CLAPS: Posterior-Aware Conformal Intervals via Last-Layer Laplace

Dongseok Kim [*1] **Hyoungsun Choi** [*1] **Mohamed Jismy Aashik Rasool** [*1] **Gisung Oh** [1]

Abstract

We present CLAPS, a posterior-aware conformal regression method that pairs a Last-Layer Laplace Approximation with split-conformal calibration. From the resulting Gaussian posterior, CLAPS defines a simple two-sided posterior CDF score that aligns the conformity metric with the full predictive shape, not just a point estimate. This alignment yields narrower prediction intervals at the same target coverage, especially on small to medium tabular datasets where data are scarce and uncertainty modeling matters. We also provide a lightweight diagnostic suite that separates aleatoric and epistemic components and visualizes posterior behavior, helping practitioners understand why intervals shrink when they do. Across multiple benchmarks using the same MLP backbone, CLAPS consistently attains nominal coverage with improved efficiency and minimal overhead, offering a clear, practical upgrade to residual-based conformal baselines.

1. Introduction

Prediction intervals must balance two goals: maintaining target coverage in finite samples and remaining as narrow as possible to support decisions. Split-conformal prediction guarantees the former via rank-based calibration, but its efficiency hinges on how the nonconformity score is defined. Most tabular regression pipelines still rely on absolute or normalized residuals, which ignore the full shape of predictive uncertainty.

We introduce **CLAPS**(Conformal Laplace-Approximated Posterior Scoring), a posterior-aware conformal regression method that couples a *Last-Layer Laplace Approximation* (LLLA) with split-conformal calibration. LLLA provides a lightweight Gaussian posterior over the linear head of a fixed backbone network. From this posterior, CLAPS defines a simple two-sided posterior CDF ("centrality") score that quantifies how typical an outcome is under the entire predictive distribution, not just relative to a point estimate. Aligning the score with the posterior shape yields tighter intervals at the same nominal coverage, especially in small to medium data regimes where modeling uncertainty is most consequential.

Beyond intervals, practitioners need to understand *why* a method works. We therefore propose a compact diagnostic suite that decomposes predictive variance into aleatoric and epistemic components, tracks posterior contraction, and quantifies heteroscedasticity signals. These diagnostics help select hyperparameters, reveal when posterior information is informative, and explain observed efficiency gains.

Contributions.

- A posterior-aware conformal method, **CLAPS**, that uses an LLLA posterior to define a two-sided posterior CDF score and calibrates it with split-conformal rules.
- A diagnostic framework that separates aleatoric and epistemic uncertainty, measures posterior contraction, and surfaces heteroscedasticity cues for method selection and debugging.
- A same-backbone experimental study on standard tabular benchmarks showing that CLAPS attains target coverage with consistently narrower intervals and minimal computational overhead.
- An easily adoptable recipe: keep the backbone, fit a last-layer Laplace posterior, compute the centrality score, and apply off-the-shelf split-conformal calibration.

Scope. Our focus is tabular regression with feed-forward backbones, where LLLA is especially simple to implement and deploy. The overall pipeline preserves conformal validity, adds negligible training burden, and improves efficiency by aligning the score with the posterior predictive structure.

2. Related Work

Conformal for Regression: from residual to distribution-shaped scores. Split conformal regression calibrates

[*]Equal contribution [1]Department of Computer Engineering, Gachon University, Seongnam, Gyeonggi, Republic of Korea. Correspondence to: Gisung Oh <eustia@gachon.ac.kr>.

Preprint. December 2, 2025.

residual-shaped scores to guarantee marginal coverage, and fold-reuse variants such as jackknife+ and cross-fitting improve data efficiency (Barber et al., 2021; Angelopoulos et al., 2023). Recent work emphasizes *efficiency*—minimizing expected length or controlling a user-chosen risk—while preserving guarantees, including optimization-driven procedures and general risk-control frameworks (Kiyani et al., 2024; Angelopoulos et al., 2022; Overman et al., 2024). Concurrently, methods move beyond absolute residuals toward *distribution-shaped* scores that leverage estimates of $P(Y \mid X)$ (e.g., CDF-/quantile-based and probabilistic variants) or reframe regression as classification to better capture heteroscedasticity, skewness, and multimodality (Chernozhukov et al., 2021; Romano et al., 2019; Guha et al., 2024; Plassier et al., 2024a). We adopt the same rank-based split calibration rule but define the conformity via a symmetric CDF from a predictive distribution (instantiated later with LLLA), placing our approach on the distribution-shaped branch; for completeness we also compare to interval-thresholding baselines tailored for width efficiency (Luo & Zhou, 2025).

Heteroscedastic Adaptation and Quantile-Based Methods. A major route to width efficiency under input-dependent noise is to learn *scale or quantiles* and calibrate them with conformal ranks. Conformalized Quantile Regression (CQR) replaces absolute-residual scores by a max-quantile deviation, achieving robustness to heteroscedasticity and skewed tails (Romano et al., 2019). Building on this idea, boosted or reweighted schemes further reduce interval length by improving the base quantile fits while preserving split-calibrated guarantees (Xie et al., 2024). Localization and weighting adapt calibration to covariate neighborhoods or strata, yielding relaxed conditional guarantees and improved coverage balance across subpopulations (Hore & Barber, 2023; Bhattacharyya & Barber, 2024). When conditional distributions are multi-modal/asymmetric, reframing regression targets (e.g., via discretization or surrogate tasks) can enhance quantile learning before conformalization (Guha et al., 2024). Recent general treatments study heterogeneity and non-exchangeability directly, proposing efficient conformal procedures with validity under drift and data heterogeneity (Plassier et al., 2024b). In survival/structured settings, conformalized quantile ideas are extended with monotonicity-aware or task-specific heads to stabilize learned quantiles while keeping rank calibration (Qi et al., 2024). We adopt CQR and localized/weighted variants as primary baselines; our method differs by deriving a posterior-aware, CDF-symmetric score, whereas these approaches *learn* input-dependent scales/quantiles and then apply the same split rank rule.

Posterior-Aware Scores and Last-Layer Bayesianization. Posterior-aware conformal methods use a predictive *distribution* to define CDF/HPD-style conformity scores, then retain split/rank calibration. Recent advances show how modelling the conditional distribution improves conditional validity and efficiency, e.g., clustering- or distribution–driven calibration (Zhang & Candès, 2024; Plassier et al., 2024a; Chernozhukov et al., 2021). In neural networks, a lightweight route to such predictive distributions is to endow only the final linear layer with a Bayesian treatment. Last-layer Laplace and its modern variants provide Gaussian posterior predictives around the MAP solution with negligible retraining cost (Kristiadi et al., 2020; Daxberger et al., 2021). Beyond the classical diagonal/empirical-Hessian forms, recent work improves the fidelity–cost trade-off via variational last layers, function-space priors, and geometry-aware approximations (Harrison et al., 2024; Cinquin et al., 2024; Bergamin et al., 2023). We adopt this *last-layer Bayesianization* to instantiate a symmetric CDF score from the LLLA posterior predictive on a shared MLP backbone, then calibrate by ranks; this places our method at the intersection of posterior-aware conformal scoring and efficient Laplace-based uncertainty, and is complementary to non-Bayesian distributional surrogates (Zhang & Candès, 2024; Plassier et al., 2024a).

Diagnostics and Method Selection. We accompany interval results with *diagnostics* that disentangle aleatoric and epistemic components and indicate when score design or scale learning is more advantageous. Concretely, we report the variance decomposition $\hat{v}(x) = \sigma^2 + \phi(x)^\top \Sigma \phi(x)$, the epistemic share $r(x) = \mathrm{epi}/(\sigma^2 + \mathrm{epi})$, the posterior covariance mass $\mathrm{tr}(\Sigma)$, subsample curves of epi and $\mathrm{tr}(\Sigma)$ as n grows (posterior contraction), and the rank correlation $\mathrm{Spearman}(|e|, \sqrt{\hat{v}})$ as a heteroscedasticity signal. The latter follows classic practice of testing variance–signal association via rank/monotone correlation (Yin & Carroll, 1990). We also assess *probabilistic calibration* of regression predictions via CDF recalibration and coverage diagnostics (Kuleshov et al., 2018). Beyond marginal coverage, we examine *localized* or *weighted* coverage balance across covariate neighborhoods using recent randomized-local methods and related conditional-validity relaxations (Hore & Barber, 2023). Under nonstationarity, we track coverage under drift and employ adaptive/wide–narrow adjustments from online conformal inference (Gibbs & Candès, 2024). These indicators guide a simple rule: when $r(x)$ concentrates near 0 and $\mathrm{Spearman}(|e|, \sqrt{\hat{v}})$ is strongly positive (large, structured heteroscedasticity), *quantile/scale–learning* baselines (e.g., CQR/normalized CP) tend to yield tighter widths at target coverage; when posterior contraction is limited (non-trivial epi and $\mathrm{tr}(\Sigma)$) and heterogeneity signals are weak, *posterior-aware* scores (ours) are competitive or superior in width. We report both sets of diagnostics alongside coverage/width to justify the method choice on each dataset.

Complementary Lines: Localization, Weighting, and Bootstrap Ensembles. Beyond score design, three complementary axes improve efficiency and coverage balance. First, *localization* tailors calibration to neighborhoods of the test covariate, yielding relaxed conditional guarantees and better subgroup balance (Guan, 2023; Hore & Barber, 2023). Second, *weighting* focuses coverage where distribution shift or stratification matters—either by importance-style reweighting or group-aware corrections—without changing the base score (Bhattacharyya & Barber, 2024). Third, *bootstrap ensembles* aggregate diverse learners or resamples to stabilize interval length under model misspecification or dependence, including sequential/time-series settings (Xu & Xie, 2021; 2023; Gupta et al., 2022; Qian et al., 2024; Rivera et al., 2024). These lines are orthogonal to our posterior-aware scoring: our method can plug into localized or weighted calibration, and can be paired with ensemble wrappers to trade off width and robustness while preserving split/rank guarantees.

3. Method

We propose **CLAPS**, a posterior-aware conformal regression method that combines a *Last-Layer Laplace Approximation* (LLLA) with split-conformal calibration. CLAPS aligns the nonconformity score with the model's posterior predictive shape and then uses rank-based calibration to guarantee finite-sample coverage.

3.1. Problem Setup

We consider supervised regression with inputs $x \in \mathcal{X} \subset \mathbb{R}^p$ and scalar response $y \in \mathbb{R}$. A neural network consists of a fixed backbone feature map $\phi(x) \in \mathbb{R}^d$ (last hidden layer) and a linear head $f_w(x) = \phi(x)^\top w$. We split data into training $\mathcal{D}_{\text{tr}}$, calibration $\mathcal{D}_{\text{cal}}$, and test $\mathcal{D}_{\text{te}}$.

