# PREFERENCE ALIGNMENT HELPS PROBABILISTIC FORECASTS ONLY WHEN THE DECISION-MAKER CANNOT ADAPT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Preference alignment has been proposed as a way to make forecasts serve downstream decisions, but it is not known when, or whether, distorting a forecast can actually improve a decision. We settle this question for probabilistic forecasting, and the answer is sharply two-sided. If the decision-maker is *adaptive*, choosing the action that minimizes expected cost under the reported distribution, then reporting the true conditional distribution is optimal and *no aligned forecast can ever do better*. For newsvendor-type decisions, which cover the standard motivating examples of safety stock and reserve sizing, this is exact: minimizing expected cost and querying the critical-ratio quantile are the same operation, so calibration already suffices. If instead the decision-maker is *non-adaptive*, applying a fixed readout it cannot re-tune to the cost structure (most commonly a system that simply consumes the point forecast), then the cost-minimizing report is provably *miscalibrated*, and alignment has real value. We instantiate the positive regime with **Prob-TPO**, which aligns full quantile functions with a distributional DPO objective, and confirm both predictions on four benchmarks (energy, retail, transformer temperature, electricity load; 5 seeds, out-of-sample). Under a fixed point-forecast readout, Prob-TPO reduces realized newsvendor cost by 6–43% relative to calibrated quantile regression, closing up to 92% of the gap to a utility-optimal oracle, while an adaptive consumer sees no benefit, exactly as the theory predicts. We further show the auxiliary and calibration weights control how much of this gain is captured, and that alignment silently breaks quantile-label semantics: an aligned "95th percentile" carries 99% coverage, so downstream consumers must re-map labels.

## 1 INTRODUCTION

Consider an inventory planner preparing replenishment orders for the upcoming week. A point forecast of 500 units tells her *what to expect*, but her decision depends on the *upper tail* of the predictive distribution: she must set safety stock to prevent stockouts with, say, 95% probability. If the 95th percentile forecast is too high, she incurs unnecessary holding costs; if too low, stockouts erode customer trust and revenue. The pinball loss that produced the forecast treats all quantile errors symmetrically, yet stockouts cost the business 3–5$\times$ more than excess inventory Ben-Tal et al. (2009).

This scenario is pervasive. Energy grid operators size reserves from prediction intervals for renewable generation, where underestimation risks blackouts at $10\times$ the cost of overestimation Hong et al. (2016). Financial risk managers compute Value-at-Risk from lower quantiles of return distributions Ben-Tal et al. (2009). In every case, the *entire predictive distribution* drives downstream decisions, and the business consequences of errors vary across quantiles.

Recent work has begun to address the misalignment between forecasting objectives and decision quality. Decision-focused learning Elmachtoub & Grigas (2022); Donti et al. (2017) integrates downstream optimization into training, but requires a fully specified and differentiable objective. Temporal Preference Optimization (TPO; Anonymous (2026)) showed that point forecasts can be aligned with decision-maker preferences through DPO Rafailov et al. (2023) and a temporal utility decomposition, achieving substantial cost reductions without requiring explicit cost specification. However, TPO operates exclusively on point predictions. When the downstream decision depends



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.

on prediction intervals, tail quantiles, or the full predictive distribution, point-forecast alignment is insufficient.

Extending preference alignment from point to distributional forecasts raises a fundamental question that has no analogue in point forecasting or language modeling: *should the aligned distribution remain statistically calibrated?* A calibrated 95th percentile should exceed the realized value 95% of the time. But if underestimation is costlier than overestimation, the utility-optimal 95th percentile is systematically higher than the calibrated one, improving decisions at the expense of statistical reliability. Practitioners navigate this tension implicitly; no existing framework makes it explicit or controllable.

We propose **Prob-TPO** (Probabilistic Temporal Preference Optimization) to address this gap. Our contributions are:

1. A *characterization of when preference alignment helps* (Section 3.3). For adaptive consumers, calibration is sufficient and alignment provably cannot help (Proposition 7); for newsvendor-type decisions this is exact, since minimizing expected cost and querying the critical-ratio quantile coincide (Corollary 8). For non-adaptive consumers with a fixed readout, the cost-minimizing report is provably miscalibrated (Proposition 9). This delimits the entire design space and explains why point-forecast alignment Anonymous (2026) works.
2. A *distributional utility framework and method*, Prob-TPO, that instantiates the positive regime: preferences over full predictive distributions via quantile-specific asymmetric costs, optimized with a distributional DPO objective over non-crossing quantile functions (Sections 3.1 and 4). We also show that under asymmetric costs no forecaster can be both calibrated and utility-optimal (Proposition 5).
3. *Empirical confirmation of both regimes* on four benchmarks, out-of-sample over 5 seeds (Section 5). Under a fixed point-forecast readout Prob-TPO cuts realized newsvendor cost by 6–43% against calibrated quantile regression, capturing up to 92% of the oracle headroom; under an adaptive consumer the advantage vanishes, as predicted. We identify the auxiliary and calibration weights as the knob governing how much of the gain is captured, and document that alignment breaks quantile-label semantics, a practical caveat with a concrete remedy.

## 2 Related Work

**Probabilistic forecasting and calibration.** Deep probabilistic forecasters produce predictive distributions via autoregressive likelihoods (DeepAR Salinas et al. (2020)), attention-based quantile heads (TFT Lim et al. (2021)), diffusion Rasul et al. (2021) or patching transformers (PatchTST Nie et al. (2023)), estimated with quantile regression Koenker & Bassett Jr (1978) and judged by proper scoring rules Gneiting & Raftery (2007); Gneiting & Katzfuss (2014). Calibration methods (recalibration Kuleshov et al. (2018), conformal prediction Romano et al. (2019); Vovk et al. (2005); Angelopoulos & Bates (2023) and interval scores Bracher et al. (2021)) ensure statistical validity but treat coverage violations symmetrically. All of these optimize decision-agnostic objectives. Our analysis explains what that costs: nothing at all for an adaptive consumer (Proposition 7), but 6–43% of realized decision cost for a fixed-readout one (Proposition 9).

**Decision-focused learning.** Smart Predict-then-Optimize Elmachtoub & Grigas (2022), task-based learning Donti et al. (2017) and the surrounding literature Mandi et al. (2024) fold downstream optimization into training, but require a specified differentiable program and mostly operate on point predictions. The classical newsvendor result Ben-Tal et al. (2009), that cost-optimal quantiles differ from nominal ones, underlies both that literature and our Corollary 8. Our contribution is orthogonal and delimiting: we identify when *any* such distortion of a probabilistic forecast can improve a decision, and show the answer depends on the consumer rather than the training method.

**Preference optimization.** RLHF Ouyang et al. (2022); Christiano et al. (2017) and DPO Rafailov et al. (2023), with variants such as IPO Azar et al. (2024) and Constitutional AI Bai et al. (2022), align generative models with preferences; TPO Anonymous (2026) brought this to point forecasting. Extending it to distributions raises a question with no analogue in language modelling: whether the

aligned object should remain calibrated. Section 3.3 shows why point-forecast alignment works: a point-forecast consumer is exactly the fixed-readout case.

## 3 Problem Formulation

We consider a probabilistic forecasting model $\hat{F}_\theta$ that, given historical context $x_{1:L}$, produces a predictive distribution over future values $y_{1:H}$ at horizon $H$. We represent the predictive distribution through its quantile function: for a finite set of quantile levels $\mathcal{Q} = \{q_1, \ldots, q_Q\} \subset (0, 1)$, the model outputs $\hat{F}^{-1}(q)$ for each $q \in \mathcal{Q}$ and each horizon $h$. A decision-maker operates in a domain characterized by a cost context $\mathcal{C}$ that encodes the asymmetric consequences of forecast errors at each quantile.

### 3.1 Distributional Utility

We define the distributional utility of a predictive distribution $\hat{F}$ evaluated against realized outcome $y^*$ under cost context $\mathcal{C}$ as:

$$U(\hat{F}, y^*, \mathcal{C}) = \sum_{q \in \mathcal{Q}} w(q) \cdot u_q\Big(\hat{F}^{-1}(q)\,,\; y^*,\; \mathcal{C}\Big)\,, \tag{1}$$

where $w(q) \geq 0$ are quantile importance weights with $\sum_q w(q) = 1$, and the per-quantile utility is:

$$u_q\Big(\hat{F}^{-1}(q)\,,\; y^*,\; \mathcal{C}\Big) = -\alpha_q^- \left[y^* - \hat{F}^{-1}(q)\right]_+ - \alpha_q^+ \left[\hat{F}^{-1}(q) - y^*\right]_+\,, \tag{2}$$

with $\alpha_q^-, \alpha_q^+ > 0$ being the quantile-specific costs of under- and over-prediction, and $[\cdot]_+ = \max(\cdot, 0)$. For lower quantiles ($q < 0.5$), overestimation is typically more costly ($\alpha_q^+ > \alpha_q^-$); for upper quantiles ($q > 0.5$), underestimation is more costly ($\alpha_q^- > \alpha_q^+$). The quantile weights $w(q)$ reflect domain-specific priorities: a retailer may assign high weight to the 95th percentile (safety stock), while an energy trader may emphasize both tails (peak and trough planning).

**Definition 1** (Distributional Utility)**.** *The multi-horizon distributional utility for a forecast $\hat{F}_{1:H}$ evaluated against outcomes $y^*_{1:H}$ is:*

$$U(\hat{F}_{1:H}, y^*_{1:H}, \mathcal{C}) = \sum_{h=1}^{H} \gamma^{h-1} \sum_{q \in \mathcal{Q}} w(q) \cdot u_q\Big(\hat{F}_h^{-1}(q),\; y^*_h,\; \mathcal{C}\Big)\,, \tag{3}$$

*where $\gamma \in (0, 1]$ is a temporal discount factor that controls the relative importance of near- versus far-horizon accuracy.*

This definition generalizes the point-forecast utility of TPO Anonymous (2026) by replacing the single-valued forecast with a weighted sum over quantile-specific costs. The additive structure across horizons and quantiles is a direct consequence of the per-quantile utility depending only on $(\hat{F}_h^{-1}(q), y^*_h)$; we state this formally for completeness.

