

Reading the Geometry: A Mechanism for When Frozen LLM Text Helps Tabular Foundation Models

Zou Qingyun¹ Yingze Li² Nuo Chen¹

¹National University of Singapore, Singapore

²Harbin Institute of Technology, Harbin, China

{qingyunzou, nuochen}@u.nus.edu

23b903046@stu.hit.edu.cn

Abstract

Tabular foundation models like TABPFN and TABICL are often augmented with frozen large language model (LLM) embeddings when tables contain free-form text. Whether this helps is highly case-dependent: across 732 (dataset, backbone, encoder) configurations, 45.9% improve, 30.9% stay neutral, and 23.2% degrade. Using a fixed, outcome-free raw score, we define high- and low-readability strata before applying target-covariance Gaussian sampling, row-text shuffling, and pre-projection whitening. Each intervention reduces the mean gain in the high-readability stratum to near zero, whereas the paired intervention effects in the low-readability stratum remain within sampling noise. The first two interventions show that neither a Gaussian sample from the target covariance nor row-misaligned text reproduces the high-stratum gain; whitening further shows sensitivity to broad pre-projection geometry. Together, the results support a dependence on row-aligned embedding structure without isolating the covariance spectrum as its unique cause. The high-stratum Gaussian-replacement effect replicates on a second backbone. A within-dataset encoder ranking, Decision-Boundary Alignment (DBA), also shows stronger rank alignment with gain than the Massive Text Embedding Benchmark (MTEB) rank in our eight-dataset diagnostic.

1 Introduction

Modern tabular foundation models (TabFMs) that perform in-context learning (ICL), such as TABPFN (Hollmann et al., 2023, 2025) and TABICL (Qu et al., 2025), predict on a new tabular dataset directly at inference time, with no per-dataset training. Many real-world tables also contain free-form text such as product descriptions or clinical notes, which these models cannot consume directly. Recent work has explored representing textual table content with frozen large lan-

guage model (LLM)-based embeddings and combining those representations with structured features (Koloski et al., 2025; Kasneci and Kasneci, 2024; Mráz et al., 2025).

Figure 1a reports the distribution of augmentation gains by dataset family, backbone, and encoder. All three outcomes are common: 45.9% are positive, 30.9% neutral, and 23.2% negative. The global average gain reported in prior work (Koloski et al., 2025; Kasneci and Kasneci, 2024) hides this structure. Two hypotheses motivate our interventions. Under a geometry-only account, the unusual shape of the encoder columns acts as an implicit regularizer. This account is motivated by qualitative probes of TABPFN’s inductive biases (McCarter, 2025) and by its reported robustness to random, label-uncorrelated columns in controlled synthetic binary tasks (Hu and Ghelichi, 2026). Under a content account, gains come from label-relevant information in the text, and encoders with higher Massive Text Embedding Benchmark (MTEB) rank (Muennighoff et al., 2023) should help most.

The interventions are informative only where text injection contributes a measurable gain. Pooling helpful tuples with neutral or harmful ones would dilute the effect of destroying the text. We therefore use S as an *ex-ante ranking instrument*, not as a causal quantity or a deployment gate: it selects the configurations on which an intervention can reveal what produced the gain.

Figure 1b shows this connection. MISMATCH measures departures of the projected embeddings from fixed analytic references, while SUSCEPTIBILITY measures the response of the specific backbone to those embedding directions. For the main intervention experiment in Table 1, we set $\gamma=1$ and rank tuples by the outcome-free raw score S_{raw} ; it does not use Δ or the true row-label assignments. Its top and bottom 30% define the high- and low-readability strata before any intervention outcome

Why and how frozen text helps tabular foundation models

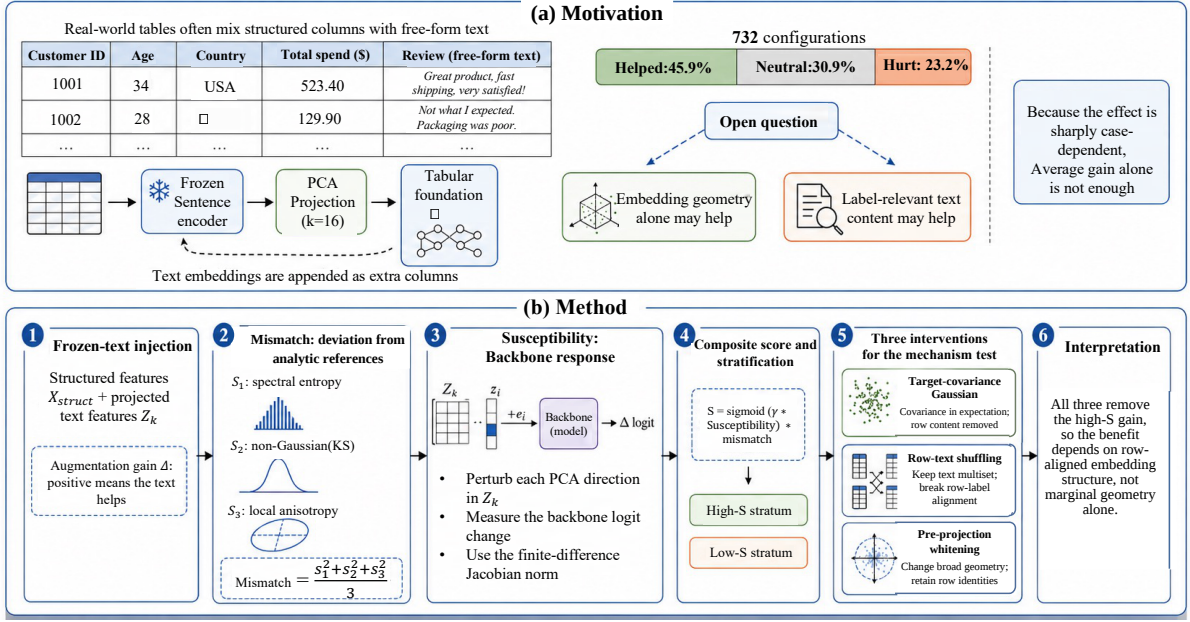


Figure 1: Overview. **(a)** Frozen-text injection is case-dependent: 45.9% of configurations improve, 30.9% remain neutral, and 23.2% degrade, so the global average hides this variation. **(b)** Our mechanism-testing pipeline. MISMATCH and SUSCEPTIBILITY form a score that ranks configurations before intervention. Table 1 uses the fixed, outcome-free raw score ($\gamma=1$) only to define high- and low-readability strata; the supervised calibration of (a, b, γ) is evaluated separately in Table 2. The interventions then test which properties of the text block account for the gain. PCA denotes principal component analysis, and KS denotes Kolmogorov–Smirnov.

is observed. The supervised parameters (a, b, γ) are fitted only for the separate leave-one-family-out (LFO) sign-stratification analysis in Table 2. With these strata fixed, we apply the interventions in §4 and compare the paired changes in Δ .

We test frozen-text gains with controlled perturbations that remove complementary properties of the embedding block.¹ The results support a dependence on row-aligned embedding structure whose usefulness covaries with encoder geometry. A fixed, outcome-free MISMATCH $\times$ SUSCEPTIBILITY score defines the intervention strata without consulting Δ , while a separately calibrated variant is evaluated as an LFO sign stratifier. We reproduce the high-stratum Gaussian-replacement effect on a second backbone and compare MTEB with Decision-Boundary Alignment (DBA) as predictors of downstream gain.

2 Background and Setup

Frozen-text injection into ICL tabular foundation models. We study the plug-and-play multimodal setup in which an unchanged ICL tab-

ular foundation model takes an augmented context $[X_{struct}, Z_k]$. Here $X_{struct} \in \mathbb{R}^{n \times p}$ holds the structured columns, and $Z_k \in \mathbb{R}^{n \times k}$ is a low-dimensional projection of frozen sentence encodings $E(T) \in \mathbb{R}^{n \times d_E}$. We adopt a frozen-text injection pipeline related to prior tabular-text studies (Koloski et al., 2025; Kasneci and Kasneci, 2024), with the following specific implementation: concatenate all text cells of a row with a separator, encode with a frozen sentence encoder E , and project to $k=16$ dimensions with a dataset-specific principal component analysis (PCA) basis. The PCA basis is fit once on all rows of a dataset before cross-validation. This label-free but transductive preprocessing matches the evaluated inference setting; it is not a strictly inductive, fold-specific PCA protocol. We measure the effect of frozen-text injection by the augmentation gain

$$\Delta(D, b, E) = \text{acc}_{\text{on}}(D, b, E, k) - \text{acc}_{\text{off}}(D, b), \quad (1)$$

measured under paired 5-fold cross-validation (CV) within dataset D (Algorithm 1, lines 5–7). A positive Δ means injection helps; a negative Δ means it hurts.

