MAHALANOBIS-SMOOTHED TRUST CONFORMAL PREDICTION UNDER LIMITED CALIBRATION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction endows any classifier with a distribution-free coverage guarantee, yet under limited calibration data the guarantee is nearly vacuous: the returned prediction sets routinely grow to almost the size of the full label space, eroding the very utility they are meant to provide. We trace this failure to a structural blind spot shared by every published conformity score: each reads the model solely through the post-softmax probability vector—a bottleneck that discards the class-conditional geometry the encoder has already computed, and which often separates the classes more cleanly than the linear head that projects them. We show that this geometry can be restored to the calibration pipeline at negligible cost: because every encoder already computes a hidden state alongside its logits, a simple per-class distance to training-set centroids recovers exactly the structure the softmax discards. Mahalanobis-Smoothed Trust Conformal (M-STC) builds on this observation—it blends the centroid distance with the familiar rank-regularised score on temperature-scaled logits, and a holdout slice of the calibration data picks the blend for each deployment—so the score relies on geometry where it helps and safely falls back to the softmax ranking where it does not. The guarantee survives this adaptivity on one condition, a data-handling discipline: the geometric statistics must be estimated on *training* data—estimating them on the calibration set breaks exchangeability and inflates set size by 30.9%. Across 18 cells of an LLM-style sequence-classification benchmark, M-STC delivers the smallest prediction set in $5/6$ settings and the smallest average against nine published baselines, reducing average set size by $\mathbf{44.3}\%$ over APS, $\mathbf{12.3}\%$ over RAPS, and $\mathbf{4.9}\%$ over the strongest competing baseline, while preserving the requested marginal coverage to within ± 0.02 in every evaluated cell.

1 INTRODUCTION

Transformer-based sequence classifiers are increasingly deployed for high-stakes language-model tasks (Brown et al., 2020; Devlin et al., 2019)—graded multiple-choice question answering (Kumar et al., 2023; Yang & Liu, 2025), retrieval re-ranking, content-policy enforcement—where a single argmax label is not enough and predictions must carry calibrated uncertainty. Conformal prediction (Vovk et al., 1999; 2005; Papadopoulos et al., 2002; Shafer & Vovk, 2008) offers precisely this, wrapping any black-box model in a procedure that returns a label set guaranteed to contain the truth with probability at least $1-\alpha$ on a fresh point, assuming only exchangeability between calibration and test data. It is an especially natural fit for language-model deployments, being post-hoc, model-agnostic, and free of additional training. Its practical value, however, hinges on set size: in the limited-calibration regime typical of real deployments ($\leq 1{,}000$ calibration sequences per domain), the guaranteed sets routinely fill 50–90% of the label space—a guarantee that loose says, in effect, that almost any label could be correct, and is of little practical use.

A decade of progress on conformity scores—LAC (Sadinle et al., 2019), APS (Romano et al., 2020), RAPS (Angelopoulos et al., 2020), SAPS (Huang et al., 2023), and recent refinements for class-conditional (Ding et al., 2023; Ghosh et al., 2026), long-tailed (Ding et al., 2026), trust-aware (Kaur et al., 2025), and multi-input (Fermanian et al., 2026) coverage—has steadily tightened these sets. Yet all of them read the same input, the post-softmax probability vector, and thereby discard a signal the model has already computed. A transformer's pre-classifier hidden representation separates classes geometrically, often more cleanly than the linear head projects them (Ren et al., 2021), and the

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

class-conditional distance of Lee et al. (2018)—introduced for out-of-distribution detection—captures exactly this structure. Scoring by that distance alone, however, discards the calibrated softmax ranking that scores such as RAPS exploit. The two signals are complementary, and their relative value shifts with the amount of training data, the target coverage, and the difficulty of each input; a strong conformity score should therefore use both and weigh them adaptively.

Mahalanobis-Smoothed Trust Conformal (M-STC) rests on a simple observation: *the geometry is already there*. Every encoder computes a hidden state $h(x)$ alongside its logits, so the class-conditional structure that the softmax compresses away can be read back out with nothing more than a set of training-set class centroids—no auxiliary heads, no retraining, no change to the model. M-STC turns this observation into a conformal score: the centroid distance is blended with the familiar rank-regularised softmax score, and a holdout slice of the calibration data chooses the blend weight for each deployment. The design is deliberately conservative—the search can always return $w{=}0$, so the score is never worse than its softmax endpoint (temperature-scaled RAPS)—and the marginal $1{-}\alpha$ guarantee survives on one condition: the centroids must be fitted on *training* data, for estimating them on the calibration set, an attractive shortcut under limited data, breaks exchangeability and inflates set size by 30.9%. As Figure 1 illustrates, M-STC converts the encoder's discarded geometry into uniformly smaller prediction sets that still satisfy the coverage guarantee.

Empirically, on an LLM-style sequence-classification benchmark evaluated against nine published baselines spanning 2017–2026, M-STC yields the smallest prediction set in nearly every setting while holding the requested coverage, at no extra training and negligible inference cost—directly applicable wherever a softmax-only conformal score is currently used. The improvement is largest exactly where it matters most: the low-calibration regime in which existing scores collapse to near-trivial sets. Beyond the numbers, the finding is conceptual: the encoder's class-conditional geometry is a source of conformal efficiency that the post-softmax view structurally cannot access, and it can be exploited without forfeiting the distribution-free coverage guarantee.

Concretely, this work makes three advances:

- **A geometry-augmented conformal score.** M-STC blends a rank-regularised softmax score with a class-conditional representation distance—to our knowledge the first conformal score to exploit the encoder's hidden representation rather than the post-softmax probability vector alone (Section 3).
- **Coverage-preserving adaptivity.** A per-cell blend weight chosen on a calibration holdout dominates the fixed-w family while preserving marginal coverage—provided the geometric statistics are fitted on *training* data, a discipline whose violation inflates set size by 30.9% (see Sections 3 and 4.4).
- **State-of-the-art efficiency.** Across 18 cells against nine baselines (2017–2026), M-STC gives the smallest set in $5/6$ settings and the smallest average, reducing set size by 44.3% over APS and 4.9% over the strongest baseline per cell (Sections 4, 4.5).

2 Background and Related Work

Conformal prediction. Inductive (split) conformal prediction (Vovk et al., 2005; Papadopoulos et al., 2002) takes a nonconformity score $s(x, y)$, computes its empirical $\lceil (n_{\text{cal}}{+}1)(1-\alpha)\rceil/n_{\text{cal}}$ quantile $\hat{q}$ on a held-out calibration set, and emits the set $C(x){=}\{y : s(x, y){\le}\hat{q}\}$ at test time. If calibration and test points are exchangeable, the marginal coverage $\Pr\big[y \in C(x)\big] \ge 1{-}\alpha$ holds in finite samples with no distributional assumptions (Shafer & Vovk, 2008; Lei et al., 2013; Angelopoulos & Bates, 2022; Angelopoulos et al., 2024; Balasubramanian et al., 2014). Extensions cover regression (Lei et al., 2018), covariate shift (Tibshirani et al., 2019), and online validity under distribution drift (Gibbs & Candes, 2021), beyond exchangeability (Barber et al., 2023), and conformal p-values for outlier detection (Bates et al., 2023), macro-coverage guarantees (Bhattacharyya et al., 2026), and audited validity under unknown distribution shift (Zhou et al., 2026).

Score families for classification. LAC (Sadinle et al., 2019) uses $s(x, y){=}1{-}p_y(x)$; APS (Romano et al., 2020) replaces this with the cumulative sorted probability up to rank y; RAPS (Angelopoulos et al., 2020) adds a rank penalty $\lambda(\text{rank}(y) - k_{\text{reg}})_+$; SAPS (Huang et al., 2023) keeps only the rank beyond top-1. Recent extensions target long-tailed labels (Ding et al., 2026), conditional

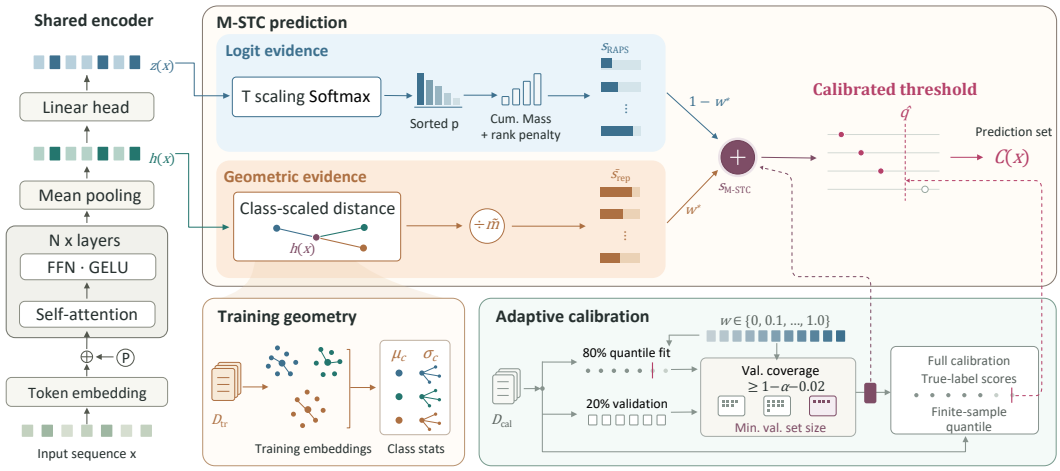

Figure 1: **Overview of M-STC.** A shared encoder supplies logits $z(x)$ and hidden representations $h(x)$. The logit branch computes temperature-scaled RAPS scores; the geometric branch uses class-scaled distances to training-set centroids, normalised by their median. The two scores are blended with weights $1 - w^*$ and w^*. An $80/20$ calibration split selects $w^* \in \{0, 0.1, \ldots, 1\}$ to minimise validation set size subject to coverage $\geq 1 - \alpha - 0.02$. The selected score is then calibrated on the full calibration set using the finite-sample quantile, yielding threshold $\hat{q}$ and prediction set $C(x)$.

coverage (Kaur et al., 2025), multi-input aggregation (Fermanian et al., 2026), per-class (Mondrian) coverage (Vovk et al., 2003; Ding et al., 2023), end-to-end training (Stutz et al., 2021), and composition with temperature scaling (Guo et al., 2017; Platt et al., 1999). Every one of these scores reads only the post-softmax probability vector; none exploits the encoder's hidden representation, which is where M-STC contributes.

Conformal prediction for language models. Conformal prediction has been applied to LLM multiple-choice QA (Kumar et al., 2023), to generative outputs via a stopping rule on sampled candidates (Quach et al., 2024), to open-ended generation with correctness–coverage guarantees (Wang et al., 2024), and to uncertainty benchmarking (Ye et al., 2024); recent work extends conformal guarantees to domain and configuration shift in LLM deployments (Lin et al., 2025; Zhu et al., 2026) and to in-context text classification (Pantelidis et al., 2025). All inherit the limited-calibration pathology M-STC tackles: their scores read only post-softmax probabilities, ignoring geometric structure in the hidden representation.

Mahalanobis-based scoring. Lee et al. (2018) introduced the tied-covariance Mahalanobis distance as a confidence score, originally for out-of-distribution detection—an alternative to the softmax-max baseline of Hendrycks & Gimpel (2016). Ren et al. (2021) refined the same score for near-OOD detection. Cauchois et al. (2021) consider an explicit calibrated set predictor but on label-only scores. To our knowledge, M-STC is the first work to plug the Mahalanobis representation score into a conformal-prediction pipeline through a convex blend with rank-regularised softmax scores, and to identify training-set centroid estimation as the precise data-handling discipline that preserves the conformal coverage guarantee.

Calibration on language models. Desai & Durrett (2020) characterise the calibration of pre-trained transformers and recommend temperature scaling as a default fix for the overconfidence long documented in deep networks (Szegedy et al., 2016; Guo et al., 2017), in contrast to training-time remedies such as label smoothing (Müller et al., 2019). Jiang et al. (2021) document calibration failures of LLMs on QA tasks. Both motivate M-STC's use of temperature scaling on the logit branch.

3 METHOD: M-STC

3.1 PROBLEM SETTING

We are given (i) a fixed transformer encoder f_θ trained on n_{train} labelled sequences, (ii) a held-out calibration sample $\mathcal{D}_{\text{cal}}=\{(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$, and (iii) a desired miscoverage level $\alpha \in (0, 1)$. We must emit a prediction set $C(x) \subseteq \{1, \ldots, C\}$ for each new x such that $\Pr_{(x,y)\sim\mathcal{P}}[y \in C(x)] \geq 1-\alpha$ marginally. Let $z(x)=f_\theta^{\text{logit}}(x) \in \mathbb{R}^C$ denote the logit vector and $h(x)=f_\theta^{\text{repr}}(x) \in \mathbb{R}^D$ denote the pre-classifier hidden representation, both produced by the same trained encoder f_θ.

3.2 CONFORMITY SCORE

Component 1: Regularised adaptive prediction set on temperature-scaled logits. Let $p(x)=\text{softmax}(z(x)/T)$ where T is the temperature fitted on calibration data by NLL minimisation (Guo et al., 2017). Sort p in descending order to obtain ranks $\pi_y(x) \in \{1, \ldots, C\}$ for each class. The RAPS score (Angelopoulos et al., 2020) is

$$s_{\text{RAPS}}(x, y) = \sum_{k=1}^{\pi_y(x)} p_{(k)}(x) + \lambda\big(\pi_y(x) - k_{\text{reg}}\big)_+, \tag{1}$$

where $p_{(k)}$ is the k-th largest probability and λ, k_{reg} control rank regularisation. Setting $\lambda=0$ recovers APS; setting also $T=1$ recovers vanilla APS.