**Theorem 2** (Optimal Quantile Characterization)**.** *The multi-horizon distributional utility decomposes as:*

$$U(\hat{F}_{1:H}, y^*_{1:H}, \mathcal{C}) = \sum_{h=1}^{H} \gamma^{h-1} U_h(\hat{F}_h, y^*_h, \mathcal{C}), \tag{4}$$

*where $U_h(\hat{F}_h, y^*_h, \mathcal{C}) = \sum_{q \in \mathcal{Q}} w(q) \cdot u_q(\hat{F}_h^{-1}(q), y^*_h, \mathcal{C})$ is the single-horizon distributional utility. The utility-maximizing quantile forecast at level $q$ and horizon $h$ satisfies:*

$$\hat{F}_h^{*-1}(q) = F_{Y_h}^{-1}\left(\frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+}\right), \tag{5}$$

*where $F_{Y_h}$ is the true conditional distribution of $Y_h$ given context.*

**Remark 3** (Newsvendor connection)**.** *Equation* (5) *is a multi-quantile generalization of the newsvendor critical ratio Ben-Tal et al. (2009): unless $\alpha_q^-/(\alpha_q^- + \alpha_q^+) = q$, the utility-optimal forecast differs from the calibrated one, which is the source of the tension in Proposition 5.*

The decomposition follows from additive separability; the optimal quantile from the first-order condition on $\mathbb{E}[u_q]$. See Appendix A.

### 3.2 CALIBRATION-PREFERENCE TRADE-OFF

A natural question is whether preference-aligned distributional forecasts can remain well-calibrated. We formalize this tension.

**Definition 4** (Expected Calibration Error)**.** *The expected calibration error (ECE) of a quantile forecaster $\hat{F}$ is:*

$$\mathrm{ECE}(\hat{F}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left| \Pr\left(Y \leq \hat{F}^{-1}(q)\right) - q \right|. \tag{6}$$

*A forecaster is perfectly calibrated if $\mathrm{ECE}(\hat{F}) = 0$, meaning the empirical coverage at each quantile level matches its nominal level.*

**Proposition 5** (Calibration-Preference Trade-off)**.** *Let $\mathcal{C}$ be a cost context with asymmetric quantile costs satisfying $\alpha_q^- \neq \alpha_q^+$ for at least one $q \in \mathcal{Q}$. Then no forecaster $\hat{F}$ can simultaneously achieve $\mathrm{ECE}(\hat{F}) = 0$ and maximize the distributional utility $U(\hat{F}, Y, \mathcal{C})$ in expectation, unless the true data-generating distribution $F_Y$ satisfies:*

$$F_Y^{-1}(q) = F_Y^{-1}\left(\frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+}\right) \quad \textit{for all } q \in \mathcal{Q} \textit{ with } \alpha_q^- \neq \alpha_q^+. \tag{7}$$

Calibration requires $\hat{F}^{-1}(q) = F_Y^{-1}(q)$, while utility maximization requires $\hat{F}^{-1}(q) = F_Y^{-1}(\alpha_q^-/(\alpha_q^- + \alpha_q^+))$; these differ when $\alpha_q^- \neq \alpha_q^+$, coinciding only under the degenerate condition (7). See Appendix A.

**Remark 6** (Severity of the Trade-off)**.** *The magnitude of the gap between calibrated and utility-optimal quantiles depends on two factors: the degree of cost asymmetry and the shape of $F_Y$ near the relevant quantile levels. For the 95th percentile with a $3\times$ underage-to-overage cost ratio, the optimal level shifts to $F_Y^{-1}(0.75)$, a substantial departure from calibration. In domains with near-symmetric costs (e.g., $\alpha_q^-/\alpha_q^+ \approx 1$), the tension is mild and calibration regularization is less necessary. This motivates a* tunable *regularizer rather than a fixed calibration constraint (Section 4.2).*

### 3.3 WHEN CAN PREFERENCE ALIGNMENT HELP?

Proposition 5 says aligned forecasts cannot stay calibrated. It does not say whether the resulting distortion *helps* a decision-maker. We now settle that question, and the answer is sharply two-sided: it depends entirely on how the forecast is consumed.

Consider a downstream decision problem with action space $\mathcal{A}$ and cost $c(a, y)$, and a forecaster that reports a predictive distribution $\hat{F}$ which the consumer turns into an action.

**Proposition 7** (Calibration is sufficient: a no-go for adaptive consumers)**.** *Suppose the consumer is* adaptive*, i.e. it selects $a^\star(\hat{F}) \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{y \sim \hat{F}}[c(a, y)]$. Then reporting the true conditional $\hat{F} = F_Y$ minimizes the realized expected cost $\mathbb{E}_{y \sim F_Y}[c(a^\star(\hat{F}), y)]$ among all reports. No preference-aligned forecast can outperform a calibrated one.*

The proof is immediate: $\mathbb{E}_{y \sim F_Y}[c(a^\star(F_Y), y)] = \min_a \mathbb{E}_{y \sim F_Y}[c(a, y)] \leq \mathbb{E}_{y \sim F_Y}[c(a, y)]$ for every action $a$, in particular for $a = a^\star(\hat{F})$.

**Corollary 8** (Newsvendor decisions are quantile-representable)**.** *For $c(a, y) = c_u[y-a]_+ + c_o[a-y]_+$ we have $a^\star(\hat{F}) = \hat{F}^{-1}(\tau^\star)$ with $\tau^\star = c_u/(c_u + c_o)$. Hence "minimize expected cost under $\hat{F}$" and "query $\hat{F}$ at level $\tau^\star$" are the* same *operation, and a calibrated forecaster queried at $\tau^\star$ is already optimal.*

Corollary 8 covers every motivating example in Section 1, so for an adaptive consumer the case for alignment collapses. The value of alignment lies elsewhere.

**Proposition 9** (Fixed readouts make distortion optimal)**.** *Suppose instead the consumer is* non-adaptive*: it applies a fixed readout $a = \hat{F}^{-1}(\tau_0)$ at a level $\tau_0$ it cannot adapt to the cost structure (e.g. a system that consumes the point forecast, $\tau_0 = 0.5$, or a mandated percentile). Then the cost-minimizing report satisfies*

$$\hat{F}^{-1}(\tau_0) = F_Y^{-1}(\tau^\star). \tag{8}$$

*Whenever $\tau_0 \neq \tau^\star$ this report is* not *calibrated, and it strictly improves the realized decision over reporting $F_Y$, by $\mathbb{E}[c(F_Y^{-1}(\tau_0), y)] - \mathbb{E}[c(F_Y^{-1}(\tau^\star), y)] > 0$ unless $F_Y^{-1}(\tau_0) = F_Y^{-1}(\tau^\star)$.*

Together, Propositions 7 and 9 characterize when preference alignment has value: *precisely when the consumer's readout is misaligned with the cost structure and cannot be changed*. This is the common case in deployed forecasting systems, where a model feeds a downstream pipeline, contract, or regulatory formula that consumes a fixed summary, most often a point forecast, and cannot be rewritten to query a cost-dependent quantile. It also explains why point-forecast alignment Anonymous (2026) works: a point-forecast consumer is exactly the $\tau_0 = 0.5$ case of Proposition 9. Section 5 confirms both predictions empirically.

### 3.4 Preference Model

Given two predictive distributions $\hat{F}_a$ and $\hat{F}_b$ evaluated against the same realized outcome $y^*$ under cost context $\mathcal{C}$, we model the probability that a decision-maker prefers $\hat{F}_a$ over $\hat{F}_b$ via the Bradley-Terry model Bradley & Terry (1952):

$$P(\hat{F}_a \succ \hat{F}_b \mid y^*, \mathcal{C}) = \sigma\Big(U(\hat{F}_a, y^*, \mathcal{C}) - U(\hat{F}_b, y^*, \mathcal{C})\Big), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. This extends the point-forecast preference model of TPO Anonymous (2026) to distributional forecasts: preference is determined by the distributional utility difference across all quantile levels.

## 4 Method: Prob-TPO

### 4.1 Architecture: Quantile Temporal Transformer

We build on the PatchTST Nie et al. (2023) architecture with three key modifications for quantile forecasting. The Quantile Temporal Transformer consists of:

1. **Patch Embedding.** The input time series $x_{1:L} \in \mathbb{R}^{L \times d}$ is split into non-overlapping patches of length $P$ and linearly projected to hidden dimension $d_h$.
2. **Positional Encoding & Transformer Encoder.** Standard sinusoidal positional encoding followed by $N_L$ transformer encoder layers with multi-head self-attention ($N_H$ heads) and feed-forward networks.
3. **Non-crossing quantile heads.** Instead of a single forecast head, we use three parallel heads:
   - **Median head:** produces $\hat{m}_h = \hat{F}^{-1}(0.5)$ for $h = 1, \ldots, H$.
   - **Upper increment head:** produces raw increments $\delta^+_{h,k}$ for upper quantiles $q > 0.5$.
   - **Lower increment head:** produces raw increments $\delta^-_{h,k}$ for lower quantiles $q < 0.5$.

**Non-crossing guarantee.** Following Chernozhukov et al. (2010); Cannon (2018), quantiles above the median are formed as the median plus a cumulative sum of Softplus increments, and those below it by subtracting one; since $\mathrm{Softplus}(\cdot) > 0$, the resulting quantile function is strictly increasing by construction, requiring no post-hoc sorting and remaining differentiable throughout. Full equations and a schematic are given in Appendix C.

### 4.2 Distributional DPO Loss

The Prob-TPO training loss combines three components:

$$\mathcal{L} = (1 - w_{\text{aux}}) \cdot \mathcal{L}_{\text{DPO}} + w_{\text{aux}} \cdot \mathcal{L}_{\text{pinball}} + \lambda_{\text{cal}} \cdot \mathcal{L}_{\text{cal}}, \tag{10}$$

where $w_{\text{aux}} \in [0, 1)$ controls the auxiliary quantile regression weight and $\lambda_{\text{cal}} \geq 0$ controls the calibration penalty.

**Distributional DPO loss.** Given a policy model $\pi_\theta$ and a frozen reference model $\pi_{\text{ref}}$, both producing quantile forecasts for the same input, we compute:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\log\sigma\left(\beta\cdot\frac{U(\hat{F}^{\theta},y,\mathcal{C})-U(\hat{F}^{\text{ref}},y,\mathcal{C})}{s}\right)\right], \tag{11}$$

where $\beta>0$ is a temperature parameter, $s=\frac{1}{2}(|C^{\theta}|+|C^{\text{ref}}|)+\epsilon$ is a normalizing scale ($C=-U$ denotes cost), and $\epsilon>0$ prevents division by zero.

**Relationship to standard DPO.** Standard DPO Rafailov et al. (2023) and the point-forecast TPO Anonymous (2026) contrast explicit winner and loser outputs. In the distributional setting, we instead contrast the policy against its own pretrained reference: the loss encourages the policy to improve its distributional utility relative to the pinball-pretrained baseline. This departure is motivated by a practical consideration: generating meaningful distributional preference pairs (where two *complete quantile functions* are compared) is substantially harder than comparing point forecasts. The reference model provides a natural "dispreferred" baseline, as it is optimized for calibration (via pinball loss) but not for decision utility. The frozen reference is also the KL anchor analogous to RLHF Ouyang et al. (2022), preventing the policy from deviating too far from a statistically well-calibrated starting point. We note this formulation is closer to self-play improvement Azar et al. (2024) than to classical DPO with explicit preference pairs, and discuss implications in Section 6.