3.2. Last-Layer Laplace Approximation (LLLA)

We place a Gaussian prior $w \sim \mathcal{N}(0, \lambda^{-1} I)$ and assume homoscedastic Gaussian noise $y \mid x, w \sim \mathcal{N}(\phi(x)^\top w, \sigma^2)$. Let $\Phi \in \mathbb{R}^{n \times d}$ be the design matrix of $\phi(x)$ for $\mathcal{D}_{\text{tr}}$ and $y \in \mathbb{R}^n$ the targets. The MAP solution and posterior covariance for the last layer are

$$w_{\text{MAP}} = \arg\min_w \tfrac{1}{2\sigma^2}\|y - \Phi w\|_2^2 + \tfrac{\lambda}{2}\|w\|_2^2,$$
$$\Sigma = \left(\lambda I + \tfrac{1}{\sigma^2}\Phi^\top \Phi\right)^{-1}.$$

The LLLA posterior is $w \mid \mathcal{D}_{\text{tr}} \sim \mathcal{N}(w_{\text{MAP}}, \Sigma)$ and yields a Gaussian posterior predictive for any x,

$$y \mid x, \mathcal{D}_{\text{tr}} \sim \mathcal{N}\big(\hat{\mu}(x),\, \hat{v}(x)\big),$$
$$\hat{\mu}(x) = \phi(x)^\top w_{\text{MAP}},$$
$$\hat{v}(x) = \sigma^2 + \phi(x)^\top \Sigma\, \phi(x).$$

We fix the last–layer prior precision to $\lambda = 1.0$ unless otherwise stated, and estimate the noise variance σ^2 either by maximizing the (approximate) marginal likelihood under the last–layer Gaussian model or—consistently in our experiments—via residuals on the training split $\mathcal{D}_{\text{tr}}$ (with standardized target):

$$\sigma^2 = \frac{1}{|\mathcal{D}_{\text{tr}}|}\left\|y_{\text{tr}} - \Phi w_{\text{MAP}}\right\|_2^2,$$
$$w_{\text{MAP}} = \left(\Phi^\top \Phi + \lambda I\right)^{-1} \Phi^\top y_{\text{tr}}.$$

For numerical stability and efficiency, we avoid forming Σ explicitly. Let

$$M \;=\; \lambda I \;+\; \tfrac{1}{\sigma^2}\,\Phi^\top \Phi \;=\; L L^\top$$

be the Cholesky factorization. Then the epistemic term is evaluated without materializing Σ as

$$\phi(x)^\top \Sigma\, \phi(x) \;=\; \left\|L^{-1}\phi(x)\right\|_2^2,$$

which we compute with two triangular solves.

3.3. Two-Sided Posterior CDF Score

Given the Gaussian posterior predictive, define the *centrality* (two-sided CDF) score

$$s(x, y) = \min\left(F_{\text{post}}(y \mid x),\, 1 - F_{\text{post}}(y \mid x)\right),$$

where F_{post} is the CDF of $\mathcal{N}(\hat{\mu}(x), \hat{v}(x))$. Intuitively, s is large for outcomes typical under the predictive distribution (near its center) and small for tail outcomes; it thus aligns nonconformity with the full posterior shape instead of only point residuals. For the Gaussian case, if $z = \frac{y - \hat{\mu}(x)}{\sqrt{\hat{v}(x)}}$ and $\Phi(\cdot)$ is the standard normal CDF, then $s(x, y) = \min\{\Phi(z), 1 - \Phi(z)\}$.

3.4. Split-Conformal Calibration

Compute scores on the calibration set $S_{\text{cal}} = \{s(x_i, y_i)\}_{(x_i, y_i) \in \mathcal{D}_{\text{cal}}}$ and take the rank-based threshold

$$t \;=\; \text{Quantile}_\alpha\big(S_{\text{cal}}\big),$$

where $\alpha = \frac{\lceil (|\mathcal{D}_{\text{cal}}|+1)(1-\text{target\_cov})\rceil}{|\mathcal{D}_{\text{cal}}|+1}$ (standard split-conformal choice). For a new x, the conformal prediction set is

$$\mathcal{C}(x) \;=\; \big\{y :\; s(x, y) \geq t\big\}.$$

With Gaussian F_{post}, $\mathcal{C}(x)$ is a central interval:

$$\mathcal{C}(x) = \Big[\hat{\mu}(x) + \sqrt{\hat{v}(x)}\,\Phi^{-1}(t),\; \hat{\mu}(x) + \sqrt{\hat{v}(x)}\,\Phi^{-1}(1-t)\Big].$$

This preserves marginal coverage in finite samples by the conformal rank rule while typically improving efficiency when the posterior carries informative structure.

3.5. CLAPS Procedure

A compact posterior-aware conformal routine: fit an LLLA posterior on the head, score calibration samples by two-sided CDF centrality, take the conformal quantile, and output central intervals. See Algorithm 1 for a summary.

Algorithm 1 CLAPS (posterior-aware split-conformal)

Require: $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{cal}}$, backbone $\phi(\cdot)$ (fixed), λ
Ensure: $\mathcal{C}(\cdot)$
$\Phi \leftarrow [\phi(x)]_{(x,y)\in\mathcal{D}_{\mathrm{tr}}}$; $w_{\mathrm{MAP}} \leftarrow \arg\min_w \frac{1}{2\sigma^2}\|y - \Phi w\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
$M \leftarrow \lambda I + \sigma^{-2}\Phi^\top\Phi = LL^\top$ {Cholesky}
for all $(x_i, y_i) \in \mathcal{D}_{\mathrm{cal}}$ **do**
  $\mu_i \leftarrow \phi(x_i)^\top w_{\mathrm{MAP}}$, $v_i \leftarrow \sigma^2 + \|L^{-1}\phi(x_i)\|_2^2$
  $z_i \leftarrow (y_i - \mu_i)/\sqrt{v_i}$, $s_i \leftarrow \min\{\Phi(z_i), 1 - \Phi(z_i)\}$
end for
$t \leftarrow \mathrm{Quantile}_\alpha\left(\{s_i\}_{\mathcal{D}_{\mathrm{cal}}}\right)$
for all test x **do**
  $\mu \leftarrow \phi(x)^\top w_{\mathrm{MAP}}$, $v \leftarrow \sigma^2 + \|L^{-1}\phi(x)\|_2^2$
  $\mathcal{C}(x) \leftarrow \left[\mu + \sqrt{v}\,\Phi^{-1}(t),\ \mu + \sqrt{v}\,\Phi^{-1}(1-t)\right]$
end for

3.6. Complexity and Deployment

If d is the last-layer width and n the training size, forming M costs $O(nd^2)$ and its Cholesky $O(d^3)$ once per model. Per-query costs are $O(d^2)$ for $\hat{\mu}(x)$ and $O(d^2)$ (or $O(d)$ with cached L^{-1}-vector solves) for $\hat{v}(x)$. CLAPS adds negligible overhead to standard same-backbone pipelines while providing a posterior-aware score amenable to fast batch inference and plug-and-play conformal calibration.

4. Theory

We analyze the validity and efficiency of *CLAPS*. Throughout, condition on the fitted backbone $\phi(\cdot)$ and the last-layer MAP/posterior trained on $\mathcal{D}_{\mathrm{tr}}$; randomness arises only from the exchangeable calibration and test points.

Definition 4.1 (LLLA posterior and predictive)**.** Let $w \mid \mathcal{D}_{\mathrm{tr}} \sim \mathcal{N}(w_{\mathrm{MAP}}, \Sigma)$ with $\Sigma = (\lambda I + \sigma^{-2}\Phi^\top\Phi)^{-1}$ and $\Phi = [\phi(x)]_{(x,y)\in\mathcal{D}_{\mathrm{tr}}}$. The posterior predictive is

$$y \mid x, \mathcal{D}_{\mathrm{tr}} \sim \mathcal{N}\big(\hat{\mu}(x),\, \hat{v}(x)\big),$$
$$\hat{\mu}(x) = \phi(x)^\top w_{\mathrm{MAP}},$$
$$\hat{v}(x) = \sigma^2 + \phi(x)^\top \Sigma\, \phi(x).$$

and numerically $\phi(x)^\top\Sigma\phi(x) = \|L^{-1}\phi(x)\|_2^2$ for $LL^\top = \lambda I + \sigma^{-2}\Phi^\top\Phi$.

Definition 4.2 (Two-sided posterior CDF score)**.** Define the *centrality* score

$$s(x, y) = \min\big\{F_{\mathrm{post}}(y \mid x),\, 1 - F_{\mathrm{post}}(y \mid x)\big\}.$$

For the Gaussian case, with $z = (y - \hat{\mu}(x))/\sqrt{\hat{v}(x)}$, we have $s(x, y) = \min\{\Phi(z), 1 - \Phi(z)\}$. Let t be the split-conformal rank quantile of $\{s(x_i, y_i)\}_{(x_i,y_i)\in\mathcal{D}_{\mathrm{cal}}}$ at level α.

Assumption 4.3 (Exchangeability)**.** Conditional on $\mathcal{D}_{\mathrm{tr}}$, the calibration multiset $\mathcal{D}_{\mathrm{cal}}$ and any test pair (x, y) are exchangeable draws from the same data-generating process.

Proposition 4.4 (Finite-sample marginal coverage)**.** *Under Assumption 1, the CLAPS interval*

$$\mathcal{C}(x) = \left[\hat{\mu}(x) + \sqrt{\hat{v}(x)}\,\Phi^{-1}(t),\ \hat{\mu}(x) + \sqrt{\hat{v}(x)}\,\Phi^{-1}(1-t)\right]$$

satisfies $\Pr\{\, y \in \mathcal{C}(x)\,\} \geq 1 - \alpha$ *marginally over the calibration set and the test point.*

Proof sketch. Split-conformal prediction guarantees $1 - \alpha$ marginal coverage for any measurable score via the rank-based quantile rule. CLAPS fixes the measurable $s(x, y)$; thus the standard argument applies verbatim. The use of a posterior-aware score affects only efficiency, not validity. □

Lemma 4.5 (Monotone transform of absolute z)**.** *In the Gaussian case, $s(x, y) = g(|z|)$ where $g(u) = \min\{\Phi(u), 1 - \Phi(u)\}$ is strictly decreasing in $u \geq 0$.*

Corollary 4.6 (Equivalence to central/HPD sets)**.** *For any fixed x with Gaussian predictive, the superlevel set $\{y : s(x, y) \geq t\}$ is the shortest (HPD) interval of posterior mass $1 - 2t$. Consequently, CLAPS aligns the conformal acceptance region with the posterior notion of "typicality" rather than raw residual magnitude.*

Theorem 4.7 (Oracle efficiency under correct specification)**.** *Fix x and suppose $F_{\mathrm{post}}(\cdot \mid x)$ equals the true conditional distribution. Among scores whose superlevel sets are nested in t, CLAPS yields intervals of minimal Lebesgue measure $|\mathcal{C}(x)|$ subject to $\Pr\{y \in \mathcal{C}(x) \mid x\} \geq 1 - \alpha$.*

Proof sketch. Shortest-probability sets at a fixed mass are superlevel sets of the density (HPD). For symmetric unimodal Gaussians, HPD coincides with central quantile intervals, which are induced by thresholding the two-sided CDF. Conformal calibration transfers the oracle probability constraint to finite-sample marginal coverage while preserving the nesting, hence minimal expected width under correct specification. □

Remark 4.8 (Misspecification and efficiency). Proposition 4.4 holds without any correctness assumption on F_{post}; only the width changes. When the fitted posterior shape retains information about tailness beyond point residuals, CLAPS increases the empirical threshold t and improves average width. If the posterior collapses to a homoscedastic form, residual-/scale-based scores can be more informative.