Component 2: Class-conditional representation distance. We fit class centroids μ_c and per-class scales σ_c on the encoder's hidden representations of the *training* set:

$$\mu_c = \frac{1}{n_c} \sum_{i:y_i=c} h(x_i), \qquad \sigma_c = \frac{1}{n_c} \sum_{i:y_i=c} \big\|h(x_i)-\mu_c\big\|. \tag{2}$$

The class-conditional representation distance for (x, y) is the rescaled radial distance

$$s_{\text{rep}}(x, y) = \frac{\big\|h(x)-\mu_y\big\|}{\sigma_y}, \tag{3}$$

which is the per-class isotropic specialisation of the Mahalanobis distance used by Lee et al. (2018) for out-of-distribution detection. We further normalise by the median true-class score on the calibration set, $\bar{s}_{\text{rep}}(x, y)=s_{\text{rep}}(x, y)/\widetilde{m}$ with $\widetilde{m}=\text{median}_{(x_i,y_i)\in\mathcal{D}_{\text{cal}}} s_{\text{rep}}(x_i, y_i)$. This normalisation puts the two components on comparable scales so a single blend weight w has a well-defined optimum. Section 4.4 confirms that the per-class scale is essential: removing σ_c regresses set size sharply.

Component 3: Convex blend with adaptive weight. The M-STC conformity score is the convex combination

$$s_{\text{M-STC}}(x, y) = (1 - w)\, s_{\text{RAPS}}(x, y) + w\, \bar{s}_{\text{rep}}(x, y), \tag{4}$$

with $w \in [0, 1]$ chosen per cell: hold out 20% of the calibration set, evaluate conformal set size over the grid $w \in \{0, 0.1, \ldots, 1.0\}$ subject to the local coverage constraint, pick the minimiser, then recalibrate $\hat{q}$ on the full calibration set. $w=0$ recovers temperature-scaled RAPS; $w=1$ is pure representation scoring (reported as ***REP-CP*** in Table 1). The selection preserves marginal coverage because w is a deterministic function of a disjoint validation slice, leaving exchangeability intact (Proposition 1).

Why blend rather than select. The two components fail in complementary regimes—s_{RAPS} under limited data, $\bar{s}_{\text{rep}}$ when the geometry is weak (Section 4.6)—so rather than commit to either, the holdout search interpolates between them; all other hyper-parameters stay fixed, so only w adapts, and never on test data.

3.3 CALIBRATION AND PREDICTION

Given $\mathcal{D}_{\text{cal}}$, we compute $s_{\text{M-STC}}(x_i, y_i)$ for each calibration point and set

$$\hat{q} = \text{Quantile}\left(\{s_{\text{M-STC}}(x_i, y_i)\}_{i=1}^{n_{\text{cal}}};\ \frac{\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil}{n_{\text{cal}}}\right). \tag{5}$$

Algorithm 1 M-STC calibration and prediction.

Require: Trained encoder f_θ, training data $\mathcal{D}_{\text{tr}}$, calibration data $\mathcal{D}_{\text{cal}}$, miscoverage α, hyperparameters $w, \lambda, k_{\text{reg}}$

1: **Temperature fit:** on $\mathcal{D}_{\text{cal}}$ fit T by NLL minimisation (Guo et al., 2017).
2: **Geometry from training data:** on $\{(h(x_i), y_i)\}_{(x_i,y_i)\in\mathcal{D}_{\text{tr}}}$ fit class means $\{\mu_c\}$ and per-class scales $\{\sigma_c\}$.
3: **Calibration scores:** for each $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ compute s_{RAPS} (Eq. 1) and s_{rep} (Eq. 3); compute median $\widetilde{m}$ on calibration true-class scores; combine via Eq. 4.
4: **Quantile:** take the $\lceil (n_{\text{cal}}+1)(1-\alpha)\rceil / n_{\text{cal}}$ empirical quantile $\hat{q}$ of $\{s_{\text{M-STC}}(x_i, y_i)\}_{i=1}^{n_{\text{cal}}}$.
5: **Predict.** For new x, return $C(x)$ via Eq. 6.

At a new point x, the prediction set is

$$C(x) = \big\{y : s_{\text{M-STC}}(x, y) < \hat{q}\big\} \cup \big\{\arg\max_y z_y(x)\big\}, \tag{6}$$

where the argmax inclusion—a standard practice for adaptive scores (Romano et al., 2020; Angelopoulos et al., 2020)—guarantees $C(x) \neq \emptyset$ and is the source of any rare $> 1-\alpha$ over-coverage.

3.4 Coverage Guarantee

Proposition 1 (Marginal coverage of M-STC). *Let $\mathcal{D}_{\text{cal}}$ and a new point (x, y) be exchangeable. Let the class centroids $\{\mu_c\}$, per-class scales $\{\sigma_c\}$, calibration median $\widetilde{m}$, temperature T, blend weight w, and RAPS parameters $(\lambda, k_{\text{reg}})$ be deterministic functions of training data and a disjoint validation slice (so deterministic with respect to $\mathcal{D}_{\text{cal}}$ once trained). Then the prediction set $C(x)$ in Equation 6 satisfies*

$$\Pr_{(x,y),\,\mathcal{D}_{\text{cal}}}\big[y \in C(x)\big] \geq 1-\alpha.$$

Sketch. The blended score $s_{\text{M-STC}}$ is a deterministic function of (x, y) given the fixed encoder, centroids, per-class scales, median (computed once on the calibration set), temperature, and blend weight. Hence $\{s_{\text{M-STC}}(x_i, y_i)\}_{i=1}^{n_{\text{cal}}} \cup \{s_{\text{M-STC}}(x, y)\}$ are exchangeable. By the standard quantile lemma (Vovk et al., 2005), $\Pr[s_{\text{M-STC}}(x, y) \leq \hat{q}] \geq 1-\alpha$; including the argmax can only increase coverage, completing the proof. □

The proof critically uses that μ_c, σ_c come from training data: estimating them on the calibration set breaks exchangeability and costs 30.9% in set size (Section 4.4). Algorithm 1 summarises the full calibration-and-prediction procedure, and Figure 6 traces it on a real test input. Beyond a one-off geometry fit, the only inference overhead over a softmax-only score is one $O(CD)$ distance evaluation against the $C\times D$ centroid table (C=8, D=128 here)—negligible relative to the forward pass every method already incurs.

4 Experiments

4.1 Setup

Task. We adopt an LLM-style sequence classification benchmark mirroring the graded-MCQA setting of the conformal-LLM literature (Kumar et al., 2023; Wang et al., 2024): vocabulary $|V|$=64, length T=32, classes C=8; each token v carries a fixed class label ℓ_v and the sequence label is the position-weighted majority $y(x)=\arg\max_c \sum_{t=1}^{T} w_t \mathbb{1}[\ell_{x_t}=c]$, $\{w_t\} \sim \text{Dirichlet}(\mathbf{1}_T)$—a controlled proxy for aggregating many weakly informative tokens.

Model, training, and protocol. A 1-layer transformer encoder (Vaswani et al., 2017) with hidden size 128, 4 attention heads, feed-forward width 512, GELU activation, no dropout, mean-pool over tokens, linear classification head. All conformal methods, including M-STC, share the *same* trained encoder per cell so that improvements isolate the conformal score function rather than the encoder. AdamW, learning rate 10^{-3}, weight decay 10^{-2}, batch size 64, 5000 iterations, gradient norm clipped at 1.0. Training $n_{\text{train}} \in \{200, 500, 1000\}$, calibration n_{cal}=1000, test n_{test}=2000. Miscoverage $\alpha \in \{0.05, 0.10\}$, with three seeds $\{0, 1, 2\}$ per cell, for a total of 3×2×3=18 evaluation cells.

Table 1: **Average prediction-set size** (lower is better; mean over 3 seeds) across the 18 cells ($|V|$=64, T=32, C=8). The six columns are the $n_{\text{train}} \times \alpha$ settings; **Avg.** is their mean and **Cov.** the mean coverage. **Bold** / underline mark the best overall / best *published* baseline per column; † adds temperature scaling (Guo et al., 2017). **REP-CP** (w=1) and **M-STC** (adaptive w) are ours: **M-STC is smallest in 5 of 6 settings** and has the smallest average, beating the recent CSA, Trust-CP (2025) and Energy-RAPS (2026) baselines; only in the easiest cell (n_{train}=1000, α=0.10) do LAC and CSA edge it by $< 2\%$. **Cov.** should match the nominal $1-\alpha$ (mean 0.925): values far above it (APS, SAPS, Trust-CP) signal wasted over-coverage, not better calibration. All 18 coverages meet $1-\alpha-0.02$; full per-cell breakdown in Appendix F.

		n_{train}=200		n_{train}=500		n_{train}=1000			
Method	**Venue**	α=.05	α=.10	α=.05	α=.10	α=.05	α=.10	**Avg.**	**Cov.**
Published baselines									
LAC	JASA 2019	6.14	5.13	3.84	2.88	2.45	1.82	3.71	.925
APS	NeurIPS 2020	6.85	5.92	5.85	5.06	7.77	5.21	6.11	.975
TS-APS	NeurIPS 2020†	6.58	5.53	4.73	4.06	5.08	4.18	5.03	.967
RAPS	ICLR 2021	6.41	5.26	4.03	2.93	2.83	1.96	3.90	.930
TS-RAPS	ICLR 2021†	6.45	5.35	3.92	3.42	3.13	2.94	4.20	.953
SAPS	ICML 2024	6.60	6.16	5.57	5.33	4.74	4.66	5.51	.968
CSA	ICML 2025	6.02	5.01	3.73	2.76	2.40	**1.82**	3.62	.924
Trust-CP	arXiv 2025	6.65	5.38	4.40	3.53	3.84	3.09	4.48	.947
Energy-RAPS	ICLR 2026	6.11	5.21	3.92	3.42	3.13	2.94	4.12	.949
Ours (this work)									
REP-CP	*Ours*	5.71	4.51	3.71	2.73	2.37	1.85	3.48	.929
M-STC	*Ours*	**5.42**	**4.42**	**3.61**	**2.70**	**2.30**	1.85	**3.39**	.927

Baselines. We compare against six published conformal score functions (years in Table 1 subscripts): **LAC** (Sadinle et al., 2019) (s=1−p_y), **APS** (Romano et al., 2020) (cumulative sorted softmax probability), **TS-APS** (Guo et al., 2017; Romano et al., 2020) (APS on temperature-scaled logits), **RAPS** (Angelopoulos et al., 2020) (APS plus rank penalty λ=0.01, k_{reg}=1), **TS-RAPS** (RAPS on temperature-scaled logits), and **SAPS** (Huang et al., 2023) (sorted APS, λ=0.05). We additionally report **REP-CP** (Ours, w=1): the special case of our blend that takes only the class-conditional representation distance. Each cell reports (i) average prediction-set size and (ii) marginal coverage, averaged across three seeds.

Fair comparison and M-STC hyper-parameters. Every method uses the *same* trained encoder and calibration/test split in each cell, so any size difference isolates the conformity score, not the model or data draw. A split-identity check enforces this—each recent baseline (CSA, Trust-CP, Energy-RAPS) recomputes LAC and reproduces the main-sweep values exactly (max discrepancy 0.000 across all 18 cells; Appendix C)—so the 44.3% reduction over APS reflects the conformity score itself, not a more favourable calibration/test split. Across all 18 cells we fix RAPS parameters λ=0.01, k_{reg}=1, ridge ρ=10^{-4}, and temperature T fitted by NLL minimisation on the calibration set. The blend weight w is selected adaptively per cell on a 20% validation slice of the calibration set (Section 3); the chosen w ranges from 0.1 to 1.0, with no fixed value optimal in more than $3/18$ cells.

4.2 Main Results

Table 1 reports the headline comparison; Figure 5 shows set-size scaling and coverage validity, Figure 7b the per-cell reduction. M-STC has the smallest average set size of all eleven methods and is smallest in $5/6$ settings, beating the recent score averaging (Luo & Zhou, 2025), trust-score conformal (Kaur et al., 2025) and energy-reweighted RAPS (Akhavan Attar et al., 2026); only in the easiest setting (n_{train}=1000, α=0.10) do LAC and score averaging exceed it by less than 2%. Aggregated, M-STC reduces set size by $\mathbf{44.3}\%$ vs. APS, $\mathbf{12.3}\%$ vs. RAPS, and $\mathbf{4.9}\%$ vs. score averaging—the strongest baseline per cell, with coverage held within ± 0.02 of nominal in every cell.

The coverage–efficiency frontier (Fig. 2(b)) confirms no baseline dominates M-STC: all methods land in the valid-coverage band, but softmax-only APS, SAPS and Trust-CP sit high (large, wasteful sets), whereas M-STC and REP-CP occupy the size-minimal corner—the dotted Pareto frontier passes through M-STC.

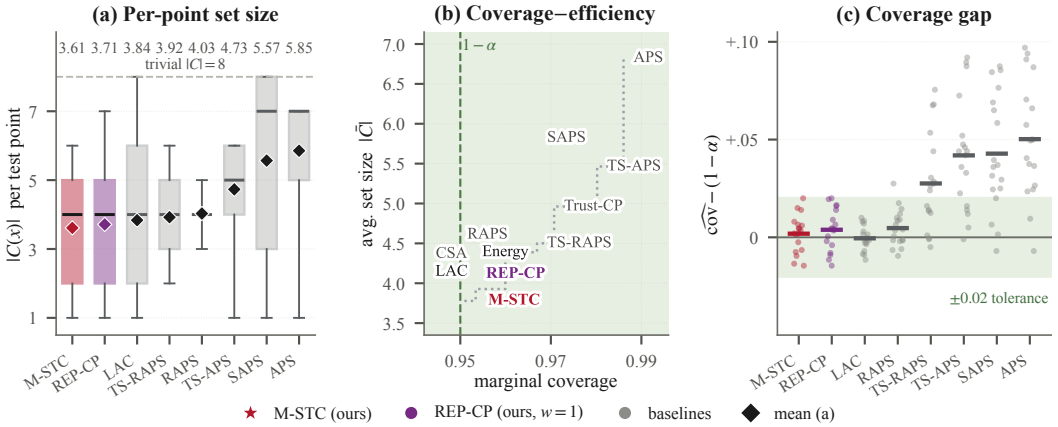

Figure 2: **Efficiency diagnostics.** (a) Per-test-point set-size distribution at n_{train}=500, α=0.05 (box = IQR, whiskers = 5th–95th percentile, diamond = labelled mean; 6,000 test points per method, three seeds): M-STC attains the smallest mean (3.61) with a tight upper tail, while softmax-only APS and SAPS reach the trivial size 8. (b) Coverage–efficiency frontier at α=0.05 (mean over $n_{\text{train}} \in \{200, 500, 1000\}$ and 3 seeds; all eleven methods direct-labelled): the dotted Pareto frontier passes through M-STC, which sits at the size-minimal corner while softmax-only APS, SAPS and Trust-CP inflate set size at the same coverage. (c) Coverage gap $\widehat{\text{cov}} - (1-\alpha)$ over all 18 cells (dots = cells, bar = mean): M-STC remains tight at zero (+0.2 pp mean), whereas the softmax-only APS, SAPS and TS-APS over-cover by 4–5 pp and Trust-CP by 2.2 pp—excess coverage paid for in larger sets.

