DIFFICULTY-CALIBRATED CONFORMAL PREDICTION FOR TRANSFER LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art machine learning models, including foundation models, are predominantly non-Bayesian, and reliable methods for quantifying their predictive uncertainty remain underdeveloped. This challenge becomes particularly important in transfer learning, where models are adapted to downstream tasks with limited data. Conformal prediction (CP) provides model-agnostic uncertainty quantification with finite-sample coverage guarantees, but conventional CP does not exploit uncertainty-related information acquired during pretraining. We introduce Difficulty-Calibrated Conformal Prediction (DC-CP), a framework that transfers source-derived task difficulty to target-task calibration. DC-CP requires neither paired source–target observations nor a shared output space and learns only a one-dimensional link between source difficulty and target prediction error from limited target data. With exact covariate-shift weights, DC-CP combined with weighted CP retains finite-sample marginal coverage under distribution shift while adapting uncertainty estimates to input difficulty. Experiments on two open-domain question-answering benchmarks for large language models and a simulation-to-real polymer science benchmark demonstrate effective difficulty transfer across tasks and domains. Our results establish a model-agnostic framework for transferring not only predictive knowledge but also uncertainty-related information to downstream tasks.

1 INTRODUCTION

The predictive and reasoning capabilities of machine learning have advanced rapidly, driven by the growth of training data and innovations in learning algorithms (LeCun et al., 2015; Kingma & Ba, 2015; Kaplan et al., 2020) and model architectures (He et al., 2016; Vaswani et al., 2017). A particularly important development has been the emergence of general-purpose and domain-specific foundation models pretrained on massive datasets (Bommasani et al., 2021; Brown et al., 2020; Lin et al., 2023). These models not only exhibit strong zero-shot performance but can also be adapted to diverse downstream tasks with limited task-specific data. Recent studies have further demonstrated scaling laws in transfer learning, where the generalization performance of transferred models improves with increasing pretraining data, following a power law (Hernandez et al., 2021; Mikami et al., 2023; Minami et al., 2025; Yoshida et al., 2025). As foundation models continue to scale, adapting pretrained models rather than building models from scratch is likely to become an increasingly standard paradigm.

Despite this progress, reliable uncertainty quantification for transferred models remains an unresolved challenge. Most foundation models and their fine-tuned variants are based on non-Bayesian neural networks and do not inherently provide calibrated predictive uncertainty. Methods such as Monte Carlo dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), quantile regression (Koenker & Bassett, 1978; Tagasovska & Lopez-Paz, 2019), and heteroscedastic neural networks (Nix & Weigend, 1994; Kendall & Gal, 2017) have been explored, but they depend on specific model architectures or training procedures and generally lack finite-sample statistical guarantees.

Conformal prediction (CP) offers a model-agnostic alternative (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Using prediction errors or nonconformity scores from a held-out calibration dataset, CP constructs prediction sets or intervals that achieve a prescribed marginal

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.

coverage under exchangeability, typically an i.i.d. assumption (Vovk et al., 2005; Lei et al., 2018). Weighted conformal prediction (WCP) (Tibshirani et al., 2019), including its extensions to counterfactual inference (Lei & Candès, 2021) and feedback covariate shift (Fannjiang et al., 2022), further extends CP to covariate shift by weighting calibration samples according to the density ratio between source and target input distributions, thereby restoring marginal coverage under covariate shift.

However, conformal uncertainty quantification remains underdeveloped for transfer learning. A fine-tuned model inherits predictive knowledge or features from its pretrained model, whereas conventional CP applied to target-domain calibration data does not exploit uncertainty information acquired during pretraining. This raises a fundamental question: can uncertainty information be transferred together with predictive knowledge and used to improve uncertainty quantification after fine-tuning?

Here, we develop a general CP framework, DC-CP (Difficulty-Calibrated Conformal Prediction), for transfer learning that transfers a source-derived scalar difficulty function to conformal calibration on a target task. Without using source-target paired data, DC-CP learns only a one-dimensional link between source difficulty and target prediction error scale and uses this link to obtain target nonconformity scores. With exact covariate-shift weights, combining DC-CP with WCP retains the usual finite-sample marginal coverage guarantee while allowing calibration to adapt to input difficulty.

DC-CP places minimal requirements on the transfer-learning pipeline. It can be applied to arbitrary pretrained and fine-tuned models without modifying their architectures or training objectives, requires neither paired source–target observations nor a shared output space, and reduces target-side adaptation to learning a one-dimensional relationship from a small target dataset. We demonstrate difficulty transfer on two open-domain question-answering benchmarks (Kwiatkowski et al., 2019; Lee et al., 2019; Joshi et al., 2017) for large language models (LLMs) and on a simulation-to-real benchmark in polymer materials science (Anonymous, 2026; Yoshida et al., 2025). Across these distinct settings, our results provide a general route toward transferring not only predictive knowledge but also uncertainty information from pretrained models to downstream tasks, establishing a model-agnostic uncertainty quantification framework for the foundation-model era.

2 Methods

2.1 Weighted Conformal Prediction

WCP extends conformal prediction to covariate shift, where the conditional distribution $P(Y|X)$ is unchanged but the input distribution differs between the calibration and test domains. Given a calibration set $\{(X_i, Y_i)\}_{i=1}^n$ drawn i.i.d. from $P_{\mathrm{cal},X} \otimes P_{Y|X}$, a test point $(X_{n+1}, Y_{n+1}) \sim P_{\mathrm{test},X} \otimes P_{Y|X}$ and a predictive model μ, we compute the nonconformity scores $V_i = V(X_i, Y_i; \mu)$; for regression, we use $V_i = |Y_i - \mu(X_i)|$. Under covariate shift, each calibration sample is weighted by the density ratio $w(x) = dP_{\mathrm{test},X}(x)/dP_{\mathrm{cal},X}(x)$ (known, or estimated by a logistic discriminator in our experiments).

For a test input x, the normalized weights are defined as

$$p_i(x) = \begin{cases} \dfrac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)}, & i = 1, \ldots, n, \\ \dfrac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}, & i = n+1. \end{cases} \tag{1}$$

The WCP threshold $\hat{q}_{1-\alpha}(x)$ is the $(1-\alpha)$-quantile, for a miscoverage level $\alpha \in (0, 1)$, of the weighted empirical distribution of the scores, yielding the prediction set

$$\begin{aligned} C(x) &= \{y : V(x, y; \mu) \le \hat{q}_{1-\alpha}(x)\}, \\ \hat{q}_{1-\alpha}(x) &= \mathrm{Quantile}_{1-\alpha}\Big(\sum_{i=1}^n p_i(x)\, \delta_{V_i} + p_{n+1}(x)\, \delta_{+\infty}\Big), \end{aligned} \tag{2}$$

where δ_v is the point mass at v. A hat marks a finite-sample quantity: $\hat{q}_{1-\alpha}(x)$ is the empirical weighted quantile computed from the n calibration scores, as opposed to the population quantiles $q_{1-\alpha}^M$ of Section 2.3.

Under covariate shift and exact density-ratio weights, WCP provides the finite-sample marginal coverage guarantee $\mathbb{P}\{Y \in C(X)\} \geq 1-\alpha$. Thus, WCP allows conformal calibration to account for source–target distributional shift while remaining independent of the underlying predictive model. We work throughout in this inductive (split) conformal setting (Papadopoulos et al., 2002).

2.2 Difficulty-Calibrated Conformal Prediction

DC-CP transfers input difficulty from a source task to conformal calibration on a target task. We write Y_S and Y_T for the outputs of the source and target tasks at an input X. Let $D(x)$ denote a scalar difficulty score derived from the source task, where larger values indicate inputs that are more difficult to predict. For example, $D(x)$ can be learned to approximate the expected prediction error of the source model,

$$D(x) \approx \mathbb{E}\left[\ell_S(Y_S, \mu_S(x)) \mid X = x\right], \tag{3}$$

where μ_S is the source model and ℓ_S is its prediction loss. For a supervised source task, we fit a gradient-boosting regressor to predict $\ell_S(Y_S, \mu_S(X))$ from X, and use its prediction as $D(X)$; other constructions are possible, and the language-model study of Section 4.2 uses base-model answer-position entropy, which needs no source labels.

We use five mutually disjoint data splits, named by task (S: source, T: target) and role. The source predictor μ_S is trained on $\mathcal{D}_S^{\text{train}}$, and the source difficulty D is learned on $\mathcal{D}_S^{\text{diff}}$ from held-out source prediction errors. The target predictor μ_T is fine-tuned on $\mathcal{D}_T^{\text{train}}$, the one-dimensional link h from difficulty to target error scale is fitted on $\mathcal{D}_T^{\text{tune}}$, and conformal calibration is performed on $\mathcal{D}_T^{\text{cal}} = \{(X_i, Y_{T,i})\}_{i=1}^n$. Source and target observations need not be paired or share the same outputs. Figure 1 shows the information flow, and Algorithm 1 gives pseudocode.

We learn a positive one-dimensional function $h : \mathbb{R} \to (0, \infty)$ that maps the source-derived difficulty $D(x)$ to the prediction-error scale of the target model. Using a target tuning set $\mathcal{D}_T^{\text{tune}}$, h is fitted from pairs

$$\left\{\big(D(X),\, \ell_T(Y_T, \mu_T(X))\big) : (X, Y_T) \in \mathcal{D}_T^{\text{tune}}\right\}. \tag{4}$$

DC-CP does not require h to be monotone. Any positive one-dimensional link fitted independently of the calibration set can be used. In this work, we use isotonic regression, so our implementation assumes a monotone relation between source difficulty and target loss. The monotone direction $\hat{\tau} \in \{\pm 1\}$ is chosen from the sign of the Spearman rank correlation between $D(X)$ and $\ell_T(Y_T, \mu_T(X))$ on $\mathcal{D}_T^{\text{tune}}$, and the fitted link is clipped below at $\varepsilon = 10^{-6}$ for positivity. As a diagnostic of how informative the source signal is, we compute for each split the tune-set Pearson correlation $\rho(D, \ell_T)$ between $D(X)$ and $\ell_T(Y_T, \mu_T(X))$, and report its mean $\bar{\rho}$ over the repeated splits; Spearman correlation is used only to choose the direction.

DC-CP then defines the normalized nonconformity score as

$$V^{\text{DC-CP}} = \frac{\ell_T(Y_T, \mu_T(X))}{h(D(X))}. \tag{5}$$

Here ℓ_T may be any nonnegative nonconformity loss: $\ell_T(y, \hat{y}) = |y - \hat{y}|$ for regression, $1 - \hat{p}_y$ for classification, where $\hat{p}_y$ is the predicted probability of class y (the class-probability score; Adaptive Prediction Sets, APS (Romano et al., 2020b; Angelopoulos et al., 2021), is its cumulative variant), and the mean token-level negative log-likelihood (NLL) of the model's generated answer for generation (Section 4.2). A score may ignore its label argument: in the language-model study it is the mean NLL of the model's own greedy answer, a deterministic function of X once μ_T is fixed.

These scores are calibrated using weighted conformal prediction: Equations (1) and (2) are applied to $V_i = \ell_T(Y_{T,i}, \mu_T(X_i))/h(D(X_i))$ for $(X_i, Y_{T,i}) \in \mathcal{D}_T^{\text{cal}}$, so $\hat{q}_{1-\alpha}(x)$ below is the weighted quantile of the DC-CP scores rather than of the raw losses. For regression with the absolute-error loss, the condition $|y - \mu_T(x)|/h(D(x)) \leq \hat{q}_{1-\alpha}(x)$ gives the prediction interval for a test input x,

$$C(x) = \mu_T(x) \pm \hat{q}_{1-\alpha}(x)\, h(D(x)). \tag{6}$$

Thus, source-derived difficulty determines the local scale of the prediction interval, while weighted conformal calibration determines the coverage level under distribution shift. Importantly, DC-CP transfers only the scalar function $D(x)$ and therefore requires neither paired source–target observations nor a shared output space between the source and target tasks.

Because D, μ_T and h are fitted on data independent of $\mathcal{D}_T^{\text{cal}}$ and of the test point, the score Equation (5) is fixed during conformal calibration, so with exact importance weights weighted conformal prediction directly gives finite-sample marginal coverage $\mathbb{P}\{Y_{T,n+1} \in C(X_{n+1})\} \geq 1-\alpha$ (Tibshirani et al., 2019); with estimated weights, as in both real-data studies below, coverage depends on the density-ratio estimate.

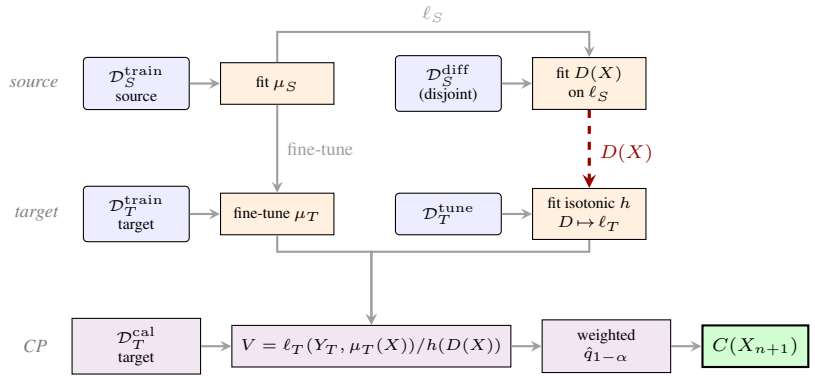

Figure 1: DC-CP information flow. The red dashed path ($D(X)$ via the isotonic link h) is the only additional source-derived quantity entering conformal calibration beyond the transferred predictor μ_T.

2.3 Why DC-CP works

Marginal validity holds for any fixed positive h and therefore does not explain when DC-CP improves on WCP. We posit the *Difficulty Transfer Hypothesis* (DTH):

$$\ell_T(Y_T, \mu_T(X)) \;=\; g(D(X))\, Z, \qquad Z \perp\!\!\!\perp X, \tag{DTH}$$

where $g : \operatorname{supp} D \to (0, \infty)$ maps the transferred source difficulty to the target loss scale, and Z is an X-independent residual noise variable.

Intuitively, DTH describes a setting in which some sources of prediction difficulty are shared across tasks. Inputs may remain difficult in both the source and target tasks because they are rare, weakly represented in the available data, intrinsically ambiguous, or not fully described by the available features. In such cases, the source-derived difficulty D can carry information about the scale of target prediction errors, although the source-to-target relationship may differ across tasks and is represented by g.

DTH can fail when the factors that make an input difficult on the target task are largely task-specific or are not captured by the transferred scalar D. In that case, DC-CP still retains marginal validity with exact weights, but its difficulty-balancing advantage may disappear.

If the estimated link is consistent up to scale, $h \to c\,g$ for some $c > 0$, then

$$\text{WCP:}\quad V^{\text{WCP}} = \ell_T = g(D)\, Z \qquad \Longrightarrow \text{ the scale } g(D) \text{ remains,}$$

$$\text{DC-CP:}\quad V^{\text{DC-CP}} = \frac{\ell_T}{h(D)} = \frac{g(D)\, Z}{h(D)} \;\to\; \frac{Z}{c} \qquad \Longrightarrow \text{ the scale is removed.}$$

The conditional coverage error (CCE) measures how far coverage within difficulty strata deviates from the nominal level. Writing $C_M(X)$ for the prediction set of method M,

$$\text{CCE}(M) \;=\; \sqrt{\mathbb{E}\big[\big(\mathbb{P}(Y_T \in C_M(X) \mid D) - (1-\alpha)\big)^2\big]}, \tag{7}$$

where the expectation is taken over $D = D(X)$ induced by $X \sim P_{\text{test},X}$; in experiments the conditional coverage is estimated within ten D-deciles of the test set. The following result describes this difference in conditional coverage.