**Auxiliary pinball loss.** To prevent distributional collapse during early DPO training, we include a standard pinball (quantile regression) loss:

$$\mathcal{L}_{\text{pinball}} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\rho_q\Big(y-\hat{F}^{-1}(q)\Big), \quad \rho_q(u)=u\cdot(q-\mathbb{1}[u<0]), \tag{12}$$

where $\rho_q$ is the check function Koenker & Bassett Jr (1978). This term anchors the quantile forecasts to statistically reasonable values while the DPO objective shifts them toward decision-optimal positions.

**Calibration regularizer.** To manage the calibration-preference trade-off (Proposition 5), we add a differentiable calibration penalty:

$$\mathcal{L}_{\text{cal}} = \sum_{q\in\mathcal{Q}}\left|\frac{1}{BH}\sum_{i=1}^{B}\sum_{h=1}^{H}\sigma\Big(T\cdot(\hat{F}_{i,h}^{-1}(q)-y_{i,h})\Big)-q\right|, \tag{13}$$

where $T>0$ is a temperature for the soft coverage indicator, $B$ is the batch size, and $H$ is the forecast horizon. As $T\to\infty$, the soft indicator approaches the hard indicator $\mathbb{1}\Big[\hat{F}^{-1}(q)\geq y\Big]$, and $\mathcal{L}_{\text{cal}}$ converges to the ECE. In practice, we use $T=10$ for smooth gradients.

**Training pipeline.** Training proceeds in three stages: pinball pretraining to a calibrated quantile forecaster; cloning and freezing that model as the reference $\pi_{\text{ref}}$; and distributional DPO fine-tuning with the composite loss (10). The reference both anchors the objective, analogously to the KL penalty in RLHF Ouyang et al. (2022), and supplies the baseline utility in (11). Pseudocode is given in Appendix C.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** We evaluate on four public domains with distinct cost asymmetries: **UCI Household Energy** Hébrail & Bérard (2012) ($\alpha^{-}{=}10\alpha^{+}$ at upper quantiles, where under-forecasting peak demand threatens grid stability Hong et al. (2016)), **UCI Online Retail II** Chen et al. (2012) (16 country-level revenue series, $\alpha^{-}{=}3\alpha^{+}$ for stockout-vs-holding costs), **ETTh1** Zhou et al. (2021) (17,420 hourly steps, $\alpha^{-}{=}5\alpha^{+}$ since under-forecasting transformer oil temperature risks equipment damage), and **Electricity/ECL** Lai et al. (2018) (40 client load series, $\alpha^{-}{=}10\alpha^{+}$). The cost ratios are modelling assumptions fixed a priori; we report a sensitivity to an alternative ratio on every dataset. All results use a chronological $80/20$ holdout in which no evaluation target is seen during training; series construction, features and splits are detailed in Appendix B.2.

**Baselines.** We compare against **Pinball** (quantile regression with the check loss Koenker & Bassett Jr (1978) on the same architecture), **DeepAR** Salinas et al. (2020) (autoregressive Gaussian likelihood), **CQR** Romano et al. (2019) (conformalized quantile regression), and **Point-TPO** Anonymous (2026) (median-only alignment, quantiles from a separate pinball model). We add two oracles: a **calibrated oracle** (domain-weighted pinball at nominal levels) and the **utility-optimal oracle**, which targets the cost-optimal shifted levels $\alpha_q^- / (\alpha_q^- + \alpha_q^+)$ of Theorem 2. The latter is learned from training data, so it approximates the achievable lower bound rather than attaining it exactly.

**Metrics.** Our primary metric is the **realized decision cost** of an actual ordering decision (Section 5.3), which shares no parameters with any training objective. We additionally report the composite **decision cost** of Definition 1, **CRPS** Gneiting & Raftery (2007), **calibration error** (mean absolute deviation between empirical and nominal coverage), the **Winkler** interval score Winkler (1972), **sharpness** and **coverage** at the $90\%$ level.

**Implementation details.** All transformer models use $d_h = 128$, $N_H = 4$ heads, $N_L = 2$ layers, patch length $P = 16$, and quantile levels $\mathcal{Q} = \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. We pretrain for $E_{\text{pre}} = 20$ epochs and fine-tune for $E_{\text{dpo}} = 10$ epochs, both with learning rate $10^{-4}$ and batch size 32. Default hyperparameters: $\beta = 0.5$, $\gamma = 0.95$, $w_{\text{aux}} = 0.3$, $\lambda_{\text{cal}} = 0.1$. All results are averaged over 5 random seeds (42, 123, 456, 789, 2024) and reported as mean $\pm$ standard deviation. Additional details in Appendix B.

**Decision cost and the utility-optimal oracle.** Throughout, *decision cost* denotes the negative distributional utility $-U(\hat{F}, y^*, \mathcal{C})$ of Definition 1, evaluated with the *same* cost context $\mathcal{C}$ used during training: the domain asymmetry ratio ($10\times$ for Energy, $3\times$ for Retail; Table 2) and domain-specific quantile weights. Reporting cost under the exact objective the model optimizes is essential: an evaluation cost with different weights or asymmetry would score every forecaster on a decision problem it was never asked to solve. Because the asymmetric cost admits a *finite* minimizer (over-forecasting is penalized at rate $\alpha_q^+ > 0$), it cannot be reduced by unboundedly inflating quantiles; the minimum is attained by the utility-optimal oracle, which forecasts the conditional quantile at the cost-optimal level $\alpha_q^- / (\alpha_q^- + \alpha_q^+)$ (Theorem 2) and lower-bounds the decision cost of every method in Table 3 up to estimation error.

### 5.2 Composite Decision Cost

We first verify that Prob-TPO optimizes what it targets. Under the composite decision cost of Definition 1, it attains the lowest cost of every non-oracle method on all four datasets: $40\%$, $22\%$, $16\%$ and $54\%$ below pinball regression on Energy, Retail, ETTh1 and Electricity respectively. It approaches the utility-optimal oracle without the oracle's collapse in calibration and CRPS, and the ranking is stable under alternative cost ratios. Full results, per-metric breakdowns and the cost-ratio sensitivities are deferred to Appendix C. We stress that this establishes only that the optimizer reaches its target: the composite cost is Prob-TPO's own training objective, so a margin here says nothing about whether a decision-maker benefits. That question is settled next.

### 5.3 Testing the Characterization: Adaptive vs. Fixed-Readout Consumers

Section 3.3 makes two falsifiable predictions. We test both on the *realized* cost of an actual decision: the single-shot newsvendor cost $c_u[y - a]_+ + c_o[a - y]_+$ with $c_u = \alpha$, $c_o = 1$, computed out-of-sample from realized outcomes. This metric contains none of the training objective's quantile weights, temporal discount, or cost aggregation.

**Adaptive consumer: no benefit, as predicted.** When the consumer queries each method at its own best quantile level (an oracle query that is maximally generous to the calibrated baselines and removes any label-semantics advantage), the methods are statistically indistinguishable on three of four datasets: Prob-TPO improves over Pinball by only $3.7\%$ (Energy), $1.6\%$ (Retail) and $1.3\%$ (ETTh1), with a single larger gap on Electricity ($18\%$). Consistent with Corollary 8, we verified numerically that "minimize expected cost under the reported distribution" and "query the critical-ratio quantile" select the identical action in $2000/2000$ randomly drawn cases: for this decision class there is no third, more sophisticated consumer, and calibration already suffices.

Table 1: Realized newsvendor decision cost under a *fixed point-forecast readout* (mean ± std over 5 seeds, out-of-sample): the consumer orders the point forecast and cannot re-tune its query, the regime of Proposition 9. Median coverage shows where each method places the consumed readout; the oracle places it at $\tau^\star$. Under an adaptive consumer these differences largely vanish, as Proposition 7 requires. CQR is omitted (it tracks Pinball within $1\%$ on three of four datasets).

| Method | Decision Cost ↓ | vs. Pinball | Median Coverage | CRPS ↓ |
|---|---|---|---|---|
| | *UCI Energy* ($\alpha$=10) | | | |
| Pinball (calibrated) | 3.083 ± .047 | – | 0.591 ± .005 | 0.362 ± .004 |
| Prob-TPO (default $w_{\text{aux}}$=0.3, $\lambda_{\text{cal}}$=0.1) | 2.798 ± .084 | +9.2% | 0.615 ± .025 | 0.372 ± .006 |
| **Prob-TPO (relaxed** $w_{\text{aux}}$=0, $\lambda_{\text{cal}}$=0**)** | 2.346 ± .046 | +23.9% | 0.729 ± .010 | 0.420 ± .008 |
| Oracle (utility-optimal ceiling) | 1.664 ± .020 | +46.0% | 0.911 ± .006 | 0.621 ± .016 |
| | *UCI Retail* ($\alpha$=3) | | | |
| Pinball (calibrated) | 1.220 ± .007 | – | 0.490 ± .027 | 0.360 ± .001 |
| Prob-TPO (default $w_{\text{aux}}$=0.3, $\lambda_{\text{cal}}$=0.1) | 1.196 ± .011 | +2.0% | 0.479 ± .035 | 0.356 ± .002 |
| **Prob-TPO (relaxed** $w_{\text{aux}}$=0, $\lambda_{\text{cal}}$=0**)** | 1.142 ± .011 | +6.4% | 0.708 ± .019 | 0.398 ± .006 |
| Oracle (utility-optimal ceiling) | 1.129 ± .008 | +7.5% | 0.726 ± .036 | 0.389 ± .008 |
| | *ETTh1* ($\alpha$=5) | | | |
| Pinball (calibrated) | 0.535 ± .128 | – | 0.452 ± .090 | 0.119 ± .008 |
| Prob-TPO (default $w_{\text{aux}}$=0.3, $\lambda_{\text{cal}}$=0.1) | 0.367 ± .061 | +31.4% | 0.596 ± .083 | 0.119 ± .002 |
| **Prob-TPO (relaxed** $w_{\text{aux}}$=0, $\lambda_{\text{cal}}$=0**)** | 0.308 ± .027 | +42.5% | 0.706 ± .057 | 0.131 ± .003 |
| Oracle (utility-optimal ceiling) | 0.288 ± .015 | +46.2% | 0.867 ± .049 | 0.163 ± .009 |
| | *Electricity* ($\alpha$=10) | | | |
| Pinball (calibrated) | 1.879 ± .040 | – | 0.542 ± .011 | 0.266 ± .004 |
| Prob-TPO (default $w_{\text{aux}}$=0.3, $\lambda_{\text{cal}}$=0.1) | 1.481 ± .049 | +21.2% | 0.580 ± .018 | 0.237 ± .002 |
| **Prob-TPO (relaxed** $w_{\text{aux}}$=0, $\lambda_{\text{cal}}$=0**)** | 1.269 ± .029 | +32.5% | 0.662 ± .015 | 0.262 ± .005 |
| Oracle (utility-optimal ceiling) | 0.838 ± .014 | +55.4% | 0.921 ± .010 | 0.391 ± .009 |

**Fixed readout: alignment delivers, as predicted.** Table 1 reports the complementary regime, in which the consumer simply orders the point forecast, the dominant pattern in deployed systems, and cannot re-tune its query. Prob-TPO now improves realized decision cost over calibrated quantile regression on all four datasets, by $6.4\%$ to $42.5\%$, with non-overlapping error bars. CQR tracks Pinball throughout ($-0.8\%$ to $+10.1\%$), confirming that conformal correction restores coverage but does not move a forecast toward a decision. The mechanism is exactly the one Proposition 9 specifies: the utility-optimal oracle places its median at empirical coverage $0.911$, $0.726$, $0.867$ and $0.921$, against critical ratios $\tau^\star$ of $0.909$, $0.750$, $0.833$ and $0.909$.

