, the model (15) falls within the scope of classical dictionary learning analyses: the sparse component is identifiable up to permutation/sign, and alternating minimization (sparse coding then dictionary update) is locally contractive when initialized within a constant-radius neighborhood of $E^\star$.

A representative recovery guarantee. To make the dependence on residual noise explicit, we state a typical theorem in the style of alternating-minimization results (the precise constants depend on the sparse coding subroutine and coefficient distribution).

Theorem 7.1 (Residual dictionary recovery, informal)**.** *Assume* (15) *and the identifiability regime above. There exists an alternating-minimization procedure that, given n residual samples and an initialization $\widehat{E}^{(0)}$ satisfying $\|\widehat{E}^{(0)} - E^\star Q\|_F \le c_1$ for some permutation/sign Q, returns $\widehat{E}$ such that*

$$\|\widehat{E} - E^\star Q'\|_F \;\le\; \delta$$

for some permutation/sign Q', with probability at least $1-\exp(-\Omega(\log m_2))$, provided

$$n \;\ge\; \widetilde{O}\Big(\frac{m_2\, s\, \log d}{\delta^2}\Big) \qquad \textit{and} \qquad \sigma_{\text{eff}}^2 \;\le\; c_2\, \delta^2. \tag{16}$$

How stage-1 errors enter. Theorem 7.1 reduces the stage-2 learning problem to a requirement on $\sigma_{\text{eff}}^2 = \mathbb{E}\|\tilde{\xi}\|_2^2$. Combining (16) with the earlier control of $\mathbb{E}\|\tilde{\xi}\|_2^2$ yields a sufficient condition of the schematic form

$$\mathbb{E}\|\xi\|_2^2 \;+\; O(\varepsilon^2\, \mathbb{E}\|\widehat{z}\|_2^2) \;+\; O(\eta^2\, \|D^\star\|_{2\to 2}^2) \;\lesssim\; \delta^2.$$

Thus, for fixed target accuracy δ, the stage-2 sample size must scale as in (16), and the stage-1 misspecification must be sufficiently small that it does not inflate the residual noise beyond the accuracy scale. Conversely, for fixed stage-1 quality (fixed ε, η), the smallest attainable δ is lower bounded by the induced σ_{eff}; in that sense, the second stage cannot exceed the fidelity of the residual signal it is given.

Initialization and practical SAE training. Theorem 7.1 is stated with a basin-of-attraction assumption. In practice, neural SAE training does not literally implement alternating minimization, but it often behaves like a smooth surrogate: the encoder approximates sparse coding and gradient updates approximate dictionary refinement. To connect practice to theory, one may (i) use a spectral or clustering-based initializer for E from residual correlations, or (ii) rely on overparameterization and mild regularization to land in a favorable region. Our claims here are therefore best read as *identifiability and sample-size guidance*: once the residuals obey (15) with small σ_{eff}, the second stage is a well-posed sparse learning problem rather than an unconstrained memorization task.

Stability across datasets and checkpoints. A useful corollary of residual identifiability is *dictionary stability* when the residual mechanism is

shared. Suppose two activation-inducing distributions (e.g. $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{chat}}$) produce residuals of the form

$$r^{(t)} = E^\star y^{(t)} + \tilde{\xi}^{(t)}, \qquad t \in \{1, 2\},$$

with the same $E^\star$ but potentially different code distributions for $y^{(t)}$ (different marginal on supports or amplitudes), and with effective noise levels bounded by a common σ_{eff}^2. Training stage-2 dictionaries $\widehat{E}^{(1)}$ and $\widehat{E}^{(2)}$ on the two residual datasets with sample sizes satisfying (16) yields

$$\min_Q \|\widehat{E}^{(1)} - \widehat{E}^{(2)} Q\|_F \;\le\; \|\widehat{E}^{(1)} - E^\star Q_1\|_F + \|E^\star Q_1 - E^\star Q_2\|_F + \|E^\star Q_2 - \widehat{E}^{(2)}\|_F \;\lesssim\; \delta,$$

for appropriate permutation/sign matrices, with high probability. The same reasoning applies across nearby model checkpoints when the residual directions $E^\star$ persist: learned residual features should align up to permutation/sign, and lack of alignment can be interpreted as evidence that the residual component itself has changed (rather than merely the first-stage decomposition).

In summary, once the residuals lie in the identifiable sparse regime, the second stage inherits standard recovery guarantees with sample complexity scaling like $\widetilde{O}(m_2 s \log d/\delta^2)$, and the resulting residual dictionary is expected to be stable across datasets and checkpoints that share the same underlying residual structure.

