Subspace Clustering and the Representational Utility of Quanta Fingerprints

Anonymous ACL submission

Abstract

The Quantization Model of Neural Scaling proposes that a language model's capabilities decompose into many small, discrete chunks known as quanta, and that these quanta can be discovered by clustering the per-token gradients of a trained model. Existing work identifies clusters with a single clustering algorithm and judges them only by whether their size distribution follows the predicted power law. We argue that this is insufficient: a clustering method should be evaluated both on its power-law fit, namely how well it recovers the predicted size distribution, and on its representational utility, namely how useful the resulting clusters are as document-level features. We introduce a controlled evaluation protocol and compare four clustering paradigms, including sparse subspace clustering, across two model sizes. The two criteria yield different rankings: on Pythia-125M, spectral clustering achieves the closest power-law fit, yet sparse subspace clustering (SSC-Lasso) produces the strongest document-level representations, outperforming spectral clustering by $+0.100$ in MCC on an AI-versus-human text classification task. Power-law fit alone is thus not a reliable criterion for selecting a quanta-discovery method; evaluating representational utility instead surfaces sparse subspace clustering as the strongest paradigm on both model sizes. Code is available at https://anonymous.4open.science/r/quanta-subspace-clustering-25B0/.

1 Introduction

Understanding why larger neural networks generalize better remains a central open question in deep learning. Michaud et al. (2023) proposed the *Quantization Model of Neural Scaling*, arguing that a network's capability is the sum of discrete skill circuits (*quanta*) whose sizes follow a power law: a handful of high-frequency quanta appear in almost every training sample, while rare quanta contribute incrementally as the model scales. Their method quanta discovery from gradients (QDG), applied to Pythia language models (Biderman et al., 2023), measures a rank-frequency slope of -1.237, broadly consistent with the power law their theory predicts. The authors suggest that "less naive clustering schemes [...] could be useful to sharpen this measurement" and describe their own QDG procedure as "neither very principled nor scalable" (Michaud et al., 2023), yet they evaluate only spectral clustering without comparing alternatives. Still, this gradient-based view has influenced recent parameter decomposition methods (Braun et al., 2025), which aims to identify functional components among a model's parameters. Here, we instead stay within the original token-clustering formulation and study the effect of the choice of clustering algorithm.

We argue that evaluating quanta discovery requires two complementary criteria. The *power-law fit* criterion measures how closely a paradigm's recovered rank frequency slope matches the power-law exponent $\beta^\star = -1.237$ reported by Michaud et al. (2023). The *representational utility* criterion instead measures how useful the resulting clusters are as document-level features, quantified by the discriminative performance of quanta fingerprints on a downstream classification task. If different clustering algorithms partition gradient space differently, any conclusion drawn from those quanta may reflect the algorithm rather than the model. Our goal is to determine whether these two criteria agree in practice, or whether a paradigm that achieves good power-law fit can still fail on representational utility, and vice versa.

To address this, we consider sparse subspace clustering (SSC), which formulates a global self-expression problem, whereby each point is reconstructed as a sparse linear combination of all others. We hypothesise that SSC will recover quanta at least as well as spectral clustering on power-law fit,

and that paradigm choice will substantially affect representational utility.

Replacing spectral clustering is non-trivial: SSC introduces hyperparameters (PCA dimension d and regularisation α) whose interaction with gradient-similarity geometry is unknown a priori (Elhamifar and Vidal, 2013), and a fair comparison requires a controlled protocol, since naive aggregation across configurations distorts the power-law slope and conflates method differences with sweep artefacts.

This work makes three contributions. First, we **replicate** the quanta-discovery findings of Michaud et al. (2023) on Pythia-19M and Pythia-125M, establishing a reproducible baseline with an envelope methodology. Second, we **compare** four clustering paradigms and evaluate three of them (Spectral, SSC-Lasso, Hierarchical) in a downstream task on representational utility: SSC-Lasso achieves the tightest envelope fit on Pythia-19M ($|\Delta| = 0.008$) while Spectral retains a slight edge on Pythia-125M ($|\Delta| = 0.004$). Third, and most importantly, we **demonstrate** that a better power-law fit does not entail higher representational utility: SSC-Lasso fingerprints outperform every evaluated paradigm on AI-versus-human text classification (MCC 0.884 on Pythia-19M, 0.960 on Pythia-125M), even though spectral clustering achieves the better envelope fit on the larger model.

2 Background

The Quantization Model Michaud et al. (2023) define a *quantum* as a set of tokens whose gradient vectors are highly similar: they activate the same internal circuit and improve together as training progresses. Their QDG pipeline proceeds in three stages. (1) A gradient-similarity matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ is computed, where C_{ij} is the cosine similarity between the per-sample gradients of tokens i and j. (2) The matrix is transformed to angular distance and clustered into $k = 400$ groups. (3) The rank-frequency distribution of cluster sizes is fit by a power law. Michaud et al. (2023) report a slope of -1.237 for Pythia-19M. The structural prediction is that this slope should be consistent across model sizes, grounding the empirical regularity of neural scaling laws (Kaplan et al., 2020) in a mechanistic hypothesis about discrete skill acquisition. Reproducing this slope is therefore both a verification of the theory and a prerequisite for any downstream analysis.

3 Methodology

We compare four clustering paradigms (§3.1–§3.2) on a shared set of per-token gradients (§3.3), and evaluate each under two criteria: power-law fit (§4) and representational utility (§5).

3.1 Quanta via Sparse Subspace Clustering

Subspace clustering methods model data as drawn from a union of low-dimensional subspaces (Parsons et al., 2004; Kriegel et al., 2009; Houle et al., 2023). Axis-parallel methods are poorly suited to gradient-similarity vectors, which are unlikely to align with coordinate axes: the relevant subspaces are arbitrarily oriented. Two traditions target such subspaces: *correlation-based* methods (e.g. 4C, ORCLUS, COPAC; Kriegel et al. 2009), which recover locally correlated clusters through neighbourhood eigenanalysis; and *self-expression* methods, which formulate subspace recovery as a global sparse-reconstruction problem.