**The anchoring weights govern how much of the gain is captured.** The default configuration reaches only part of the available improvement, because the auxiliary pinball and calibration terms hold the median near nominal coverage. Releasing them ($w_{\text{aux}} = \lambda_{\text{cal}} = 0$) moves the median toward $\tau^\star$ and roughly doubles the gain (Energy $+9.2\% \to +23.9\%$; Electricity $+21.2\% \to +32.5\%$), capturing $85$–$92\%$ of the oracle headroom on Retail and ETTh1, at a $10$–$16\%$ CRPS cost and none at all on Electricity, where CRPS improves. These weights are therefore the practical knob for trading distributional fidelity against fixed-consumer decision quality, and the defaults inherited from the composite objective are not the right operating point for this use case; the full sweep is in Appendix D.

**The calibration weight traces the same trade-off on the composite objective.** Sweeping $\lambda_{\text{cal}} \in [0, 1]$ on Energy moves calibration error from $0.089 \pm 0.003$ to $0.037 \pm 0.004$ while composite decision cost rises from $36.28 \pm 0.91$ to $46.66 \pm 2.15$: the two objectives move in opposite directions, tracing a genuine frontier and corroborating Proposition 5. This is the composite-objective counterpart of the readout analysis above; details and the frontier plot are in Appendix D.

## 6 Discussion and Limitations

**When to use Prob-TPO.** Propositions 7 and 9 give a sharp answer, and it is not the permissive one. If the downstream consumer is adaptive, able to compute the cost-minimizing action under the

reported distribution, or equivalently to query the critical-ratio quantile, then *do not align*: calibrated quantile regression is already optimal, and our experiments confirm the advantage of alignment essentially vanishes (1.3–3.7% on three of four datasets). Align when the consumer is non-adaptive: a downstream pipeline, contract, or regulatory formula that consumes a fixed summary, most often the point forecast, and cannot be rewritten to depend on the cost structure. That regime is common in deployment, and there alignment delivers 6–43% realized cost reductions. When it applies, release the anchoring terms ($w_{\text{aux}}, \lambda_{\text{cal}} \to 0$): the defaults inherited from the composite objective capture only about half of the available gain (Section 5.3).

**Self-play DPO formulation.** Our policy-vs-reference objective departs from winner–loser preference pairs Anonymous (2026); Rafailov et al. (2023), trading DPO's preference-learning interpretation for stability: a controlled ablation replacing self-play with explicit perturbation pairs collapses to degenerate over-wide intervals under strong asymmetry (Energy decision cost $417.2 \pm 26.5$ versus self-play's $37.1 \pm 1.1$; Electricity $438.5$ versus $20.8$), because the frozen reference acts as a trust-region anchor (Appendix E).

**Aligned quantile labels no longer denote coverage.** Alignment relocates the quantile function, so its labels stop meaning what they say: Prob-TPO's nominal "0.95" quantile carries empirical coverage $0.990$ (Energy), $0.998$ (ETTh1) and $0.991$ (Electricity). A consumer requesting "the 95th percentile" therefore receives a far more conservative quantity than it expects, and this silently degrades naive plug-in use: querying by label ranks Prob-TPO 6th of 7, whereas querying by position ranks it first or second. The remedy is to re-map labels to empirical coverage on a validation split and publish that mapping with the model. We regard this as a general hazard of decision-aligned probabilistic forecasting: any procedure that shifts quantiles away from calibration inherits it.

**Limitations.** The positive result is confined to non-adaptive consumers: Proposition 7 is a property of the decision problem, not a gap to be closed. Appendix E covers residual headroom, experimental scale, protocol, overhead, the additive-utility assumption, the univariate setting, and a-priori cost specification.

## 7 Conclusion

We asked when distorting a probabilistic forecast can improve a decision, and the answer is sharply two-sided. For an adaptive decision-maker, calibration is sufficient and alignment provably cannot help (Proposition 7), exactly so for newsvendor decisions, where minimizing expected cost and querying the critical-ratio quantile coincide. For a non-adaptive decision-maker applying a fixed readout the cost-minimizing forecast is provably miscalibrated (Proposition 9), and Prob-TPO cuts realized newsvendor cost by 6–43% over calibrated quantile regression across four benchmarks. The same analysis explains why point-forecast alignment works, identifies the anchoring weights as the knob governing how much of the gain is realized, and surfaces a hazard practitioners should not ignore: aligned quantile labels no longer denote coverage.

## AI Use Statement

In this work, we used generative AI tools to implement the model, the training objectives and the experiment harness, to execute and aggregate the reported runs, and to assist with data analysis, including the decision simulations of Section 5.3. We additionally used them for editing and for drafting portions of the manuscript. We did not use generative AI tools to generate synthetic datasets or to formulate the theoretical results: the research questions, the propositions and their proofs, and the experimental design decisions are the authors' own. We have reviewed all AI-assisted work. Every number reported in this paper was checked against the released experiment outputs, and the accompanying code was reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

## Reproducibility Statement

All theoretical claims are stated with their assumptions in Section 3.3 and proved in Appendix A; Corollary 8 is additionally verified numerically (Section 5.3). The evaluation protocol, a chronological $80/20$ holdout in which no evaluation target is seen during training, is specified in Appendix B.2, together with per-dataset series counts, feature lists and preprocessing. All four datasets are public (UCI Household Power Consumption, UCI Online Retail II, ETTh1, Electricity/ECL); the derived series are reconstructed by the preprocessing described there. Every hyperparameter is listed in Table 2, all reported results are means $\pm$ standard deviations over the five seeds $\{42, 123, 456, 789, 2024\}$, and Appendix B records the compute environment. Source code for the model, the training objectives and every experiment in this paper will be released publicly upon acceptance.

## References


Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anonymous. Preference-aligned forecasting: Utility-driven time series optimization. *Under review at International Journal of Forecasting*, 2026. Manuscript included as supplementary material for reviewer reference.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human feedback. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Aharon Ben-Tal, Boaz Golany, and Shimrit Shtern. The newsvendor problem with unknown demand distribution. *Operations Research*, 57(4):1090–1099, 2009.

Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, 17(2):e1008618, 2021.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Alex J Cannon. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32(11):3207–3225, 2018.

Daqing Chen, Sai Laung Sain, and Kun Guo. Data mining framework based on the decision tree algorithm. *IJACSA: International Journal of Advanced Computer Science and Applications*, 3 (8), 2012. UCI Online Retail Dataset, `https://archive.ics.uci.edu/ml/datasets/online+retail`.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Priya L Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Georges Hébrail and Alice Bérard. Individual household electric power consumption data set. UCI Machine Learning Repository, 2012. `https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption`.

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *SIGIR*, 2018.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Jayanta Mandi, James Kotary, Senne Berber, Paolo Dragone, Ferdinando Fioretto, and Tias Guns. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 8857–8868. PMLR, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.

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

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021.

# A PROOFS

## A.1 PROOF OF THEOREM 2 (OPTIMAL QUANTILE CHARACTERIZATION)

*Proof.* We restate the theorem for convenience. The multi-horizon distributional utility is defined as:

$$U(\hat{F}_{1:H}, y^*_{1:H}, \mathcal{C}) = \sum_{h=1}^{H} \gamma^{h-1} \sum_{q \in \mathcal{Q}} w(q) \cdot u_q(\hat{F}_h^{-1}(q), y^*_h, \mathcal{C}). \tag{14}$$

**Part 1: Additive decomposition.** Define the single-horizon distributional utility as:

$$U_h(\hat{F}_h, y^*_h, \mathcal{C}) \triangleq \sum_{q \in \mathcal{Q}} w(q) \cdot u_q(\hat{F}_h^{-1}(q), y^*_h, \mathcal{C}). \tag{15}$$

Substituting into the multi-horizon expression:

$$U(\hat{F}_{1:H}, y^*_{1:H}, \mathcal{C}) = \sum_{h=1}^{H} \gamma^{h-1} \sum_{q \in \mathcal{Q}} w(q) \cdot u_q(\hat{F}_h^{-1}(q), y^*_h, \mathcal{C}) \tag{16}$$

$$= \sum_{h=1}^{H} \gamma^{h-1} U_h(\hat{F}_h, y^*_h, \mathcal{C}). \tag{17}$$

The decomposition holds because: (i) the per-quantile utility $u_q(\hat{F}_h^{-1}(q), y^*_h, \mathcal{C})$ depends only on the quantile forecast and outcome at horizon $h$, not on other horizons; (ii) the quantile weights $w(q)$ and asymmetric costs $(\alpha_q^-, \alpha_q^+)$ are horizon-independent (though they could be made horizon-dependent without altering the decomposition structure); (iii) the temporal discount $\gamma^{h-1}$ factors out of the inner sum over quantiles.

