A Pure Single-Hypersphere SVM for Native Sparse Multiclass Text

Brian Mingus Cognatory, LLC
brian@cognatory.com July 28, 2026

Abstract. This paper specifies and measures the active `PureHypersphereSVM`: one center, one nonnegative scalar squared radius, one high-dimensional joint hyperball, and one hypersphere boundary. A fixed regular simplex encodes the labels; a sparse linear input map is tensored with that code; and multiclass distances are evaluated exactly without materializing a Kronecker tensor. Training uses one primal containment/exclusion/ranking objective and an exact one-dimensional conditional solve for the radius at every center iterate. The frozen `native-sparse-text-v1` development suite comprises 12 publisher-supplied sparse matrices, five stratified folds per matrix, and 242,573 out-of-fold predictions. With equal dataset weight, the verified development result is **87.95% accuracy**, **86.38% macro-F1**, and **83.903/100 dummy-normalized skill**. These are development OOF measurements, not official-test, holdout, or state-of-the-art claims. The experiment performs no Cognatory tokenization, TF–IDF, feature hashing, column selection, or feature rescaling. Closed-set accuracy and macro-F1 do not by themselves certify hyperball feasibility or membership calibration.

1 Scope and contribution

The model studied here is deliberately narrow. For every input and every possible label it constructs one point in a shared high-dimensional joint space, measures that point against one center, and uses one hypersphere boundary. The deployed decision system has no secondary classifier, conditional dispatch, or ensemble. This is a direct multiclass construction rather than a decomposition into binary machines [2]. Its target regime is native high-dimensional sparse text, where linear large-margin methods have a long empirical history [3].

The contribution of this document is fourfold. First, it states the exact geometry in the active `pure_hypersphere` model, geometry, and feature modules. Second, it explains how the tensor-product geometry reduces to sparse matrix multiplication while remaining mathematically identical to one hypersphere. Third, it records the immutable, fail-closed development protocol used for the first multi-dataset ratchet. Fourth, it reports the resulting 60-fold OOF baseline and its complete confusion atlas. Earlier experiments or architectures are outside the evidence base of this paper.

The term "SVM" is used in its large-margin sense. Rows whose containment, exclusion, or ranking constraints are active are recorded as support diagnostics; they are not asserted to be nonzero dual coefficients. The implementation is a primal joint-space machine, and this paper does not claim a dual solver that the code does not contain.

2 One joint hypersphere

2.1 Fixed regular-simplex labels

Let the observed class set be $\mathcal{Y} = \{1, \dots, K\}$ with $K \geq 2$. The implementation creates fixed vectors $e_y \in \mathbb{R}^{K-1}$ satisfying

$$\sum_{y=1}^{K} e_y = 0, \qquad e_y^\top e_{y'} = \begin{cases} 1, & y = y', \\ -1/(K-1), & y \neq y'. \end{cases} \tag{1}$$

Thus every label has the same norm and every distinct pair has the same angle. The codes span the complete centered $(K-1)$-dimensional label space without assigning a privileged direction to any class. This is the regular-simplex coding used by the verified baseline; the label geometry is not learned.

2.2 Joint input–label map

Let $x \in \mathbb{R}^d$ denote one row of a publisher's sparse numeric matrix and let b be a shared bias scale. Define $\bar{x} = [x^\top, b]^\top \in \mathbb{R}^{d+1}$, $D = (d+1)(K-1)$, and the joint point

$$z(x, y) = \bar{x} \otimes e_y \in \mathbb{R}^D. \tag{2}$$

The learned joint-space geometry is one center $c \in \mathbb{R}^D$ and one scalar $\rho \geq 0$, where ρ is the squared radius. For $\rho > 0$, its D-dimensional closed hyperball and $(D-1)$-dimensional hypersphere boundary are

$$\mathbb{B}^D(c, \rho) = \{z \in \mathbb{R}^D : \|z - c\|_2^2 \leq \rho\}, \tag{3}$$

$$\mathbb{S}^{D-1}(c, \rho) = \partial \mathbb{B}^D(c, \rho) = \{z \in \mathbb{R}^D : \|z - c\|_2^2 = \rho\}. \tag{4}$$