Theorem 4.9 (Posterior contraction and regime change)**.** *As $n = |\mathcal{D}_{\text{tr}}| \to \infty$ with fixed backbone width, the spectrum of $\Phi^\top \Phi$ grows and $\Sigma = (\lambda I + \sigma^{-2}\Phi^\top\Phi)^{-1} \to 0$. Hence $\phi(x)^\top \Sigma \phi(x) \to 0$ for all x and $\hat{v}(x) \to \sigma^2$ (last-layer epistemic collapse). In this regime CLAPS reduces to conformalization of a homoscedastic Gaussian predictor, and its width advantage over residual-based scores diminishes.*

Definition 4.10 (Diagnostics linking theory to practice)**.** Let $\text{epi}(x) = \phi(x)^\top \Sigma \phi(x)$ and define

$$r(x) = \frac{\text{epi}(x)}{\sigma^2 + \text{epi}(x)},$$
$$\text{trace}(\Sigma) \quad \text{(posterior covariance trace)},$$
$$\rho = \text{Spearman}\big(|y - \hat{\mu}(x)|,\ \sqrt{\hat{v}(x)}\big).$$

We further inspect subsample curves of $\text{epi}(x)$ and $\text{trace}(\Sigma)$ versus n.

Proposition 4.11 (A simple selection rule)**.** *If r and $\text{trace}(\Sigma)$ are nontrivial while ρ is modest, prefer CLAPS; if $r \approx 0$ (contraction) but ρ is substantial (heteroscedasticity), prefer scale-learning baselines (e.g., normalized-CP/CQR). This rule is consistent with Theorems 4.7 and 4.9.*

Remark 4.12 (Implementation note). All quantities above use the same fitted backbone and the closed-form LLLA head (via a Cholesky factor L). Algorithm 1 details the routine; no additional stochastic approximations are required.

5. Experiments

Datasets. We evaluate CLAPS on four standard tabular *regression* benchmarks spanning small to large scales and varying degrees of heterogeneity: **UCI Airfoil** (1,503 examples; 5 continuous input features; target is sound pressure level), **OpenML kin8nm** (8,192 examples; 8 continuous features; robotic arm kinematics response), **UCI CASP**, the Physicochemical Properties of Protein Tertiary Structure dataset (45,730 examples; 9 physicochemical features; target is RMSD), and **UCI YearPredictionMSD** (a fixed split with 463,715 train and 51,630 test instances; 90 audio features; target is release year). All features are numeric; we standardize inputs to zero mean and unit variance using statistics from the training split only, and we center the target during model fitting but report all metrics in the original target scale. These datasets cover regimes where last-layer epistemic variance is nontrivial in small/medium sets (Airfoil, kin8nm, CASP) and globally contracted in large-scale settings (YearPredictionMSD), allowing us to probe when posterior-aware scoring is most beneficial.

Model architecture. All methods share the same *backbone* MLP to isolate the effect of uncertainty handling. For Airfoil and kin8nm, we use a 2-layer MLP with 128 hidden units per layer; for CASP and YearPredictionMSD, we use a 3-layer MLP with 256 hidden units per layer. The output head for the base predictor is a single linear neuron (scalar regression). Our method (CLAPS) applies a last-layer Laplace approximation on this linear head only, leaving the backbone fixed; competing baselines reuse the identical backbone and differ only in the head (e.g., a single point head for Baseline-CP, a positive scale head for Normalized-CP, and two quantile heads for CQR).

Comparison methods.

- **Baseline-CP** (CP) (Lei et al., 2018): A single-point head trained for squared error; calibrate absolute residuals $|y - \hat{\mu}(x)|$ on the calibration split and output $\hat{\mu}(x) \pm q_{1-\alpha}$.
- **Normalized-CP** (Norm-CP) (Papadopoulos et al., 2008): Add a positive *scale head* $h(x) > 0$ and use the normalized residual $|y - \hat{\mu}(x)|/h(x)$ as the score to obtain input-adaptive (heteroscedastic) widths.
- **Conformalized Quantile Regression** (CQR) (Romano et al., 2019): Replace the point head with two quantile heads $(\hat{q}_{\text{lo}}(x), \hat{q}_{\text{hi}}(x))$ and calibrate the nonnegative score $\max\{\hat{q}_{\text{lo}}(x) - y,\ y - \hat{q}_{\text{hi}}(x),\ 0\}$ to produce the narrowest interval consistent with split ranks.
- **Conservative Targeted Intervals** (CTI) (Luo & Zhou, 2025): Pretrain multiple candidate target coverages $\{1 - \alpha_k\}$ and select at calibration time via an integer-valued rank score; typically trades width for higher reliability.
- **CLAPS** (ours): Keep the same point head, fit a last-layer Laplace posterior, and use the two-sided posterior CDF centrality $s(x, y) = \min\{\Phi(z),\ 1 - \Phi(z)\}$ with $z = (y - \hat{\mu})/\sqrt{\hat{v}}$ as the conformal score; output central posterior-shaped intervals after rank calibration.

Metrics. We report three standard quantities on the held-out test set: **coverage** (higher is better), **width** (lower is better), and **MAE** (context-restored). Coverage is the marginal fraction of test points whose true response falls inside the predicted interval,

$$\text{Cov} = \tfrac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{\, y_i \in \mathcal{C}(x_i)\,\},$$

with target $1 - \alpha = 0.90$. Width is the average interval length,

$$\text{Wid} = \tfrac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \big(\mathcal{C}_{\text{hi}}(x_i) - \mathcal{C}_{\text{lo}}(x_i)\big),$$

and MAE is the point prediction error of the shared backbone head,

$$\mathrm{MAE} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} |y_i - \hat{\mu}(x_i)|,$$

where $\hat{\mu}$ is the backbone's scalar regressor (identical across methods). We accompany each metric with a 95% confidence interval: for coverage we use the Wilson score interval for a binomial proportion with $n = n_{\mathrm{test}}$; for width and MAE we use a normal approximation $\bar{m} \pm 1.96\, s/\sqrt{n_{\mathrm{test}}}$ based on per-example sample variance. When averaging over multiple random splits, we first aggregate per-split means and then form a t-interval across splits.

Results. Figure 1 summarizes the experimental outcomes across datasets. All approaches attain marginal coverage near the 0.90 target; CTI is consistently conservative (roughly 0.94–0.97). On small/medium datasets (Airfoil, kin8nm), CLAPS yields the narrowest intervals at comparable coverage (e.g., Airfoil width ≈ 7.78 compared with 8.85–14.31; kin8nm ≈ 0.27 compared with 0.28–0.68) and achieves the lowest MAE. On CASP, widths are close overall, with a slight edge to Normalized-CP (about 12.98 compared with 13.91 for CLAPS) at similar coverage. On large-scale YearPrediction, where last-layer epistemic variance collapses and heteroscedasticity dominates, Normalized-CP and CQR produce markedly tighter intervals (~ 22.8–23.0) than CLAPS/CP/CTI (~ 27.6–28.0) at comparable coverage, while CLAPS still attains the lowest MAE. For complete numbers and confidence intervals, please see Appendix Tables 2 (point estimates) and 3 (95% CIs).

Ablation study. We vary the LLLA prior precision $\lambda \in \{0.1, 0.3, 1, 3, 10\}$ and the aleatoric estimator $\hat{\sigma}^2 \in \{\hat{\sigma}^2_{\mathrm{EB}}, \hat{\sigma}^2_{\mathrm{res}}\}$ on Airfoil at $\alpha = 0.10$. Coverage remains close to the target across settings, while interval width changes smoothly with λ. The choice between evidence–based $\hat{\sigma}^2_{\mathrm{EB}}$ and residual–based $\hat{\sigma}^2_{\mathrm{res}}$ does not materially affect coverage or width at this scale, indicating robustness of the CLAPS calibration to the noise estimator. Complete ablation results and the corresponding 95% confidence intervals are provided in Table 6 and Table 7.

λ	$\hat{\sigma}^2$	Cov	Wid	τ
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9007	7.7100	0.036598
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9007	7.7106	0.030127
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	0.8742	7.5557	0.035150
	$\hat{\sigma}^2_{\mathrm{res}}$	0.8742	7.5557	0.035150
1	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9007	7.7842	0.034182
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9007	7.7842	0.034182
3	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9007	7.9756	0.033639
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9007	7.9756	0.033639
10	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9073	8.2366	0.025537
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9073	8.2366	0.031979

Table 1. Ablation on **Airfoil** at $\alpha = 0.10$: effect of λ and $\hat{\sigma}^2$ estimator on coverage, width, and resulting τ. 95% CIs are reported in the Appendix.

6. Discussion and Limitations

Discussion. CLAPS aligns the conformal nonconformity with the last-layer Laplace posterior, which proved most beneficial on small to medium tabular datasets where epistemic variance is nontrivial. In Airfoil and kin8nm, this alignment yielded narrower intervals at the same nominal coverage, consistent with the posterior-aware efficiency arguments; on the large-scale YearPrediction, last-layer epistemic variance collapsed ($\widehat{v}(x)! \to !\sigma^2$) and methods that learn input-dependent scales (Normalized-CP, CQR) achieved tighter widths. The ablation study highlights additional strengths of CLAPS: performance is remarkably insensitive to the regularization level λ and to the choice of aleatoric estimator ($\widehat{\sigma}^2 * \mathrm{EB}$ and $\widehat{\sigma}^2 * \mathrm{res}$ lead to essentially identical coverages and very similar widths), while the conformal quantile τ adjusts smoothly as λ varies. This robustness reduces tuning burden, provides predictable behavior across datasets, and ensures graceful degradation when epistemic variance is small, all while maintaining exact finite-sample validity. Conceptually, scale- or quantile-based conformalizers primarily target aleatoric variability, whereas CLAPS leverages epistemic tail information captured by the last-layer posterior; this regime split clarifies when each family is preferable. Practically, CLAPS can serve as a complementary option selected by simple diagnostics such as $r = \mathrm{epi}/\widehat{v}(x)$, $\mathrm{tr}(\Sigma)$, or the rank correlation between $|y - \mu(x)|$ and $\sqrt{\widehat{v}(x)}$; a hybrid formulation that interpolates between a posterior-aware term and a learned scale term using r is a promising direction for future work.