8 Theory III: Lower Bounds and the Necessity of Hierarchy

We now formalize two complementary senses in which a single-stage sparse model is intrinsically limited under the two-level generative process. The first is an *irreducible reconstruction floor* arising from misspecification: if a k-sparse model of size m_1 is effectively dedicated to the $D^\star z$ component, then the additive residual component $E^\star y$ cannot be removed beyond its projection onto the representable subspace. The second is an *MDL/code-length lower bound*: even when a single-stage model is permitted to expand, matching the accuracy of a hierarchical representation typically forces either (i) a much larger dictionary (approaching "one-latent-per-pattern"), or (ii) a larger effective support size, both of which increase description length.

A misspecification floor for single-stage (m_1, k) representations. Fix any encoder–decoder pair (ϕ, ψ) where $\phi : \mathbb{R}^d \to \mathbb{R}^{m_1}$ obeys $\|\phi(a)\|_0 \le k$ almost surely and $\psi(x) = \widehat{D}x$ for some $\widehat{D} \in \mathbb{R}^{d \times m_1}$ with unit-norm columns (we allow $\widehat{D}$ to be arbitrary, not necessarily equal to $D^\star$). Denote the orthogonal projector onto $\operatorname{span}(\widehat{D})$ by $\Pi_{\widehat{D}}$, and write $a = u + v + \xi$ with $u := D^\star z$

and $v := E^\star y$. Then for any reconstruction $\widehat{a} = \psi(\phi(a)) \in \mathrm{span}(\widehat{D})$ we have the deterministic inequality

$$\|a - \widehat{a}\|_2^2 \;\geq\; \|(I - \Pi_{\widehat{D}})(u+v)\|_2^2 \;-\; 2\langle (I - \Pi_{\widehat{D}})(u+v),\, \xi\rangle. \tag{17}$$

Taking expectations and using $\mathbb{E}\,\xi = 0$ (and independence from (z, y)) yields

$$\mathbb{E}\|a - \widehat{a}\|_2^2 \;\geq\; \mathbb{E}\|(I - \Pi_{\widehat{D}})(u+v)\|_2^2. \tag{18}$$

In the regime motivating our construction, $\widehat{D}$ is chosen so that $\mathrm{span}(\widehat{D})$ well-approximates $\mathrm{span}(D^\star)$ while having insufficient capacity to additionally approximate $\mathrm{span}(E^\star)$. A convenient abstraction is to assume that, for all v in the support of $E^\star y$, the projection error satisfies $\|(I - \Pi_{\widehat{D}})v\|_2 \geq (1 - \epsilon_E)\|v\|_2$ for some $\epsilon_E \in [0, 1)$. If moreover $\mathbb{E}\langle u, v\rangle = 0$ (as in the generative model with independent mean-zero coefficients and incoherent dictionaries), then expanding (18) gives

$$\mathbb{E}\|a - \widehat{a}\|_2^2 \;\geq\; \mathbb{E}\|(I - \Pi_{\widehat{D}})u\|_2^2 \;+\; (1 - \epsilon_E)^2\, \mathbb{E}\|v\|_2^2. \tag{19}$$

Thus, even if the first term is made negligible by learning $\widehat{D}$ close to $D^\star$, the residual energy $\mathbb{E}\|E^\star y\|_2^2$ remains as a floor up to the representability factor $(1 - \epsilon_E)^2$. This is precisely the regime where the second stage is not merely helpful but necessary: it provides additional representational directions *outside* $\mathrm{span}(\widehat{D})$.

Why "increase m_1" is not a free substitute. One might attempt to lower ϵ_E in (19) by enlarging m_1 so that $\mathrm{span}(\widehat{D})$ approximates both $\mathrm{span}(D^\star)$ and $\mathrm{span}(E^\star)$. However, for fixed sparsity budget k, this trades one obstruction for another: representing $a = u + v$ with a single k-sparse code requires that the sparse supports corresponding to u and v be jointly expressible within k nonzeros. In particular, under generic position/incoherence assumptions, representing u typically consumes $\Omega(k)$ degrees of freedom already; adding an independent s-sparse component forces either increased sparsity (roughly $k + s$ nonzeros) or increased approximation error. The hierarchical model avoids this collision by allocating distinct sparsity budgets (k, s) to distinct subproblems (first reconstruct u, then reconstruct v from residuals).