We first fix notation. F_Z denotes the cumulative distribution function of Z; $n_{\text{tune}} = |\mathcal{D}_T^{\text{tune}}|$ and $n_{\text{cal}} = |\mathcal{D}_T^{\text{cal}}|$ are the tune- and calibration-set sizes; and $q_{1-\alpha}^{\text{WCP}}$ and $q_{1-\alpha}^{\text{DC-CP}}$ are the population

$(1-\alpha)$-quantiles of the WCP and DC-CP scores under the target distribution, written without a hat to distinguish them from the finite-sample threshold $\hat{q}_{1-\alpha}(x)$ of Equation (2).

Theorem 1 (CCE separation under difficulty transfer; informal)**.** *Assume DTH,*

$$\ell_T(Y_T, \mu_T(X)) \;=\; g(D(X))\, Z, \qquad Z \perp\!\!\!\perp X,$$

and suppose that the link h, fitted on the n_{tune} target tuning samples, is consistent up to scale,

$$h \;\to\; c\, g \qquad \text{as } n_{\text{tune}} \to \infty,$$

for some $c > 0$.

The WCP threshold is then

$$q_{1-\alpha}^{\text{WCP}} \;=\; \text{Quantile}_{1-\alpha}\big(g(D(X))\, Z\big),$$

the population $(1-\alpha)$-quantile of the WCP score under the target distribution, and

$$\mathbb{P}\big(Y_T \in C_{\text{WCP}}(X) \mid D = d\big) = F_Z\left(\frac{q_{1-\alpha}^{\text{WCP}}}{g(d)}\right),$$

which generally varies with d when g is nonconstant, even though WCP retains marginal coverage.

For DC-CP, as $n_{\text{tune}} \to \infty$,

$$V^{\text{DC-CP}} \;=\; \frac{g(D)\, Z}{h(D)} \;\to\; \frac{Z}{c},$$

and, as the calibration size $n_{\text{cal}} \to \infty$, its population threshold $q_{1-\alpha}^{\text{DC-CP}} = \text{Quantile}_{1-\alpha}\big(\ell_T / h(D)\big)$ satisfies

$$q_{1-\alpha}^{\text{DC-CP}} \;\to\; \frac{F_Z^{-1}(1-\alpha)}{c}.$$

Hence the difficulty-conditional coverage converges to $1-\alpha$, and therefore

$$\text{CCE(DC-CP)} \;\to\; 0.$$

Under the additional regularity conditions given in Section B.4, the CCE of WCP is lower-bounded by

$$\text{CCE(WCP)} \;\geq\; \kappa \sqrt{\text{Var}\big(g(D)\big)}$$

for a problem-dependent constant $\kappa > 0$.

With an exact link $h = c\, g$ the limits are equalities, so the conditional coverage of DC-CP is $1-\alpha$ at every difficulty level; the explicit constants and the proof are in Section B.4. This contrast is visible in the experiments: WCP coverage drifts across D-deciles, whereas DC-CP substantially flattens this drift (Figures 3 and 4), and the fitted h tracks the target loss scale across benchmarks (Figures 5 to 8). Section B.3 explains why regressing ℓ_T on D targets $c\, g$ and when the isotonic estimator is consistent. Section 3 compares DC-CP with the target-side scale and quantile models of locally weighted conformal prediction (LWCP) and conformalized quantile regression (CQR). DC-CP is particularly useful under covariate shift or in high-dimensional settings, where a target-side model must extrapolate from limited target data. When D contains no information about the target loss scale, the population link is constant and DC-CP reduces to WCP; with finite tuning data, estimating h may instead add noise.

3 Related work

WCP (Tibshirani et al., 2019), extended to non-exchangeable data (Barber et al., 2023) and to robust and PAC guarantees (Cauchois et al., 2024; Park et al., 2022), restores marginal coverage under covariate shift by weighting calibration scores with a density ratio, usually estimated by a probabilistic discriminator (Shimodaira, 2000; Sugiyama et al., 2008; Bickel et al., 2009). It does not explicitly model a difficulty-dependent error scale. Adaptive methods learn this relationship from target data. LWCP (Lei et al., 2018), like earlier normalized scores (Papadopoulos et al., 2008; Boström & Johansson, 2020), divides the error by a p-dimensional target-side scale model, whereas

CQR (Romano et al., 2019) conformalizes p-dimensional target-side quantile regressors. Under covariate shift, both must extrapolate in input space from $P_{\mathrm{cal},X}$ to $P_{\mathrm{test},X}$. DC-CP instead learns a one-dimensional link on a difficulty transferred from the source (Table 1); all methods use the same weighted conformal quantile and retain marginal validity with exact weights. CCE measures balance across difficulty strata rather than distribution-free conditional coverage; related conditional notions are discussed in Section A.3.

In transfer learning (Pan & Yang, 2010; Zhuang et al., 2021), uncertainty calibration is typically rebuilt from target data. Existing approaches that incorporate source or auxiliary information into conformal calibration either require paired source responses at calibration and test points (Equation (8)), assume a shared label space (Wang & Qiao, 2025), or transfer source data or auxiliary signals into target-side models (Zhang et al., 2026; Seedat et al., 2023; Min et al., 2026). DC-CP instead transfers only a scalar source-derived difficulty and requires neither paired responses nor a shared output space.

Table 1: The three baselines and DC-CP. All four use the same importance weights and the weighted quantile $q = \hat{q}_{1-\alpha}(x)$ of Equation (2), applied to their own score. The regression set is shown for the absolute-error loss; p is the input dimension (4, 21, 190 and 1024 in our benchmarks). DC-CP fits D on $\mathcal{D}_S^{\mathrm{diff}}$ and h on $\mathcal{D}_T^{\mathrm{tune}}$.

Method	Score V	Regression set $C(x)$	Learned for adaptation	Fitted on
WCP	ℓ_T	$\mu_T(x) \pm q$	nothing	–
LWCP	$\ell_T / \hat{\sigma}(X)$	$\mu_T(x) \pm q\,\hat{\sigma}(x)$	$\hat{\sigma} : \mathbb{R}^p \to (0, \infty)$	$\mathcal{D}_T^{\mathrm{tune}}$
CQR	$\max\{\hat{Q}_T^{\mathrm{lo}}(X) - Y_T,\, Y_T - \hat{Q}_T^{\mathrm{hi}}(X)\}$	$[\hat{Q}_T^{\mathrm{lo}}(x) - q,\, \hat{Q}_T^{\mathrm{hi}}(x) + q]$	$\hat{Q}_T^{\mathrm{lo}}, \hat{Q}_T^{\mathrm{hi}} : \mathbb{R}^p \to \mathbb{R}$	$\mathcal{D}_T^{\mathrm{tune}}$
DC-CP	$\ell_T / h(D(X))$	$\mu_T(x) \pm q\,h(D(x))$	$h : \mathbb{R} \to (0, \infty)$ (isotonic here)	$\mathcal{D}_S^{\mathrm{diff}}, \mathcal{D}_T^{\mathrm{tune}}$

4 Results

WCP and LWCP share the transferred predictor μ_T with DC-CP, which alone uses source-side calibration information beyond it; CQR is included as a target-side quantile-regression baseline (Section 3). We first isolate the mechanism synthetically, contrasting the covariate-shift regime that motivates the paper against a matched no-shift control. We then report two real-data settings: answer retention in question answering on two benchmarks, with an 8B-parameter language model under topic shift, where the target-side covariate is a learned sentence embedding; and a polymer benchmark with a held-out chemical class. SARCOS (Section D.1) illustrates a regime in which a target-side scale model is sufficient on its own.

4.1 Difficulty-stratified coverage under covariate shift

We evaluate Ackley and Levy functions in $d = 4$ under covariate shift and split the test set at the median distance to the target fine-tuning center. We call the nearer half target in-distribution (tgt ID) and the farther half target out-of-distribution (tgt OOD).

All four methods are marginally valid (Table 2), so the comparison concerns how coverage is distributed. WCP over-covers on tgt ID (0.972 on Ackley, 1.000 on Levy) and under-covers on tgt OOD (0.832 and 0.801). LWCP reduces this imbalance only slightly, whereas DC-CP is closest to nominal in both halves ($0.908/0.901$ on Ackley and $0.928/0.889$ on Levy). Relative to LWCP, DC-CP reduces $|\mathrm{cov} - 0.9|$ in all four cells, with the largest gain on Levy tgt OOD ($+0.038$). Levy is also where the source signal is more informative ($\bar{\rho} = 0.52$ vs. 0.27), which is consistent with the mechanism, although $\bar{\rho}$ is a diagnostic rather than the quantity in Theorem 1.

This comes with an efficiency trade-off. LWCP attains the narrowest marginal width on Levy (1.44 vs. 2.45 for DC-CP), but does so by under-covering the hard half at 0.823 while over-covering the easy half at 0.979. CQR instead inflates widths broadly. Marginal width alone therefore conceals the conditional imbalance that the two-cell split exposes.

Without shift the advantage over LWCP largely vanishes: the same paired quantity on tgt OOD is -0.002 on Ackley and $+0.006$ on Levy, against $+0.007$ and $+0.038$ under shift (Table 8). This suggests that the main limitation is poor target-region coverage induced by the shift, rather than tune-set size alone: $\hat{\sigma}(X)$ is fitted on tune points drawn from $P_{\mathrm{cal},X}$, which samples the shifted test region

Table 2: Coverage under covariate shift on the target in-distribution (tgt ID) and out-of-distribution (tgt OOD) halves of the test set. Mean coverage ± s.d. / median width, 100 seeds, nominal 0.9; setup in Section C.2. **Bold**: closest to nominal, a statement about calibration and not about efficiency or significance. The marginal row is a validity check rather than a comparison, since all four methods are marginally valid by construction, and is left unbolded. Paired comparisons in the text use the per-seed difference in $|\text{cov} - 0.9|$; standard errors and paired t statistics are in Section D.3. No-shift control: Table 8.

Func	Cell	WCP cov/w	CQR cov/w	LWCP cov/w	DC-CP cov/w
Ackley	tgt ID	.972±.02/2.58	.826±.06/5.22	.947±.03/2.16	**.908±.04**/2.29
	tgt OOD	.832±.04/2.50	.995±.01/4.74	.861±.03/2.65	**.901±.05**/3.53
	marginal	.902±.02/2.53	.910±.03/4.98	.904±.03/2.35	.904±.03/2.93
Levy	tgt ID	1.00±.00/2.67	.854±.05/9.28	.979±.02/1.09	**.928±.05**/1.34
	tgt OOD	.801±.04/2.59	.964±.02/9.10	.823±.04/2.04	**.889±.05**/4.12
	marginal	.900±.02/2.62	.909±.03/9.20	.901±.02/1.44	.908±.03/2.45

sparsely, and an n_{tune} sweep under shift (Section D.4) shows the gap persisting to $n_{\text{tune}} = 400$. $D(X)$, by contrast, is estimated once from the source residual pool and never re-estimated per target.

We also test Theorem 1 directly. Because Ackley and Levy were not constructed to satisfy DTH, we first use a synthetic setting where it holds exactly and the difficulty spread can be controlled (Section C.1). Figure 2a shows the separation $\text{CCE}(\text{WCP}) - \text{CCE}(\text{DC-CP})$. As the difficulty spread increases, WCP becomes increasingly imbalanced, whereas the remaining DC-CP error mainly comes from estimating h with finite data. This error decreases with the tune-set size: at $\gamma = 2$ in the DTH-exact setup of Section C.1, the DC-CP CCE falls from 0.052 to 0.014 as n_{tune} increases from 100 to 1600 (Tables 5 and 6). Accordingly, the crossover at which DC-CP outperforms WCP moves toward smaller difficulty spreads as the tune set grows.

Figure 2b repeats the analysis on the same Ackley and Levy runs. Since DTH implies

$$\mathbb{E}[\ell_T \mid D] = g(D)\,\mathbb{E}[Z],$$

we use $\sqrt{\text{Var}(\mathbb{E}[\ell_T \mid D])}$ as a proxy for the variation of the difficulty scale. Across the 200 runs of Table 2, the separation increases with this quantity ($r = 0.54$), showing the same qualitative trend as panel (a) but with a smaller effect, although DTH is not imposed in these experiments.

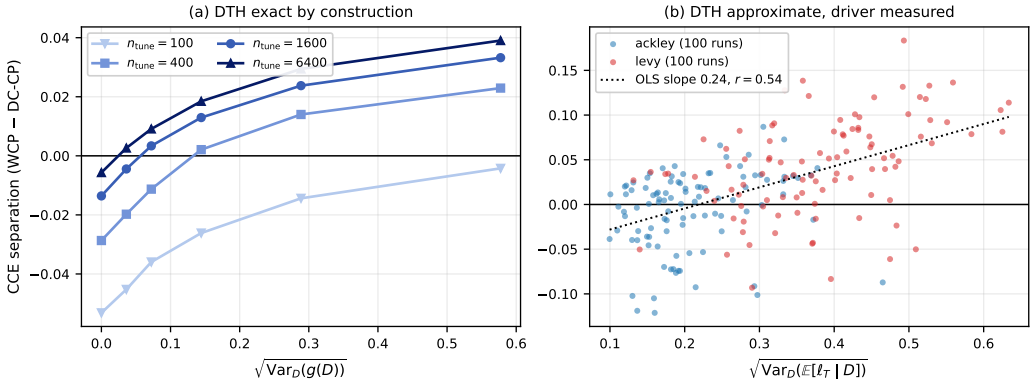

Figure 2: Difficulty-scale variation and the WCP–DC-CP separation. (a) DTH exact by construction (Section C.1): DC-CP has an estimation error from fitting h, which decreases as the tune set grows, so the crossing moves left as $\text{CCE}(\text{DC-CP}) \to 0$. (b) Ackley and Levy, where DTH is not imposed by construction and the difficulty-scale variation is estimated empirically.

4.2 MAIN RESULT I: LLM UNCERTAINTY AS TRANSFERABLE DIFFICULTY

We test whether source-derived difficulty transfers to a high-dimensional language-model setting, using NQ-Open (Kwiatkowski et al., 2019; Lee et al., 2019) and TriviaQA (Joshi et al., 2017) under topic shift. Llama 3.1 8B-Instruct is fine-tuned with low-rank adaptation (LoRA) (Hu et al., 2022); we use base-model answer-position entropy as $D(X)$, which needs no source labels, and the fine-tuned model's mean negative log-likelihood (NLL) of its own greedy answer as the retention score. Importance weights come from a discriminator on E5 text embeddings (Wang et al., 2022), and LWCP fits $\hat{\sigma}(X)$ on the same 1024-d embedding; CQR does not apply because the score is not a regression residual. Because the score does not use the gold answer, coverage here refers to retention rather than correctness, and with estimated weights the finite-sample guarantee does not hold exactly. Full setup: Sections C.5 and D.9.

At $\alpha = 0.1$ ($n_{\text{tune}} = 425$, $n_{\text{cal}} = 440$, $n_{\text{test}} = 1099$, 30 seeds), hardest-decile retention on NQ-Open increases from 0.500 (WCP) and 0.535 (LWCP) to $\mathbf{0.795}$ with DC-CP, while CCE falls from 0.145 and 0.134 to $\mathbf{0.059}$. TriviaQA shows the same pattern: hardest-decile retention rises from 0.674 and 0.675 to $\mathbf{0.886}$, and CCE falls from 0.102 and 0.099 to $\mathbf{0.038}$ (Table 14, Figure 3). The mechanism is the one seen in Section 4.1: on NQ-Open, $\hat{\sigma}(X)$ correlates 0.957 with the score on the tune set but only 0.12 on the shifted test set, whereas $h(D)$ retains 0.45 (0.950, 0.043 and 0.474 on TriviaQA), so the 1024-d scale model does not extrapolate beyond the tune topics while the one-dimensional difficulty does. Because DC-CP retains more difficult questions, retained exact match on NQ-Open falls slightly, from 0.347 to 0.331; ablations over n_{tune}, the definition of D and the shift severity are in Section D.8.