¹Code and results: https://github.com/zanehp/Tabular_Foundation_Models.

Backbones and encoders. We use two ICL tabular foundation models as backbones b : TABPFN v2.6 (the default checkpoint resolved by the 7.1.1 package, following the TabPFN-2.5 model family (Grinsztajn et al., 2025)) and TABICL v2 (Qu et al., 2025); both are held strictly frozen. We use four sentence encoders E spanning the MTEB leaderboard (Muennighoff et al., 2023): BGE-LARGE-EN (Xiao et al., 2024), E5-LARGE-V2 (Wang et al., 2022), GTE-LARGE (Li et al., 2023), and SENTENCE-T5-BASE (Ni et al., 2022). These four define the 732-tuple main matrix. The PCA-controlled text-only diagnostic in §5.4 adds three encoders but does not change the main matrix.

Dataset families. We curate 58 tabular-with-text datasets organized into six families (the exact source inventory is in Appendix M): fam1 *tabular-text* (OpenML tasks with synthesized text columns at four informativeness levels: high, zero, adversarial, redundant); fam2 *text-classification* (HuggingFace English text classification); fam3 *multi-aspect* (multi-text product reviews); fam4 *crosslingual* (Chinese, German, French, Japanese, and Spanish Amazon reviews and Cross-lingual Natural Language Inference (XNLI) subsets); fam5 *specialty* (medical, legal, financial domains); fam6 *synthetic* (structural causal model (SCM)-style tabular data with verbalized structure). Evaluating these dataset variants with the available backbone–encoder combinations gives the 732 (dataset, backbone, encoder) tuples used throughout.

Sign labels. We label each tuple by $\text{sign}(\Delta)$ with a gray zone of $\varepsilon=0.5\%$ accuracy: $+1$ if $\Delta > +\varepsilon$, -1 if $\Delta < -\varepsilon$, and 0 otherwise. The 0 class is excluded from the primary area under the receiver operating characteristic curve (AUROC) computation and reported as retained mass.

3 The MISMATCH $\times$ SUSCEPTIBILITY Framework

Our framework decomposes the question “can the backbone read this encoder’s geometry on this dataset?” into two factors. MISMATCH is computed from the projected text columns alone and compares their spectrum, marginal distribution, and local geometry with fixed analytic references. SUSCEPTIBILITY measures how strongly the backbone’s predictions respond to perturbations along the encoder’s directions. Neither factor uses Δ or the true row–label correspondence; SUSCEPTIBILITY

retains the task class count but replaces support-label assignments with fixed random pseudo-labels. Their raw product ranks tuples for the intervention experiment, while a separately calibrated version is evaluated as a sign stratifier.

3.1 MISMATCH: geometric deviation from analytic references

Let $\mathbf{Z}_k$ be the k -dim projection of row-wise normalized text embeddings. We compute three closed-form statistics. Their raw forms define the intervention strata, and their training-family z -scores are used only in the LFO analysis.

Spectral entropy S_1 . The first substatistic measures spectral concentration. Let $\sigma_1, \dots, \sigma_k$ be the singular values of $\mathbf{Z}_k$. We normalize their squared values and compute spectral entropy:

$$p_i = \frac{\sigma_i^2}{\sum_{\ell=1}^k \sigma_\ell^2}, \quad i = 1, \dots, k, \quad (2)$$

$$S_1 = - \sum_{i=1}^k p_i \log p_i.$$

which is maximized at $\log k$ when the spectrum is isotropic and decreases as the embedding concentrates onto fewer directions. Prior analyses have found anisotropy and concentration along dominant directions in contextual token and word embeddings (Ethayarajh, 2019; Gao et al., 2019), motivating the possibility that S_1 falls below $\log k$. The signed offset $S_1 - \log k$ below records how far the spectrum lies below this isotropic reference.

Flattened marginal Kolmogorov–Smirnov (KS) statistic S_2 . Spectral shape alone misses differences in the per-coordinate distribution, so the second substatistic targets non-Gaussianity directly. After standardizing each coordinate, we flatten $\mathbf{Z}_k$ over rows and dimensions and compute a one-dimensional Kolmogorov–Smirnov statistic against $\mathcal{N}(0, 1)$:

$$S_2 = \sup_z |F_{\text{emp}}(z) - \Phi(z)|. \quad (3)$$

This statistic summarizes marginal non-Gaussianity, including heavy tails and bimodality. A sliced-Wasserstein variant (Appendix B) gives the same rankings.

Local anisotropy S_3 . Global spectral and marginal measures can still miss directional structure that emerges only at the neighborhood scale, so we add a local probe. For each row i , let $\mathbf{C}_i$ be the empirical covariance over the $K=10$ nearest

neighbors of row i in $\mathbf{Z}_k$. We report the median regularized condition number,

$$S_3 = \text{median}_i \frac{\lambda_{\max}(\mathbf{C}_i + \lambda I)}{\lambda_{\min}(\mathbf{C}_i + \lambda I)}, \quad (4)$$

with regularizer $\lambda = 10^{-3} \text{tr}(\mathbf{C}_{\text{global}})$. A large S_3 indicates thin local neighborhoods, whereas the analytic isotropic reference has condition number one.

The outcome-free raw coordinate used to select intervention strata is

$$M_{\text{raw}} = (S_1 - \log k) + S_2 + \log(1 + S_3). \quad (5)$$

For LFO sign stratification, we standardize each statistic using training-family moments and define

$$M_z = \frac{1}{3}(S_1^z + S_2^z + S_3^z). \quad (6)$$

We refer to M_{raw} or M_z as **MISMATCH** when the raw or LFO context is clear.

Connection to classical anisotropy. Classical natural language processing (NLP) work studies embedding anisotropy through several complementary summaries. On the 336 tuples with matching raw-embedding caches, raw-space IsoScore correlates with **MISMATCH** ($\rho = +0.483$) and continuous Δ ($\rho = +0.630$). Yet it transfers poorly as a family-held-out sign stratifier: raw-space IsoScore reaches 0.467 AUROC, compared with 0.574 for **MISMATCH**; PCA-space IsoScore and participation ratio each reach 0.523. Thus **MISMATCH** overlaps with classical anisotropy summaries but is not reducible to one of them. See Appendix H.

3.2 SUSCEPTIBILITY: backbone response

MISMATCH sees only the encoder. To measure whether the backbone b responds to the projected text, we estimate a finite-difference *row-uniform* response in one mini-batch. We draw up to 200 support rows and 32 query rows from the training pool; for smaller datasets, the support size is reduced to the available rows after reserving the queries. For each PCA dimension $j \in \{1, \dots, k\}$, we perturb the support text block by $\pm \eta \mathbf{1}_n \mathbf{e}_j^\top$, where $\mathbf{1}_n$ is the all-ones vector and $\eta = 0.05 \|\mathbf{Z}_k\|_F / k$. Structured columns and query features remain fixed. We then read the pre-softmax query logits:

$$\mathbf{g}_j = \frac{1}{2\eta} [b(\mathbf{Z}_k + \eta \mathbf{1}_n \mathbf{e}_j^\top) - b(\mathbf{Z}_k - \eta \mathbf{1}_n \mathbf{e}_j^\top)]. \quad (7)$$

The perturbation is applied identically to every row, so **SUSCEPTIBILITY** measures the backbone’s average response to a global shift along PCA direction j , not per-row local sensitivity. Stacking the

$\mathbf{g}_j$ gives a row-uniform Jacobian $\mathbf{J} \in \mathbb{R}^{k \times 32 \times |\mathcal{Y}|}$, which we summarize by its mean per-query Frobenius norm,

$$\bar{J} = \frac{1}{32} \sum_{q=1}^{32} \|\mathbf{J}[:, q, :]\|_F, \quad (8)$$

and use $\bar{J}$ as raw **SUSCEPTIBILITY**. The LFO analysis standardizes it using training-family moments.

3.3 Composite S

S is the ranking device that links the framework to Table 1. For that experiment, we combine Eqs. 5 and 8 with a fixed gate:

$$S_{\text{raw}} = \sigma(\bar{J}) M_{\text{raw}}. \quad (9)$$

Here σ is the logistic function and the gating coefficient γ is fixed to one. We rank all tuples by S_{raw} and freeze the top and bottom 30% as the high- and low-readability strata before running the interventions. Thus Table 1 does not select on Δ .