At $\rho = 0$, both sets degenerate to the singleton $\{c\}$. This is not a three-dimensional sphere. For example, News20 multiclass has $d = 130{,}107$ and $K = 20$, hence $D = 2{,}472{,}052$. For a labeled example (x_i, y_i), the true joint point is penalized when it lies outside the hyperball. Every incorrect label creates a different joint point for the same input; the closest such point is penalized whenever its squared distance is less than $\rho + m_{\text{out}}$—that is, whenever it has not cleared the global exclusion margin beyond the same boundary. These are soft penalties rather than guaranteed feasibility. No class-specific center or radius is independently parameterized; every label-slice region is induced by the same joint center and global boundary.

2.3 Factored distance, not multiple heads

Let $c = \operatorname{rvec}(A)$ under the row-vectorization convention satisfying $\langle c, \bar{x} \otimes e_y \rangle = \bar{x}^\top A e_y$, with $A \in \mathbb{R}^{(d+1)\times(K-1)}$. Because every e_y has unit norm, the algebraically exact joint distance is

$$d(x, y) = \|z(x, y) - c\|_2^2 = \|\bar{x}\|_2^2 + \|A\|_F^2 - 2\bar{x}^\top A e_y. \tag{5}$$

Equation 5 is the implementation's central scaling identity. It evaluates all K distances through a sparse product $\bar{X}A$ and a

dense product with the simplex matrix. The D Kronecker coordinates are never materialized. Its float32 evaluation is subject to ordinary roundoff and cancellation near the center, including small negative computed values for a mathematically nonnegative squared distance. With $f_y(x) = \bar{x}^\top A e_y$, the columns of A are not independently trained heads; together with the fixed simplex they parameterize one zero-sum multiclass affine scorer induced by the joint center. Indeed, $\sum_y A e_y = 0$, and the full-rank simplex spans every such zero-sum score family. For $w_y = A e_y$, each fixed-label slice can also be written

$$d(x, y) = \|\bar{x} - w_y\|_2^2 + \|A\|_F^2 - \|w_y\|_2^2. \tag{6}$$

The slice centers and effective thresholds are therefore induced views of the one joint center and one global ρ, not separately stored class models.

Prediction and hyperball membership are derived from the same signed hyperspherical score:

$$s(x, y) = \rho - d(x, y), \tag{7}$$

$$\widehat{y}(x) = \arg\min_{y \in \mathcal{Y}} d(x, y) = \arg\max_{y \in \mathcal{Y}} f_y(x), \tag{8}$$

$$M_y(x) = \mathbf{1}\{d(x, y) \le \rho\}, \tag{9}$$

$$\mathrm{inside}(x) = \max_y M_y(x) = \mathbf{1}\{d(x, \widehat{y}) \le \rho\}. \tag{10}$$

The common ρ cancels when labels are ranked, so closed-set prediction is exactly affine linear and always returns a class even when every candidate lies outside the hyperball. It remains operative in membership and in the training constraints. Thus accuracy and macro-F1 alone do not test the radius. The radius is one global joint-space acceptance boundary, not an independently fitted per-class threshold; an abstaining classifier would require an explicit reject convention based on $\mathrm{inside}(x)$. Exact distance ties follow the implementation's stable class order and select the first entry in the sorted stored class array, hence the smallest observed class label.

3 Large-margin hypersphere objective

For training row i, define the true-pair distance and exact closest-wrong distance as

$$d_i^+ = d(x_i, y_i), \qquad d_i^- = \min_{y \ne y_i} d(x_i, y). \tag{11}$$

Let $w_i \ge 0$, $W = \sum_i w_i > 0$, $C_+ > 0$, $C_-, C_r \ge 0$, $m_{\mathrm{out}}, m_{\mathrm{rank}} \ge 0$, and $p \in \{1, 2\}$. The soft constrained primal is

$$\begin{aligned} \min_{c, \rho, \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\zeta}} \quad & \rho + \frac{C_+}{W} \sum_i w_i \xi_i + \frac{C_-}{W} \sum_i w_i \eta_i + \frac{C_r}{W} \sum_i w_i \zeta_i^p \\ \text{s.t.} \quad & d(x_i, y_i) \le \rho + \xi_i, \\ & d(x_i, y) \ge \rho + m_{\mathrm{out}} - \eta_i, \quad y \ne y_i, \\ & d(x_i, y) - d(x_i, y_i) \ge m_{\mathrm{rank}} - \zeta_i, \quad y \ne y_i, \\ & \rho, \xi_i, \eta_i, \zeta_i \ge 0. \end{aligned} \tag{12}$$

