Distribution-Robust Functional Subset Projection for Structured Neural Network Width Compression

Kenju Tomita
Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA
Ephemerent LLC, Tarrytown, NY, USA
Corresponding author: kenju.tomita@gmail.com

August 5, 2026

Abstract

Purpose: Post-training width compression can mistake components for redundant when calibration data underweights a rare environment, or treat redundant components as different because of finite-sample noise. We study whether functional, multi-environment subset selection and hierarchical grouping can separate these effects.

Methods: Distribution-robust functional subset projection (DR-FSP) represents each hidden unit by its activation profile and outgoing write vector, selects a retained functional basis, and globally refits the writes under target-energy-normalized mean or worst-environment reconstruction. A nested average-linkage hierarchy is evaluated with raw fingerprints and an optional bootstrap-Gaussian regularizer.

Results: In a six-coordinate shift construction over 20 seeds, source and prevalence-weighted selection retain 0/120 stable coordinates, whereas minimax selection retains 120/120 and matches a 729-subset blockwise-exact reference. In a 100-unit controlled layer, the raw and Gaussian hierarchies with the same 5% energy floor recover 50/50 dominant families at width 50; 70 nonzero measured directions show that this is an operating point rather than algebraic rank. A repeated GPT-2 Small diagnostic across nine layer–split trials is mixed: the raw hierarchy is competitive at mild reduction and Gaussian refinement is worse on average.

Conclusion: The evidence supports a controlled mechanism study of robust functional selection and refitting, not a state-of-the-art language-model compressor or a claim of hardware speedup, universal shift robustness, or depth compression.

Keywords: structured neural network compression; distribution shift; robust optimization; functional representations; neuron merging; language model compression

1 Introduction

A structured compression decision is a counterfactual claim: after a hidden component is deleted or merged, can the retained network still produce the computation that matters at deployment? Raw parameter magnitude is an incomplete answer. A hidden unit both detects an input-dependent scalar feature and writes a vector into the next representation. For environment e, unit i has empirical contribution

$$F_{e,i} = a_{e,i} v_i^\top, \tag{1}$$

where $a_{e,i} \in \mathbb{R}^{n_e}$ is its activation profile over n_e calibration examples and $v_i \in \mathbb{R}^d$ is its outgoing write vector. This is a rank-one vector-valued operator. The operator perspective is closely related to neuron merging and to HOPE's Hilbert–Schmidt formulation of pruning, merging, and block removal [1, 2].

Finite calibration creates two logically different problems. First, two units can appear redundant because the calibration objective underweights an important environment. Second, two genuinely redundant units can appear distinct because their measured functional fingerprints vary under resampling. The first problem concerns *which basis survives*; the second concerns *how that basis is estimated*. Conflating them can make an improved refit look like an improved selector, or make a denoiser look useful when a separate threshold is doing the work.

This paper develops and audits a unified pipeline for these questions. The primary method uses raw vector-valued functional geometry, hierarchical grouping, and a multi-environment compensated refit. A bootstrap-Gaussian refinement is included as an optional statistical regularizer. Importantly, repeated GPT-2 experiments do not support presenting that regularizer as uniformly beneficial. The negative result is retained and used to narrow the paper's claims.

Contributions.

1. We formulate target-energy-normalized multi-environment subset projection and derive the weighted ridge normal equations with coefficients $q_e/(n_e d_e)$.
2. We separate robust subset selection from robust refitting in a factorial controlled experiment and compare greedy selection with a 729-subset blockwise-exact reference.
3. We construct compact vector-valued functional fingerprints and one nested hierarchy that can be cut at several widths, followed by a global refit of the selected representatives.
4. We audit an optional bootstrap-Gaussian regularizer, an explicit energy floor, and a raw spectral-rank baseline separately. This distinguishes measured rank, dominant functional families, and a chosen rate–distortion operating width.
5. We report a repeated GPT-2 Small diagnostic across three layers and three disjoint corpus offsets. Fixed Gaussian refinement is heterogeneous and worse on average; raw functional geometry remains competitive but does not dominate simpler baselines.
6. We release a reproducibility archive with code, raw per-trial outputs, tests, dependency records, the exact model revision, and SHA-256 checksums.

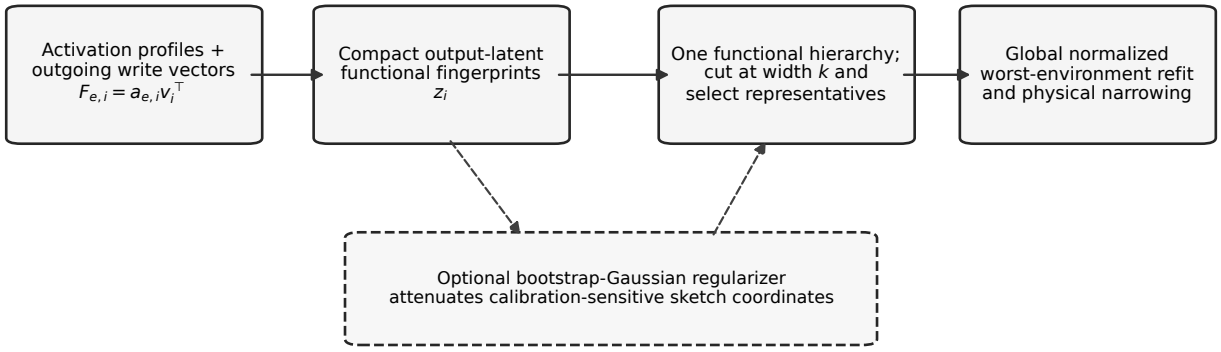

Fig. 1: DR-FSP with hierarchical functional grouping. The primary path uses raw vector-valued functional fingerprints, one hierarchy, representative selection, and a global worst-environment refit. Bootstrap-Gaussian refinement is optional and is evaluated as an ablation rather than assumed beneficial.

