# Some models are useful, but for how long?

Kentaro Hoffman[1], Steven Salerno[2], Jeff Leek[2] ,Tyler McCormick[1],

[1]. University of Washington [2]. Fred Hutch Cancer Center

How long does a generative AI model last before it becomes more economical to refit it? Does Prediction-Powered Inference help extend a model's life?

Not only are AI models expensive to create, but there are also substantial costs to keep models up-to-date. This leaves maintainers with three options:

1. **Retain** the existing model with its decreased performance
2. **Recalibrate** the existing model
3. **Refit** a new model from scratch

**Goal:** Determine which option is most economical at various points in a model's life cycle.

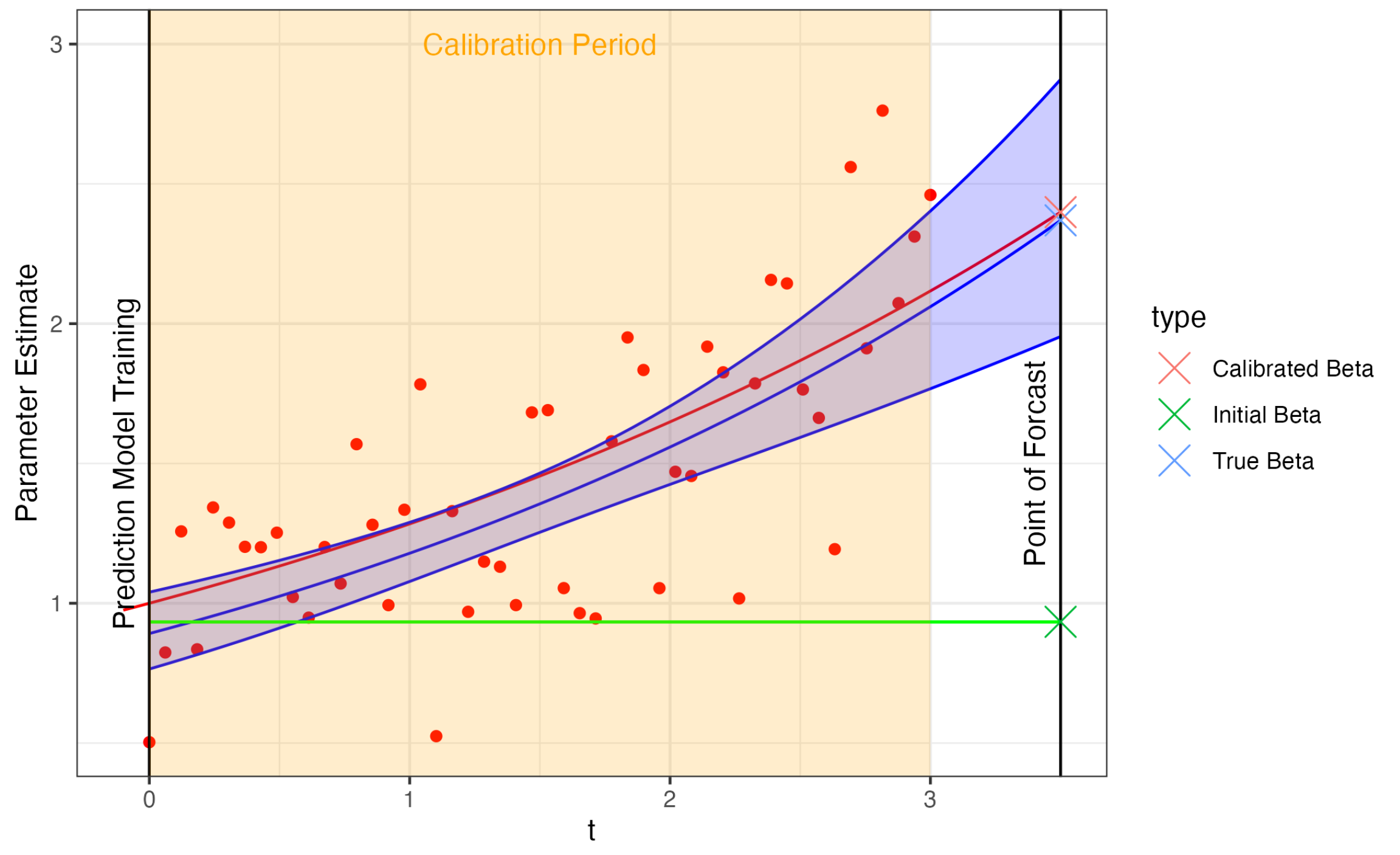
## Prediction-Powered Inference++ Recalibration