We adopt sparse subspace clustering (SSC; Elhamifar and Vidal 2013) as a representative of the self-expression family: rather than relying on local correlation structure, it reconstructs each token from the others under a global sparsity penalty, which we transfer to gradient clustering by solving the following self-expression per token $\mathbf{y}_i$:

$$\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_1 \quad \text{s.t.} \quad \mathbf{y}_i = \mathbf{Y}_{\setminus i}\, \mathbf{c}_i, \tag{1}$$

where $\mathbf{Y}_{\setminus i}$ is the data matrix with column i removed. Intuitively, a token whose gradient lies in a d-dimensional subspace can be reconstructed from as few as d other tokens of that same subspace, whereas drawing on tokens from other subspaces would require additional non-zero coefficients. Minimizing $\|\mathbf{c}_i\|_1$ therefore favors within-subspace reconstructions. The sparse coefficients encode subspace membership, and, under mild constraints, non-zero coefficients appear only between points on the same subspace. The affinity matrix $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^\top$ is then passed to spectral clustering. If quanta span structured subspaces of gradient space, the self-expression coefficients should be sparse and intra-subspace, making SSC a principled fit for quanta discovery that recovers subspace structure through the affinity rather than optimising local graph cuts on raw similarity. We also evaluate SSC-OMP, which replaces the ℓ_1 penalty with orthogonal matching pursuit: rather than penalizing coefficient magnitude, it greedily selects the most correlated basis vector at each step, yielding sparser

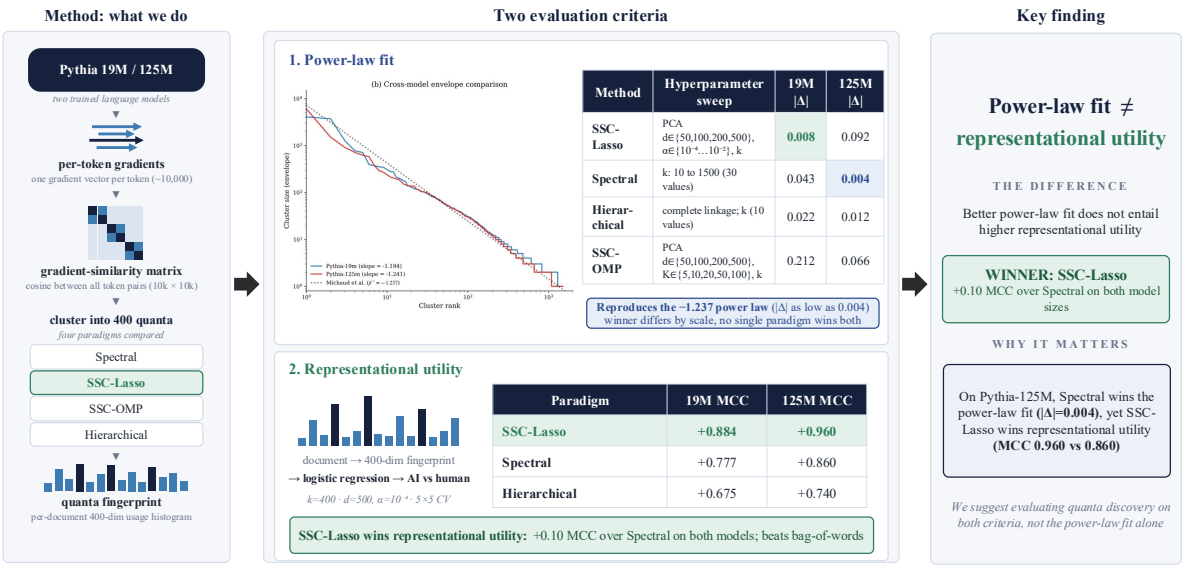

Figure 1: Per-token gradients from Pythia-19M/125M are clustered into 400 quanta by four paradigms (left) and evaluated under two criteria (centre): envelope fit against the target slope $\beta^* = -1.237$ of Michaud et al. (2023), and representational utility of quanta fingerprints on AI-versus-human classification. Spectral wins on power-law fit on Pythia-125M, while SSC-Lasso wins on representational utility on both models.

solutions at lower computational cost but without the convex recovery guarantees of the Lasso.

3.2 Clustering Paradigms

We evaluate four clustering paradigms representing distinct algorithmic families on the same gradient-similarity data. **Spectral** clustering (normalised-cut, von Luxburg 2007) on the angular-similarity matrix serves as the direct baseline, replicating the method of Michaud et al. (2023), with k swept over 30 values from $k = 10$ to $k = 1500$. **SSC-Lasso** (SSC with ℓ_1 self-expression, Eq. 1) is the primary subspace method under investigation, motivated by its global self-expression formulation and convex recovery guarantees. We sweep $d \in \{50, 100, 200, 500\}$ (four values) and α over five log-spaced values from 10^{-4} to 10^{-2}, and k for each of the resulting 20 (d, α) pairs. **SSC-OMP** (You et al., 2016) (orthogonal matching pursuit, $d \in \{50, 100, 200, 500\}$, sparsity $K \in \{5, 10, 20, 50, 100\}$) is included as a greedy alternative within the same subspace family, isolating the effect of the ℓ_1 penalty. **Hierarchical** clustering (complete linkage, Müllner 2011) provides a tree-based reference with no subspace assumption, sweeping k over 10 values within the same $k = 10$ to $k = 1500$ range.

Both SSC variants share the same second stage: they cluster the affinity $\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^T$ induced by their self-expression coefficients, using spectral clustering as a final step. This spectral step is not what distinguishes them from the Spectral baseline. The baseline applies spectral clustering directly to the angular-similarity matrix, so tokens are linked by raw gradient similarity; the SSC variants instead link tokens when one reconstructs the other from within a shared subspace, and differ from each other only in how those reconstruction coefficients are computed — convex $\ell 1$ minimisation (SSC-Lasso) versus greedy orthogonal matching pursuit (SSC-OMP). The subspace assumption therefore enters through the affinity, not the clustering.

3.3 Experimental Setup

We use the Pythia model suite (Biderman et al., 2023), evaluating Pythia-19M and Pythia-125M at training checkpoint 143,000.[1] Both models are trained on The Pile (Gao et al., 2021), an 800 GB corpus of diverse English text. Following Michaud et al. (2023), we select 10,000 tokens on which the model achieves near-zero cross-entropy loss (< 0.1 nats), compute per-token gradients, and construct the cosine similarity matrix. Using correctly predicted tokens ensures the gradient signal reflects internal representations. Each token is thus represented by its gradient in parameter space: the points

[1]We follow the naming convention of Michaud et al. (2023): Pythia-19M and Pythia-125M correspond to the current EleutherAI releases Pythia-70M-v0 and Pythia-160M-v0, respectively.