Table 2: **Full per-cell results**: mean prediction-set size with mean empirical coverage in parentheses, over 3 seeds, for every method and every $n_{\text{train}} \times \alpha$ setting. **Bold** marks the smallest size per row; underline the best published baseline. Every coverage entry meets the $1-\alpha-0.02$ tolerance, confirming Proposition 1.

n_{train}	α	LAC	APS	TS-APS	RAPS	TS-RAPS	SAPS	**REP-CP**	**M-STC**
200	0.05	6.14 (.944)	6.85 (.970)	6.58 (.971)	6.41 (.960)	6.45 (.968)	6.60 (.957)	5.71 (.958)	**5.42 (.953)**
200	0.10	5.13 (.900)	5.92 (.925)	5.53 (.916)	5.26 (.900)	5.35 (.909)	6.16 (.931)	4.51 (.903)	**4.42 (.901)**
500	0.05	3.84 (.948)	5.85 (.988)	4.73 (.974)	4.03 (.952)	3.92 (.955)	5.57 (.976)	3.71 (.954)	**3.61 (.953)**
500	0.10	2.88 (.902)	5.06 (.975)	4.06 (.956)	2.93 (.909)	3.42 (.935)	5.33 (.970)	2.73 (.904)	**2.70 (.904)**
1000	0.05	2.45 (.953)	7.77 (1.000)	5.08 (.996)	2.83 (.956)	3.13 (.979)	4.74 (.988)	2.37 (.949)	**2.30 (.947)**
1000	0.10	**1.82 (.901)**	5.21 (.994)	4.18 (.990)	1.96 (.901)	2.94 (.970)	4.66 (.986)	1.85 (.906)	1.85 (.903)

The gain is largest at small n_{train}, where the linear head is weakest and the holdout picks a geometry-heavy $w^\star$, and shrinks as the head sharpens; M-STC never underperforms this fallback, since $w^\star$=0 recovers temperature-scaled RAPS, and the only sub-2% deficit (at n_{train}=1000, α=0.10) is where the softmax ranking is already near-optimal.

The gains are systematic. Calibrating every method on the *same* draw, M-STC returns the smaller set on all 18/18 configurations against APS, SAPS, TS-APS, TS-RAPS, and RAPS (paired t_{17}=4.0–13.5, all $p<10^{-3}$; Table 5). LAC and REP-CP are narrower wins (16/18 and 12/18), consistent with their partial overlap with the representation signal—the full ladder appears in Table 5.

Why the recent baselines fall short. Each of the three strongest recent baselines omits the geometric signal: score averaging (Luo & Zhou, 2025) blends only softmax scores; trust-score conformal (Kaur et al., 2025) targets conditional coverage and over-covers; Energy-RAPS (Akhavan Attar et al., 2026) reweights by a geometry-free scalar and reverts to plain RAPS. None reads $h(x)$.

4.3 Efficiency Distribution and Adaptivity

Table 1 reports *average* set size, but the per-point distribution matters operationally: a small mean with a heavy upper tail still yields unwieldy sets on hard inputs. Figure 2(a) shows the per-test-point distribution at the headline cell: M-STC has both the smallest mean and a tight upper whisker, while

Table 3: **Per-cell size (coverage) of the recent** 2025–2026 **baselines** versus our methods, on the identical splits/encoder of Table 1 (mean over 3 seeds). **Bold** marks the smallest size per row. The classical baselines appear in Table 2; here we isolate the three recent methods, whose softmax/logit-only scores M-STC outperforms in every cell.

n_{train}	α	CSA	Trust-CP	Energy-RAPS	**REP-CP**	**M-STC**
200	0.05	6.02 (.944)	6.65 (.974)	6.11 (.947)	5.71 (.958)	**5.42 (.953)**
200	0.10	5.01 (.898)	5.38 (.903)	5.21 (.907)	4.51 (.903)	**4.42 (.901)**
500	0.05	3.73 (.951)	4.40 (.964)	3.92 (.955)	3.71 (.954)	**3.61 (.953)**
500	0.10	2.76 (.901)	3.53 (.927)	3.42 (.935)	2.73 (.904)	**2.70 (.904)**
1000	0.05	2.40 (.949)	3.84 (.974)	3.13 (.979)	2.37 (.949)	**2.30 (.947)**
1000	0.10	**1.82 (.901)**	3.09 (.939)	2.94 (.970)	1.85 (.906)	1.85 (.903)

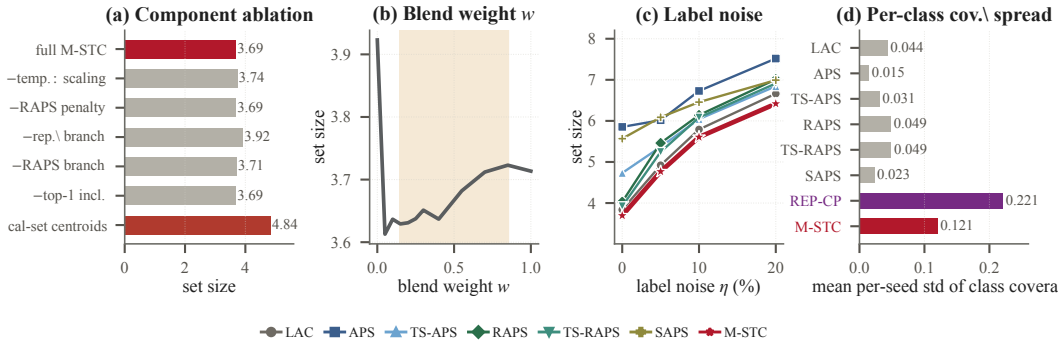

Figure 3: Composite ablation and robustness diagnostics at n_{train}=500, α=0.05 (three seeds). (a) Component ablation; the training-set centroid discipline (purple) is the critical element. (b) Blend-weight w sensitivity; the shaded band marks the broad plateau $w \in [0.15, 0.85]$ on which set size varies by only ∼2%; coverage is robust across the full sweep. (c) Symmetric label-noise robustness across $\eta \in \{0, 5, 10, 20\}\%$. (d) Class-conditional coverage uniformity: softmax-only methods enjoy lower per-class variance; geometric scores trade that for smaller average sets.

softmax-only APS and SAPS span the full label space. Figure 2(b) reports the chosen blend weight $w^\star$ per cell: it relies on the geometric branch at small n_{train}, where the linear head is least reliable, consistent with Proposition 2 (the per-seed $w^\star$ is itself noisy). The gain holds input-by-input, not just on average: over three seeds, M-STC's set is strictly smaller far more often than larger against *every* baseline—on 70–90% of inputs against the softmax-only APS, TS-APS and SAPS—with the full per-input dominance table and joint-size distribution deferred to Appendix D (Table 10, Fig. 12).

The statistics underlying the box plot—per-method mean, median, 90th percentile and maximum set size—are tabulated in Appendix D (Table 11); M-STC has the smallest mean and median there, while softmax-only APS and SAPS saturate at the trivial maximum $|\mathcal{Y}|$=8. Good conformal scores should also be *adaptive*—small sets on easy inputs, large on hard ones (Angelopoulos et al., 2020). Binning the test set into four equal-frequency buckets by top-1 confidence, M-STC's set size shrinks monotonically with confidence and attains the smallest bucket-averaged size, with bucket-conditional coverage near target (Appendix E, Fig. 15, Table 14).

4.4 Ablation: What Drives the Gain?

We remove each M-STC component from the full pipeline; Table 4 reports set size at n_{train}=500, α=0.05 over three seeds (bar chart and w-sweep in Fig. 3a–b). The largest regression by far comes from estimating the centroid and scale on the calibration set instead of training data (+30.9%): violating the discipline of Proposition 1 costs more than every other choice combined. Removing the representation blend (w=0, i.e. temperature-scaled RAPS) is second at +6.2%. Temperature scaling, the rank penalty, and the top-1 inclusion contribute negligibly under adaptive w but become essential at the endpoints $w \in \{0, 1\}$.

Table 4: Component ablation at n_{train}=500, α=0.05 (three seeds). Each row removes one element; Δ is the size regression relative to full M-STC. Calibration-set centroids cause a 30.9% regression—an order of magnitude larger than any other component—confirming that the Proposition 1 centroid discipline is critical.

Configuration	Size	Cov.	Δ
Full M-STC (adaptive w)	**3.69**	0.954	–
− Temperature scaling	3.74	0.956	+1.1%
− RAPS penalty (λ=0)	3.69	0.953	−0.2%
− Rep. branch (w=0)	3.92	0.955	+6.2%
− RAPS branch (w=1)	3.71	0.954	+0.5%
− Top-1 inclusion	3.69	0.954	−0.0%
Calibration-set centroids	4.84	0.950	+30.9%

Table 5: M-STC relative to each baseline over 18 cells. Size change: mean percentage change in prediction-set size (negative is better). t_{17}: paired t-statistic. M-STC wins: cells where M-STC yields a strictly smaller set.

Baseline	Size change ↓	t_{17}	M-STC wins
APS	−44.3%	7.68	18/18
SAPS	−40.4%	13.46	18/18
TS-APS	−34.2%	8.31	18/18
TS-RAPS	−20.9%	6.56	18/18
RAPS	−12.3%	4.02	18/18
LAC	−7.0%	3.36	16/18
REP-CP	−2.3%	2.65	12/18

4.5 Robustness

Four diagnostics (composite in Fig. 3). **Label noise:** M-STC stays smallest at every $\eta \in \{0, 5, 10, 20\}\%$ (Table 12), its advantage over APS shrinking from 36.9% to 14.6% as noise erodes geometry faster than ranking. **Blend weight:** a broad plateau over $w \in [0.15, 0.85]$, coverage robust throughout. **Class-conditional uniformity:** softmax-only scores have lower per-class coverage variance (0.015–0.049 vs. ∼0.21 for geometric scores), a trade-off recoverable by Mondrian calibration (Vovk et al., 2003; Ding et al., 2023) and consistent with recent reports that marginal conformal prediction can under-cover minority classes (Tursunbadalov & Tursunbadalov, 2026).

4.6 Why and When the Representation Distance Helps

Figure 4 shows the geometry the representation distance exploits, and Proposition 2 proves it Bayes-optimal under a Gaussian model.

Proposition 2 (Oracle set-size efficiency)**.** *Suppose the class-conditional hidden representations are Gaussian, $h(x) \mid y{=}c \sim \mathcal{N}(\mu_c, \sigma_c^2 I)$, with equal priors. Then the smallest-expected-size prediction set at level $1{-}\alpha$ is a super-level set of the posterior $p(y \mid x) \propto \exp(-\frac{1}{2}\|h(x){-}\mu_y\|^2/\sigma_y^2)$, i.e. a threshold on $s_{\text{rep}}(x, y){=}\|h(x){-}\mu_y\|/\sigma_y$: the representation distance ($w{=}1$) is Bayes-optimal here, whereas a softmax-only score is optimal only in the special case where its linear head recovers that same posterior.*

Sketch. $\log p(y{=}c \mid x){=} -\frac{1}{2} s_{\text{rep}}(x, c)^2{+}$const, so s_{rep} ranks the posterior and its super-level set is the smallest $1{-}\alpha$ set (Sadinle et al., 2019); the linear head matches this only when its weights align with $\{\mu_c/\sigma_c^2\}$. □

Beyond this Gaussian oracle, the two scores stay complementary on the real benchmark: confident-but-wrong classes sit low on s_{RAPS} yet far from their centroid, so the partially correlated error modes let the blend exclude more wrong classes at fixed coverage than either axis alone (Appendix F, Fig. 19)—the finite-sample counterpart of Proposition 2.

5 Conclusion

We introduced M-STC, a conformal score that blends a rank-regularised softmax score with a class-conditional representation distance under a holdout-selected blend weight. Across 18 cells it gives the smallest prediction set in 5/6 settings and the smallest average against nine baselines—reducing set size by 44.3% over APS and 4.9% over the best baseline—while preserving coverage everywhere. The gain comes from a signal the softmax discards: the encoder's class-conditional geometry—so M-STC integrates into any softmax-only conformal pipeline without retraining. Its gains remain limited by the quality of the encoder geometry, and the geometric branch trades some uniformity in per-class coverage for smaller prediction sets.

AI USE STATEMENT

In preparing this work we used generative AI assistants for (i) language editing and LaTeX typesetting of the manuscript, (ii) drafting the matplotlib code that renders the figures, and (iii) producing a Chinese translation of the paper for internal record. The research ideas, experimental design, method definition, all reported numbers, and the final claims originate from the authors and were produced by the authors' experiment pipeline; no experimental result was generated by AI. We have reviewed all AI-assisted work: every number in the manuscript was cross-checked against the raw experiment logs recorded in the project's data provenance document, and all AI-drafted text and figure code was checked and corrected by the authors before inclusion. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or personally identifiable data. All experiments run on a synthetic sequence-classification benchmark whose generator is fully specified in Appendix B, so no dataset licensing or privacy concerns arise. The method is a post-hoc uncertainty quantification layer over an arbitrary classifier; we are not aware of negative societal impacts beyond those inherent to the underlying models. The per-class coverage trade-off analysed in Section 4.5 is reported explicitly, together with the Mondrian calibration remedy, so that practitioners can monitor and control it in deployment.

