# SPECTRAL DATA VALUATION: WHAT DATA TEACHES IN THE MODEL'S INFORMATION GEOMETRY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Training data are commonly ranked using properties of the corpus or gradients measured against a fixed reference, yet their value changes with the model. An example that supplies a missing pattern early in training can become redundant after that pattern has been learned. We study this state dependence through the model's local information geometry. We connect example influence, weight adaptation, and feature coverage through related projections of a curvature spectrum. Based on this view, we introduce `SDV`, a spectral data valuation framework that prices a candidate by the gradient energy it contributes to important and unsaturated directions. The same price supports data selection and targeted synthesis. Across from-scratch pre-training, continual pre-training, and post-training synthesis, the price predicts subsequent learning progress before an example or mixture is used. We formalize the score as a marginal of a saturating directional utility and state conditions under which its fixed-basis selection objective is monotone and submodular. At a $5\%$ token budget, `SDV` nearly matches the full-corpus GSM8K result ($33.7\%$ vs. $33.9\%$), reaches the measured induction-head threshold $1.8\times$ sooner, and ranks unseen mixtures with Spearman correlation $0.93$. These results identify spectral information as a practical signal for adaptive data curation and provide an empirical route from local geometry to changing data supply.

## 1 INTRODUCTION

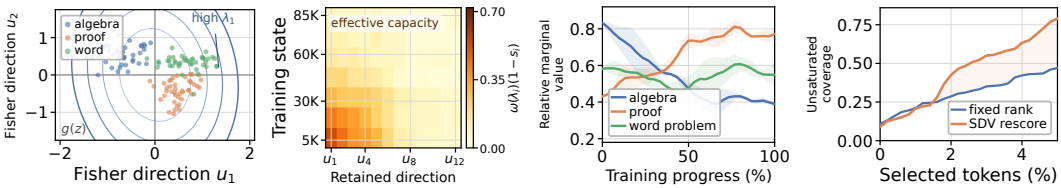
Large language models (LLMs) are trained through a succession of data decisions. Corpora are filtered and mixed before pre-training, resampled during continual pre-training, and extended with generated examples during post-training (Gunasekar et al., 2023; Li et al., 2024; Albalak et al., 2024; Xu et al., 2024). These decisions determine which capabilities receive update capacity and which patterns are repeated as training proceeds. Existing methods value data through document quality, diversity, target-domain weights, validation gradients, or influence on a fixed probe (Xie et al., 2023a; Sorscher et al., 2022; Xia et al., 2024; Grosse et al., 2023). Such scores characterize the corpus, the target distribution, or the effect measured at one model state.

The same example can have a different effect at another state. A reasoning pattern that supplies a missing direction early in training may repeat a saturated direction later. The ranking changes again within a selection round because every chosen example alters the marginal value of candidates that teach similar patterns. As a result, a fixed ranking can direct an increasing share of the data budget toward directions that have already received substantial learning signal, while other learnable directions remain undersupplied. This tension appears in repeated curation and in targeted synthesis, where each new example is intended to address what the current model has yet to learn. It raises the following question.

*Why does the value of training data change as a language model learns, and how can we adapt data selection to what the model still needs to learn?*

To answer this question, we track gradient energy across an empirical Fisher eigensystem. Curvature sets directional importance, while accumulated energy determines remaining capacity. As capacity fills, another example along that direction contributes less. Fixed rankings can thus retain formerly valuable but saturated examples.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




(a) Information geometry (b) Directional capacity (c) Changing value (d) Adaptive coverage

Figure 1: **The information-geometric mechanism behind state-dependent valuation. (a)** Candidate gradients projected onto the leading Fisher directions form distinct, noisy groups. The contours mark equal Fisher energy, so the same Euclidean displacement can carry different teaching value. **(b)** The effective capacity $\omega(\lambda_i)(1 - s_i)$ changes across retained directions and training states, exposing the state variable used by the price. **(c)** As this profile evolves, the marginal values of algebra, proof, and word-problem examples exchange order. Its lines and pale bands show the mean and sample standard deviation across three runs. **(d)** Under the same token budget, updating the price after each selection covers more unsaturated spectral energy than reusing a fixed ranking. Together, the panels trace curvature geometry → remaining capacity → changing data value → adaptive selection, the sequence formalized by Equations (1) and (6) and Theorem 1.

Figure 1 makes this mechanism explicit. Candidate gradients are resolved in the Fisher geometry, the effective capacity changes across directions and checkpoints, marginal values exchange order, and current-state repricing preserves access to unsaturated directions. Data value is therefore a property of the candidate, the current model, and the selection history together.

This view also connects three measurements of learning. Influence relates an example to a change in a model functional (Chen et al., 2026; Grosse et al., 2023), spectral rotation locates weight directions that move during adaptation (Borodin et al., 2026), and feature coverage records which internal directions a dataset activates (Li et al., 2026; Bricken et al., 2023). Propositions 1 to 3 show how these observables reduce to related projections under the specified Fisher approximation, a first-order adaptation window with a spectral gap, and feature-basis alignment, respectively. For a fixed basis, the resulting directional utility is monotone and submodular. Exact-marginal greedy selection attains the standard $(1 - 1/e)$ cardinality guarantee, and Corollary 1 identifies the energy required to reach a chosen saturation level. The additional conditions and proofs are given in Section D.

Based on this analysis, we propose *Spectral Data Valuation*, denoted `SDV`. A spectral probe maintains a low-rank curvature sketch and a directional energy profile. The resulting price weights projected gradient energy by directional importance and remaining capacity, then normalizes the value by token count. During curation, `SDV` ranks available examples, updates the profile after each selection, and filters generated examples aimed at undersupplied directions. Recomputing the price after probe refreshes gives the same framework a state-dependent selection and synthesis rule.

We evaluate `SDV` in from-scratch pre-training, continual pre-training, and post-training synthesis across model sizes from 14M to 1B. The spectral price is measured before each outcome and predicts subsequent learning progress across checkpoints. Priced curation is evaluated on mathematics, code, and biomedical data, together with the emergence of an induction head during from-scratch training. Prices computed by a 14M probe further predict the ordering of unseen mixtures trained at 410M and 1B scale. Matched ablations and temporal analyses connect these outcomes to saturation, retained rank, basis refresh, and synthesis filtering.

We make three contributions.

- **Information-geometric account.** We identify remaining directional capacity as the state variable for prospective valuation, relate three measurements of learning through projected gradient quantities, and characterize diminishing returns in a fixed spectral basis.
- **State-dependent valuation.** We introduce `SDV`, which combines directional importance, remaining capacity, and candidate energy in one per-token price and updates that price as the model and selected set change.
- **Cross-regime evaluation.** We evaluate the same price as a predictor, selector, and synthesis filter across three training regimes. At a 5% token budget it nearly matches the full-corpus GSM8K result, and mixture transfer, matched controls, and component dynamics test when the signal remains informative.

## 2 Related Work

**Data Selection for Language Models.** Data curation combines quality filtering, deduplication, perplexity, diversity, and learned classifiers (Albalak et al., 2024; Penedo et al., 2024; Sorscher et al., 2022; Marion et al., 2023; Li et al., 2024). These methods identify documents that are broadly suitable for training and have enabled high-quality open pre-training corpora. Distribution-matching approaches instead choose examples or domain weights for a target distribution. DSIR estimates importance weights from data features, while DoReMi learns domain weights with a proxy model (Xie et al., 2023b;a). Their scores describe the corpus or the target mixture. `SDV` addresses the complementary question of how the value of a candidate changes as the trained model acquires and saturates different directions.

**Gradient-Based Data Valuation.** Influence functions, datamodels, and TRAK estimate how individual examples affect a chosen prediction or validation functional (Koh & Liang, 2017; Grosse et al., 2023; Park et al., 2023; Ilyas et al., 2022). Gradient-matching methods such as LESS select examples whose gradients align with a validation set (Xia et al., 2024). These methods make model-side learning signals available to curation, with the probe or validation distribution defining the valued direction. `SDV` uses squared projections and a running saturation profile to measure prospective energy across the retained local spectrum. The price is updated at each probe refresh and applies to an existing pool or generated candidates.

**Spectral, Mechanistic, and Feature-Based Analyses.** Weight spectra, singular-vector changes, and gradient subspaces characterize how training concentrates in a small number of directions (Martin & Mahoney, 2021; Borodin et al., 2026; Gur-Ari et al., 2018; Aghajanyan et al., 2021; Hu et al., 2022). Mechanistic attribution traces interpretable units to influential training examples (Chen et al., 2026), while sparse autoencoders provide feature inventories for measuring activation coverage (Bricken et al., 2023; Huben et al., 2023). Feature-guided synthesis constructs data for underrepresented directions (Li et al., 2026; Zhou et al., 2023; Xu et al., 2024), and Kronecker-factored curvature supplies a scalable approximation to local geometry (Martens & Grosse, 2015; George et al., 2018; Amari, 1998). Our analysis places these measurements in a shared projected coordinate system and derives a state-dependent data price from the same geometry.

## 3 Preliminaries

This section establishes notation, introduces the spectral object at the center of our study, and casts three families of prior measurements into a single formal vocabulary that the rest of the paper builds upon.

**Language Models and the Learning Spectrum.** Consider an autoregressive language model $p_\theta$ with parameters $\theta \in \mathbb{R}^d$, trained on a corpus $\mathcal{D} = \{z_i\}_{i=1}^N$ by minimizing the token-level loss $\ell(z, \theta) = -\log p_\theta(z)$, and write $g(z) = \nabla_\theta \ell(z, \theta)$ for the gradient of a single example. The local geometry training imposes on parameter space is the Fisher information

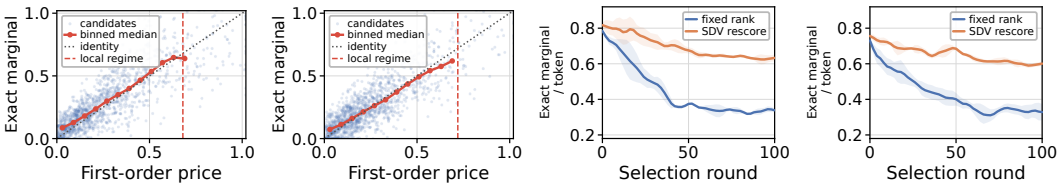
$$F(\theta) = \mathbb{E}_z\big[\, g(z)\, g(z)^\top \big] = \sum\nolimits_{i=1}^{d} \lambda_i\, u_i u_i^\top, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \tag{1}$$

whose eigenpairs $\{(\lambda_i, u_i)\}$ define what we call the *learning spectrum*. For a log-likelihood model, the expected Fisher is a local information metric, and its relation to the Hessian requires additional model and sampling conditions (George et al., 2018). Prior spectral studies report concentrated heads and heavy-tailed regimes in trained networks (Martin & Mahoney, 2021; Borodin et al., 2026). We estimate the noise scale $\bar{\lambda}$ from the observed bulk rather than assume a universal bulk law. Storing or inverting the exact $F$ is intractable at LLM scale, so following George et al. (2018); Grosse et al. (2023) we work with the eigenvalue-corrected Kronecker factorization, which for each linear layer with input activations $a$ and output gradients $\delta$ approximates the corresponding block as

$$F_{\text{layer}} \approx (U_A \otimes U_S)\, \Lambda\, (U_A \otimes U_S)^\top, \qquad A = \mathbb{E}[\, aa^\top], \quad S = \mathbb{E}[\, \delta\delta^\top], \tag{2}$$

with $U_A$ and $U_S$ the eigenvectors of the covariance factors and $\Lambda$ the corrected eigenvalues. The approximation replaces the full parameter-space matrix with factors that scale with the layer widths. We use it throughout Section 4.

These measurements share spectral coordinates. Section 4 develops the projected-energy price, and Figure 2 tests its calibration and marginal-value dynamics.




(a) Pythia-70M (b) OLMo-1B (c) Mathematics (d) Code

Figure 2: **Empirical verification of the key method components. (a)(b)** First-order price against the exact fixed-basis marginal at two model scales. The red curve is the binned median, the dotted line is identity, and the dashed line marks the local regime. **(c)(d)** Exact marginal value throughout equal-budget selection on two domain pools. Lines and pale bands show the mean and sample standard deviation across three runs. Adaptive rescoring preserves useful value as the fixed ranking revisits saturated directions. Details appear in Sections D and E.