2 Related Work

Projection and compensation. Neuron Merging compensates deleted neurons through retained neighbors [1]. HOPE models neurons as rank-one Hilbert–Schmidt operators and unifies pruning, merging, and residual-pathway eviction in functional space [2]. Model folding, DOTResize, mechanism sparsification, and GRAIL likewise replace uncompensated deletion with projection, aggregation, transport, or post-hoc linear reconstruction [3–6]. DR-FSP does not claim novelty for projection itself. Its narrower focus is environment-indexed basis selection, explicit separation of selection from refitting, and hierarchical reuse of a functional geometry across target widths.

Structured language-model compression. OSSCAR optimizes one-shot structured layer reconstruction through combinatorial search [7]. LLM-Pruner, FLAP, and SliceGPT remove structured channels or residual dimensions [8–10]. SparseGPT and Wanda are influential one-shot or saliency-based references [11, 12]. These systems differ in granularity, recovery compute, kernel structure, and whether they preserve an ordinary dense narrower layer. Our GPT-2 study is diagnostic and is not a complete comparison with this mature literature.

Calibration shift and worst-group objectives. Calibration data can materially change post-training pruning and quantization results [13]. Group distributionally robust optimization protects the worst named group but can overfit noisy minority estimates without sufficient regularization [14]. DR-FSP adopts the same worst-environment principle for a discrete hidden-channel subset followed by a dense refit.

Noisy low-rank structure. Probabilistic PCA, singular-value shrinkage, ScreeNOT, Bayesian-bootstrap weighting, and random projections provide relevant statistical ingredients [15–19]. Our Gaussian step is only a working coordinatewise regularizer. It is not an optimal shrinker and does not satisfy the assumptions required by those theories.

3 Multi-Environment Functional Subset Projection

For environment $e \in \{1, \dots, E\}$, let

$$A_e \in \mathbb{R}^{n_e \times m}, \qquad V \in \mathbb{R}^{m \times d}, \qquad Y_e = A_e V. \tag{2}$$

Rows index calibration examples and columns index hidden units. We retain $S \subseteq \{1, \dots, m\}$ with $|S| = k < m$ and fit $B \in \mathbb{R}^{k \times d}$ so that $A_{e,S}B$ approximates the original vector-valued layer contribution Y_e.

3.1 Normalized environment risk

Different environments can have different output energies. Define

$$d_e = \frac{1}{n_e}\|Y_e\|_F^2 + \epsilon, \qquad R_e(S, B) = \frac{\|Y_e - A_{e,S}B\|_F^2}{n_e d_e}. \tag{3}$$

R_e is dimensionless, equals zero under exact reconstruction, and is approximately one for the zero predictor. Values above one are valid: the compressed layer then creates more squared error than outputting zero relative to that environment's target energy.

We distinguish

$$J_{\text{source}}(S, B) = R_1(S, B), \tag{4}$$

$$J_{\text{mean}}(S, B) = \sum_e \pi_e R_e(S, B), \tag{5}$$

$$J_{\max}(S, B) = \max_e R_e(S, B), \tag{6}$$

where π is a deployment-prevalence vector. The gap

$$G_\pi(S, B) = \max_e R_e(S, B) - \sum_e \pi_e R_e(S, B) \tag{7}$$

measures how much a prevalence-weighted average hides the worst named environment.

3.2 Weighted dense refit

For a fixed subset S and environment weights $q \in \Delta^{E-1}$, minimize

$$\mathcal{L}_q(B; S) = \sum_e q_e R_e(S, B) + \lambda\|B - B_0\|_F^2. \tag{8}$$

Proposition 1 (Target-energy-normalized ridge refit)**.** *Let $\alpha_e = q_e/(n_e d_e)$. For $\lambda > 0$, the unique minimizer satisfies*

$$\left(\sum_e \alpha_e A_{e,S}^\top A_{e,S} + \lambda I\right) B_q^*(S) = \sum_e \alpha_e A_{e,S}^\top Y_e + \lambda B_0. \tag{9}$$

For $\lambda = 0$, a Moore–Penrose solution yields a minimum-norm least-squares minimizer when the weighted design is rank deficient.

The factor d_e^{-1} is required because the optimized risk is normalized by target energy. The implementation is tested against an independently stacked weighted least-squares system. For a fixed subset, approximate minimax refitting alternates the exact solve in Eq. 9 with exponentiated-gradient ascent on q and returns the lowest observed maximum-risk iterate. This is an approximation, not a general saddle-point certificate.

3.3 Selection versus refitting

Selection chooses the activation basis $A_{e,S}$; refitting changes only its outgoing coefficients B. If selection omits a stable activation direction, no coefficient update can generally synthesize that missing direction. Our first controlled experiment crosses source, mean, and minimax selection with mean and minimax refitting to isolate this distinction.