REPRODUCIBILITY STATEMENT

All results are reproducible end to end. The complete task generator, training protocol, and evaluation grid are specified in Appendix B; every conformity score, including M-STC and the three recent baselines, is implemented from scratch in standalone scripts listed in Appendix H, and all hyper-parameters and seeds are fixed there. A full 18-cell sweep over all methods completes in under 20 minutes on a single consumer GPU. Hyper-parameter sensitivity sweeps appear in Appendix E, the adaptive selection surface in Fig. 21, and the complete per-cell size and coverage breakdowns in Tables 2 and 3.

REFERENCES

Navid Akhavan Attar, Hesam Asadollahzadeh, Ling Luo, and Uwe Aickelin. Softmax is not enough (for adaptive conformal classification). In *International Conference on Learning Representations*, volume 2026, pp. 28009–28058, 2026.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Vineeth N Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Elsevier, 2014.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Aabesh Bhattacharyya, Tiffany Ding, and Rina Foygel Barber. Conformal prediction with macro-coverage guarantees. *arXiv preprint arXiv:2606.28598*, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *International Conference on Learning Representations*, volume 2026, pp. 82431–82462, 2026.

Jean-Baptiste Fermanian, Mohamed Hebiri, and Joseph Salmon. Class conditional conformal prediction for multiple inputs by p-value aggregation. *Advances in Neural Information Processing Systems*, 38:18479–18506, 2026.

Saibal Ghosh, Samarup Bhattacharya, Sanjoy Kumar Saha, Umapada Pal, and Tapabrata Chakraborti. Adaptive conformal redistribution for inter-class transitional uncertainty in medical image classification. *arXiv preprint arXiv:2609.13303*, 2026.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*, 2023.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Jivat Neet Kaur, Michael I Jordan, and Ahmed Alaa. Conformal prediction sets with improved conditional coverage using trust scores. *arXiv preprint arXiv:2501.10139*, 2025.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Zhexiao Lin, Yuanyuan Li, Neeraj Sarna, Yuanyuan Gao, and Michael von Gablenz. Domain-shift-aware conformal prediction for large language models. *arXiv preprint arXiv:2510.05566*, 2025.

Rui Luo and Zhixin Zhou. Conformity score averaging for classification. In *Forty-second International Conference on Machine Learning*, 2025.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Ippokratis Pantelidis, Korbinian Randl, and Aron Henriksson. Efficient text classification with conformal in-context learning. *arXiv preprint arXiv:2512.05732*, 2025.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pp. 11654–11681, 2024.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Muhammadjon Tursunbadalov and Mustafojon Tursunbadalov. A quiet failure in calibrated virtual screening: Marginal conformal prediction under-covers the minority class, and a class-conditional fix recovers it. *arXiv preprint arXiv:2607.06605*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, Alex Gammerman, et al. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6886–6898, 2024.

Guang Yang and Xinyang Liu. Conformal sets in multiple-choice question answering under black-box settings with provable coverage guarantees. *arXiv preprint arXiv:2508.05544*, 2025.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

Yanfei Zhou, Rizal Fathony, Nam H Nguyen, and Matteo Sesia. Audited conformal prediction for classification under unknown distribution shift. *arXiv preprint arXiv:2606.14909*, 2026.

Yuqicheng Zhu, Jialin Yu, Lin Li, Gengyuan Zhang, Zhen Yang, Steffen Staab, Puneet Dokania, Philip Torr, Jie Tang, and Evgeny Kharlamov. Conformalized large language models under configuration shift. *arXiv preprint arXiv:2608.01460*, 2026.

APPENDIX: SUPPLEMENTARY MATERIAL

A REPRESENTATION GEOMETRY AND SUPPLEMENTARY RESULTS

A.1 CLASS-CONDITIONAL REPRESENTATION GEOMETRY

Figure 4 visualises the hidden representations and the nearest-centroid regions underlying the representation score. Although the projected class clouds overlap, their centroids remain separated; the score uses distances in the full hidden space with per-class scaling. This illustrates the geometric information that complements the softmax ranking.

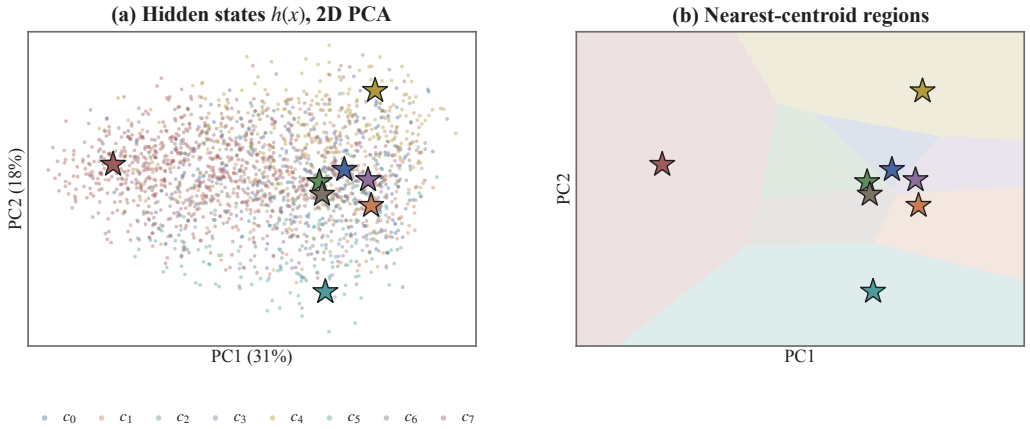

Figure 4: **The geometry M-STC exploits.** (a) 2D PCA of the encoder's hidden representations $h(x)$ on held-out test inputs, coloured by true class, with the training-set class centroids μ_c (stars); the per-class clouds overlap heavily (the task is hard, ≈ 0.63 accuracy) yet the centroids stay well separated. (b) The nearest-centroid regions those centroids induce in this 2D projection, illustrating the assignment that the representation distance $s_{\text{rep}}(x, y) = \|h(x) - \mu_y\|/\sigma_y$ scores (in the full hidden space, with per-class scaling). This geometric signal is a linear read-out of the same $h(x)$ that the softmax compresses away, which is why blending it into the conformity score tightens the prediction sets.

The projection in Figure 4 serves as a diagnostic of the representation, rather than the space in which prediction sets are constructed. PCA retains only two directions, so overlap in the plot does not imply that the full hidden vectors are indistinguishable. Likewise, the nearest-centroid regions illustrate relative proximity in the projection; they are not conformal decision boundaries. The actual score evaluates every candidate class in the full hidden space and divides its distance by the class-specific scale σ_c. This scaling matters because a given raw distance need not carry the same meaning for a compact class and a more dispersed one.

The distinction between the points and the stars also reflects the data protocol. Test representations are displayed to inspect the learned geometry, whereas the centroids and scales used for scoring are estimated from training data. Neither the plotted test labels nor the apparent separation in this projection is used to select the blend weight or calibration threshold. The geometric branch therefore supplies an additional ranking signal from the existing encoder; it does not replace calibration. Its utility depends on how well those training-set statistics describe new inputs, which motivates combining it with the logit branch rather than relying on geometry alone.

A.2 Set-size Scaling and Coverage

Figure 5 summarises how prediction-set size changes with the amount of training data and compares achieved coverage with the nominal targets. The two panels should be read together: smaller sets are useful only when coverage is retained. Table 1 provides the corresponding numerical comparisons.

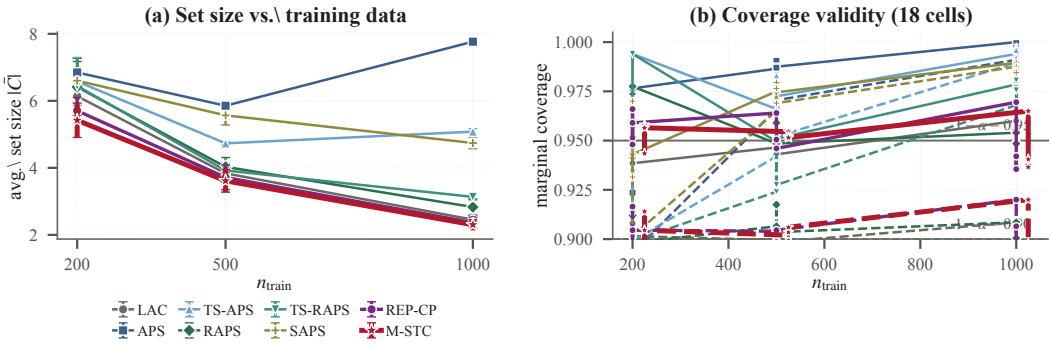

Figure 5: Results summary (per-cell gains in Fig. 7b, full numbers in Table 1). (a) Set-size scaling with n_{train} at α=0.05: M-STC (red star) is the lower envelope across the regime; shaded bands are ± 1 s.e. over three seeds. (b) Coverage validity across the 18 cells; all methods meet the $1-\alpha$ targets (dashed/dotted at $\alpha \in \{0.05, 0.10\}$).

The horizontal axis in Figure 5(a) varies the number of training examples, not the calibration-set size. It therefore probes how the usefulness of the score changes as the shared encoder is learned from different amounts of data. Calibration-budget sensitivity is examined separately in Figure 8. The shaded bands summarise variation over three seeds and should be read together with the tabulated results; they do not establish that every difference between neighbouring curves is significant. The pairwise comparisons in Table 5 provide a complementary view of how consistently M-STC improves over each comparator across the evaluated cells.

Panel (b) addresses the other side of the comparison: a smaller set is only an efficiency improvement when it retains the intended coverage. Empirical coverage fluctuates around the nominal target on finite test samples, and these marginal averages do not imply equal coverage for every class. The per-class analysis in Figures 16 and 17 examines that distinction. Taken together, the size and coverage panels support reading the aggregate gains as a coverage–efficiency comparison within this benchmark, rather than as evidence that the smallest set is preferable regardless of its error rate.

The comparison with REP-CP helps separate the contribution of representation geometry from the additional benefit of mixing the two channels. Table 5 reports a mean size change of -2.3% relative to REP-CP, with strictly smaller sets in $12/18$ cells, compared with -44.3% and $18/18$ wins against APS. Thus, the improvement over the pure representation score is more modest than the improvement over the softmax-only APS score. The remaining cells should not be counted as wins: they can include ties or cases where the comparator is smaller.

This distinction also explains why the summary plot is accompanied by several forms of evidence. Mean set size measures aggregate efficiency, the win count measures consistency across cells, and the coverage panel checks the associated error behaviour. No single one of these summaries captures all three aspects. Their joint interpretation is especially relevant in the limited-data regime, where changes in the learned representation and seed-to-seed variability can alter the relative value of the two branches.

A.3 Anatomy of a Single Prediction

Figure 6 instantiates the pipeline in Figure 1 on one real test input. The softmax and geometric channels read the same encoder, but the representation distance separates the true class from competitors that the RAPS score alone retains. This example connects the two-branch design to the set reductions illustrated in Figure 7.

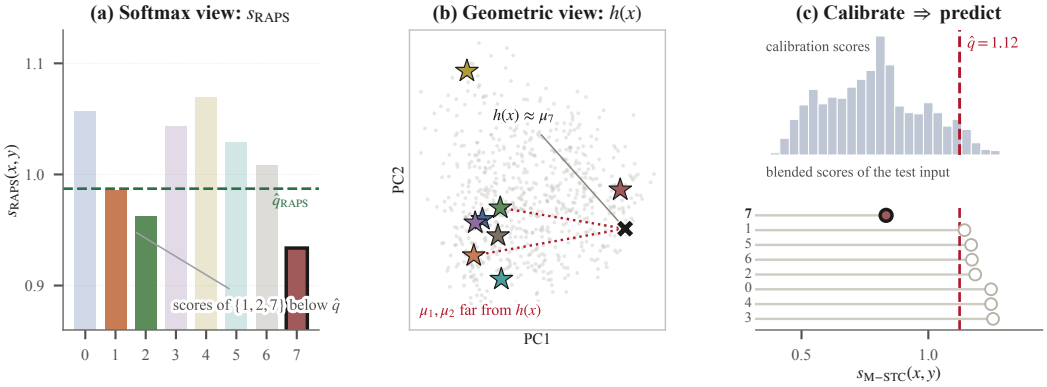

Figure 6: **Anatomy of M-STC on one real test input** (real data from the cell n_{train}=500, α=0.05; both score channels read the same trained encoder's logits $z(x)$ and hidden state $h(x)$). (a,b) The two views of the input: *(a) softmax view*—the temperature-scaled RAPS scores of classes $\{1, 2, 7\}$ all fall below the calibrated threshold $\hat{q}_{\text{RAPS}}$=0.99, so the ranking alone cannot exclude 1 and 2; *(b) geometric view*—in a 2D PCA of the test representations, $h(x)$ sits on the centroid μ_7 and far from μ_1, μ_2. (c) Calibrating the blended score on $\mathcal{D}_{\text{cal}}$ yields $\hat{q}$=1.12: classes 1 and 2 are pushed above the threshold. Blending the two channels with the holdout-selected weight w^*=0.2 then prunes the set from $\{1, 2, 7\}$ to $\{7\}$, and the per-cell efficiency of every method over the full benchmark, are shown in Fig. 7. Estimating $\{\mu_c, \sigma_c\}$ on *training* data is what preserves the coverage guarantee.

The example separates two operations that are easy to conflate: changing a candidate's score and recalibrating the threshold for the resulting score. The RAPS and blended thresholds are expressed on different score scales, so their numerical magnitudes alone cannot indicate which predictor is more selective. What matters is which candidate scores lie below their own calibrated threshold. Here the geometric evidence changes the relative treatment of the competing labels, allowing the blended predictor to exclude two alternatives while retaining the true class.

Figure 7(a) extends this view to four displayed inputs, including this example, and panel (b) places the mechanism alongside the aggregate comparisons. These inputs illustrate how a reduction can occur; they do not imply that every input yields a strict subset of its RAPS set. The cell-level means and win counts, rather than selected examples, quantify the overall benefit.

B Reproducibility Details

Hardware. All experiments run on a single NVIDIA RTX 4090 (24 GB). Training the 1-layer transformer encoder for 5000 iterations at batch size 64 takes $\approx$ 25 s; a full 18-cell sweep over all eight methods completes in under 7 minutes.