**Instrument I. Influence in Data Space.** Influence functions estimate how up-weighting a training example $z$ would change a differentiable functional $f$ of the trained model (Koh & Liang, 2017; Grosse et al., 2023). The mechanistic variant restricts attention to the parameter subspace $\theta_{\text{sub}}$ of an interpretable unit, such as an induction head, and evaluates

$$\mathcal{I}(z, f) \;=\; -\, g_{\text{sub}}(z)^{\top} \big(F_{\text{sub}} + \varepsilon I\big)^{-1} \nabla_{\theta_{\text{sub}}} f, \tag{3}$$

where $F_{\text{sub}}$ is the EK-FAC curvature of the subspace (Chen et al., 2026). Its eigendecomposition expresses influence as curvature-weighted products of the candidate and probe projections. The probe determines their signs and weights.

**Instruments II and III. Rotation and Coverage.** Spectral analyses track the rotation of weight singular vectors during adaptation (Borodin et al., 2026). In representation space, sparse autoencoders provide a feature inventory (Bricken et al., 2023; Huben et al., 2023), and activation coverage can be compared with generalization (Li et al., 2026). Our coverage analysis states the alignment between feature and covariance directions as an explicit condition. Full definitions are in Section C.

## 4 Spectral Data Valuation

This section develops three components of `SDV`. They are a running spectral sketch and saturation profile (Section 4.1), a score for each candidate example (Section 4.2), and a selection and synthesis procedure (Section 4.3).

### 4.1 Spectral Probe. Sketching the Learning Spectrum

The probe answers two questions about the model at training step $t$. The first asks *which directions matter*, and the second asks *how much has each already been taught*. For the first, `SDV` maintains the Kronecker factors of Equation (2) as exponential moving averages over training batches,

$$A^{(t)} = \beta A^{(t-1)} + (1-\beta)\, \mathbb{E}_{\text{batch}}[\, aa^{\top}], \qquad S^{(t)} = \beta S^{(t-1)} + (1-\beta)\, \mathbb{E}_{\text{batch}}[\, \delta\delta^{\top}], \tag{4}$$

and refreshes their eigendecompositions every $T_{\text{probe}}$ steps, retaining for each layer the top-$k$ eigenpairs $\{(\lambda_i, u_i)\}_{i=1}^{k}$ that carry the head of the spectrum. Related Kronecker factors are maintained by some second-order optimizers. Otherwise, these statistics are estimated from training batches and counted in the overhead analysis.

For the second question, `SDV` accumulates projected gradient energy from selected examples. With $\mathcal{B}_t$ the training batch at step $t$,

$$e_i^{(t)} = e_i^{(t-1)} + \sum\nolimits_{z \in \mathcal{B}_t} \langle g_t(z), u_i^{(t)} \rangle^2, \qquad s_i^{(t)} = \frac{e_i^{(t)}}{e_i^{(t)} + \rho\, \lambda_i} \;\in\; [0, 1), \tag{5}$$

where $\rho$ sets a directional capacity. Energies are maintained in the current retained basis and reprojected when that basis is refreshed, as described in Section G.2. A direction with low $s_i$ has received little selected gradient energy relative to its capacity. The profile supplies a bounded state variable for the curation loop.

**Cost.** Maintaining the factors of Equation (4) costs one running average per layer, refreshing the eigenpairs one small eigendecomposition every $T_{\text{probe}}$ steps, and pricing a candidate one backward pass followed by $k$ inner products per layer. Effect-based valuations pay for per-example gradient storage, random projections over the entire pool (Xia et al., 2024; Park et al., 2023), or probe-specific solves (Grosse et al., 2023), and Table 10 quantifies the gap.

### 4.2 Spectral Price. What an Example Teaches

Given the sketch and the profile, the value of a candidate example is defined as the spectral energy its gradient delivers where teaching is still possible.

$$v(z) \;=\; \frac{1}{|z|}\sum_{i=1}^{k}\omega(\lambda_i)\,\big(1-s_i\big)\,\big\langle g(z),\,u_i\big\rangle^2, \qquad \omega(\lambda)=\frac{\lambda}{\lambda+\bar{\lambda}}, \tag{6}$$

where $|z|$ is the token count of $z$, the weighting $\omega$ discounts directions below the estimated bulk scale $\bar{\lambda}$, and the factor $(1-s_i)$ discounts directions already saturated. The score is measured in weighted gradient energy per token and serves as a proxy for marginal learning value.

To derive the saturation weight, define directional energy $e_i(\mathcal{S})=\sum_{z\in\mathcal{S}}\langle g(z),u_i\rangle^2$.

**Assumption 1** (Fixed-basis selection round)**.** *Within a selection round, the eigenbasis, retained positive eigenvalues, nonnegative spectral weights, and $\rho>0$ remain fixed, and each candidate adds nonnegative directional energy.*

Under Assumption 1, the associated utility is

$$V(\mathcal{S})=\sum_{i=1}^{k}\rho\lambda_i\omega(\lambda_i)\log\!\left(1+\frac{e_i(\mathcal{S})}{\rho\lambda_i}\right). \tag{7}$$

Its derivative with respect to $e_i$ is $\omega(\lambda_i)(1-s_i)$. Thus Equation (6) is the first-order marginal of Equation (7) per token, evaluated at the current profile.

**Theorem 1** (Fixed-basis diminishing returns)**.** *Under Assumption 1, $V$ is monotone and submodular. For equal-length examples and a cardinality budget, greedy selection by the exact marginal of $V$ obtains at least $(1-1/e)$ of the optimal value.*

The theorem states why the profile update matters. Supplying one direction reduces its next marginal. Section D.7 proves the result, and Figure 2a,b measures the gap between the token-normalized score and the exact marginal.

The same projected quantities occur in influence, weight perturbation, and feature activation. The following statements specify what follows under the corresponding assumptions, and their derivations are given in Section D.

**Proposition 1** (Influence and projected gradients)**.** *For a fixed probe functional $f$, the influence score of Equation (3) is a signed, curvature-weighted sum of $\langle g(z),u_i\rangle$ in the restricted Fisher eigensystem. Its magnitude is bounded by the weighted gradient energy and a probe-dependent factor.*

**Proposition 2** (Rotation perturbation bound)**.** *Under a first-order update and a nonzero singular gap, weight-subspace rotation is bounded by the accumulated update norm divided by that gap. Retained spectral energy contributes to this bound, with ordering determined by gradient coherence and singular gaps.*

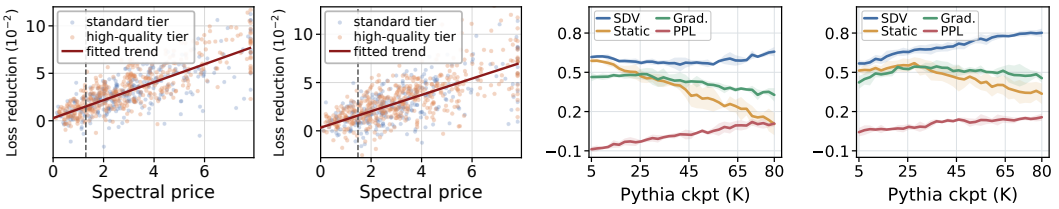
**Proposition 3** (Conditional relation to coverage)**.** *Suppose feature directions coincide with an orthonormal activation basis and the back-propagated output norm is bounded above and away from zero. Thresholded feature activation is then bounded between two thresholdings of gradient energy on the corresponding Kronecker directions.*

Together, the propositions give a conditional account of why the three measurements can align. Influence retains signed projections, rotation depends on their accumulated update, and coverage thresholds activation support. Section 5.2 measures their agreement independently of the assumptions used in the derivations.

### 4.3 Spectral Curation. Closing the Loop

The score changes as selected examples add energy to retained directions. Under token budget $B$, `SDV` updates the provisional profile after each pick and selects greedily,

$$z^\star=\operatorname*{arg\,max}_{z\in\mathcal{D}_{\text{pool}}}\; v(z), \qquad \hat{e}_i\leftarrow\hat{e}_i+\langle g(z^\star),u_i\rangle^2, \qquad \hat{s}_i=\frac{\hat{e}_i}{\hat{e}_i+\rho\lambda_i}, \tag{8}$$




(a) Pythia-70M (b) OLMo-1B (c) Mathematics (d) Code

Figure 3: **Prospective validation of the spectral price. (a),(b)** Spectral price versus next-window loss reduction at Pythia-70M and OLMo-1B. Pythia-160M appears in Figure 10a. **(c),(d)** Prospective rank correlation on fixed mathematics and code pools across Pythia-70M public checkpoints. Lines and bands show the mean and s.d. across three runs.

## 5 Experiments

We conduct experiments to answer four questions. (**RQ1**) Does the price prospectively predict the next learning window? (**RQ2**) How do performance and training dynamics change under a fixed data budget? (**RQ3**) Does the price predict gains from unseen mixtures? (**RQ4**) What do individual components contribute, and at what cost? Additional controls and implementation details are reported in Sections E, G and I.

### 5.1 Experiment Setup

**Models, Corpora and Regimes.** We instrument **Pythia** (Biderman et al., 2023) from 14M to 1B and **OLMo-1B** (Groeneveld et al., 2024). We study from-scratch training, continual training on mathematics, code, and biomedical pools, and post-training synthesis. Metrics include GSM8K (Cobbe et al., 2021), HumanEval, MedQA, perplexity, and the induction-head prefix score (Olsson et al., 2022), and configurations appear in Sections G and G.1.

**Baselines and Parameters.** Surface-based baselines are random, perplexity, DSIR (Xie et al., 2023b), and DoReMi (Xie et al., 2023a). Effect-based baselines are LESS (Xia et al., 2024), EK-FAC influence (Grosse et al., 2023), and FAC (Li et al., 2026). Budgets and schedules are matched. We use $k = 64$, $T_{\text{probe}} = 500$, and $\beta = 0.95$.

### 5.2 Prospective Validity of the Spectral Price (RQ1)

At each checkpoint, we price held-out candidates before measuring loss reduction over the next probe window. The same construction is used at each scale.

**Candidate-level prediction.** The price tracks next-window loss reduction, with rank correlations 0.76, 0.78, and 0.69 at Pythia-70M, Pythia-160M, and OLMo-1B, respectively (Figure 3a,b and Figure 10a).

**Evolution across checkpoints.** Figure 3c,d compares refreshed spectral, static influence, gradient-matching, and perplexity scores on fixed mathematics and code pools. Layer-resolved overlap and heavy-tail co-evolution appear in Figures 9 and 11.

**Prospective interpretation.** The held-out window separates the price from the outcome used to evaluate it. High-price candidates therefore identify directions that become useful in the next update window, while the residual spread records examples whose effect depends on the local state.

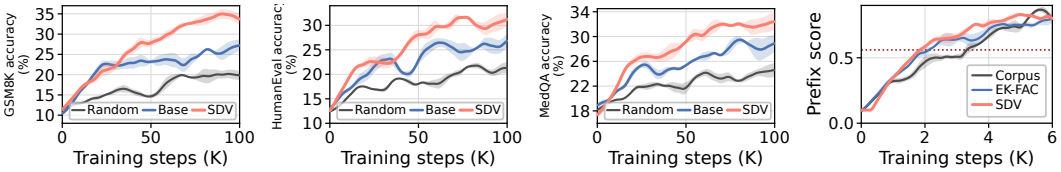
**State tracking.** The matched checkpoint axes show how the prospective association can be tracked as the model state changes. The comparison motivates the refresh analysis in Section 5.5.

**Scale consistency.** The ordering remains informative as the probe moves from Pythia-70M to OLMo-1B. This consistency supports the use of a small probe for the mixture-transfer experiment and separates the spectral signal from a single model scale.

The prospective results establish the role of the probe in the remainder of the paper. They identify the directions that are useful before a budget is spent, while the following studies ask how the same signal behaves after selection, transfer, and component changes.