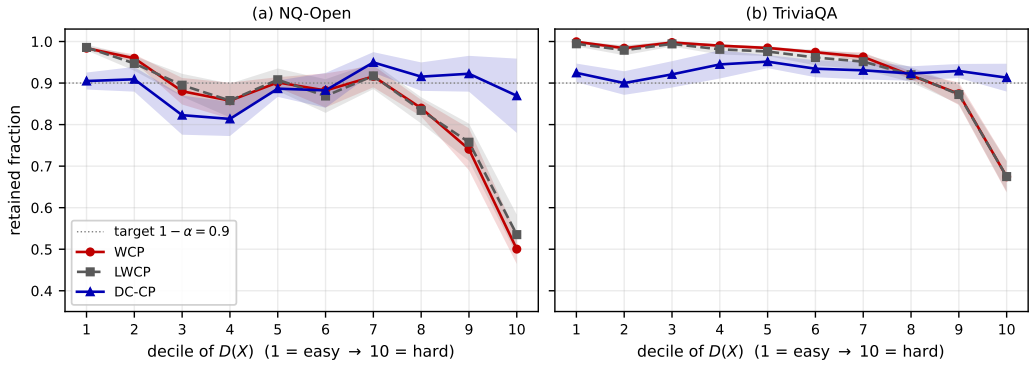

Figure 3: Retained fraction per $D(X)$-decile under topic shift (Llama 3.1 8B LoRA, $\alpha = 0.1$, $n_{\text{tune}} = 425$, $n_{\text{cal}} = 440$, $n_{\text{test}} = 1099$, 30 seeds; ribbons: ± 1 s.d.). (b) applies the same protocol to TriviaQA. In both, WCP and LWCP lie on top of each other and fall together on the hardest decile, while DC-CP remains substantially closer to nominal.

4.3 MAIN RESULT II: CROSS-TASK TRANSFER UNDER A POLYMER-CLASS HOLDOUT

We next test cross-task transfer when source and target outputs differ, on a simulation-to-real liquid-crystal polymer (LCP) benchmark (Anonymous, 2026). Source models are trained on ~ 60k simulated polymers from PolyOmics (Yoshida et al., 2025), whereas the targets are experimental melting-temperature (T_m) regression and nematic-phase classification on shared 190-dimensional kernel-mean descriptors (Kusaba et al., 2023). Whole polymer classes are held out to induce OOD shift, so DC-CP can transfer only source-derived difficulty, not paired residuals. Source properties, holdouts and the classification analogue of LWCP are described in Section C.3.

DC-CP reaches near-nominal coverage on both T_m holdouts and is closest to nominal in both nematic classification settings (Table 3), although on polyether WCP is equally close with narrower intervals. In particular, DC-CP attains 0.919 coverage with the smallest set size on held-out polyimides. These results show that difficulty transfer remains possible even when source and target tasks have incompatible output spaces.

Table 3: Simulation-to-real transfer-learning LCP results at $1-\alpha = 0.9$, 30 seeds, mean coverage $\pm$ s.d., over four polymer-class holdouts. Regression width is the median over seeds, in K; classification size is the mean over seeds, in $[0, 2]$. **Bold**: coverage closest to nominal, ties both bold; a statement about calibration, not efficiency. On the classification rows LWCP is the analogue described in Section C.3; CQR is omitted.

Task	Holdout (n_{test})	WCP	CQR	LWCP	DC-CP
Reg. T_m	polyimide, cls 13 (43)	$.878 \pm .12/284$	$.898 \pm .09/368$	$.867 \pm .17/356$	$\mathbf{.899 \pm .12}/331$
cov / w (K)	polyether, cls 7 (34)	$\mathbf{.902 \pm .10}/238$	$.999 \pm .01/299$	$.942 \pm .08/380$	$\mathbf{.902 \pm .10}/311$
Cls. nematic	polyimide, cls 13 (320)	$.960 \pm .06/1.88$	n/a	$.930 \pm .13/1.85$	$\mathbf{.919 \pm .12}/1.83$
cov / size	not_13 reverse-OOD (137)	$.845 \pm .13/1.50$	n/a	$.849 \pm .13/1.52$	$\mathbf{.862 \pm .12}/1.60$

5 Discussion and conclusion

This work asks whether transfer learning can transfer not only a predictor, but also information useful for calibrating its uncertainty. DC-CP provides one such mechanism: it transfers a source-derived scalar encoding prediction difficulty and uses a small target tune set to learn how that scalar relates to the target loss scale. Because the transferred quantity is a score rather than a label or residual, source and target tasks need not share an output space or provide paired responses.

The experiments identify when this transfer is useful. Under covariate shift, a target-side scale model must extrapolate from the region covered by the target tuning data into the shifted deployment region, whereas a source-derived difficulty score can remain informative there. This produces the largest gains in our synthetic and language-model experiments. Without shift, target-side adaptation is often competitive or better once enough target data are available. The polymer experiment illustrates a complementary case: even when direct source–target residual transfer is impossible, difficulty itself can still transfer across tasks with incompatible outputs.

The method has two main limitations.

(i) *Difficulty transfer is not guaranteed.* DC-CP relies on the source-derived scalar containing information about variation in the target loss scale. The Difficulty Transfer Hypothesis formalizes this through a one-dimensional scale function $g(D)$. When this relationship is weak or violated, marginal validity remains, but the difficulty-balancing advantage can disappear, and at finite tune size estimating h can add noise, so the tune-set association between D and the target score should be checked before use.

(ii) *DC-CP is deliberately low-dimensional.* The target link h is only a one-dimensional function of the transferred source difficulty. This makes it possible to adapt uncertainty with limited target data, but it cannot capture target-specific structure absent from D.

Taken together, these results suggest that calibration need not always be rebuilt from scratch after transfer learning. A source task can provide a reusable notion of prediction difficulty, and a small amount of target data can be sufficient to translate that notion into a new task-specific uncertainty scale.

AI use statement

In this work, we used generative AI tools to implement methods, to propose and refine hypotheses, to design and give feedback on experiments, and to interpret results: Claude Code was used to write and run the experiment code, to draft and edit parts of this paper, and to translate drafts between English and Japanese, and ChatGPT was used to assist with writing.

Generative AI tools were also used to review draft mathematical arguments, to identify possible gaps, and to suggest candidate corrections; all mathematical statements and proofs in the final paper were independently checked, derived, and rewritten by the authors. We have not used generative AI tools to originate the theoretical model or conceptual framework, to generate synthetic datasets, or to clean or reformat datasets; qualitative and thematic data analysis and the formulation of survey or interview questions are not applicable to this work.

We have reviewed all AI-assisted work: every numerical claim in this paper was re-derived from the result files rather than transcribed from model output, and AI-proposed hypotheses were adopted only where an experiment

we ran supported them. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

DC-CP is a calibration method: it takes an existing predictor and a small target tune set and, with auxiliary models fitted on independent data and exact covariate-shift weights (Section 2.2), returns prediction sets with finite-sample marginal coverage. It does not generate, modify, or amplify the underlying predictor, and it inherits the failure modes of that predictor; conformal validity is a coverage guarantee, not a fairness, factuality, or robustness guarantee. Practitioners should not read a nominal $1 - \alpha$ coverage level as evidence that a downstream decision is safe. The LLM extension uses a publicly released model (Llama 3.1 8B-Instruct) on public benchmarks (NQ-Open and TriviaQA); no human subjects, private data, or new datasets are introduced. The polymer benchmark combines public OpenLCP descriptors and PolyOmics simulations with license-restricted PoLyInfo metadata; the subset of results that depends only on public data is reported separately in Section D.6. Further discussion is in Section F.4.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated with their assumptions and proved in full: the algorithm and the marginal-validity argument in Section A, and the population-level CCE separation theorem, together with the Difficulty Transfer Hypothesis and the link-consistency condition it requires, in Section B. Complete experimental protocols for the synthetic, polymer, SARCOS, and LLM settings, including model architectures, splits, and hyperparameters, are given in Section C, with additional results, sweeps, and ablations in Section D. Data provenance and licensing, including which results are reproducible from public data alone, are described in Section F.1; a public-data-only reproduction of the main polymer claim is given in Section D.6. Code that reproduces every table and figure, together with the per-seed outputs behind them, is submitted as anonymous supplementary material (Section F.2), and the compute required is disclosed in Section F.3.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: a gentle introduction. *Found. Trends Mach. Learn.*, 16:494–591, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Jitkomut Lei, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

Anonymous. Design and phase control of liquid crystal polymers using machine learning. Under review; author names withheld for double-blind review, 2026.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Inf. Inference*, 10:455–482, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Ann. Stat.*, 51:816–845, 2023.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10:2137–2155, 2009.

Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In *Proc. Symp. Conformal Probabilistic Predict. Appl. (COPA)*, pp. 114–133, 2020.

Tom B. Brown et al. Language models are few-shot learners. In *Adv. Neural Inf. Process. Syst.*, volume 33, pp. 1877–1901, 2020.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: confident predictions even when distributions shift. *J. Am. Stat. Assoc.*, 119:3033–3044, 2024.

Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proc. Natl Acad. Sci. USA*, 119: e2204569119, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1050–1059, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Adv. Neural Inf. Process. Syst.*, volume 30, pp. 4878–4887, 2017.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *J. R. Stat. Soc. B*, 87:1100–1126, 2025.

Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110:33–50, 2023.

Yoshihiro Hayashi, Junichiro Shiomi, Junko Morikawa, and Ryo Yoshida. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Comput. Mater.*, 8:222, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv:2102.01293*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.

Zirui Hu, Zheng Zhang, Yingjie Wang, Leszek Rutkowski, and Dacheng Tao. CoFact: Conformal factuality guarantees for language models under covariate shift. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2026.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, pp. 1601–1611, 2017.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Adv. Neural Inf. Process. Syst.*, volume 30, pp. 5574–5584, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

Minoru Kusaba, Yoshihiro Hayashi, Chang Liu, Araki Wakiuchi, and Ryo Yoshida. Representation of materials by kernel mean embedding. *Phys. Rev. B*, 108:134107, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguist.*, 7:452–466, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Adv. Neural Inf. Process. Syst.*, volume 30, pp. 6402–6413, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, pp. 6086–6096, 2019.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. B*, 76:71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113:1094–1111, 2018.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *J. R. Stat. Soc. B*, 83:911–938, 2021.

Zeming Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023.

Hiroaki Mikami, Kenji Fukumizu, Shogo Murai, Shuji Suzuki, Yuta Kikuchi, Taiji Suzuki, Shin-ichi Maeda, and Kohei Hayashi. A scaling law for syn2real transfer: How much is your pre-training effective? In *Mach. Learn. Knowl. Discov. Databases (ECML PKDD 2022)*, pp. 477–492, 2023.

Yinjie Min, Liuhua Peng, and Changliang Zou. Enhanced localized conformal prediction with imperfect auxiliary information. *J. Am. Stat. Assoc.*, 2026. to appear; arXiv:2606.08551.

Shunya Minami, Kenji Fukumizu, Yoshihiro Hayashi, and Ryo Yoshida. Transfer learning with affine model transformation. In *Adv. Neural Inf. Process. Syst.*, volume 36, pp. 17296–17329, 2023.

Shunya Minami, Yoshihiro Hayashi, Stephen Wu, Kenji Fukumizu, Hiroki Sugisawa, Masashi Ishii, Isao Kuwajima, Kazuya Shiratori, and Ryo Yoshida. Scaling law of Sim2Real transfer learning in expanding computational materials databases for real-world predictions. *npj Comput. Mater.*, 11:146, 2025.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proc. IEEE Int. Conf. Neural Netw. (ICNN)*, volume 1, pp. 55–60, 1994.

Shingo Otsuka, Isao Kuwajima, Junko Hosoya, Yibin Xu, and Masayoshi Yamazaki. PoLyInfo: Polymer database for polymeric materials design. In *Proc. Int. Conf. Emerging Intelligent Data Web Technol. (EIDWT)*, pp. 22–29, 2011.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. Stanford University, 2013.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22:1345–1359, 2010.

Harris Papadopoulos, Konstantinos Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Proc. Eur. Conf. Mach. Learn. (ECML)*, pp. 345–356, 2002.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proc. IASTED Int. Conf. Artif. Intell. Appl. (AIA)*, pp. 64–69, 2008.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. PAC prediction sets under covariate shift. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Adv. Neural Inf. Process. Syst.*, volume 32, pp. 3543–3553, 2019.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J. Candès. With malice toward none: assessing uncertainty via equalized coverage. *Harv. Data Sci. Rev.*, 2, 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Adv. Neural Inf. Process. Syst.*, volume 33, pp. 3581–3591, 2020b.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, pp. 10160–10177, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, 2008.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference*, 90:227–244, 2000.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Ann. Inst. Stat. Math.*, 60:699–746, 2008.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Adv. Neural Inf. Process. Syst.*, volume 32, pp. 6414–6425, 2019.

Ryan J. Tibshirani, Rina Foygel-Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Adv. Neural Inf. Process. Syst.*, volume 32, pp. 2530–2540, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inf. Process. Syst.*, volume 30, pp. 5998–6008, 2017.

Sethu Vijayakumar and Stefan Schaal. Locally weighted projection regression: An $O(n)$ algorithm for incremental real time learning in high dimensional space. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1079–1086, 2000.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92:349–376, 2012.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. In *Proc. Int. Conf. Mach. Learn. (ICML) Workshop on On-line Learning*, 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Baozhen Wang and Xingye Qiao. Conformal prediction under generalized covariate shift with posterior drift. In *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, pp. 4888–4896, 2025.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv:2212.03533*, 2022.

Ryo Yoshida, Yoshihiro Hayashi, Hidemine Furuya, Ryohei Hosoya, et al. Omics-scale polymer computational database transferable to real-world artificial intelligence applications. *arXiv:2511.11626*, 2025.

Ce Zhang, Ting Li, Jinhan Xie, Linglong Kong, and Bei Jiang. Transfer conformal predictive inference for regression. *J. Mach. Learn. Res.*, 27:1–68, 2026.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proc. IEEE*, 109:43–76, 2021.

A Algorithm and method details

A.1 Algorithm

Algorithm 1: DC-CP at a test point X_{n+1}.

Input: disjoint splits $\mathcal{D}_S^{\text{train}}, \mathcal{D}_S^{\text{diff}}, \mathcal{D}_T^{\text{train}}, \mathcal{D}_T^{\text{tune}}, \mathcal{D}_T^{\text{cal}}$; fitted predictors μ_S, μ_T; weight $w(x) = P_{\text{test},X}(x)/P_{\text{cal},X}(x)$; level α.

Output: prediction set $C(X_{n+1})$ with $\mathbb{P}(Y_{T,n+1} \in C(X_{n+1})) \geq 1 - \alpha$.