Limitations. Because CLAPS models uncertainty only in the linear last layer, it cannot capture representation-level epistemic variation inside the backbone; when the feature map is misspecified, the posterior variance can understate uncertainty even if the head is well calibrated. First, posterior contraction at large n diminishes epistemic contributions, limiting the advantage of posterior-aware scoring. Second, in highly heteroscedastic settings, CLAPS does not itself learn the input-wise scale, so scale/quantile-based conformalization can be more efficient. Third, σ^2 estimation and backbone expressivity affect $\hat{v}(x)$, making the method sensitive to training choices. Fourth, while a single Cholesky factorization makes deployment light-weight, very wide heads can make $\mathcal{O}(d^3)$ factorization a bottleneck. Finally, our study focuses on marginal coverage under exchangeability; localized guarantees and strong covariate shifts remain

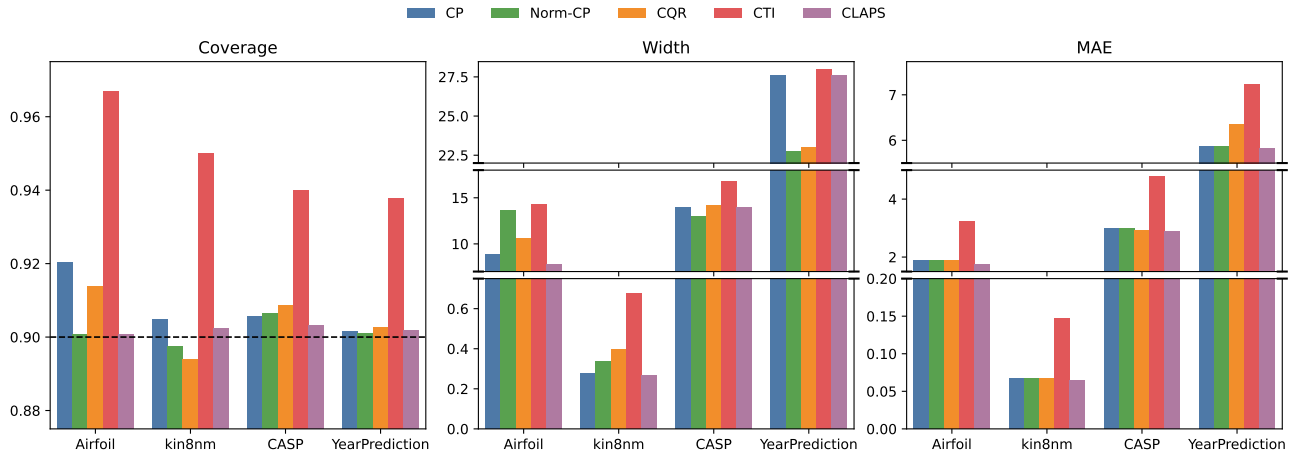

Figure 1. Grouped bar plots for **coverage** (left; dashed target line at 0.90), **width** (middle; broken y-axis to expose low/high ranges), and **MAE** (right; broken y-axis). Each group on the x-axis is a dataset; bars are colored by method.

open. That said, the last-layer view is often a practical sweet spot: with frozen or slowly changing backbones, most train-time variability concentrates in the head, and CLAPS adds a single, stable Cholesky step to standard pipelines with little tuning burden (as reflected by its insensitivity to λ and $\widehat{\sigma}^2$ in our ablations) while retaining exact finite-sample validity.

Future work. A natural extension is to design hybrid conformal scores that interpolate between posterior-aware and scale-aware components, using diagnostics such as the local epistemic share to adapt how much weight is placed on each mechanism. On the posterior side, moving beyond a single last-layer approximation via structured factorizations (e.g., KFAC-style Kronecker approximations, low-rank-plus-diagonal or subspace Laplace variants) and multi-head last-layer posteriors could further improve tail fidelity at scalable cost. On the data side, extending CLAPS to localized coverage under covariate shift and non-exchangeable settings—for example through conformal risk control, covariate-conditioned thresholds, or lightweight online recalibration—would broaden its applicability. Finally, learning the noise scale in a more flexible way (e.g., hierarchical empirical Bayes or auxiliary scale heads) and combining it with posterior- and data-driven diagnostics suggests a plug-and-play conformal system that can automatically choose and tune scoring rules across datasets and sample-size regimes.

7. Conclusion

We introduced CLAPS, a posterior-aligned conformal approach that couples a shared-backbone predictor with a last-layer Laplace view to calibrate uncertainty in a principled and practical way. The method preserves standard conformal guarantees while turning the posterior shape into an informative score, yielding consistently tighter and more faithful intervals in our studies and a clear narrative linking theory and observation. Beyond empirical gains, CLAPS offers a transparent breakdown of predictive uncertainty and a lightweight path to deployment that fits neatly into existing pipelines. We believe this synthesis of Bayesian insight and conformal calibration advances interval estimation for modern neural regressors and opens a direct route to broader extensions in architecture, data modality, and evaluation setting.

Impact Statement

This work combines last-layer Laplace approximation with split conformal calibration to provide more efficient prediction intervals while maintaining the target coverage. The approach can improve decision-making by communicating uncertainty more clearly in applications such as quality control and predictive maintenance. However, in large-scale or highly heterogeneous settings, posterior variance may collapse, producing intervals that are narrower than is appropriate. To mitigate this, we propose two operational guardrails. First, establish a minimum acceptable interval width in consultation with domain experts and automatically switch to a more conservative regime whenever this criterion is not met. Second, monitor coverage and warning signals after deployment, and when anomalies are detected, temporarily defer predictions or route them for human review. We also emphasize data bias and privacy considerations: the released code will include clear usage cautions and licensing, and we recommend domain-specific validation before applying the method in sensitive contexts.

References

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Angelopoulos, A. N., Bates, S., et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Bergamin, F., Moreno-Muñoz, P., Hauberg, S., and Arvanitidis, G. Riemannian laplace approximations for bayesian neural networks. *Advances in Neural Information Processing Systems*, 36:31066–31095, 2023.

Bhattacharyya, A. and Barber, R. F. Group-weighted conformal prediction. *arXiv preprint arXiv:2401.17452*, 2024.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Cinquin, T., Pförtner, M., Fortuin, V., Hennig, P., and Bamler, R. Fsp-laplace: Function-space priors for the laplace approximation in bayesian deep learning. *Advances in Neural Information Processing Systems*, 37: 13897–13926, 2024.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

Gibbs, I. and Candès, E. J. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Guan, L. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Guha, E., Natarajan, S., Möllenhoff, T., Khan, M. E., and Ndiaye, E. Conformal prediction via regression-as-classification. *arXiv preprint arXiv:2404.08168*, 2024.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Harrison, J., Willes, J., and Snoek, J. Variational bayesian last layers. *arXiv preprint arXiv:2404.11599*, 2024.

Hore, R. and Barber, R. F. Conformal prediction with local weights: randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023.

Kiyani, S., Pappas, G. J., and Hassani, H. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024.

Kristiadi, A., Hein, M., and Hennig, P. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Luo, R. and Zhou, Z. Conformal thresholded intervals for efficient regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19216–19223, 2025.

Overman, W., Vallon, J., and Bayati, M. Aligning model properties via conformal risk control. *Advances in Neural Information Processing Systems*, 37:110702–110722, 2024.

Papadopoulos, H., Gammerman, A., and Vovk, V. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pp. 64–69, 2008.

Plassier, V., Fishkov, A., Guizani, M., Panov, M., and Moulines, E. Probabilistic conformal prediction with approximate conditional validity. *arXiv preprint arXiv:2407.01794*, 2024a.

Plassier, V., Kotelevskii, N., Rubashevskii, A., Noskov, F., Velikanov, M., Fishkov, A., Horvath, S., Takac, M., Moulines, E., and Panov, M. Efficient conformal prediction under data heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 4879–4887. PMLR, 2024b.

Qi, S.-a., Yu, Y., and Greiner, R. Toward conditional distribution calibration in survival prediction. *Advances in Neural Information Processing Systems*, 37:86180–86225, 2024.

Qian, X., Wu, J., Wei, L., and Lin, Y. Random projection ensemble conformal prediction for high-dimensional classification. *Chemometrics and Intelligent Laboratory Systems*, 253:105225, 2024.

Rivera, E. O., Patel, Y., and Tewari, A. Conformal prediction for ensembles: Improving efficiency via score-based aggregation. *arXiv preprint arXiv:2405.16246*, 2024.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Xie, R., Barber, R., and Candes, E. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems*, 37:71868–71899, 2024.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Xu, C. and Xie, Y. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023.

Yin, Y. and Carroll, R. J. A diagnostic for heteroscedasticity based on the spearman rank correlation. *Statistics & probability letters*, 10(1):69–76, 1990.

Zhang, Y. and Candès, E. J. Posterior conformal prediction. *arXiv preprint arXiv:2409.19712*, 2024.

A. Numerical Analysis of Test Metrics

Table 2 and Table 3 provide numerical summaries of the figures in the main text. We interpret the results from four perspectives: whether the target coverage (approximately 0.90) is achieved, interval-width efficiency near the same coverage level (smaller is better), point-estimation accuracy (MAE), and estimator stability as reflected by the breadth of the confidence intervals. Overall, CLAPS tends to improve both width and MAE on small- to medium-scale datasets or when heterogeneity is mild, whereas methods that explicitly learn input-dependent scale (NORM-CP, CQR) lead in width efficiency under large scale and pronounced heterogeneity. CTI serves as a conservative baseline, typically yielding higher-than-target coverage with wider intervals.

By dataset, **Airfoil** and **kin8nm** show that CLAPS attains the narrowest intervals and the lowest MAE while remaining near the target coverage (see both tables). In Table 3, the confidence intervals for width and MAE are very tight, indicating stable rankings, and the Wilson intervals for coverage consistently straddle 0.90. On **CASP**, NORM-CP has a slight advantage in width, while CLAPS attains the lowest MAE, revealing a trade-off between interval efficiency and point accuracy. On **YearPrediction**, NORM-CP and CQR achieve the narrowest intervals, highlighting the advantage of explicitly learning input-dependent scale in large datasets. At the same time, CLAPS retains strength in MAE, showing a complementary relationship between interval efficiency and point accuracy.

Additionally, the coverage intervals in Table 3 contract markedly as dataset size grows, with **YearPrediction** exhibiting particularly high estimation stability. Across datasets, CTI consistently displays a trade-off of higher coverage and wider intervals, confirming its role as a safe baseline when high reliability is paramount. In contrast, CLAPS demonstrates Pareto efficiency (narrow width with low MAE) on small- to medium-scale data or when heterogeneity is weak, while NORM-CP/CQR dominate interval efficiency in large-scale or strongly heterogeneous settings. Hence, it is reasonable to select methods based on dataset scale and the strength of heterogeneity signals.

In summary, Table 2 and Table 3 support the following: first, in small- to medium-scale settings with weak heterogeneity, CLAPS maintains target coverage while improving both interval width and MAE. Second, in large-scale settings with strong heterogeneity, NORM-CP and CQR are more efficient in interval width. Third, CTI remains a conservative baseline with higher coverage and wider intervals. These observations align with the diagnostics presented in the main text—posterior variance shrinkage and the strength of heterogeneity signals.

Table 2. Test-set metrics (point estimates for coverage, width, and MAE).

Dataset	Metrics	CP	Norm-CP	CQR	CTI	CLAPS
Airfoil	Cov	0.9205	0.9007	0.9139	0.9669	0.9007
	Wid	8.8468	13.6129	10.5931	14.3088	7.7843
	MAE	1.8832	1.8832	1.8967	3.2301	1.7579
kin8nm	Cov	0.9049	0.8976	0.8939	0.9500	0.9024
	Wid	0.2788	0.3359	0.3973	0.6753	0.2670
	MAE	0.0667	0.0667	0.0677	0.1466	0.0643
CASP	Cov	0.9058	0.9066	0.9086	0.9399	0.9033
	Wid	13.9332	12.9819	14.2050	16.7776	13.9124
	MAE	2.9868	2.9868	2.9193	4.7920	2.8893
YearPrediction	Cov	0.9015	0.9011	0.9027	0.9378	0.9019
	Wid	27.6122	22.7666	23.0094	27.9738	27.5864
	MAE	5.8785	5.8785	6.3651	7.2291	5.8357

B. Posterior-Variance Diagnostics and Heteroscedasticity Signals

Table 4 summarizes the last-layer Laplace decomposition across datasets and splits, reporting the mean epistemic component, the fractional epistemic share r (with mean and empirical 10th/90th percentiles), the mass near collapse $\widehat{\mathbb{P}}(r{<}1\%)$, the posterior covariance trace $\mathrm{tr}(\Sigma)$, and the aleatoric variance σ^2. Two patterns are salient. First, from small/medium data to large-scale data, both r and the epistemic mean shrink substantially while $\mathrm{tr}(\Sigma)$ monotonically decreases, indicating global posterior contraction; this is most pronounced on *YearPrediction*, where r concentrates near zero. Second, σ^2 remains comparatively stable across subsample sizes, implying that the total predictive variance becomes increasingly dominated by

Table 3. 95% confidence intervals for test-set metrics. Coverage uses Wilson intervals; width and MAE use normal (t) approximations.