One shared exclusion slack and one shared ranking slack are used per row. Eliminating them takes the worst incorrect label exactly, yielding the implemented core objective

$$\begin{aligned} \mathcal{L}(c, \rho) = \rho &+ \frac{C_+}{W} \sum_i w_i \left[d_i^+ - \rho \right]_+ \\ &+ \frac{C_-}{W} \sum_i w_i \left[\rho + m_{\mathrm{out}} - d_i^- \right]_+ \\ &+ \frac{C_r}{W} \sum_i w_i \left[m_{\mathrm{rank}} + d_i^+ - d_i^- \right]_+^p . \end{aligned} \tag{13}$$

The leading ρ term explicitly shrinks the squared radius. It is the standard SVDD radius surrogate [1], not the literal D-dimensional volume $\pi^{D/2} \rho^{D/2} / \Gamma(D/2 + 1)$. Radius and volume are monotone-equivalent for a hard feasible problem, but fixed soft-margin tradeoffs would differ under a literal volume penalty.

The three hinge families have direct geometric meanings:

- *containment*: a correct joint point outside the hyperball pays a penalty;
- *exclusion*: the closest incorrect joint point pays a penalty until it clears the global squared-distance margin beyond the boundary;
- *ranking*: an additional structured loss asks that the correct point be closer than the closest incorrect point by m_{rank} [2].

Zero exclusion slack implies that every incorrect joint point for that row is beyond the shared margin; positive slack means outside placement is penalized, not guaranteed. At fixed c, all containment and exclusion slacks can vanish for one shared radius if and only if

$$\max_i d_i^+ \le \min_i \left(d_i^- - m_{\mathrm{out}} \right), \tag{14}$$

and all ranking slacks can vanish if and only if

$$\min_i \left(d_i^- - d_i^+ \right) \ge m_{\mathrm{rank}}. \tag{15}$$

The first condition couples every row through one global ρ; the second is the rowwise closest-wrong ranking condition. A zero-slack hard solution requires both. All wrong-label comparisons are exhaustive over the observed class set rather than sampled. The reported baseline uses $p = 2$ for the ranking hinge, $C_+ = 1/\nu = 10$, $C_- = 1$, $C_r = 10{,}000$, $m_{\mathrm{out}} = 0.05$, $m_{\mathrm{rank}} = 1$, and uniform training weights.

3.1 Exact scalar radius solve

For fixed c, the ranking term is independent of ρ and the remainder of Equation 13 is convex and piecewise linear in $\rho \ge 0$. Away from ties, its slope is

$$g(\rho) = 1 - \frac{C_+}{W} \sum_{d_i^+ > \rho} w_i + \frac{C_-}{W} \sum_{d_i^- - m_{\mathrm{out}} < \rho} w_i. \tag{16}$$

The slope can change only at the breakpoints d_i^+ and $d_i^- - m_{\mathrm{out}}$. The implementation stably orders those breakpoints and returns zero if the initial subgradient is nonnegative; otherwise it returns the first breakpoint at which the cumulative

slope becomes nonnegative. This is a global minimizer of the exact scalar subproblem conditional on fixed c, not a global solution of the joint center–radius problem. Writing $J(c) = \min_{\rho \geq 0} \mathcal{L}(c, \rho)$, the center problem remains nonconvex and is handled by a finite run of full-batch Adam.

The frozen source snapshot underlying Table 1 detached the selected breakpoint before the center update. At an ordinary-hinge knot, the automatic derivative of the zero-valued hinge is also zero; consequently that historical Adam step was a partial subgradient with ρ held fixed and was not always a KKT-consistent subgradient of $J(c)$. The loss value and conditional radius were exact, but the run is not a certificate of a global minimizer of Equation 13. The corrected active in-memory and streamed trainers assign the residual radial KKT mass to tied boundary events in their backward passes. The frozen results are retained as historical evidence for their exact source snapshot and are not retroactively attributed to that correction.