An MDL lower bound: support counting and description length. We now make the "not a free substitute" statement quantitative in an MDL language. Consider a family of residual codes y with random supports of size s over m_2 atoms; the number of possible supports is $\binom{m_2}{s}$. A single-stage code $x \in \mathbb{R}^{m_1}$ with $\|x\|_0 \leq k$ can realize at most $\binom{m_1}{k}$ distinct support patterns. If we demand that distinct residual supports induce (with nontrivial

probability) distinguishable reconstructions—a mild nondegeneracy condition when coefficients are continuous and the dictionary is incoherent—then any single-stage scheme that "covers" the residual variability must satisfy the combinatorial constraint

$$\binom{m_1}{k} \gtrsim \binom{m_2}{s}, \tag{20}$$

which in the sparse regime implies (up to logarithmic factors) $k \log m_1 \gtrsim s \log m_2$. This is already a code-length statement: the standard MDL proxy for a k-sparse code has a leading term proportional to $\log \binom{m_1}{k} \approx k \log(m_1/k)$, corresponding to encoding its support. In a hierarchical representation, we encode the stage-1 support and the residual support separately, yielding a total leading term approximately

$$\log \binom{m_1}{k} + \log \binom{m_2}{s} \approx k \log(m_1/k) + s \log(m_2/s), \tag{21}$$

with amplitude quantization terms added similarly per nonzero. In contrast, forcing a single-stage model to capture both components either increases k (raising the first term) or increases m_1 until (20) holds, which raises $\log m_1$ and hence the support-encoding cost. The hierarchical construction therefore admits a Pareto regime: for fixed m_1 (chosen to capture $D^\star$) and moderate m_2 with $s \ll k$, we simultaneously decrease reconstruction error (by fitting $E^\star y$ explicitly) and decrease description length relative to any single-stage alternative that achieves comparable error.

Interpretation for practice. The lower bounds above justify the architectural separation enforced by residual learning: if $E^\star y$ carries nontrivial energy outside the span learned by the stage-1 model, then any single-stage (m_1, k) solution faces either an irreducible error floor (19) or an MDL penalty increase implied by (20). The hierarchy is therefore not an aesthetic choice but the minimal mechanism that (i) removes residual energy and (ii) does so with reusable structure rather than per-example memorization.

9 Results (Expected) and Ablations (Planned)

Residual dictionary interpretability. Our primary qualitative expectation is that the stage-2 dictionary E (trained on residuals $r = a - \widehat{D}\widehat{z}$) yields latents whose semantics are more *specific* and less *entangled* than (i) the highest-error latents in an equivalently sized single-stage SAE and (ii) "SAE+error-node" baselines that append extra latents to absorb reconstruction error without an explicit residual interface. Concretely, we expect many E-atoms to align with directions that are systematically underfit by stage-1 due to sparsity collisions, including (a) rare but high-magnitude

activation bursts, (b) high-frequency "formatting" or delimiter structure in chat-formatted inputs, and (c) safety-relevant concepts that are distributionally shifted between $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{safety}}$. We will operationalize interpretability via (1) top-activating examples and n-gram/metadata enrichment, (2) clusterability of latents by co-activation graphs (e.g. modularity on the latent correlation network), and (3) stability of the learned atoms across random seeds, quantified by nearest-neighbor matching in cosine similarity (up to sign/permutation) for both D and E. Our expected pattern is that E exhibits higher cross-seed stability than an "error-node" augmentation of a single SAE at comparable reconstruction error, because residual training enforces reuse across samples rather than per-example absorption.

Faithfulness gains versus compute and parameter count. We anticipate that hierarchical reconstructions $\hat{a} = Dz + Ey$ improve faithfulness metrics $\text{Faith}(D, E)$ relative to single-stage baselines at matched decoder parameter count $O(d(m_1 + m_2))$, particularly under activation replacement on out-of-distribution inputs (notably $\mathcal{D}_{\text{chat}}$ and $\mathcal{D}_{\text{safety}}$ when trained primarily on $\mathcal{D}_{\text{pre}}$). The core expected phenomenon is that adding m_2 residual atoms yields a larger reduction in downstream loss under replacement than allocating the same additional parameters to increasing m_1, because the former targets the misspecification component while the latter tends to dilute the stage-1 basis and increases feature competition under fixed k. We will report faithfulness as (i) relative change in downstream loss (or logit KL) when replacing activations by reconstructions, and (ii) layerwise sensitivity curves as we vary replacement probability. We will also plot compute–faithfulness Pareto frontiers by varying training steps and dictionary sizes, with the expected ordering

$$\text{HR-SAE} \succ \text{single SAE of matched params} \succ \text{single SAE of matched } m_1$$

on held-out mixtures, where "$\succ$" denotes uniformly better tradeoffs in (replacement loss increase, MDL proxy).