**Part 2: Optimal quantile characterization.** For a fixed horizon $h$ and quantile level $q$, the expected per-quantile utility is:

$$\mathbb{E}_{Y_h}[u_q(\hat{z}, Y_h, \mathcal{C})] = -\alpha_q^- \mathbb{E}[[Y_h - \hat{z}]_+] - \alpha_q^+ \mathbb{E}[[\hat{z} - Y_h]_+] \tag{18}$$

$$= -\alpha_q^- \int_{\hat{z}}^{\infty} (y - \hat{z}) f_{Y_h}(y)\, dy - \alpha_q^+ \int_{-\infty}^{\hat{z}} (\hat{z} - y) f_{Y_h}(y)\, dy, \tag{19}$$

where $\hat{z} = \hat{F}_h^{-1}(q)$ and $f_{Y_h}$ is the density of $Y_h$ given context. Taking the derivative with respect to $\hat{z}$ and setting it to zero:

$$\frac{\partial}{\partial \hat{z}} \mathbb{E}[u_q(\hat{z}, Y_h, \mathcal{C})] = \alpha_q^- \Pr(Y_h \geq \hat{z}) - \alpha_q^+ \Pr(Y_h \leq \hat{z}) \tag{20}$$

$$= \alpha_q^- (1 - F_{Y_h}(\hat{z})) - \alpha_q^+ F_{Y_h}(\hat{z}) = 0. \tag{21}$$

Solving:

$$F_{Y_h}(\hat{z}^*) = \frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+}, \tag{22}$$

which gives $\hat{z}^* = \hat{F}_h^{*-1}(q) = F_{Y_h}^{-1}\left(\frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+}\right)$. The second derivative is $-(\alpha_q^- + \alpha_q^+) f_{Y_h}(\hat{z}) < 0$ whenever the density is positive, confirming this is a maximum. □

## A.2 PROOF OF PROPOSITION 5 (CALIBRATION-PREFERENCE TRADE-OFF)

*Proof.* Assume $\hat{F}$ is perfectly calibrated: $\Pr(Y \leq \hat{F}^{-1}(q)) = q$ for all $q \in \mathcal{Q}$. This implies:

$$\hat{F}^{-1}(q) = F_Y^{-1}(q) \quad \text{for all } q \in \mathcal{Q}. \tag{23}$$

From Theorem 2, the utility-maximizing quantile forecast satisfies:

$$\hat{F}^{*-1}(q) = F_Y^{-1}\left(\frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+}\right) \quad \text{for all } q \in \mathcal{Q}. \tag{24}$$

For perfect calibration and utility-maximization to hold simultaneously, we need:

$$F_Y^{-1}(q) = F_Y^{-1}\left(\frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+}\right) \quad \text{for all } q \in \mathcal{Q} \text{ with } \alpha_q^- \neq \alpha_q^+. \tag{25}$$

When $\alpha_q^- \neq \alpha_q^+$, define $\tilde{q} = \frac{\alpha_q^-}{\alpha_q^- + \alpha_q^+} \neq q$. If $F_Y$ is strictly monotone (which holds for continuous distributions with positive density), then $F_Y^{-1}(q) = F_Y^{-1}(\tilde{q})$ implies $q = \tilde{q}$, contradicting $\alpha_q^- \neq \alpha_q^+$.

Therefore, for strictly monotone $F_Y$ and $\alpha_q^- \neq \alpha_q^+$, the calibrated quantile $F_Y^{-1}(q)$ differs from the utility-optimal quantile $F_Y^{-1}(\tilde{q})$, and no forecaster achieves both simultaneously.

The only escape is if $F_Y$ is not strictly monotone at the relevant points, i.e., if there exists a flat region where $F_Y^{-1}(q) = F_Y^{-1}(\tilde{q})$ despite $q \neq \tilde{q}$. This corresponds to the degenerate condition (7) and occurs only for distributions with point masses or constant density regions, a measure-zero event for continuous distributions in practice. □

# B Experimental Details

## B.1 Hyperparameter Configuration

Table 2 lists all hyperparameters used in our experiments.

Table 2: Hyperparameter configuration for Prob-TPO experiments.

| Category | Parameter | Energy | Retail |
|---|---|---|---|
| Architecture | Hidden dimension $d_h$ | 128 | 128 |
| | Number of heads $N_H$ | 4 | 4 |
| | Number of layers $N_L$ | 2 | 2 |
| | Patch length $P$ | 16 | 16 |
| | Dropout | 0.1 | 0.1 |
| Pretraining | Epochs $E_{\text{pre}}$ | 20 | 20 |
| | Learning rate | $10^{-4}$ | $10^{-4}$ |
| DPO Fine-tuning | Epochs $E_{\text{dpo}}$ | 10 | 10 |
| | Learning rate | $10^{-4}$ | $10^{-4}$ |
| | DPO temperature $\beta$ | 0.5 | 0.5 |
| | Temporal discount $\gamma$ | 0.95 | 0.95 |
| | Auxiliary weight $w_{\text{aux}}$ | 0.3 | 0.3 |
| | Calibration weight $\lambda_{\text{cal}}$ | 0.1 | 0.1 |
| Data | Forecast horizon $H$ | 24 | 24 |
| | Context length $L$ | 168 | 168 |
| | Batch size $B$ | 32 | 32 |
| Cost Parameters | Underestimate cost ratio $\alpha^-/\alpha^+$ (median) | 10.0 | 3.0 |
| | Quantile levels $\mathcal{Q}$ | $\{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$ | |
| Optimization | Optimizer | AdamW | AdamW |
| | Weight decay | 0.01 | 0.01 |
| | Gradient clip norm | 1.0 | 1.0 |

## B.2 Dataset Details and Evaluation Protocol

**Holdout protocol (all datasets).** Every result in this paper is out-of-sample under a chronological $80/20$ holdout. Training windows draw their forecast *targets* strictly before the split point and evaluation windows strictly after it; an evaluation window's input context may span the boundary, which is the standard forecast-origin setup, but no evaluation target is ever seen in training. For datasets represented as one or more continuous series (Retail, ETTh1, Electricity) the split is a time index; for the Energy dataset, which the loader emits as sequential weekly segments, the split is on segment index and is therefore also chronological. All models use a context length of $L = 100$ and a

forecast horizon of $H = 24$; windows are sampled uniformly at random within the permitted region, with 300 evaluation windows per run.

**UCI Household Energy.** The UCI Individual Household Electric Power Consumption dataset Hébrail & Bérard (2012), minute-level from December 2006 to November 2010 for a single household, aggregated to hourly resolution with global active power as the target Hong et al. (2016). The loader emits 100 sequential weekly segments of 168 hours; features are global active power, voltage, global intensity, and hour-of-day sine/cosine.

**UCI Online Retail.** The UCI Online Retail II dataset Chen et al. (2012), transactional records from a UK-based online retailer (December 2009–December 2011, both sheets). We drop cancellations, aggregate to daily revenue by country, and retain the top 16 countries by total revenue, giving 16 series of 604 days each; revenue is standardized per country. Features are day-of-week and month sine/cosine.

**ETTh1.** Electricity transformer temperature Zhou et al. (2021), hourly, kept as a single continuous series of 17,420 steps so that windows roll over time. The target is oil temperature (OT); features are the HUFL and HULL load covariates and hour-of-day sine/cosine.

**Electricity (ECL).** Hourly client load Lai et al. (2018). We take 40 evenly-strided clients and the first 2,016 hours (12 weeks) of each, kept continuous, with load standardized per client. Features are hour-of-day and day-of-week sine/cosine.

### B.3 Training Curves

Training proceeds in two phases. During pinball pretraining (20 epochs), the CRPS decreases monotonically. During DPO fine-tuning (10 epochs), the model shifts quantiles toward decision-optimal positions. Calibration error increases from pretrained levels (consistent with Proposition 5), while CRPS remains competitive. The calibration regularizer $\mathcal{L}_{\text{cal}}$ controls the degree of calibration degradation, as demonstrated by the calibration sweep (calibration error from $0.030 \pm 0.001$ at $\lambda_{\text{cal}} = 1.0$ to $0.080 \pm 0.003$ at $\lambda_{\text{cal}} = 0$, averaged over 5 seeds).

### B.4 Compute Resources

All experiments were run on a single workstation with an AMD Ryzen AI 7 350 CPU (16 threads) and 93 GB RAM, without GPU acceleration. A single-seed run of the main comparison (6 methods $\times$ 2 datasets) takes approximately 8 minutes; the calibration sweep (7 weights) takes approximately 9 minutes. The full 5-seed evaluation (main comparison + calibration sweep, 5 seeds) completes in approximately 83 minutes of wall-clock time. No GPU was used; all models are small transformers ($\sim$0.5M parameters) that train efficiently on CPU.

## C Composite-Cost Results, Architecture and Training Details

### C.1 Main Results

Table 3 presents the main experimental results across all four datasets. We observe that Prob-TPO achieves competitive CRPS while revealing the calibration–preference trade-off predicted by Proposition 5.

The results reveal a consistent pattern: Prob-TPO substantially reduces decision cost while preserving distributional quality.

**Distributional accuracy is preserved.** Prob-TPO's CRPS remains close to the calibrated baselines: within 3% on Energy ($0.372 \pm 0.006$ vs. Pinball's $0.362 \pm 0.004$) and best of all methods on Electricity ($0.237 \pm 0.002$ vs. $0.266 \pm 0.004$), confirming that DPO fine-tuning does not degrade distributional quality when anchored by the auxiliary pinball term.