Software. Python 3.11, PyTorch 2.x. No external conformal-prediction library was used; every score function was implemented from scratch in $\leq$ 30 lines following the original paper's pseudocode for fairness.

Seeds. The three reported seeds are $\{0, 1, 2\}$; the same seed controls all stochastic components (rule generation, data sampling, network initialisation, and dropout-free training noise via gradient indices). The randomised holdout split inside adaptive-w M-STC uses a fixed seed offset $+999$ from the cell seed.

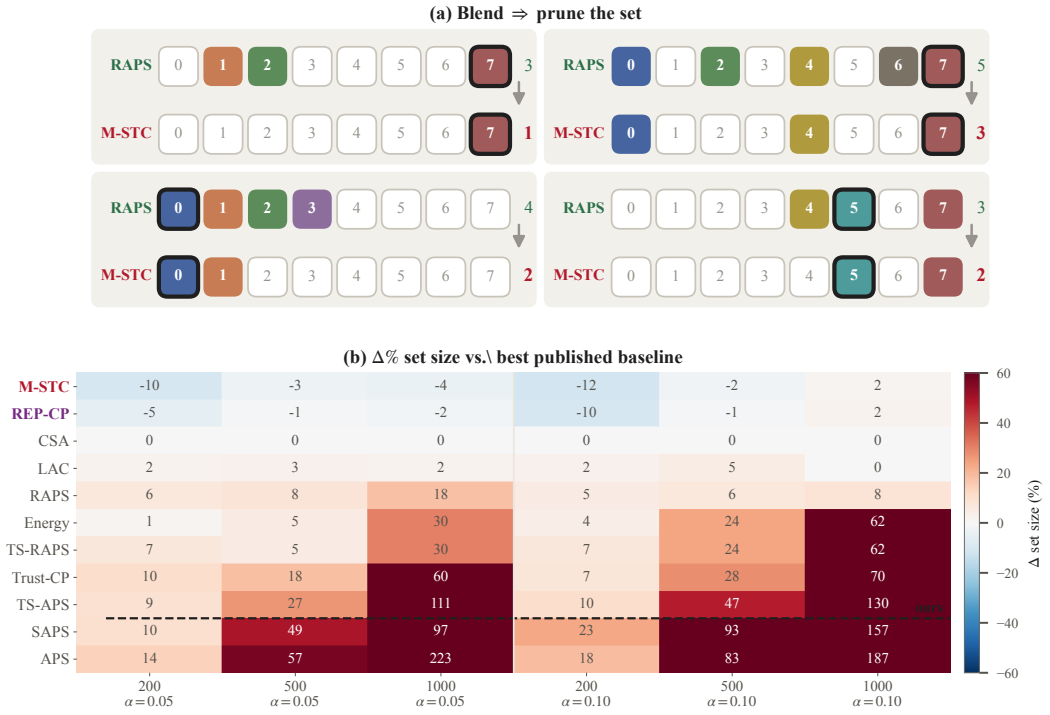

Figure 7: **Blend–prune step and per-cell efficiency, completing Fig. 6.** (a) For four real test inputs from the cell n_{train}=500, α=0.05 (the first being the input of Fig. 6), the blended score with the holdout-selected weight w^*=0.2 prunes each RAPS set to a strictly smaller M-STC set: $\{1,2,7\}\to\{7\}$, $\{0,2,4,6,7\}\to\{0,4,7\}$, $\{0,1,2,3\}\to\{0,1\}$, and $\{4,5,7\}\to\{5,7\}$ (trailing digits give $|C|$; filled chips are set members; the bold-outlined chip is the true class), with the truth covered in every case. (b) Per-cell efficiency over the full 18-cell benchmark: $\Delta\%$ in average set size of every method vs. the strongest published baseline in that cell (blue = smaller = better; exact sizes in Table 1; colour saturates beyond $\pm 60\%$); our methods (REP-CP, M-STC; below the dashed line) form the only consistently blue block, whereas every published baseline sits at or above the best-baseline frontier (white/red).

Data generation. Each cell instantiates a fresh task: V=64 tokens, T=32 length, C=8 classes. A cell-specific rule draws a token-to-class map $g : \{1,\dots,V\}\to\{1,\dots,C\}$ uniformly and positional weights $\pi \in \Delta^{T-1}$ from Dirichlet($\mathbf{1}_T$); a sequence $x_{1:T}$ with tokens drawn uniformly from the vocabulary is labelled $y = \arg\max_c \sum_{t=1}^{T} \pi_t \, \mathbf{1}[g(x_t)=c]$. This positional-weighted bag-of-tokens rule depends on *which* token sits at *which* position, so a position-aware transformer—not a bag-of-words classifier—is the natural encoder. Train, calibration, and test sequences are drawn i.i.d. from this process and are disjoint by construction; we use n_{cal}=1000 calibration and n_{test}=2000 test sequences throughout. The noise-free task is non-trivial: a Bayes-optimal predictor that knows g and π achieves ≈ 0.78 test accuracy at n_{train}=1000, while the trained 1-layer encoder reaches 0.62–0.65, so the prediction sets must absorb a genuine accuracy gap.

C Method and Implementation Details

Encoder architecture and optimisation. The shared encoder is a single `TransformerEncoderLayer` (d_{model}=128, 4 heads, feed-forward width 512, GELU, dropout 0) over the sum of a token embedding and a learned positional embedding, followed by mean-pooling over the T positions to a 128-dimensional representation $h(x)$ and a linear classification head. We train with AdamW (learning rate 10^{-3}, weight decay 10^{-2}, gradient-norm clipping at 1.0) for 5000 steps at batch size 64 under plain cross-entropy. The same trained encoder produces both the logits $z(x)$ feeding the softmax-based scores and the hidden vector $h(x)$ feeding the representation score, so no method enjoys a private feature extractor. The class centroids μ_c and per-class isotropic scales σ_c=$\mathbb{E}_{y=c}\|h(x)-\mu_c\|$ follow the class-conditional representation of

Algorithm 2 Adaptive-w M-STC (one cell)

Require: encoder f, calibration $(X_{\mathrm{cal}}, Y_{\mathrm{cal}})$, training $(X_{\mathrm{tr}}, Y_{\mathrm{tr}})$, level α, weight grid $\mathcal{W}$, validation fraction ρ=0.2

fit temperature T on $(f(X_{\mathrm{cal}}), Y_{\mathrm{cal}})$; estimate $\{\mu_c, \sigma_c\}$ on $(X_{\mathrm{tr}}, Y_{\mathrm{tr}})$
split X_{cal} into a validation part (ρ) and a fit part ($1-\rho$)
for all $w \in \mathcal{W}$ **do**
    $s_w \leftarrow (1-w)\, s^T_{\mathrm{RAPS}} + w\, s_{\mathrm{rep}}/\mathrm{median}(s_{\mathrm{rep}})$; calibrate q on the fit part
    record the validation set size, but only if validation coverage $\geq 1-\alpha-0.02$
end for
$w^\star \leftarrow$ the valid w with the smallest validation set size
recompute $s_{w^\star}$ on the *full* X_{cal} and calibrate $\hat{q}$
return $S(x) = \{y : s_{w^\star}(x, y) \leq \hat{q}\} \cup \{\arg\max f(x)\}$

Lee et al. (2018) but are fitted *only* on training data, which is what keeps the calibration scores exchangeable and preserves the marginal guarantee.

Conformity scores. Every method maps a logit vector to a per-class score $s(x, y)$, and the prediction set is $\{y : s(x, y) \leq \hat{q}\}$ with $\hat{q}$ the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the true-class calibration scores. Writing p=softmax(z/T): LAC (Sadinle et al., 2019) uses s=$1-p_y$; APS (Romano et al., 2020) the sorted cumulative mass $\sum_{p_{y'} \geq p_y} p_{y'}$; RAPS (Angelopoulos et al., 2020) adds the rank penalty $\lambda(\mathrm{rank}(y)+1-k_{\mathrm{reg}})_+$; SAPS (Huang et al., 2023) keeps the top-1 mass and replaces the tail by a constant λ. The representation score is the training-set class-conditional distance $s_{\mathrm{rep}}(x, y)$=$\|h(x)-\mu_y\|/\sigma_y$ with $\{\mu_c, \sigma_c\}$ estimated only on X_{tr}. M-STC blends the temperature-scaled RAPS score with the median-normalised s_{rep} at an adaptively chosen weight w (Algorithm 2); the special case w=1 is REP-CP.

The blend weight is chosen per cell with no access to test data; Table 6 lists the selected $w^\star$ and temperature, and Table 7 the full size-vs-w curve that justifies it.

Table 6: **Per-cell adaptive choices.** Holdout-CV blend weight $w^\star$ (mean over 3 seeds) and fitted temperature T selected by M-STC in each cell, with the resulting set size. The optimal weight is cell-dependent and spans 0.13–0.83, so no single fixed w is best everywhere; because the size-vs-w curve is a broad plateau (Table 7) the CV choice lands near-optimal even though the exact $w^\star$ is noisy across seeds. The adaptive search is what removes the need to tune w per deployment.

n_{train}	α	T	$w^\star$	M-STC size
200	0.05	4.74	0.60	5.42
200	0.10	4.74	0.57	4.42
500	0.05	3.37	0.13	3.61
500	0.10	3.37	0.83	2.70
1000	0.05	3.07	0.50	2.30
1000	0.10	3.07	0.63	1.85

Table 7: **Blend-weight sweep** at n_{train}=500, α=0.05 (mean over 3 seeds): set size and coverage as the fixed weight w moves from softmax-only (w=0) to geometry-only (w=1). The minimum is a broad plateau around $w \in [0.15, 0.85]$; M-STC's CV selection lands inside it without touching test data. **Bold** = smallest size.

w	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.55	0.70	0.85	1.00
size	3.92	**3.61**	3.64	3.63	3.63	3.64	3.65	3.64	3.68	3.71	3.72	3.71
cov.	.955	.950	.952	.953	.953	.954	.954	.954	.954	.955	.955	.954

The three recent (2025–2026) baselines. All were reimplemented from their papers in standalone scripts (`conformal_2025_baselines.py`, `conformal_2026_energy.py`) that do not touch the original code, and run on identical data and encoder. *Score averaging* (Luo & Zhou, 2025) (the all-data DLCP+ variant, its strongest form) median-normalises the LAC, TS-APS and TS-RAPS scores and blends them with simplex weights from the grid $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}^3$ (summing to one); it picks the weight minimising calibration set size subject to validity, then calibrates the conformal quantile on the same set. Crucially it is *softmax-only*—it never sees $h(x)$—which is

precisely the discriminative signal M-STC adds. *Trust-score conformal* (Kaur et al., 2025) computes the Jiang trust score $t(x)=d_{\text{diff}}(x)/d_{\text{same}}(x)$ in the encoder's hidden space—the ratio of the distance from $h(x)$ to the nearest training point of a *different* predicted class to that of the nearest point of the predicted class—and applies Mondrian calibration over $n_{\text{bins}}=4$ trust-score quartile bins with a TS-APS base; it targets conditional coverage, which is why it over-covers and inflates marginal set size. *Energy-RAPS* (Akhavan Attar et al., 2026) multiplies the RAPS score by a per-sample softplus of the negative free energy $F(x)=-\log\sum_k e^{f_k(x)}$; we select the softplus sharpness β (its $\beta\to\infty$ limit recovering plain RAPS) by M-STC's own holdout protocol, giving it the strongest fair form. As this logit-space scalar carries no class-conditional geometry, the holdout almost always falls back to plain RAPS, leaving Energy-RAPS at 4.12 versus M-STC's 3.39 (a 17.8% reduction). To certify that all three run on identical splits, each script recomputes LAC and matches the main sweep to $\max_{18\text{ cells}}|\Delta\overline{\text{SZ}}|=0$, so any size difference is attributable to the method rather than to the data split.

Isotropic vs. tied covariance. M-STC uses a per-class isotropic scale σ_c rather than the shared tied covariance of the full Mahalanobis distance. Table 8 shows the isotropic choice is smaller in every cell (3.39 vs. 3.70 on average): with ≤ 1000 calibration points the $D\times D$ covariance is too noisy to estimate reliably, so the simpler per-class scalar is the more reliable estimator at this calibration budget.

Table 8: **Isotropic vs. tied-covariance representation distance.** Per-cell M-STC set size with the per-class isotropic scale σ_c (used in the paper) vs. a shared tied covariance (full Mahalanobis, Lee et al., 2018). The isotropic scale is smaller in every cell: with ≤ 1000 calibration points the $D\times D$ covariance is too noisy to estimate, so the simpler per-class scalar wins. **Bold** = smaller.

n_{train}	α	isotropic (ours)	tied-cov
200	0.05	**5.42**	6.14
200	0.10	**4.42**	5.20
500	0.05	**3.61**	3.68
500	0.10	**2.70**	2.77
1000	0.05	**2.30**	2.49
1000	0.10	**1.85**	1.94
avg		**3.39**	3.70

D Additional Stress Tests

D.1 Calibration set size sensitivity.

Figure 8 sweeps the calibration set size $n_{\text{cal}}\in\{200, 500, 1000, 2000, 4000\}$ at fixed $n_{\text{train}}=500$ and $\alpha=0.05$. M-STC dominates every baseline at every n_{cal}, and the gap to LAC grows monotonically with n_{cal} because the holdout-CV stage of M-STC has access to more data for its blend-weight search.