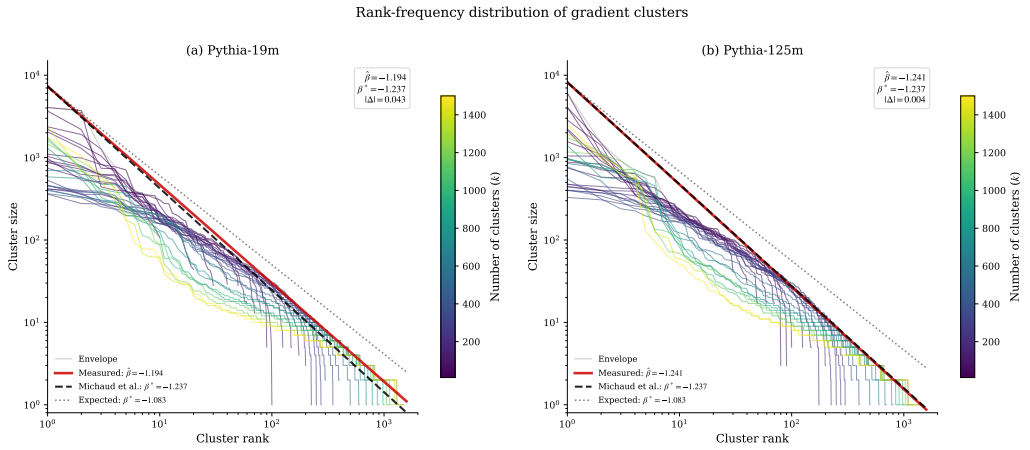

Figure 2: Rank-frequency envelope for spectral clustering on Pythia-19M and Pythia-125M, sweeping 30 cluster counts from $k = 10$ to $k = 1500$ (colours). The envelope slope (fit over the full k-sweep) is -1.194 (19M) and -1.241 (125M), matching the target slope $\beta^* = -1.237$ of Michaud et al. (2023). The dotted reference line (-1.083) is the finite-sample corrected expectation for $n = 10{,}000$ tokens and $k = 400$ clusters, included as a lower bound on what tail truncation alone would predict.

we cluster are tokens, the axes are (a PCA projection of) the model parameters, and gradient similarity serves as a proxy for two tokens recruiting the same circuit (rather than a direct measurement).

Tokens are drawn from consecutive sequences in The Pile test split; no explicit filtering of special tokens or whitespace is applied beyond the zero-loss threshold, which implicitly excludes most trivially predicted tokens. The token index set is constructed once, following the original pipeline design, and shared across model sizes; gradients are then computed per model on this shared set. Because cross-entropy values differ across model sizes, the near-zero-loss criterion is not guaranteed to hold identically for both models. Within each model, the same index set is used across all clustering methods, so method comparisons are not confounded by token-selection differences.

4 Power-Law Fit

4.1 Envelope Protocol

Comparing clustering methods fairly requires accounting for their different hyperparameters (k, PCA dimension d, regularisation α, linkage criterion, etc.). A naive comparison at fixed $k = 400$ would conflate hyperparameter sensitivity with paradigm quality. For each method we therefore sweep its admissible hyperparameter settings and, at each rank, record the *maximum cluster size* across the sweep. We then fit a log-log slope to ranks 100–1000 following Michaud et al. (2023) and report slope $\hat{\beta}$ and $|\Delta| = |\hat{\beta} - \beta^*|$, where $\beta^* = -1.237$ is their reported reference slope. We use this log-log slope as a descriptive comparison statistic rather than a formal estimate of a power-law exponent, for which maximum-likelihood fitting with goodness-of-fit testing would be required (Clauset et al., 2009). The rank window $[100, 1000]$ likewise avoids the largest clusters (dominated by high-frequency function words) and the smallest (dominated by discretisation noise). This procedure is applied identically to all methods, ensuring each paradigm is evaluated at its best achievable rank-frequency curve[2] The max-envelope asks whether a paradigm can recover the power law under some admissible setting, a capacity rather than average-case measure.

Because SSC-Lasso has a larger hyperparameter space (20 (d, α) pairs) than Spectral or Hierarchical, we evaluate every paradigm at its own best configuration, following standard practice in tuned comparisons. The winning SSC-Lasso configuration $(d = 500, \alpha = 0.0001)$ on Pythia-19M achieves $|\Delta| = 0.008$ on its own k-sweep, so the envelope win is not an artefact of combining multiple configurations, though it remains hyperparameter-sensitive: only the tuned (d, α) configuration reaches the target slope.

[2]We do not use Silhouette coefficient or other internal cluster-validation indices as quality measures. Such indices reward compact, well-separated, roughly equal-density clusters, which is the opposite of the long-tailed power-law distribution the Quantization Model predicts.

Method	Pythia-19M		Pythia-125M	
	$\hat{\beta}$	$\lvert\Delta\rvert$	$\hat{\beta}$	$\lvert\Delta\rvert$
SSC-Lasso	$\mathbf{-1.245}$	$\mathbf{0.008}$	-1.329	0.092
Hierarchical	-1.215	0.022	-1.249	0.012
Spectral	-1.194	0.043	$\mathbf{-1.241}$	$\mathbf{0.004}$
SSC-OMP	-1.449	0.212	-1.303	0.066

Table 1: Envelope slopes and absolute deviations from the target slope of Michaud et al. (2023) ($\beta^* = -1.237$). Best result per model in **bold**.

4.2 Reproducing the Quanta Power Law

Figure 2 shows the rank-frequency envelope obtained by sweeping k for spectral clustering on both models. The recovered slopes are -1.194 (Pythia-19M) and -1.241 (Pythia-125M), matching the target $\beta^* = -1.237$ and within the uncertainty of at least ± 0.2 stated by Michaud et al. (2023). Figure 3 shows the expected block structure in the gradient-similarity matrix. Together, these results indicate that the QDG can be replicated on both model sizes, establishing the baseline for the systematic comparison that follows.

4.3 Envelope Fit Across Clustering Methods

Table 1 summarizes envelope slopes for all four methods on both models. On Pythia-19M, SSC-Lasso achieves the closest match to the target of Michaud et al. (2023) ($|\Delta| = 0.008$), followed by Hierarchical (0.022) and Spectral (0.043). The optimal SSC-Lasso configuration is $d = 500$, $\alpha = 0.0001$. On Pythia-125M the leading method changes: Spectral now wins ($|\Delta| = 0.004$), followed by Hierarchical (0.012); SSC-Lasso degrades to $|\Delta| = 0.092$.