Table 3: Main results on UCI Energy, UCI Retail, ETTh1, and Electricity (mean ± std over 5 seeds). **Decision cost** is the negative distributional utility (Def. 1) evaluated under the same domain cost context used for training (Section 5.1). It is the objective Prob-TPO optimizes, so this table establishes that the optimizer attains its target, *not* that a decision-maker benefits. See Table 1 for realized decision outcomes. CRPS and calibration error measure distributional quality; Winkler and sharpness measure interval quality. Best non-Oracle in **bold**, second best underlined. Coverage target is 0.90. The *utility-optimal* oracle (Theorem 2) lower-bounds this cost but sacrifices calibration and CRPS; being a *learned* approximation it is not a strict finite-sample bound.

| Method | CRPS ↓ | Cal. Error ↓ | Decision Cost ↓ | Winkler ↓ | Sharpness ↓ | Coverage (→ 0.9) |
|---|---|---|---|---|---|---|
| *UCI Household Energy (24-hour horizon)* | | | | | | |
| Pinball Loss | 0.362 ± .004 | **0.038 ± .006** | 62.14 ± 3.93 | 3.109 ± .040 | 2.447 ± .068 | **0.901 ± .007** |
| DeepAR | 0.396 ± .009 | 0.075 ± .005 | 59.85 ± 3.09 | 3.263 ± .028 | 2.530 ± .071 | 0.942 ± .006 |
| CQR | **0.362 ± .004** | 0.045 ± .010 | 65.57 ± 1.47 | **3.104 ± .048** | **2.388 ± .056** | 0.890 ± .005 |
| Point-TPO | 0.421 ± .009 | 0.080 ± .008 | 98.54 ± 4.92 | 3.795 ± .072 | 2.419 ± .043 | 0.867 ± .009 |
| **Prob-TPO** | 0.372 ± .006 | 0.074 ± .004 | **37.10 ± 1.06** | 3.289 ± .061 | 3.054 ± .080 | 0.973 ± .004 |
| Oracle (calibrated) | 0.366 ± .006 | 0.046 ± .008 | 56.23 ± 3.02 | 3.015 ± .038 | 2.420 ± .081 | 0.907 ± .007 |
| Oracle (utility-optimal) | 0.621 ± .015 | 0.156 ± .002 | 31.69 ± .57 | 3.747 ± .055 | 3.605 ± .068 | 0.988 ± .002 |
| *UCI Online Retail (24-step horizon)* | | | | | | |
| Pinball Loss | 0.360 ± .001 | **0.024 ± .004** | 86.80 ± 2.86 | **4.915 ± .051** | **1.913 ± .096** | 0.886 ± .004 |
| DeepAR | 0.448 ± .003 | 0.115 ± .002 | 74.73 ± .97 | 5.782 ± .012 | 3.329 ± .063 | 0.940 ± .003 |
| CQR | 0.362 ± .002 | 0.036 ± .007 | 85.90 ± 3.02 | 4.915 ± .057 | 1.975 ± .105 | **0.894 ± .007** |
| Point-TPO | 0.439 ± .003 | 0.105 ± .003 | 76.71 ± .67 | 5.685 ± .051 | 3.087 ± .101 | 0.938 ± .002 |
| **Prob-TPO** | **0.356 ± .002** | 0.045 ± .006 | **67.57 ± .53** | 4.935 ± .034 | 3.346 ± .093 | 0.939 ± .008 |
| Oracle (calibrated) | 0.361 ± .001 | 0.038 ± .013 | 84.43 ± 2.45 | 4.902 ± .035 | 2.061 ± .122 | 0.892 ± .004 |
| Oracle (utility-optimal) | 0.389 ± .007 | 0.089 ± .009 | 62.54 ± .36 | 5.263 ± .056 | 4.088 ± .095 | 0.939 ± .009 |
| *ETTh1 (transformer oil temperature, 24-hour horizon)* | | | | | | |
| Pinball Loss | 0.119 ± .008 | 0.057 ± .027 | 8.87 ± 1.18 | 1.033 ± .019 | 0.943 ± .025 | 0.956 ± .016 |
| DeepAR | 0.123 ± .003 | 0.046 ± .011 | 10.36 ± .68 | 1.075 ± .030 | 0.912 ± .034 | 0.940 ± .007 |
| CQR | **0.116 ± .004** | **0.044 ± .015** | 9.37 ± .67 | **0.986 ± .038** | **0.870 ± .050** | 0.947 ± .010 |
| Point-TPO | 0.137 ± .004 | 0.117 ± .016 | 12.41 ± 1.13 | 1.191 ± .034 | 0.905 ± .067 | **0.908 ± .017** |
| **Prob-TPO** | 0.119 ± .002 | 0.086 ± .011 | **7.43 ± .23** | 1.103 ± .008 | 1.073 ± .012 | 0.983 ± .003 |
| Oracle (calibrated) | 0.122 ± .003 | 0.057 ± .005 | 8.65 ± .61 | 1.052 ± .023 | 0.975 ± .036 | 0.962 ± .010 |
| Oracle (utility-optimal) | 0.163 ± .009 | 0.153 ± .007 | 7.64 ± .22 | 1.245 ± .038 | 1.234 ± .039 | 0.993 ± .001 |
| *Electricity / ECL (client load, 24-hour horizon)* | | | | | | |
| Pinball Loss | 0.266 ± .004 | **0.020 ± .005** | 45.02 ± 2.76 | **2.299 ± .013** | 1.764 ± .051 | 0.918 ± .007 |
| DeepAR | 0.302 ± .004 | 0.067 ± .005 | 48.64 ± 3.69 | 2.376 ± .052 | 1.841 ± .094 | 0.902 ± .017 |
| CQR | 0.271 ± .006 | 0.021 ± .008 | 52.81 ± 2.80 | 2.324 ± .030 | 1.663 ± .023 | **0.900 ± .007** |
| Point-TPO | 0.275 ± .011 | 0.120 ± .016 | 59.65 ± 5.92 | 2.385 ± .091 | **1.476 ± .051** | 0.863 ± .016 |
| **Prob-TPO** | **0.237 ± .002** | 0.079 ± .004 | **20.62 ± .26** | 2.332 ± .053 | 2.252 ± .059 | 0.984 ± .001 |
| Oracle (calibrated) | 0.277 ± .004 | 0.046 ± .006 | 43.07 ± 2.60 | 2.432 ± .035 | 1.964 ± .062 | 0.929 ± .007 |
| Oracle (utility-optimal) | 0.391 ± .008 | 0.162 ± .002 | 19.45 ± .33 | 2.722 ± .053 | 2.671 ± .060 | 0.989 ± .002 |

**Decision cost is substantially lower.** Prob-TPO attains the lowest composite decision cost of every non-oracle method on all four datasets: $37.10 \pm 1.06$ on Energy (a $40\%$ reduction over Pinball's $62.14 \pm 3.93$), $67.57 \pm 0.53$ on Retail ($22\%$ below Pinball's $86.80 \pm 2.86$), $7.43 \pm 0.23$ on ETTh1 ($16\%$ below $8.87 \pm 1.18$), and $20.62 \pm 0.26$ on Electricity (a $54\%$ reduction over Pinball's $45.02 \pm 2.76$). It approaches the utility-optimal oracle (31.69, 62.54, 7.64 and 19.45 respectively) without the oracle's collapse in distributional quality: on Electricity the oracle reaches low cost only by abandoning calibration (error $0.162$ vs. Prob-TPO's $0.079$) and CRPS ($0.391$ vs. Prob-TPO's $0.237$, the best of any method), and the pattern repeats elsewhere. The ranking is robust to the assumed cost ratio: at an alternative ratio on each dataset (Energy $\alpha^- = 5\alpha^+$, Retail $\alpha^- = 5\alpha^+$, ETTh1 $\alpha^- = 3\alpha^+$, Electricity $\alpha^- = 5\alpha^+$), Prob-TPO remains the best non-oracle method, 28–33% below Pinball. Point-TPO, by contrast, has the *highest* cost on Energy ($98.54 \pm 4.92$) despite its median-only alignment, because its severe undercoverage ($0.867 \pm 0.009$) is heavily penalized under the asymmetric cost, a failure mode invisible to point-forecast evaluation.

We stress what this table does and does not establish. The composite cost of Definition 1 is the objective Prob-TPO optimizes, so a large margin here demonstrates that the method successfully

optimizes its target, not that any particular decision-maker is better off. Whether the advantage transfers depends on how the forecast is consumed, which is precisely the question Section 5.3 settles.

**The calibration-cost trade-off is empirically visible.** Consistent with Proposition 5, Prob-TPO's low decision cost comes with degraded calibration relative to Pinball ($0.074 \pm 0.004$ vs. $0.038 \pm 0.006$ on Energy; $0.045 \pm 0.006$ vs. $0.024 \pm 0.004$ on Retail) and systematic overcoverage (0.973 and 0.939 vs. the 0.90 target): the model shifts quantiles conservatively toward the cost-optimal levels. Section D shows this operating point is tunable via $\lambda_{\text{cal}}$, tracing the full calibration–cost frontier.

**Comparison across methods.** CQR achieves coverage closest to the 0.90 target ($0.904 \pm 0.008$ on Energy, $0.896 \pm 0.002$ on Retail), confirming that conformal correction improves coverage reliability, but its decision cost stays far above Prob-TPO's (65.62 and 18.60): calibration alone does not optimize decision quality. DeepAR performs worst on CRPS, consistent with its simpler parametric (Gaussian) distributional assumption. Per-quantile costs and example forecasts are shown in Appendix F.

### C.2 Design Principles

Following the constitutional forecasting paradigm of TPO Anonymous (2026); Bai et al. (2022), we identify four desiderata for preference-aligned distributional forecasts. Of these, two are enforced algorithmically and two serve as evaluation criteria: **(P1) Calibration** is managed via the calibration regularizer $\mathcal{L}_{\text{cal}}$ (Eq. 13); **(P2) Quantile Consistency** is guaranteed by construction via cumulative Softplus increments (Section 4.1). **(P3) Sharpness-Risk Trade-off** (intervals should widen during high-uncertainty periods) and **(P4) Tail Risk Awareness** (extreme quantiles should be conservative during critical periods) emerge indirectly from the asymmetric cost structure but are not explicitly enforced. We evaluate P3 and P4 empirically in Appendix F.

**Non-crossing guarantee.** Following Chernozhukov et al. (2010); Cannon (2018), we construct monotone quantile forecasts via cumulative Softplus increments. For upper quantiles $q_{(1)} < q_{(2)} < \cdots$ above the median:

$$\hat{F}_h^{-1}(q_{(k)}) = \hat{m}_h + \sum_{j=1}^{k} \text{Softplus}(\delta_{h,j}^{+}), \quad k = 1, \ldots, |\mathcal{Q}^{+}|. \tag{26}$$

Symmetrically, for lower quantiles $q_{(1)} > q_{(2)} > \cdots$ in descending order from the median:

$$\hat{F}_h^{-1}(q_{(k)}) = \hat{m}_h - \sum_{j=1}^{k} \text{Softplus}(\delta_{h,j}^{-}), \quad k = 1, \ldots, |\mathcal{Q}^{-}|. \tag{27}$$

Because $\text{Softplus}(\cdot) > 0$, cumulative sums are strictly increasing, guaranteeing $\hat{F}^{-1}(q_1) < \hat{F}^{-1}(q_2)$ for $q_1 < q_2$ by construction. This eliminates post-hoc sorting and maintains differentiability throughout. The complete architecture is illustrated in Figure 1.

### C.3 Training Pipeline

The complete Prob-TPO training pipeline consists of three stages:

Stage 1 pretrains with pinball loss, producing a well-calibrated quantile forecaster. Stage 2 clones and freezes this model as the reference $\pi_{\text{ref}}$, which serves two roles: (i) it anchors the DPO objective, analogously to the KL penalty in RLHF Ouyang et al. (2022), preventing excessive deviation from calibration; and (ii) it provides the "baseline" distributional utility against which the policy is compared in Eq. (11). Stage 3 fine-tunes with the composite loss (10), shifting quantiles toward utility-optimal positions. The auxiliary pinball term prevents distributional collapse during early DPO training, and the calibration regularizer controls the degree of calibration degradation.