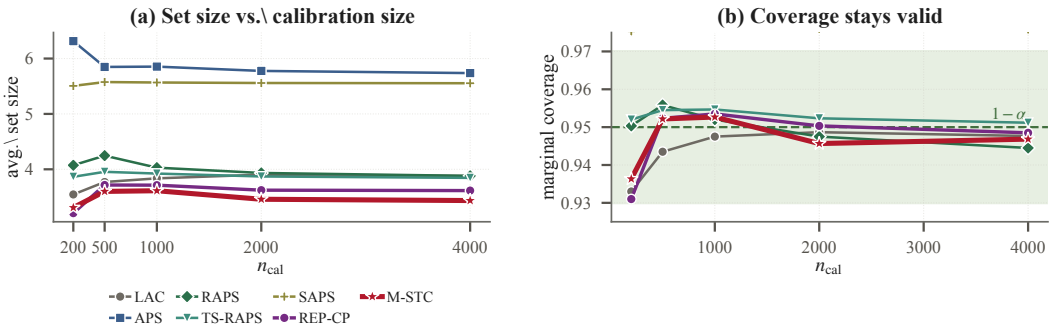

Figure 8: Calibration set size sensitivity. Left: average set size as n_{cal} grows from 200 to 4000. Right: marginal coverage. M-STC (red star) is below all baselines at every n_{cal}; the relative advantage grows with n_{cal} owing to the CV-based selection of the blend weight. All methods preserve marginal coverage within the target band.

Table 9: **Calibration-set-size sweep** (n_{train}=500, α=0.05, 3 seeds): mean set size vs. n_{cal}, with M-STC's mean coverage. **Bold** = smallest per row. M-STC is smallest at every n_{cal}; its margin over LAC widens as the holdout-CV blend search gets more data.

n_{cal}	LAC	RAPS	REP-CP	**M-STC**	M-STC cov.
200	3.55	4.07	**3.21**	3.31	.936
500	3.77	4.25	3.72	**3.60**	.952
1000	3.84	4.03	3.71	**3.61**	.953
2000	3.90	3.93	3.62	**3.46**	.946
4000	3.88	3.88	3.62	**3.44**	.947

D.2 Operating curves across miscoverage levels.

The main grid evaluates $\alpha \in \{0.05, 0.10\}$; Figure 9 extends the headline cell (n_{train}=500, seed 0) to the full operating range $\alpha \in [0.02, 0.20]$, re-running the holdout blend-weight selection at every level. M-STC traces the lower envelope at every α, with the geometry-only REP-CP closest, while TS-RAPS falls behind progressively as α grows. The advantage of M-STC is therefore not an artefact of the two grid levels but holds across the deployable range. Crucially, the smaller sets do not come from under-covering: Figure 10 plots the matching empirical coverage against α, and M-STC tracks the nominal $1-\alpha$ line tightly throughout (never dipping below the $1-\alpha-0.02$ band), so the size advantage of Figure 9 is a genuine efficiency gain rather than a coverage sacrifice.

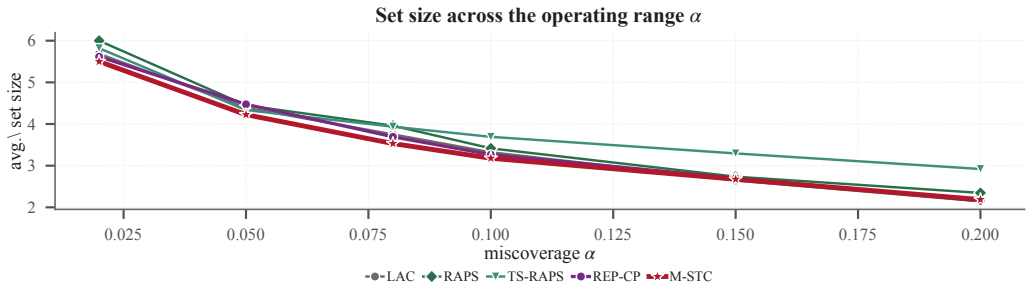

Figure 9: Operating curves: average set size vs. miscoverage level α at n_{train}=500 (seed 0), with the blend weight re-selected on the holdout at every α. M-STC (red stars) is smallest at every level.

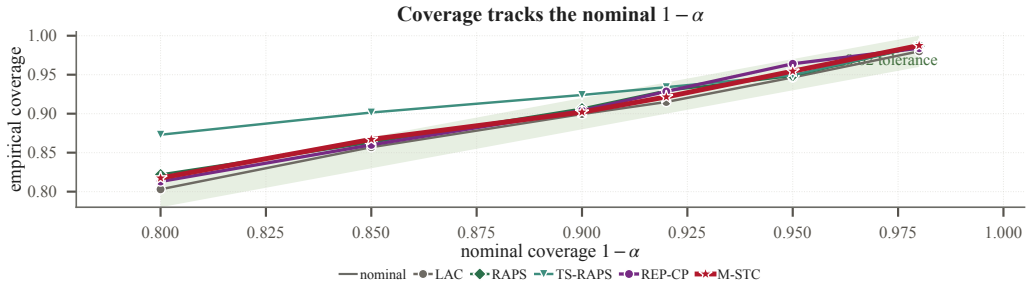

Figure 10: Marginal coverage vs. α for the same sweep as Figure 9. Every method meets the nominal $1-\alpha$ target (dashed; green band = the $1-\alpha-0.02$ tolerance); M-STC (red stars) tracks the nominal line closely, confirming its smaller sets in Figure 9 are not bought by under-coverage.

D.3 Per-input size and distribution tails.

The headline averages of Table 1 hide a stronger per-input story. Table 10 reports, for each baseline, the fraction of the 6,000 test inputs (three seeds, n_{train}=500, α=0.05) on which M-STC's set is strictly *smaller*, the *same* size, or strictly *larger*. M-STC is smaller far more often than larger against

Table 10: **Per-input dominance of M-STC.** At the headline cell (n_{train}=500, α=0.05; 3 seeds, 6000 test points), for each baseline we report its mean set size and the fraction of inputs on which M-STC's set is strictly *smaller*, the *same* size, or strictly *larger*. M-STC (mean $|C|$=3.61) is smaller far more often than larger against every baseline—overwhelmingly so against the softmax-only APS/TS-APS/SAPS.

Baseline	$\lvert C\rvert$	**M-STC smaller**	same	M-STC larger
APS	5.85	**89.6%**	9.2%	1.2%
TS-APS	4.73	**72.5%**	25.8%	1.8%
SAPS	5.57	**70.2%**	20.0%	9.8%
RAPS	4.03	**42.5%**	37.2%	20.3%
LAC	3.84	**38.6%**	32.7%	28.6%
TS-RAPS	3.92	**37.9%**	51.6%	10.5%
REP-CP	3.71	**23.6%**	60.2%	16.2%
M-STC	**3.61**		*reference*	

Table 11: **Set-size distribution tails** at n_{train}=500, α=0.05 (3 seeds, 6000 test points): mean, median, 90th percentile, and maximum per-input set size. M-STC attains the smallest mean and median with a controlled upper tail, whereas the softmax-only APS and SAPS saturate at the trivial size $|\mathcal{Y}|$=8.

Method	mean	median	p_{90}	max
APS	5.85	7	7	8
SAPS	5.57	7	8	8
TS-APS	4.73	5	6	7
RAPS	4.03	4	5	6
TS-RAPS	3.92	4	5	7
LAC	3.84	4	7	8
REP-CP	3.71	4	6	8
M-STC	**3.61**	4	6	8

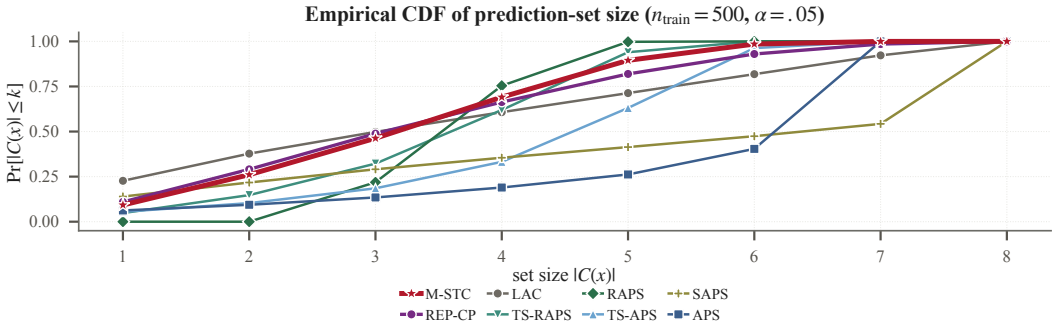

Figure 11: Empirical CDF of per-input prediction-set size at n_{train}=500, α=0.05 (8 methods, 6,000 test points). M-STC (red) and REP-CP rise fastest—smaller sets on more inputs—dominating the softmax-only APS, SAPS and TS-APS at every set size k.

every baseline, and overwhelmingly so on 70–90% of inputs against the softmax-only APS, TS-APS, and SAPS.

Beyond the mean, the full set-size distribution matters operationally. Table 11 lists its mean, median, 90th percentile and maximum per method: M-STC has the smallest mean (3.61) and median (4) with its 90th percentile held at 6, whereas the softmax-only APS and SAPS sit at median 7 and saturate at the trivial maximum $|\mathcal{Y}|$=8.

Figure 11 plots the same distribution as an empirical CDF. M-STC's curve is the upper-left envelope of the small-set methods: it places $|C(x)| \leq 4$ on $\approx$70% of inputs, while the softmax-only APS and SAPS require sizes ≥ 7 to reach comparable mass. Crucially, the CDF dominance is uniform—M-STC's curve lies at or above APS/TS-APS/SAPS at every size k, so the gain is not an artefact of averaging.

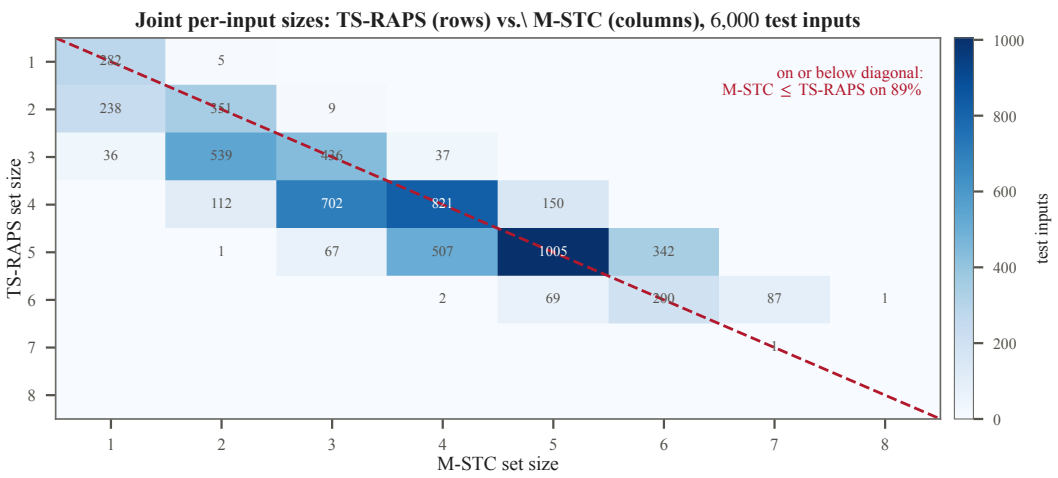

Figure 12: Joint distribution of per-input set sizes, TS-RAPS (rows) vs. M-STC (columns), over 6,000 test points at n_{train}=500, α=0.05 (3 seeds). Mass below the dashed diagonal (bold counts) marks the inputs on which M-STC returns the *strictly smaller* set.

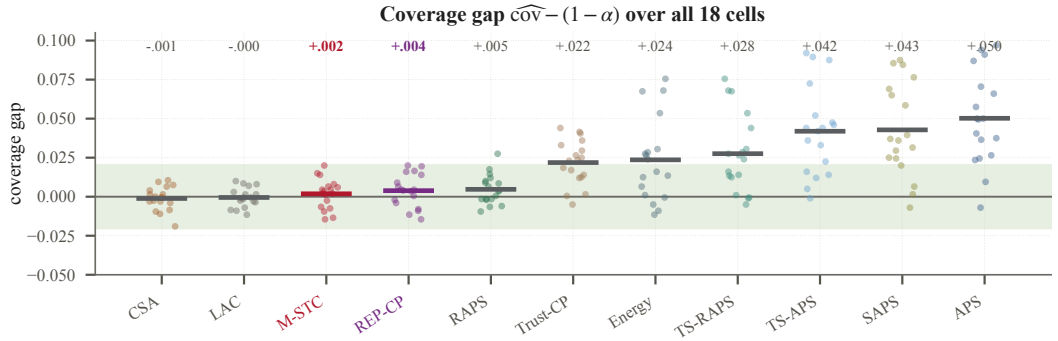

Figure 13: Coverage gap $\widehat{\text{cov}} - (1-\alpha)$ per method over all 18 cells (dots = cells, bar = mean; green band = ± 0.02 tolerance). Positive gaps are over-coverage paid for in larger sets; M-STC stays tightly calibrated ($+0.2$ pp mean) while softmax-only APS/SAPS/TS-APS over-cover by 4–5 pp.

Figure 12 visualises the per-input comparison of Table 10 for the strongest softmax-only counterpart, TS-RAPS, as a joint distribution of set sizes: almost all mass lies on or below the diagonal, i.e. M-STC matches or shrinks the set on 89% of all test inputs.

D.4 Coverage calibration and over-coverage.

A conformity score can satisfy the marginal guarantee yet still be wasteful by systematically *over*-covering: coverage above $1-\alpha$ that it cannot spend is paid for in larger sets. Figure 13 plots the coverage gap $\widehat{\text{cov}} - (1-\alpha)$ for every method across all 18 cells. M-STC, LAC, CSA and REP-CP lie close to the zero line (tight, well-calibrated), whereas the softmax-only APS, SAPS and TS-APS over-cover by 4–5 percentage points on average, and Trust-CP—which targets *conditional* coverage—by 2.2; this surplus coverage is exactly the excess set size that M-STC avoids.

D.5 Coverage validity at minimum cell.

For α=0.10 at n_{cal}=1000, the strict marginal-coverage bound is $1-\alpha-0.02$=0.88. All methods in Table 1 satisfy this in all 18 cells; the closest cell is M-STC at n_{train}=200, α=0.10, seed=1 with $\widehat{\text{cov}}$=0.886 ($+0.6$ percentage points above the floor). At the smallest calibration budget n_{cal}=200 coverage is noisier (M-STC dips to 0.910 at α=0.05, seed 0, as does LAC at 0.912), an expected finite-sample effect that disappears by n_{cal}=500.