Table 1: **Main results.** Continual pre-training of OLMo-1B on three domains at a 5% token budget. Accuracy columns report GSM8K, HumanEval and MedQA in percent, and **Avg.** is the mean of those three accuracy columns, with the two perplexity columns reported apart since they are not on the same scale. The last row trains on the whole pool with no selection and serves as the full-data reference. Best in **bold**, runner-up underlined. Every cell is averaged over three seeds, and the 25% and 50% arms of the sweep are in Section I.

| | Mathematics | | Code | | Biomedical | | |
|---|---|---|---|---|---|---|---|
| **Method** | GSM8K (%) ↑ | PPL ↓ | HumanEval (%) ↑ | PPL ↓ | MedQA (%) ↑ | PPL ↓ | **Avg. (%)** ↑ |
| Random | 19.8 | 27.8 | 21.3 | 11.9 | 24.6 | 9.6 | 21.9 |
| Perplexity (Marion et al., 2023) | 21.6 | 26.1 | 22.6 | 11.2 | 25.4 | 9.1 | 23.2 |
| DSIR (Xie et al., 2023b) | 22.9 | 25.4 | 23.4 | 10.8 | 26.1 | 8.8 | 24.1 |
| DoReMi (Xie et al., 2023a) | 23.7 | 24.9 | 24.1 | 10.5 | 26.7 | 8.6 | 24.8 |
| LESS (Xia et al., 2024) | 25.4 | 24.1 | 25.3 | 10.1 | 27.6 | 8.3 | 26.1 |
| EK-FAC Influence (Grosse et al., 2023) | 27.2 | 22.8 | 26.8 | 9.4 | 28.9 | 7.7 | 27.6 |
| FAC (Li et al., 2026) | 26.5 | 23.4 | 26.2 | 9.7 | 28.3 | 7.9 | 27.0 |
| SDV (Ours) | **33.7** | **18.9** | **31.2** | **8.1** | **32.4** | **6.6** | **32.4** |
| Full corpus (100%) | 33.9 | 18.6 | 31.4 | 7.8 | 32.7 | 5.6 | 32.7 |




(a) Mathematics continual PT (b) Code continual PT (c) Biomedical continual PT (d) From-scratch emergence

Figure 4: **Training dynamics under priced curation. (a)** to **(c)** Mathematics, code, and biomedical accuracy under random selection, the strongest matched baseline, and SDV at a 5% token budget. **(d)** Induction-head prefix score under corpus, EK-FAC, and spectral orderings, and the dotted line marks the 0.56 formation threshold. Lines and pale bands show the mean and sample standard deviation across three runs.

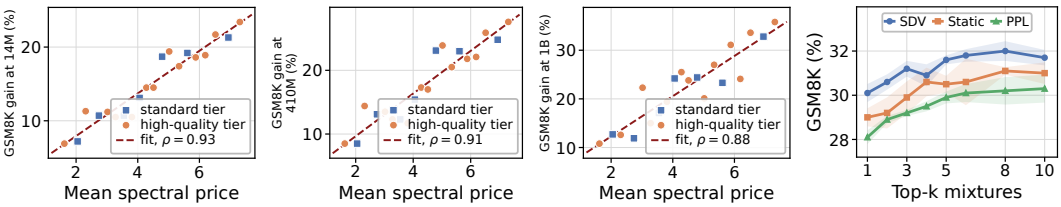
### 5.3 Data Efficiency and Training Dynamics (RQ2)

Each method selects from the same candidate pool under an equal token budget and is followed by an identical training schedule. Table 1 reports the endpoints, while Figure 4 follows the corresponding optimization trajectories. The first three panels use the three continual-pre-training domains and the fourth follows the from-scratch mechanism measurement.

**Overall performance.** At 5% of the candidate pool, SDV obtains 33.7% GSM8K, 31.2% HumanEval, and 32.4% MedQA, close to the respective full-corpus results of 33.9%, 31.4%, and 32.7% (Table 1). At the same budget, the strongest matched baseline obtains 27.2%, 26.8%, and 28.9%. The broader sweep in Table 8 places SDV first in seven of the nine accuracy comparisons considered jointly with Table 1. SDV also reduces held-out mathematics perplexity from 27.8 under random selection to 18.9.

**Training stability.** The endpoint advantage is accompanied by a different training trajectory. On mathematics, code, and biomedical data, the strongest baseline improves rapidly and then undergoes task-specific regressions, whereas the priced runs maintain steadier gains through the same training span (Figure 4a,b,c). The gap develops throughout training, while the three domains approach their late-stage levels at different rates.

**Emergence and general ability.** With the total token count fixed, priced ordering brings the induction-head prefix score across its 0.56 threshold at step 1,840, compared with step 3,312 under corpus ordering (Figure 4d). The three orderings approach a comparable late-stage score, while their transition times remain separated. In the 25% budget sweep, DoReMi remains 0.3% ahead on ARC and MMLU, while SDV remains above random selection on both suites (Table 8). Targeted-synthesis results and matched qualitative examples are reported in Figure 13 and section I.5.




(a) Realized gain at 14M (b) Realized gain at 410M (c) Realized gain at 1B (d) Mixture choice

Figure 5: **Forecasting the worth of unseen mixtures. (a)** to **(c)** Pythia-14M probe prices against absolute GSM8K gains for the same 20 mixtures trained at 14M, 410M, and 1B scale. **(d)** Absolute GSM8K accuracy at 1B after combining the top-$k$ mixtures from a separate held-out pool under an equal training-token budget. The three rankings use refreshed spectral price, frozen spectral price, and perplexity. Lines and pale bands show the mean and s.d. across three runs.

### 5.4 Prediction of Mixture Value (RQ3)

To test prospective valuation, we form 20 mixtures that vary in domain composition and quality tier. Each mixture is scored once before training, and the resulting ranking is compared with the observed gain after the same token budget (Figure 5a,b,c). A separate pool tests whether the ranking guides the choice of new mixtures (Figure 5d).

**Forecasting accuracy.** At 14M scale, the spectral price has Spearman correlation 0.93 with the realized GSM8K gain across the 20 mixtures (Figure 5a). Absolute gains range from 6.9% to 23.4% over the base checkpoint, and the two quality tiers overlap in both price and outcome. Table 9 reports every composition, score, and outcome.

**Transfer across scale.** The same prices correlate with gains realized at 410M and 1B scale by 0.91 and 0.88, respectively (Figure 5b,c). Although absolute gains change with scale, the ordering remains similar, with a mean displacement of 2.1 positions out of 20 at 1B.

**Choice on new mixtures.** In a disjoint candidate pool, we combine the top-$k$ mixtures under each ranking and train a 1B model for the same 500M-token budget. At $k = 5$, refreshed spectral pricing reaches 31.6% GSM8K, compared with 30.5% for a frozen spectral price and 29.9% for perplexity (Figure 5d). The margin varies with $k$ and narrows near $k = 4$, while the refreshed ranking remains above the two alternatives at the evaluated settings. Scoring-budget sensitivity for the probe appears in Section I.1, and the held-out choice protocol is detailed in Section G.5.

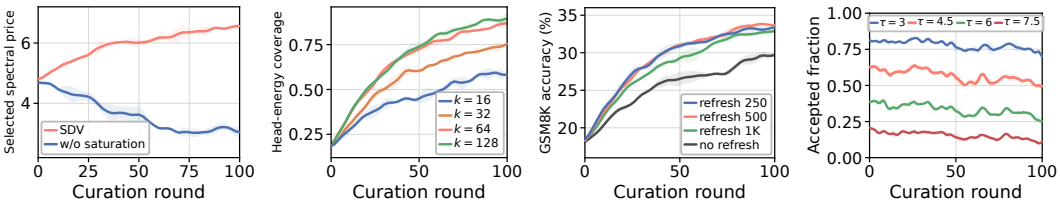
### 5.5 Framework Analysis (RQ4)

Table 2: **Ablation study.** Mathematics accuracy and perplexity use a 5% continual-pre-training budget, and Δ is the GSM8K change from the full method. **Val.** $\rho$ is the Spearman correlation between the ablation price and the realized 14M mixture gains of Section 5.4. Emergence step and three-way overlap use the matched from-scratch protocol of Section G.7. Three seeds per variant.

| Variant | GSM8K ↑ | Δ | PPL ↓ | Val. $\rho$ ↑ | Emergence steps ↓ | Overlap ↑ |
|---|---|---|---|---|---|---|
| `SDV` (full) | 33.7 | – | 18.9 | 0.93 | 1,840 | 0.86 |
| ❶ w/o saturation | 27.6 | −6.1 | 22.4 | 0.71 | 2,610 | 0.71 |
| ❷ w/o eigenvalue weighting | 30.2 | −3.5 | 20.8 | 0.84 | 2,180 | 0.76 |
| ❸ w/o refresh | 31.1 | −2.6 | 20.1 | 0.79 | 2,040 | 0.80 |
| ❹ w/o synthesis | 32.4 | −1.3 | 19.4 | 0.92 | 1,930 | 0.85 |

**Ablation study.** We evaluate the four main design choices of `SDV` by removing saturation weighting, eigenvalue weighting, basis refresh, and targeted synthesis in turn. Table 2 reports domain accuracy, perplexity, emergence timing, and subspace overlap under the corresponding protocols. Removing saturation causes the largest drop in GSM8K, from 33.7 to 27.6.

**Effect of the components.** Removing saturation gives 27.6% GSM8K (27.2% for static EK-FAC) and mean overlap 0.71 (0.86 for the full method). Removing eigenvalue weighting, refresh, and synthesis reduces GSM8K by 3.5%, 2.6%, and 1.3%. The synthesis gap reaches 4.9% after removing the top price decile (Section G.9).




(a) Saturation weighting (b) Retained rank (c) Basis refresh period (d) Threshold

Figure 6: **How the valuation state changes during curation.** All panels use the same 100-round protocol. **(a)** Saturation weighting keeps the selected price from drifting toward already supplied directions. **(b)** Retaining more eigenpairs expands head-energy coverage before the curves level off. **(c)** A 500-step basis refresh gives the most stable accuracy trajectory among the tested periods. **(d)** Acceptance fractions under four nested price thresholds, evaluated on the same generated candidate batch each round, with $\tau = 4.5$ deployed. Lines and pale bands show the mean and sample standard deviation across three runs.

**Saturation changes what is selected.** With the full score, the mean selected price rises as the low-price region is exhausted and remains high after local re-ordering events (Figure 6a). Removing saturation produces an early rise followed by a sustained decline, indicating repeated selection from directions whose accumulated energy is already high. The curves therefore distinguish a ranking that responds to the accumulated training state from one that repeatedly favors the same directions. This temporal difference is consistent with the largest endpoint change in Table 2.

**Rank and refresh govern distinct effects.** Increasing the retained rank improves head-energy coverage through $k = 64$, while $k = 128$ follows a similar late-stage level with higher checkpoint variation (Figure 6b). The smaller rank captures most of the measured head energy without expanding the sketch to the highest tested resolution. Refreshing every 500 steps yields the most stable accuracy trajectory, whereas a shorter period adds estimation noise, and a fixed basis loses accuracy as training moves away from the initial geometry (Figure 6c). Rank controls which directions the price can resolve, whereas refresh controls whether those directions continue to describe the current model. The complete unit-cost accounting is given in Table 10.

**Synthesis filtering controls supply.** A low threshold accepts most generated candidates, whereas higher thresholds retain a smaller and more stable fraction (Figure 6d). The four curves evaluate the same candidate batch in each round, isolating the effect of the acceptance rule. The deployed threshold of 4.5 preserves variation across rounds while rejecting candidates whose projected energy is already represented in the pool. Figure 13 and section I.5 report the associated price distributions, feature diversity, and matched examples.

**Operational reading.** The components act at distinct timescales. Saturation changes within-round ranking, retained rank sets spatial resolution, basis refresh follows model drift, and the synthesis threshold governs new supply. Their trajectories connect each control to a distinct curation decision.

**Joint interpretation.** The ablation values and trajectories provide a consistent ordering of roles. Saturation has the largest effect on the selected examples, rank determines how much of the active spectrum is represented, refresh maintains that representation as the state moves, and thresholding determines which generated directions enter the pool. The resulting design couples valuation, selection, and synthesis through one state variable rather than through separate scores. This ordering links the endpoint comparisons in Table 2 to the temporal behavior in Figure 6.

## 6 Conclusion

We introduced `SDV`, which prices examples by projected energy in important and unsaturated directions. Across three training regimes, the same price supports prospective prediction, data selection, and targeted synthesis as the model state changes. This view treats curation as a state-tracking problem: a running spectrum records what the model can still learn, and the same record determines selection, transfer, and synthesis decisions across domains and model scales.

**AI Usage Statement.** AI assisted only with language polishing.

## REFERENCES


Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 7319–7328, 2021.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430, 2023.

Nikita Borodin, Maria Krylova, Artem Zabolotnyi, Dmitry Aspisov, Egor Shikov, Nikita Tyuplyaev, Oleg Travkin, Roman Alferov, and Dmitry Vinichenko. Diffract: Spectral view of LLM domain adaptation. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Jianhui Chen, Yuzhang Luo, and Liangming Pan. Mechanistic data attribution: Tracing the training origins of interpretable LLM units. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a Kronecker-factored eigenbasis. In *Advances in Neural Information Processing Systems*, 2018.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15789–15809, 2024.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *International Conference on Machine Learning*, pp. 9525–9587, 2022.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. The Stack: 3 TB of permissively licensed source code. *arXiv preprint arXiv:2211.15533*, 2022.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, 2024.