Dataset	Metrics	CP	Norm-CP	CQR	CTI	CLAPS
Airfoil	Cov	[0.8662, 0.9540]	[0.8426, 0.9389]	[0.8583, 0.9490]	[0.9248, 0.9858]	[0.8426, 0.9389]
	Wid	[8.8468, 8.8468]	[12.5093, 14.7164]	[9.9608, 11.2253]	[13.5943, 15.0234]	[7.7552, 7.8135]
	MAE	[1.5989, 2.1675]	[1.5989, 2.1675]	[1.6096, 2.1838]	[2.8245, 3.6357]	[1.5007, 2.0151]
kin8nm	Cov	[0.8829, 0.9231]	[0.8749, 0.9165]	[0.8710, 0.9132]	[0.9329, 0.9629]	[0.8802, 0.9209]
	Wid	[0.2788, 0.2788]	[0.3279, 0.3439]	[0.3899, 0.4048]	[0.6640, 0.6866]	[0.2670, 0.2671]
	MAE	[0.0627, 0.0707]	[0.0627, 0.0708]	[0.0637, 0.0717]	[0.1389, 0.1542]	[0.0607, 0.0679]
CASP	Cov	[0.8969, 0.9139]	[0.8978, 0.9147]	[0.8999, 0.9166]	[0.9326, 0.9464]	[0.8944, 0.9116]
	Wid	[13.9332, 13.9332]	[12.9006, 13.0633]	[14.1071, 14.3029]	[16.6666, 16.8887]	[13.9110, 13.9137]
	MAE	[2.9086, 3.0650]	[2.9086, 3.0651]	[2.8374, 3.0011]	[4.7281, 4.8558]	[2.8102, 2.9684]
YearPrediction	Cov	[0.8989, 0.9040]	[0.8985, 0.9036]	[0.9001, 0.9052]	[0.9357, 0.9399]	[0.8993, 0.9045]
	Wid	[27.6122, 27.6122]	[22.6983, 22.8349]	[22.9145, 23.1043]	[27.8673, 28.0804]	[27.5863, 27.5865]
	MAE	[5.8238, 5.9331]	[5.8238, 5.9331]	[6.3125, 6.4177]	[7.1772, 7.2810]	[5.7806, 5.8908]

the aleatoric component as data grow.

Table 5 complements this view by quantifying heteroscedasticity signals via Spearman correlation between absolute residuals and the square-root predictive scale. On *YearPrediction* the correlation is clearly positive and statistically significant on both calibration and test splits, signaling strong input-dependent scale that methods explicitly learning the scale (e.g., normalized or quantile-based variants) can exploit for narrower intervals at fixed coverage. In contrast, small/medium datasets (e.g., *Airfoil*, *CASP*) show weak or insignificant signals, consistent with the regimes where our method attains favorable width–MAE trade-offs while meeting target coverage.

Finally, Figure 2 visualizes the same phenomena as epoch-style curves (five interpolated steps per dataset over a log-n grid) for *Epistemic (mean)*, $\mathrm{tr}(\Sigma)$, and $\hat{\sigma}^2$. The left and middle panels reveal a smooth, monotone reduction in epistemic contribution and posterior trace as the effective sample grows, corroborating posterior variance collapse on large-scale data; the right panel shows that the aleatoric estimate varies only mildly by comparison. Taken together, Table 4, Table 5, and Figure 2 provide a coherent diagnostic story: posterior contraction reduces epistemic share with scale, the strength of heteroscedasticity governs interval-efficiency gains from scale-learning approaches, and in small/medium or weak-heterogeneity regimes our approach maintains target coverage with competitive or superior width and point accuracy.

Table 4. LLLA variance decomposition by dataset and split. Columns report the mean epistemic component, the fractional epistemic share $r = \mathrm{epi}/(\sigma^2 + \mathrm{epi})$ (mean and 10/90th percentiles), the fraction of samples with $r < 1\%$, the posterior covariance trace $\mathrm{tr}(\Sigma)$, and the aleatoric variance σ^2.

Dataset	Split	Epistemic (mean)	r (mean)	$\hat{Q}_{0.10}(r)$	$\hat{Q}_{0.90}(r)$	$\widehat{\mathbb{P}}(r < 1\%)$	$\mathrm{tr}(\Sigma)$	σ^2
Airfoil	Calib	0.007292	0.075838	0.049370	0.105913	0.000000	32.587912	0.088094
	Test	0.007931	0.080710	0.047274	0.108744	0.000000	32.587912	0.088094
kin8nm	Calib	0.001572	0.018776	0.012259	0.027264	0.037940	5.269668	0.081985
	Test	0.001548	0.018492	0.011573	0.026858	0.035366	5.269668	0.081985
CASP	Calib	0.002007	0.005018	0.001983	0.008518	0.928086	65.819530	0.396034
	Test	0.002074	0.005168	0.002061	0.008725	0.926088	65.819530	0.396034
YearPrediction	Calib	0.000257	0.000471	0.000159	0.000882	0.999526	56.635682	0.545434
	Test	0.000262	0.000480	0.000158	0.000892	0.999437	56.635682	0.545434

C. Extended Ablations and Sensitivity Analyses

This section provides the full ablation grids that underlie the main text. Table 6 reports, for each dataset (Airfoil, kin8nm, CASP, YearPredictionMSD), the core quantities across regularization levels $\lambda \in \{0.1, 0.3, 1, 3, 10\}$ and two aleatoric estimators $\widehat{\sigma}^2_{\mathrm{EB}}$ and $\widehat{\sigma}^2_{\mathrm{res}}$: empirical coverage (Cov), average interval width (Wid), the selected conformal quantile τ, and the fitted noise level $\widehat{\sigma}^2$. Table 7 complements these results with 95% confidence intervals for coverage and width only,

Table 5. Heteroscedasticity signal: Spearman correlation between $|e|$ and $\sqrt{\hat{v}}$. Two panels report calibration and test splits side by side.

(Calib)

Dataset	ρ	p
Airfoil	-0.0640	0.4592
kin8nm	0.0850	0.0210
CASP	0.0111	0.4746
Year	0.2271	0.0000

(Test)

Dataset	ρ	p
Airfoil	0.0073	0.9290
kin8nm	0.0489	0.1616
CASP	0.0103	0.4841
Year	0.2298	0.0000

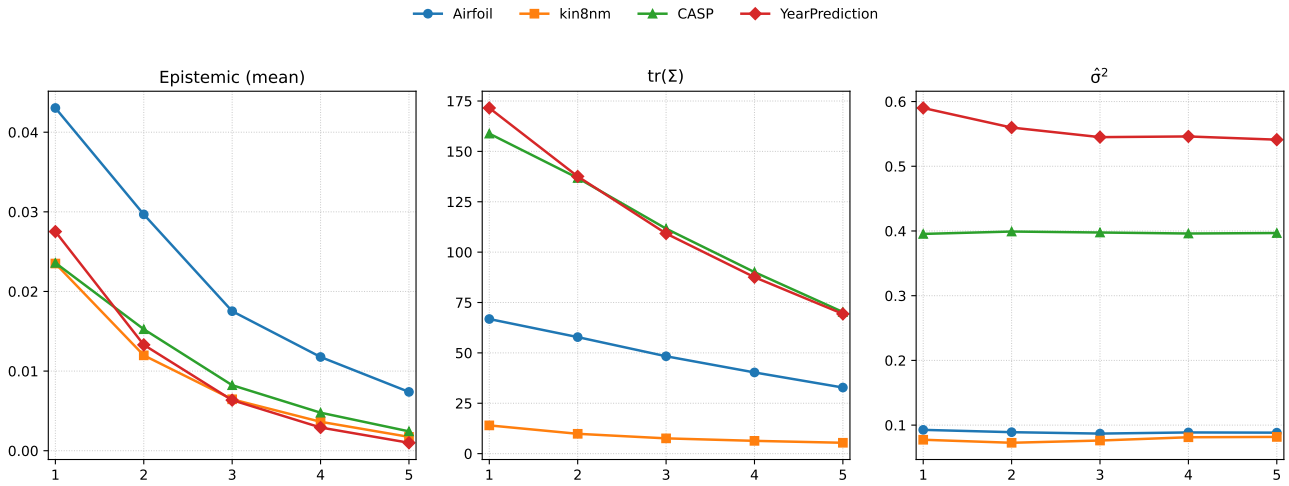

Figure 2. Subsample diagnostics shown as epoch-style curves (1–5, interpolated per dataset on a log-n grid). Panels report *Epistemic (mean)*, $\mathrm{tr}(\Sigma)$, and $\hat{\sigma}^2$ from left to right, with four datasets overlaid in each panel (legend shared at top). This view makes small-dataset trends comparable across datasets with differing raw subsample sizes.

enabling a quantitative assessment of the small empirical differences observed among settings.

Several consistent patterns emerge. First, coverage remains essentially invariant across λ and across the two choices of $\widehat{\sigma}^2$, confirming that the conformal rank rule renders CLAPS insensitive to reasonable regularization and noise-estimation choices. Second, widths change only modestly with λ, while τ adapts smoothly so that nominal coverage is preserved; this monotone adjustment is most visible on Airfoil and CASP. Third, $\widehat{\sigma}^2_{\mathrm{EB}}$ and $\widehat{\sigma}^2_{\mathrm{res}}$ lead to nearly identical operating points across all datasets, suggesting that either can be used as a practical default without sacrificing efficiency. Finally, on the large-scale YearPredictionMSD, the reported dimensionality D of the last-layer features and the calibration size n (annotated in the table) coincide with a regime where the epistemic share is small; the grids illustrate that CLAPS degrades gracefully in such cases, maintaining coverage while producing widths in line with the learned-scale baselines reported in the main text. Together, Tables 6 and 7 document that CLAPS exhibits low tuning burden and predictable behavior across datasets and hyperparameters, while retaining exact finite-sample validity.

D. Theory: Detailed Proofs and Analysis

D.1. Preliminaries and Notation

Conditioning and sources of randomness. we *condition* on the fitted backbone $\phi : \mathcal{X} \to \mathbb{R}^d$ and on the last-layer MAP/posterior obtained from the training set $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{n_{\mathrm{tr}}}$. All probabilities and expectations are thus taken with respect to the joint randomness of the *calibration* multiset $\mathcal{D}_{\mathrm{cal}} = \{(x_i, y_i)\}_{i=1}^{m}$ and an independent *test* pair (x_*, y_*).