The LFO analysis in Table 2 is separate. Within each fold, we standardize the factors using training-family moments and jointly fit (a, b, γ) on the signed training-family tuples. The raw gated score and its probability-style calibration are

$$S_\gamma = M_z \sigma(\gamma \bar{J}^z), \quad \hat{p}_\gamma = \sigma(b + a S_\gamma). \quad (10)$$

The three parameters minimize mean binary cross-entropy plus the implemented $10^{-3} \gamma^2$ penalty. The held-out family is transformed with the frozen moments and parameters. Thus the per-tuple factors are outcome-free, but the LFO calibration is supervised. Table 2 reports both $\hat{p}_\gamma$ and the uncalibrated S_γ . Because the fitted slope a can be negative, the outer calibration can reverse the within-fold ordering, so their AUROCs need not coincide. The calibrated linear ablation is $\hat{p}_{\text{lin}} = \sigma(b_{\text{lin}} + w_M M_z + w_J \bar{J}^z + w_{MJ} M_z \bar{J}^z)$.

4 Experimental Protocol

Algorithm 1 states the end-to-end workflow. True outcome assignments enter the per-fold paired accuracy that defines Δ and the cross-tuple logistic fit of (a, b, γ) on calibration families (boxed). For **SUSCEPTIBILITY**, backbone calls retain only the task class count and replace the support assignments with fixed random pseudo-labels. The intervention strata in §5.3 use S_{raw} with a fixed gate and do not consult $\text{sign}(\Delta)$.

Algorithm 1 End-to-end TabFusion protocol. Boxes mark uses of true outcome assignments; all other score construction is outcome-free.

```

1: Per-dataset  $D$  (outcome-free score construction):
2:    $\tilde{\mathbf{Z}} \leftarrow \text{row-}\ell_2\text{-normalize } E(\mathbf{T})$ 
3:    $V_k \leftarrow \text{PCA}(\tilde{\mathbf{Z}}, k=16)$ ;  $\mathbf{Z}_k \leftarrow \tilde{\mathbf{Z}}V_k$ 
4:    $(S_1, S_2, S_3, \bar{J}) \leftarrow \text{stats on } \mathbf{Z}_k$  (random pseudo-labels for  $\bar{J}$ )
5: Per-tuple  $\Delta(D, b, E)$  (labels per fold):
6:   for each of 5 paired CV folds:
7:      $\Delta_{\text{fold}} \leftarrow \boxed{\text{acc}_{\text{on}}} - \boxed{\text{acc}_{\text{off}}}$ 
8:    $\Delta \leftarrow \text{mean}(\Delta_{\text{fold}})$ 
9: Cross-tuple LFO (Table 2):
10:  for each held-out family  $f$ :
11:    train  $\leftarrow$  tuples with family  $\neq f$ 
12:     $(\mu_{\text{tr}}, s_{\text{tr}}) \leftarrow$  z-moments of  $(S_1, S_2, S_3, \bar{J})$ 
13:     $S_\gamma \leftarrow M_z \sigma(\gamma \bar{J}^z)$ 
14:    fit  $(a, b, \gamma)$  jointly by regularized BCE of  $\boxed{\text{sign}(\Delta)}$ 
15:    under  $\sigma(b + a S_\gamma)$ 
16:    freeze  $(\mu_{\text{tr}}, s_{\text{tr}}, a, b, \gamma)$ , apply  $\hat{p}_\gamma$  to  $f$ 
17: Intervention strata (Table 1):
18:    $S_{\text{raw}} \leftarrow \sigma(\bar{J})M_{\text{raw}}$  (fixed gate)
19:   high/low  $\leftarrow$  top/bottom 30% by  $S_{\text{raw}}$ 

```

Configurations, splits, and scoring. We evaluate the 732 (dataset, backbone, encoder) tuples introduced in §2; Algorithm 1 gives the per-dataset Δ procedure and cross-tuple LFO split. Tuples with $|\Delta| \leq \varepsilon$ form a separately reported 0 class. Under LFO, we compare the calibrated multiplicative predictor $\hat{p}_\gamma$ with the calibrated linear predictor $\hat{p}_{\text{lin}}$, calibrated one-factor predictors based on MISMATCH or SUSCEPTIBILITY, and the raw gated score S_γ without its outer logistic calibration. Following Efron and Tibshirani (1986), we use dataset-grouped bootstrap confidence intervals (CIs; $B=2000$) and weighted-mean AUROC over the four multi-class held-out families; FAM2 and FAM5 are excluded because their test folds are single-class.

Interventions. We probe the mechanism behind S with complementary interventions, each paired against the original Δ within tuple to control for dataset difficulty and backbone choice: **(I1) Target-covariance Gaussian:** draw a zero-mean Gaussian block with target covariance $\hat{\Sigma}$ estimated from $\mathbf{Z}_k$. The finite sample matches this covariance in expectation, not exactly, and removes row-specific content. **(I2) Shuffle text within column:** permute rows of $\mathbf{T}$ before re-encoding, preserving the marginal text distribution and global spectrum while destroying row-label correspondence. **(I3) Pre-PCA whitening:** apply zero-phase component analysis (ZCA) whitening ($\mathbf{X} \mapsto \mathbf{X} \hat{\Sigma}^{-1/2}$) to the raw d_E -

dimensional encoder output before PCA projection. This flattens the covariance spectrum but generally also changes angles, cosine similarities, and neighborhood geometry. I3 therefore tests sensitivity to a broad pre-projection geometric change; it does not isolate the spectrum while holding all directional content fixed. We apply each intervention to the high- and low- S strata and report paired Wilcoxon tests with dataset-grouped bootstrap CIs.

5 Results

We organize the findings around five research questions (RQs).

5.1 RQ1: bidirectionality varies by family

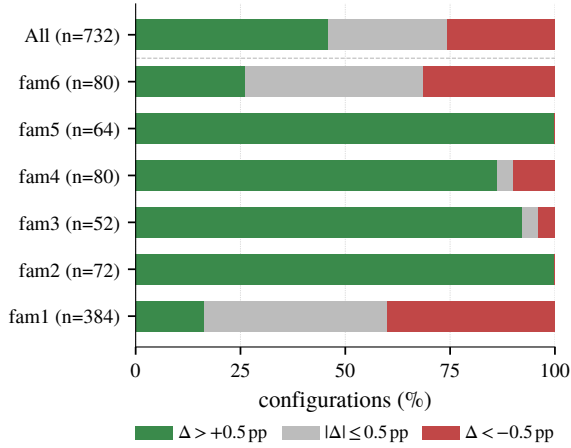
Frozen-text injection is bidirectional in aggregate and sharply family-conditional (Figure 2a). Pure text-classification (fam2) and specialty (fam5) families are all-positive in our matrix. Their structured columns carry no signal by construction, so their text-off baselines remain near chance. The remaining families contain both positive- and negative- Δ tuples, with fam1 and fam6 showing the broadest sign mix. Some (family, encoder, backbone) cells cross into the opposite sign. The global-mix bar hides this family-level heterogeneity. We therefore analyze the mechanism within strata; a single deployment threshold would suppress the differences among families.

5.2 RQ2: S as a mechanism probe

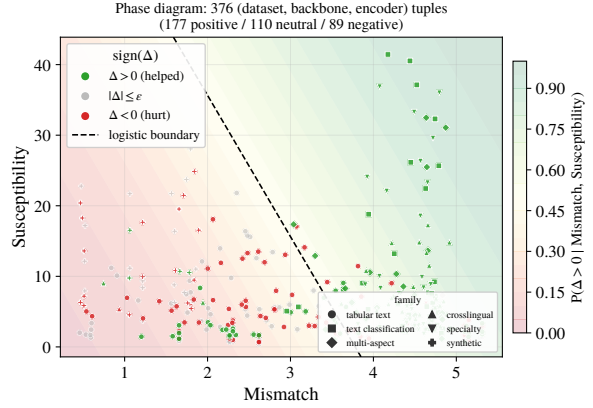
LFO sign prediction. Table 2 quantifies the predictive limit of S : the best LFO AUROC is 0.626, below our fixed 0.68 deployment threshold. Under the leave-one-family-out grouped bootstrap, the calibrated geometry-based variants occupy a narrow AUROC band, and none differs statistically from the calibrated linear composite. Performance varies sharply by family: AUROC is high on the imbalanced fam3 fold and near chance on some structurally two-sided families, which rules out a reliable shared cutoff.

Subsequent analyses use S only to select high- and low-readability strata for intervention (§5.3). We evaluate DBA separately for deployment-oriented model selection in §5.5.