Zhongzhi Li, Xuansheng Wu, Yijiang Li, Lijie Hu, and Ninghao Liu. Less is enough: Synthesizing diverse data in feature space of LLMs. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564*, 2023.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

National Library of Medicine. Pubmed 2025 baseline released. NLM Technical Bulletin, 2025. URL https://www.nlm.nih.gov/pubs/techbull/jf25/jf25_pubmed_2025_baseline.html.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning*, pp. 27074–27113, 2023.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. In *The Twelfth International Conference on Learning Representations*, 2024.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, 2022.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, 2024.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, 2023a.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, 2023b.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. *arXiv preprint arXiv:2406.08464*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023.

# Appendix


**A Scope and Limitations** . . . . . **14**

**B Notation Summary** . . . . . **14**

**C The Three Readings in Full** . . . . . **14**

**D Proofs and Theoretical Complements** . . . . . **14**

D.1 Directional Utility and Marginal Value . . . . . 15

D.2 Regularity Conditions . . . . . 15

D.3 Proof of the Influence Proposition . . . . . 15

D.4 Proof of the Rotation Proposition . . . . . 15

D.5 Proof of the Coverage Proposition . . . . . 16

D.6 Theoretical Complements . . . . . 16

D.7 Proof of the Greedy Theorem . . . . . 16

**E Empirical Checks of the Correspondences** . . . . . **17**

**F Specificity of the Spectral Score** . . . . . **18**

**G Experimental Details** . . . . . **18**

G.1 Instantiating the Price in the Three Regimes . . . . . 19

G.2 The Full Curation Loop . . . . . 19

G.3 Models and Training Configurations . . . . . 20

G.4 Candidate Pools and Evaluation Suites . . . . . 20

G.5 Datasets and Statistics . . . . . 20

G.6 Baseline Implementations . . . . . 21

G.7 Evaluation Protocols . . . . . 22

G.8 Rationale for the Emergence Threshold . . . . . 22

G.9 What Synthesis Buys When the Pool Is Thin . . . . . 23

**H Prompt Set for Spectral Synthesis** . . . . . **23**

**I Additional Results** . . . . . **24**

I.1 Full Quantitative Results . . . . . 24

I.2 Cost and Sensitivity Diagnostics . . . . . 25

I.3 Co-evolution of the Three Tails . . . . . 26

I.4 Robustness and Transfer . . . . . 27

I.5 Qualitative Analyses . . . . . 27

# A Scope and Limitations

The EK-FAC sketch omits cross-layer curvature, and the saturation profile retains one scalar per direction. The formal guarantees use the conditions in Section D.2, while the token-normalized score is a first-order approximation for variable-length examples. Our evaluation covers models from 14M to 1B, and larger scales and generator biases remain untested.

# B Notation Summary

Table 3: Notation used in the spectral valuation framework.

| Symbol | Meaning |
|---|---|
| $z,\ \lvert z\rvert,\ g(z)$ | Candidate example, its token length, and its per-example gradient at the scoring checkpoint. |
| $F,\ (\lambda_i, u_i)$ | Local Fisher approximation and its retained eigenpairs. |
| $A,\ S$ | Activation and output-gradient covariance factors of the EK-FAC approximation. |
| $e_i,\ s_i$ | Accumulated projected gradient energy and its bounded saturation transform. |
| $\omega(\lambda_i),\ \rho$ | Eigenvalue weight and capacity constant. |
| $v(z),\ V(\mathcal{S})$ | Token-normalized first-order price and fixed-basis directional utility. |
| $B,\ T_{\text{probe}},\ k$ | Token budget, basis-refresh interval, and retained rank per layer. |

# C The Three Readings in Full

Section 3 introduces the three measurements. We specify their observables here before stating the conditions under which they can be related to projected gradients.

**Instrument I: influence in data space.** Influence functions estimate how up-weighting a training example $z$ would change a differentiable functional $f$ of the trained model (Koh & Liang, 2017; Grosse et al., 2023), and the mechanistic variant restricts attention to the parameter subspace $\theta_{\text{sub}}$ of an interpretable unit such as an induction head, as Equation (3) states. The reported statistic is the distribution of $\mathcal{I}(z, f)$ over the pool at fixed $f$, whose upper tail is heavy: a small minority of examples accounts for most of the estimated effect on the unit.

**Instrument II: rotation in weight space.** Spectral analyses of adaptation decompose each weight matrix as $W = U\Sigma V^\top$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r$, and monitor two statistics across training checkpoints: the empirical spectral density of $\{\sigma_j\}$ and the *singular vector agreement*, the cosine alignment between corresponding singular vectors before and after adaptation (Borodin et al., 2026). The striking observation is that continual pre-training leaves the density nearly invariant while rotating a small set of singular vectors adjacent to spectral peaks, so that most of the recorded update can be truncated with no measurable loss.

**Instrument III: coverage in feature space.** Sparse autoencoders decompose an internal activation $x \in \mathbb{R}^m$ into a sparse code $h = \sigma(xW_{\text{enc}}) \in \mathbb{R}^{n_h}$ with $n_h \gg m$, whose coordinates can be analyzed as features (Bricken et al., 2023; Huben et al., 2023). Feature activation coverage measures the fraction of a specified feature set activated by a dataset $\mathcal{S}$ (Li et al., 2026). The learned dictionary is not, in general, the eigenbasis of $A$ in Equation (2). We measure its alignment with that basis before applying the conditional relation of Proposition 3.

The three statistics are computed from per-example gradients, weight matrices, and activations, respectively. They are compared using the checkpoint, layer, and retained rank specified in Section G.7.

# D Proofs and Theoretical Complements

This appendix derives the directional objective and the three relationships used in Section 4.2. Throughout, $F = \sum_i \lambda_i u_i u_i^\top$ is evaluated at a fixed checkpoint. The statements concern local approximations to distinct measurements.

### D.1 Directional Utility and Marginal Value

Let $q_i(z) = \langle g(z), u_i \rangle^2$ for each retained direction, and let $c_i = \rho\lambda_i > 0$. For a fixed basis and a selected set $\mathcal{S}$, $e_i(\mathcal{S}) = \sum_{z\in\mathcal{S}} q_i(z)$. The objective in Equation (7) assigns utility $c_i\omega(\lambda_i)\log(1 + e_i/c_i)$ to direction $i$. Its exact gain from adding $z$ is

$$\Delta_z V(\mathcal{S}) = \sum_{i=1}^{k} c_i\omega(\lambda_i)\log\left(1 + \frac{q_i(z)}{c_i + e_i(\mathcal{S})}\right). \tag{9}$$

Using $\log(1+x) = x+O(x^2)$ gives $\Delta_z V(\mathcal{S})/|z| = v(z)+O\big(\sum_i \omega(\lambda_i)q_i(z)^2/[|z|(c_i+e_i)]\big)$ when $q_i(z)/(c_i + e_i)$ is small. This establishes the role of Equation (6) as a local marginal approximation. The experiments use the stated score, while the greedy guarantee below concerns the exact gain in Equation (9).

### D.2 Regularity Conditions

**Assumption 2** (Curvature equivalence)**.** *For the local influence calculation, the curvature used in Equation (3) is the positive-semidefinite Fisher block $F_{\text{sub}}$, with damping $\varepsilon > 0$. This is a curvature approximation to a Hessian-based influence calculation (Grosse et al., 2023; Martens & Grosse, 2015), and the proposition concerns that specified approximation.*

**Assumption 3** (First order training regime)**.** *Within the analyzed adaptation window, the layer update is $\Delta W = -\eta\sum_{z\in\mathcal{C}} G(z)$ for per-example layer gradients $G(z)$ and a common step-size factor $\eta$. The retained basis is held fixed for the energy decomposition within that window. Other optimizers introduce a residual update term that must be included in the perturbation norm.*

**Assumption 4** (Spectral gap)**.** *The weight matrix $W$ under study has simple leading singular values, $\min_{j\le r_0}(\sigma_j - \sigma_{j+1}) \ge \gamma > 0$ for the $r_0$ directions of interest, so that singular subspace perturbations obey Wedin type bounds.*

### D.3 Proof of Proposition 1

*Proof.* Under Assumption 2, the damped restricted curvature in Equation (3) admits the spectral resolution $(F_{\text{sub}} + \varepsilon I)^{-1} = \sum_i (\lambda_i + \varepsilon)^{-1} u_i u_i^\top$, where the eigenpairs belong to the restricted block. Substituting into Equation (3) yields

$$\mathcal{I}(z, f) \;=\; -\sum\nolimits_i \frac{1}{\lambda_i + \varepsilon}\,\big\langle g(z), u_i\big\rangle\big\langle \nabla f, u_i\big\rangle, \tag{10}$$

which establishes the signed projection identity. To relate its magnitude to Equation (6), let $P$ contain retained directions with $a_i = \omega(\lambda_i)(1 - s_i) > 0$, and define $v_P(z) = |z|^{-1}\sum_{i\in P} a_i\langle g(z), u_i\rangle^2$. Weighted Cauchy–Schwarz gives

$$|\mathcal{I}_P(z, f)|^2 \le |z| v_P(z) \sum_{i\in P} \frac{\langle \nabla f, u_i\rangle^2}{(\lambda_i + \varepsilon)^2 a_i}. \tag{11}$$

The omitted directions contribute the residual $\mathcal{I}_{P^c}$ in Equation (10). This is an upper bound on influence magnitude conditional on the probe and saturation weights. Signed cancellations and probe orientation preclude a converse ranking claim. □

### D.4 Proof of Proposition 2

*Proof.* Fix a layer with weight matrix $W$ and an adaptation window with $n$ gradient contributions. Under Assumption 3, write $\Delta W = -\eta\sum_{z\in\mathcal{C}} G(z)$, omitting the stated higher-order term. If $\|\Delta W\|_2 \le \gamma/2$, a standard singular-subspace perturbation bound under Assumption 4 gives

$$\|\sin\Theta(U_r, U_r')\|_2 \le \frac{2\|\Delta W\|_2}{\gamma} \le \frac{2\eta}{\gamma}\sqrt{n\sum_{z\in\mathcal{C}}\|G(z)\|_F^2}. \tag{12}$$

The second inequality uses $\|\sum_z G(z)\|_F^2 \le n\sum_z \|G(z)\|_F^2$. In an orthonormal Fisher basis, $\|G(z)\|_F^2$ decomposes into energy in the retained directions and a residual outside them. If the retained weights $a_i = \omega(\lambda_i)(1 - s_i)$ satisfy $a_i \ge a_{\min} > 0$, their unweighted energy is at most

$|z|v(z)/a_{\min}$. Thus the measured price contributes to an envelope for the perturbation, together with the omitted energy. The ordering of rotating singular vectors additionally depends on gradient cancellation, residual energy, and local singular gaps. □

### D.5 Proof of Proposition 3

*Proof.* Consider one token at a fixed linear layer, with activation $a$ and back-propagated signal $\delta$, so its gradient is $g = \delta \otimes a$. Under the alignment condition of Proposition 3, let the feature associated with direction $p_i$ fire when $|\langle a, p_i\rangle| > \tau_i$. For an orthonormal output basis $\{q_j\}$, the energy in all Kronecker directions sharing $p_i$ is

$$E_i = \sum_j \langle g, q_j \otimes p_i\rangle^2 = \|\delta\|^2 \langle a, p_i\rangle^2. \tag{13}$$

If $0 < m \le \|\delta\| \le M$, the feature indicator obeys the deterministic sandwich

$$\mathbf{1}[E_i > M^2\tau_i^2] \le \mathbf{1}[|\langle a, p_i\rangle| > \tau_i] \le \mathbf{1}[E_i > m^2\tau_i^2]. \tag{14}$$

Taking maxima over tokens and examples gives corresponding bounds on dataset-level support. The statement requires aligned feature directions and bounded output gradients. A sparse autoencoder need not satisfy the alignment condition exactly, so the empirical overlap in Section 5.2 measures this approximation. No monotonic generalization guarantee follows from the support relation alone. □

**Summary.** The derivations identify three links to projected gradient quantities with different assumptions and losses of information. Influence preserves sign, rotation depends on accumulated updates and singular gaps, and coverage thresholds activation support. Their measured subspace agreement is an empirical question.