Source-side D: fit a gradient-boosting machine (GBM) regressor of $\ell_S(Y_S, \mu_S(X))$ on $\mathcal{D}_S^{\text{diff}}$ $\Rightarrow D(\cdot) \approx \mathbb{E}[\ell_S \mid \cdot]$ (or use base-model entropy for an LLM);

Target-side h: *(implementation used in this work)* on $\mathcal{D}_T^{\text{tune}}$, fit the isotonic regression of $\ell_T(Y_T, \mu_T(X))$ on $D(X)$ over $(X, Y_T) \in \mathcal{D}_T^{\text{tune}}$, implemented with `IsotonicRegression(increasing="auto")`, which chooses the monotone direction from the Spearman correlation; clip $h \geq \varepsilon$;

Calibration scores: $V_i \leftarrow \ell_T(Y_{T,i}, \mu_T(X_i))/h(D(X_i))$ for $(X_i, Y_{T,i}) \in \mathcal{D}_T^{\text{cal}}$;

Weighted quantile: compute $\hat{q}_{1-\alpha}(X_{n+1})$ from $\{V_i\}_{i=1}^n \cup \{+\infty\}$ with test-atom weights $p_i \propto w(X_i),\ p_{n+1} \propto w(X_{n+1})$;

Output: $C(X_{n+1}) = \{y : \ell_T(y, \mu_T(X_{n+1})) \leq \hat{q}_{1-\alpha}(X_{n+1}) \cdot h(D(X_{n+1}))\}$;

A.2 Marginal validity

Every score in this paper is fitted on data independent of the calibration set and of the test point. Conditional on those data it is a fixed function of the observation, and since the covariate shift acts on X alone, marginal coverage with exact importance weights follows directly from the weighted-exchangeability argument of Tibshirani et al. (2019); averaging over the independent data gives the unconditional statement. For DC-CP this includes the data-driven direction $\hat{\tau}$, which depends on $\mathcal{D}_T^{\text{tune}}$ alone. The same argument covers CV-CP (Section E.1) and the ratio score used as a motivating contrast in Section 3,

$$V^{\text{Ratio-CP}} = \frac{\ell_T(Y_T, \mu_T(X))}{\ell_S(Y_S, \mu_S(X))} \qquad (V^{\text{Ratio-CP}} = +\infty \text{ if } \ell_S = 0), \tag{8}$$

which, unlike DC-CP, needs a paired source response Y_S at every calibration and test point.

A.3 Relation to conditional coverage

CCE measures how evenly coverage is spread across difficulty strata; it is not a conditional-coverage guarantee. Exact distribution-free conditional coverage is impossible with finite samples (Vovk, 2012; Lei & Wasserman, 2014; Barber et al., 2021). Methods that target weaker notions, such as Mondrian conformal prediction over predefined categories (Vovk et al., 2003), equalized coverage over protected groups (Romano et al., 2020a), coverage over a declared class of covariate shifts (Gibbs et al., 2025) and localized conformal prediction (Guan, 2023), change the calibration step and can be applied on top of the DC-CP score.

A.4 Methods comparison table

Table 4: Methods compared in this work. $\hat{Q}_T^{\mathrm{lo}}, \hat{Q}_T^{\mathrm{hi}}$ are the target-side $\alpha/2$ and $1-\alpha/2$ conditional quantile regressors of $Y_T \mid X$ (GBM in our experiments). ℓ_T, ℓ_S are the target and source losses; only CV-CP uses the signed residuals $R_T = Y_T - \mu_T(X)$ and $R_S = Y_S - \mu_S(X)$, with $\hat{\beta}$ the fitted source-to-target residual coefficient (Section E.1).

Method	Score V_i	Weights	Source role	Paired data	$Y_{S,n+1}$ obs
WCP (Tibshirani et al., 2019)	ℓ_T	$w(x)$	only via μ_T	no	no
CQR (Romano et al., 2019)	$\max\{\hat{Q}_T^{\mathrm{lo}}(X)-Y_T,\ Y_T-\hat{Q}_T^{\mathrm{hi}}(X)\}$	$w(x)$	none (no μ_T)	no	no
LWCP (Lei et al., 2018)	$\ell_T/\hat{\sigma}(X)$	$w(x)$	only via μ_T	no	no
Ratio-CP (motivating contrast)	ℓ_T/ℓ_S	$w(x)$	per-point Y_S	yes	yes
DC-CP (ours)	$\ell_T/h(D(X))$	$w(x)$	unpaired source	no	no
CV-CP (ours, Section E.1)	$\lvert R_T - \hat{\beta} R_S\rvert$	$w(x)$	per-point Y_S	yes	yes

B Population intuition for difficulty transfer

This appendix gives the population-level intuition behind DC-CP in four steps. Marginal validity (Section 2.2) does not depend on any content here. Here $Q_{1-\alpha}(Y \mid X)$ denotes the conditional $(1-\alpha)$-quantile of Y given X, while $\mathrm{Quantile}_{1-\alpha}(\cdot)$ is the quantile operator of Equation (2).

B.1 Difficulty Transfer Hypothesis

DTH.

There exists a continuous function $g : \operatorname{supp} D \to (0, \infty)$ and a random variable Z, independent of X, such that

$$\ell_T(Y_T, \mu_T(X)) = g(D(X))\, Z.$$

We also consider a relaxed form of DTH that allows for deviations from this exact factorization:

$$\mathbb{E}[|\log \ell_T(Y_T, \mu_T(X)) - \log g(D(X)) - \log Z|] \le \varepsilon,$$

for some $\varepsilon \ge 0$, where the expectation is taken under the joint target law of (X, Y_T). For this log-scale formulation we assume $\ell_T > 0$ and $Z > 0$ almost surely; zero-valued losses can be handled by a fixed positive offset.

DTH characterizes the transfer of difficulty as a scale relationship. Under the exact DTH, the conditional distribution of the target loss ℓ_T given X is obtained by rescaling a common distribution F_Z by the source-derived factor $g(D(X))$. Consequently, the $(1-\alpha)$ conditional quantile factorizes as

$$Q_{1-\alpha}(\ell_T \mid X) = g(D(X)) F_Z^{-1}(1-\alpha).$$

Equivalently, normalizing the target loss by $g(D(X))$ removes its input-dependent scale:

$$\frac{\ell_T}{g(D(X))} = Z,$$

so that the normalized score has a distribution independent of X. This makes it possible to calibrate the normalized scores using a single conformal threshold while allowing the resulting prediction sets to adapt to input-dependent difficulty through $g(D(X))$.

This observation motivates the DC-CP score $\ell_T/h(D)$, where h is learned from target tuning data. If $h \to c\, g$ for some constant $c > 0$ (Section B.3), then

$$\frac{\ell_T}{h(D)} \to \frac{Z}{c}.$$

Thus, the normalized score becomes asymptotically independent of X, up to a constant scale factor that is absorbed by the conformal threshold. Under the relaxed DTH, this invariance holds approximately rather than exactly.

Finally, the exact DTH implies the conditional mean relationship

$$\mathbb{E}[\ell_T \mid X] = \mathbb{E}[Z]\, g(D(X)),$$

with an approximate relationship under the relaxed form. This provides a simpler empirical diagnostic for assessing whether source-derived difficulty captures systematic variation in target prediction error.

B.2 SHARED LATENT DIFFICULTY DERIVATION

We next provide a sufficient condition under which DTH arises naturally. Suppose that the source and target prediction losses share a common latent notion of input difficulty. Specifically, let $\Lambda = \lambda(X)$ denote a latent difficulty associated with input X, and assume that Λ controls the scales of the source and target losses through task-specific functions a_S and a_T, respectively. The following proposition shows that, under this shared latent-scale model, the source-derived difficulty can be used to characterize the scale of the target loss.

Proposition 2 (DTH from shared latent difficulty)**.** *Let X denote the input, and suppose that the source and target losses admit the representations*

$$\ell_S = a_S(\Lambda)Z_S, \qquad \ell_T = a_T(\Lambda)Z_T, \qquad \Lambda = \lambda(X),$$

where Λ is a latent input difficulty shared between the source and target tasks, and

$$(Z_S, Z_T) \perp\!\!\!\perp X.$$

We assume $Z_S, Z_T \geq 0$ almost surely and $0 < \mathbb{E}[Z_S] < \infty$.

Define

$$D^\star(X) = \mathbb{E}[\ell_S \mid X] = \mathbb{E}[Z_S]\, a_S(\Lambda),$$

and let $D(X)$ denote an estimate of $D^\star(X)$.

Assume that the target scale can be written as a function of the source scale:

$$a_T(\Lambda) = \widetilde{g}(a_S(\Lambda))$$

for some positive function $\widetilde{g}$.

Define

$$g(d) := \widetilde{g}\left(\frac{d}{\mathbb{E}[Z_S]}\right).$$

If

$$\mathbb{E}_{X\sim P_{\text{test},X}} \left|\log D(X) - \log D^\star(X)\right| \leq \varepsilon_D,$$

and $u \mapsto \log g(e^u)$ is $L_{\log g}$-Lipschitz on the range of $\log D$ and $\log D^\star$, then the target loss satisfies the log-scale DTH with $Z = Z_T$ and

$$\varepsilon \leq L_{\log g}\varepsilon_D.$$

In particular, if $D = D^\star$, then the exact DTH holds:

$$\ell_T = g(D(X))Z_T.$$

Proof. Because Z_S is independent of X,

$$D^\star(X) = \mathbb{E}[\ell_S \mid X] = \mathbb{E}[Z_S]\, a_S(\Lambda).$$

Therefore,

$$a_S(\Lambda) = \frac{D^\star(X)}{\mathbb{E}[Z_S]}.$$

By the assumed factorization of the target scale,

$$a_T(\Lambda) = \widetilde{g}(a_S(\Lambda)) = \widetilde{g}\left(\frac{D^\star(X)}{\mathbb{E}[Z_S]}\right) = g(D^\star(X)).$$

Hence,

$$\ell_T = g(D^\star(X))Z_T.$$

If $D = D^\star$, this is the exact DTH.

When D is estimated, the Lipschitz assumption gives

$$\left|\log g(D(X)) - \log g(D^\star(X))\right| \leq L_{\log g}\left|\log D(X) - \log D^\star(X)\right|.$$

Since

$$\log \ell_T = \log g(D^\star(X)) + \log Z_T,$$

taking expectation under the joint target law yields

$$\mathbb{E}[|\log \ell_T - \log g(D(X)) - \log Z_T|] \leq L_{\log g}\varepsilon_D.$$

□

B.3 Link consistency and the isotonic special case

Throughout this subsection, $P_{\mathrm{tune},D}$ and $P_{\mathrm{test},D}$ denote the laws of the transferred difficulty $D(X)$ when X follows the tuning distribution $P_{\mathrm{tune},X}$ and the test distribution $P_{\mathrm{test},X}$:

$$P_{\mathrm{tune},D} := \mathcal{L}\big(D(X)\big),\ X \sim P_{\mathrm{tune},X}, \qquad P_{\mathrm{test},D} := \mathcal{L}\big(D(X)\big),\ X \sim P_{\mathrm{test},X}.$$

The link h is fitted on $\mathcal{D}_T^{\mathrm{tune}}$, so $P_{\mathrm{tune},D}$ is the distribution the estimator sees, and it is kept distinct from the calibration distribution even though the two coincide in our experiments.

The theory does not require a particular estimator for the one-dimensional link h. It only requires that the fitted link recover the target loss scale g up to a constant factor:

$$\|\log h_n - \log g - \log c\|_{L^2(P_{\mathrm{test},D})} \xrightarrow{p} 0 \qquad \text{for some } c > 0.$$

This is assumption (A3) in Section B.4. The constant c is irrelevant for DC-CP because multiplying all normalized scores by the same constant also multiplies the conformal threshold by the same amount.

Regression on D is what estimates this target scale. Under DTH,

$$\ell_T = g(D)Z, \qquad Z \perp\!\!\!\perp X.$$

Since $D = D(X)$, Z is also independent of D. Therefore,

$$\mathbb{E}[\ell_T \mid D] = \mathbb{E}[Z]\, g(D).$$

Thus, regressing the target loss ℓ_T on the transferred difficulty D estimates $g(D)$ up to the constant $c = \mathbb{E}[Z]$, which is exactly what DC-CP needs.

Isotonic regression is sufficient for this purpose. DC-CP itself does not require g to be monotone, but in this work we estimate h by least-squares isotonic regression. Therefore, when g is monotone, standard isotonic-regression consistency results give

$$\|h_n - c\, g\|_{L^2(P_{\mathrm{tune},D})} \xrightarrow{p} 0, \qquad c = \mathbb{E}[Z],$$

under standard regularity conditions, namely bounded support, a Lipschitz monotone g, and $\mathbb{E}[Z^2] < \infty$ with $\mathbb{E}[Z] > 0$.

Because g is bounded away from zero under (A1), and the fitted isotonic link is clipped below at a positive value as in Algorithm 1, this ordinary L^2 consistency also implies the log-scale consistency used in (A3).

One condition remains: the tuning and test difficulties must overlap. The link h_n is learned from the tuning distribution, while the theorem evaluates coverage under the test distribution. We therefore assume that the test distribution does not contain difficulty regions that are absent from the tuning distribution. Formally,

$$P_{\mathrm{test},D} \ll P_{\mathrm{tune},D}, \qquad \left\| \frac{dP_{\mathrm{test},D}}{dP_{\mathrm{tune},D}} \right\|_\infty \le W$$

for some $W < \infty$. Then, writing $\Delta_n = \log h_n - \log g - \log c$,

$$\mathbb{E}_{P_{\mathrm{test},D}}[\Delta_n^2] \le W\, \mathbb{E}_{P_{\mathrm{tune},D}}[\Delta_n^2] \to 0.$$

Hence, consistency learned on the tuning distribution also holds on the test distribution, which gives assumption (A3). Importantly, this overlap condition is required only for the one-dimensional difficulty $D(X)$, not for the full input space X.

B.4 Formal statement and proof of Theorem 1

For method $M \in \{\mathrm{WCP}, \mathrm{DC\text{-}CP}\}$ with set $C_M(X)$, $\mathrm{CCE}(M)$ is as in Equation (7). Throughout this subsection, expectations and variances involving D are taken with respect to the distribution of $D(X)$ induced by $X \sim P_{\mathrm{test},X}$. The following theorem is the formal version of Theorem 1 and provides its proof under explicit regularity conditions. We emphasize that the result holds **under DTH and link consistency** (assumptions (A1)–(A3) below); both are population-level conditions and the conclusion is a population-limit statement, not a finite-sample claim.

Theorem 3 (CCE separation under DTH)**.** *Assume the following three conditions.*

(A1) *DTH (Equation (DTH)):*

$$\ell_T \;=\; g(D(X))\,Z, \qquad Z \perp\!\!\!\perp X,$$

with $g : \operatorname{supp} D \to (0,\infty)$ continuous and bounded:

$$0 \;<\; \inf_d g(d) \;\le\; \sup_d g(d) \;<\; \infty.$$

(A2) *The marginal $1-\alpha$ quantile $q^{\mathrm{WCP}}_{1-\alpha} > 0$ of ℓ_T exists, and F_Z has density $f_Z \ge m > 0$ on the interval*

$$I \;:=\; \Big[\, \tfrac{q^{\mathrm{WCP}}_{1-\alpha}}{\sup_d g(d)},\;\; \tfrac{q^{\mathrm{WCP}}_{1-\alpha}}{\inf_d g(d)} \,\Big].$$

In addition, F_Z is continuous and strictly increasing at $F_Z^{-1}(1-\alpha)$, so that this quantile is unique.

(A3) *Link consistency up to scale: with h_n the link fitted on a tune set of size $n = n_{\mathrm{tune}}$, for some constant $c > 0$,*

$$\big\|\log h_n - \log g - \log c\big\|_{L^2(P_{\mathrm{test},D})} \;\xrightarrow{p}\; 0 \qquad \text{as } n \to \infty,$$

where $P_{\mathrm{test},D}$ is the law of $D(X)$ under the deployment distribution $P_{\mathrm{test},X}$ and the convergence is over the randomness of the tune set (Section B.3 gives sufficient conditions for the isotonic implementation used in this work, with $c = \mathbb{E}[Z]$).