MISMATCH dominates the phase plane. Figure 2b plots every TabICL tuple in the raw ($M_{\text{raw}}, \bar{J}$) plane used for intervention selection, color-coded by $\text{sign}(\Delta)$. Most of the visible enrichment lies along the M_{raw} axis. The high- M_{raw}



(a) Per-family decomposition of Δ across the 732 configurations. Top bar (above the dashed line) is the global mix.



(b) (MISMATCH, SUSCEPTIBILITY) plane for the 376 TabICL tuples. Green: $\Delta > +0.005$; red: $\Delta < -0.005$; gray: $|\Delta| \leq 0.005$. Dashed: logistic boundary fit on signed tuples only. Marker shape encodes dataset family.

Figure 2: The phenomenon (left, a) and the score that sorts it (right, b). Panel a abbreviates percentage points as pp.

region contains more $+\Delta$ tuples (green), while the low- M_{raw} region contains more $-\Delta$ (red) and zero (gray) tuples. The fitted logistic boundary (dashed) is close to vertical: nearly all of its discriminative weight falls on MISMATCH, and SUSCEPTIBILITY adds only a small lift along the secondary axis. This enrichment is sufficient for contrasting the high- and low-readability strata under intervention.

Why SUSCEPTIBILITY remains in the framework. Table 2 shows that SUSCEPTIBILITY alone scores near chance (AUROC 0.553) and that the calibrated multiplicative predictor does not outperform calibrated MISMATCH alone. In our four-encoder panel, MISMATCH carries most of the predictive signal because every tested encoder imposes substantial anisotropy; SUSCEPTIBILITY therefore seldom identifies a “backbone-blind” tuple.

We retain SUSCEPTIBILITY because the framework must represent both encoder geometry and backbone response. Recomputing SUSCEPTIBILITY for each backbone lets the score reflect cross-backbone changes, whereas MISMATCH alone summarizes dataset–encoder geometry and is agnostic to the backbone. The inclusive analysis provides separate empirical support: when the gray zone is retained, SUSCEPTIBILITY has the highest AUROC for the $-\Delta$ class (0.623) and the highest per-fold correlation with continuous Δ ($\rho = +0.240$); see Appendix L. This complementary evidence motivates retaining two factors, although SUSCEPTIBILITY adds little to the sign-only AUROC in Table 2.

5.3 RQ3: interventions probe content and geometry dependence

We partition the TabICL intervention tuples into a high- S stratum (dominated by $+\Delta$) and a low- S stratum (dominated by ≈ 0 or $-\Delta$); Table 1 lists the per-stratum sizes. For each stratum, we compare the original Δ_{real} against the I1–I3 outcomes Δ_{int} , paired within tuple.

The high- S gain disappears under intervention. On the high- S stratum, the paired mean differences are -0.179 , -0.181 , and -0.181 for I1, I2, and I3, respectively (Table 1). Each estimate is significant under paired Wilcoxon, and 70.7–73.2% of the originally positive tuples exit the $+\Delta$ zone. Figure 3a shows the corresponding tuple-level shifts below the $y=x$ line.

The interventions alter complementary properties of the embedding block. Target-covariance Gaussian replacement (I1) tests whether samples from the estimated second-order distribution reproduce the gain, while shuffling (I2) preserves the embedding multiset but breaks row–label correspondence. Their failure to preserve the high-stratum gain supports a dependence on row-specific embedding content beyond marginal covariance. Pre-PCA whitening (I3) changes both the covariance spectrum and angular or neighborhood geometry; its effect shows sensitivity to pre-projection geometry but does not isolate any one geometric component. The interventions do not establish linguistic semantics as the unique cause, as discussed under Limitations.

Intervention	Stratum	n	Mean difference	95% confidence interval (grouped bootstrap)	Wilcoxon p	Flip rate
Target-cov. Gaussian (I1)	high- S	112	-0.179	$[-0.249, -0.116]$	7.3×10^{-16}	59/82 (72.0%)
	low- S	112	+0.001	$[-0.002, +0.004]$	0.38	n/a
Shuffle text (I2)	high- S	112	-0.181	$[-0.255, -0.120]$	4.6×10^{-16}	60/82 (73.2%)
	low- S	112	+0.001	$[-0.001, +0.003]$	0.49	n/a
Pre-PCA whitening (I3)	high- S	112	-0.181	$[-0.249, -0.118]$	8.8×10^{-16}	58/82 (70.7%)
	low- S	112	+0.000	$[-0.001, +0.002]$	0.98	n/a

Table 1: Intervention effects on TABICL tuples. Bootstrap CIs are dataset-grouped ($B=2000$). Flip rate is the fraction of originally $+\Delta$ tuples ($\Delta_{\text{real}} > 0.005$ accuracy) that exit the $+\Delta$ zone after intervention.

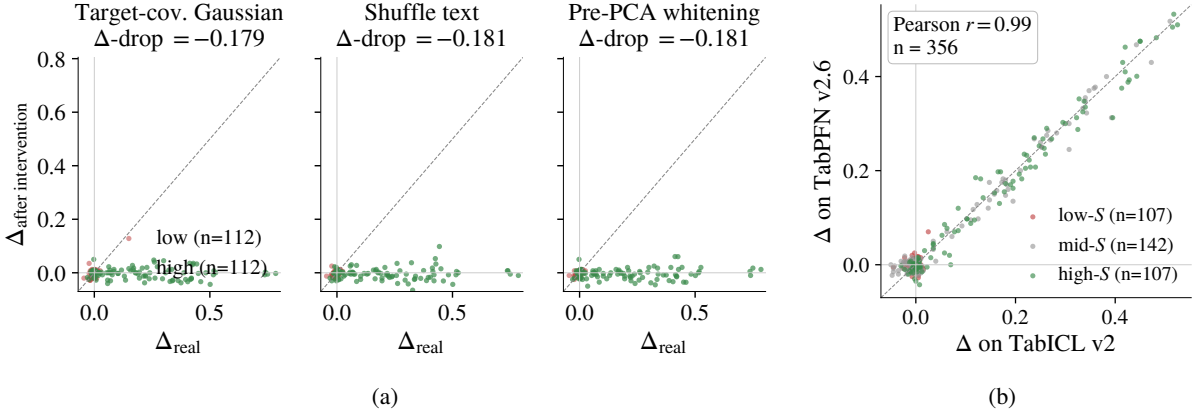


Figure 3: **(a)** Paired Δ before and after I1–I3. Green: high- S ; red: low- S ; dashed: $y=x$. Mean shifts and confidence intervals are reported in Table 1. **(b)** Per-tuple Δ on TabICL v2 versus TabPFN v2.6 over 356 matched (encoder, dataset, info-level) configurations. Dashed: $y=x$. Coloring by S -stratum.

Score	AUROC	95% CI	p_{linear}
$\hat{p}_{\text{lin}}$ (cal.)	0.626	$[\cdot533, \cdot682]$	n/a
$\hat{p}_{\gamma}$ (cal.)	0.602	$[\cdot519, \cdot650]$	0.42
MISMATCH (cal.)	0.594	$[\cdot512, \cdot645]$	0.28
SUSCEPTIBILITY (cal.)	0.553	$[\cdot402, \cdot679]$	0.33
S_{γ} (raw)	0.437	$[\cdot384, \cdot526]$	0.01

Table 2: LFO grouped-bootstrap AUROC ($B=2000$) for $\text{sign}(\Delta)$ prediction across four multi-class held-out families ($n=389$ signed tuples). The first four rows are probability-style predictions calibrated on each training fold; the last row retains the jointly fitted gate but omits the outer slope and intercept. p_{linear} is the paired bootstrap two-sided p -value against $\hat{p}_{\text{lin}}$. A strict 3-fold variant excluding the imbalanced FAM3 fold is in Appendix B.

Low- S intervention effects remain near zero.

On the low- S stratum, the interventions induce no detectable mean change: paired mean differences range from 0.000 to +0.001, with Wilcoxon p -values from 0.38 to 0.98 (Table 1). The low- S points in Figure 3a remain close to the $y=x$ line. These tuples have no measurable baseline gain, leaving little effect for an intervention to remove.

Replication on TabPFN v2.6. The high-stratum Gaussian-replacement effect is not a TabICL-specific artifact. On a stratified TabPFN v2.6 subsample (Appendix D), the high- S mean gain again falls to near zero: the paired difference is -0.098 ($p=0.0044$), about 55% of the TabICL effect. The low- S stratum starts near zero and shifts by a statistically detectable but practically small $+0.004$ ($p=0.0013$). Thus the high-stratum destructive effect transfers across architectures, whereas the exact low-stratum null does not.