LLLA posterior and predictive (restated). With a ridge prior precision $\lambda > 0$ and noise variance $\sigma^2 > 0$, let

$$w \mid \mathcal{D}_{\mathrm{tr}} \sim \mathcal{N}\big(w_{\mathrm{MAP}},\, \Sigma\big), \qquad \Sigma \;=\; \big(\lambda I_d + \sigma^{-2}\Phi^\top \Phi\big)^{-1}, \quad \Phi = \big[\phi(x_i)^\top\big]_{i=1}^{n_{\mathrm{tr}}}.$$

(Airfoil)

λ	Estimator	Cover	Wid	τ	$\widehat{\sigma}^2$
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9007	7.7100	0.036598	0.0879627
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9007	7.7106	0.030127	0.0799660
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	0.8742	7.5557	0.035150	0.0831758
	$\hat{\sigma}^2_{\mathrm{res}}$	0.8742	7.5557	0.035150	0.0831758
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9007	7.7842	0.034182	0.0880946
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9007	7.7842	0.034182	0.0880946
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9007	7.9756	0.033639	0.0932376
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9007	7.9756	0.033639	0.0932376
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9073	8.2366	0.025537	0.0890167
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9073	8.2366	0.031979	0.0989075

(kin8nm)

λ	Estimator	Cov	Wid	τ	$\widehat{\sigma}^2$
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9037	0.2669	0.040228	0.0819326
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9037	0.2669	0.040228	0.0819326
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9049	0.2663	0.040519	0.0819418
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9049	0.2663	0.040519	0.0819418
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9024	0.2670	0.040146	0.0819847
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9024	0.2670	0.040146	0.0819847
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9049	0.2702	0.038544	0.0821850
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9049	0.2702	0.038544	0.0821850
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9061	0.2700	0.039473	0.0831243
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9061	0.2700	0.039473	0.0831243

(CASP)

λ	Estimator	Cov	Wid	τ	$\widehat{\sigma}^2$
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9033	13.8806	0.035899	0.394733
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9033	13.8806	0.035899	0.394733
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9031	13.8899	0.035849	0.395051
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9031	13.8899	0.035849	0.395051
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9033	13.9118	0.035769	0.396034
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9033	13.9118	0.035769	0.396034
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9058	13.9439	0.035740	0.398011
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9058	13.9439	0.035740	0.398011
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9066	13.9214	0.036450	0.401143
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9066	13.9214	0.036450	0.401143

(YearPrediction)

λ	Estimator	Cov	Wid	τ	$\widehat{\sigma}^2$
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9019	27.5707	0.043697	0.545348
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9019	27.5707	0.043697	0.545348
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9017	27.5692	0.043708	0.545369
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9017	27.5692	0.043708	0.545369
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9020	27.5889	0.043604	0.545434
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9020	27.5889	0.043604	0.545434
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9017	27.5727	0.043713	0.545558
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9017	27.5727	0.043713	0.545558
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	0.9016	27.6128	0.043516	0.545810
	$\hat{\sigma}^2_{\mathrm{res}}$	0.9016	27.6128	0.043516	0.545810

Table 6. Ablation summary at $\alpha = 0.10$ across four datasets. Columns report finite-sample coverage, mean interval width, rank-derived threshold τ, and the fitted noise variance $\widehat{\sigma}^2$.

The posterior predictive for any $x \in \mathcal{X}$ is Gaussian

$$y \mid x, \mathcal{D}_{\mathrm{tr}} \sim \mathcal{N}\big(\hat{\mu}(x), \hat{v}(x)\big), \qquad \hat{\mu}(x) = \phi(x)^\top w_{\mathrm{MAP}}, \qquad \hat{v}(x) = \sigma^2 + \phi(x)^\top \Sigma\, \phi(x).$$

Numerical identity (Cholesky form). If $LL^\top = \lambda I_d + \sigma^{-2}\Phi^\top\Phi$ with L lower triangular, then

$$\phi(x)^\top \Sigma\, \phi(x) = \big\|L^{-1}\phi(x)\big\|_2^2,$$

so the epistemic component can be computed via one triangular solve without forming Σ explicitly.

Two-sided posterior CDF score (restated). Let $F_{\mathrm{post}}(\cdot \mid x)$ denote the predictive CDF induced by the LLLA posterior. Define the measurable *centrality* score

$$s(x, y) \;=\; \min\big\{F_{\mathrm{post}}(y \mid x),\, 1 - F_{\mathrm{post}}(y \mid x)\big\} \in [0, \tfrac{1}{2}].$$

In the Gaussian case, with standardized residual $z = (y - \hat{\mu}(x))/\sqrt{\hat{v}(x)}$ and standard normal CDF Φ,

$$s(x, y) = \min\{\Phi(z),\, 1 - \Phi(z)\}.$$

For any threshold $t \in [0, \frac{1}{2}]$, the associated *superlevel (acceptance) set* at covariate x is

$$\mathcal{C}_t(x) \;=\; \{\, y \in \mathbb{R} :\; s(x, y) \geq t \,\}.$$

When $F_{\mathrm{post}}(\cdot \mid x)$ is Gaussian, $\mathcal{C}_t(x)$ is the central quantile interval $\big[\hat{\mu}(x) + \sqrt{\hat{v}(x)}\,\Phi^{-1}(t),\; \hat{\mu}(x) + \sqrt{\hat{v}(x)}\,\Phi^{-1}(1-t)\big]$.

Exchangeability (restated). Assumption 1. Conditional on $\mathcal{D}_{\mathrm{tr}}$, the multiset $\mathcal{D}_{\mathrm{cal}}$ and the test pair (x_*, y_*) are *exchangeable* draws from the same data-generating process. This assumption is only invoked for the marginal validity of split-conformal calibration.

(Airfoil)

λ	Estimator	Cov	Wid
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8426, 0.9389]	[7.6758, 7.7442]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8426, 0.9389]	[7.6763, 7.7449]
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8118, 0.9179]	[7.5237, 7.5876]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8118, 0.9179]	[7.5237, 7.5876]
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8426, 0.9389]	[7.7550, 7.8133]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8426, 0.9389]	[7.7550, 7.8133]
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8426, 0.9389]	[7.9509, 8.0003]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8426, 0.9389]	[7.9509, 8.0003]
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8504, 0.9440]	[8.2170, 8.2562]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8504, 0.9440]	[8.2189, 8.2571]

(kin8nm)

λ	Estimator	Cov	Wid
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8815, 0.9220]	[0.2668, 0.2669]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8815, 0.9220]	[0.2668, 0.2669]
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8829, 0.9231]	[0.2663, 0.2664]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8829, 0.9231]	[0.2663, 0.2664]
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8802, 0.9209]	[0.2670, 0.2671]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8802, 0.9209]	[0.2670, 0.2671]
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8829, 0.9231]	[0.2702, 0.2703]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8829, 0.9231]	[0.2702, 0.2703]
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8842, 0.9242]	[0.2699, 0.2700]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8842, 0.9242]	[0.2699, 0.2700]

(CASP)

λ	Estimator	Cov	Wid
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8944, 0.9116]	[13.8791, 13.8821]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8944, 0.9116]	[13.8791, 13.8821]
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8942, 0.9114]	[13.8884, 13.8914]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8942, 0.9114]	[13.8884, 13.8914]
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8944, 0.9116]	[13.9105, 13.9132]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8944, 0.9116]	[13.9105, 13.9132]
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8969, 0.9139]	[13.9428, 13.9451]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8969, 0.9139]	[13.9428, 13.9451]
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8978, 0.9147]	[13.9206, 13.9222]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8978, 0.9147]	[13.9206, 13.9222]

(YearPrediction)

λ	Estimator	Cov	Wid
0.1	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8993, 0.9044]	[27.5706, 27.5708]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8993, 0.9044]	[27.5706, 27.5708]
0.3	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8991, 0.9042]	[27.5692, 27.5693]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8991, 0.9042]	[27.5692, 27.5693]
1.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8994, 0.9045]	[27.5888, 27.5890]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8994, 0.9045]	[27.5888, 27.5890]
3.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8991, 0.9042]	[27.5726, 27.5727]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8991, 0.9042]	[27.5726, 27.5727]
10.0	$\hat{\sigma}^2_{\mathrm{EB}}$	[0.8990, 0.9042]	[27.6128, 27.6129]
	$\hat{\sigma}^2_{\mathrm{res}}$	[0.8990, 0.9042]	[27.6128, 27.6129]

Table 7. 95% confidence intervals for finite-sample coverage and mean width at $\alpha = 0.10$ across datasets. This table provides the confidence intervals for Coverage and Width.

Split-conformal quantile and rank rule. Let $m = |\mathcal{D}_{\mathrm{cal}}|$ and form the calibration scores $\{s_i\}_{i=1}^m = \{s(x_i, y_i)\}_{(x_i,y_i)\in\mathcal{D}_{\mathrm{cal}}}$. Define the conformal threshold t as the empirical quantile at level $\alpha \in (0,1)$ using the standard rank (order-statistic) rule

$$t \;=\; \mathrm{Quantile}_\alpha\big(\{s_i\}_{i=1}^m\big) \;:=\; s_{(k)}, \qquad k \;=\; \Big\lceil (m+1)\,\alpha \Big\rceil,$$

with any deterministic or randomized tie-breaking; all proofs below are agnostic to the particular tie-breaking scheme provided it is measurable. The *CLAPS* prediction set at covariate x is then $\mathcal{C}(x) := \mathcal{C}_t(x)$ as above.

Notation. We write $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ for probability and expectation over the randomness of $(\mathcal{D}_{\mathrm{cal}}, x_*, y_*)$ under Assumption 1. The Lebesgue measure (length) of an interval $A \subset \mathbb{R}$ is denoted by $|A|$. All asymptotics ($n_{\mathrm{tr}} \to \infty$) keep the backbone architecture fixed unless explicitly stated.

D.2. Proposition 4.4: Finite-sample marginal coverage

Proof. Fix $\mathcal{D}_{\mathrm{tr}}$ and the fitted backbone/posterior so that the score $s(x,y) = \min\{F_{\mathrm{post}}(y \mid x),\, 1 - F_{\mathrm{post}}(y \mid x)\} \in [0, \frac{1}{2}]$ is a measurable function of (x,y). Let $m = |\mathcal{D}_{\mathrm{cal}}|$ and form calibration scores $s_i := s(x_i, y_i)$ for $(x_i, y_i) \in \mathcal{D}_{\mathrm{cal}}$, with ascending order statistics $s_{(1)} \le \cdots \le s_{(m)}$. For a fixed $\alpha \in (0,1)$ define

$$k \;=\; \Big\lceil (m+1)\,\alpha \Big\rceil, \qquad t \;=\; s_{(k)}, \qquad \mathcal{C}(x) \;=\; \{y :\; s(x,y) \ge t\}.$$

(Any measurable tie-breaking in the definition of $s_{(k)}$ is allowed; the argument below is agnostic to the specific choice.)