3.2 Constraint-active support diagnostics

After the best center is restored and the exact conditional radius is recomputed, row i is reported as a support diagnostic when any one of the following is active or violated within the configured numerical tolerance:

$$d_i^+ - \rho \geq -\varepsilon, \tag{17}$$

$$\rho + m_{\text{out}} - d_i^- \geq -\varepsilon, \tag{18}$$

$$m_{\text{rank}} + d_i^+ - d_i^- \geq -\varepsilon. \tag{19}$$

The implementation records containment, exclusion, and ranking indices separately. These tolerance-selected rows are not claimed to be certified nonzero dual support vectors or a KKT certificate. The baseline retains their indices and labels but does not duplicate the corresponding sparse feature rows in model state.

4 Sparse primal implementation

The native input is converted only to a finite, canonical, sorted float32 CSR matrix. Duplicate sparse coordinates are summed; no dense vocabulary-sized array is formed. Forward and backward center products use complete, unsampled CSR–dense products subject to float32 roundoff. The custom sparse backward path uses the canonical transpose and a cached row-block plan; bounded intermediate allocations and ordered partial sums change workspace use, not the algebra in Equation 5. Prediction omits the unused transpose and processes rows in bounded batches.

The development baseline is the strictly linear path: linear feature weight 1, interaction-sketch weight 0, fixed simplex labels, and disabled topology loss. The center is the only gradient-updated geometric parameter; the shared squared radius is re-optimized exactly, conditionally on that center, at every iterate. Full-batch Adam uses a learning rate of 0.005, at most 200 epochs, at least 80 epochs, and patience 60. The shared bias scale is 0.25; the center starts from an equally class-weighted mean of joint training points. Training is otherwise sample-unweighted. The run seed is 20260727, deterministic PyTorch algorithms are requested, and TF32 is disabled for the reference CUDA run.

4.1 Machine-checked purity

Every fold serializes a purity manifest. The report validator requires model type `PureHypersphereSVM`, exactly one center, exactly one radius, simplex rank $K - 1$, the hyperspherical decision rule in Equation 8, feature-map weights consistent with the effective configuration, and empty lists for every auxiliary classifier, routing mechanism, and ensemble. The static purity signature must match across all 60 folds. For this report its SHA-256 fingerprint begins `cbf73b61d51c`; the complete digest is stored in the aggregate report.

This manifest is evidence about serialized decision structure, not a slogan. A fold that changes the decision rule or activates an unreported feature path cannot enter the aggregate.

5 Immutable development protocol

5.1 Native publisher matrices

The suite consumes numeric sparse matrices released by their dataset publishers or public dataset hosts. Some contain counts and some contain weights already chosen upstream. Cognatory does not reconstruct text features: there is no Cognatory tokenization, TF–IDF fitting, feature hashing, column selection, rescaling, or shared artificial vocabulary. Parsing preserves the published column space. Schema checks bind row count, column count, class count, archive and member digests, canonical CSR bytes, and label bytes. The only task-specific label conversion is the publisher-defined IMDB rating polarity map (ratings 1–4 versus 7–10).

The frozen manifest is `native-sparse-text`, version 1. It contains exactly 12 datasets. Each artifact and each fold is digest-bound, and future suite versions may append datasets but may not silently remove or replace an incumbent dataset. The manifest's canonical fingerprint begins `559550d483ca`.

5.2 Five-fold OOF reconstruction

For every dataset, a seeded stratified five-fold partition is created using split seed 20260727. Each fold's train and validation indices are sorted, unique, disjoint, complementary, and hashed. Across the five folds, every row must occur in validation exactly once. This produces 60 fold results and 242,573 OOF predictions over 36,143,382 stored nonzeros.

Each fold writes an NPZ sidecar containing validation indices, truth, prediction, fold-majority dummy prediction, and inside-hyperball membership. Array and file digests are verified with pickle loading disabled. The reporter reconstructs each dataset's OOF vector, checks it against the frozen label digest, and derives the confusion matrix and metrics from those reconstructed predictions. It rejects missing datasets, missing folds, duplicate paths, configuration drift, source-code drift, or a changed purity signature. Behavioral source files are hashed before and after fitting.