- Prediction-Powered Inference++ is a procedure for performing valid statistical inference when experimental data is supplemented with predictions from a black-box.
- It estimates parameters $\hat{\theta}$ from:

$$\hat{\theta} = argmin_{\theta}\, L_n(\theta) + \lambda(\tilde{L}_n(\theta) - L_n^f(\theta))$$

- $[L_n(\theta), \tilde{L}_n(\theta), L_n^f(\theta)]$ are the loss function using the black-box predicted values, the experimentally observed values, and the black-box predictions of the experimental data.

## Estimating MSE



While the generative model initially gives imputed outcomes that give correct downstream inference, as time goes on, the imputation and downstream inference worsens. To correct for this, we can employ additional experimental data collection that occurs during a "calibration period".

### Full AI Model refitting Algorithm

```
1. For b in 1, ..., B
   (a) Take a bootstrap $(X^b, Y^b)$ of size $(c - c_{model})/(c_{unlab})$ from the data that was held out during initial model training
   (b) Compute $\hat{Var}(\hat{\beta}^b) = (X^{bT}\hat{W}^b X^b)^{-1}$ where $\hat{\beta}^b = \text{argmin}_{\beta} g(Y_t^b, X_t^b; \beta)$
2. $MSE(\hat{\beta}_{refit}) =$Median$(\hat{Var}(\hat{\beta}^1_{refit}), ..., \hat{Var}(\hat{\beta}^b_{refit}))$
```