4 Hierarchical Functional Compression

4.1 Compact operator fingerprints

The full operator $F_{e,i} = a_{e,i} v_i^\top$ can be large. Let $P \in \mathbb{R}^{d \times r}$ contain the top r right-singular directions of concatenated layer outputs, and let $G_e \in \mathbb{R}^{p \times n_e}$ have independent entries

$$(G_e)_{\ell t} \sim \mathcal{N}\left(0, \frac{1}{p n_e}\right). \tag{10}$$

We define

$$z_{e,i} = \frac{\operatorname{vec}\left[(G_e a_{e,i})(P^\top v_i)^\top\right]}{\sqrt{d_e}} \in \mathbb{R}^{pr}, \tag{11}$$

and concatenate environments to form z_i.

Proposition 2 (Expected projected functional inner product)**.**

$$\mathbb{E}_G[z_{e,i}^\top z_{e,j}] = \frac{a_{e,i}^\top a_{e,j}}{n_e d_e}\, v_i^\top P P^\top v_j. \tag{12}$$

Thus two units are similar only when both their activation profiles and their write directions align in the retained output subspace. The sketch is computational; it does not decode a human-readable concept.

4.2 Raw hierarchy and representatives

The primary hierarchy uses cosine distance

$$D_{ij}^{\text{raw}} = 1 - \frac{z_i^\top z_j}{\|z_i\|_2 \|z_j\|_2 + \epsilon}. \tag{13}$$

Average linkage produces one dendrogram. Cutting it at width k yields k clusters. One representative is selected from each cluster by centrality in functional space, and all retained representatives are jointly refit with Eq. 9. Cluster partitions are nested, but the selected neuron identities need not be nested because the medoid is recomputed after each cut.

4.3 Optional bootstrap-Gaussian regularizer

To estimate calibration sensitivity, rows receive independent exponential weights normalized to mean one, producing bootstrap fingerprints $z_{ij}^{(b)}$. Define

$$\bar{z}_{ij} = \frac{1}{B_{\text{boot}}} \sum_b z_{ij}^{(b)}, \qquad s_{ij}^2 = \frac{1}{B_{\text{boot}} - 1} \sum_b (z_{ij}^{(b)} - \bar{z}_{ij})^2. \tag{14}$$

The sample variance is *not* divided by B_{boot}: the number of bootstrap draws controls Monte Carlo precision, not the amount of information in the calibration set. With variance temperature γ,

$$\omega_{ij}^2 = \gamma s_{ij}^2, \qquad \widehat{\tau}_j^2 = \max\left\{ \text{Var}_i(\bar{z}_{ij}) - \frac{1}{m} \sum_i \omega_{ij}^2, 10^{-12} \right\}, \tag{15}$$

we use the working normal-means shrinkage

$$\mu_{ij} = \frac{\widehat{\tau}_j^2}{\widehat{\tau}_j^2 + \omega_{ij}^2} \bar{z}_{ij}, \qquad \nu_{ij} = \frac{\widehat{\tau}_j^2 \omega_{ij}^2}{\widehat{\tau}_j^2 + \omega_{ij}^2}. \tag{16}$$

This is a regularizing approximation, not a calibrated posterior over the population operator. The Gaussian hierarchy uses an uncertainty-adjusted cosine based on μ_i and ν_i. Because the real-model evidence is heterogeneous, the Gaussian geometry is reported as an ablation rather than the default method.

4.4 Energy policies and three notions of width

Zero-write units can be removed exactly. A separate declared energy policy may remove low-energy nonzero units. In the controlled 100-unit experiment we use a 5% fraction of the median positive energy; this threshold is intentionally *not* transferred to GPT-2 because it was not independently validated there.

We distinguish three quantities:

1. **Measured nonzero rank**: the number of nonzero singular directions in the empirical sketch.
2. **Dominant-family width**: the number of planted high-energy functional families in the controlled construction.
3. **Operating width**: a chosen point on a rate–distortion curve.

A tree-gap stopping heuristic and a raw singular-value-gap baseline are diagnostics, not certified rank estimators.

4.5 Computational cost

Forming projected fingerprints costs approximately $O(\sum_e pn_e m + mdr)$ before pairwise comparisons. Exact pairwise distances and average-linkage clustering require $O(m^2 Epr)$ time and $O(m^2)$ memory. The global dense refit scales with the retained width and total calibration rows. The current prototype is therefore suitable for single-layer studies and requires blocking, screening, approximate nearest-neighbor search, or low-rank updates for substantially wider models.

5 Controlled Experiments

5.1 Rare-environment selection and refitting

The first construction has six latent coordinates, each with two fragile candidates and one stable candidate. Two common environments receive prevalence 0.499 each; a failure environment receives prevalence 0.002. Each environment nevertheless has 320 calibration examples, so prevalence is separated from sample count. There are 20 seeds and six coordinate decisions per seed, for 120 stable-selection opportunities.