The aggregate also binds the exact source snapshot that produced these numbers. The active implementation may receive numerical, optimization, or performance changes after that snapshot, but such changes do not retroactively inherit this result. A run from changed source is a new ratchet candi-

date until the complete frozen suite has been regenerated and validated.

5.3 Equal-dataset ratchet score

Accuracy and macro-F1 are first computed from each dataset's pooled OOF confusion matrix. Suite accuracy and macro-F1 are then the arithmetic means of the 12 dataset values, so a 3,000-row corpus and a 72,000-row corpus receive equal weight.

The ratchet also compares each metric m with the corresponding pooled OOF metric b from fold-specific training-majority predictions:

$$q(m;b) = \frac{m-b}{1-b}, \qquad S_j = 50\{q(F_j;b_{F,j}) + q(A_j;b_{A,j})\}. \tag{20}$$

The reported skill is $12^{-1}\sum_j S_j$. It is a transparent dummy-normalized development index, not a probability and not a leaderboard metric.

6 Verified development results

Table 1 reports one pooled OOF confusion result per dataset. The column count is the publisher schema retained by the loader; Sector has 49,087 observed active columns inside its 55,197-column schema, while the other matrices activate every published column. No number in the table comes from an official test partition.

Table 1: Frozen `native-sparse-text-v1` development OOF baseline. Accuracy and macro-F1 are percentages. Skill is the dummy-normalized index in Equation 20. The final row gives equal weight to each dataset.

Dataset	Rows	Columns	K	Accuracy	Macro-F1	Skill
Farm Ads	4,143	54,877	2	90.01	89.95	81.587
IMDB sentiment	25,000	89,527	2	84.87	84.87	73.518
LA1S	3,204	13,195	6	87.20	84.15	82.357
LA2S	3,075	12,432	6	87.87	85.39	83.504
LEDGAR	70,000	19,996	100	85.47	79.08	81.863
New3S	9,558	26,832	44	82.25	82.36	81.576
News20 binary	19,996	1,355,191	2	96.92	96.92	94.608
News20 multiclass	11,314	130,107	20	92.36	92.36	92.129
OHSCAL	11,162	11,465	10	76.22	75.13	73.333
Real-sim	72,309	20,958	2	97.46	97.01	93.341
SCOTUS	6,400	126,405	13	80.78	75.41	75.135
Sector	6,412	55,197	105	93.93	93.91	93.882
Equal-dataset mean	—	—	—	**87.95**	**86.38**	**83.903**

The result is broad rather than uniform. News20 binary and Real-sim exceed 97% macro-F1 or nearly so, while the legal SCOTUS task and medical OHSCAL task remain materially harder. LEDGAR's 100 classes produce 85.47% accuracy but 79.08% macro-F1, revealing class-sensitive error that accuracy alone would conceal. These gaps are exactly why the ratchet preserves datasets and shows native-class confusion structure instead of collapsing the suite to one scalar. Figure 1 exposes that native-class structure across the complete suite.

Across the 60 folds, measured fit time totaled 496.36 seconds and median fold fit time was 4.65 seconds on the recorded CUDA environment; the largest recorded peak allocated GPU memory was 357,806,592 bytes. These are provenance diagnostics, not hardware-normalized performance claims: the accelerator was shared during the run, and elapsed time should not be used as a comparative benchmark.