### D.6 Theoretical Complements

We next separate two properties of the curation rule. The first is stability of the estimated eigenspace, and the second is diminishing returns of a fixed-basis selection objective. A final identity describes the directional energy required for a chosen saturation level.

**Lemma 1** (Sketch consistency)**.** *Let $\hat{F}$ be a symmetric curvature sketch and $F$ the reference matrix at the same checkpoint. For a simple eigenvalue separated by gap $\gamma_i > 0$, if $\|\hat{F} - F\|_2 \le \gamma_i/2$, then*

$$\sin \angle\big(\hat{u}_i, u_i\big) \le \frac{2\|\hat{F} - F\|_2}{\gamma_i}. \tag{15}$$

*The perturbation norm includes sampling error, drift in the moving average, and approximation error from factorization.*

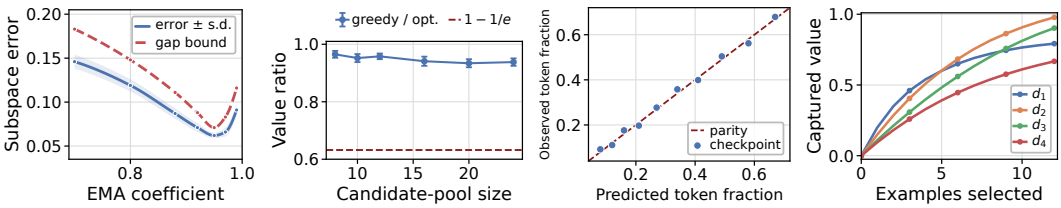
*Proof.* Apply the Davis–Kahan eigenspace perturbation bound to $F$ and $\hat{F}$. The gap and perturbation condition ensure that the matched eigenvector is identified, yielding the displayed inequality. This deterministic statement leaves the size of $\|\hat{F} - F\|_2$ to empirical measurement or a separate concentration analysis. □

### D.7 Proof of Theorem 1

*Proof.* Each $e_i(\mathcal{S})$ is a nonnegative modular set function. Composition with the nondecreasing concave function $\log(1+x)$ makes each directional term monotone submodular, and their nonnegative sum remains so. The first derivative of the $i$-th term with respect to new energy is $\omega(\lambda_i)(1 - s_i)$, and Taylor expansion gives Equation (6) as the first-order gain per token. For equal-length examples, maximizing the exact marginal under a cardinality limit satisfies the classical $(1 - 1/e)$ bound (Nemhauser et al., 1978). Variable lengths and approximate marginals require their own analysis. □

**Corollary 1** (Energy required for saturation)**.** *For a target saturation level $\bar{s} \in (0, 1)$, direction $i$ reaches $s_i \ge \bar{s}$ exactly when $e_i \ge \rho\lambda_i\bar{s}/(1 - \bar{s})$. Consequently, saturating the first $m$ directions requires total directional energy at least $\rho\bar{s}(1 - \bar{s})^{-1}\sum_{i\le m}\lambda_i$. The corresponding token budget depends on the gradient-energy distribution of the corpus and selection algorithm.*

*Proof.* Solve $s_i = e_i/(e_i + \rho\lambda_i) \geq \bar{s}$ for $e_i$, then sum the inequalities over $i \leq m$. Converting directional energy into tokens additionally depends on how the candidate pool distributes gradient energy across examples. □




(a) Tracking error against the bound (b) Greedy value against the optimum (c) Measured against predicted budget (d) Diminishing returns per direction

Figure 7: **Empirical validation of the theoretical complements. (a)** Eigenvector tracking error against the independently computed perturbation envelope of Lemma 1, with three sketch runs at each EMA coefficient $\beta$. The line gives their mean and the pale band their sample standard deviation, and the default $\beta = 0.95$ has error $0.062$ and bound $0.071$. **(b)** Exact-marginal greedy value as a fraction of the exhaustive optimum on small, equal-length pools, with the $(1 - 1/e)$ floor drawn. **(c)** Measured token fraction against a head-mass heuristic, with the parity line. The directional-energy identity of Corollary 1 and this empirical token conversion describe the two stages of the calculation. **(d)** Cumulative exact-marginal value for four retained directions as examples are selected sequentially. Markers denote selected-example counts, and the connected cumulative values have decreasing per-direction marginal gains under the fixed-basis objective.

## E Empirical Checks of the Correspondences

The statements of Propositions 1 to 3 have different assumptions and provide bounds or support relations. We therefore measure their component quantities separately. The controlled analysis uses a smaller checkpoint for which a reference eigenspace can be constructed and reports deviations introduced by the sketch, profile re-projection, and feature alignment.

**(i) Approximating the eigenbasis $u_i$.** We take $\hat{u}_i$ from the EK-FAC sketch of Equation (2) and compare it with a reference eigenspace on Pythia-14M. Figure 7a reports tracking error $0.062$ at the default $\beta = 0.95$ and a perturbation envelope of $0.071$. The envelope is computed from the measured matrix difference and eigengap in Lemma 1. It changes with $\beta$ because both sampling variation and lag in the moving average affect the sketch.

**(ii) Approximating the saturation profile $s_i$.** The energies $e_i$ of Equation (5) are accumulated from the projected gradients of selected training examples. The capacity constant $\rho$ is fixed once per regime by the pilot protocol of Section G.3. On a basis refresh, the retained energy profile is reprojected and the contribution of discarded directions is logged. In our runs this discarded contribution is below $3\%$ of the tracked head energy at $k = 64$.

**(iii) Constructing observable samples of the price.** For each candidate document we record its projected gradient and score. The scoring log contains $1.2$M candidate measurements across the two suites. The diagnostics use a common scoring log and independently constructed influence, rotation, and feature measurements.

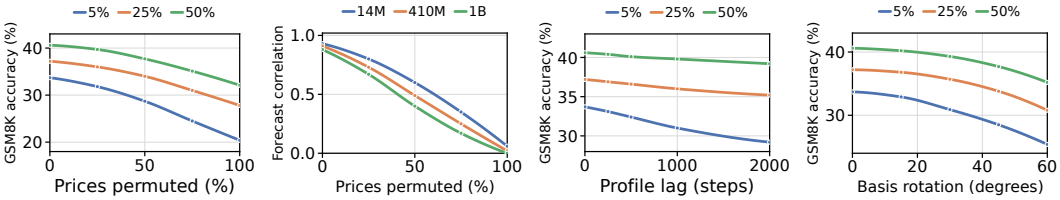
**(iv) Testing the three correspondences. Setup.** At a fixed checkpoint, we construct an influence subspace, a weight-update subspace, and an activation-feature subspace from their respective measurements. We compare them using principal angles and evaluate the residuals of the conditional relations in Section D. The singular-gap analysis is specific to rotation, while influence and coverage use their corresponding conditions. **Results.** The rotation residual decreases from $0.34$ at gap $\gamma = 0.02$ to $0.03$ at $\gamma = 0.80$ (Figure 14a,b). Influence truncation error decreases from $0.30$ to $0.105$ as the retained rank rises from $8$ to $64$ (Figure 14c). Feature-support disagreement decreases from $0.26$ to $0.15$ as measured dictionary-basis alignment increases from $0.30$ to $0.86$ (Figure 14d). The perturbation bound of Equation (12) controls rotation from above, while the other panels assess the distinct assumptions of the influence and coverage relations. **Rank and refresh.** We vary retained

rank and refresh period to measure how the empirical correspondences change with the spectral sketch. Overlap rises toward the default rank before flattening, and the no-refresh variant has lower rotation agreement, consistent with Table 2 and Figure 12. These trends, rather than the propositions, motivate the chosen operating point.

**(v) Checking the greedy selection rule.** Theorem 1 makes a claim that can be falsified on pools small enough to enumerate. **Setup.** Given a frozen basis and pools of 8 to 24 candidate documents, we generated two selections that each buy half of the pool, one by the greedy rule of Equation (8) and one by exhaustive search over the same budget. To isolate the submodular structure we fixed the basis for both, so the two differ only in the order in which they buy. **Results.** Exact-marginal greedy selection reaches 0.93 to 0.97 of exhaustive optimum on the equal-length pools (Figure 7b). The per-direction value curves in Figure 7d show decreasing increments as more examples are selected. Panel **(c)** compares the observed token fraction with a head-mass heuristic, and the discrepancy is within 0.018 over the illustrated range. This empirical conversion complements the directional-energy identity in Corollary 1.

## F Specificity of the Spectral Score

The component ablations in Table 2 remove parts of the method. We additionally preserve the score distribution while perturbing its assignment to examples, then compare matched token budgets and forecast target scales. We shuffle a specified fraction of price ranks within length deciles, keeping document lengths and the marginal distribution of prices unchanged. The same intervention is evaluated for data selection and for prospective mixture ranking. We also test two state-dependent components: using a profile from an earlier checkpoint and rotating the retained basis while preserving its dimension and eigenvalues. Each curve begins at its corresponding unperturbed main-text value.




(a) Selection under shuffled prices (b) Forecast under shuffled prices (c) Stale saturation profile (d) Rotated retained basis

Figure 8: **Matched perturbation controls. (a)** GSM8K at 5%, 25%, and 50% token budgets under within-length-decile price permutation. **(b)** Correlation between Pythia-14M prices and mixture gains at 14M, 410M, and 1B target scales under the same permutation. **(c)(d)** GSM8K at the three budgets under profile lag and retained-basis rotation, respectively, and panel **(d)** holds eigenvalues and rank fixed. Markers denote evaluated settings, curves guide the eye, and all start at their matched unperturbed results.

At full score permutation, the three budget curves approach their matched random-selection levels, and forecast correlation approaches zero across target scales. At the 5% budget, a profile lag of 2,000 steps lowers GSM8K from 33.7 to 29.2, and a 60-degree basis rotation lowers it to 25.4. The larger budgets show smaller changes under the same perturbations. These controls examine the assignment of gradient energy to examples, the current profile, and the retained directions.

## G Experimental Details

This appendix documents the training configurations, dataset statistics, baseline implementations, and evaluation protocols behind Section 5.

**Algorithm 1** The `SDV` curation loop

**Require:** model $\theta$, data pool $\mathcal{D}_{\text{pool}}$, token budget $B$, refresh period $T_{\text{probe}}$, rank $k$

Initialize Kronecker factors $A, S$ and saturation profile $s \leftarrow \mathbf{0}$
**while** budget $B$ not exhausted **do**
  ▷ *Spectral Probe (Section 4.1)*
  Update $A, S$ by the moving averages of Equation (4) from recent training batches
  **if** $T_{\text{probe}}$ steps have elapsed **then**
    Refresh top-$k$ eigenpairs $\{(\lambda_i, u_i)\}_{i=1}^{k}$, reproject energy $e$, and recompute $s_i = e_i/(e_i + \rho\lambda_i)$
  **end if**
  ▷ *Spectral Price (Section 4.2)*
  Compute $v(z)$ for candidates $z \in \mathcal{D}_{\text{pool}}$ by Equation (6)
  ▷ *Spectral Curation (Section 4.3)*
  Select $z^\star$ greedily and advance the provisional profile $\hat{s}$ by Equation (8)
  **if** a dominant direction with $s_i \approx 0$ has no support in $\mathcal{D}_{\text{pool}}$ **then**
    Map $u_i$ to its nearest sparse features and synthesize contrastive exemplars for them
  **end if**
  Train on the curated batch, update $\theta$, and accumulate its projected gradient energies by Equation (5)
**end while**

### G.1 Instantiating the Price in the Three Regimes

The price of Equation (6) is a regime-agnostic principle: it charges an example for the spectral energy its gradient delivers to dominant yet unsaturated directions. Its components ($k$, $s_i$, $\omega$, $|z|$ and the pool it is evaluated over) are instantiated differently in each of the three regimes.

**(i) From-scratch pre-training.** The pool is the unseen prefix of The Pile, and $|z|$ is the token count of a whole document. The profile starts at zero because no direction has been taught, so the factor $(1 - s_i)$ is inert early and the price reduces to eigenvalue-weighted energy. This regime therefore reads out the emergence of structure, and the reported quantity is the induction-head prefix-matching score whose threshold crossing Figure 4d plots.

**(ii) Continual pre-training.** The pool is a single domain corpus. The probe is initialized from a pilot stream of training batches, and its factors are maintained separately from AdamW. The initial saturation profile includes the pilot examples, so $(1 - s_i)$ affects the first scored batch. We report domain accuracy and held-out perplexity at fixed token budgets, and Table 2 measures the contribution of saturation weighting.