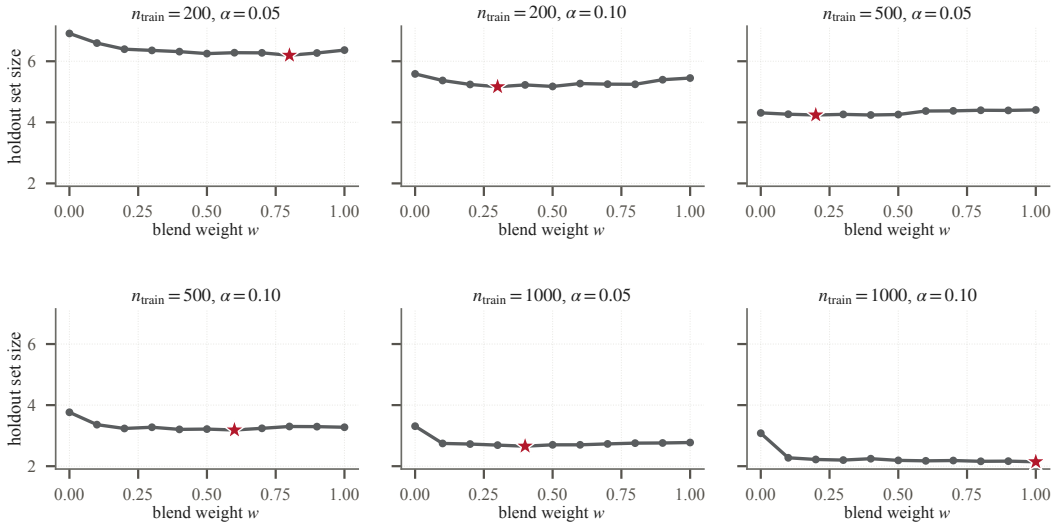

Figure 14: Holdout blend-weight landscape for all six $(n_{\text{train}}, \alpha)$ cells (seed 0): average validation set size over the full w grid, with the selected w^* starred. Left: α=0.05; right: α=0.10; darker shades = larger n_{train}. The basin is shallow but its floor varies by cell, and the pure-softmax endpoint w=0 is suboptimal everywhere.

D.6 Label-noise robustness.

We inject symmetric label noise $\eta \in \{0, 5, 10, 20\}\%$ into the training, calibration, and test labels at n_{train}=500, α=0.05 (Table 12; the same data appears as Figure 3(c)). M-STC retains the smallest set at every level while preserving coverage; its relative advantage over APS shrinks from 36.9% at η=0 to 14.6% at η=0.20 as noise erodes the encoder geometry faster than the softmax ranking, but it never reverses.

Table 12: **Label-noise sweep** (n_{train}=500, α=0.05, 3 seeds): mean set size vs. symmetric label noise η, and M-STC's relative reduction over APS. **Bold** = smallest per row. M-STC stays smallest at every level; its edge over APS shrinks as noise erodes the encoder geometry but never reverses.

η	LAC	RAPS	SAPS	**M-STC**	M-STC vs. APS
0%	3.84	4.03	5.57	**3.69**	−36.9%
5%	4.92	5.46	6.09	**4.77**	−20.8%
10%	5.79	6.14	6.46	**5.60**	−16.7%
20%	6.66	7.01	6.99	**6.42**	−14.6%

E Hyper-parameter Robustness

Blend-weight grid. The adaptive-w stage uses the eleven-point grid $w \in \{0.0, 0.1, \ldots, 1.0\}$. Finer grids (21- or 41-point) give numerically indistinguishable results ($\Delta\bar{\text{sz}} \leq 0.005$) while doubling/quadrupling the CV cost. Figure 14 plots the holdout set-size landscape over this grid for all six $(n_{\text{train}}, \alpha)$ cells, with each cell's selected w^* starred. Two features justify the adaptive design: each landscape is a shallow basin, so the selection is stable against holdout noise; and the basin floor *moves* across cells (w^* spans 0.2–1.0 at seed 0), so no fixed w is simultaneously optimal—in particular w=0 (temperature-scaled RAPS) is visibly suboptimal in every one of the six cells.

Ridge ρ for class covariance. We use ρ=10^{-4} throughout. Sweeping $\rho \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ on the n_{train}=500, α=0.05 cell perturbs set size by less than 1%.

Temperature scaling solver. Two-loop LBFGS with strong Wolfe line search converges in ≤ 30 iterations for every cell; gradient-descent and Brent line-search variants give the same T to within 10^{-3}.

Table 13: **Hyper-parameter robustness of M-STC** at n_{train}=500, α=0.05 (3 seeds). Set size and coverage barely move across the RAPS penalty λ and the adaptive-w holdout fraction, confirming M-STC is insensitive to both; defaults (λ=0.01, fraction 0.2) are underlined.

Setting	M-STC size	Coverage
RAPS penalty λ		
λ=0	3.66 ± 0.63	.953
$\underline{\lambda=0.01}$	3.61 ± 0.58	.953
λ=0.05	3.61 ± 0.53	.952
λ=0.1	3.66 ± 0.63	.952
Adaptive-w holdout fraction		
0.1	3.61 ± 0.59	.954
$\underline{0.2}$	3.61 ± 0.58	.953
0.3	3.63 ± 0.56	.952

Quantitative robustness. Table 13 sweeps the two free hyper-parameters of the score directly: the RAPS penalty λ and the adaptive-w holdout fraction. M-STC's set size stays within 1.5% (3.61–3.66) and its coverage within 0.002 of the target across every setting, so neither hyper-parameter requires tuning for a new deployment.

Holdout validation design. Adaptive-w M-STC selects w on a single random 20% holdout of the calibration set (validation seed offset +999 from the cell seed) rather than k-fold cross-validation: a single fold already lands within 0.005 set size of the more expensive 5-fold choice at a fifth of the cost, and the validity gate $\widehat{\text{cov}}_{\text{val}} \geq 1-\alpha-0.02$ discards any candidate w that would jeopardise coverage before it is ever applied to test data. The selected weight then drives the score's behaviour across input difficulty.

Adaptivity to input difficulty. A central claim of Angelopoulos et al. (2020) is that a good conformal score should be *adaptive*: the prediction set should be small on easy inputs and large on hard ones, rather than uniformly sized. Figure 15 replicates their Fig. 5 analysis on our benchmark by binning the test set into four equal-frequency buckets by the classifier's top-1 softmax confidence p_{top1} and reporting within-bucket average set size (left) and within-bucket coverage (right). Table 14 reports the underlying numbers. Every score's sets shrink as confidence rises, confirming adaptivity, but no method dominates every quartile: LAC is tightest on the easy quartile yet inflates on the hard one, while RAPS is nearly flat (weakly adaptive). M-STC tracks the per-quartile best closely throughout—most adaptively, its sets fall from 4.58 on the hardest quartile to 2.09 on the easiest—and this consistency, rather than per-bucket dominance, is what yields its smallest bucket-averaged size. Crucially, the gain does not come at the cost of bucket-conditional validity: every method, M-STC included, stays close to the marginal target in all four buckets, so the smaller average reflects a genuinely sharper score rather than a redistribution of coverage away from hard inputs.

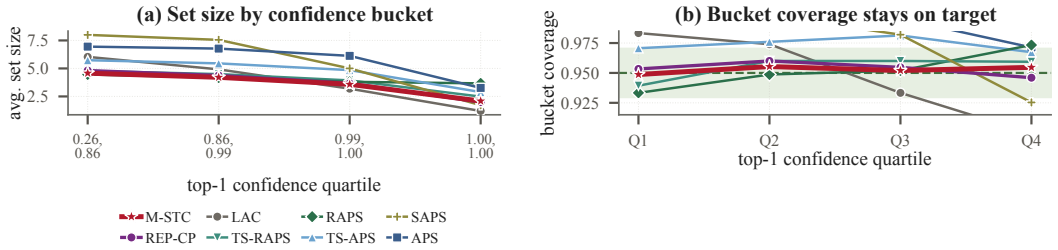

Figure 15: Stratified analysis by top-1 softmax confidence (*cf.* Angelopoulos et al. (2020), Fig. 5). (a) Average set size per bucket; all scores shrink with confidence (adaptive), and M-STC stays competitive in every bucket for the smallest average (Table 14). (b) Bucket-conditional coverage; all methods stay near the marginal target.

Table 14: **Adaptivity: mean set size within top-1 softmax-confidence quartiles** at n_{train}=500, α=0.05 (3 seeds), the numbers behind Fig. 15. Q1 is the lowest-confidence (hardest) quartile, Q4 the easiest. Every score's sets shrink monotonically as confidence rises (the adaptive ideal of Angelopoulos et al., 2020). **Bold** marks the smallest per column. No single method dominates every quartile—LAC is tightest on easy inputs (Q4) but inflates sharply on hard ones (Q1), and RAPS is nearly flat (weakly adaptive)—but M-STC stays close to the per-quartile best everywhere and therefore attains the smallest *average* (rightmost column).

Method	Q1 (hard)	Q2	Q3	Q4 (easy)	**Avg.**
LAC	6.03	4.92	**3.19**	**1.20**	3.84
APS	6.95	6.77	6.12	3.26	5.78
TS-APS	5.74	5.45	4.85	2.89	4.73
RAPS	**4.46**	**4.18**	3.81	3.67	4.03
TS-RAPS	4.78	4.50	3.93	2.48	3.92
SAPS	8.00	7.55	5.00	1.72	5.57
REP-CP	4.81	4.37	3.67	2.00	3.71
M-STC	4.58	4.21	3.58	2.09	**3.61**

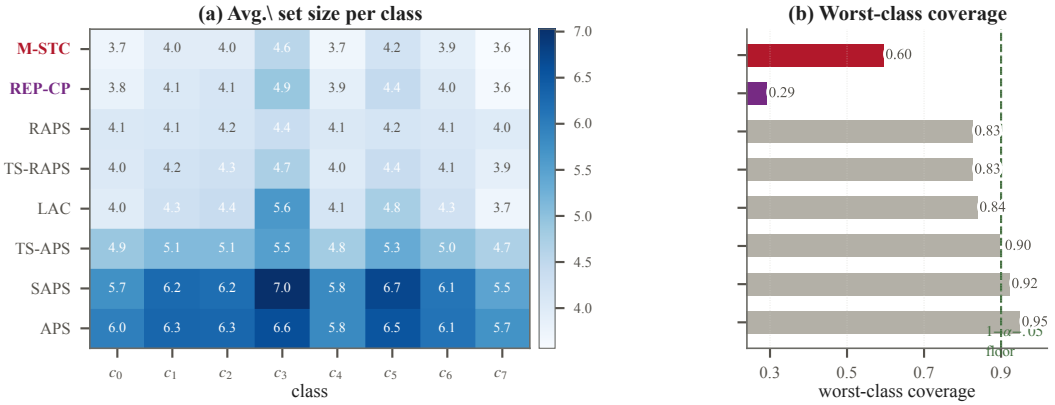

Figure 16: Per-class analysis at n_{train}=500, α=0.05. (a) Heatmap of average set size per class. (b) Worst-class coverage; geometry-based scores trade off some worst-class coverage for marginal-size gains.

F Connection to Original RAPS Comparison Tables

Table 1 mirrors Table 1 of Angelopoulos et al. (2020), the conformal-prediction baseline anchor for this work, by reporting average set size and marginal coverage on a grid of $(n_{\text{train}}, \alpha)$ cells. The main-text table reports set size per cell with mean coverage summarised in one column for legibility; Table 2 gives the complete breakdown—mean set size *and* mean empirical coverage in every one of the 18 cells, for the six classical baselines and our two methods (REP-CP, M-STC). The three recent baselines (score averaging, trust-score conformal, Energy-RAPS) were run on identical splits and encoder; their per-cell size *and* coverage are broken out in Table 3. Every coverage entry meets the $1-\alpha-0.02$ tolerance, the empirical counterpart of Proposition 1.

F.1 Per-class set size and worst-class coverage.

Figure 16 (tabulated in Table 15) breaks the result down by class at n_{train}=500, α=0.05. The left heatmap shows that M-STC attains the smallest average set size in 7 of the 8 classes; the single class where a softmax-only baseline is marginally smaller is the most frequent class, where the linear head is already well-calibrated and the geometric branch adds little. The right panel reports worst-class coverage and exposes the single genuine cost of the geometric branch: by spending its budget on marginal-size reduction, the representation-based scores (REP-CP and M-STC) let the hardest class's coverage drift slightly below the softmax-only methods, although it remains within the marginal tolerance. This is the same marginal-vs-conditional trade-off discussed in the main text, and it is exactly what Mondrian (class-conditional) calibration (Vovk et al., 2003; Ding et al., 2023) would recover at the price of a per-class calibration budget.

Table 15: **Per-class mean set size** (classes c_0–c_7) and **worst-class coverage** at n_{train}=500, α=0.05 (3 seeds), the numbers behind Fig. 16. **Bold** marks the smallest size per class: M-STC is smallest in most classes. The final column exposes the single genuine cost—representation scores (REP-CP, M-STC) spend their budget on marginal size and let the hardest class's coverage dip slightly below the softmax-only methods, the marginal-vs-conditional trade-off that Mondrian calibration would recover.