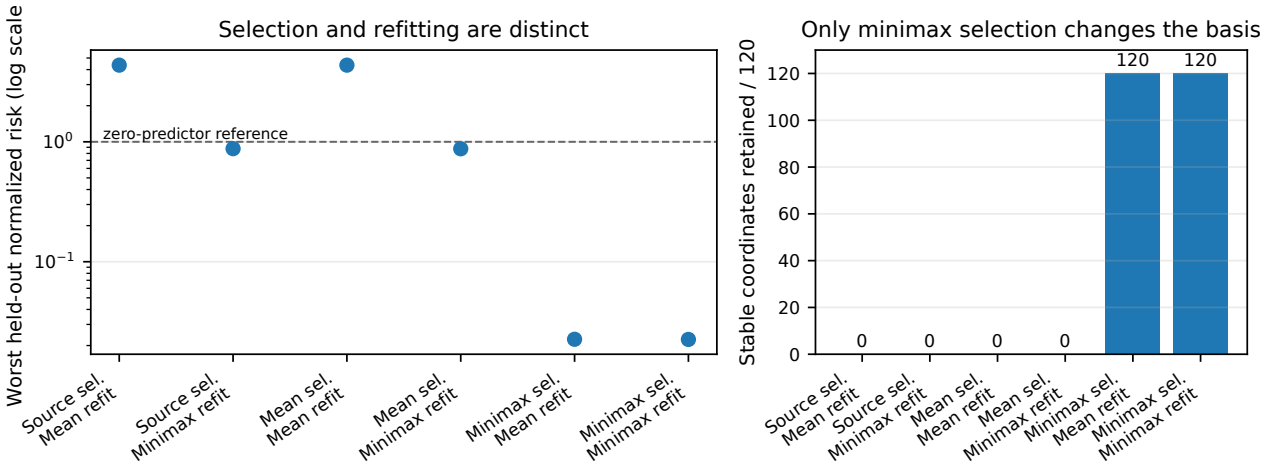

Fig. 2: Rare-environment factorial. Error bars are sample standard deviations over 20 seeds. A minimax refit substantially reduces the damage of a fragile subset but cannot reconstruct a stable activation direction omitted during selection.

Source and prevalence-weighted mean selection retain 0/120 stable coordinates. Minimax selection retains 120/120, reaches worst held-out risk 0.02249 ± 0.00042, and matches the 729-subset blockwise-exact solution in all 20 seeds. Source/mean and mean/mean risks exceed one because their fitted fragile coefficients amplify the shifted nuisance and become worse than the zero predictor. This establishes a mechanism in the constructed problem, not its prevalence in natural models.

5.2 Overcomplete 100-unit layer

The second construction contains 50 dominant family representatives, 25 exact positive-rescaling clones, 15 nuisance-bearing near-clones, five zero-write distractors, and five low-energy nonzero auxiliary directions. Consequently, it has 50 dominant families but 70 measured nonzero directions. All selectors retain 50 units and receive the same minimax outgoing refit. Results are mean $\pm$ sample standard deviation over 50 seeds.

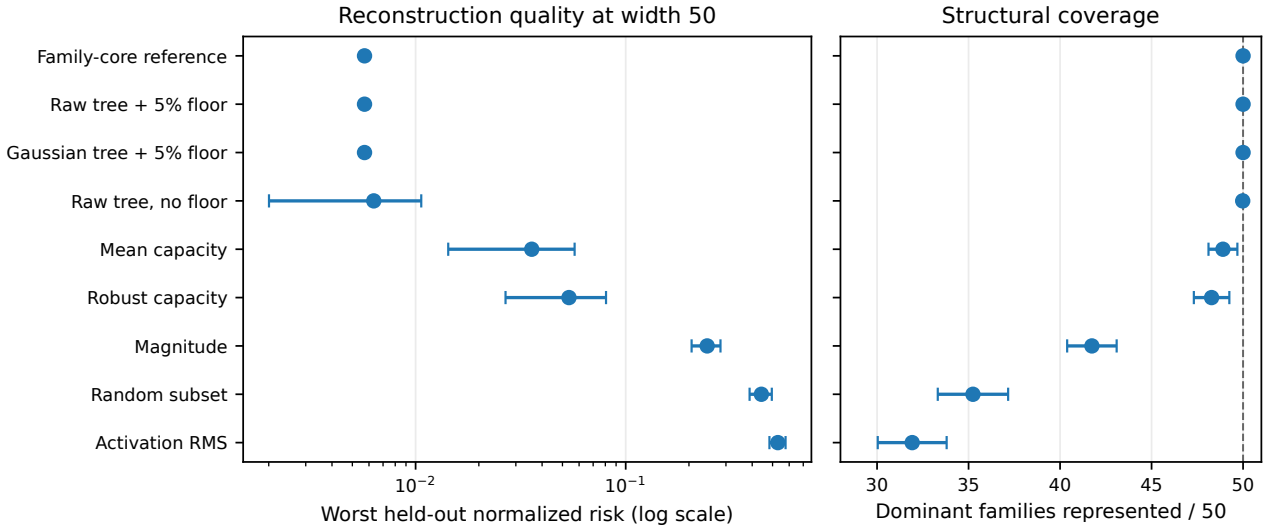

Fig. 3: Controlled 100-to-50 compression. Functional hierarchies with the same declared 5% energy floor recover all 50 dominant families. Activation strength alone retains null and low-energy units because it ignores the outgoing write.

Table 1: Controlled 100-to-50 results over 50 seeds. The reference is the planted family-core subset and is not an available natural-model selector.