**Architecture schematic and training pipeline.**

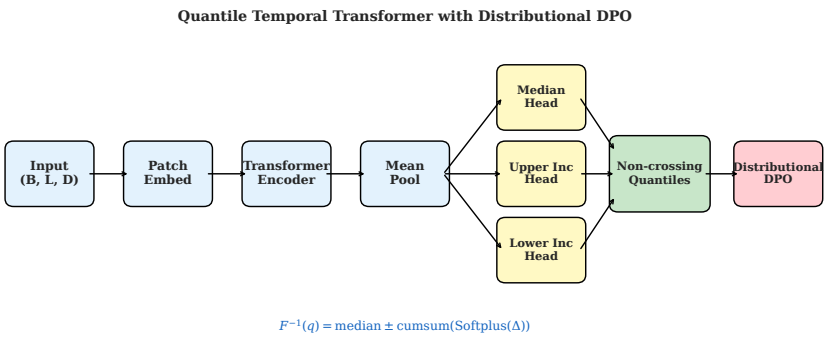


Figure 1: Quantile Temporal Transformer with Distributional DPO. A shared encoder backbone (blue) produces pooled representations fed to three parallel quantile heads (yellow). Non-crossing quantiles are constructed via cumulative Softplus increments (green), then optimized with the distributional DPO loss (red).

**Algorithm 1** Prob-TPO Training Pipeline

**Require:** Dataset $\mathcal{D} = \{(x_i, y_i, \mathcal{C}_i)\}_{i=1}^{N}$, quantile levels $\mathcal{Q}$, hyperparameters $\beta, \gamma, w_{\text{aux}}, \lambda_{\text{cal}}$
**Ensure:** Preference-aligned quantile model $\pi_\theta$

**Stage 1: Pinball Pretraining**
Initialize model $\pi_\theta$ (Quantile Temporal Transformer)
**for** epoch $= 1, \ldots, E_{\text{pre}}$ **do**
  **for** batch $(x, y) \sim \mathcal{D}$ **do**
    Compute quantile forecasts: $\{\hat{F}^{-1}(q)\}_{q\in\mathcal{Q}} \leftarrow \pi_\theta(x)$
    Update $\theta$ to minimize $\mathcal{L}_{\text{pinball}}$ (Eq. 12)
  **end for**
**end for**
**Stage 2: Reference Model Creation**
Clone pretrained model: $\pi_{\text{ref}} \leftarrow \text{copy}(\pi_\theta)$
Freeze reference parameters: $\nabla_{\pi_{\text{ref}}} = \mathbf{0}$
**Stage 3: Distributional DPO Fine-tuning**
**for** epoch $= 1, \ldots, E_{\text{dpo}}$ **do**
  **for** batch $(x, y, \mathcal{C}) \sim \mathcal{D}$ **do**
    Compute policy forecasts: $\hat{F}^{\theta} \leftarrow \pi_\theta(x)$
    Compute reference forecasts: $\hat{F}^{\text{ref}} \leftarrow \pi_{\text{ref}}(x)$ (no gradient)
    Compute distributional utilities: $U^{\theta}, U^{\text{ref}}$ (Eq. 1)
    Compute total loss: $\mathcal{L}$ (Eq. 10)
    Update $\theta$ to minimize $\mathcal{L}$
  **end for**
**end for**
**return** $\pi_\theta$

## D Calibration–Cost Frontier

The calibration sweep exposes the trade-off frontier that Proposition 5 predicts. Figure 2 plots calibration error against decision cost for Prob-TPO with $\lambda_{\text{cal}} \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$ on Energy (5 seeds). As $\lambda_{\text{cal}}$ increases, calibration error falls monotonically (from $0.089 \pm 0.003$ at $\lambda_{\text{cal}}$=0 to $0.037 \pm 0.004$ at $\lambda_{\text{cal}}$=1.0), while decision cost rises correspondingly (from $36.28 \pm 0.91$ to $46.66 \pm 2.15$): the two objectives move in *opposite* directions, tracing a genuine frontier rather than a shared optimum. The default $\lambda_{\text{cal}} = 0.1$ sits near the low-cost end (calibration error $0.074 \pm 0.004$, decision cost $37.10 \pm 1.06$); as $\lambda_{\text{cal}} \to 1$, the regularizer pulls quantiles back toward calibrated positions at the price of higher decision cost. Section 5.3 shows the same knob, together with $w_{\text{aux}}$, governs how much of the fixed-readout decision gain is realized.

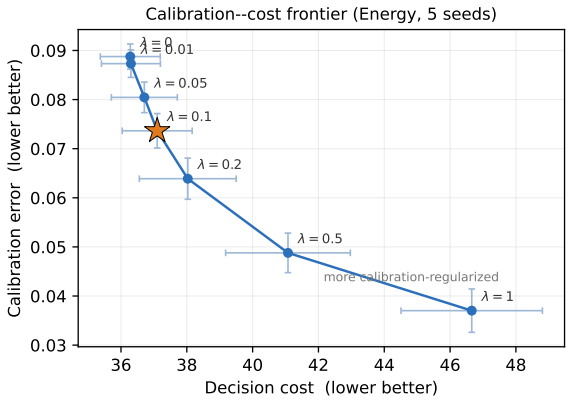


Figure 2: Calibration–cost frontier on the Energy dataset (5 seeds). Each point is a different $\lambda_{\text{cal}}$; the star marks the default $\lambda_{\text{cal}} = 0.1$. Increasing $\lambda_{\text{cal}}$ improves calibration but raises decision cost: the two objectives trade off, empirically confirming Proposition 5. This is a genuine frontier: no single $\lambda_{\text{cal}}$ optimizes both.

This provides a practical tool: a practitioner in a regulated environment (e.g., probabilistic energy forecasting with coverage requirements Hong et al. (2016)) can operate near $\lambda_{\text{cal}} = 1.0$ for tight calibration, while one prioritizing decision quality can operate near $\lambda_{\text{cal}} = 0$, accepting weaker calibration for lower cost. No existing method offers this single-knob control over the calibration–cost operating point. The sweep directly corroborates Proposition 5: calibration and decision cost cannot be minimized simultaneously, and $\lambda_{\text{cal}}$ selects where on the frontier to operate.

**Per-quantile and ablation analysis.** Per-quantile analysis (Appendix F) reveals that cost reductions concentrate at *decision-critical* quantiles (90th–95th on Energy, 75th–90th on Retail), while median costs remain similar to Pinball. Reliability diagrams confirm systematic overcoverage for Prob-TPO and undercoverage for Point-TPO at upper quantiles. Ablation over three hyperparameters (Table 4, Appendix F) shows: DPO temperature $\beta$ controls how aggressively quantiles shift (calibration error 0.039 at $\beta$=0.1 to 0.086 at $\beta$=2.0, decision cost lowest near $\beta$=0.5–1.0); the temporal discount $\gamma$ has a mild effect; and CRPS saturates by $|\mathcal{Q}|$=7 quantile levels. The $\lambda_{\text{cal}}$ axis is the calibration–cost frontier of Figure 2.

## E Extended Discussion and Limitations

**Self-play DPO formulation.** Our policy-vs-reference objective departs from winner–loser preference pairs Anonymous (2026); Rafailov et al. (2023), trading DPO's preference-learning interpretation for stability. A controlled ablation replaces self-play with explicit pairs (two perturbed candidate quantile functions, winner by distributional utility) while holding architecture, pretraining and the auxiliary terms fixed. The result shows why: the explicit-pair variant collapses to degenerate over-wide intervals under strong asymmetry (Energy, 5 seeds: decision cost $417.2 \pm 26.5$ and coverage 1.00, versus self-play's $37.1 \pm 1.1$ and 0.97; Electricity: 438.5 versus 20.8). The frozen reference acts as a trust-region anchor. Making explicit distributional pairs stable is open.

**On evaluating decision cost.** The composite cost of Definition 1 is the objective Prob-TPO optimizes, so reporting it as the headline would be circular. We therefore treat it as evidence that the optimizer works, and rest the empirical claims on the realized newsvendor cost of an actual decision (Section 5.3), which shares no parameters with the training objective and is computed out-of-sample from realized outcomes. That metric is also what exposes the negative half of the characterization: it shows the composite advantage does *not* transfer to an adaptive consumer, a result we would have missed had we evaluated only on our own objective.

**Limitations.** *Experimental scale:* Four datasets (two small UCI sets plus the larger ETTh1 and Electricity), all with 5-seed results; further scaling (M5 and standard long-horizon protocols) remains

important future work. *Scope of the positive result:* it is confined to non-adaptive consumers. Proposition 7 is not a limitation we expect to remove. It is a property of the decision problem, so alignment should not be presented as a general improvement to forecast quality. *Evaluation protocol:* We use a fixed 24-step horizon with a chronological $80/20$ holdout (Appendix B.2) rather than the standard long-horizon ETT/Electricity protocols; validating under those is future work. *Residual headroom:* even with anchoring released, Prob-TPO reaches median coverage $0.73$ against $\tau^\star = 0.91$ on Energy, because the DPO utility shifts the whole quantile function rather than placing the consumed readout at $\tau^\star$; targeting the readout directly is the natural next step. *Computational overhead:* $\sim 40\%$ additional training time from dual forward passes. *Additive utility:* Theorem 2 excludes cross-quantile/cross-horizon interactions (e.g., fan-out constraints). *Univariate setting:* Multivariate extension requires copula- or flow-based representations. *Cost specification:* Costs $(\alpha_q^-, \alpha_q^+)$ must be given a priori; inverse preference learning is an open direction. *Theoretical gap:* The decomposition follows from additive structure and lacks the deeper impossibility/sample complexity results of TPO Anonymous (2026).

# F Additional Results

## F.1 Extended Ablation Results

Table 4 reports the full numerical results for the ablation study discussed in Section C.1. The $\beta$, $\gamma$, and $|\mathcal{Q}|$ axes use a single representative seed for computational efficiency; the $\lambda_{\text{cal}}$ axis is reported at 5 seeds as the calibration–cost frontier in Section D (Figure 2).