The deviations in Table 1 vary substantially across methods and models. We report slopes as point estimates, since the envelope is a max-aggregate rather than a sampling distribution; token-level bootstrap resampling (Appendix B) indicates slope variability of order ± 0.03, comparable to the $|\Delta|$ gaps between the closest methods and well within the ± 0.2 uncertainty stated by Michaud et al. (2023). The fine-grained ranking is therefore indicative rather than definitive.

4.4 Hyperparameter Sensitivity and Stability

Per-method envelope curves and the SSC-Lasso hyperparameter landscapes for both models are in Appendix A. The optimal region on Pythia-19M is $d = 500$, $\alpha = 0.0001$, and on Pythia-125M it shifts to $d = 200$, $\alpha = 0.01$, consistent with the interpretation that larger models have more compact gradient subspaces, though this remains a post-hoc explanation of the hyperparameter shift.

Bootstrap stability across the three paradigms is reported in Appendix B. Normalised mutual information (NMI) is consistently high across methods (0.67–0.77), indicating stable macro structure. The adjusted Rand index (ARI) varies substantially: Hierarchical is most stable (0.36–0.44), SSC-Lasso moderate (0.33–0.38), and Spectral notably lower (0.12–0.13).

Envelope fit measures whether a clustering paradigm recovers the predicted rank-frequency geometry, but it does not test whether the resulting clusters carry per-quantum identity sharp enough to be useful as document representations. We therefore ask whether quanta fingerprints (normalised per-cluster usage distributions) can discriminate AI-generated from human-written text, and whether that capacity depends on the clustering paradigm.

5 Representational Utility

The previous section established how the paradigms perform on the power-law fit criterion. We now evaluate the representational utility criterion, operationalized as whether the discovered clusters carry sufficient per-quantum identity to be useful as document representations. Each document's quanta fingerprint (Section 5.1) encodes which gradient-derived clusters its tokens activate, and paradigm choice determines how these 400-dimensional vectors are constructed and whether they carry discriminative per-cluster structure or merely aggregate statistics. We show representational utility on same-model AI-versus-human text discrimination as a proof-of-concept.

5.1 Quanta Fingerprints

A document's *quanta fingerprint* is the normalised histogram of cluster assignments over its constituent tokens: each token mapped to cluster c contributes $1/|\text{doc}|$ to dimension c, yielding a 400-dimensional unit probability vector.

For the representational evaluation, we classify AI-generated versus human-written documents. Human documents are drawn from The Pile test split ($n = 784$). AI-generated text is produced by the same Pythia model under evaluation: we sample 600 documents from The Pile test prompts using pure multinomial sampling (temperature 1.0, no top-k truncation, at most 256 new tokens). Doc-

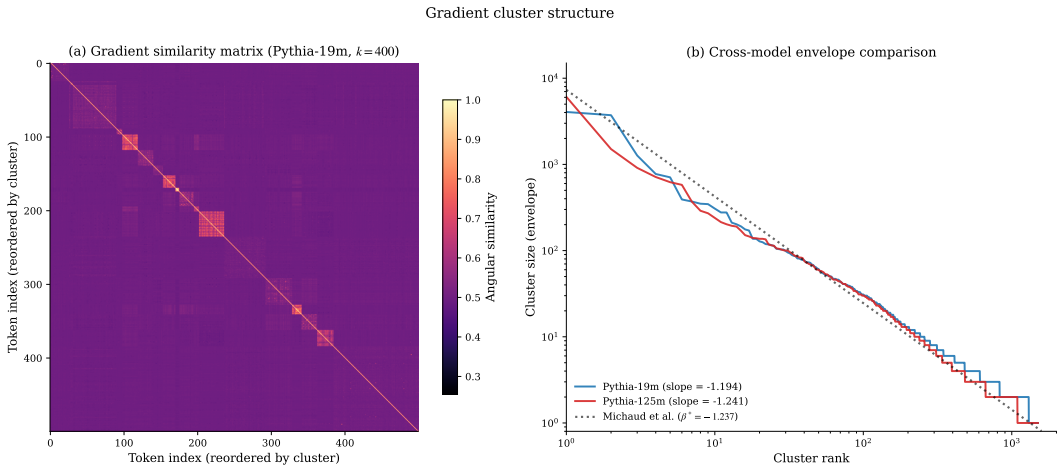

Figure 3: Gradient cluster structure. **(a)** Token-level gradient-similarity matrix for Pythia-19M (first 500 tokens reordered by cluster label at $k = 400$; entries are pairwise angular similarities between individual gradient vectors). **(b)** Cross-model envelope comparison against the target slope of Michaud et al. (2023) ($\beta^* = -1.237$).

uments yielding no zero-loss tokens are discarded, and the same pool of 600 AI documents is shared across all clustering methods. After filtering for documents with at least three zero-loss tokens assigned to a cluster, 497–498 AI documents are retained across the three paradigms (Spectral, SSC-Lasso, and Hierarchical). Classification uses logistic regression (Pedregosa et al., 2011) with stratified 5-fold cross-validation repeated over 5 random seeds (0–4) for fold assignment; reported MCC is the mean over all 25 evaluations.

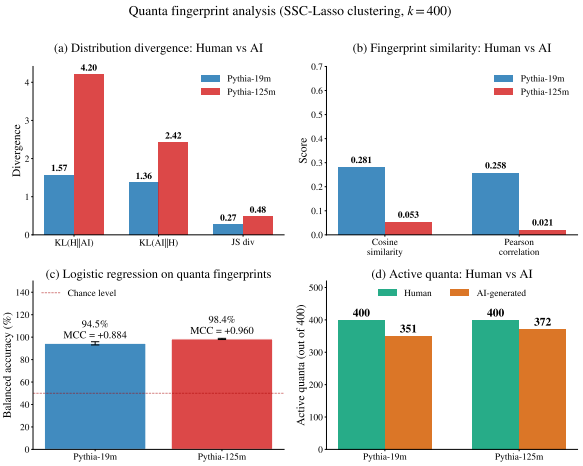

Figure 4: Quanta fingerprint analysis (SSC-Lasso). (a) KL and JS divergence between the mean AI and mean human fingerprint vectors. (b) Cross-group fingerprint similarity (cosine similarity and Pearson correlation). (c) SSC-Lasso logistic regression balanced accuracy and MCC on Pythia-19M and Pythia-125M. (d) Number of clusters with non-zero aggregate usage; human documents activate all 400 quanta, while AI-generated text leaves some unused.