Define the explicit problem-dependent constant

$$\kappa_g \;:=\; \frac{q^{\mathrm{WCP}}_{1-\alpha}}{(\sup_d g(d))^2} \;>\; 0.$$

Then, with the population threshold ($n_{\mathrm{cal}} \to \infty$), the DC-CP procedure based on $V = \ell_T / h(D)$ satisfies, for the estimated link h_n as $n \to \infty$,

$$\mathrm{CCE}(\text{DC-CP}) \;\xrightarrow{p}\; 0 \qquad \textit{(in probability over the tune sample)}.$$

Moreover,

$$\mathrm{CCE}(\mathrm{WCP}) \;\ge\; \kappa_g\, m \sqrt{\mathrm{Var}\big(g(D)\big)},$$

so $\kappa = \kappa_g m$ in Theorem 1.

Proof.

Step 1 (WCP lower bound; cell-coverage formula).
WCP thresholds ℓ_T at the marginal $1-\alpha$ quantile $q^{\mathrm{WCP}}_{1-\alpha}$. By (A1),

$$\mathbb{P}\big(Y_T \in C_{\mathrm{WCP}}(X) \mid D = d\big) \;=\; \mathbb{P}\big(g(D)Z \le q^{\mathrm{WCP}}_{1-\alpha} \,\big|\, D = d\big) \;=\; F_Z\big(q^{\mathrm{WCP}}_{1-\alpha}/g(d)\big),$$

and, since $q^{\mathrm{WCP}}_{1-\alpha}$ is the population $(1-\alpha)$ quantile and the distribution of ℓ_T is continuous at $q^{\mathrm{WCP}}_{1-\alpha}$ under (A2) (because $q^{\mathrm{WCP}}_{1-\alpha}/g(d) \in I$ for every d), $\mathbb{E}[F_Z(q^{\mathrm{WCP}}_{1-\alpha}/g(D))] = 1-\alpha$.

Step 2 (derivative lower bound).
Define

$$\psi(u) \;:=\; F_Z\big(q^{\mathrm{WCP}}_{1-\alpha}/u\big), \qquad u \in [\inf_d g(d),\, \sup_d g(d)].$$

Differentiating,

$$\psi'(u) \;=\; -\,f_Z\big(q^{\mathrm{WCP}}_{1-\alpha}/u\big) \cdot \frac{q^{\mathrm{WCP}}_{1-\alpha}}{u^2}.$$

For $u \in [\inf_d g(d), \sup_d g(d)]$ the argument $q^{\mathrm{WCP}}_{1-\alpha}/u \in I$, so (A2) yields $f_Z(q^{\mathrm{WCP}}_{1-\alpha}/u) \ge m$; combined with $u^2 \le (\sup_d g(d))^2$,

$$|\psi'(u)| \;\ge\; \frac{m\, q^{\mathrm{WCP}}_{1-\alpha}}{(\sup_d g(d))^2} \;=\; \kappa_g\, m \qquad \text{uniformly on } [\inf_d g(d), \sup_d g(d)].$$

Step 3 (variance lower bound via pairwise identity and the mean-value theorem (MVT)).
Let D, D' be i.i.d. copies. By the MVT and Step 2,

$$\big|\psi(g(d)) - \psi(g(d'))\big| \;\geq\; \kappa_g\, m\, \big|g(d) - g(d')\big|.$$

The pairwise-variance identity $\mathrm{Var}(Y) = \frac{1}{2}\mathbb{E}[(Y - Y')^2]$ applied to $Y = \psi(g(D))$ then gives

$$\mathrm{Var}\big(\psi(g(D))\big) \;\geq\; (\kappa_g\, m)^2\, \mathrm{Var}\big(g(D)\big).$$

Step 4 (translate to CCE).
The WCP coverage at difficulty D is $\psi(g(D))$, whose mean is $1 - \alpha$ by Step 1, so

$$\mathrm{CCE}(\mathrm{WCP})^2 \;=\; \mathrm{Var}\big(\psi(g(D))\big) \;\geq\; (\kappa_g\, m)^2\, \mathrm{Var}\big(g(D)\big),$$

and taking square roots gives the stated WCP lower bound.

Step 5 (DC-CP convergence).
We first consider the ideal case $h = cg$. Then

$$V = \frac{g(D)Z}{c\, g(D)} = \frac{Z}{c},$$

so the score no longer depends on difficulty. Its population $(1-\alpha)$-quantile is

$$q_{1-\alpha}^{\text{DC-CP}} = \frac{F_Z^{-1}(1-\alpha)}{c},$$

and therefore, for every d,

$$\mathbb{P}\big(Y_T \in C_{\text{DC-CP}}(X) \mid D = d\big) = 1 - \alpha.$$

Hence, with the exact link $h = cg$,

$$\mathrm{CCE}(\text{DC-CP}) = 0.$$

For the estimated link h_n, assumption (A3) gives

$$\frac{h_n(D)}{g(D)} \xrightarrow{p} c.$$

Therefore,

$$V_n = Z\frac{g(D)}{h_n(D)} \xrightarrow{p} \frac{Z}{c}.$$

By the continuity and uniqueness of the $(1-\alpha)$-quantile in (A2),

$$q_{1-\alpha}^{\text{DC-CP}} \xrightarrow{p} \frac{F_Z^{-1}(1-\alpha)}{c}.$$

It follows that

$$\mathbb{P}\big(Y_T \in C_{\text{DC-CP}}(X) \mid D\big) = F_Z\left(q_{1-\alpha}^{\text{DC-CP}}\frac{h_n(D)}{g(D)}\right) \xrightarrow{p} 1 - \alpha.$$

Since the conditional coverage probability is bounded in $[0, 1]$, the corresponding squared coverage error also converges to zero in expectation over D. Thus,

$$\mathbb{E}\Big[\big(\mathbb{P}\big(Y_T \in C_{\text{DC-CP}}(X) \mid D\big) - (1-\alpha)\big)^2\Big] \xrightarrow{p} 0,$$

and hence

$$\mathrm{CCE}(\text{DC-CP}) \xrightarrow{p} 0.$$

□

When is DTH plausible? DTH is strong but partially testable: it requires the conditional target-loss distribution to vary with X through the transferred one-dimensional scalar $D(X)$, up to a common residual distribution. Under (A1), WCP's D-conditional coverage $F_Z(q_{1-\alpha}^{\mathrm{WCP}}/g(d))$ decreases with $g(d)$, and the body diagnostics (Figures 2 and 4) instantiate this prediction. If D contains no information about the target loss scale, then $g(D)$ is constant and DC-CP reduces at the population level to WCP. More general violations of DTH, such as X-dependent heteroscedasticity not captured by $g(D)$, cannot in general be removed by a one-dimensional link.

C Experimental setup

C.1 DTH-exact synthetic setup

Ackley and Levy were not constructed to satisfy DTH. Their tune-set correlations $\bar{\rho}(D, \ell_T)$, 0.27 and 0.52, show only moderate source-target association and are reported as diagnostics (Figure 2b). Here we instead construct a world in which (A1) holds by construction for every $\gamma \geq 0$ and every constant in Theorem 3 is analytic; $\gamma = 0$ is included as a null control with no difficulty variation, where g is constant. Take $X \sim U([0,1]^4)$, true difficulty $D^\star = \bar{X} := \frac{1}{4}\sum_{j=1}^4 X_j$, and

$$\ell_T = g(D^\star)\,|Z_0|, \quad g(d) = 1 + \gamma d, \qquad \ell_S = e^{D^\star}\,|Z_0'|, \qquad Z_0, Z_0' \sim \mathcal{N}(0,1) \perp\!\!\!\perp X.$$

Then $Z = |Z_0|$ is half-normal, so $F_Z(z) = 2\Phi(z) - 1$ and $f_Z = 2\phi$ is decreasing on $z \geq 0$, with Φ and ϕ the standard normal distribution and density functions; hence $m = 2\phi(q_{1-\alpha}^{\mathrm{WCP}})$, $\kappa_g = q_{1-\alpha}^{\mathrm{WCP}}/(1+\gamma)^2$ and $\mathrm{Var}(g(D))^{1/2} = \gamma\,\mathrm{sd}(D^\star)$, and the exact population value $\mathrm{sd}[F_Z(q_{1-\alpha}^{\mathrm{WCP}}/g(D))]$ is available for comparison. The source link e^d differs from the target link $1+\gamma d$ on purpose: D is then recovered from source residuals only up to a reparametrization, which h must absorb. There is no covariate shift here, since the theorem concerns DTH and mixing the two would confound what is being checked. We use $n_{\mathrm{src}} = 4000$, $n_{\mathrm{cal}} = 5000$, $n_{\mathrm{test}} = 20000$ and 50 reps, giving a binomial noise floor of 0.007 at ten deciles, which we subtract.

For comparison, we also report the first-order delta-method approximation

$$\mathrm{CCE(WCP)} \;\approx\; f_Z\big(q_{1-\alpha}^{\mathrm{WCP}}/\bar{g}\big)\,\frac{q_{1-\alpha}^{\mathrm{WCP}}}{\bar{g}^{\,2}}\,\mathrm{sd}\big(g(D)\big), \qquad \bar{g} = \mathbb{E}[g(D)],$$

valid when $u \mapsto F_Z(q_{1-\alpha}^{\mathrm{WCP}}/u)$ is twice continuously differentiable on the range of g. This approximation describes the magnitude of the imbalance, whereas Theorem 3 gives the formal lower bound, which is a worst-case one and can be conservative when the range of g is wide.

Figure 2a plots the γ sweep from these runs and is discussed in Section 4.1. Table 5 gives the numbers behind it, including this first-order size; Table 6 gives the n_{tune} sweep at $\gamma = 2$. For panel (b), the driver is estimated within each run by dividing the test set into 50 equal-count D-bins and computing the one-way ANOVA between-group component of ℓ_T, corrected for within-bin noise by subtracting $(K-1)\,\mathrm{MSW}/N$, where K is the number of bins, N the number of test points and MSW the mean square within bins. Across the 200 runs of Table 2, an ordinary least-squares fit of the separation on the square root of this estimate has slope 0.24 (correlation $r = 0.54$).

Table 5: DTH-exact world, 50 reps, the numbers behind Figure 2. The bound of Theorem 3 against the exact population value it bounds and the first-order size above, with DC-CP at three tune budgets. $\gamma = 0$ is a null control with constant g, i.e. $\mathrm{Var}(g) = 0$, where DC-CP has nothing to correct and pays estimation error only.

			WCP		DC-CP, $n_{\text{tune}} =$		
γ	$\sqrt{\mathrm{Var}(g)}$	bound	exact	first order	100	400	1600
0	.000	.00000	.0000	.0000	.054	.030	.016
0.25	.036	.00615	.0109	.0109	.054	.029	.015
0.5	.072	.00633	.0194	.0196	.052	.027	.014
1	.144	.00333	.0318	.0325	.052	.024	.014
2	.289	.00035	.0460	.0483	.052	.023	.014
4	.577	.00000	.0585	.0635	.052	.024	.014

Table 6: DTH-exact world at $\gamma = 2$, 50 reps: the numbers behind Figure 2. WCP does not use the tune set. A3 is the centered log-ratio error $\mathbb{E}[(\log h - \log g - \log \hat{c})^2]$ with $\log \hat{c}$ the mean log-ratio, the quantity assumption (A3) drives to zero (here $\mathbb{E}[Z] = \sqrt{2/\pi}$, so the uncentred error would floor at 0.05).

n_{tune}	20	50	100	200	400	800	1600
CCE(DC-CP)	.104	.073	.052	.034	.023	.017	.014
CCE(WCP)	.037	.037	.037	.037	.037	.037	.037
A3	.242	.101	.052	.031	.028	.014	.011

C.2 Synthetic Ackley/Levy setup

$\mathcal{X} = [-3, 3]^4$. Functions Ackley/Levy. Source outputs $Y_S = f(X)$ and target outputs $Y_T = f(X)$. Source points are drawn from $\mathcal{N}(\mathbf{0}, 0.7^2 I)$ truncated to $\mathcal{X}$, and target fine-tuning points from $\mathcal{N}(0.3 \cdot \mathbf{1}, 0.7^2 I)$ truncated to $\mathcal{X}$. μ_S: multilayer perceptron (MLP; 2 ReLU layers, width 64) on $n_S = 2000$ source points (70% for μ_S, 30% held-out for fitting D). μ_T: 3-epoch fine-tune of μ_S on $n_T = 100$ target points. Body Table 2 draws calibration and tune from $P_{\text{cal},X} = U([-3, 3]^4)$ and test from $P_{\text{test},X} = \mathcal{N}(\mu_{\text{shift}}\mathbf{1}, I)$ at $\mu_{\text{shift}} = 1$, truncated to the box, with the true $w = P_{\text{test},X}/P_{\text{cal},X}$; Table 8 is the matched control with $P_{\text{test},X} = P_{\text{cal},X}$ and $w \equiv 1$. Cells are defined by a median-distance split, so the decomposition measures difficulty stratification rather than the importance weighting itself. $n_{\text{cal}} = 1000$, $n_{\text{tune}} = 100$; n_{cal} must be large enough that the weighted quantile does not fall on the $+\infty$ atom once the importance weights concentrate (effective calibration size ≈ 104 at $\mu_{\text{shift}} = 1$). Test size $n_{\text{test}} = 4000$, 100 i.i.d. seeds. The n_{tune} sweep of Table 10 (Section D.5) varies $n_{\text{tune}} \in \{10, 15, 20, 30, 50, 100, 200, 400\}$ at the same $n_{\text{cal}} = 1000$. The baselines share `scikit-learn` gradient-boosting defaults across all synthetic runs: CQR uses two `GradientBoostingRegressors` with `loss="quantile"` at $\alpha/2$ and $1 - \alpha/2$ (200 trees, max depth 3, learning rate 0.05), and the LWCP scale model $\hat{\sigma}(X)$ is a `GradientBoostingRegressor` (80 trees, max depth 3, learning rate 0.05) fitted on $\ell_T(Y_T, \mu_T(X))$ over the tune set, floored at 10^{-6}.

C.3 Polymer setup

Source pretrain. A multilayer perceptron ($190 \to 512 \to 256 \to 128 \to 64 \to 32 \to 1$, ReLU) is pretrained on $\approx 60{,}000$ molecular-dynamics simulated polymers from PolyOmics (Yoshida et al., 2025). WCP, LWCP and DC-CP use the same transferred target predictor μ_T; DC-CP additionally uses the fixed source-derived difficulty $D(X)$ of Equation (3), which is where μ_S enters. CQR instead fits target-side conditional quantiles directly. The WCP/LWCP/DC-CP comparison in Table 3 therefore isolates the calibration rule around a common target predictor, while CQR serves as a target-side adaptive baseline. The simulation-to-real transfer-learning scaling law of Anonymous (2026) identifies refractive index (along with radius of gyration) as the source whose transfer is strongest for nematic classification, and attributes this to aromatic polarizability being the electronic feature most predictive of nematogenic structure.

The T_m source. The T_m holdouts use a separately pretrained μ_S, with the same architecture and PolyOmics training set but TC_bond as the regression target. It is the bond-stretching term in the decomposition of the simulated thermal conductivity (Hayashi et al., 2022), and is among the simulated source properties with the strongest transfer performance for T_m (Anonymous, 2026).

Classifier fine-tuning. The final layer is reinitialized for classification, and the model is fine-tuned with BCEWithLogitsLoss, $\mathrm{pos_weight}$ set to the empirical $(1-p)/p$ ratio (with a multiplier 2), Adam learning rate 10^{-3}, batch size 32, and early stopping with patience 5 on the held-out target val set. Coverage in Table 3 is computed against the nematic phase label; if neither class is included, i.e. both nonconformity scores exceed the CP threshold, the argmax class is added to the set. This only enlarges the set, so it cannot lower coverage.