What the geometry looks like. A two-dimensional t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) case study on AG News (Appendix I, Figure 4) makes the high/low- S contrast visible: I1 collapses the four class clusters of the high- S variant into a single blob after sampling from the covariance target estimated on the original block, while the low- S variant has no class structure to lose. The visualization is illustrative; Table 1 provides the quantitative evidence.

5.4 RQ4: PCA-controlled text-only quality is more informative than MTEB

In the four-encoder main panel, MTEB rank does not separate mean downstream gain. Table 3 aggregates Δ and the two geometry indicators per encoder, and its bottom row gives each one’s Spearman correlation with MTEB rank: the correlation with mean Δ is zero in this small panel, while the correlations with MISMATCH and SUSCEPTIBILITY are positive. Because rank 1 is the strongest MTEB encoder, the positive signs mean that lower-ranked MTEB encoders have larger mean MISMATCH and SUSCEPTIBILITY in this four-encoder panel. Those aggregate differences do not translate into larger per-encoder mean gains. This is consistent with the intervention evidence (§5.3) that Δ depends on row-aligned embedding structure within a dataset, not on an encoder-level geometry average. With $n=4$ encoders, the comparison is directional and cannot establish general independence.

PCA-16 text-only ranking. The encoders have native dimensions from 384 to 4096, so we also compare them after the same PCA reduction used by the tabular pipeline. On eight English text-classification datasets, we fit a 5-fold logistic-regression classifier using only each encoder’s PCA-16 text features and no structured columns. Accuracy ranges from .763 for ALL-MINILM-L6-V2 (Sentence Transformers, 2021a) to .873 for E5-LARGE-V2; Appendix G gives the full ranking. This PCA-controlled text-only ranking correlates with mean downstream Δ at $\rho = +0.93$ ($p=0.003$), compared with $\rho = +0.75$ ($p=0.052$) for MTEB score. The diagnostic uses labels from the same tasks, so it is a supervised, task-aligned comparison and does not provide independent mechanistic evidence.

5.5 RQ5: cross-encoder portability and transfer limits

The calibrated multiplicative model has limited encoder-specific dependence in our panel. Fitting (a, b, γ) on one encoder and evaluating the resulting predictor on another reduces mean AUROC by only 0.002 (on-diagonal 0.653 versus off-diagonal 0.650), within our fixed 0.05 tolerance; several off-diagonal pairs perform better. This portability suggests limited encoder-specific dependence in the fitted model within our panel.

Strata defined from TabICL also separate

Encoder	MTEB	Δ	Mis.	Susc.	S
BGE	1	+0.078	3.33	6.07	3.15
GTE	2	+0.075	3.21	8.83	3.13
E5	3	+0.085	3.48	8.99	3.39
ST5	4	+0.077	3.42	8.40	3.33
ρ versus MTEB rank		+0.000	+0.60	+0.40	+0.60

Table 3: Per-encoder aggregates over 183 (dataset, backbone) configurations. Encoder shorthands: BGE = BGE-LARGE-EN, GTE = GTE-LARGE, E5 = E5-LARGE-V2, ST5 = SENTENCE-T5-BASE. Metric columns: **Mis.** = MISMATCH, **Susc.** = SUSCEPTIBILITY, S the composite score.

TabPFN gains similarly. Per-tuple Δ on TabICL v2 versus TabPFN v2.6 tracks $y=x$ on the high- S cluster, and per-stratum means are closely aligned between the two backbones (Figure 3b; Appendix E for matched-tuple detail).

On the low- S stratum, pre-PCA whitening moves 14/39 (35.9%) negative- Δ tuples into the neutral or positive zone, but its mean paired effect is only +0.0004 ($p=0.98$). We therefore treat this crossing count as descriptive rather than evidence of systematic rescue (Appendix J). On the high- S stratum, the same broad geometric change removes the gain (cf. I3 in Table 1).

The calibrated multiplicative model trained on the English and synthetic families produces held-out crosslingual AUROC 0.630 (95% CI [0.538, 0.840]), below our fixed 0.65 transfer threshold (Appendix C). We do not claim that this gap is closed without a separately reported in-language calibration experiment.

When target training labels are available, *Decision-Boundary Alignment* (DBA) provides a supervised selection heuristic. DBA is the fraction of rows whose leave-one-out nearest neighbor in cosine space on $E(\mathbf{T})$ shares the same label. Across our four encoders on an English text-classification panel (Appendix F), DBA has positive rank correlation with observed gain on all eight datasets and a higher mean correlation than the external MTEB rank. Because each dataset has only four encoders, this evidence is preliminary.

6 Related Work

TABPFN initiated the in-context tabular foundation model line (Hollmann et al., 2023, 2025); later systems include TABICL and TABDPT (Qu et al., 2025; Ma et al., 2025). Retrained multimodal extensions include TABSTAR (Arazi et al., 2025).

Our setting instead keeps the backbone unchanged and measures its response to injected text geometry.

Prior studies of frozen-text injection report average gains (Koloski et al., 2025; Kasneci and Kasneci, 2024). Mráz et al. (2025) reported both gains and losses across datasets and model settings without characterizing their mechanism. McCarter (2025) probed TABPFN’s inductive biases qualitatively, and Hu and Ghelichi (2026) found robustness to random, label-uncorrelated columns in controlled synthetic binary tasks. The latter result is compatible with a geometry-only account, which our complementary interventions probe (§5.3).

MISMATCH draws on the cone effect (Ethayarajh, 2019; Gao et al., 2019) and quantitative anisotropy measures (Rudman et al., 2022), recast here as a backbone-readability test. The intervention design is motivated by shortcut-learning concerns (Geirhos et al., 2020), mechanistic-interpretability analyses (Olah et al., 2020), and convergent-validity logic (Campbell and Fiske, 1959). By conditioning on the bidirectional structure, the experiments probe marginal covariance, row alignment, and broader pre-projection geometry through complementary perturbations that do not perfectly isolate each factor.

7 Conclusion

The MISMATCH $\times$ SUSCEPTIBILITY framework provides an outcome-free per-tuple mechanism score for frozen-text injection in ICL tabular models. The intervention results support a dependence on row-aligned embedding structure whose usefulness covaries with encoder geometry. A Gaussian draw from the estimated covariance target and row-shuffled content do not reproduce the high-stratum gains; pre-PCA whitening shows additional sensitivity to broad geometric change but does not isolate the covariance spectrum as the unique cause. Mean- Δ reporting obscures the high-/low- S split. In our seven-encoder diagnostic, PCA-controlled, task-aligned text quality tracks gain more closely than MTEB. Classical isotropy summaries correlate with gain globally but do not replace MISMATCH under family-held-out sign evaluation. Tests on vision-language adapters and time-series models with categorical metadata are needed to determine whether the framework extends beyond ICL tabular models.

Limitations

Our evaluation covers two ICL tabular backbones and four main encoders. The best LFO result is only 0.626 AUROC; thus, S is a mechanism stratifier, not a deployment or model-selection gate.

The interventions establish dependence on row-aligned embedding structure, but do not uniquely prove that linguistic semantics is causal; correlated lexical, stylistic, or annotation artifacts remain possible explanations. ZCA whitening changes angular and neighborhood geometry together with the covariance spectrum, so I3 cannot identify spectrum alone as the causal component. The dataset-level PCA is also fit before cross-validation and is therefore transductive, although it uses no labels; a strict inductive deployment should fit PCA within the available training rows. The public datasets have heterogeneous licenses and may contain sensitive user-authored text. We redistribute no raw data, performed no new record-level privacy audit, and discourage consequential deployment without separate validation.

The cached-embedding MISMATCH $\times$ SUSCEPTIBILITY sweep is bounded by 6.1 RTX 4090 graphics processing unit (GPU)-hours, but missing historical logs prevent a defensible end-to-end total. Appendix K reports checkpoint identifiers, parameter counts, and software versions.

Acknowledgments

We thank the National University of Singapore and its School of Computing for their support. We used ChatGPT to assist with polishing the manuscript; all technical content, experiments, and claims were authored and verified by the authors.