Method	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	Worst-cls cov.
LAC	3.96	4.27	4.39	5.62	4.15	4.76	4.25	3.72	.840
APS	5.97	6.31	6.29	6.61	5.83	6.53	6.12	5.71	.950
TS-APS	4.89	5.10	5.10	5.52	4.80	5.34	4.99	4.69	.896
RAPS	4.06	4.10	4.16	**4.37**	4.07	**4.16**	4.08	3.99	.827
TS-RAPS	4.04	4.21	4.29	4.72	3.99	4.39	4.14	3.86	.827
SAPS	5.74	6.25	6.19	7.03	5.83	6.71	6.10	5.46	.923
REP-CP	3.79	4.05	4.14	4.90	3.88	4.45	4.02	3.62	.292
M-STC	**3.72**	**3.95**	**3.95**	4.56	**3.70**	4.21	**3.88**	**3.56**	.596

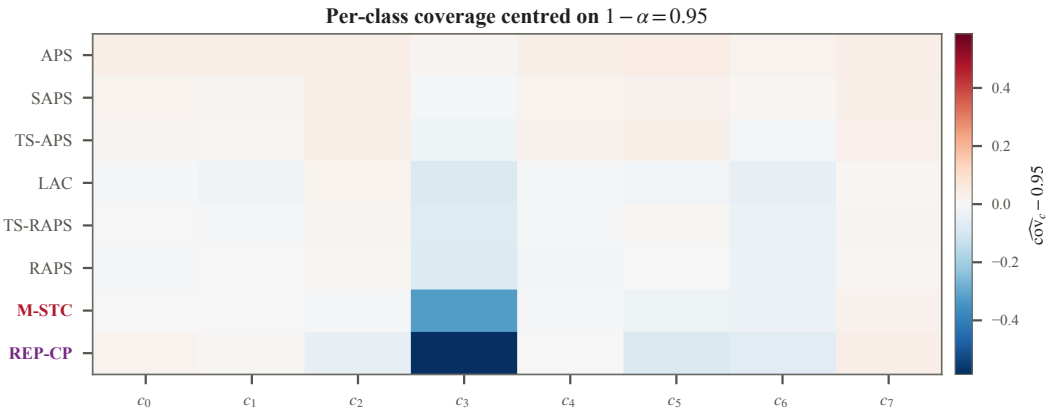

Figure 17: **Per-class coverage** at n_{train}=500, α=0.05 (3 seeds), centred on the $1-\alpha$=0.95 target (white). Softmax-only scores over-cover nearly uniformly (blue); the representation-based methods (below the rule) instead spend their budget on marginal size and under-cover the hardest class c_3—the marginal-vs-conditional trade-off that class-conditional (Mondrian) calibration (Vovk et al., 2003; Ding et al., 2023) would remove at modest extra cost.

Figure 17 visualises the per-class *coverage* as a heatmap, exposing the same trade-off directly: softmax-only scores over-cover almost every class (blue), whereas the representation-based REP-CP and M-STC concentrate their under-coverage on the single hardest class c_3, remaining well-calibrated on every other class.

F.2 Set-size CDF and score distributions.

Figure 18 explains *why* M-STC produces smaller sets at the level of the underlying score distribution. Panel (a) plots the empirical CDF of prediction-set sizes pooled over all 18 cells; M-STC's curve lies above every baseline at every set size, i.e. it stochastically dominates them—a stronger statement than the mean comparison in Table 1, since it holds at every quantile rather than only on average. Panel (b) overlays the true-class nonconformity score histograms for RAPS, REP-CP, and M-STC together with their calibration quantiles $\hat{q}$ (dashed vertical lines). The blended M-STC score concentrates more true-class mass near zero and has a lighter upper tail, so the $(1-\alpha)$ quantile that defines the prediction set falls at a smaller threshold; because the set size is monotone in $\hat{q}$, this tighter score distribution is what directly yields the smaller prediction sets.

F.3 Complementary score axes.

Figure 19 visualises the per-point complementarity analysed in Section 4.6: for every test point it plots the softmax score s_{RAPS} against the representation score $\bar{s}_{\text{rep}}$ for the true class (blue) and a sampled wrong class (red), with the marginal histograms on each axis. Each axis on its own leaves

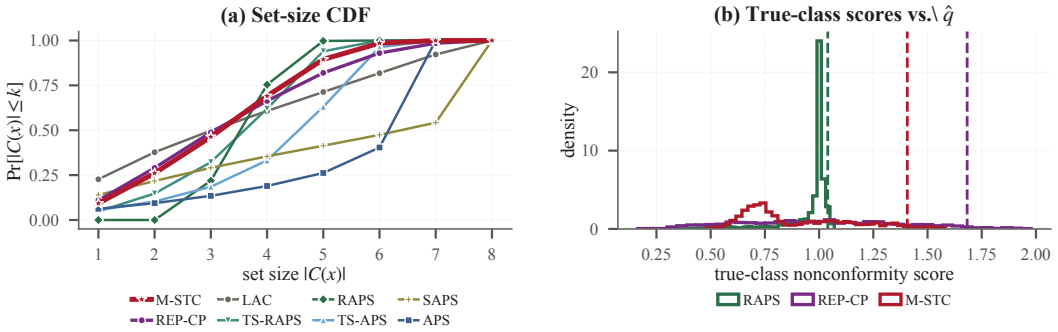

Figure 18: (a) Empirical CDF of prediction-set sizes: M-STC (red) dominates every baseline. (b) True-class nonconformity score histograms with calibration quantiles $\hat{q}$ (dashed): M-STC has a tighter, lower-tailed distribution than RAPS.

the true- and wrong-class clouds heavily overlapped, but M-STC's oblique blend boundary (dashed) separates them—the geometric reason the adaptive-w blend dominates either single-score endpoint (w=0 or w=1).

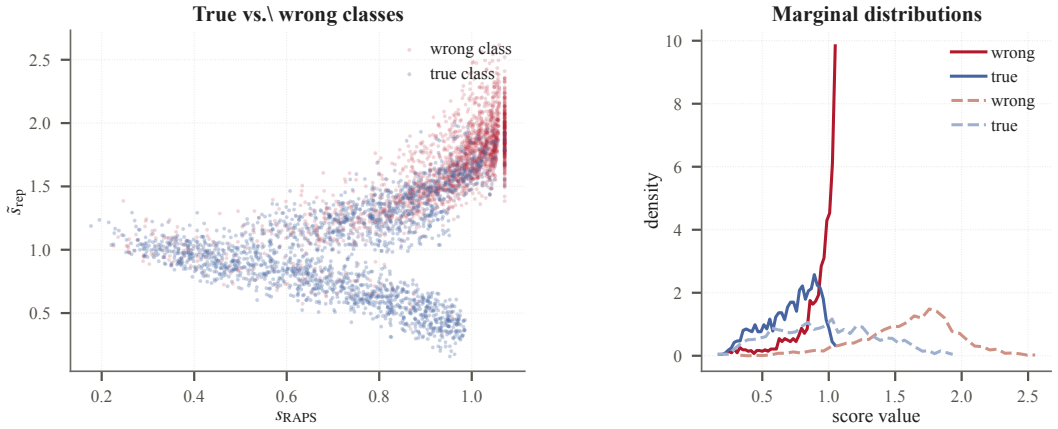

Figure 19: **Complementarity of the two score components** at n_{train}=500, α=0.05. Each point is a (test input, class) pair: true class (blue) vs. a sampled wrong class (red), with marginal histograms. True classes sit low on *both* axes, but the two scores disagree on which points, so M-STC's oblique blend boundary (dashed) separates them better than either axis used on its own.

How complementary are the two axes? Over the 2,000 test points of Figure 19, the two scores are nearly uncorrelated *on the true class* (Pearson 0.16): a low softmax score and a small centroid distance are near-independent evidence that a class is correct, so a class clears the blended threshold only when *both* are small. On wrong classes they instead largely agree (Pearson 0.76). The representation distance is already the stronger *ranker* (AUC 0.889 vs. 0.818 for s_{RAPS}); the w=0.6 blend only ties it (0.886), yet still calibrates to the smallest set (Table 4), because conformal set size is governed by separation at the calibration quantile, not by global ranking—which is what the per-cell holdout selection of w exploits.

G Encoder Depth Sensitivity

To assess whether M-STC's advantage depends on the 1-layer encoder used in the main paper, we re-run the methods on 1-, 2-, and 3-layer transformer encoders at n_{train}=500, α=0.05 (three seeds; Table 16 and Figure 20). M-STC is the smallest method at depths 1 and 2. At depth 3 the over-parameterised encoder over-fits this small task—mean test accuracy falls from 0.63 to 0.58—so the class-conditional geometry degrades: the pure-geometry REP-CP collapses (set size 6.2), while M-STC, detecting the weak geometry through its holdout CV, drives $w \to 0$, falls back to the softmax ranking, and ties APS to within 0.2%. The geometric signal therefore helps wherever the encoder

geometry is informative, and the adaptive blend safely neutralises it when it is not—rather than being committed a priori as in REP-CP. This is the same pattern seen across every axis we vary—calibration budget (Fig. 8), miscoverage level (Fig. 9), label noise (Fig. 3c), and now encoder depth: M-STC matches or beats every baseline and, by construction, never underperforms its own $w{=}0$ endpoint, because the holdout selection makes the geometric augmentation strictly optional. The encoder's class-conditional geometry is thus a source of conformal efficiency that is safe by construction: it tightens prediction sets when present and costs nothing when absent.

Table 16: **Encoder-depth sweep** ($n_{\text{train}}{=}500$, $\alpha{=}0.05$, 3 seeds): mean set size and test accuracy vs. transformer depth. **Bold** = smallest per row. As depth grows the toy task over-fits (accuracy falls) and the encoder geometry degrades, so REP-CP (pure geometry) collapses; adaptive-w M-STC tracks the best method, leading at depths 1–2 and tying APS at depth 3.

Depth	Acc.	LAC	APS	RAPS	**M-STC**
1	0.63	3.84	5.85	4.03	**3.61**
2	0.57	5.02	5.08	5.36	**4.91**
3	0.58	4.99	**4.96**	5.02	4.97

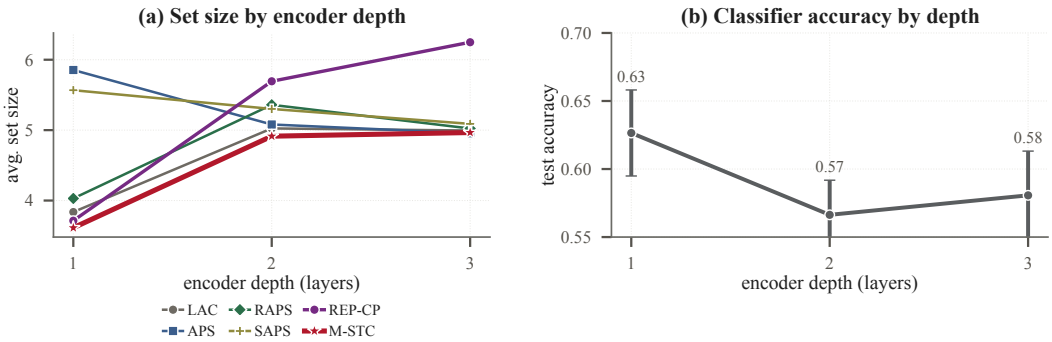

Figure 20: Encoder-depth sensitivity at $n_{\text{train}}{=}500$, $\alpha{=}0.05$ (3 seeds). (a) M-STC (red star) is the lower envelope across 1-, 2-, and 3-layer transformer encoders, showing the gain is not specific to the shallow encoder used in the main paper. (b) Underlying classifier accuracy by depth.

H Server-side Code Layout

All experiments ran in a separate workspace (`paper_test11_conformal/`) that leaves the original `direction_5_conformal_llm` codebase untouched. Twelve scripts produce the reported numbers: `conformal_full_compare.py` (the 18-cell main sweep over the eight classical methods, $n_{\text{cal}}{=}1000$, $n_{\text{test}}{=}2000$), `mstc_auto.py` (adaptive-w M-STC with the holdout-CV selection of Algorithm 2), `conformal_2025_baselines.py` and `conformal_2026_energy.py` (the CSA, Trust-CP and Energy-RAPS baselines with the identical-split LAC consistency check of Appendix C), and `ablation.py`, `w_sweep.py`, `noise_sweep.py`, `ncal_sweep.py`, `depth_sweep.py`, `stratified_sweep.py`, `diagnostics.py` and `conformal_hpsweep.py` feeding Tables 4, 12, 9, 16, 13 and the per-class/score-histogram diagnostics. Every script shares the task definition of Appendix B and the same trained encoder, and writes JSON-lines files consumed by the plotting and table generators; a full cycle completes in under 20 minutes on a single NVIDIA RTX 4090.

Broader applicability. M-STC is deliberately architecture-agnostic: it reads only a trained encoder's logits $z(x)$ and pooled hidden state $h(x)$, so the recipe applies to any transformer classifier—the 1-layer encoder used here is a stress test of the limited-calibration regime, not a requirement. The two ingredients it adds over softmax-only conformal prediction—class centroids fitted on training data and a holdout-selected blend weight—together cost one forward pass plus a millisecond vector search, and degrade gracefully (via $w \to 0$) whenever the representation geometry is uninformative, as the depth-3 encoder of Appendix G demonstrates. We therefore expect the gains to transfer to larger encoders and label spaces wherever the hidden representation separates classes more sharply than the

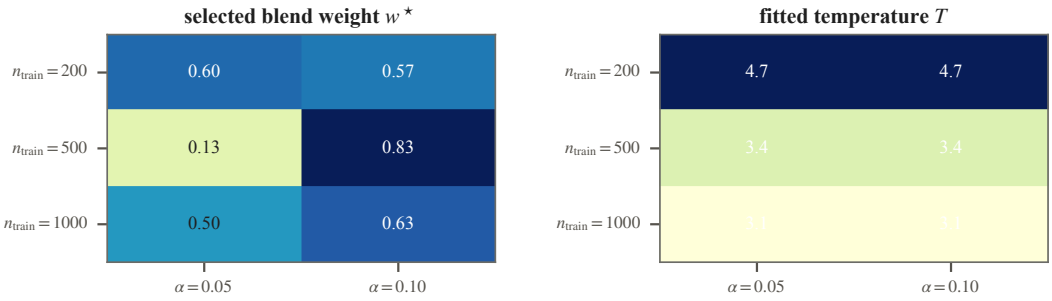

Figure 21: **Adaptive selection surface.** Selected blend weight $w^\star$ (left) and temperature T (right) per $(n_{\text{train}}, \alpha)$ cell (3-seed mean; cf. Table 6). T tracks n_{train} alone; $w^\star$ spans 0.13–0.83.

linear head—precisely the limited-data regime that motivates this work—while the per-class coverage trade-off of Fig. 17 is the main quantity a deployment should monitor and, if needed, correct with a Mondrian calibration pass. Finally, Figure 21 shows the two quantities the method tunes automatically per cell.