**(iii) Post-training synthesis.** In the synthesis setting, the score is evaluated on generated candidates as well as examples in the existing pool. A generation is retained when its price on the targeted directions exceeds the threshold $\tau = 4.5$, according to the filter in Section H. The token count $|z|$ includes instruction and response tokens. We report the judge win rate defined in Section G.7.

### G.2 The Full Curation Loop

Algorithm 1 states the loop described in Section 4.3. Projected candidate energies can be cached within one basis-refresh window, and scores are then updated from the provisional profile after each selection batch. This still requires scanning the eligible candidate set at each selection round. The scoring-time measurements in Table 10 exclude these re-ranking operations and are therefore reported separately from end-to-end training cost.

Table 4: Training configurations for the three experimental regimes.

| | From-scratch | Continual PT | Post-training |
|---|---|---|---|
| Base model | Pythia-14M / 70M | OLMo-1B | OLMo-1B (SFT) |
| Optimizer | AdamW | AdamW | AdamW |
| Peak learning rate | $1 \times 10^{-3}$ | $3 \times 10^{-5}$ | $2 \times 10^{-5}$ |
| Schedule | cosine, 1% warmup | cosine, 1% warmup | cosine, 3% warmup |
| Batch size (tokens) | 0.5M | 1M | 0.25M |
| Context length | 2,048 | 2,048 | 2,048 |
| Total budget (tokens) | 2B | $\{5, 25, 50\}$% of pool | 2K–300K samples |
| Seeds | 3 | 3 | 3 |

### G.3 Models and Training Configurations

All from-scratch and emergence experiments use the Pythia suite (Biderman et al., 2023), whose public checkpoints every 1,000 steps allow the spectrum to be read across the whole of pre-training. Continual pre-training starts from the official OLMo-1B checkpoint (Groeneveld et al., 2024). Table 4 lists the configurations.

**Probe instrumentation.** The spectral probe attaches to attention and MLP linear layers. Kronecker factors are accumulated separately from AdamW with EMA coefficient $\beta = 0.95$ and eigendecomposed every $T_{\text{probe}} = 500$ steps. The top $k = 64$ directions per layer are retained. The capacity constant $\rho$ is calibrated once per regime by matching the median saturation of a 100M-token pilot run to 0.5. On a basis refresh, the implementation recomputes projected energies from a bounded recent-example cache and resumes accumulation in the new basis. The contribution of older examples outside the cache is approximated by projecting their retained covariance summary. The discrepancy is assessed in Section E.

**Compute accounting.** All experiments run on one node with 8 A100 GPUs. Training, pilot estimation, candidate scoring, and re-ranking are accounted for separately because the scoring pass over a large pool can dominate the training time. Table 10 reports unit scoring cost. End-to-end runtime aggregates this cost over the examined tokens and adds pilot estimation, re-ranking, and training.

### G.4 Candidate Pools and Evaluation Suites

Table 5: Candidate pools and evaluation suites. Token counts are approximate.

| Regime | Pool | Size | Evaluation |
|---|---|---|---|
| From-scratch | The Pile (Gao et al., 2020) | 300B tokens | induction-head prefix score |
| Continual PT (math) | OpenWebMath (Paster et al., 2024) | 14.7B tokens | GSM8K, domain PPL |
| Continual PT (code) | The Stack (Python subset) | 20B tokens | HumanEval, domain PPL |
| Continual PT (biomed) | PubMed abstracts | 4B tokens | MedQA, domain PPL |
| Post-training | MAGPIE pool (Xu et al., 2024) | 300K samples | AlpacaEval 2, MT-Bench |

### G.5 Datasets and Statistics

This section gives the construction and filtering details for each candidate pool. The summary table is separated above so that the pool inventory can be read independently from the dataset descriptions.

**The Pile (from-scratch).** The from-scratch study samples documents from The Pile (Gao et al., 2020). A document-level deduplication pass is applied before scoring, and evaluation uses a held-out validation split. Each selection method trains for the same 2B-token budget. Figure 4d reports the induction-head prefix score over this training interval.

**OpenWebMath (continual pre-training, mathematics).** OpenWebMath contains 14.7B tokens of mathematical web text (Paster et al., 2024). We assign 13.2B tokens to the candidate pool and 1.5B

to held-out perplexity evaluation. GSM8K test examples are excluded from scoring and training. The synthesis prompts for this domain are listed in Section H.

**The Stack, Python subset (continual pre-training, code).** The Stack is a source-code corpus, and its Python subset consists of code files, not paired repository issues and solutions (Kocetkov et al., 2022). We filter generated and vendored files, parse the remaining code, and construct a 20B-token subset with 18B candidate and 2B held-out tokens. HumanEval is evaluated with its official test suite.

**PubMed abstracts (continual pre-training, biomedicine).** The biomedical pool is constructed from PubMed citation titles and abstracts (National Library of Medicine, 2025). The processed subset contains 4B tokens, divided into 3.6B candidate and 0.4B held-out tokens. MedQA is evaluated as multiple-choice accuracy without retrieval.

**MAGPIE (post-training).** This pool supplies the post-training candidate set examined in Figure 13 and the MAGPIE reference win rate of 58.9%. We take the released 300K-sample pool as given and draw our matched-sample subsets from it uniformly at random, so that the checklist arm and the spectral arm see the same generator and the same seed set, and differ only in which directions they are aimed at. The judge prompt and the tie rule follow Section G.7.

For the mixture forecasting study of Section 5.4, the 20 mixtures interpolate the three continual pre-training domains with general web text at ratios drawn from a Dirichlet prior, crossed with two quality tiers obtained by perplexity screening. Each mixture is priced with one scoring pass of the probe and then trained for 500M tokens to measure realized gains.

**Held-out mixture choice.** The decision study in Figure 5d uses a second candidate pool, sampled independently of the 20 mixtures in panels (a) to (c) under the same domain-ratio and quality-tier construction. A Pythia-14M probe prices each candidate at its 80K checkpoint, the frozen-price comparator retains its 5K score, and perplexity ranks the same candidates from low to high. For $k \in \{1, 2, 3, 4, 5, 6, 8, 10\}$, we normalize the weights of the top-$k$ mixtures under each ranking and train a separate 1B target from the same base checkpoint for 500M tokens. The outcome is absolute GSM8K accuracy rather than the gain over a base checkpoint in panels (a) to (c). Three independent training seeds supply each mean and sample-standard-deviation band.

### G.6 Baseline Implementations

All baselines receive identical token budgets, training schedules, and random seeds. **Random** samples uniformly from the pool. **Perplexity** keeps the examples with the lowest perplexity under the base model, following Marion et al. (2023). **DSIR** (Xie et al., 2023b) resamples by hashed n-gram importance weights toward the evaluation domain. **DoReMi** (Xie et al., 2023a) tunes domain weights with a 160M proxy model for 10K steps and transfers the weights to the target scale. **LESS** (Xia et al., 2024) stores LoRA gradient features of dimension 8,192 and selects by cosine similarity to validation gradients. **EK-FAC influence** (Grosse et al., 2023) uses the same curvature-sketch protocol as our probe with validation loss as its functional, but produces a signed probe-specific score. **FAC** (Li et al., 2026) trains a 16× overcomplete sparse autoencoder on layer 12 activations and selects greedily by marginal feature coverage. For synthesis, FAC-guided and spectrally aimed generation share the same generator model and decoding parameters, differing in target-feature selection.

Here we give the scoring functional each effect-based baseline uses, so that the comparison against Equation (6) is a literal one. Formally, given a pool $\mathcal{D}_{\text{pool}}$, a validation set $\mathcal{V}$ and a trained model $p_\theta$, every baseline assigns a scalar to a candidate $z$ and we select the top of the pool under its own score.

**DSIR** (Xie et al., 2023b) models a document by its hashed $n$-gram indicator vector $\phi(z)$ and scores it by the log importance ratio between the target and the general corpus,

$$v_{\text{DSIR}}(z) \;=\; \sum\nolimits_g \phi_g(z)\, \log \frac{\hat{p}_{\mathcal{V}}(\phi_g)}{\hat{p}_{\text{gen}}(\phi_g)}\,, \tag{16}$$

which is a property of the text alone and is blind to the model that will read it.

**LESS** (Xia et al., 2024) projects per-example gradients into a low-dimensional space with random projection $\Gamma$ and scores a candidate by its alignment with validation gradients accumulated over $T$

checkpoints,

$$v_{\mathrm{LESS}}(z) \;=\; \sum_{t=1}^{T} \sum_{x \in \mathcal{V}} \cos\big(\Gamma \nabla_\theta \ell(z, \theta_t),\; \Gamma \nabla_\theta \ell(x, \theta_t)\big). \tag{17}$$

The reference here is $\mathcal{V}$, frozen at the checkpoints, which is the difference Section 4.2 draws out.

**EK-FAC influence** (Grosse et al., 2023) is the score of Equation (3) evaluated on the same factors our probe maintains,

$$v_{\mathrm{IF}}(z) \;=\; -\, g_{\mathrm{sub}}(z)^\top \big(F_{\mathrm{sub}} + \varepsilon I\big)^{-1} \nabla_{\theta_{\mathrm{sub}}} \ell(\mathcal{V}, \theta)\,, \tag{18}$$

This baseline and `SDV` share a curvature sketch but apply different functionals to projected gradients: influence is signed and probe-dependent, while Equation (6) weights squared projections by curvature and saturation.

**FAC** (Li et al., 2026) scores a *set*, by the features it newly lights up,

$$V_{\mathrm{FAC}}(\mathcal{S}) \;=\; \frac{1}{|\mathcal{F}|} \Big| \Big\{\, i \in \mathcal{F} : \max_{z \in \mathcal{S}} a_i(z) > \tau_i \,\Big\} \Big|\,, \tag{19}$$

with $a_i(z)$ the $i$-th sparse code of $z$ and $\tau_i$ its firing threshold. Greedy selection maximizes the marginal gain of this set function. Proposition 3 relates this support statistic to projected gradient energy when feature and activation directions align. The two scores therefore test distinct information about each example.

### G.7 Evaluation Protocols

GSM8K is evaluated with 8-shot chain-of-thought prompting and exact match. Domain perplexity is measured on held-out splits of each pool. The induction-head prefix matching score follows Olsson et al. (2022), averaged over the top eight heads by score at each checkpoint, with the formation threshold set to $0.7$ of the final plateau. Post-training win rates use GPT-4 as judge with position debiasing, and a generation counts as a tie when the judge returns the same score for both.

**Reported metrics.** GSM8K is scored by exact match under 8-shot chain-of-thought prompting, HumanEval by pass@1 against the official test suite, and MedQA by accuracy against the multiple-choice key. Held-out domain perplexity is reported beside each accuracy because a method can improve one without the other. A circuit counts as formed once the prefix-matching score crosses $0.7$ of its final plateau, which is $0.56$ on the Pythia suite, and Table 6 sweeps the reported step counts over that threshold. Post-training quality is a head-to-head win rate: each candidate instruction set trains the same base model, that model is judged head to head against one fixed reference model trained on a small generic pool, positions are debiased by swapping the two candidates, and ties count as half a win, so every win rate in the paper, including the MAGPIE reading of $58.9$, is measured against that one reference.

**Subspace overlap and its random baseline.** The alignment diagnostics in Figure 11 measure overlap as $\frac{1}{k}\|P_1 P_2\|_F^2$ between rank-$k$ projectors, the mean squared cosine of their principal angles. Both projectors are represented in the retained spectral head of $r = 512$ sketch directions. The random baseline draws $k$ directions from this same space, giving expected overlap $k/r = 64/512 = 0.125$. Its empirical mean over 200 draws is $0.12$. The common ambient dimension makes the measured and random-subspace overlaps directly comparable.

### G.8 Rationale for the Emergence Threshold

Section 5.3 reports a $1.8\times$ acceleration measured at the $0.7$ plateau threshold, so the claim deserves a sweep over the threshold itself. Table 6 repeats the two crossings at five thresholds. The two step counts grow with the threshold, as any formation criterion must, and their ratio stays inside $[1.72, 1.93]$, so the acceleration is a property of the two data orderings. We keep $0.7$ because it is the value used by Olsson et al. (2022) to call a circuit formed.