Method	Worst risk	Families / 50
Family-core reference	0.005713 ± 0.000112	50.00 ± 0.00
Raw hierarchy + 5% floor	0.005713 ± 0.000112	50.00 ± 0.00
Gaussian hierarchy + 5% floor	0.005713 ± 0.000112	50.00 ± 0.00
Raw hierarchy, no floor	0.006322 ± 0.004317	49.98 ± 0.14
Mean capacity	0.035714 ± 0.021415	48.90 ± 0.79
Robust capacity	0.053700 ± 0.026905	48.28 ± 0.97
Magnitude	0.244223 ± 0.038333	41.74 ± 1.35
Random subset	0.442333 ± 0.053582	35.24 ± 1.92
Activation RMS	0.529916 ± 0.046693	31.92 ± 1.88

At known width 50, raw and Gaussian trees with the same floor are identical. The fixed-width result therefore supports write-aware geometry, energy filtering, and global compensation; it does not support a unique Gaussian advantage.

5.3 Stopping-rule audit

The raw no-floor sketch and its singular-value-gap baseline identify 70 nonzero measured directions. Raw hierarchy plus the 5% floor stops near 65. Gaussian shrinkage without the floor stops near 55. Only their combination stops near the planted 50-family core (50.02 ± 0.14 at 48 rows per environment). This interaction is shown in Fig. 4.

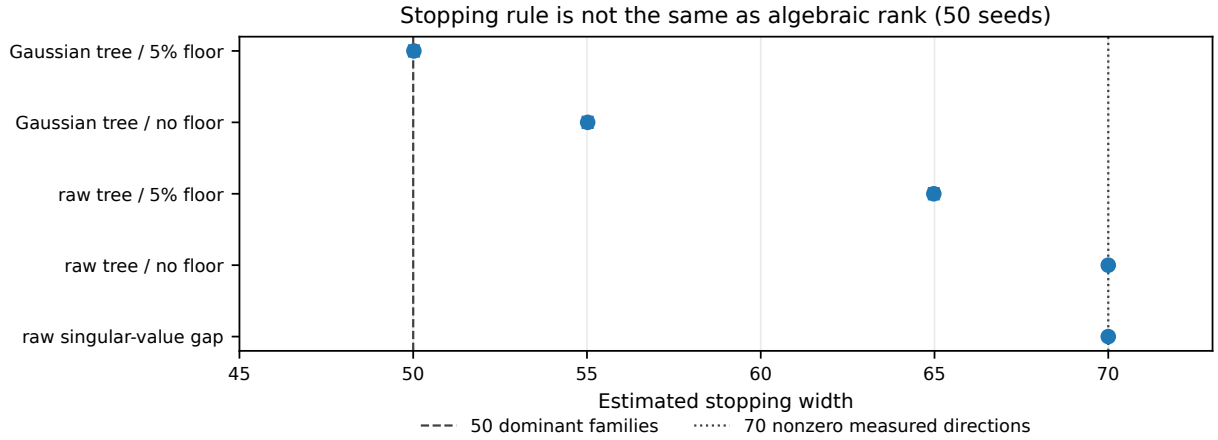

Fig. 4: The controlled stopping width is not algebraic rank. The raw spectrum and raw no-floor hierarchy identify 70 nonzero measured directions, whereas Gaussian shrinkage plus a declared energy floor identifies a lower 50-family operating point.

The result should be interpreted as a rate–distortion policy in a planted construction, not discovery of a universal intrinsic neural dimension. Moderate variance temperatures from 1/72 through 1/18 are stable in this benchmark; $\gamma = 1/4$ degrades family recovery. Varying the bootstrap draw count from 8 to 72 changes numerical precision but no longer changes the shrinkage variance mechanically.

6 Repeated GPT-2 Small Diagnostic

6.1 Protocol

We evaluate GPT-2 Small at resolved revision `607a30d783dfa663caf39e06633721c8d4cfcd7e`. The study uses MLP layers 2, 5, and 8; three disjoint offsets per layer; and prose, dialogue, and code as named environments.

Each layer–split trial uses 256 calibration tokens and 512 held-out tokens per environment. Code calibration and evaluation come from different CPython v3.13.0 modules. The functional sketch uses output rank 24, four sample probes, and 24 bootstrap draws. We physically rebuild the selected MLP at 25%, 50%, and 60% local width reduction and apply the same normalized minimax refit to all subsets.

The comparison includes magnitude, Wanda-style saliency, robust functional capacity, raw functional hierarchy, and fixed Gaussian hierarchy. There are nine layer–split trials for each method and ratio. No threshold or method is chosen on the final evaluation data; however, the experiment remains a modest fixed-corpus diagnostic rather than a competitive benchmark.