Data and splits. The Open LCP dataset (Anonymous, 2026) contains 457 polymers with nematic phase labels and PoLyInfo (Otsuka et al., 2011) `polymer_class`, with 190-dimensional kernel-mean force-field descriptors (Kusaba et al., 2023) per polymer. The descriptor is the kernel mean embedding of a polymer's force-field atom features, constructed by the published procedure of Kusaba et al. (2023); the descriptor table, the nematic labels and the class assignments are the inputs our pipeline consumes, and everything needed to reproduce the results reported here is specified in this appendix. Two classification holdouts are reported in Table 3: class 13 (polyimide, $n_{\text{test}} = 320$, positive-class rate 0.17); not_13 (reverse-OOD: train on polyimide, test on everything else, $n_{\text{test}} = 137$, positive-class rate 0.29). The two T_m regression holdouts are polyimide (class 13) and polyether (class 7). Classes are given by the 21-class PoLyInfo taxonomy as implemented by RadonPy's `polyinfo_classifier` (Hayashi et al., 2022); every holdout uses 30 seeds and a grouped split. Features use OpenLCP∪PoLyInfo; an OpenLCP-only check is in Section D.6, and its matched nematic class-13 reproduction agrees within 1 s.d. on every method.

Difficulty function and calibrator. $D(X)$ is a `scikit-learn` `GradientBoostingRegressor` (100 trees, max depth 3, learning rate 0.05, minimum 20 samples per leaf, seeded per split) fitted on $\mathcal{D}_S^{\text{diff}}$, regressing the source loss $\ell_S(Y_S, \mu_S(X)) = |Y_S - \mu_S(X)|$ on the same 190-d descriptors. The calibrator h is `IsotonicRegression` with `out_of_bounds="clip"` and automatic direction selection, its output floored at 10^{-6}. The LWCP scale model $\hat{\sigma}(X)$ is a `GradientBoostingRegressor` (80 trees, max depth 3, learning rate 0.05) fitted on the target score over the tune set, the absolute residual for T_m and $1 - \hat{p}_y$ for nematic classification, where it is therefore an analogue of LWCP rather than the original regression method; `min_samples_leaf` is at the library default: the classification tune sets fall to 14 examples, and the minimum leaf size of 20 used for D would force every tree to a single leaf, making $\hat{\sigma}$ constant and collapsing LWCP onto WCP by construction.

Importance weights. The class holdout induces a strong covariate-distribution shift in descriptor space, which we treat using importance weighting; all four methods use the same estimated $w(X) = P_{\text{test},X}(X)/P_{\text{cal},X}(X)$. For each seed, an L_2-regularized logistic discriminator (`C`$= 0.01$) is fitted on the standardized 190-d descriptors to separate the held-out class (label 1, the test covariates, which are observed) from the pooled calibration and tune covariates (label 0); its probabilities are clipped to $[10^{-3}, 1-10^{-3}]$, converted to a density ratio by $\hat{w} = \frac{p}{1-p} \cdot \frac{\pi_0}{\pi_1}$ with π_0, π_1 the class priors in the pooled sample, and normalized to mean 1 on the calibration set. The ratio is then clipped to $[1/u,\, u]$ with $u = \frac{1}{2}\, n_{\text{cal}}\, \alpha/(1-\alpha)$, which keeps the test atom p_{n+1} below α so that the weighted quantile never falls on $+\infty$; no cross-fitting is used. The same clipped weights enter WCP, CQR, LWCP and DC-CP, so differences between methods are not attributable to the importance-weight estimator. These are estimated and deliberately regularized importance weights, so the exact finite-sample coverage guarantee of Section 2.2 does not apply to the polymer experiment; the clipping serves numerical stability and a common comparison across methods, not the guarantee, which is checked with the true w on the synthetic benchmark (Table 2).

Tune-set correlation and CCE. The tune-set correlation $\bar{\rho}(D, \ell_T)$ is weak and of unstable sign: it averages -0.22 on polyimide and $+0.18$ on not_13, with s.d. 0.33 across seeds, and is negative in 70% of polyimide seeds against 33% of not_13 seeds (on the T_m holdouts it averages $+0.16$ and is negative in 17–20% of seeds). The monotone direction is thus a property of the individual tune set rather than of the holdout, which is why DC-CP selects it from the data. We do not report CCE on

this benchmark: the T_m test sets contain only 43 and 34 polymers, leaving three to four observations per D-decile and a binomial noise floor (~ 0.15) larger than the observed effect, and the tune sets are below the $n_{\text{tune}} \approx 200$ crossover of Section C.1, making the fitted isotonic h coarse in D.

C.4 SARCOS SETUP

We use the SARCOS robot-arm inverse-dynamics benchmark (Vijayakumar & Schaal, 2000; Rasmussen & Williams, 2006), following the transfer-learning construction of Minami et al. (2023): 21-dim joint state (position/velocity/acceleration of 7 joints) $\to$ torque at one chosen joint. The source model μ_S is an MLP (2 ReLU layers, width 64, 300 Adam iterations) predicting the other six joint torques (multi-output) from 28,000 training rows; the source-derived difficulty $D(X) \approx \mathbb{E}[\ell_S(Y_S, \mu_S(X)) \mid X]$, with $\ell_S(y, \hat{y}) = \|y - \hat{y}\|_2/\sqrt{6}$, is a `GradientBoostingRegressor` (200 trees, max depth 3, learning rate 0.05, minimum 20 samples per leaf) fitted on a 1000-row held-out source-residual split. Target torque models μ_T are kernel ridge regressors (`scikit-learn KernelRidge`, RBF kernel, $\lambda = 10^{-2}$, $\gamma = 1/21$) trained on $n_{\text{tgt}} \in \{5, 10, 20, 30, 50\}$ samples (no covariate shift; $w \equiv 1$). CQR and LWCP use the same gradient-boosting settings as on the synthetic benchmark. The purpose of SARCOS is not covariate-shift evaluation but task-to-task transfer of difficulty information: the source predictor is trained on a different output (other six torques), and DC-CP tests whether the resulting $D(X)$ helps locally calibrate the held-out target joint even when all CP weights are unit. The CP set ($n_{\text{cal}} = 100$) and tune set ($n_{\text{tune}} = 20$) are drawn from the remaining pool; 20 random seeds, evaluated on the standard SARCOS test set (4449 rows).

C.5 LLM SETUP

NQ-Open questions (Kwiatkowski et al., 2019; Lee et al., 2019) ($N = 3610$ from the `nq_open` validation split) are embedded with `intfloat/E5-large-v2` (Wang et al., 2022) and clustered into 15 k-means groups. Of these, six clusters (1099 questions) are held out as the target deployment set. LoRA fine-tuning of Llama 3.1 8B-Instruct uses $r = 8$ ($\alpha_{\text{LoRA}} = 16$, dropout 0.05), attention modules only, bf16 base weights, 3 epochs, AdamW at lr 2×10^{-4} with linear decay and no warm-up, batch size 8, sequence length 160, loss on answer tokens only; the nonconformity score is the mean per-token NLL of the model's own greedy answer, so it is a function of the input alone and the gold answers enter only through the exact-match numbers reported alongside. The fine-tuning (FT) split $\mathcal{D}_T^{\text{train}}$ has 300 in-source-topic examples and is fixed across all calibration-split seeds. The base-model entropy $D(X) = \mathbb{H}[p_{\text{base}}(\cdot \mid X)]$ is evaluated at the final token of the chat-template-rendered prompt (the position where greedy decoding would emit its first answer token); this requires a single forward pass of the base model and no generation.

Importance weights. The topic-cluster holdout is a covariate shift in embedding space, and all three methods use the same estimated $w(X)$. For each seed, an L_2-regularized logistic discriminator (`scikit-learn` default `C`$= 1$) is fitted on the L_2-normalized 1024-d E5-large-v2 text embeddings to separate the calibration covariates (label 0) from the held-out test covariates (label 1), which enter unlabelled as deployment covariates; the tune set does not enter the discriminator. Its log-odds are used directly as $\log w$ up to an additive constant, which cancels in the normalized weighted quantile of Equations (1) and (2), so no class-prior correction is applied. No probability or weight clipping is used and there is no cross-fitting: the discriminator is fitted on the same calibration and test points to which the weights are applied. The $+\infty$ atom is kept explicitly, and a test point whose weighted quantile falls on it is retained unconditionally; this happens for 0.05% of test points at six held-out clusters (footnote to Table 13). The effective calibration size under these weights is ≈ 190 of $n_{\text{cal}} = 440$. As on the polymer benchmark, these are estimated weights, so the exact coverage guarantee of Section 2.2 does not apply; WCP's marginal retention of 0.846 on NQ-Open is the corresponding empirical deviation.

Scope of the retention guarantee. The score does not use the gold answer, so the guarantee concerns the marginal probability of retention, not the correctness of retained answers; this differs from conformal factuality methods, which use factuality labels to control hallucination or factuality error (Hu et al., 2026). Thresholding a confidence score to abstain is the standard selective-prediction construction (Geifman & El-Yaniv, 2017; Hendrycks & Gimpel, 2017; Ren et al., 2023); what DC-CP calibrates is the difficulty profile of that abstention, not the accuracy of what survives it.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 SARCOS TASK-TO-TASK RESULTS

Table 7: SARCOS per-joint results at $n_{\text{tgt}} = n_{\text{tune}} = 20$, 20 seeds, nominal $1-\alpha = 0.9$. Bold marks the narrowest median width among entries with cov ≥ 0.895 (i.e. not under-covering). Full SARCOS protocol in Section C.4; analogous synthetic n_{tune} sweep in Table 10.

Joint	$\bar{\rho}$	WCP cov / w	CQR cov / w	LWCP cov / w	DC-CP cov / w
Torque1	+0.42	0.896 / 51.6	0.898 / 65.3	0.905 / 65.5	0.899 / **45.4**
Torque2	+0.26	0.889 / 34.4	0.894 / 47.8	0.908 / 39.3	0.897 / **32.9**
Torque3	+0.11	0.894 / 21.2	0.898 / 32.8	0.899 / **22.6**	0.898 / 25.1
Torque4	+0.38	0.897 / 29.1	0.906 / 44.7	0.892 / 26.6	0.901 / **25.9**
Torque5	+0.28	0.890 / 2.3	0.896 / 3.0	0.897 / **2.2**	0.899 / 2.8
Torque6	+0.28	0.912 / **4.6**	0.913 / 5.4	0.912 / 5.8	0.909 / 5.1
Torque7	+0.32	0.896 / **5.9**	0.896 / 8.3	0.911 / 6.3	0.904 / 6.4

At small target tune size $n_{\text{tune}} = 20$ all four methods reach marginal nominal coverage. DC-CP gives the narrowest width on the three high-magnitude joints (Torque1/2/4) where $D(X)$ has the largest dynamic range; WCP wins on Torque6/7 where intervals are essentially homoscedastic and adaptive methods only add noise; LWCP wins on Torque3 and Torque5 where the source channel is weakly informative. The no-shift synthetic sweep (Table 10) likewise shows that a well-supported target-side scale model can be preferable when it is fitted in the same region in which it is applied.

D.2 SYNTHETIC DECILE ANALYSIS

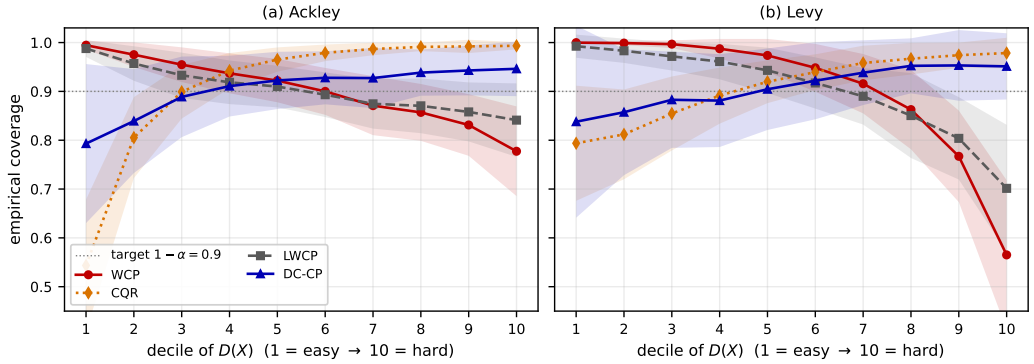

Figure 4: Empirical signature of DTH under covariate shift. Coverage by deciles of $D(X)$ at the Table 2 setting ($\mu_{\text{shift}} = 1$, true w, $n_{\text{cal}} = 1000$, $n_{\text{tune}} = 100$, 100 reps, $1-\alpha = 0.9$; curves are rep means, ribbons ± 1 s.d.). WCP drifts monotonically, over-covering at low D and under-covering at high D, as the DTH coverage $F_Z(q_{1-\alpha}^{\text{WCP}}/g(d))$ predicts for an increasing g, and the drift is severe on Levy (1.000 to 0.565). The plotted profiles carry the systematic part of CCE, on which DC-CP is the best method on Levy (0.041, against 0.065 for CQR, 0.088 for LWCP and 0.133 for WCP) and second on Ackley (0.048, against LWCP's 0.044), while over-correcting at the easy end there (0.793 on decile 1). The ribbons carry the rest, and they are widest for DC-CP. Section D.2 separates the two.

Figure 4 is drawn at the Table 2 setting, so the per-decile view and the two-cell table describe the same runs. Two things are visible. The monotone WCP drift that Theorem 3 predicts for an increasing g is present in both functions and is stronger on Levy, where the source signal is also more informative ($\bar{\rho} = 0.52$ against 0.27); this is qualitatively consistent with the mechanism, although $\bar{\rho}$ is a diagnostic rather than the quantity appearing in the bound.

Reading one number off the figure needs care, because CCE estimated per rep and then averaged, the convention of Table 10 and of the LLM tables, mixes two quantities. With $\widehat{\text{cov}}_k$ the observed

coverage in decile k of one rep, $\mathbb{E}[(\widehat{\mathrm{cov}}_k - (1-\alpha))^2]$ is a squared systematic bias plus $\mathrm{Var}(\widehat{\mathrm{cov}}_k)$, and the variance is genuine run-to-run variation of the fitted procedure plus binomial noise from the 400 test points per decile (s.d. 0.015 here, a minor term). Theorem 3 bounds the population coverage $\mathbb{P}(Y_T \in C_M(X) \mid D = d)$, so it is the systematic part that the theorem concerns; the per-replicate quantity additionally reflects the finite-sample variability of the fitted procedure. Separating them is the point of this figure:

	Ackley				Levy			
	WCP	CQR	LWCP	DC-CP	WCP	CQR	LWCP	DC-CP
systematic	.065	.135	**.044**	.048	.133	.065	.088	**.041**
run-to-run	.050	.053	**.047**	.080	.067	**.060**	.064	.101
per rep	.079	.141	**.064**	.086	.144	**.084**	.105	.102

DC-CP has the lowest systematic difficulty-stratified imbalance on Levy and is within 0.004 of the lowest on Ackley, consistent with the scale-cancellation mechanism of Theorem 3; it also has much the largest run-to-run spread, because h is refitted on 100 tune points in every rep. Adding the two, it ties LWCP on Levy and falls behind on Ackley, and CQR, whose wide intervals move little between reps, takes the per-rep column on Levy despite the worst systematic profile on Ackley. This is the same trade-off Figure 2a plots in the theorem's own coordinates, a systematic benefit that grows with the difficulty spread against an estimation price set by the tune budget, and at $n_{\mathrm{tune}} = 100$ that price is not yet small. It is also why the body claims a coverage advantage on the target-distance axis (Table 2) rather than a CCE advantage on this benchmark.