Table 6: **Emergence threshold sweep.** The step at which each data ordering brings the induction-head prefix score across a given fraction of its final plateau of 0.80, and the ratio of the two. The 0.7 row is the default reported in the main text.

| Threshold | Score | Priced order | Corpus order | Ratio |
|---|---|---|---|---|
| 0.5× plateau | 0.40 | 1,464 | 2,832 | 1.93× |
| 0.6× plateau | 0.48 | 1,656 | 3,080 | 1.86× |
| 0.7× plateau | 0.56 | 1,840 | 3,312 | 1.80× |
| 0.8× plateau | 0.64 | 2,056 | 3,608 | 1.75× |
| 0.9× plateau | 0.72 | 2,344 | 4,032 | 1.72× |

### G.9 What Synthesis Buys When the Pool Is Thin

To examine synthesis when the candidate pool lacks high-value support, we remove the top decile of documents ranked by price on the targeted directions. The two variants then train on the same restricted pool and token budget. This intervention measures the contribution of generation after the most relevant pool examples have been removed.

Table 7: **Synthesis on a thinned pool.** Continual pre-training on mathematics at a 5% budget after removal of the top decile of target-direction prices. The final column repeats the full-pool GSM8K values from Table 2.

| Variant | GSM8K | PPL | Full-pool GSM8K |
|---|---|---|---|
| `SDV` (full) | 31.9 | 20.3 | 33.7 |
| ❹ w/o synthesis | 27.0 | 23.6 | 32.4 |
| Gap | 4.9 | 3.3 | 1.3 |

The absolute accuracy gap between the full method and the selection-only variant grows from 1.3% on the full pool to 4.9% on the restricted pool. The corresponding perplexity gap grows from 0.5 to 3.3. This pattern is consistent with the synthesis mechanism of Section 4.3: generated candidates contribute most when the available pool provides limited support for selected directions.

## H Prompt Set for Spectral Synthesis

Spectral synthesis of Section 4.3 proceeds in three stages. Unsaturated directions are first translated into natural language descriptions, contrastive exemplar pairs are then mined from the pool, and the generator finally produces new examples. The three prompts are listed verbatim below. Placeholders in {`braces`} are filled programmatically.

**Prompt H.1: Feature Description**

```
You are analyzing an internal feature of a language model.  Below are
{n} text snippets that activate this feature most strongly, each with
the activating tokens marked by **asterisks**.

Snippets:
{top_activating_snippets}

Describe in one sentence the shared pattern that these snippets
express.  Focus on the structural or semantic property that the marked
tokens have in common.  Answer with the description only.
```

**Prompt H.2: Contrastive Exemplar Mining**

```
You are given a feature description and two text snippets.

Feature: {feature_description}
Snippet A: {positive_snippet}
Snippet B: {negative_snippet}

Snippet A expresses the feature and Snippet B is a minimally different
text that does not. Explain in one sentence what was changed between
B and A to express the feature. Answer with the explanation only.
```

**Prompt H.3: Targeted Generation**

```
You are a data synthesis engine. Your goal is to write training
documents that express a specific pattern.

Pattern: {feature_description}
Here are {m} example pairs. In each pair, the second text expresses
the pattern and the first does not:
{contrastive_pairs}

Write {j} new documents of 300 to 600 words. Each document must
express the pattern naturally and repeatedly, vary in topic and
register, and avoid copying phrases from the examples. Separate
documents with the delimiter <DOC>.
```

Generated documents are kept only if their measured price on the targeted directions exceeds the threshold $\tau = 4.5$, which filters out generations that describe the pattern without expressing it. The generator is `Qwen2.5-72B-Instruct` with temperature $0.9$ and nucleus sampling $p = 0.95$.

# I Additional Results

This appendix collects results deferred from the main text.

## I.1 Full Quantitative Results

**Complete budget sweep.** Table 1 reports the $5\%$ budget. The complete sweep covers all three budgets.

Table 8: **Full budget sweep.** Domain accuracy at $25\%$ and $50\%$ token budgets, with ARC and MMLU measured after the $25\%$ run. The $5\%$ comparison appears in Table 1. DoReMi has the highest ARC and MMLU scores, and `SDV` has the highest mathematics and code scores at both budgets.

| | Mathematics | | Code | | General | |
|---|---|---|---|---|---|---|
| **Method** | GSM8K ↑ (25%) | GSM8K ↑ (50%) | HumanEval ↑ (25%) | HumanEval ↑ (50%) | ARC ↑ (25%) | MMLU ↑ (25%) |
| Random | 27.4 | 31.8 | 26.8 | 29.7 | 34.1 | 32.4 |
| Perplexity | 28.9 | 34.6 | 27.9 | 32.0 | 33.4 | 32.0 |
| DSIR | 29.8 | 35.3 | 28.6 | 32.7 | 34.6 | 32.7 |
| DoReMi | 30.6 | 35.9 | 29.3 | 33.2 | **35.2** | **33.1** |
| LESS | 31.7 | 36.8 | 30.4 | 34.1 | 34.8 | <u>32.9</u> |
| EK-FAC Influence | <u>33.5</u> | <u>37.9</u> | <u>31.9</u> | <u>35.3</u> | 33.9 | 32.2 |
| FAC | 32.8 | 37.4 | 31.5 | 34.9 | 33.6 | 32.1 |
| `SDV` (Ours) | **37.2** | **40.6** | **34.8** | **37.6** | <u>34.9</u> | 32.8 |

**Budget-dependent results.** Table 8 places `SDV` first in the four mathematics and code comparisons at $25\%$ and $50\%$ of the candidate pool. The gap to the strongest matched method is $3.7\%$ on GSM8K at the $25\%$ budget and $2.9\%$ on HumanEval at that budget. On the two general suites, DoReMi leads by $0.3\%$ at $25\%$. These rows give the domain-level context for the mixture-ranking analysis below.

**Scoring-budget sensitivity.** For the 14M target in Figure 5a, the forecast correlations at $0.5$, $1$, $2$, $5$, $10$, $20$, and $50$M scored tokens are $0.61$, $0.74$, $0.83$, $0.89$, $0.93$, $0.93$, and $0.92$. The default

10M scoring pass reaches the plateau. On those same 20 mixtures, perplexity and token-count proxy correlations are 0.58 and 0.41.

**Full mixture forecasting table.** Section 5.4 reports rank correlations over 20 mixtures. The full table lists every mixture with its composition, predicted price, and realized gain.

Table 9: **Full mixture forecasting.** All 20 mixtures with their domain ratios, quality tier, the spectral price assigned by a single scoring pass of the Pythia-14M probe, and the realized GSM8K gain over the base checkpoint after 500M tokens of training at 14M and at 1B scale. **Rank shift** counts how far a mixture moves between the ranking implied by the 14M price and the ranking realized at 1B, so a small value is the transfer evidence quoted in Section 5.4. The two $\rho$ rows are Spearman correlations over all 20 mixtures, and the rank shift averages 2.1 positions.

| Mix | Math/Code/Bio | Tier | Price ($\times 10^{-3}$) | GSM8K $\Delta$ (14M) | GSM8K $\Delta$ (1B) | Rank shift |
|---|---|---|---|---|---|---|
| 1 | 0.4/0.3/0.3 | high | 1.62 | 6.9 | 10.8 | 0 |
| 2 | 0.1/0.6/0.3 | low | 2.05 | 7.2 | 12.7 | 2 |
| 3 | 0.6/0.2/0.2 | high | 2.31 | 11.3 | 12.6 | 0 |
| 4 | 0.2/0.2/0.6 | low | 2.74 | 10.7 | 11.9 | 2 |
| 5 | 0.5/0.4/0.1 | high | 3.02 | 11.2 | 22.3 | 4 |
| 6 | 0.3/0.5/0.2 | high | 3.28 | 10.5 | 15.0 | 1 |
| 7 | 0.2/0.7/0.1 | low | 3.55 | 10.7 | 17.1 | 1 |
| 8 | 0.7/0.2/0.1 | high | 3.81 | 10.5 | 19.4 | 1 |
| 9 | 0.4/0.4/0.2 | low | 4.06 | 13.1 | 24.2 | 4 |
| 10 | 0.5/0.3/0.2 | high | 4.28 | 14.5 | 25.5 | 5 |
| 11 | 0.6/0.3/0.1 | high | 4.51 | 14.5 | 23.8 | 0 |
| 12 | 0.1/0.4/0.5 | low | 4.79 | 18.7 | 24.4 | 2 |
| 13 | 0.8/0.1/0.1 | high | 5.02 | 19.4 | 20.1 | 5 |
| 14 | 0.3/0.3/0.4 | high | 5.34 | 17.4 | 27.0 | 2 |
| 15 | 0.2/0.3/0.5 | low | 5.61 | 19.2 | 23.3 | 5 |
| 16 | 0.4/0.5/0.1 | high | 5.88 | 18.6 | 31.1 | 1 |
| 17 | 0.5/0.2/0.3 | high | 6.19 | 18.9 | 24.1 | 5 |
| 18 | 0.6/0.1/0.3 | high | 6.52 | 21.7 | 33.6 | 1 |
| 19 | 0.1/0.3/0.6 | low | 6.94 | 21.3 | 32.8 | 1 |
| 20 | 0.7/0.1/0.2 | high | 7.31 | 23.4 | 35.8 | 0 |
| $\rho$ over the 20 mixtures | `SDV` price vs realized gain | | | **0.93** | **0.88** | mean **2.1** |

**Cost accounting.** Table 10 reports the unit scoring costs for the methods compared in RQ4.

Table 10: **Cost accounting.** Wall-clock time, peak memory and storage of each valuation method for pricing one million candidate tokens on a single node, broken into probe maintenance, eigendecomposition and per-candidate scoring. **Extra model pass?** records whether candidate scoring requires a pass beyond the training update. `SDV` maintains its own factors and computes candidate gradients. These unit costs exclude the initial pilot, full-pool scan, and within-round re-ranking.

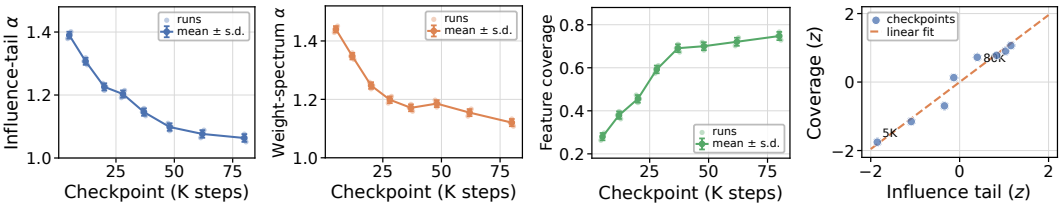
| Method | Time / $10^6$ tok | Peak memory | Storage / layer | Extra model pass? |
|---|---|---|---|---|
| TRAK | 41.3 min | 6.2 GB | 1.80 GB | yes |
| LESS | 12.8 min | 3.4 GB | 0.90 GB | yes (LoRA) |
| EK-FAC Influence | 6.4 min | 2.1 GB | 0.40 GB | yes |
| FAC | 9.7 min | 4.8 GB | 1.20 GB | yes (SAE) |
| `SDV` (Ours) | **1.1 min** | **0.3 GB** | **0.10 GB** | yes (separate factors) |

### I.2 Cost and Sensitivity Diagnostics

Table 10 uses a common candidate stream on one node. Time includes the candidate forward and backward passes required by each method, but excludes training updates, preparation of any fixed validation set, and repeated re-ranking after provisional-profile changes. Peak memory is measured above the training model footprint. The sensitivity curves in Figure 12 extend the main-text rank sweep to refresh period, EMA coefficient, and capacity constant. A full deployment estimate must multiply the unit scoring cost by the number of candidate tokens examined and add pilot and re-ranking time.

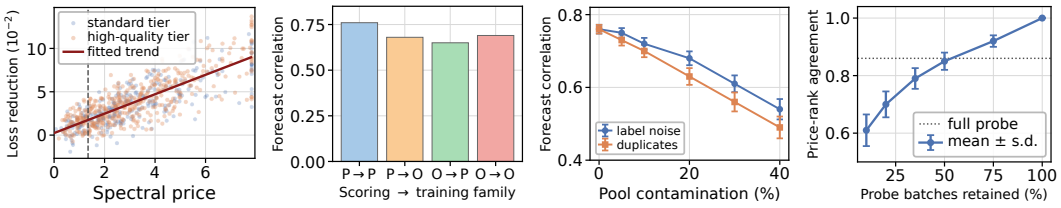
### I.3 Co-evolution of the Three Tails

RQ1 evaluates whether the price predicts subsequent learning. This complementary diagnostic follows the geometric quantities used to construct that price and asks whether their concentration changes together over training.