References

- Alan Arazi, Eilam Shapira, and Roi Reichart. 2025. [TabSTAR: A tabular foundation model for tabular data with text fields](#). In *Advances in Neural Information Processing Systems 38*, pages 190695–190748.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.
- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. [XLM-T: Multilingual lan-](#)

- guage models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 258–266, Marseille, France. European Language Resources Association.
- Barry Becker and Ronny Kohavi. 1996. [Adult \[dataset\]](#). UCI Machine Learning Repository. License: CC BY 4.0. OpenML ID 1590.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Donald T. Campbell and Donald W. Fiske. 1959. [Convergent and discriminant validation by the multitrait-multimethod matrix](#). *Psychological Bulletin*, 56(2):81–105.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A benchmark dataset for legal language understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. [Modeling wine preferences by data mining from physicochemical properties](#). *Decision Support Systems*, 47(4):547–553.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4040–4054. Association for Computational Linguistics.
- Robert Detrano, Andras Jánosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. 1989. [International application of a new probability algorithm for the diagnosis of coronary artery disease](#). *The American Journal of Cardiology*, 64(5):304–310.
- Bradley Efron and Robert Tibshirani. 1986. [Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy](#). *Statistical Science*, 1(1):54–75.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. [Representation degeneration problem in training natural language generation models](#). In *International Conference on Learning Representations (ICLR)*.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2:665–673.
- Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, and 7 others. 2025. [TabPFN-2.5: Advancing the state of the art in tabular foundation models](#). *arXiv preprint arXiv:2511.08667*.
- Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. [Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports](#). *Journal of Biomedical Informatics*, 45(5):885–892.
- Hans Hofmann. 1994. [Statlog \(German credit data\) \[dataset\]](#). UCI Machine Learning Repository. License: CC BY 4.0. OpenML ID 31 (credit-g).
- Noah Hollmann, Samuel Müller, Katharina Eggenberger, and Frank Hutter. 2023. [TabPFN: A transformer that solves small tabular classification problems in a second](#). In *International Conference on Learning Representations (ICLR)*.
- Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. 2025. [Accurate predictions on small data with a tabular foundation model](#). *Nature*, 637(8045):319–326.
- Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. [Toward semantics-based answer pinpointing](#). In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*.
- James Hu and Mahdi Ghelichi. 2026. [Noise immunity in in-context tabular learning: An empirical robustness analysis of TabPFN’s attention mechanisms](#). *arXiv preprint arXiv:2604.04868*.
- IBM Sample Data Sets. 2018. Telco customer churn [dataset]. <https://www.kaggle.com/datasets/>

- [blastchar/telco-customer-churn](#). OpenML ID 42178 (telco-customer-churn); no associated publication.
- Gjergji Kasneci and Enkelejda Kasneci. 2024. [Enriching tabular data with contextual LLM embeddings: A comprehensive ablation study for ensemble classifiers](#). *arXiv preprint arXiv:2411.01645*.
- Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. [The multilingual Amazon reviews corpus](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568. Association for Computational Linguistics.
- Ron Kohavi. 1996. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 202–207. AAAI Press.
- Boshko Koloski, Andrei Margeloiu, Xiangjian Jiang, Blaž Škrlj, Nikola Simidjievski, and Mateja Jamnik. 2025. [LLM embeddings for deep learning on tabular data](#). *arXiv preprint arXiv:2502.11596*.
- Ken Lang. 1995. NewsWeeder: Learning to filter net-news. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339. Morgan Kaufmann.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.
- Xin Li and Dan Roth. 2002. [Learning question classifiers](#). In *COLING 2002: The 19th International Conference on Computational Linguistics*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#). *arXiv preprint arXiv:2308.03281*.
- Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. 2025. [TabDPT: Scaling tabular foundation models on real data](#). In *Advances in Neural Information Processing Systems* 38, pages 191387–191417.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. [Good debt or bad debt: Detecting semantic orientations in economic texts](#). *Journal of the Association for Information Science and Technology*, 65(4):782–796.
- Julian McAuley and Jure Leskovec. 2013. [Hidden factors and hidden topics: Understanding rating dimensions with review text](#). In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*, pages 165–172. ACM.
- Calvin McCarter. 2025. [What exactly has TabPFN learned to do?](#) *arXiv preprint arXiv:2502.08978*.
- Donald Michie, David J. Spiegelhalter, and Charles C. Taylor, editors. 1994. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.
- Rami M. Mohammad, Fadi Thabtah, and Lee McCuskey. 2012. An assessment of features related to phishing websites using an automated technique. In *Proceedings of the International Conference for Internet Technology and Secured Transactions (IC-ITST)*, pages 492–497, London, UK. IEEE.
- Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. [A data-driven approach to predict the success of bank telemarketing](#). *Decision Support Systems*, 62:22–31.
- Sérgio Moro, Raul M. S. Laureano, and Paulo Cortez. 2011. Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In *Proceedings of the European Simulation and Modelling Conference (ESM’2011)*, pages 117–121, Guimarães, Portugal. EUROSIS.
- Pete Mowforth and Barry Shepherd. n.d. [Statlog \(vehicle silhouettes\) \[dataset\]](#). UCI Machine Learning Repository. License: CC BY 4.0. OpenML ID 54 (vehicle).
- Martin Mráz, Breenda Das, Anshul Gupta, Lennart Purrucker, and Frank Hutter. 2025. [Towards benchmarking foundation models for tabular data with text](#). In *Foundation Models for Structured Data Workshop at the International Conference on Machine Learning (ICML)*.
- MTEB Project. 2026. [MTEB legacy leaderboard](#). Hugging Face Spaces. Archived 56-task leaderboard; accessed August 31, 2026.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.
- Jianmo Ni, Gustavo Hernandez Ábrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. [Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. [Zoom in: An introduction to circuits](#). *Distill*.
- Bo Pang and Lillian Lee. 2005. [Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.
- Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. 2025. [TabICL: A tabular foundation model for in-context learning on large data](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50817–50847. PMLR.
- William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. [IsoScore: Measuring the uniformity of embedding space utilization](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. [CARER: Contextualized affect representations for emotion recognition](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.
- Sentence Transformers. 2021a. [all-MiniLM-L6-v2 model card](#). Hugging Face. Accessed August 31, 2026.
- Sentence Transformers. 2021b. [all-mpnet-base-v2 model card](#). Hugging Face. Accessed August 31, 2026.
- Alen Shapiro. 1983. [Chess \(king-rook vs. king-pawn\) \[dataset\]](#). UCI Machine Learning Repository. License: CC BY 4.0. OpenML ID 3 (kr-vs-kp).
- Eva Sharma, Chen Li, and Lu Wang. 2019. [BIG-PATENT: A large-scale dataset for abstractive and coherent summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.
- Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care (SCAMC)*, pages 261–265. IEEE Computer Society Press.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- UCI Machine Learning Repository. 1981. [Mushroom \[dataset\]](#). UCI Machine Learning Repository. Records drawn from *The Audubon Society Field Guide to North American Mushrooms* (1981); donated by J. Schlimmer. License: CC BY 4.0. OpenML ID 24.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-SNE](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2014. [OpenML: Networked science in machine learning](#). *ACM SIGKDD Explorations Newsletter*, 15(2):49–60.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *arXiv preprint arXiv:2212.03533*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916. Association for Computational Linguistics.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general Chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.
- I-Cheng Yeh and Che-hui Lien. 2009. [The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients](#). *Expert Systems with Applications*, 36(2):2473–2480.
- Zeroshot. 2022a. Twitter financial news sentiment [dataset]. <https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment>. License: MIT. No associated publication.
- Zeroshot. 2022b. Twitter financial news topic [dataset]. <https://huggingface.co/datasets/zeroshot/twitter-financial-news-topic>. License: MIT. No associated publication.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 649–657.

A Statistical Methodology

Paired Wilcoxon signed-rank tests. For every row of Table 1 we compute the two-sided paired Wilcoxon signed-rank statistic between Δ_{real} and Δ_{int} at the tuple level. On the high- S stratum every p -value is at most 8.8×10^{-16} (target-covariance Gaussian $p=7.3 \times 10^{-16}$; shuffle-text $p=4.6 \times 10^{-16}$; pre-PCA whitening $p=8.8 \times 10^{-16}$). On the low- S stratum the p -values are 0.38, 0.49, and 0.98 respectively, all consistent with the null of zero paired difference. The intervention analysis includes only the fixed high- and low- S strata; no mid-stratum intervention result is reported.

Dataset-grouped bootstrap. All LFO AUROCs are reported with dataset-level grouped bootstrap CIs ($B=2000$). Resampling whole datasets preserves the intra-dataset correlation across backbones and encoders and avoids spuriously narrow intervals.

B Robustness Checks

Sliced Wasserstein variant of S_2 . We replace S_2 (the flattened one-dimensional KS statistic) with the sliced Wasserstein distance against $\mathcal{N}(0, I_k)$ using 50 random projections. The Spearman rank correlation of S_2 and its sliced-Wasserstein variant across all 732 tuples exceeds 0.85. All LFO AUROCs in Table 2 change by less than 0.005 when the variant is used.