D.3 No-shift control and the source axis

Table 8 repeats the body diagnostic with $P_{\mathrm{test},X} = P_{\mathrm{cal},X}$ and $w \equiv 1$, and adds the full source×target 2×2. It serves two purposes. First, it is the control for the body claim: absent shift, the DC-CP advantage over LWCP disappears, and on Ackley LWCP is nominally the better method. Paired comparisons use the per-seed difference in $|\mathrm{cov} - 0.9|$, reported as mean ± standard error over seeds with the paired t statistic. Without shift the tgt OOD values are -0.002 ± 0.002 (t=−1.0) on Ackley and $+0.006 \pm 0.003$ (t=2.1) on Levy; under shift they are $+0.007 \pm 0.004$ (t=1.9) and $+0.038 \pm 0.005$ (t=8.1), with tgt ID values $+0.013 \pm 0.003$ (t=4.6) and $+0.028 \pm 0.003$ (t=8.1), and all four shifted comparisons against WCP give $t > 8$. Second, it shows why the body stratifies on the target axis alone. Since μ_S is fitted on $\mathcal{N}(\mathbf{0}, 0.7^2 I)$ and μ_T fine-tuned on $\mathcal{N}(0.3 \cdot \mathbf{1}, 0.7^2 I)$, the two distances are nearly collinear: the off-diagonal cells hold 6% of the mass each here (3% under shift) and the diagonal reproduces the tgt ID / tgt OOD split. Separating the two axes would require source and target regions that differ substantially, as in the polymer benchmark.

Table 8: No-shift control ($P_{\text{test},X} = P_{\text{cal},X}$, $w \equiv 1$), 100 seeds, same world and sample sizes as Table 2. Top: the target-only split used in the body. Bottom: the full source×target 2×2, where src ID/OOD splits the test set in the same way at the median distance to the source training center, cell mass in parentheses. **Bold** as in Table 2; it is spread across methods here, which is the point of the control.

Func	Cell	WCP cov/w	CQR cov/w	LWCP cov/w	DC-CP cov/w
Target-only split (as in the body)					
Ackley	tgt ID	.962±.01/3.65	**.884±.03**/3.83	.918±.02/2.76	.918±.02/3.27
	tgt OOD	.838±.02/3.65	**.914±.03**/3.56	.880±.02/3.94	.884±.03/4.22
	marginal	.900±.01/3.65	.899±.01/3.69	.899±.01/3.23	.901±.01/3.78
Levy	tgt ID	.997±.00/7.11	.966±.02/9.37	.937±.02/2.43	**.925±.03**/4.00
	tgt OOD	.800±.02/7.11	.834±.03/10.78	.864±.03/6.11	**.876±.03**/8.30
	marginal	.898±.01/7.11	.900±.01/9.77	.900±.01/4.11	.900±.01/6.29
Full source×target 2×2					
Ackley	src ID / tgt ID (0.44)	.970±.01/3.65	.870±.03/3.86	.923±.02/2.69	**.919±.02**/3.17
	src ID / tgt OOD (0.06)	.924±.03/3.65	.994±.01/3.70	.889±.04/3.25	**.908±.04**/3.69
	src OOD / tgt ID (0.06)	**.904±.05**/3.65	.996±.01/3.57	.884±.05/3.50	.906±.05/4.00
	src OOD / tgt OOD (0.44)	.827±.02/3.65	**.903±.03**/3.54	.879±.02/4.06	.881±.03/4.31
Levy	src ID / tgt ID (0.44)	.998±.00/7.11	.964±.02/9.34	.943±.02/2.30	**.929±.03**/3.80
	src ID / tgt OOD (0.06)	.906±.04/7.11	.890±.04/10.51	.866±.06/5.34	**.898±.05**/7.42
	src OOD / tgt ID (0.06)	.993±.01/7.11	.983±.03/9.53	.889±.07/3.45	**.899±.07**/5.34
	src OOD / tgt OOD (0.44)	.785±.02/7.11	.827±.03/10.81	.863±.03/6.21	**.873±.03**/8.40

D.4 TUNE-SET SIZE UNDER COVARIATE SHIFT

The body claims that the regime boundary is covariate shift rather than tune-set size. Table 9 supports this: it repeats the tgt OOD comparison of Table 2 at $\mu_{\text{shift}} = 1$ while sweeping n_{tune} over two decades. The gap to LWCP does not close as n_{tune} grows: Δ stays between $+0.026$ and $+0.030$ on Ackley and widens from $+0.040$ to $+0.065$ on Levy. The limitation is not the quantity of tune data but its location: every tune point is drawn from $P_{\text{cal},X}$, so $\hat{\sigma}(X)$ is fitted where the shifted test mass is sparse. This reverses the no-shift behavior of Table 10, where LWCP has the lower CCE at every tune size.

Table 9: Coverage on tgt OOD (nominal 0.9) under covariate shift ($\mu_{\text{shift}} = 1$, $n_{\text{cal}} = 1000$, 30 seeds) as the tune set grows. Δ is DC-CP − LWCP.

Func	n_{tune}	WCP	LWCP	DC-CP	Δ
Ackley	20	.839	.865	.896	+.030
	50	.842	.866	.892	+.026
	100	.841	.869	.894	+.026
	400	.841	.868	.897	+.030
Levy	20	.808	.819	.860	+.040
	50	.808	.827	.878	+.051
	100	.808	.825	.891	+.065
	400	.807	.832	.897	+.065

A practical caveat for reproducing these numbers: n_{cal} must be large enough that the weighted $(1-\alpha)$ quantile does not fall on the $+\infty$ atom. At $n_{\text{cal}} = 100$ the effective calibration size under this shift drops into the single digits and every method returns infinite intervals, making the cell coverages trivially 1; we therefore use $n_{\text{cal}} = 1000$ throughout the shifted experiments (effective size ≈ 104 at $\mu_{\text{shift}} = 1$).

D.5 n_{tune} SWEEP

Table 10 sweeps the target tune set size that both DC-CP (used to fit the 1-D isotonic h) and LWCP (used to fit a d=4-input GBM $\hat{\sigma}(X)$ on this synthetic) share, in the no-shift setting. Without shift every tune point lies where the test mass lies, so $\hat{\sigma}(X)$ is fitted in the region it is applied to, and LWCP has the lower CCE at every tune size in both functions. DC-CP closes the gap as n_{tune} grows ($0.106 \to 0.038$ on Ackley against LWCP's $0.053 \to 0.036$) but does not overtake, and its intervals stay wider on Levy. This is the control condition for Table 9: the ordering there is the reverse, and it is the shift, not the tune budget, that produces it.

Table 10: CCE over ten D-deciles, computed within each rep and averaged over reps (the "per rep" row of Section D.2), and median marginal width as n_{tune} varies, no shift, $n_{\text{cal}} = 1000$, 100 reps. Smaller is better for CCE; bold marks the lowest in each function.

	Ackley CCE / width				Levy CCE / width			
n_{tune}	WCP	CQR	LWCP	DC-CP	WCP	CQR	LWCP	DC-CP
10	.064/3.6	.115/4.3	**.053**/5.5	.106/5.0	.124/7.1	.114/15.3	**.081**/8.3	.105/10.7
15	.064/3.6	.110/4.5	**.051**/4.9	.104/5.6	.124/7.1	.107/16.0	**.076**/6.8	.100/10.9
20	.064/3.6	.108/4.4	**.046**/4.5	.097/4.6	.125/7.1	.105/16.2	**.074**/5.8	.099/7.1
30	.064/3.6	.091/4.3	**.042**/4.1	.087/4.1	.124/7.1	.106/14.0	**.067**/5.2	.093/6.8
50	.065/3.6	.094/4.1	**.041**/3.6	.073/3.9	.123/7.1	.105/12.8	**.063**/4.6	.080/6.5
100	.064/3.6	.090/3.7	**.039**/3.2	.058/3.8	.124/7.1	.097/9.8	**.053**/4.1	.068/6.3
200	.064/3.6	.080/3.2	**.037**/3.0	.045/3.7	.124/7.1	.087/6.6	**.047**/3.7	.057/6.0
400	.065/3.6	.068/2.6	**.036**/2.9	.038/3.6	.125/7.1	.076/4.4	**.043**/3.5	.050/6.0

D.6 OPENLCP-ONLY REPRODUCTION

This subsection re-runs the polymer experiments after stripping the proprietary PoLyInfo feature lookup, using public OpenLCP descriptors (Anonymous, 2026) only.

(a) T_m regression (class 9 holdout, smaller scale). The public OpenLCP T_m subset (116 polymers, no class 13) limits the regression reproduction to a class 9 holdout ($n_{\text{test}} = 11$, $n_{\text{pool}} = 105$); source = TC_bond on PolyOmics, features OpenLCP only, 30 seeds, $\bar{\rho}(D, \ell_T) = +0.11$.

Table 11: Public-data-only T_m regression at $1-\alpha = 0.9$ (OpenLCP class 9 holdout, $n_{\text{test}} = 11$, $n_{\text{pool}} = 105$, 30 seeds, source = TC_bond on PolyOmics, features from OpenLCP only). Coverage is mean $\pm$ s.d., width is the median over seeds of the per-seed median interval (Kelvin). Bold marks the entry closest to nominal. The small n_{test} makes per-seed coverage discretized in steps of $1/11 \approx 0.09$, hence the large s.d. on all methods.

WCP cov / w	CQR cov / w	LWCP cov / w	DC-CP cov / w
$0.830 \pm 0.155/208$	$0.855 \pm 0.119/264$	$\mathbf{0.882 \pm 0.166}/410$	$0.864 \pm 0.144/336$

All four methods under-cover at this small scale; DC-CP, CQR and LWCP all improve marginal coverage over WCP (LWCP closest to nominal at widest intervals, DC-CP balanced). At $n_{\text{test}} = 11$ this is a public-data consistency check rather than a reproduction; it runs in the same qualitative direction as the cls 7/cls 13 rows of Table 3.

(b) Nematic-phase classification (polyimide holdout, class 13). All 457 polymers used in the polyimide experiment are in OpenLCP (PoLyInfo overlaps on only 6 entries also in OpenLCP), so the holdout set is unchanged and only the feature lookup differs.

All three methods agree with the polyimide row of Table 3 within one s.d.: WCP 0.960, LWCP 0.931/0.930 (open-only / OpenLCP+PoLyInfo), DC-CP 0.919; $\bar{\rho}(D, \ell_T) = -0.22$ in both runs. The polymer classification result is therefore not driven by proprietary PoLyInfo entries.

Table 12: Public-data-only reproduction of the polyimide row of Table 3 at $1-\alpha = 0.9$. Source pretrain: PolyOmics. Target fine-tune: OpenLCP ($N = 457$, polyimide test $= 320$). Features: OpenLCP only (no PoLyInfo). 30 seeds, source $=$ refractive index. CQR is omitted (the binary nematic score $1 - \hat{p}_y$ is not a regression residual).

WCP cov	WCP size	LWCP cov	LWCP size	DC-CP cov	DC-CP size
0.960 ± 0.056	1.88	0.931 ± 0.130	1.85	$\mathbf{0.919 \pm 0.116}$	**1.83**

D.7 What does $h(D)$ actually learn?

These diagnostics show how the fitted isotonic link relates the source-derived difficulty $D(X)$ to the target nonconformity score. Under DTH (Section B.1), the population target of h is proportional to the conditional scale $g(D)$ (Section B.3) on which the lower bound of Theorem 3 depends; on the real benchmarks the plots should be read as empirical monotone trends rather than as evidence that DTH holds exactly. The cross-task polymer and LLM panels are especially informative: $D(X)$ and ℓ_T live in incomparable units (a dimensionless refractive-index residual vs. a binary classification score; base-model entropy in nats vs. a continuous nonconformity score), and yet $h(D)$ tracks the observed monotone trend in every case.

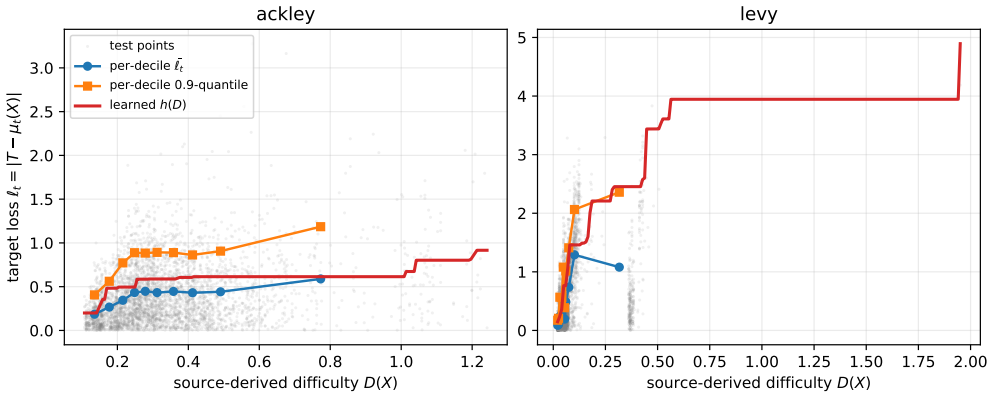

Figure 5: Synthetic Ackley/Levy, test set drawn from the shifted deployment law of Table 2. Test-set scatter (D, ℓ_T) with overlaid per-decile mean (blue), per-decile 0.9-quantile (orange), and learned isotonic h (red). h is fitted on the tune set, which is drawn from $P_{\mathrm{cal},X}$, and is applied here to the shifted test points: $h(D)$ still tracks the observed monotone trend on the shifted test set.

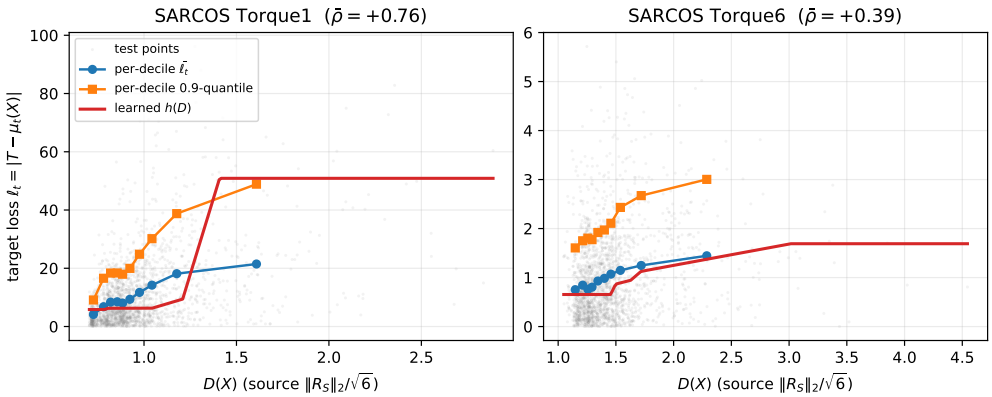

Figure 6: SARCOS robot arm, one of the 20 seeds behind Table 7 ($n_{\text{tgt}} = n_{\text{tune}} = 20$, $n_{\text{cal}} = 100$, same splits and models), for Torque 1 (source-informative) and Torque 6 (weaker). Each panel uses the seed whose tune-set $\bar{\rho}$ is closest to its 20-seed mean ($+0.42$ for Torque 1 and $+0.28$ for Torque 6, as in Table 7), chosen for that closeness rather than for the outcome. Even when source and target torques are different joints, $h(D)$ tracks the observed monotone trend in both panels.