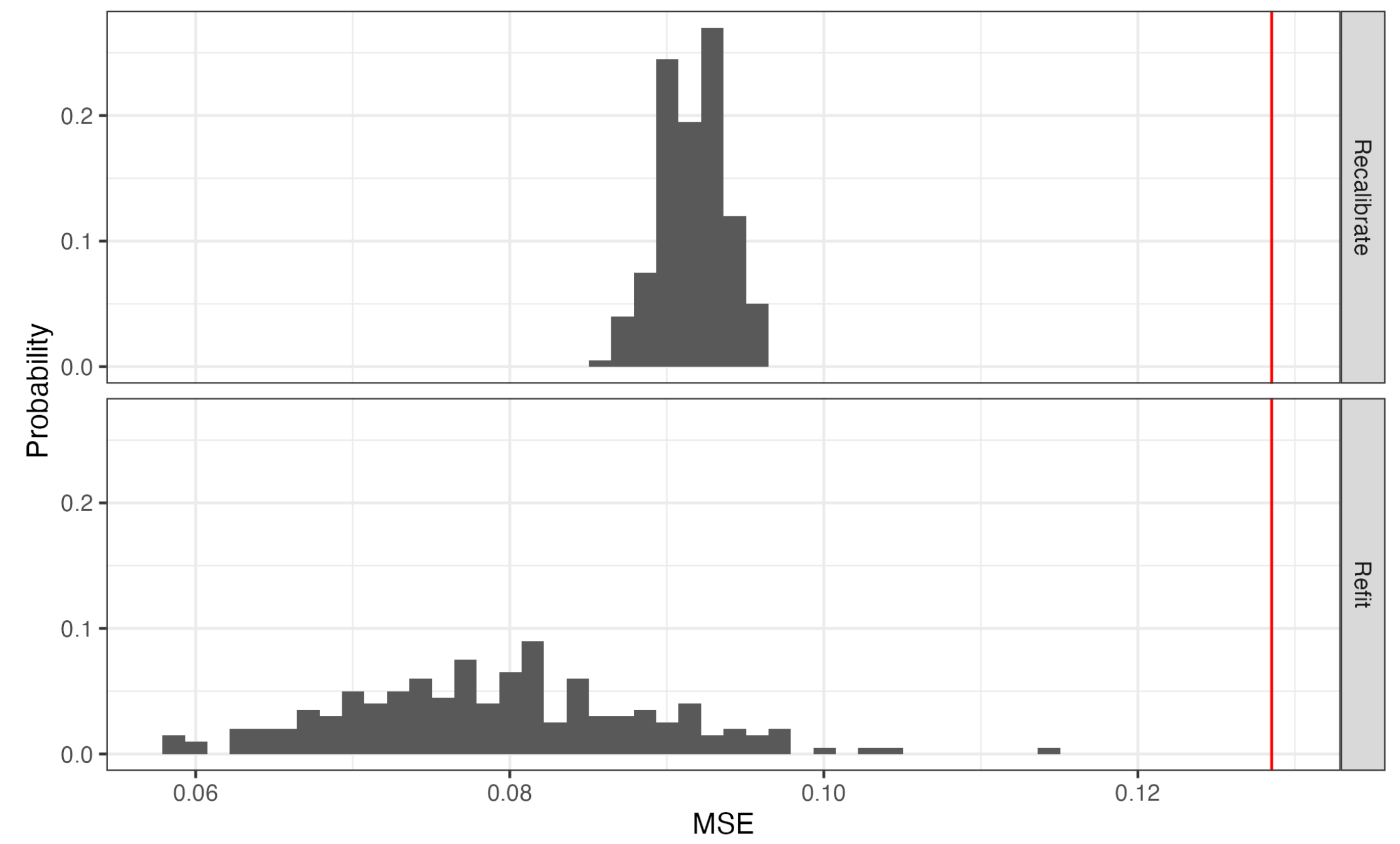
### AI Model Recalibration Algorithm (Grid Search)

```
1. For each calibration point $t \in \{\tau_1, ..., \tau_T\}$, perform PPI++ and compute $\hat{W}_{i,refit}, \hat{D}_{i,refit}$, and $\hat{\gamma}_i$
2. Fit time series for $\hat{W}_{i,refit}, \hat{D}_{i,refit}$, and $\hat{\gamma}_i$.
3. Using said time series, estimate the forcasted distribution for for $\hat{W}_{i,refit}, \hat{\Delta}_{i,refit}$, and $\gamma_i$ at $t_{future}$.
4. For each sample size ratios $\zeta \in (0, ..., 1)$:
   (a) For $b = 1, ...B$:
       i. Draw $\hat{\gamma}^b$, $\hat{W}_{i,refit}$, and $\hat{D}_{i,refit}$ from their forecast distributions
       ii. Draw $(1 - \zeta) * \frac{c}{c_{unlab}}$ samples of $X_{unlab}$ and $(\zeta * \frac{c}{c_{lab}}$ samples of $X_{lab}$ with replacement. Call these $X^b_{unlab}$ and $X^b_{lab}$ respectively
       iii. Let $MSE^b_{ref}(\eta) = (X^{bT}_{unlab}\hat{W}^b X_{unlab^b})^{-1} + \gamma^{b2}(X^{bT}_{lab}\hat{D}^b_{refit}X^b_{lab})^{-1}$
   (b) Let $MSE_{ref}(\eta) = Median(MSE^1_{ref}(\eta), ..., MSE^B_{ref}(\eta))$
5. $MSE_{ref} = \min_{\eta}(MSE_{ref}(\eta))$
```

- Seeks to estimate the optimal PPI++ correction factor at the point of forecast.
- Requires learning the rate of decay/increase for point estimate and correct factor variances.
- Estimates the optimal ratio of labeled to unlabeled data at the point of forecast as well as the MSE using such data will achieve.