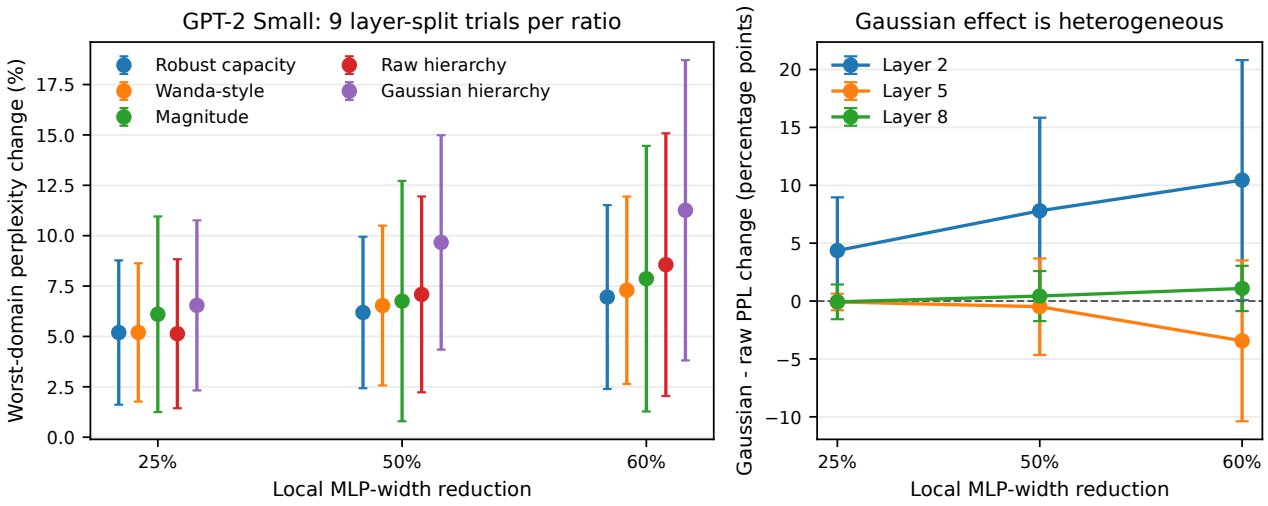

Fig. 5: Repeated GPT-2 Small diagnostic. Left: mean worst-domain perplexity change with sample-standard-deviation error bars over nine layer–split trials. Right: Gaussian-minus-raw differences by layer; positive values mean Gaussian refinement is worse.

Table 2: Worst-domain relative perplexity change (%, mean ± sample SD) over nine GPT-2 layer–split trials per compression ratio. Lower is better.

Method	25%	50%	60%
Robust capacity	5.19 ± 3.58	6.19 ± 3.76	6.95 ± 4.56
Wanda-style	5.20 ± 3.43	6.53 ± 3.97	7.29 ± 4.65
Magnitude	6.10 ± 4.85	6.75 ± 5.97	7.87 ± 6.59
Raw hierarchy	5.13 ± 3.70	7.09 ± 4.86	8.56 ± 6.52
Gaussian hierarchy	6.54 ± 4.22	9.66 ± 5.32	11.26 ± 7.45

6.2 Result and interpretation

Raw hierarchy is competitive at 25% reduction but does not dominate scalar selectors at 50% or 60%. Robust capacity has the lowest aggregate mean at the two more aggressive ratios. Gaussian refinement is worse than raw hierarchy by 1.41, 2.58, and 2.70 percentage points at 25%, 50%, and 60% respectively, but the corresponding paired bootstrap intervals over nine trials each include zero. Averaging the three ratios within each layer–split trial, Gaussian minus raw is 2.23 percentage points with a paired bootstrap 95% interval $[-1.09, 6.16]$. The effect is strongly layer dependent: Gaussian refinement harms layer 2, is sometimes beneficial in layer 5, and is mixed in layer 8.

Against robust capacity, the trial-averaged Gaussian disadvantage is 3.04 percentage points with 95% interval $[0.14, 6.68]$. The data therefore do not support using the Gaussian regularizer as a default real-model geometry. This negative result is central to the final claim boundary: bootstrap variability can be informative in controlled data, but coordinatewise shrinkage can distort angles that matter for natural-model merging.

A 25%, 50%, or 60% local reduction in one MLP removes only approximately 0.95%, 1.90%, or 2.28% of total GPT-2 parameters. We do not infer proportional latency, memory-bandwidth, or energy savings.

7 Discussion

7.1 What is supported

The controlled studies support three specific statements. First, worst-environment selection can preserve a stable functional basis that prevalence-weighted selection deletes. Second, selection and refitting are distinct interventions: a robust refit can reduce damage from a fragile subset but cannot generally recreate a missing activation direction. Third, vector-valued functional grouping with an explicit rate–distortion policy can recover planted dominant families in an overcomplete layer more reliably than activation-only, magnitude, or random selection.

The GPT-2 study supports implementation portability and reveals a useful negative result. Raw functional hierarchy is viable but not uniformly superior. Fixed Gaussian shrinkage is not reliable enough to be the default. A future regularizer should be selected on independent validation data, operate on covariance-aware subspaces rather than independent coordinates, or abstain when calibration uncertainty is too large.

7.2 What is not supported

We do not establish state-of-the-art LLM compression, a universal intrinsic dimension, semantic neuron labels, unknown-shift robustness, multi-block joint compression, or transformer-depth reduction. Width compression removes parallel channels inside a block. Depth compression removes sequential transformations and requires separate machinery such as layer-level fusion or eviction [2, 20].

7.3 Limitations and falsification criteria

DR-FSP protects only named calibration environments. Worst-environment optimization can chase noisy group estimates. The output-latent basis is data dependent, and the coordinate-independence assumption in the Gaussian regularizer is false in general. Average linkage is greedy and $O(m^2)$. A one-representative-per-cluster rule cannot express arbitrary subspace replacements. Layer reconstruction is not downstream loss, and the GPT-2 evaluation is limited to one model, three layers, three corpus offsets, and no recovery fine-tuning.

The practical hypothesis would be weakened if larger repeated evaluations show no consistent advantage for raw functional grouping over simpler selectors at equal recovery compute, if selected subsets are unstable under ordinary calibration resampling, or if downstream task loss is poorly predicted by the normalized layer risk. The Gaussian hypothesis is already weakened by the present GPT-2 results and should be treated as open rather than confirmed.