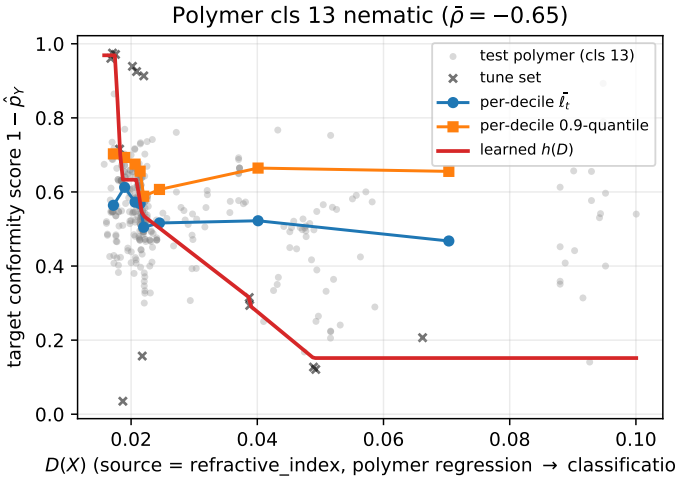

Figure 7: Polymer class 13 nematic classification, single seed (source = refractive index, target = nematic phase). Direction selection picks $\hat{\tau} = -1$ and the learned $h(D)$ tracks the observed monotone trend, with ℓ_S a dimensionless refractive-index absolute residual and $1 - \hat{p}_y$ a probability. This seed draws $\bar{\rho} = -0.65$ against the cross-seed mean -0.22 reported in Section C.3.

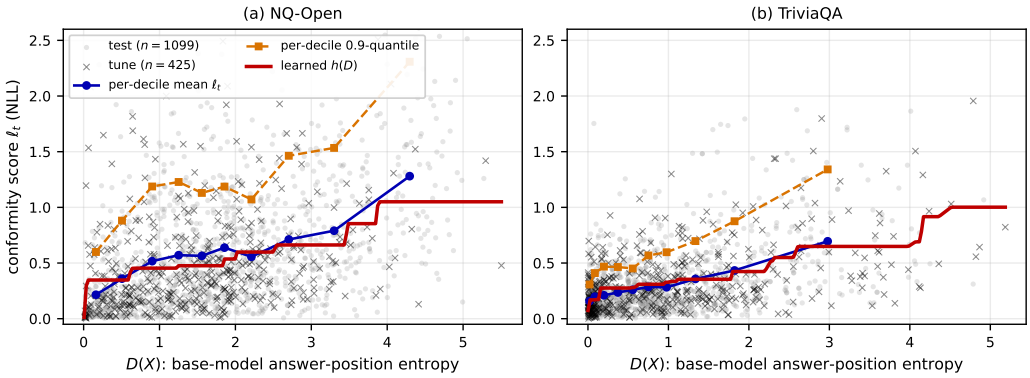

Figure 8: What $h(D)$ learns on the two QA benchmarks, LoRA arm, main configuration. $D(X)$ is base-model answer-position entropy and ℓ_T the mean per-token NLL of the generated answer. Direction selection chooses $\hat{\tau} = +1$ in both. (a) NQ-Open, seed 12: the per-decile mean score rises from 0.22 to 1.28 and the isotonic h tracks it. (b) TriviaQA, seed 11: the fine-tuned model is much stronger, so the scores are compressed into 0.16 to 0.70, yet the monotone link is if anything cleaner. Each panel uses the seed whose tune-set $\bar{\rho}$ is closest to its own 30-seed mean, 0.288 against 0.289 (s.d. 0.040) on NQ-Open and 0.500 against 0.500 (s.d. 0.033) on TriviaQA, chosen for that closeness rather than for the outcome.

D.8 LLM ABLATIONS

Split convention. All ablations use the same pipeline and split convention as the main experiment (§4.2): n_{cal}=440, tune and calibration sets drawn from the in-source-topic pool with the fine-tuning questions excluded, and the same 30 calibration seeds, so the main row of each panel reproduces the body numbers exactly; Figure 8 follows the same convention.

Table 13 sweeps three axes while holding n_{cal} and the split convention fixed at the body's values, so that only the swept axis moves.

The advantage over LWCP is flat in n_{tune} and stable across shift severity. The n_{tune} axis is the sharpest: LWCP never improves with more tune data, staying level with WCP from 200 to 1500 tune points, while DC-CP holds a 0.24–0.26 worst-decile lead throughout. More target-side data does not rescue the scale model, because its failure is extrapolation under shift rather than sample size.

The magnitude of the advantage depends strongly on the choice of D. Among the source-side definitions of D, it broadly tracks $|\bar{\rho}|$, being negligible at $\bar{\rho} \approx 0.10$, small at 0.15–0.20, large at 0.29 and largest at 0.55. The fine-tuned-model row additionally includes the fine-tuned model's own answer-position entropy as a target-derived diagnostic. It uses the same one-dimensional isotonic score form as DC-CP but lies outside the source-transfer setting that defines the proposed method, and we include it only as a diagnostic over the choice of D. It is more strongly associated with the target NLL ($\bar{\rho} = 0.69$) and still yields a large reduction in difficulty-stratified imbalance relative to WCP and LWCP under its own D-stratification, although it over-retains marginally (0.94 against 0.89 for the main row, at nominal 0.90). The score form is thus not tied to a particular definition of D; the main experiments use source-side uncertainty throughout.

Two further rows mark the limits of this dependence. The un-instructed-prompt row of panel (b) is the failure mode: with $\bar{\rho}$ close to zero the isotonic h is nearly but not exactly flat, and dividing by a noisy near-constant costs 0.04 of worst-decile retention relative to either baseline (-0.037, $t = -4.9$). The safety valve is thus a population-limit statement: at finite n_{tune} an uninformative D returns WCP plus the estimation noise in h, not WCP exactly. The effect is small and confined to this regime, but it is the practical reason to check $\bar{\rho}$ on the tune set before deploying. Conversely, the 4-shot chat variant has negative $\bar{\rho}$ and DC-CP still helps (t=8.8), the data-driven direction selection having picked the decreasing branch.

Table 13: LLM ablations, $\alpha = 0.1$, $n_{\text{cal}} = 440$, 30 seeds, mean over seeds. Cells give worst-decile retention / CCE; the last column gives the paired DC-CP − LWCP gap in worst-decile retention, with the paired t as a superscript. Deciles are those of the D named in each row, so the WCP and LWCP columns move with it down panel (b). Only the swept axis moves; n_{cal} and the split convention match the body throughout.

Setting	$\bar{\rho}$	WCP	LWCP	DC-CP	Δ vs LWCP^t
(a) n_{tune} sweep (D = answer-position entropy)					
200	+0.30	.484/.151	.542/.130	.783/.067	$+0.242^{26.0}$
400	+0.28	.497/.146	.537/.132	.782/.062	$+0.245^{23.3}$
425 *(main)*	+0.29	.500/.145	.535/.134	.795/.059	$+0.260^{23.3}$
800	+0.28	.503/.143	.538/.131	.798/.055	$+0.260^{27.3}$
1200	+0.28	.505/.143	.542/.131	.790/.059	$+0.248^{29.4}$
1500	+0.28	.489/.149	.531/.137	.785/.060	$+0.254^{33.6}$
(b) alternative difficulty signals ($n_{\text{tune}} = 425$)					
answer-position entropy *(main)*	+0.29	.500/.145	.535/.134	.795/.059	$+0.260^{23.3}$
fine-tuned-model entropy	+0.69	.348/.212	.411/.190	.827/.065	$+0.416^{35.3}$
prompt perplexity (question)	+0.15	.699/.094	.740/.079	.784/.061	$+0.044^{5.2}$
prompt perplexity (full prompt)	+0.18	.701/.091	.750/.075	.778/.064	$+0.028^{3.2}$
entropy, un-instructed prompt	+0.10	.788/.064	.785/.062	.748/.080	-0.037 ($t = -4.9$)
entropy, 4-shot plain	+0.55	.390/.192	.449/.170	.835/.059	$+0.386^{31.3}$
entropy, 4-shot chat	−0.24	.737/.096	.735/.090	.806/.052	$+0.071^{8.8}$
(c) shift severity (held-out topic clusters)					
2	+0.33	.428/.159	.457/.154	.806/.054	$+0.349^{19.8}$
6 *(main)*	+0.29	.500/.145	.535/.134	.795/.059	$+0.260^{23.3}$
8†	+0.21	.772/.057	.763/.061	.888/.035	$+0.125^{18.7}$

†At eight held-out clusters 6.5% of test points receive an infinite threshold and are retained unconditionally, which inflates marginal retention for all three methods; the fraction is 0.0005 at six clusters and 0.016 at two. Effective calibration size is 172–232 of $n_{\text{cal}} = 440$ in every row.

D.9 NQ-Open and TriviaQA

Everything but the dataset is held fixed. We take the `rc.nocontext` validation split of TriviaQA (Joshi et al., 2017) and subsample it to $n = 3610$ questions with a fixed seed, so that the two studies share the sample size and the comparison is not confounded by calibration-set size; answer aliases are kept and exact match is scored against the full alias set. The same E5-large-v2 text embedding model, the same 15-way k-means procedure with six held-out clusters ($n_{\text{test}} = 1099$), the same $n_{\text{tune}} = 425$ and $n_{\text{cal}} = 440$, the same LoRA recipe and the same 30 calibration seeds are used throughout.

Table 14: The two question-answering studies of Section 4.2, 30 seeds, mean ± s.d. Retention is at $1 - \alpha = 0.9$. The penultimate row gives the paired DC-CP − LWCP gap with its t statistic, and the last row gives the tune-set correlation $\bar{\rho}(D, \ell_T)$.

	NQ-Open		TriviaQA	
	worst decile	CCE	worst decile	CCE
WCP	.500±.036	.145±.016	.674±.038	.102±.007
LWCP	.535±.048	.134±.018	.675±.037	.099±.007
DC-CP	**.795±.056**	**.059±.015**	**.886±.027**	**.038±.010**
DC-CP − LWCP	$+.260^{t=23.3}$	$-.075^{t=-21.2}$	$+.211^{t=32.0}$	$-.061^{t=-26.4}$
$\bar{\rho}(D, \ell_T)$	+0.29		+0.50	

Two things are worth noting. The fine-tuned model is much stronger on TriviaQA (WCP-retained exact match 0.773 against 0.347; unconditional test exact match 0.743 against 0.306), so the nonconformity scores are more concentrated and there is less difficulty variation available; the separation survives anyway, and $\bar{\rho}$ is in fact larger. And LWCP is not merely close to WCP but numerically

indistinguishable from it, agreeing to three decimals on worst-decile retention and to four on exact match, which is the sharpest form of the observation that a target-side scale model fitted on the tune topics provides little transferable information about the score in the shifted topics. The diagnostic of Section 4.2 reproduces: $\hat{\sigma}$ correlates 0.950 with the score on the tune set and 0.043 on the shifted test set, while the correlation for $h(D)$ remains at 0.474. The coincidence is not a degenerate scale model: on NQ-Open $\hat{\sigma}(X)$ has coefficient of variation 0.24 on the test set, against 0.35 for $h(D)$, so it varies with the input but in a direction that no longer tracks the score.

E Extensions

E.1 CV-CP: Control-Variate Conformal Prediction

CV-CP is a complementary score-design tool for the narrow regime where paired source-target triples $(X_i, Y_{S,i}, Y_{T,i})$ are available and $Y_{S,n+1}$ is observed at test time. It is not part of the central DC-CP argument; we include it only to record that, when paired data exist, an even simpler additive correction works.

Let $R_{T,i} = Y_{T,i} - \mu_T(X_i)$, $R_{S,i} = Y_{S,i} - \mu_S(X_i)$. Partition the target calibration data into disjoint tune subset $\mathcal{D}_T^{\text{tune}}$ and CP subset $\mathcal{D}_T^{\text{cal}}$, and use a linear control variate (Owen, 2013):

$$\hat{\beta} = \left.\frac{\widehat{\text{Cov}}(R_T, R_S)}{\widehat{\text{Var}}(R_S)}\right|_{\mathcal{D}_T^{\text{tune}}}, \qquad V_i = |R_{T,i} - \hat{\beta} R_{S,i}|,$$

$$C(X_{n+1}) = \mu_T(X_{n+1}) + \hat{\beta}(Y_{S,n+1} - \mu_S(X_{n+1})) \pm \hat{q}_{1-\alpha}.$$

If $\hat{\beta}$ is computed from data disjoint from the calibration set and the test point and the true importance weight is used, CV-CP retains finite-sample marginal coverage by the argument of Section A.2. With $\rho = \text{Corr}(R_T, R_S)$ the variance of the optimally adjusted residual shrinks by the factor $(1-\rho^2)$, and CV-CP reduces to WCP at $\rho = 0$. The paired-(X, Y_S, Y_T) and observed-$Y_{S,n+1}$ requirements exclude CV-CP from the simulation-to-real transfer-learning polymer benchmark and the LLM study, so empirical tables are omitted.

F Reproducibility

F.1 Data availability

The synthetic Ackley/Levy benchmark, the SARCOS robot-arm dataset, and the NQ-Open and TriviaQA question sets used in the language-model studies are all public. The polymer benchmark of Table 3 uses PoLyInfo-derived feature records that cannot be redistributed publicly due to licensing restrictions; Section D.6 re-runs the polyimide nematic-phase row from public data only (OpenLCP + PolyOmics) and shows the claim is preserved. Both datasets required for reproduction are public: the OpenLCP descriptor table, nematic labels and class assignments used for the public-data reproduction are included in the anonymous supplementary material, and the PolyOmics simulated-polymer dataset used for source pretraining is publicly available (Yoshida et al., 2025).

F.2 Code availability

Code for all experiments is submitted as anonymous supplementary material and will be released as a public repository upon publication. It carries a single `requirements.txt`, the per-experiment drivers, and the per-seed outputs behind every table and figure, so each figure can be redrawn and every number checked without rerunning anything. The only dataset included is the public OpenLCP table (457 polymers, 49 KB), redistributed under its own permissive licence; the PolyOmics source properties are fetched from their public release, and the licensed PoLyInfo path is an alternative implementation that the code carries commented out, so the default run is the OpenLCP-only configuration of Section D.6. The language-model study ships no question, answer or model weight; only per-seed summaries.

F.3 Compute disclosure

All synthetic, polymer, SARCOS, and small-scale analysis experiments run on a single workstation with an Apple M-series chip (no discrete GPU); each table in this paper takes from a few minutes to a few hours on this hardware. The LLM experiment is the only one needing a GPU. On a single RTX 6000 Ada (48 GB), embedding and clustering the 3610 NQ-Open questions, scoring base-model entropy over all of them, LoRA fine-tuning, and greedy generation with NLL scoring together take ≈ 15 GPU-minutes; the TriviaQA study repeats the same pipeline at the same size. All per-question quantities are cached, so the conformal layer, the 30-seed splits and every ablation re-run on CPU in seconds. Total compute to reproduce the paper from scratch is well under one GPU-hour plus ≈ 8 workstation hours.

F.4 Ethics and broader impact

DC-CP is a calibration method that takes an existing predictor and a small target tune set and, with auxiliary models fitted on independent data and exact covariate-shift weights (Section 2.2), returns prediction sets with finite-sample marginal coverage. It does not generate, modify, or amplify the underlying predictor, and inherits the failure modes of its base predictor; conformal validity is a coverage guarantee, not a fairness, factuality, or robustness guarantee. The LLM extension uses a publicly released model (Llama 3.1 8B-Instruct) on public benchmarks (NQ-Open and TriviaQA); no human subjects, private data, or new datasets are introduced. The polymer benchmark uses public OpenLCP descriptors and PolyOmics simulations together with PoLyInfo metadata that is license-restricted; results that depend on the public-only subset are explicitly reported in Section D.6.