Dataset dependence and cross-mixture generalization. We plan a controlled analysis over mixtures of $\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{chat}}, \mathcal{D}_{\text{safety}}$ both for training and evaluation. Our expectation is asymmetric generalization: (i) D trained on $\mathcal{D}_{\text{pre}}$ remains largely reusable across mixtures (capturing high-variance "core" features), whereas (ii) E is more sensitive to the training mixture and preferentially captures distribution-specific residual structure. We will quantify this by training (D, E) on each mixture and measuring (a) reconstruction error and MDL on each evaluation distribution, (b) faithfulness under replacement on each evaluation distribution, and (c) residual predictability: the fraction of residual variance explained by E on held-out data, i.e. $R^2(r, Ey)$. A key expected diagnostic is that R^2 for residuals transfers poorly when E

is trained on a narrow distribution, even if $\|a - \hat{a}\|_2^2$ remains similar; this gap signals that residual structure is not purely noise but distribution-dependent signal.

Intervention robustness and causal tests. We will evaluate robustness of latent interventions by comparing three procedures: (1) direct latent ablation in stage-1 ($z_j \leftarrow 0$), (2) direct latent ablation in stage-2 ($y_\ell \leftarrow 0$), and (3) matched perturbations in single-stage baselines. We expect interventions on E-latents to exhibit more localized downstream effects (smaller collateral changes in unrelated logits) when E is trained on residuals with a sparsity constraint $s \ll d$, reflecting reduced entanglement. We will additionally test compositionality by activating multiple E-latents and checking approximate additivity of induced changes in downstream loss or logits. A planned negative control is to intervene on randomly rotated latent bases with identical reconstruction error; we expect intervention effects to be less stable under rotations for non-hierarchical baselines.

Planned ablations. We will ablate: (i) residual interface versus no interface (training stage-2 on a rather than r), (ii) conditional residual prediction (CR-SAE: $y = \mathrm{Enc2}(r, z)$) versus unconditional ($y = \mathrm{Enc2}(r)$), (iii) MDL penalty strength β and the separation of λ_1, λ_2, (iv) joint fine-tuning of (D, E) versus freezing D after stage-1, and (v) migration controls (penalties discouraging E from re-learning directions already in $\mathrm{span}(D)$, measured via $\mu(D, E)$ and subspace overlap). Our expectation is that (i) the residual interface is necessary for interpretability of E at fixed m_2, (ii) CR-SAE improves faithfulness on chat/safety distributions by allowing E to specialize to residual modes conditioned on coarse stage-1 context, and (iii) overly small β yields degenerate high-support residual codes that improve $\|a - \hat{a}\|_2^2$ but harm faithfulness and stability.

Negative results and failure modes. We anticipate several regimes where hierarchy does not help. First, when $\mathbb{E}\|E^\star y\|_2^2$ is negligible relative to noise, stage-2 learns near-random atoms and yields no faithfulness improvement beyond variance reduction. Second, if stage-1 is substantially misspecified (large ε), then residuals contain structured leakage from $D^\star z$ and stage-2 may "steal" features, reducing interpretability and increasing $\mu(D, E)$. Third, under severe distribution shift, E may overfit idiosyncratic residual patterns (especially with weak MDL), producing a low reconstruction error but poor replacement faithfulness. We will treat these as first-class outcomes by reporting (a) subspace overlap and cross-coherence diagnostics, (b) seed instability, and (c) gaps between reconstruction metrics and faithfulness metrics, which we take as evidence that the learned codes are not aligned with the model's causal features.

10 Discussion

The hierarchical residual construction is motivated by a simple observation about mechanistic interpretability practice: one seldom needs a *single* globally optimal coordinate system for activations, but rather a coordinate system that (i) captures the highest-leverage, repeatedly reused structure and (ii) exposes the remaining structure in a form that is still reusable, compressible, and intervention-friendly. The two-stage interface $a \approx Dz + Ey$ makes this separation explicit. In particular, the residual channel $r := a - \widehat{D}\widehat{z}$ is not treated as an unstructured error term but as a second object of study, with its own sparsity budget, dictionary, and (crucially) its own inductive bias for reuse across samples.