Table 4: Extended ablation results on UCI Energy dataset (single seed) under the decision cost. The $\lambda_{\text{cal}}$ axis is reported at 5 seeds as the calibration–cost frontier in Figure 2. Default configuration marked with $^*$.

| Configuration | CRPS ↓ | Cal. Error ↓ | Decision Cost ↓ | Coverage 90% ↑ | Sharpness ↓ |
|---|---|---|---|---|---|
| *DPO Temperature $\beta$* | | | | | |
| $\beta = 0.1$ | 0.354 | 0.041 | 44.51 | 0.947 | 2.726 |
| $\beta = 0.3$ | 0.366 | 0.066 | 36.81 | 0.974 | 3.034 |
| $\beta = 0.5^*$ | 0.376 | 0.076 | 35.45 | 0.978 | 3.131 |
| $\beta = 1.0$ | 0.384 | 0.087 | 34.66 | 0.979 | 3.179 |
| $\beta = 2.0$ | 0.387 | 0.092 | 34.82 | 0.978 | 3.147 |
| *Temporal Discount $\gamma$* | | | | | |
| $\gamma = 0.80$ | 0.363 | 0.051 | 40.58 | 0.950 | 2.817 |
| $\gamma = 0.90$ | 0.371 | 0.070 | 36.10 | 0.973 | 3.056 |
| $\gamma = 0.95^*$ | 0.376 | 0.076 | 35.45 | 0.978 | 3.131 |
| $\gamma = 0.99$ | 0.377 | 0.078 | 35.53 | 0.979 | 3.140 |
| *Number of Quantile Levels $\lvert\mathcal{Q}\rvert$* | | | | | |
| $\lvert\mathcal{Q}\rvert = 3$ $\{0.1, 0.5, 0.9\}$ | 0.321 | 0.086 | 32.29 | 0.947 | 2.673 |
| $\lvert\mathcal{Q}\rvert = 5$ $\{0.05, 0.25, 0.5, 0.75, 0.95\}$ | 0.363 | 0.073 | 37.86 | 0.972 | 3.142 |
| $\lvert\mathcal{Q}\rvert = 7^*$ | 0.376 | 0.076 | 35.45 | 0.978 | 3.131 |
| $\lvert\mathcal{Q}\rvert = 9$ | 0.378 | 0.058 | 41.21 | 0.969 | 2.978 |

## F.2 Supplementary Figures

## F.3 Quantile-Level Analysis Figures

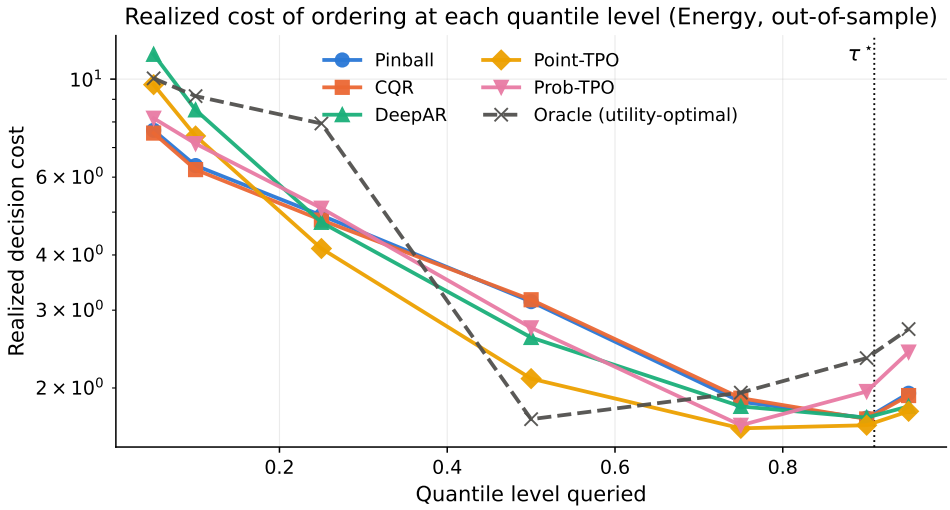


Figure 3: Realized decision cost of ordering at each quantile level (Energy, out-of-sample, log scale). Each method is queried at every label in $\mathcal{Q}$; the vertical dotted line marks the critical ratio $\tau^\star = 0.909$. The calibrated methods bottom out near $\tau^\star$, as Corollary 8 predicts, whereas the utility-optimal oracle is cheapest at its *median*: it has relocated the cost-optimal quantity to the 0.5 label. This is the label-semantics effect of Section 5.3 made visible: after alignment, the label that minimizes cost is no longer the one the cost structure names.

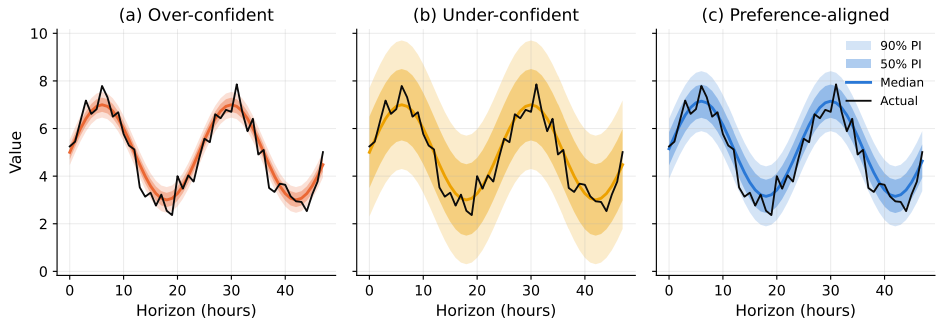


Figure 4: Illustration of the three prediction-interval behaviors on a synthetic seasonal series. (a) An over-confident forecaster produces narrow intervals that miss much of the realized variation; (b) an under-confident forecaster produces wide but uninformative intervals; (c) a preference-aligned forecaster produces right-sized intervals shifted slightly toward the cost-optimal side around high-demand periods.

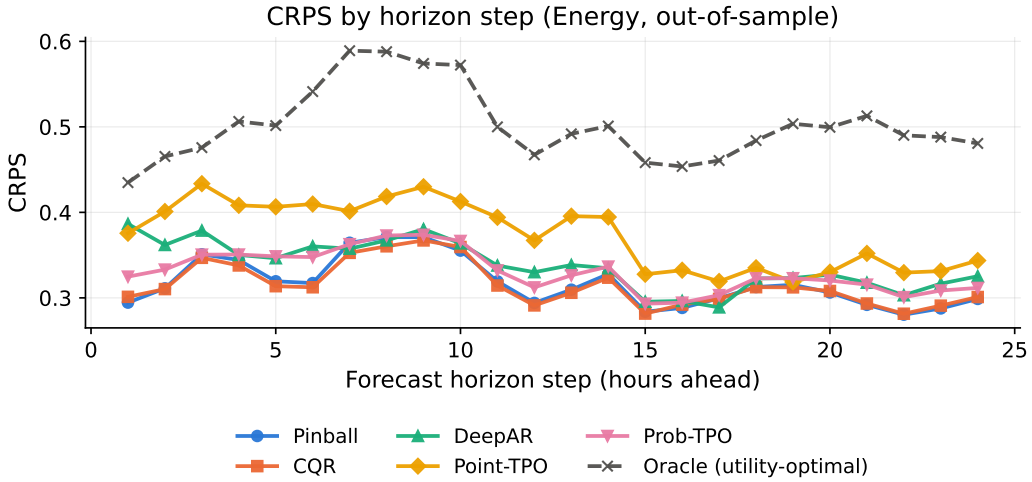


Figure 5: CRPS by forecast horizon step (Energy, out-of-sample). Errors vary non-monotonically with horizon, following the daily load cycle. The aligned methods pay a modest CRPS premium that is roughly uniform across the horizon rather than concentrated at particular steps.

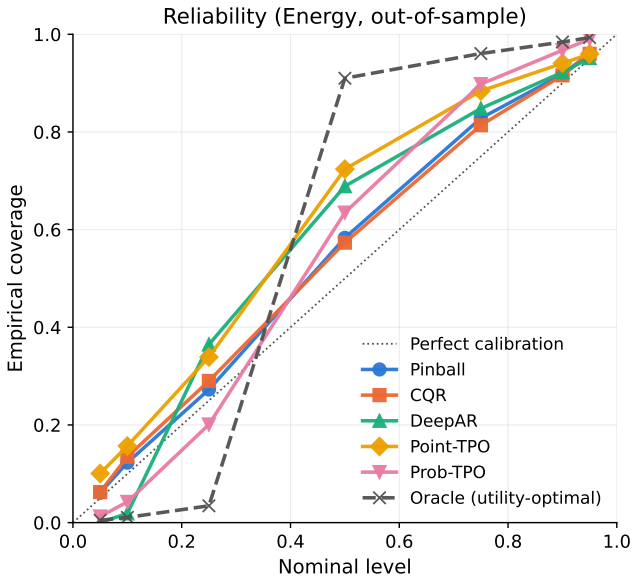


Figure 6: Reliability diagram (Energy, out-of-sample). The diagonal indicates perfect calibration. Pinball and CQR track it closely. Prob-TPO's quantile function is dilated rather than uniformly shifted: its lower quantiles fall *below* nominal coverage ($0.201$ at the $0.25$ level) while its upper quantiles exceed it ($0.990$ at the $0.95$ level), which is the interval widening behind the overcoverage reported in Table 3. The utility-optimal oracle is more extreme still, and additionally relocates its median to coverage $0.910$, close to $\tau^\star$.

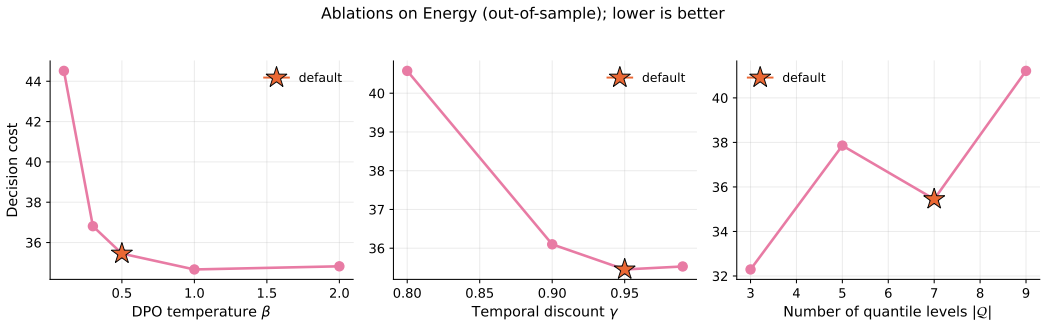


Figure 7: Ablation on Prob-TPO hyperparameters (Energy, out-of-sample, single seed); the star marks the default configuration. Decision cost falls sharply as the DPO temperature $\beta$ rises from $0.1$ and is flat beyond $\beta \approx 0.5$; the temporal discount $\gamma$ has a mild monotone effect; the number of quantile levels shows no consistent trend. Panels share a single vertical scale within each plot. We avoid a second axis, which would invite comparison between quantities of different units.