## Economic Allocation

Given Distributional estimates of MSE, how do we decide which option is optimal for us?



**Refit** has the lower average MSE but higher variability while **recalibrate** is the opposite. The redline is average **retain** MSE. Deciding which is optimal requires specifying one's tradeoff between expected return and volatility.

### Arrow-Pratt Utility function

$$U(r) = E_P(r) - \frac{\lambda}{2}\sigma_P^2(r) - \frac{\theta}{2}\sigma_\mu^2(r).$$

- Extension of the workhorse of asset allocation in the financial industry
- Balances between expected return of asset $r$, $E_p(r)$, volatility due to stochastic noise, $\sigma_P(r)$, and volatility due to estimation error $\sigma\_\mu(r)$
- Larger $\lambda$ makes one more averse to MSE volatility from the data generation process, while larger $\theta$ makes one more averse to MSE volatility from calibration data variability

### Optimal Allocation

$$\lambda \begin{bmatrix} \sigma_P^2(MSE_{ref}) & \sigma_P(MSE_{ref}, MSE_{rec}) \\ \sigma_P(MSE_{ref}, MSE_{rec}) & \sigma_P^2(MSE_{rec}) \end{bmatrix} \begin{bmatrix} \hat{w}_{ref} \\ \hat{w}_{rec} \end{bmatrix} + \theta \begin{bmatrix} 0 & 0 \\ 0 & \sigma_\mu^2(MSE_{rec}) \end{bmatrix} = \begin{bmatrix} E_p(MSE_{ref}) - MSE_{ret)} \\ E_p(MSE_{rec}) - MSE_{ret)} \end{bmatrix}$$

- System of equations to compute the optimal allocation of assets (sample size) between **Refit** and **Recalibrate**

$$\text{Decision Rule} = \begin{cases} \text{Refit is Optimal} & \text{If } \hat{w}_{ref} > 0 \text{ and } \hat{w}_{ref} > \hat{w}_{ref} \\ \text{Recalibrate is Optimal} & \text{If } \hat{w}_{rec} > 0 \text{ and } \hat{w}_{ref} < \hat{w}_{rec} \\ \text{Retain is Optimal} & \text{If } \hat{w}_{rec} < 0 \text{ and } \hat{w}_{ref} < 0 \end{cases}$$