Strict 3-fold LFO (excluding FAM3). The main Table 2 uses the inclusive 4-fold LFO with FAM3 multi-aspect ($n_- = 2$ negatives) treated as multi-class. For a more conservative check we exclude FAM3 and recompute on the 3 folds (FAM1, FAM4, FAM6; $n=339$ signed tuples). The linear composite remains best, while the remaining calibrated variants lie within 0.006 AUROC and no pairwise difference is significant: $\hat{p}_{\text{lin}}=0.584 [0.518, 0.657]$, $\hat{p}_{\gamma}=0.558 [0.502, 0.620]$ ($p_{\text{linear}}=0.46$), $\text{SUSCEPTIBILITY}=0.553 [0.402, 0.688]$ ($p=0.56$), and $\text{MISMATCH}=0.552 [0.497, 0.616]$ ($p=0.32$). Excluding FAM3 lowers the linear composite by 0.041 AUROC. The linear-composite AUROC on the FAM3 fold is 0.906 in the inclusive analysis but is based on only two negative tuples. The intervention strata do not depend on either LFO fold definition.

C Crosslingual transfer detail

We fit (a, b, γ) on 448 tuples from the English and synthetic families and evaluate the resulting predictor on 77 signed fam4 tuples without target-family recalibration. The held-out set contains 69 positive and 8 negative tuples. AUROC is 0.630 with a dataset-grouped 95% CI of $[0.538, 0.840]$, below the fixed 0.65 transfer threshold. Because no in-language calibration-size sweep is reported, we make no sample-efficiency claim for closing this gap.

D TabPFN v2.6 intervention replication

We sample 30 high- S tuples (top 30% by raw S in the TabPFN v2.6 main matrix) and 30 low- S tuples (bottom 30%). We then reload the texts deterministically with seed 42, re-encode with the same encoder, project to $k=16$ PCA, and apply target-covariance Gaussian replacement. We recompute Δ via 5-fold CV on TabPFN v2.6.

Stratum	n	$\bar{\Delta}_{\text{real}}$	$\bar{\Delta}_{\text{int}}$	difference [CI]
high- S	30	+0.095	−0.003	−0.098 [−.170, −.036]
low- S	30	−0.004	+0.000	+0.004 [+0.002, +0.007]

Table 4: Target-covariance Gaussian intervention on TabPFN v2.6. Paired Wilcoxon $p=0.0044$ (high- S) and $p=0.0013$ (low- S). The low-stratum change is statistically detectable but only 0.4 percentage points.

E Cross-architecture replication detail

Table 5 reports stratum means for TabICL v2 and TabPFN v2.6 on the joined 356-tuple subsample.

Stratum	n	$\bar{\Delta}_{\text{TabICL}}$	$\bar{\Delta}_{\text{TabPFN}}$
high- S	107	+0.146	+0.141
mid- S	142	+0.069	+0.069
low- S	107	−0.003	−0.001

Table 5: Cross-architecture stratum-mean Δ . Stratification is induced by the same raw S in both backbones (computed once on TabICL).

Per-tuple cross-backbone correlations: Spearman $\rho = +0.687$ ($p=4.4 \times 10^{-51}$), Pearson $r = +0.989$ ($p=2.9 \times 10^{-299}$), sign-agreement on non-zero tuples = 80.2%. Per-stratum Spearman: high- S +0.95, mid- S +0.67, low- S +0.08 (the last is consistent with the noise floor when both backbones produce $\Delta \approx 0$). The TabPFN v2.6 intervention replication covers target-covariance Gaussian

replacement only. Replicating shuffle-text and pre-PCA whitening is left to future work because the paired sweeps exceed the compute budget of the present study.

F Per-dataset DBA versus MTEB

Within each of the eight fam2 English text-classification datasets, we rank the four encoders by (i) their DBA scores and (ii) MTEB rank, then compute the Spearman correlation against the encoder ranking induced by the observed Δ . The mean across datasets is the headline number reported in §5.5.

Dataset	ρ_{DBA}	ρ_{MTEB}
ag_news	+0.95	−1.00
amazon_polarity	+0.32	−0.32
dbpedia_14	+1.00	+0.80
imdb	+0.40	−0.40
rotten_tomatoes	+0.80	−0.40
sst2	+0.80	−0.40
trec_coarse	+0.80	+0.80
yelp_polarity	+0.20	+0.80
mean	+0.658	−0.015

Table 6: Per-dataset Spearman correlation between the encoder ranking by DBA (left) or by MTEB (right) and the observed encoder ranking by Δ . Each row uses $n=4$ encoders. Mean across datasets reported as $\bar{\rho}$ in §5.5.

G PCA-16 text-only encoder ranking

We evaluate the four main encoders plus ALL-MINILM-L6-V2 (Sentence Transformers, 2021a), ALL-MPNET-BASE-V2 (Sentence Transformers, 2021b), and E5-MISTRAL-7B (Wang et al., 2024). For each of the eight fam2 English text-classification datasets, we project the frozen embeddings to 16 dimensions and fit a 5-fold stratified logistic-regression classifier on those text features alone. No structured columns are used. Table 7 reports accuracy averaged over the eight tasks.

H Anisotropy comparison

For the anisotropy comparison, we join the raw-embedding caches to the main outcomes using the information-level suffix as part of the dataset identifier, yielding 336 matched tuples from fam1–fam5. Table 8 reports per-encoder means for raw-space IsoScore and for IsoScore and participation ratio (PR) computed on the same PCA-16 block used by MISMATCH (Rudman et al., 2022).

Association and held-out prediction. Raw-space IsoScore correlates with MISMATCH at $\rho = +0.483$ ($p=4.9 \times 10^{-21}$) and with continuous Δ at $\rho = +0.630$ ($p=1.3 \times 10^{-38}$). On the PCA-16 block, IsoScore and PR yield the same reported correlations on this joined subset: $\rho = -0.081$ with MISMATCH ($p=0.14$) and $\rho = -0.428$ with Δ ($p=2.3 \times 10^{-16}$). These signs differ across representation spaces, so no single isotropy scalar supplies a representation-invariant explanation.

For the paper’s stratification goal, we evaluate family-held-out sign prediction on the 141 signed tuples in this joined subset. MISMATCH alone reaches 0.574 AUROC, raw-space IsoScore 0.467, PCA-16 IsoScore 0.523, and PCA-16 PR 0.523; MISMATCH $\times$ SUSCEPTIBILITY reaches 0.576. Thus classical isotropy has a substantial global association with continuous gain, but it does not replace MISMATCH for cross-family sign stratification.

I Visual case study (AG News)

We re-encode AG News (400 rows, four topical classes) with BGE-LARGE-EN on its real-text and adversarial-filler variants, then apply target-covariance Gaussian replacement (I1) to both (Figure 4). The high- S original shows four separated class clusters in two-dimensional t-SNE; I1 collapses them into a single blob after drawing from a zero-mean Gaussian with covariance estimated from the original block. The finite draw matches that covariance only approximately. The low- S variant has no class structure to begin with and is visually unchanged. Because t-SNE is sensitive to its projection settings, Figure 4 is only an illustration; Table 1 provides the quantitative evidence.

J Pre-PCA whitening sign transitions

The low- S stratum contains 39 negative- Δ tuples. Pre-PCA whitening moves 14 (35.9%) into the neutral or positive zone, but the paired mean difference over all 112 low- S tuples is only +0.0004 with $p=0.98$. Target-covariance Gaussian replacement and shuffling produce 18/39 and 17/39 such crossings, respectively. The count is therefore descriptive and not specific evidence that whitening systematically rescues harmful injection. On the high- S stratum, whitening instead removes the mean gain (I3 in Table 1), showing that the same broad geometric change has strongly stratum-dependent consequences.

Encoder	Native dimension	PCA dimension	Text-only accuracy	MTEB score	$\bar{\Delta}_8$
E5-LARGE-V2	1024	16	0.873	62.3	0.460
BGE-LARGE-EN	1024	16	0.869	64.2	0.443
E5-MISTRAL-7B	4096	16	0.868	66.6	0.444
SENTENCE-T5-BASE	768	16	0.849	57.0	0.432
GTE-LARGE	1024	16	0.848	63.1	0.435
ALL-MPNET-BASE-V2	768	16	0.834	57.8	0.395
ALL-MINILM-L6-V2	384	16	0.763	56.3	0.321

Table 7: Text-only performance after the same PCA-16 reduction used in the tabular pipeline. $\bar{\Delta}_8$ is the encoder’s mean downstream gain on the same eight datasets. MTEB scores are legacy 56-task averages from the archived leaderboard (accessed August 31, 2026) (MTEB Project, 2026). Text-only accuracy correlates with $\bar{\Delta}_8$ at $\rho = +0.93$ ($p=0.003$); MTEB score reaches $\rho = +0.75$ ($p=0.052$).