7.4 Broader impact and ethics

Structured compression can reduce storage, arithmetic, and potentially deployment energy, but it can also erase capabilities that are rare in the calibration distribution. A misleading average metric may therefore create false confidence about safety-critical or minority-domain behavior. Recommended mitigations are domain-stratified evaluation, worst-domain reporting, uncertainty intervals, and an explicit option not to compress when named environments cannot be represented adequately. The experiments use public-domain literary text, publicly licensed CPython source, and pretrained public models; no human-subject data are collected.

8 Reproducibility

The reproducibility materials are provided as Online Resource 1 and at `https://github.com/kenjugmail/DR-FSP`.Online Resource 1 includes the exact controlled seeds, corrected 50-seed and 20-seed outputs, all nine GPT-2 layer–split trials, the resolved model revision, source-corpus manifest, dependency records, plotting scripts, tests, and SHA-256 checksums. Unless otherwise stated, "±" denotes sample standard deviation across independent seeds or layer–split trials, not standard error. The public repository contains the same reproducibility materials.

8.1 Use of AI-assisted tools

Large-language-model tools were used for editorial drafting, code review, and experiment orchestration. The author independently defined the research questions, implemented or audited the methods and numerical outputs, and takes full responsibility for the manuscript, analyses, and artifact. No language model is an author.

9 Conclusion

DR-FSP treats structured width compression as functional-basis selection followed by a globally compensated multi-environment refit. Controlled experiments show that worst-environment selection can prevent a concrete form of false redundancy and that hierarchical vector-valued geometry can recover planted functional families in an overcomplete layer. The final audit also produces a negative result: fixed bootstrap-Gaussian refinement does not improve the repeated GPT-2 study and is worse on average. The scientifically defensible conclusion is therefore narrower than a new state-of-the-art compressor. Functional geometry and robust selection are promising mechanisms; Gaussian refinement remains an optional, unconfirmed regularizer; and broader model, task, and hardware evidence is required.

10 Statements and Declarations

Funding

No funding was received for conducting this study or preparing this manuscript.

Competing interests

Kenju Tomita is the founder and owner of Ephemerent LLC. The work was conducted independently; Ephemerent LLC had no role in study design, data collection, analysis, or the decision to publish. The author declares no other financial or non-financial competing interests relevant to this article.

Ethics approval

Not applicable. The study used synthetic data, public-domain text, publicly licensed source code, and a publicly available pretrained model. No human participants, human data, or animals were involved.

Consent to participate and consent to publish

Not applicable.

Data, materials and code availability

The code, raw numerical outputs, figures, tests, dependency records, model revision, and SHA-256 checksums supporting the findings are provided in Online Resource 1 and at `https://github.com/kenjugmail/DR-FSP`. Pretrained model weights and third-party corpora are not redistributed and remain subject to their respective licenses.

Author contributions

Kenju Tomita: conceptualization, methodology, software, formal analysis, investigation, visualization, writing–original draft, and writing–review and editing. The author approved the final manuscript and accepts responsibility for all aspects of the work.

A Derivation of the normalized refit

Writing $\alpha_e = q_e/(n_e d_e)$, Eq. 8 expands as

$$\mathcal{L}_q(B;S) = \sum_e \alpha_e \operatorname{tr}\big[(Y_e - A_{e,S}B)^\top (Y_e - A_{e,S}B)\big] \tag{17}$$

$$+ \lambda \operatorname{tr}\big[(B - B_0)^\top (B - B_0)\big]\,. \tag{18}$$

Using $\nabla_B \|Y - AB\|_F^2 = 2A^\top(AB - Y)$, setting the gradient to zero, and collecting terms yields Eq. 9. The Hessian is positive definite for $\lambda > 0$.

B Derivation of the expected sketch inner product

Using the vectorization identity,

$$z_{e,i}^\top z_{e,j} = \frac{[(G_e a_{e,i})^\top (G_e a_{e,j})][(P^\top v_i)^\top (P^\top v_j)]}{d_e}. \tag{19}$$

Because $\mathbb{E}[G_e^\top G_e] = I/n_e$, taking expectation over G_e gives Eq. 12.