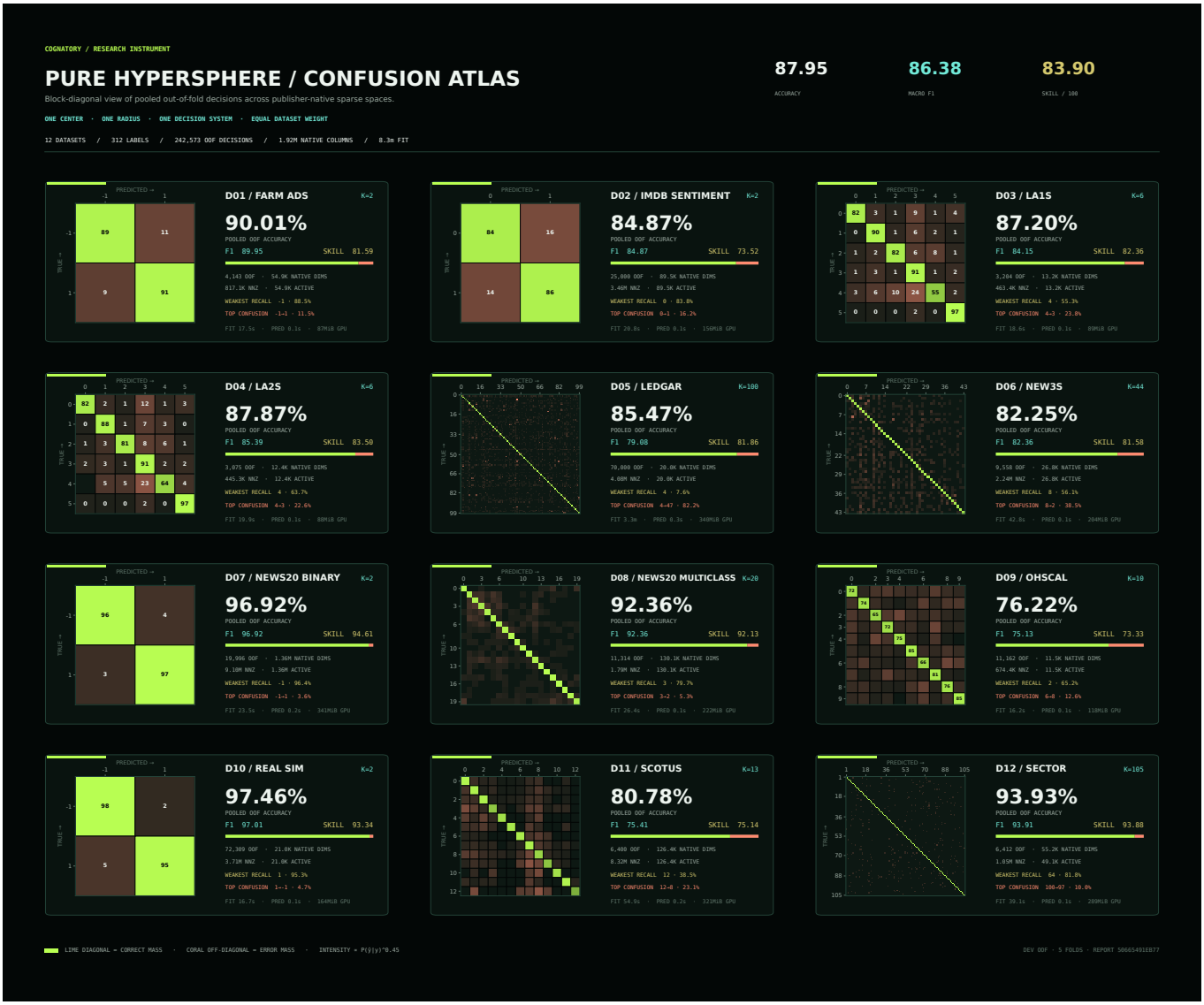

Figure 1: Vector-native meta-confusion atlas for the verified development OOF report. Every dataset retains its native class set (up to $K = 105$). Rows are true labels, columns are predicted labels, and cells are row-normalized. The design distinguishes correct diagonal mass from off-diagonal confusion and annotates each panel with its metrics, weakest recall, and strongest observed confusion. The atlas is generated from the same hardened aggregate report as Table 1; no synthetic preview or obsolete single-dataset image is used.

7 What the result does and does not establish

The 87.95% accuracy, 86.38% macro-F1, and 83.903 skill values establish a reproducible development baseline for this exact source/configuration/manifest tuple. They show that a fixed-simplex, single-center linear hypersphere can produce useful multiclass predictions on heterogeneous native sparse matrices without feature construction by Cognatory. They do not establish general superiority to other classifier families, state-of-the-art performance, calibrated open-set rejection, or official-test generalization.