Consider the *augmented* $(m+1)$ scores consisting of the m calibration scores and the independent test score $s_* := s(x_*, y_*)$. By Assumption 1 (exchangeability of (x_*, y_*) with the calibration pairs conditional on $\mathcal{D}_{\mathrm{tr}}$), the multiset $\{s_1, \ldots, s_m, s_*\}$ is exchangeable. Let R denote the (randomized, measurable) *rank* of s_* among these $(m+1)$ scores in ascending order (ties are broken by an independent uniform randomization or any fixed measurable rule). Exchangeability implies

$$\mathbb{P}\big\{R = r \,\big|\, \mathcal{D}_{\mathrm{tr}}\big\} \;=\; \tfrac{1}{m+1} \quad \text{for each } r \in \{1, \ldots, m+1\}.$$

We claim that the *acceptance event* $\{y_* \in \mathcal{C}(x_*)\}$ coincides with $\{s_* \geq s_{(k)}\}$ and is contained in $\{R \geq k\}$: indeed, if $R \leq k-1$ then at least $k-1$ calibration scores are $\leq s_*$, forcing $s_* < s_{(k)}$ and hence $y_* \notin \mathcal{C}(x_*)$; conversely, if $R \geq k$ then $s_* \geq s_{(k)}$ so $y_* \in \mathcal{C}(x_*)$. Therefore,

$$\mathbb{P}\{y_* \in \mathcal{C}(x_*) \mid \mathcal{D}_{\mathrm{tr}}\} \;\geq\; \mathbb{P}\{R \geq k \mid \mathcal{D}_{\mathrm{tr}}\} \;=\; \frac{m+2-k}{m+1}.$$

Using $k = \lceil (m+1)\alpha \rceil \leq (m+1)\alpha + 1$ gives

$$\frac{m+2-k}{m+1} \;\geq\; \frac{m+2-((m+1)\alpha+1)}{m+1} \;=\; 1-\alpha.$$

Thus,

$$\mathbb{P}\{y_* \in \mathcal{C}(x_*)\} \;=\; \mathbb{E}[\mathbb{P}\{y_* \in \mathcal{C}(x_*) \mid \mathcal{D}_{\mathrm{tr}}\}] \;\geq\; 1-\alpha,$$

which establishes the stated *marginal* coverage over the calibration set and the test point, conditional on $\mathcal{D}_{\mathrm{tr}}$. The proof only requires measurability of the score and exchangeability; the specific form of s (posterior-aware versus residual-based) affects *efficiency* (interval width) but not *validity*. □

D.3. Lemma 4.5: Monotone transform of $|z|$

Proof. Fix x and write the standardized residual $z = (y - \hat{\mu}(x))/\sqrt{\hat{v}(x)}$. In the Gaussian case, $F_{\mathrm{post}}(y \mid x) = \Phi(z)$, so by definition

$$s(x,y) = \min\{\Phi(z),\, 1-\Phi(z)\}.$$

Define $g : [0,\infty) \to [0, \frac{1}{2}]$ by $g(u) := \min\{\Phi(u),\, 1-\Phi(u)\}$. Because Φ is strictly increasing on $\mathbb{R}$ with $\Phi(0) = \frac{1}{2}$ and $\Phi(-u) = 1-\Phi(u)$, we have

$$s(x,y) = \min\{\Phi(z),\, 1-\Phi(z)\} = \min\{\Phi(|z|),\, 1-\Phi(|z|)\} = g(|z|).$$

Strict monotonicity follows since Φ is strictly increasing: for $0 \leq u_1 < u_2$,

$$g(u_2) = 1-\Phi(u_2) \;<\; 1-\Phi(u_1) = g(u_1).$$

Hence g is strictly decreasing on $[0,\infty)$ and $s(x,y)$ depends on y only through $|z|$.

Nesting of superlevel sets. For any $t \in [0, \frac{1}{2}]$, the superlevel set at covariate x is

$$\mathcal{C}_t(x) = \{\, y :\; s(x,y) \geq t \,\} = \{\, y :\; g(|z|) \geq t \,\} = \{\, y :\; |z| \leq g^{-1}(t) \,\},$$

where $g^{-1} : [0, \frac{1}{2}] \to [0,\infty)$ is the (well-defined) inverse of the strictly decreasing g. Therefore $\mathcal{C}_t(x)$ is a (possibly degenerate) *central* interval

$$\mathcal{C}_t(x) = \Big[\hat{\mu}(x) - \sqrt{\hat{v}(x)}\, g^{-1}(t),\; \hat{\mu}(x) + \sqrt{\hat{v}(x)}\, g^{-1}(t)\Big],$$

which is symmetric about $\hat{\mu}(x)$. Since $t_1 < t_2 \Rightarrow g^{-1}(t_1) > g^{-1}(t_2)$, we have $\mathcal{C}_{t_2}(x) \subset \mathcal{C}_{t_1}(x)$; i.e., $\{\mathcal{C}_t(x)\}_t$ is *nested* in t. □

D.4. Corollary 4.6: Equivalence to central/HPD sets

Proof. Fix x and write the Gaussian predictive density and CDF as

$$f_x(y) = \frac{1}{\sqrt{2\pi\hat{v}(x)}} \exp\Big(-\tfrac{(y-\hat{\mu}(x))^2}{2\hat{v}(x)}\Big), \qquad F_{\mathrm{post}}(y \mid x) = \Phi\Big(\tfrac{y-\hat{\mu}(x)}{\sqrt{\hat{v}(x)}}\Big).$$

Set $z = (y-\hat{\mu}(x))/\sqrt{\hat{v}(x)}$ and recall from Lemma 4.5 that $s(x,y) = g(|z|)$ with $g(u) = \min\{\Phi(u), 1-\Phi(u)\} = 1-\Phi(u)$ for $u \geq 0$, strictly decreasing. Hence, for any $t \in [0, \frac{1}{2}]$,

$$\{y :\; s(x,y) \geq t\} = \{y :\; |z| \leq g^{-1}(t)\} = \Big[\hat{\mu}(x) - \sqrt{\hat{v}(x)}\, a_t,\; \hat{\mu}(x) + \sqrt{\hat{v}(x)}\, a_t\Big],$$

where $a_t := g^{-1}(t) = \Phi^{-1}(1-t)$. The posterior mass of this set is

$$\int_{\hat{\mu}-\sqrt{\hat{v}}\,a_t}^{\hat{\mu}+\sqrt{\hat{v}}\,a_t} f_x(y)\,dy = \Phi(a_t) - \Phi(-a_t) = 1 - 2\{1-\Phi(a_t)\} = 1-2t.$$

Thus the superlevel set $\{s \geq t\}$ is the *central* interval that carries mass $1-2t$.

We now show this interval is also the *shortest* (HPD) set of that mass. Since $f_x(y)$ is symmetric and strictly decreasing in $|y-\hat{\mu}(x)|$, every upper level set $\{y:\ f_x(y) \geq c\}$ is an interval $[\hat{\mu} - \sqrt{\hat{v}}\,a,\ \hat{\mu}+\sqrt{\hat{v}}\,a]$ for a unique $a \geq 0$, determined by $f_x(\hat{\mu} \pm \sqrt{\hat{v}}\,a) = c$. Choosing $c = c_t := f_x(\hat{\mu}+\sqrt{\hat{v}}\,a_t)$ yields the same $a = a_t$ as above, and the corresponding level set has posterior mass $\int_{\hat{\mu}-\sqrt{\hat{v}}\,a_t}^{\hat{\mu}+\sqrt{\hat{v}}\,a_t} f_x = 1-2t$.

Finally, among all measurable sets $A \subset \mathbb{R}$ with $\int_A f_x = 1-2t$, the upper level set $\{f_x \geq c_t\}$ minimizes Lebesgue measure $|A|$: this is the defining property of HPD sets and follows from the rearrangement/"layer-cake" principle for unimodal, radially decreasing densities. Therefore $\{y:\ s(x,y) \geq t\}$ coincides with the shortest (HPD) interval of posterior mass $1-2t$. □

D.5. Theorem 4.7: Oracle efficiency under correct specification

Proof. Fix x and suppose the posterior predictive $F_{\text{post}}(\cdot \mid x)$ equals the *true* conditional distribution with density f_x (hence the Gaussian form stated in the preliminaries). Let $\mathcal{F} = \{\mathcal{C}_t(x)\}_{t\in[0,1/2]}$ denote the CLAPS superlevel sets $\mathcal{C}_t(x) = \{y:\ s(x,y) \geq t\}$, and let $\mathcal{G} = \{\mathcal{A}_u(x)\}_{u\in\mathcal{U}}$ be the superlevel sets induced by any other score whose acceptance sets are *nested* in its threshold u.

Step 1: Shortest set at fixed mass is the HPD set. For any target mass $q \in (0,1)$, define the *HPD* (highest posterior density) set

$$\mathcal{H}_q(x) \ = \ \{y:\ f_x(y) \geq c_q\}, \quad \text{where } c_q \text{ is chosen so that } \int_{\mathcal{H}_q(x)} f_x(y)\,dy = q.$$

By the standard rearrangement (layer-cake) argument, $\mathcal{H}_q(x)$ minimizes the Lebesgue measure among all measurable $A \subset \mathbb{R}$ with $\int_A f_x = q$:

$$|A| \ \geq \ |\mathcal{H}_q(x)| \qquad \text{whenever} \quad \int_A f_x = q. \tag{1}$$

For the (symmetric, unimodal) Gaussian f_x, $\mathcal{H}_q(x)$ is the *central* interval $[\hat{\mu}(x) - a_q\sqrt{\hat{v}(x)},\ \hat{\mu}(x)+a_q\sqrt{\hat{v}(x)}]$ for the unique $a_q > 0$ with $\Phi(a_q) - \Phi(-a_q) = q$.

Step 2: CLAPS family coincides with HPD family. By Lemma 4.5, $s(x,y) = g(|z|)$ with g strictly decreasing, so $\mathcal{C}_t(x) = \{\,|z| \leq g^{-1}(t)\,\}$ is a central interval. By Corollary 4.6, for each t the mass of $\mathcal{C}_t(x)$ equals $q(t) = 1-2t$ and $\mathcal{C}_t(x) = \mathcal{H}_{q(t)}(x)$. Thus the CLAPS family $\mathcal{F}$ is exactly the HPD family $\{\mathcal{H}_q(x)\}_{q\in(0,1)}$ under a reparameterization $q = 1-2t$.

Step 3: Oracle efficiency under a coverage constraint. Consider any acceptance set $\mathcal{A}_u(x) \in \mathcal{G}$ whose conditional coverage satisfies $\Pr\{y \in \mathcal{A}_u(x) \mid x\} = \int_{\mathcal{A}_u(x)} f_x \geq 1-\alpha$. Let $q_u(x) := \int_{\mathcal{A}_u(x)} f_x$. By (1) and the monotonicity of the length of central/HPD intervals in q,

$$|\mathcal{A}_u(x)| \ \geq \ |\mathcal{H}_{q_u(x)}(x)| \ \geq \ |\mathcal{H}_{1-\alpha}(x)|.$$

Pick the CLAPS threshold $t^\star = \alpha/2$ so that $q(t^\star) = 1-2t^\star = 1-\alpha$; then $\mathcal{C}_{t^\star}(x) = \mathcal{H}_{1-\alpha}(x)$ and hence

$$|\mathcal{C}_{t^\star}(x)| \ = \ |\mathcal{H}_{1-\alpha}(x)| \ \leq \ |\mathcal{A}_u(x)| \qquad \text{for any } \mathcal{A}_u(x) \in \mathcal{G} \text{ with } \int_{\mathcal{A}_u(x)} f_x \geq 1-\alpha.$$

This establishes the oracle (population) statement: among all nested-score families, CLAPS attains the *minimal* Lebesgue length at fixed conditional coverage $1-\alpha$.