A direct implication for mechanistic interpretability pipelines is that feature discovery can be organized into *tiers* rather than a single monolithic SAE training run. In the tiered view, stage-1 latents serve as a coarse but stable basis for the "core" activation geometry, while stage-2 latents serve as a refinement basis that targets systematic misspecification modes. This suggests a workflow in which we first fit a conservative, stable D under a stringent code-length constraint, and only then allocate additional capacity to E on the residuals. Conceptually, this resembles building an atlas: D provides a global chart capturing dominant directions, and E adds local coordinates for directions that are rare, distribution-specific, or suppressed by feature competition under fixed k.

For downstream analyses that rely on interventions, the separation is also methodological. If we intervene on z-coordinates, we are perturbing directions that the first-stage encoder/decoder deem globally salient; if we intervene on y-coordinates, we are perturbing structure that is salient *conditional on* the first stage having already explained what it can. This conditionality is useful even when E is learned unconditionally from residuals, because the residual itself depends on the stage-1 explanation. In practice, this allows a more disciplined interpretation of intervention results: changes attributable to E-latents are, by construction, changes that cannot be cheaply represented in the stage-1 code, and therefore are less likely to be artifacts of arbitrary basis choice within $\mathrm{span}(D)$.

The monitoring and control perspective is similar. Many proposed safety monitors are implicitly single-stage: they attempt to compress or classify activations directly. A residualized representation offers an alternative decomposition of the monitoring problem into (i) monitoring the stable, high-coverage representation z for broad behaviors and (ii) monitoring the residual code y for "edge" behaviors that are systematically underexplained by D. Since residual structure is expected to be more sensitive to distribution shift, it is a plausible locus for detecting changes between $\mathcal{D}_{\mathrm{pre}}$ and deployment-time mixtures. In settings where one seeks to *control* behavior by constraining internal states, the hierarchy also suggests a way to trade

off fidelity and constraint strength: one may enforce strict constraints on y (e.g., sparsity, amplitude caps, or outright suppression of certain atoms) while leaving z largely unconstrained, thereby targeting fine-grained modes without broadly disrupting computation.

At the same time, the framework clarifies limits. First, the residual interface does not eliminate identifiability issues; it relocates them. The learned E is only meaningful insofar as (a) stage-1 errors are small enough that r is well-approximated by $E^\star y$ rather than leakage of $D^\star z$, and (b) the residual dictionary satisfies its own incoherence/sparsity conditions. When ε is nontrivial, the term $(D^\star P - \widehat{D})\widehat{z}$ can dominate $\tilde{\xi}$, and the second stage may preferentially model the leaked component. In that regime, E ceases to be a refinement dictionary and becomes a compensator for stage-1 misspecification, which undermines the intended interpretation. Diagnostics such as $\mu(D, E)$ and subspace overlap therefore are not ancillary; they are necessary to justify a hierarchical reading of the learned atoms.

Second, the residual channel does not resolve the fundamental tension between reconstruction and faithfulness. Replacement faithfulness concerns whether $\hat{a}$ preserves the causal role of a in the network, which can fail even at low $\|a - \hat{a}\|_2^2$ when small perturbations occur in high-sensitivity directions. Hierarchical training may improve this by allocating capacity to systematically missed directions, but it cannot guarantee faithfulness without an explicit objective that references the downstream computation. This suggests that the most principled use of the hierarchy is as a *representation* on top of which one can layer faithfulness-sensitive selection or regularization, rather than as a standalone solution.

Geometrically, the hierarchy is a constrained approximation to nonlinear activation structure using linear pieces. Even if the true activation distribution concentrates near a curved manifold, a single global dictionary with a fixed sparsity budget must represent curvature as a collection of competing linear directions. Residualization can be interpreted as an iterative linearization: stage-1 captures a dominant subspace/union-of-subspaces structure, and stage-2 captures directions corresponding to the remaining curvature, higher-order interactions, or context-dependent deviations. This viewpoint motivates conditional residual models (CR-SAEs): allowing y to depend on z is a minimal way to encode nonlinearity, since z functions as a coarse state variable that selects which residual modes are relevant. In other words, conditioning is a proxy for a mixture-of-linear models in which z indexes the mixture component, and E provides the component-wise correction.