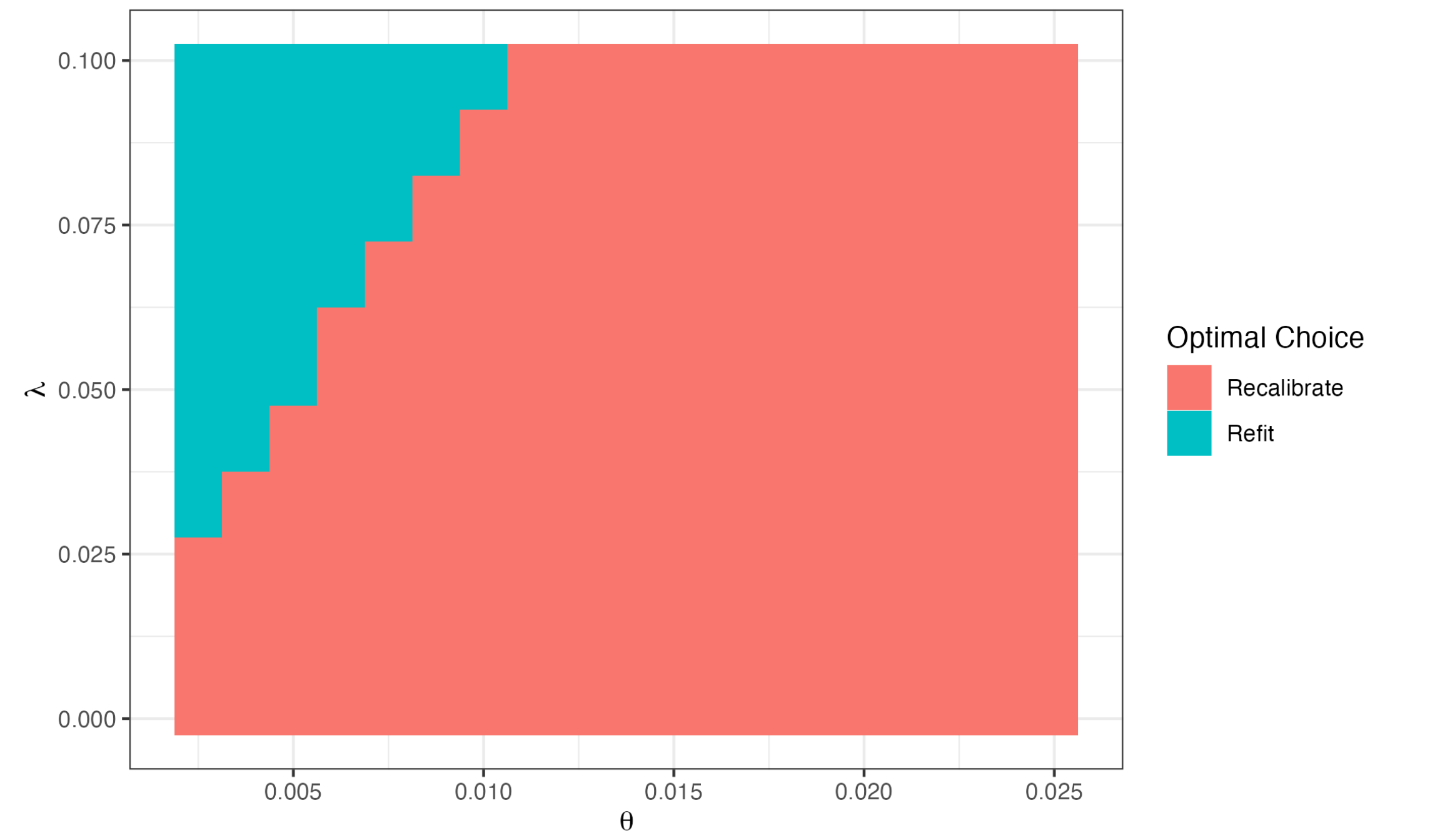
$$\hat{w}_{PPI++} = \frac{BD - HA}{CD - H^2}\hat{w}_{ref} = \frac{CA - HB}{CD - H^2}$$

$$A = -1 * (E_p(MSE_{ref}) - \overline{MSE_{ret}})$$
$$B = -1 * (E_p(MSE_{PPI++}) - \overline{MSE_{ret}})$$
$$C = \lambda\sigma_P(MSE_{ref})$$
$$D = \lambda\sigma_P(MSE_{PPI++}) + \theta\sigma_\mu(MSE_{ref})$$
$$H = \lambda\sigma_P(MSE_{ref}, MSE_{PPI++})$$