	Pythia-19M		Pythia-125M	
Paradigm	Bal. acc.	MCC	Bal. acc.	MCC
TF-IDF	0.897	+0.791	0.891	+0.781
Spectral	0.901	+0.777	0.938	+0.860
Hierarchical	0.830	+0.675	0.873	+0.740
SSC-Lasso	**0.945**	**+0.884**	**0.984**	**+0.960**

Table 2: AI-versus-human classification performance. Bal. acc. is balanced accuracy from single-seed 5-fold CV; MCC is the mean Matthews correlation coefficient across 5 seeds × 5-fold CV. Best result per model in **bold**. TF-IDF is a bag-of-words (BoW) baseline; Spectral is the QDG baseline of Michaud et al. (2023); Hierarchical uses complete linkage. All paradigms use the same pool of 600 generated AI documents, of which $n_{\text{AI}} \approx 497$ are retained after filtering.

5.2 AI-versus-Human Text Classification

Using the experimental protocol described in Section 5.1, we evaluate Spectral, SSC-Lasso, and Hierarchical on AI-versus-human text classification. SSC-Lasso leads on both models, improving over Spectral by +0.107 MCC on Pythia-19M and +0.100 MCC on Pythia-125M. The ranking SSC-Lasso > Spectral > Hierarchical holds on both models. Table 2 reports MCC for all evaluated paradigm × model combinations.

This confirms that the representational utility ranking differs from the envelope ranking: SSC-Lasso is not the best envelope method on Pythia-125M, but yields the strongest quanta fingerprint representation on both models (Figure 4). The Pythia-125M gain is qualified by the catch-all structure analyzed in Section 5.4.

5.3 How Do Quanta Fingerprints Differ?

Why might paradigm choice affect fingerprint utility so substantially? The token-category composition of clusters offers a structural explanation. Figure 5 shows the token-category distribution at $k = 400$ across the three paradigms. SSC-Lasso concentrates the largest share of tokens in Content clusters (77%/89% on Pythia-19M/125M), the highest of any paradigm; Spectral is intermediate (66%/68%) and Hierarchical the lowest (56%/60%), distributing correspondingly more of their tokens across surface categories such as punctuation and formatting (shown individually in Figure 12). The cross-model category correlation under SSC-Lasso is $r = 0.997$, indicating stable high-level structure across model sizes (Figure 12). Individual SSC-Lasso clusters are themselves linguistically coherent, each dominated by a recognisable syntactic or discourse function (Appendix E, Table 7). The Content-concentration ordering across paradigms (SSC-Lasso $>$ Spectral $>$ Hierarchical) matches their fingerprint MCC ordering (Table 2).

5.4 Interpreting the Fingerprint Signal

The SSC-Lasso catch-all cluster and its surface interpretation On Pythia-125M, SSC-Lasso assigns 74.9% of the 10,000 gradient tokens to a single cluster (cluster 1), which is active in 98% of human documents but only 6% of model-generated documents. Inspection of the token composition reveals that cluster 1 is dominated by punctuation and formatting tokens (Table 4, Appendix C).

This composition indicates that cluster 1 is a formatting catch-all rather than a coherent skill-circuit quantum: its discriminative signal (MCC $+0.950$ alone; Table 5) most plausibly reflects surface structural differences between human-authored Pile text and Pythia-sampled text, such as higher punctuation density and more varied newline structure, rather than mechanistically meaningful gradient-space geometry. Removing cluster 1 and renormalising the remaining 399-dimensional fingerprints to sum to one still yields MCC $+0.910$, indicating that substantial discriminative signal exists in the remaining clusters beyond this surface-driven component. This renormalised 399-cluster fingerprint exceeds the matched TF-IDF bag-of-words baseline (MCC $+0.781$; Table 6), confirming that the residual signal is not reducible to word-frequency patterns alone. No analogous catch-all structure emerges under Spectral or Hierarchical clustering, consistent with their substantially lower Max 1-q MCC values (Table 5, Appendix C).

Per-cluster signal beyond surface statistics Hierarchical clustering ranks second in envelope fit on both models (Table 1) yet ranks last in fingerprint MCC on both models (MCC $+0.675$ on 19M, $+0.740$ on 125M). Its robustness analysis (Table 5) confirms that surface features account for a substantial share of Hierarchical's signal ($\Delta = 0.068$ on 19M, 0.092 on 125M); per-cluster identity does contribute, but weakly relative to Spectral ($\Delta = 0.193$/0.239) and SSC-Lasso on 19M ($\Delta = 0.224$). On Pythia-19M, where no catch-all cluster emerges, SSC-Lasso's self-expression objective produces clusters with substantially more discriminative per-cluster structure ($\Delta = 0.224$ vs. Spectral's 0.193), confirming subspace signal beyond surface statistics.

6 Discussion

Better power-law fit does not entail higher representational utility The central finding of this work is that power-law fit does not imply higher representational utility across clustering paradigms. The clearest evidence is on Pythia-125M, where Spectral achieves the tightest envelope fit ($|\Delta| = 0.004$) yet SSC-Lasso produces the strongest quanta fingerprint classifier (MCC $+0.960$ vs. Spectral's $+0.860$). This gap persists even after removing the dominant formatting cluster ($+0.910$ vs. $+0.860$, Section 5.4), so it does not reduce to surface structure. Hierarchical ranks second in envelope fit on both models yet last in fingerprint MCC on both models ($+0.675$/$+0.740$ vs. SSC-Lasso's $+0.884$/$+0.960$), consistent with complete-linkage agglomeration producing high-density clusters with limited per-cluster discriminative identity. Taken together, these results show that a clustering paradigm can reproduce the target slope while generating clusters with limited representational utility, a failure mode invisible to envelope evaluation alone. Quanta discovery pipelines should therefore be evaluated under both criteria.

Envelope fit is a noisy target In a controlled experiment where the true power-law exponent was known by construction, the pipeline of Michaud et al. (2023) recovered a slope of roughly -1.1 rather than the imposed -1.4, which is off by 0.3 compared to the known underlying distribution, cautioning against selecting on envelope fit alone.