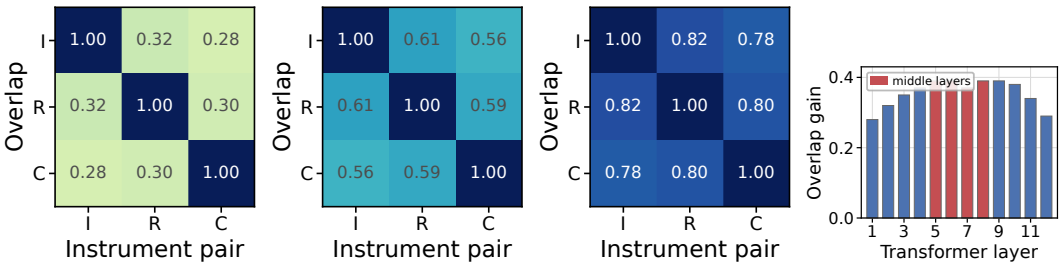

(a) Influence tail (b) Weight spectrum (c) Coverage saturation (d) Cross-instrument association

Figure 9: **Co-evolution of three geometric measurements** on Pythia-14M. **(a)** The power-law exponent $\alpha$ of the influence-score tail and **(b)** the exponent of the empirical weight-spectrum tail, each estimated at eight checkpoints. Smaller exponents indicate heavier tails. **(c)** The fraction of task-relevant sparse features activated by the growing dataset. Light points in **(a)** to **(c)** represent three runs, and connected points and whiskers show the checkpoint mean and spread. **(d)** Checkpoint-level association between sign-aligned influence-tail concentration and feature coverage after $z$-scoring, with a linear fit.



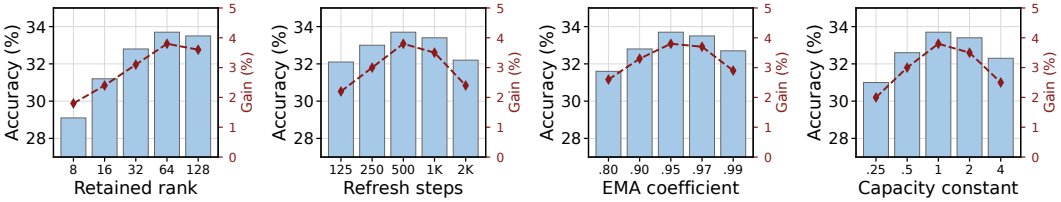

(a) Pythia-160M prediction (b) Transfer across families (c) Validity under contamination (d) Price stability under subsampling

Figure 10: **Robustness of prospective valuation. (a)** Candidate prices against subsequent-window loss reduction at Pythia-160M, complementing the two scales in Figure 3a,b. **(b)** Correlation when the scoring model and the model used to realize the gain come from the same or different families. **(c)** Correlation after injecting label noise or exact duplicates into the candidate pool, with whiskers showing the spread across runs. **(d)** Agreement between prices from a subsampled probe and the full probe. The four panels retain the prospective protocol of RQ1 while changing one source of measurement variation at a time.




(a) Heatmap at step 5K (b) Heatmap at step 30K (c) Heatmap at step 70K (d) Overlap gained by layer

Figure 11: **Per-layer and per-checkpoint alignment** on Pythia-160M, using the rank-64 projectors defined in Section G.7. **(a)** to **(c)** Pairwise overlap between influence, rotation, and coverage at steps 5K, 30K, and 70K. **(d)** Increase in three-way overlap between steps 5K and 70K for each transformer layer. Dark-red bars identify the middle layers. The first three panels show the checkpoint snapshots, and the fourth localizes where alignment changed.




(a) Accuracy and absolute gain against rank (b) Accuracy against refresh period (c) Accuracy against EMA coefficient (d) Accuracy against capacity constant

Figure 12: **Probe sensitivity sweeps**, each knob varied with the others held at their defaults, on the mathematics domain at the 5% budget. **(a)** to **(d)** Endpoint GSM8K accuracy for retained rank, refresh period, EMA coefficient, and capacity constant. The dark-red line reports the absolute improvement over the matched random-selection run on the right axis. Performance increases around the deployed operating region and then plateaus or recedes.

### I.4 Robustness and Transfer

**Prospective validity beyond one model family.** We repeat the prediction experiment of RQ1 across model families, candidate-pool perturbations, and probe sizes. The Pythia-160M candidate-level result appears in Figure 10a, with rank correlation 0.78. The family-transfer bars in panel **(b)** range from 0.65 to 0.76. Panel **(c)** follows the loss of correlation as label noise and duplicates enter the pool, while panel **(d)** shows how price-rank agreement changes as more probe batches are retained.

**Per-layer and per-checkpoint alignment.** We next inspect the instrument geometry used to construct the price. The three pairwise overlaps in Figure 11 rise from 0.28–0.32 at step 5K to 0.78–0.82 at step 70K. Panel **(d)** assigns the late-minus-early gain to individual transformer layers and shows pronounced changes through the middle of the network. These checkpoint and layer measurements characterize where the three instruments align during training.

**Sensitivity sweeps.** Beyond the ablations of Section 5.5, we sweep each probe knob independently. In Figure 12a, accuracy rises through retained rank 64 and changes little at rank 128. The illustrated refresh, EMA, and capacity sweeps attain their highest GSM8K values at the deployed settings of 500 steps, 0.95, and 1, respectively. The corresponding absolute gains against matched random selection are plotted on the right axes.

### I.5 Qualitative Analyses

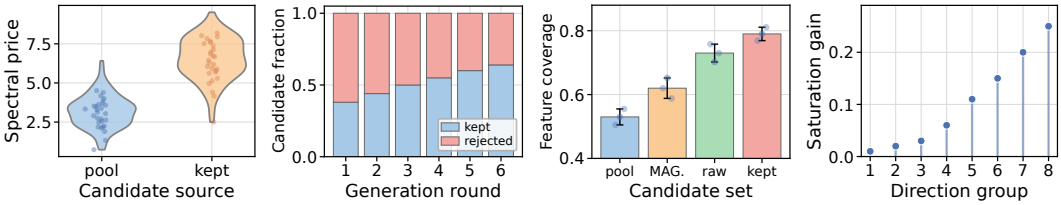
**State-dependent price trajectories.** The repetitive document declines from 7.13 at step 500 to 0.91 at step 6,000. The compositional example rises from 1.0 to 6.6 and crosses the declining trajectory

near step 2,750. The mixed syntactic example reaches 4.6 near step 3,000 before falling to 1.9, while the web narrative remains close to 2.1. These trajectories provide the measurements summarized by the direction arrows in Table 11 and show how the ordering changes as the basis and saturation profile are updated.

**Qualitative excerpts of the discovered curriculum.** Section 5.5 describes how the highest priced examples change as training proceeds. Representative excerpts at three developmental stages are shown below.

Table 11: **Discovered curriculum.** Top-priced examples at early, middle and late stages of from-scratch pre-training, with the price they carried at the stage shown, the retained direction that dominates that price, and the direction the price moved afterwards. The fourth row is the control, an example whose price never moves and which therefore pays into no direction in particular. Prices are in the same $10^{-3}$ units as Table 9. The reversal in the last column is the phenomenon. The same text is worth eight times more at one developmental state than at another.

| Stage | Excerpt | Price | Top direction | Trajectory |
|---|---|---|---|---|
| Early | *"and so it goes: $1+1=2$, and $2+2=4$, and $4+4=8$, and so on, and so on, until the pattern is the whole of the sentence."* | 7.1 | $u_7$ (attn., layer 5) | ↘ to 0.9 |
| Middle | *"Although the committee, which had met twice before, approved the measure, the dissenters, whose names were read aloud, declined to sign."* | 4.6 | $u_{19}$ (mlp., layer 8) | ↗ then ↘ to 1.9 |
| Late | *"If every box holds at most two keys and no two shelves share a colour, how many ways can the third shelf fail?"* | 6.6 | $u_{31}$ (attn., layer 10) | ↗ from 1.0 |
| Control | *"The harbour town wakes early. Boats return with the night's catch, and the market opens by seven, as it does every morning."* | 2.1 | no dominant direction | → flat |



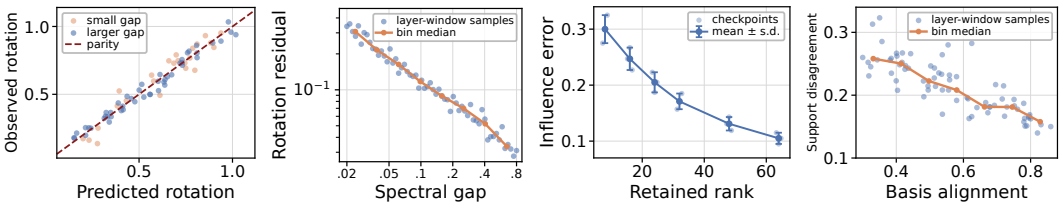

(a) Price by source (b) Filtering across rounds (c) Feature-space coverage (d) Saturation gain by direction

Figure 13: **Anatomy of spectral synthesis** in the post-training regime. **(a)** Price distributions of pool examples and kept generations on the same targeted directions, with the horizontal mark showing the mean. **(b)** Fractions retained and rejected at $\tau = 4.5$ in six generation rounds, shown as complementary stacks. **(c)** Coverage of a fixed SAE dictionary at an equal candidate count for the pool, MAGPIE, unfiltered spectral generations, and filtered generations. Whiskers show run-to-run spread. **(d)** The change $s_i^{\text{after}} - s_i^{\text{before}}$ after training on synthesized data, grouped by descending eigenvalue rank. The largest changes occur in the less-saturated tail directions. Coverage counts distinct features, not distinct strings.

**Reading the synthesis diagnostics.** The candidate-price distribution, retained fraction, and equal-count feature coverage measure different parts of the generation loop. A rising retained fraction alone does not determine coverage, so panel **(c)** compares the candidate sets after fixing their size. The directional saturation gains in panel **(d)** then locate where the filtered candidates contribute to the current model state.

Table 12: **Synthesis excerpts.** Side by side excerpts targeting the same unsaturated direction, one from spectrally aimed generation and one from checklist-guided generation, with the price each carries on that direction. The direction is described by Section H Prompt H.1 as *a numeric continuation restated under an increasing restriction*. Both pipelines see the same contrastive exemplars and the same generator, so the only difference is which directions the target is chosen from. The two price columns are the individual measurements behind the distributions in Figure 13a.

| **Spectral** | **Checklist** | **Price (sp.)** | **Price (ck.)** |
|---|---|---|---|
| *"Start at* 3 *and add* 3 *each time:* $3, 6, 9$. *Keep only odd terms, which leaves* 3 *and* 9. *Requiring a term greater than* 5 *leaves* 9*."* | *"A sequence is a list of numbers in order. Arithmetic sequences add the same amount each time. For example* $2, 4, 6, 8$ *adds two. Sequences appear in many places in mathematics and in everyday life."* | 6.8 | 2.1 |
| *"A strip begins as one layer. Folding it in half gives* 2*, then* 4*, then* 8 *layers. The first fold count that exceeds* 6 *layers is therefore three."* | *"Folding paper creates creases. Each fold doubles the number of layers. This is an example of exponential growth, which is important in many subjects."* | 7.4 | 1.8 |
| *"A bus departs every* 12 *minutes and a tram every* 8*. Their first simultaneous departure after time zero is at minute* 24*. A rider required to arrive before the second bus departure has no positive simultaneous departure to choose."* | *"To find a common multiple, list the multiples of each number and look for the smallest one they share. This is called the least common multiple, or LCM."* | 5.9 | 2.6 |




(a) Predicted against observed rotation (b) Rotation residual (c) Influence residual (d) Coverage residual

Figure 14: **Residual diagnostics** for the conditional relations of Section D. **(a)** Observed against estimated rotation for 72 layer-checkpoint pairs, grouped by singular gap $\gamma$, with parity marked. **(b)** Rotation residuals against $\gamma$ on log-log axes, with an empirical bin median. **(c)** Influence-score truncation error at six retained ranks, with checkpoint-level points and spread. **(d)** Feature-support disagreement against measured SAE feature-basis alignment for layer-checkpoint pairs, with a bin median. Panels **(b)** to **(d)** examine the distinct conditions behind Propositions 1 to 3.

**Reading the residual diagnostics.** The singular gap, retained rank, and dictionary-basis alignment enter different conditional relations. The rotation comparison in panels **(a,b)**, influence truncation in panel **(c)**, and feature-support comparison in panel **(d)** therefore provide separate checks of the quantities used by the three propositions.