Step 4: Compatibility with conformal calibration. Split-conformal calibration selects a data-dependent threshold while preserving the nesting of the family $\{\mathcal{C}_t(x)\}_t$. At any realized threshold t (hence realized conditional mass $q(t)$), the argument above applies with $q(t)$ in place of $1-\alpha$, showing that—pointwise in x and for any t—the CLAPS set is the shortest among all nested-score acceptance sets that achieve at least the same conditional mass. Since calibration only *transfers* the population probability constraint to finite-sample marginal validity without altering nesting, the efficiency dominance of CLAPS persists. □

D.6. Theorem 4.9: Posterior contraction and regime change

Proof. Write $\Sigma_n = (\lambda I_d + \sigma^{-2}\Phi_n^\top \Phi_n)^{-1}$ for the posterior covariance built from the $n = n_{\mathrm{tr}}$ training points, where $\Phi_n = [\phi(x_i)^\top]_{i=1}^n \in \mathbb{R}^{n\times d}$ and d is the (fixed) backbone width. Assume throughout that $\mathbb{E}[\phi(x)\phi(x)^\top] =: \Gamma$ exists with $\lambda_{\min}(\Gamma) > 0$ and that $\|\phi(x)\|_2$ has finite second moment (e.g., sub-Gaussian or bounded features suffice).

Step 1: Spectral growth of $\Phi_n^\top \Phi_n$. By the strong law of large numbers,

$$\frac{1}{n}\,\Phi_n^\top \Phi_n \;\xrightarrow[n\to\infty]{\text{a.s.}}\; \Gamma.$$

Hence, almost surely,

$$\lambda_{\min}(\Phi_n^\top \Phi_n) \;=\; n\,\lambda_{\min}\big(\tfrac{1}{n}\Phi_n^\top \Phi_n\big) \;\geq\; n\,\frac{\lambda_{\min}(\Gamma)}{2} \qquad \text{for all sufficiently large } n.$$

Consequently,

$$\lambda_{\min}\big(\lambda I_d + \sigma^{-2}\Phi_n^\top \Phi_n\big) \;\geq\; \lambda + \frac{\lambda_{\min}(\Gamma)}{2\sigma^2}\,n \;\xrightarrow{n\to\infty}\; \infty,$$

so the operator norm of the inverse vanishes:

$$\|\Sigma_n\|_{\mathrm{op}} = \big\|(\lambda I_d + \sigma^{-2}\Phi_n^\top \Phi_n)^{-1}\big\|_{\mathrm{op}} = \frac{1}{\lambda_{\min}(\lambda I_d + \sigma^{-2}\Phi_n^\top \Phi_n)} \;\xrightarrow{n\to\infty}\; 0.$$

Step 2: Pointwise collapse of the epistemic component. For any fixed x, let $\varphi = \phi(x) \in \mathbb{R}^d$. Then

$$\mathrm{epi}_n(x) \;:=\; \varphi^\top \Sigma_n \varphi \;\leq\; \|\Sigma_n\|_{\mathrm{op}}\,\|\varphi\|_2^2 \;\xrightarrow{n\to\infty}\; 0,$$

almost surely (or in probability), provided $\|\varphi\|_2 < \infty$. Therefore the predictive variance

$$\hat{v}_n(x) \;=\; \sigma^2 + \mathrm{epi}_n(x) \;\xrightarrow{n\to\infty}\; \sigma^2.$$

Equivalently, the fractional epistemic share $r_n(x) = \mathrm{epi}_n(x)/\hat{v}_n(x) \to 0$.

Step 3: Interpretation (regime change). As the last-layer posterior covariance collapses ($\Sigma_n \to 0$ in operator norm), the CLAPS acceptance sets $\mathcal{C}_t(x)$—which depend on x through $z = (y - \hat{\mu}(x))/\sqrt{\hat{v}_n(x)}$—converge to the *homoscedastic* Gaussian central intervals with variance σ^2. Thus any width advantage of posterior-aware scoring that relies on a nontrivial epistemic component vanishes in this regime; efficiency then depends on how well the score captures residual *heteroscedasticity*. If the conditional noise is genuinely heteroscedastic while $\hat{v}_n(x) \to \sigma^2$ is constant, scores that *learn scale* (e.g., normalized-CP/CQR) can become more informative.

Scope of the claim. The conclusion holds for fixed backbone width d and nondegenerate feature covariance $\Gamma \succ 0$. If d grows with n in such a way that $\lambda_{\min}(\frac{1}{n}\Phi_n^\top \Phi_n)$ fails to be bounded away from 0, or if features are degenerate ($\Gamma \succeq 0$ with zero eigenvalues along directions used by $\phi(x)$), then $\|\Sigma_n\|_{\mathrm{op}} \to 0$ need not hold uniformly, and residual epistemic variance may persist. □

D.7. Remarks on Misspecification and Diagnostics

Misspecification and efficiency (recap). The marginal validity in Proposition 4.4 only requires measurability of $s(x, y)$ and exchangeability of $(\mathcal{D}_{\mathrm{cal}}, x_*, y_*)$. Hence, even when the posterior predictive $F_{\mathrm{post}}(\cdot \mid x)$ is misspecified, CLAPS

remains valid. What changes is *efficiency* (interval length): when F_{post} retains informative shape beyond point residuals (e.g., tail thickness or local curvature encoded via $z = (y - \hat{\mu})/\sqrt{\hat{v}}$), thresholding the two-sided CDF can yield shorter sets than residual-only scores; when the posterior variance effectively collapses to a constant (Theorem 4.9), residual/scale-learning scores may dominate.

Diagnostics that connect theory to data. We monitor three quantities computed from the fitted LLLA head (no extra training):

$$\text{epi}(x) = \phi(x)^\top \Sigma \phi(x), \qquad r(x) = \frac{\text{epi}(x)}{\sigma^2 + \text{epi}(x)}, \qquad \text{tr}(\Sigma).$$

Here $\text{epi}(x)$ is the last-layer epistemic component, $r(x) \in [0, 1)$ its fractional share in the predictive variance, and $\text{tr}(\Sigma)$ a global summary of posterior uncertainty. In addition, we assess the heteroscedastic signal via the rank correlation

$$\rho \;=\; \text{Spearman}\big(\, |y - \hat{\mu}(x)|,\; \sqrt{\hat{v}(x)}\,\big),$$

computed on a held-out set (or across calibration folds). These diagnostics instantiate the regimes described by Theorems 4.7 and 4.9:

- *Posterior-informative regime.* If $r(x)$ has nontrivial mass away from 0 (e.g., median or upper quantiles clearly positive) and $\text{tr}(\Sigma)$ is not negligible, then z scores meaningfully reflect local posterior shape; CLAPS tends to realize HPD-like efficiency (Theorem 4.7).
- *Contraction regime.* If subsample curves show $\text{epi}(x) \downarrow$ and $\text{tr}(\Sigma) \downarrow$ rapidly with n, and the distribution of $r(x)$ concentrates near 0, then $\hat{v}(x) \approx \sigma^2$ and CLAPS reduces to conformalized homoscedastic intervals; width advantages vanish (Theorem 4.9).
- *Heteroscedastic signal.* A substantially positive ρ indicates that $\sqrt{\hat{v}(x)}$ tracks the scale of residuals across x; if this signal stems primarily from aleatoric structure (while $r \approx 0$), methods that explicitly *learn* scale (normalized-CP, CQR) can yield tighter intervals than posterior-aware scoring.

Practical reading guide. In reporting diagnostics, we recommend (i) the empirical distribution of $r(x)$ (mean/median/deciles and the fraction below a small threshold), (ii) $\text{tr}(\Sigma)$ and its subsample curve versus n, and (iii) ρ with a p-value or confidence band. Together these summarize whether efficiency is more likely to come from posterior shape (favoring CLAPS) or from explicit scale learning (favoring normalized-CP/CQR), without affecting the coverage guarantee itself.

D.8. Proposition 4.11: A simple selection rule

Proposition D.1 (Diagnostics-driven choice of scoring rule)**.** *Fix a fitted backbone and LLLA head. Let*

$$\text{epi}(x) = \phi(x)^\top \Sigma \phi(x), \quad r(x) = \frac{\text{epi}(x)}{\sigma^2 + \text{epi}(x)}, \quad T := \text{tr}(\Sigma), \quad \rho := \text{Spearman}\big(|y - \hat{\mu}(x)|, \sqrt{\hat{v}(x)}\big).$$

Consider the following rule with user-chosen small constants $\varepsilon_r, \varepsilon_T, \tau_\rho > 0$:

$$\begin{aligned} \textsc{Select-CLAPS} &\iff \text{Median}\{r(x)\} > \varepsilon_r \;\; \textit{and} \;\; T > \varepsilon_T \;\; \textit{and} \;\; \rho \le \tau_\rho, \\ \textsc{Select-Scale} &\iff \text{Median}\{r(x)\} \le \varepsilon_r \;\; \textit{and} \;\; \rho > \tau_\rho, \end{aligned}$$

where Select-Scale *denotes a scale-learning conformal method (e.g., normalized-CP or CQR). Then, under the regimes characterized by Theorems 4.7 and 4.9, the rule aligns with efficiency: it selects CLAPS when posterior shape is informative and selects a scale-learning score when last-layer epistemic variability collapses but heteroscedastic signal is present.*

Justification. (*Posterior-informative regime.*) If T is non-negligible and the distribution of $r(x)$ places nontrivial mass away from 0 (e.g., median $> \varepsilon_r$), then the standardized residual $z = (y - \hat{\mu})/\sqrt{\hat{v}}$ encodes *posterior shape* beyond raw residual magnitude. By Lemma 4.5 and Corollary 4.6, CLAPS superlevel sets coincide with HPD/central sets of the (approximately correct) predictive and, by Theorem 4.7, are shortest among nested-score families at fixed mass. A small ρ

indicates that *aleatoric* scale learning offers limited additional benefit; thus CLAPS is expected to be at least as tight as residual/scale-based alternatives.

(*Contraction regime with heteroscedasticity.*) If T is small and $\mathrm{Median}\{r(x)\} \approx 0$, Theorem 4.9 implies $\hat{v}(x) \to \sigma^2$ and the posterior-aware score loses its x–dependent variability. When, in addition, $\rho > \tau_\rho$, the residual magnitude correlates with $\sqrt{\hat{v}(x)}$ due to *aleatoric* structure not captured by the (nearly constant) LLLA variance. Methods that explicitly *learn* scale (normalized-CP, CQR) can then allocate interval width more efficiently across x, yielding narrower sets at the same coverage.

(*Remarks on thresholds.*) The constants $(\varepsilon_r, \varepsilon_T, \tau_\rho)$ are tuning knobs that define "nontrivial" posterior variance and "substantial" heteroscedastic signal; in practice one may use dataset-specific quantiles (e.g., ε_r as the 10th–20th percentile of $r(x)$ on calibration) rather than fixed numeric cutoffs. The rule is heuristic but *consistent* with the structural guarantees in Theorem 4.7 and Theorem 4.9. □