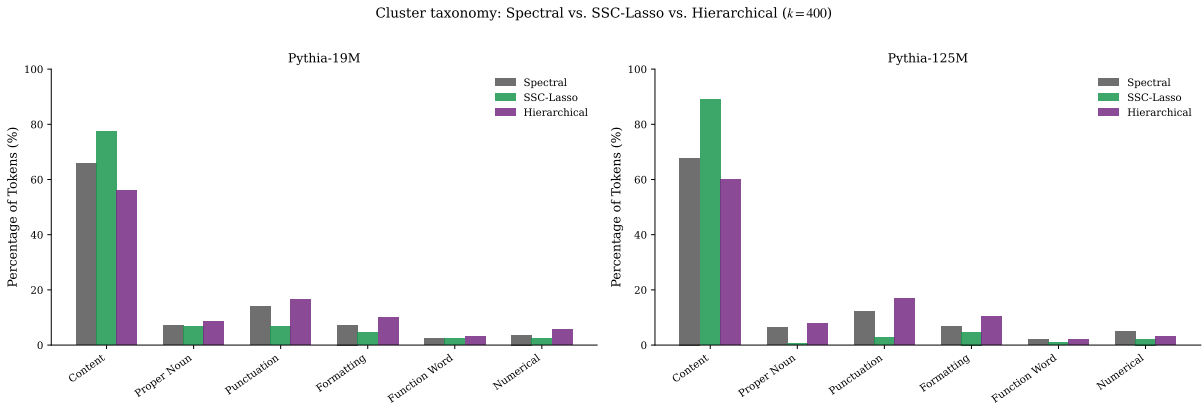

Figure 5: Cluster taxonomy comparison across three paradigms at $k = 400$ for Pythia-19M (left) and Pythia-125M (right). SSC-Lasso concentrates 77%/89% of tokens in the Content category (Pythia-19M/125M), the highest of any paradigm; Spectral is intermediate (66%/68%), and Hierarchical the lowest (56%/60%), placing correspondingly more tokens in surface categories such as punctuation and formatting.

Subspace geometry across model scales The SSC self-expression objective assumes each point lies on a linear subspace spanned by its clustermates. On Pythia-19M this assumption is well supported: the optimal configuration uses $d = 500$ and $\alpha = 0.0001$, and the recovered affinity graph has sufficient connectivity to yield both a tighter envelope fit than Spectral ($|\Delta| = 0.008$ vs. 0.043) and substantially stronger fingerprint performance ($+0.884$ vs. $+0.777$). On Pythia-125M, gradient subspaces appear more compact: the optimal SSC-Lasso configuration shifts to lower PCA dimensionality ($d = 200$) and heavier regularisation ($\alpha = 0.01$), producing affinity graphs so sparse that they frequently degenerate under bootstrap subsampling. The stability analysis (Table 3) therefore holds the configuration fixed at ($d = 500$, $\alpha = 0.0001$) for both models, where Pythia-125M is less stable than Pythia-19M but still more stable than Spectral. That the envelope-optimal Pythia-125M configuration cannot be resampled robustly is itself evidence of a tension between SSC's sparsity bias and the denser correlation structure of larger models, which likely underlies the reversal in envelope ranking between scales. SSC-OMP's markedly weaker envelope fit on both models (Table 1) suggests that the convex ℓ_1 relaxation, rather than sparsity per se, drives the subspace advantage: greedy support selection lacks the recovery guarantees that make the Lasso coefficients respect subspace membership.

Stability, scale, and generalisation A separate caveat applies across scales. Cross-model cluster agreement is low for every method (ARI 0.017–0.058 when aligning 19M and 125M assignments), even though the high-level category distribution is highly stable (Pearson $r = 0.997$, Section 5.3). This is a noteworthy negative result: specific quanta appear to be re-discovered rather than shared across model sizes, with stability holding only at the coarser level of token categories.

7 Conclusion

Agreement with the predicted power-law slope is not sufficient for evaluating quanta-discovery methods. On Pythia-125M, Spectral clustering achieves the closest envelope fit ($|\Delta| = 0.004$), yet SSC-Lasso yields the highest quanta fingerprint performance (MCC $+0.960$ vs. $+0.860$); Hierarchical reinforces the mismatch, ranking second in envelope fit on both models but last in fingerprint MCC. Power-law fit and representational utility therefore capture distinct properties of a clustering and should both enter method evaluation. SSC-Lasso attains the tightest envelope fit on Pythia-19M ($|\Delta| = 0.008$) and the strongest fingerprint classifiers on both models; on Pythia-125M this gain is partly driven by a formatting catch-all cluster, but removing it still yields MCC $+0.910$, consistent with genuine cluster-specific structure rather than a purely surface artefact. Beyond AI-text detection, the fingerprint representation is a compact, model-grounded feature vector that may transfer to authorship attribution, register classification, and domain-shift detection; establishing whether the catch-all is a Pythia-specific artefact and whether the envelope-ranking reversal persists at larger scales are the main open questions.

Limitations

The following caveats should be noted. First, experiments cover only two model sizes, and the scale-dependent reversal in envelope ranking (Section 4.3) warns against extrapolating to larger models. Second, the AI-generated text is produced by the same Pythia model under evaluation, so the fingerprint signal could partly reflect Pythia-specific artefacts, including tokenisation patterns, sampling biases, and prompt-format effects, rather than a generalisable notion of model-generated text. Third, the Pythia-125M fingerprint gain should be interpreted with care: it is partly driven by a formatting-heavy catch-all cluster (Section 5.4) that distinguishes human Pile text from Pythia output largely through punctuation and newline structure rather than clearly interpretable quantum identity. The post-ablation comparison in Section 5.4 nevertheless shows that the observed differences are not entirely explained by this cluster. Fourth, the fingerprint classifier uses a fixed resolution of $k = 400$ clusters, and the optimal value of k for classification need not coincide with that which best recovers the power-law envelope. Fifth, our subspace-clustering evaluation covers only the global self-expression family (SSC-Lasso and SSC-OMP); correlation-based methods such as 4C, ORCLUS, and COPAC (Kriegel et al., 2009) impose locality assumptions rather than global sparsity, and whether these families offer better scale-robustness remains an open question.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 2397–2430.

Dan Braun, Lucius Bushnaq, Stefan Heimersheim, Jake Mendel, and Lee Sharkey. 2025. Interpretability in parameter space: Minimizing mechanistic description length with attribution-based parameter decomposition. *Preprint*, arXiv:2501.14926.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Ehsan Elhamifar and René Vidal. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Michael E. Houle, Martin Kiermeier, and Arthur Zimek. 2023. Clustering high-dimensional data. In Lior Rokach, Oded Maimon, and Erez Shmueli, editors, *Machine Learning for Data Science Handbook*, pages 219–237. Springer.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(1):1:1–1:58.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. 2023. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36. ArXiv:2303.13506.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *Preprint*, arXiv:1109.2378.

Lance Parsons, Ehtesham Haque, and Huan Liu. 2004. Subspace clustering for high dimensional data: A review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Chong You, Daniel P. Robinson, and René Vidal. 2016. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3918–3927.