Encoder	Iso _{raw}	Iso ₁₆	PR ₁₆
BGE-LARGE-EN	0.804	0.854	0.760
GTE-LARGE	0.802	0.856	0.765
E5-LARGE-V2	0.804	0.854	0.772
SENTENCE-T5-BASE	0.796	0.860	0.772

Table 8: Encoder-level anisotropy aggregates over 336 joined tuples. Raw and PCA-16 quantities are not directly comparable in magnitude because their ambient dimensions differ.

K Computational cost

On a single consumer GPU (NVIDIA RTX 4090), the full MISMATCH $\times$ SUSCEPTIBILITY computation per (dataset, backbone, encoder) tuple takes under 30 seconds: < 5 s for MISMATCH and ≈ 25 s for the $2k$ backbone forward passes that produce SUSCEPTIBILITY. Across 732 tuples this is at most 21,960 GPU-seconds, or 6.1 RTX 4090 GPU-hours, once embeddings are cached. Historical end-to-end timing logs were not retained, so this measured score-sweep bound does not include embedding extraction, the paired 5-fold Δ runs, exploratory analyses, or intervention reruns.

The frozen backbones contain 10,732,042 parameters (TabPFN v2.6 default checkpoint) and 27,552,258 parameters (TabICL classifier v2). The four main frozen encoders contain 335,142,400 parameters each for BGE-large-en-v1.5, E5-large-v2, and GTE-large, and 110,218,368 for Sentence-T5-base. The three diagnostic-only encoders contain 22,713,728 (all-MiniLM-L6-v2), 109,486,978 (all-mpnet-base-v2), and 7,110,660,096 (E5-Mistral-7B-instruct) parameters. No pretrained model parameter is updated.

The recorded environment is Python 3.12.11, Compute Unified Device Architecture (CUDA) 12.8, PyTorch 2.10.0, NumPy 2.2.6, pandas 2.3.1, SciPy 1.15.2, scikit-

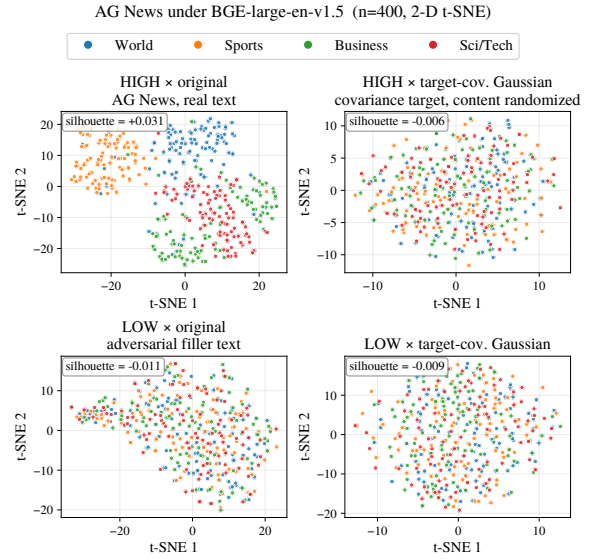


Figure 4: Case-study t-distributed stochastic neighbor embedding (t-SNE) of 1024-dimensional BGE-LARGE-EN embeddings on AG News ($n=400$). Top: real text; bottom: adversarial filler. Silhouette scores are computed in the original 1024-dimensional space.

learn 1.6.1, Transformers 4.56.0, Sentence Transformers 4.0.1, TabPFN 7.1.1, and TabICL 2.1.1. The resolved checkpoints are tabPFN-v2.6-classifier-v2.6_default.ckpt and tabICL-classifier-v2-20260212.ckpt. Versioned result files retain the earlier string tabPFN-2.5 as a legacy label; the checkpoint identifier above is authoritative. The code release provides the exact relevant package set and the TabICL commit hash.

All LFO standardization moments and calibration parameters, including (a, b, γ) , are fit on the calibration families only. Each PCA basis is fit without labels on all rows of its own dataset, including held-out datasets, as described in §2; the held-out family’s Δ outcomes and sign labels are never used for calibration.

L Inclusive evaluation including the gray zone

The main Table 2 excludes $|\Delta| \leq \varepsilon$ tuples; the retained-mass loss on the TabICL matrix is $1 - 266/376 = 29.3\%$. On all 376 TabICL tuples, including the gray zone, we report the per-fold weighted AUROC for both target classes and a per-fold weighted Spearman of the score against the continuous Δ (Table 9). Scoring uses the same jointly fitted model as Table 2 (i.e., $\hat{p}_\gamma = \sigma(b + aS_\gamma)$ with (a, b, γ) trained on the signed training fold), applied to all test-fold tuples.

Score	AUROC ₊₁	AUROC ₋₁	$\rho_{\text{per-fold}}$
$\hat{p}_{\text{lin}}$ (cal.)	0.587	0.596	+0.194
$\hat{p}_\gamma$ (cal.)	0.574	0.565	+0.106
MISMATCH (cal.)	0.549	0.558	+0.094
SUSCEPTIBILITY (cal.)	0.572	0.623	+0.240

Table 9: Inclusive evaluation on all 376 TabICL tuples (+1: 177, 0: 110, -1: 89; retained mass 0.707). AUROC₊₁ is one-versus-rest for the $+\Delta$ class with the gray zone in the negative class; AUROC₋₁ uses the same convention for the $-\Delta$ class, scoring with the negative predicted probability. $\rho_{\text{per-fold}}$ is the Spearman rank correlation of the probability-style score against continuous Δ , computed per fold and weighted by fold size; per-fold computation avoids the cross-fold scale issue that contaminates a pooled correlation.

Adding the gray zone lowers the positive-class AUROC of $\hat{p}_{\text{lin}}$ from 0.626 (Table 2) to 0.587. Gray-zone tuples have geometric profiles that overlap the $+\Delta$ class.

For the $-\Delta$ class, SUSCEPTIBILITY alone has the highest AUROC (0.623) and the highest per-fold correlation with continuous Δ ($\rho = +0.240$). Its contribution is concentrated on the low- S end of the matrix, where backbone sensitivity separates harmful injection from near-zero effects.

We also fit a least-squares linear rescaling of the probability-style score to continuous Δ , calibrated on the signed training fold. Its per-fold weighted mean-squared error (MSE) exceeds the per-fold mean- Δ baseline. We therefore report the predictor only as a sign stratifier, without treating it as a magnitude regressor.

M Dataset source inventory

The valid main matrix contains 48 externally sourced datasets and ten datasets generated by our synthetic-data code. The twelve FAM1 tables are hosted by OpenML and retain their original tabu-

lar features; their platform, canonical papers, and dataset records are cited here (Vanschoren et al., 2014; Kohavi, 1996; Becker and Kohavi, 1996; Moro et al., 2014, 2011; UCI Machine Learning Repository, 1981; IBM Sample Data Sets, 2018; Cortez et al., 2009; Hofmann, 1994; Michie et al., 1994; Yeh and Lien, 2009; Mowforth and Shepherd, n.d.; Mohammad et al., 2012; Smith et al., 1988; Detrano et al., 1989; Shapiro, 1983).

The ten FAM2 English text-classification datasets use their canonical benchmark sources (Zhang et al., 2015; Maas et al., 2011; Socher et al., 2013; Wang et al., 2019; Pang and Lee, 2005; Lehmann et al., 2015; McAuley and Leskovec, 2013; Li and Roth, 2002; Hovy et al., 2001; Saravia et al., 2018; Lang, 1995). The seven FAM3 multi-aspect datasets are sourced from Yelp, SNLI, TweetEval, and GoEmotions (Zhang et al., 2015; Bowman et al., 2015; Barbieri et al., 2020; Demszky et al., 2020). The ten FAM4 crosslingual datasets use Multilingual Amazon Reviews, multilingual TweetEval, and XNLI (Keung et al., 2020; Barbieri et al., 2022; Conneau et al., 2018). The nine FAM5 specialty datasets use their financial, biomedical, legal, patent, and social-media sources (Malo et al., 2014; Zeroshot, 2022a,b; Gurulingappa et al., 2012; Chalkidis et al., 2022; Sharma et al., 2019; Barbieri et al., 2020). The remaining ten FAM6 datasets are generated by our structural-causal-model-style data generator and therefore have no external dataset citation. Exact OpenML identifiers and Hugging Face paths are provided in the released dataset manifest.