Finally, the discussion points toward generalizations. One can iterate the construction to multiple residual stages, or replace the linear residual dictionary by a structured family (e.g., grouped atoms, low-rank blocks, or convolutional structure) to better match known symmetries in activations. One can also couple the hierarchy to explicit code-length accounting, treating $\mathrm{MDL}(z, y)$ as a first-class quantity rather than a proxy regularizer. The

central claim we take forward is modest: when activations decompose into reusable components at different scales of frequency and rarity, enforcing that decomposition via an explicit residual interface is a principled way to improve both interpretability and operational usefulness, while making failure modes visible through coherence, stability, and faithfulness diagnostics.

11 Conclusion and Open Problems

We formalized and studied a hierarchical residual interface for activation decomposition in which a first sparse dictionary captures a reusable "core" component and a second sparse dictionary captures structure that remains systematic after subtracting the first explanation. The central technical point is that, under a two-level sparse generative model and sufficiently accurate first-stage recovery, the residual inherits a sparse structure in an independent residual dictionary up to an effective noise term whose magnitude can be controlled by first-stage estimation errors. This permits a second stage with standard sparse dictionary learning guarantees, and it yields a principled regime in which a two-stage model reduces reconstruction error below the misspecification floor of a single-stage (m_1, k) model while also improving code-length tradeoffs under an MDL-style penalty. Empirically, the resulting HR-SAE/CR-SAE constructions are meant to be usable as modular components in interpretability pipelines: record activations, fit D, residualize, fit E, and then analyze and intervene on z- and y-coordinates separately.

The broader goal suggested by this work is to move from "feature lists" toward mechanism-level decompositions: representations whose components are not only sparse and reusable, but also stable under reasonable changes in data distribution and informative under interventions. The residual hierarchy is one step in this direction because it makes the failure modes legible: if the second stage is forced to explain leaked first-stage structure, then coherence and subspace-overlap diagnostics should detect that the purported refinement basis has become a compensator. In other words, the hierarchy supplies explicit interfaces at which we can test whether we are compressing genuine reusable structure or merely allocating capacity to error correction.

We close with open problems that, in our view, delimit the next technical steps.

Residual hierarchies beyond two stages. The two-stage construction is the simplest nontrivial instance of an iterative residualization scheme. A natural extension is a multi-stage model

$$a \approx D^{(1)}z^{(1)} + D^{(2)}z^{(2)} + \cdots + D^{(T)}z^{(T)}, \qquad r^{(t)} := a - \sum_{\tau \le t} D^{(\tau)}z^{(\tau)},$$

with stage-wise sparsity budgets and code-length accounting. The main theoretical questions are (i) whether error accumulation across stages can be

controlled so that later residuals remain sparse in new dictionaries rather than devolving into modeling earlier leakage, and (ii) whether one can obtain sample-complexity bounds that scale reasonably in T without requiring exponentially stringent coherence conditions. On the algorithmic side, one wants training procedures that prevent "feature migration" (the same direction appearing in multiple stages) while still allowing later stages to represent genuine refinements. This calls for explicit regularizers coupling stages (e.g., penalties on $\mu(D^{(t)}, D^{(\tau)})$ for $\tau < t$) and for stability diagnostics that can be monitored online.

Interaction with faithfulness objectives. Our formal analysis concerns reconstruction and description length, while mechanistic utility depends on faithfulness under interventions (e.g., activation replacement). A concrete open problem is to characterize when minimizing $\|a - \hat{a}\|_2^2$ plus MDL is sufficient to control downstream loss increase, and when it is not. This likely requires sensitivity bounds for the downstream computation (local Lipschitz constants or Jacobian spectra) and a way to weight reconstruction errors by causal importance. A promising direction is to incorporate an explicit faithfulness term into training (possibly at a small number of probe layers) and to understand whether hierarchical capacity allocation reduces the number of faithfulness-sensitive directions that must be tracked.

On-manifold patching and residualization. Recent "on-manifold" interventions aim to restrict patched activations to lie on (or near) the activation manifold, thereby reducing unnatural states. The residual hierarchy suggests a decomposition of this constraint: one may demand that Dz lie on a high-coverage manifold chart while allowing Ey to parameterize deviations within a controlled family. The open question is whether one can formalize an "on-manifold" notion in terms of code constraints (support patterns, amplitude priors, or conditional models) and prove that residual patching in (z, y)-space yields smaller distributional shift than patching in a-space. Technically, this seems to require connecting sparse generative assumptions to manifold curvature and to the geometry of conditional residual distributions.