The present evaluation is single-label multiclass even when a source corpus has a richer history elsewhere. It does not report extreme multi-label ranking metrics, probability calibration, tail-label propensity metrics, or a production holdout. Inside-hyperball membership is preserved in every prediction artifact, but it is not evaluated or calibrated in this report. Because ρ cancels from the nearest-label ranking, accuracy and macro-F1 are invariant to the fitted radius and do not validate containment, exclusion, rejection, or the hard feasibility conditions in Equations 14–15.

8 Research-audit boundary and next work

Three literature audits—ultra-scale exact optimization, learned shared geometry, and pure extreme multi-label classification—were used to set boundaries for the next ratchets. Their defensible conclusions are directions to test, not retroactive descriptions of the baseline.

8.1 Exact-preserving scale work

The highest-confidence scaling path is row-sharded CSR with one resident center, exact coverage of every shard, exact all-label wrong-pair scans, exact global radius statistics, and deterministic full-pass verification before stopping. Active-row caches may schedule work efficiently only if the final violation scan is exact. Sampled wrong-label maxima, approximate retrieval, unchecked shrinking, and independently trained shard models would change the objective or invalidate its exact full-pass evaluation. A stochastic sparse warm start followed by a deterministic bundle or cutting-plane finisher is theoretically attractive, but it has not been implemented or measured in this report [5, 6].

8.2 Shared geometry that remains one hypersphere

A shared positive-semidefinite input transform can be absorbed into one explicit map and therefore remains compatible with Equation 2. Diagonal or diagonal-plus-low-rank transforms are especially plausible for sparse text because they can reweight rare coordinates and capture a small number of correlated directions without a dense full-covariance map. Any future learned label geometry must remain one shared, equal-norm or explicitly regularized embedding used inside the same joint distance. Fixed simplex codes are a principled prior, not a generalization theorem [4]. None of these learned-geometry variants contributes to Table 1.

8.3 Pure multi-label extension

A pure multi-label system could define every label decision by the same predicate $\mathbf{1}\{d(x, y) \leq \rho\}$ and rank labels by the same $-d(x, y)$. That preserves one center and one global radius. It also exposes a hard geometric tension: documents with many separated positive-label codes require a larger common containment region, which can admit nearby negatives. Therefore a future multi-label study must report both ranking and globally thresholded set metrics, stratify by label cardinality, and treat global-radius feasibility as a hypothesis to test. This paper makes no multi-label performance claim.

9 Sealed official-test policy and limitations

The current loader rejects filenames resembling official test artifacts, and the recorded fold payloads assert that no official test data were accessed. Development ratchets may change the model only against frozen OOF manifests. An official test run will occur only after the implementation, hyperparameters, dataset mapping, metrics, and reporting code are declared final in a new signed protocol. Test truth will be materialized once; no model choice will be revised from that result. Until then, every number in this paper is labeled development OOF.

Several limitations remain. The 12 datasets are not statistically independent: LA1S and LA2S are related corpora, and suite versioning records that correlation. Equal dataset weighting is a policy choice rather than an estimate of deployment prevalence. The baseline uses publisher-prepared numeric matrices, so it evaluates the classifier rather than an end-to-end text feature pipeline. The optimizer is a deterministic full-batch Adam reference, not a convergence-certified large-scale solver or a certificate of the joint global optimum. The report does not yet retain training containment and exclusion violation rates or the global feasibility gap. Finally, one scalar radius can be restrictive for heterogeneous acceptance behavior; purity forbids concealing that limitation behind independently fitted per-class thresholds.

10 Conclusion

The active system is mathematically and operationally one high-dimensional hypersphere bounding one hyperball. A fixed regular simplex and a sparse linear input map create one joint space; one center and one scalar squared radius define its geometry; and containment, exclusion, and ranking constraints define its soft margin. Factored evaluation makes the construction practical without independently trained classifier heads. The immutable 12-dataset, 60-fold protocol supplies a reproducible closed-set baseline: 87.95% equal-dataset development OOF accuracy, 86.38% macro-F1, and 83.903 skill. The next credible gains must improve that same model, add direct hypersphere feasibility and membership evidence, and pass that same evidence discipline.

References

[1] D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.

[2] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML*, 1998.

[4] Y. Mroueh, T. Poggio, L. Rosasco, and J.-J. Slotine. Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems*, 2012.

[5] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *International Conference on Machine Learning*, 2007.

[6] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77:27–59, 2009.