A Method-Comparison Figures

Figures 6 and 7 overlay all four methods' envelopes on Pythia-19M and Pythia-125M; Figures 8 and 9

show the SSC-Lasso slope as a function of (d, α) on each model; and Figures 10 and 11 give the per-method rank-frequency fits at each method's optimal configuration.

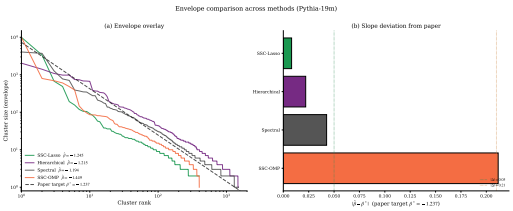

Figure 6: Envelope overlay for all four methods on Pythia-19M. SSC-Lasso achieves the tightest fit ($|\Delta| = 0.008$); Hierarchical ($|\Delta| = 0.022$) and Spectral ($|\Delta| = 0.043$) are the closest runners-up.

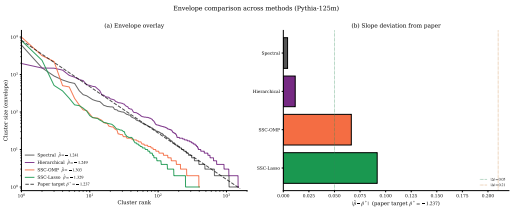

Figure 7: Envelope overlay on Pythia-125M. Spectral wins ($|\Delta| = 0.004$); SSC-Lasso over-steepens to -1.329.

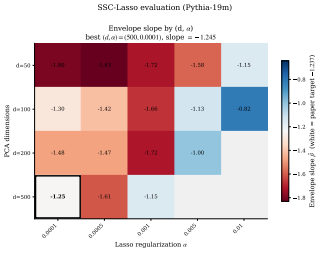

Figure 8: SSC-Lasso envelope slope as a function of d and α on Pythia-19M. The optimal configuration ($d = 500, \alpha = 0.0001$) achieves $|\Delta| = 0.008$. Blank cells indicate configurations where the affinity graph became too sparse to produce a valid envelope.

B Bootstrap Stability: Full Statistics

We assess clustering stability by resampling: 50 iterations at 80% token subsampling, with SSC-Lasso held at ($d = 500$, $\alpha = 0.0001$) for both models, and Spectral and Hierarchical at their canonical configurations. Table 3 reports NMI and ARI (mean $\pm$ std and median) for all three paradigms. The envelope-optimal SSC-Lasso configuration on Pythia-125M ($d = 200$, $\alpha = 0.01$) is too sparse to resample reliably, so the stability

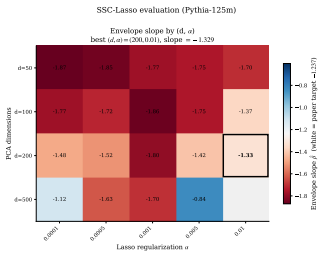

Figure 9: SSC-Lasso hyperparameter sensitivity on Pythia-125M. Optimal region shifts to $d = 200$, $\alpha = 0.01$; best slope -1.329 ($|\Delta| = 0.092$). Blank cells indicate configurations where the affinity graph became too sparse to produce a valid envelope.

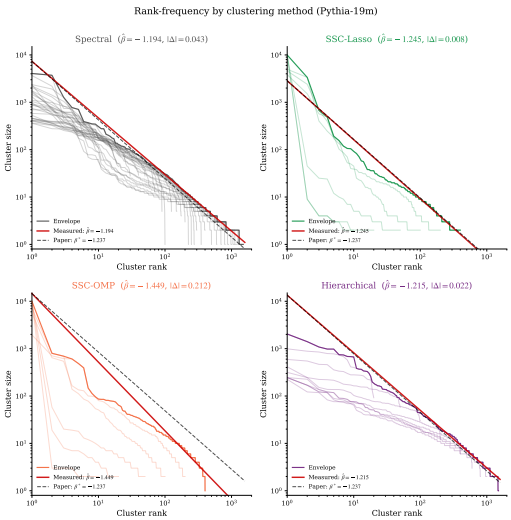

Figure 10: Individual rank-frequency envelopes for all four methods on Pythia-19M at optimal hyperparameter settings.

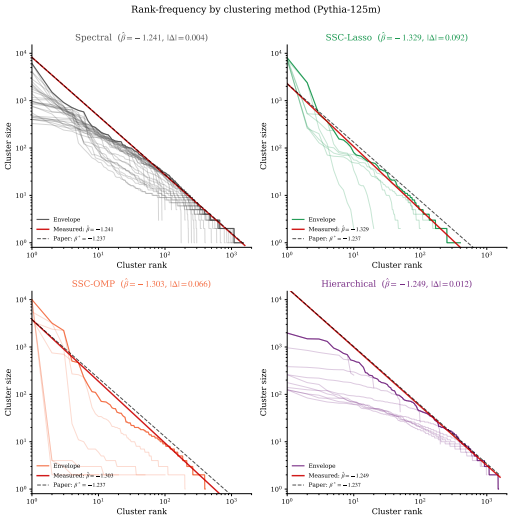

Figure 11: Individual rank-frequency envelopes for all four methods on Pythia-125M at optimal hyperparameter settings.

Table 3: Bootstrap stability for all three paradigms ($k = 400$, 80% token subsampling, 50 iterations). SSC-Lasso uses ($d = 500$, $\alpha = 0.0001$) for both models; its envelope-optimal Pythia-125M configuration ($d = 200$, $\alpha = 0.01$) is too sparse to resample reliably. Spectral and Hierarchical use their canonical configurations.