Cross-layer and cross-module features. A single-layer decomposition is not, by itself, a mechanism decomposition, since mechanisms typically span multiple layers and modules. One needs notions of feature correspondence across layers (or time steps) and tests for whether a latent in one layer predicts, explains, or causally mediates computation in another. The residual viewpoint offers a concrete handle: one can ask whether residual latents y at layer ℓ become core latents z at layer $\ell + 1$, or whether the hierarchy aligns across layers after accounting for known linear maps (e.g., attention

output projections). An open problem is to define and estimate cross-layer "transport" maps between latent spaces that respect sparsity and preserve intervention semantics, and to determine identifiability conditions for such maps under distribution shift.

Conditioning as minimal nonlinearity. Conditional residual models (CR-SAEs) treat y as dependent on z, which is a minimal route to representing mixtures of linear structures. A theoretical gap is to characterize when conditioning is necessary (e.g., when $E^\star$ itself varies with a discrete or continuous context variable) and to bound the additional sample complexity induced by conditioning. One expects a tradeoff: conditioning can improve representation efficiency but can also reduce identifiability by expanding the effective model class. Establishing regimes where conditioning yields provable gains without sacrificing interpretability remains open.

MDL as an operational quantity. We used MDL-style penalties as a conceptual and practical tool to prevent degenerate memorization solutions. A more complete treatment would (i) specify a concrete coding scheme for supports and amplitudes, (ii) relate the resulting description length to generalization across distributions $\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{chat}}, \mathcal{D}_{\text{safety}}$, and (iii) test whether MDL improvements predict faithfulness or monitoring performance. The open problem is to make "code length" not merely a regularizer but a measurable artifact with predictive value for downstream interpretability tasks.

Taken together, these problems point toward a program in which residual hierarchies are not an endpoint but an organizing principle: expose structure stage by stage, attach explicit diagnostics to each interface, and couple reconstruction with objectives that reflect causal use in the network. The residual interface is attractive precisely because it is simple enough to analyze and to instrument, yet expressive enough to capture repeated structure beyond what a single sparse code can represent at fixed budget.

12 12. Conclusion and Open Problems: toward mechanism-level decompositions; residual hierarchies beyond two stages; interaction with on-manifold patching and cross-layer features.

We regard the hierarchical residual decomposition as an interface specification rather than merely an estimator: it imposes an explicit contract between stages ("Stage 1 explains what it can with a sparse reusable code; Stage 2 is permitted to spend capacity only on what remains") and thereby makes several otherwise latent ambiguities testable. In particular, it becomes meaningful to ask whether a learned refinement dictionary is expressing a

genuinely new set of sparse directions, or whether it is functioning as an error-corrector for systematic defects of the first stage. This distinction is central if one aims for mechanism-level decompositions, where latent coordinates should support stable interpretation and predictable interventions. The present guarantees only address reconstruction and (via MDL) a coarse notion of compressibility; the open questions below concern the additional structure required to turn a good compressor into a good mechanistic coordinate system.

Beyond two stages: controlling accumulation and preventing degeneracy. An obvious extension is a T-stage residual cascade with dictionaries $D^{(1)}, \dots, D^{(T)}$ and residuals

$$r^{(t)} \;:=\; a - \sum_{\tau=1}^{t} D^{(\tau)} z^{(\tau)}, \qquad \|z^{(t)}\|_0 \le k_t.$$

Even under a multi-level generative model $a = \sum_{t=1}^{T} D^{(t)\star} z^{(t)\star} + \xi$, the difficulty is that estimation errors compound: the effective noise injected into $r^{(t)}$ depends on all earlier dictionary and code errors. A concrete theoretical problem is to prove a stage-wise analogue of Theorem 1 with an error recursion of the form

$$r^{(t)} \;=\; D^{(t+1)\star} z^{(t+1)\star} + \tilde{\xi}^{(t)}, \qquad \mathbb{E}\|\tilde{\xi}^{(t)}\|_2^2 \;\le\; \sigma^2 + \sum_{\tau \le t} \mathrm{Err}_\tau,$$

where each Err_τ scales in a controlled way with $\|\widehat{D}^{(\tau)} - D^{(\tau)\star}\|$ and moments of $\widehat{z}^{(\tau)}$. One would like conditions under which $\sum_{\tau \le t} \mathrm{Err}_\tau$ remains bounded for moderate t, rather than forcing coherence assumptions that become exponentially stringent in T.