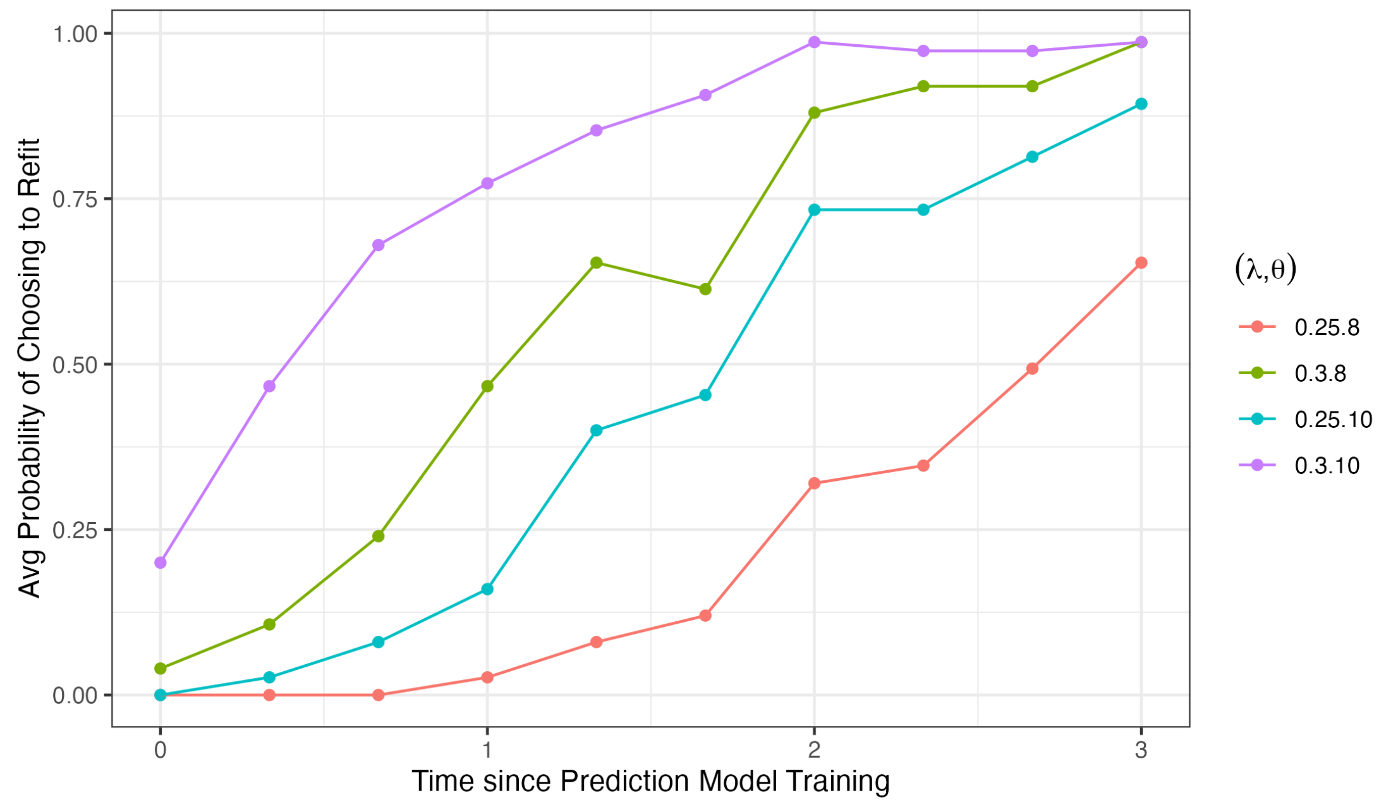
- Analytical solution to determine optimal allocation of budget to refit, $\hat{w}_{ref}$, and recalibrate, $\hat{w}_{rec}$

## Let's Save some Money!

Optimal economic decision for utility functions with various levels of $(\lambda, \theta)$ aversion.



Refitting becomes more economical over time



Recalibration becomes preferable when:

1. Cost of model fitting increases
2. Cost of labeled data increases
3. Point of forecast becomes closer to time of model fitting
4. More calibration data becomes available

**Take Away: Determining how to maintain a model is complex, but it can be done using model recalibration procedures and financial asset allocation**

Kentaro Hoffman, Tyler McCormick, Roy Van der Weide (2024+) Some models are useful, but for how long?
Kentaro Hoffman, Stephen Salerno, Awan Afiaz, Jeffrey T. Leek, & Tyler H. McCormick. (2024). Do We Really Even Need Data?.
Anastasios N. Angelopoulos, John C. Duchi, & Tijana Zrnic. (2024). PPI++: Efficient Prediction-Powered Inference.