Method	Model	NMI (mean±std)	NMI (med.)	ARI (mean±std)	ARI (med.)
Hierarchical	Pythia-19M	0.772 ± 0.002	0.772	0.440 ± 0.022	0.438
Hierarchical	Pythia-125M	0.752 ± 0.002	0.752	0.357 ± 0.012	0.357
SSC-Lasso	Pythia-19M	0.731 ± 0.020	0.732	0.381 ± 0.094	0.367
SSC-Lasso	Pythia-125M	0.685 ± 0.014	0.687	0.329 ± 0.042	0.330
Spectral	Pythia-19M	0.687 ± 0.005	0.688	0.134 ± 0.009	0.132
Spectral	Pythia-125M	0.670 ± 0.006	0.670	0.124 ± 0.007	0.125

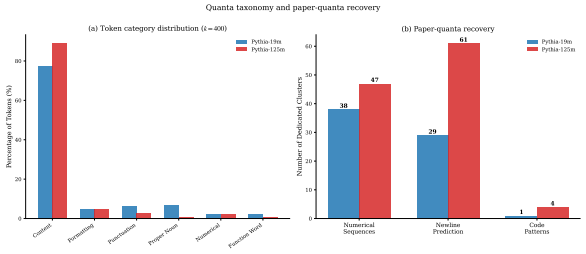

Figure 12: (a) Percentage of tokens in each linguistic category for SSC-Lasso on Pythia-19M and Pythia-125M, with each cluster assigned its dominant token category. (b) Number of dedicated clusters per category that recover a paper quantum.

analysis holds SSC-Lasso fixed at ($d = 500$, $\alpha = 0.0001$) throughout. Figure 13 shows the full NMI and ARI distributions across the 50 iterations for SSC-Lasso on both models; equivalent distributions for Spectral and Hierarchical are omitted for space, but the full per-iteration data is available in the project repository.

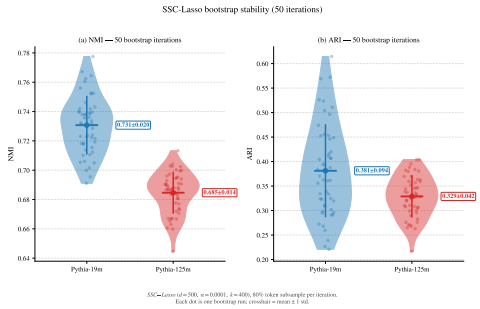

Figure 13: NMI and ARI distributions across 50 bootstrap iterations for SSC-Lasso on Pythia-19M and Pythia-125M. Mean ± std annotated right of each distribution.

C Fingerprint Robustness Analysis

Table 4 lists the token composition of the Pythia-125M SSC-Lasso catch-all cluster discussed in Section 5.4. Table 5 reports a surface-feature baseline using only three document-level statistics (Shannon entropy, active-quanta count, peak probability), maximum single-quantum MCC, and a label-permutation baseline for all six paradigm × model combinations.

Table 4: Top-10 tokens in SSC-Lasso cluster 1 (Pythia-125M). Cluster contains 7,491 of 10,000 gradient tokens (74.9%).

Rank	Token	% of cluster
1	\n (newline)	5.0
2	.	2.1
3	-	1.5
4	’	1.4
5	,	1.2
6	:	0.7
7	the	0.6
8	of	0.5
9	</ (HTML tag)	0.5
10	(	0.5

Hierarchical shows distributed signal on both models ($\Delta = 0.068$ on 19M, 0.092 on 125M), meaning per-cluster identity contributes beyond surface statistics. However, its per-cluster signal is markedly weaker than Spectral ($\Delta = 0.193/0.239$) and SSC-Lasso on 19M ($\Delta = 0.224$), explaining Hierarchical's lower full MCC despite a comparable surface baseline. SSC-Lasso on 125M is one-cluster-led: cluster 1 alone achieves MCC +0.950, yet removing it and renormalising the remaining 399 features still yields MCC +0.910, indicating substantial distributed signal beyond the catch-all cluster. SSC-Lasso on 19M shows the strongest distributed signal across all methods ($\Delta = 0.224$), with no single cluster dominating (Max 1-q = +0.711).

Table 5: Robustness checks. Surface (3-statistic aggregate baseline), Δ = Full$-$Surface (per-cluster signal beyond surface statistics), Max 1-q (best single-cluster MCC), and Perm. (label-permutation baseline, $\approx$0 expected). Full MCC for each paradigm is reported in Table 2. ‡ one-cluster-led (Max 1-q $\geq 0.95\times$ Full).

Paradigm	Model	Surface	Δ	Max 1-q	Perm.	Flag
Spectral	19M	$+0.584$	$+0.193$	$+0.369$	$+0.031$	distributed
SSC-Lasso	19M	$+0.659$	$+0.224$	$+0.711$	-0.023	distributed
Hierarchical	19M	$+0.607$	$+0.068$	$+0.510$	$+0.031$	distributed
Spectral	125M	$+0.621$	$+0.239$	$+0.359$	-0.004	distributed
SSC-Lasso	125M	$+0.871$	$+0.089$	$+0.950$	$+0.049$	‡
Hierarchical	125M	$+0.648$	$+0.092$	$+0.351$	-0.001	distributed

D Bag-of-Words Baseline

To assess whether quanta fingerprints capture signal beyond surface word statistics, we trained a TF-IDF logistic regression classifier on the same human documents and the shared 600-document AI corpus used for the SSC-Lasso fingerprint experiment (784 human and 599 qualifying AI documents; the bag-of-words baseline keeps any non-empty document and so retains more AI documents than the $\approx$497 used by the fingerprint, which requires at least three zero-loss tokens), using a 10,000-feature vocabulary, ℓ_2 regularisation, and the same 5-fold CV protocol. Table 6 compares it against the SSC-Lasso fingerprint classifier.

E SSC-Lasso Quanta Examples

Table 7 shows four representative quanta from the SSC-Lasso solution on Pythia-19M (d=500, α=0.0001, k=400), illustrating the linguistic coherence of individual clusters. Each cluster is dominated by a single token type corresponding to a recognisable syntactic or discourse function.

Table 6: Bag-of-words (TF-IDF, same human documents, shared AI corpus, and CV protocol as the SSC-Lasso fingerprint classifier) vs. SSC-Lasso fingerprint (Table 2). MCC: mean across 5 seeds $\times$ 5 folds (both methods). Balanced accuracy: 5-seed CV mean for BoW; single-seed value from Table 2 for the fingerprint.

	Pythia-19M		Pythia-125M	
Method	Bal. acc.	MCC	Bal. acc.	MCC
TF-IDF (BoW)	0.897	+0.791	0.891	+0.781
SSC-Lasso fingerprint	**0.945**	**+0.884**	**0.984**	**+0.960**

Table 7: Four representative SSC-Lasso quanta from Pythia-19M (d=500, α=0.0001, k=400), ordered by cluster size rank (rank 1 = largest cluster). Cluster composition is computed over the same 10,000 zero-loss gradient tokens used throughout.

Rank (size)	Top tokens (share)	Inferred linguistic skill
14 (56)	\n (100%)	Paragraph-break prediction
18 (36)	." (78%), !" (14%), . . . " (8%)	Dialogue sentence ending
23 (29)	’ (100%)	Apostrophe in contractions
28 (26)	such (92%)	Enumeration marker (*such as* . . .)