On the algorithmic side, multi-stage models create new degeneracies that are absent in the two-stage case: "feature migration" (the same direction reappearing across stages) and "stage collapse" (later stages learning to undo or renormalize earlier reconstructions). A principled open problem is to define constraints or penalties that enforce a meaningful factorization without sacrificing fit. Candidates include explicit cross-stage incoherence control (penalizing $\mu(D^{(t)}, D^{(\tau)})$), orthogonality-on-average constraints under the activation distribution, or MDL budgets that are allocated per stage (so that late stages cannot cheaply encode what earlier stages failed to encode). A satisfactory theory would relate such regularizers to identifiability and to empirical stability of learned latents.

Mechanism-level decompositions: stability and intervention semantics. A mechanistic representation should be stable under benign distribution shift and should support interventions with predictable effects.

This suggests metrics beyond reconstruction, for instance: (i) support stability of z and y under changes from $\mathcal{D}_{\text{pre}}$ to $\mathcal{D}_{\text{chat}}$, (ii) invariance of latent-to-downstream influence, and (iii) agreement of latent correspondences across training runs up to permutation/sign. An open theoretical direction is to connect these desiderata to properties of the generative model (support randomness, separation, conditional independence) and to properties of the underlying network (local linearity of the mapping from activations to loss). Concretely, one would like conditions ensuring that the mapping $a \mapsto (z(a), y(a))$ is not only sparse but also locally Lipschitz on the data manifold, which would rule out encoders that are compressive yet fragile.

On-manifold patching: residual coordinates as a controllable deviation family. When one replaces an activation a by a reconstruction $\hat{a}$, one risks producing off-manifold states even if $\|a-\hat{a}\|_2$ is small. The residual hierarchy suggests a decomposition of this risk: the stage-1 component Dz may be viewed as a coarse chart of typical activations, while the residual component Ey parameterizes a restricted family of deviations. A concrete open problem is to formalize an "on-manifold" constraint in terms of code distributions. For example, if (z, y) are modeled by a prior (or conditional prior) $p(z)p(y \mid z)$, then one can ask for bounds of the form

$$\mathrm{TV}\big(\mathcal{L}(a), \mathcal{L}(\hat{a})\big) \;\le\; f\big(\mathbb{E}L(a,\hat{a}),\; \mathrm{KL}\big(\mathcal{L}(z,y) \,\|\, p(z)p(y \mid z)\big)\big)\,,$$

where $\mathcal{L}$ denotes law under the data distribution. Establishing such bounds would connect reconstruction and code regularity to distributional shift under patching. It would also motivate training objectives that explicitly fit the empirical code distribution (e.g., via a tractable conditional model for y given z) so that patched codes can be sampled from a learned on-manifold surrogate rather than chosen adversarially.

Cross-layer and cross-module structure: transport maps between latent spaces. Single-layer decompositions are insufficient if mechanisms are distributed across layers. The residual framework introduces a natural family of questions about correspondence: if (z_ℓ, y_ℓ) decompose layer-ℓ activations, when does $z_{\ell+1}$ depend sparsely on z_ℓ (core-to-core propagation), and when do residual latents become core latents at the next layer (refinements becoming canonical)? One formal approach is to posit a sparse transport model

$$z_{\ell+1} \;\approx\; T_\ell z_\ell + U_\ell y_\ell,$$

with T_ℓ, U_ℓ structured (e.g., sparse, block-sparse, or low-rank), and to study identifiability of (T_ℓ, U_ℓ) jointly with dictionaries across layers. The open problem is to define transport estimation procedures that are invariant to the permutation/sign symmetries within each layer while remaining sensitive to genuine mechanistic alignment.

A related challenge is robustness under distribution shift: correspondences learned on $\mathcal{D}_{\text{pre}}$ may fail on $\mathcal{D}_{\text{chat}}$ if new circuits activate. The hierarchy suggests a diagnostic: if the mapping from early-layer core latents to later-layer residual latents grows in magnitude under shift, then the decomposition is witnessing new computation not captured by the pretraining-aligned core. Formalizing such diagnostics requires connecting changes in latent usage statistics to changes in downstream behavior, ideally with guarantees that are not artifacts of encoder nonuniqueness.

Summary. The residual hierarchy offers a concrete scaffold for separating reusable structure from structured remainder, but the main conceptual work remains: to turn this scaffold into a representation that is stable, cross-layer coherent, and safe to intervene upon. Advancing the theory will likely require importing tools from sparse recovery with model mismatch, representation alignment under symmetries, and distribution-shift control via generative modeling of codes.