C Additional controlled compression path

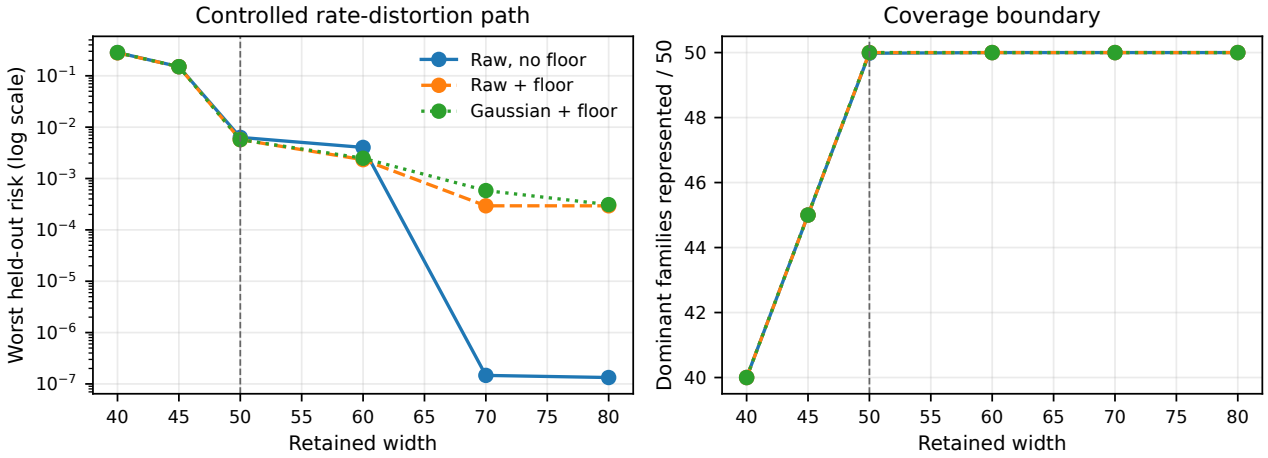

Fig. 6: Controlled rate–distortion path. The raw no-floor hierarchy reconstructs nearly exactly at width 70 because the sketch contains 70 nonzero directions. Error rises sharply below width 50 because distinct dominant families must then be merged.

D Additional statistics

Table 3: Paired GPT-2 comparisons after averaging the three compression ratios within each of the nine layer–split trials. Positive values mean the row method is worse. Intervals are percentile bootstrap 95% intervals over the nine trial units.

Comparison	Mean difference (pp)	95% interval
Gaussian − raw hierarchy	2.23	$[-1.09, 6.16]$
Gaussian − robust capacity	3.04	$[0.14, 6.68]$
Gaussian − Wanda-style	2.81	$[-0.40, 6.67]$
Raw hierarchy − robust capacity	0.82	$[-0.06, 1.82]$
Raw hierarchy − Wanda-style	0.59	$[-0.45, 1.66]$

E Symbol glossary

Table 4: Core notation.

Symbol	Meaning
e, E, n_e	Environment index, number of named environments, and examples in environment e
m, d	Original hidden width and outgoing representation dimension
A_e, V, Y_e	Activation matrix, outgoing matrix, and original layer contribution $Y_e = A_e V$
$a_{e,i}, v_i, F_{e,i}$	Unit activation profile, write vector, and rank-one functional contribution
S, k, B	Retained indices, retained width, and refit outgoing matrix
d_e, R_e	Target energy and normalized environment reconstruction risk
P, r	Output-latent basis and retained output rank
G_e, p, z_i	Gaussian sample projection, probe count, and compact functional fingerprint
$s_{ij}^2, \gamma, \omega_{ij}^2$	Bootstrap sample variance, variance temperature, and effective shrinkage variance
μ_i, ν_i	Gaussian-refined mean and coordinatewise variance
π, q	Deployment prevalence and optimization environment weights

References

[1] Woojeong Kim, Suhyun Kim, Mincheol Park, and Geonseok Jeon. Neuron merging: Compensating for pruned neurons. In *Advances in Neural Information Processing Systems*, 2020.

[2] Hossein Mobahi and Peter L. Bartlett. Hilbert operator for progressive encoding (HOPE): A mathematical framework for deconstructing learned representations in deep networks. *arXiv preprint arXiv:2607.21366*, 2026.

[3] Dong Wang, Haris Šikić, Lothar Thiele, and Olga Saukh. Forget the data and fine-tuning! just fold the network to compress. In *International Conference on Learning Representations*, 2025. arXiv:2502.10216.

[4] Neha Verma, Kenton Murray, and Kevin Duh. DOTResize: Reducing LLM width via discrete optimal transport-based neuron merging. *arXiv preprint arXiv:2507.04517*, 2025.

[5] Amir Asiaee. Efficient discovery of approximate causal abstractions via neural mechanism sparsification. *arXiv preprint arXiv:2602.24266*, 2026.

[6] Wenwu Tang, Dong Wang, Lothar Thiele, and Olga Saukh. GRAIL: Post-hoc compensation by linear reconstruction for compressed networks. In *Conference on Parsimony and Learning*, volume 328 of *Proceedings of Machine Learning Research*, pages 881–895, 2026.

[7] Xiang Meng, Shibal Ibrahim, Kayhan Behdin, Hussein Hazimeh, Natalia Ponomareva, and Rahul Mazumder. OSSCAR: One-shot structured pruning in vision and language models with combinatorial optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35354–35377, 2024.

[8] Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, 2023.

[9] Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models. *arXiv preprint arXiv:2312.11983*, 2023.

[10] Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, 2024.

[11] Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[12] Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024.

[13] Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10100–10111, 2024. doi: 10.18653/v1/2024.acl-long.544.

[14] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[15] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196.

[16] Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

[17] David L. Donoho, Matan Gavish, and Elad Romanov. ScreeNOT: Exact MSE-optimal singular value thresholding in correlated noise. *The Annals of Statistics*, 51(1):122–148, 2023. doi: 10.1214/22-AOS2232.

[18] Donald B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

[19] Dimitris Achlioptas. Database-friendly random projections: Johnson–lindenstrauss with binary coins. In *Journal of Computer and System Sciences*, volume 66, pages 671–687, 2003.

[20] Jinuk Kim, Marwa El Halabi, Mingi Ji, and Hyun Oh Song. LayerMerge: Neural network depth compression through layer pruning and merging. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23825–23842, 2024.