HIFI: A HIGHLY EFFICIENT FRAMEWORK FOR STATISTICAL INFERENCE WITH FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models provide powerful representations for unstructured data, such as text and images, and parameter-efficient fine-tuning (PEFT) makes their adaptation to downstream tasks increasingly practical. For neural estimators that lack a tractable sampling distribution, resampling-based procedures offer a flexible route to statistical inference, but they often require hundreds of model fits, making them computationally prohibitive at foundation-model scale. We introduce HIFI, a two-stage fine-tuning framework for statistical inference on parameters such as conditional means. Stage 1 adapts a foundation model once on a larger sample, capturing the complex structure of the data and controlling estimation bias through a single fit. Stage 2 applies a resampling inference procedure to a smaller independent sample, concurrently fitting multiple lightweight PEFT estimators to the residual to quantify estimation variability without relearning the full task. HIFI accommodates a broad class of resampling procedures, including ensemble subsampling and the nonparametric bootstrap. Across vision and text experiments, HIFI scales to pretrained models ranging from tens of millions to more than ten billion parameters while providing accurate point estimates and well-calibrated confidence intervals across diverse architectures, and reduces training compute by up to 91.9% relative to direct resampling. An application to gene-expression inference illustrates the practical value of the framework for uncertainty-aware scientific inference.

1 INTRODUCTION

Pretrained foundation models have changed how predictive models are built, especially for text, images, and other unstructured data (Bommasani et al., 2021). Rather than estimating every model parameter from a task-specific sample, practitioners can adapt a pretrained backbone using a comparatively small number of trainable parameters (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Hu et al., 2022; Wang et al., 2025). Parameter-efficient fine-tuning (PEFT) therefore makes powerful representations available when full fine-tuning is prohibitively expensive. In many scientific applications, however, an accurate point estimate is not the final goal, because reliable scientific interpretation requires quantifying the uncertainty associated with that estimate. We therefore seek a confidence interval for the population conditional mean $\mu_0(x) = \mathbb{E}(Y \mid X = x)$ at a fixed input x.

Several approaches have been developed for inference[1] on conditional means estimated by neural networks. Existing methods include approaches that exploit specialized analytical structure (Liu et al., 2022; Kallus & McInerney, 2022; Sluijterman et al., 2025), sample-splitting procedures such as HulC (Kuchibhotla et al., 2024), and resampling methods (Fei & Li, 2024; Meng & Li, 2025; Wu & Politis, 2025; Arie & Gorfine, 2024). We focus on the resampling-based class because it requires relatively little model-specific analytical structure and therefore provides a flexible route to inference for complex neural estimators. These procedures construct uncertainty estimates from multiple fitted estimators, but doing so introduces two central challenges. First, fitting multiple estimators does not by itself eliminate systematic estimation error, whose control is typically required for valid inference. Second, obtaining many sufficiently accurate fits can be computationally expensive. This cost becomes particularly severe for large foundation models, since each additional fit still requires substantial forward–backward computation even when only a small subset of parameters is updated.

[1]Throughout this paper, "inference" refers to statistical inference unless otherwise specified.

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.

Our contributions. We introduce HIFI, a highly efficient two-stage PEFT framework for statistical inference with foundation models, illustrated in Figure 1. Stage 1 uses a larger sample and a higher-capacity adaptation to capture the complex, task-relevant structure of the data, controlling estimation bias through a single fit. Stage 2 then uses a smaller independent sample to resample and ensemble lightweight PEFT estimators of the remaining residual, quantifying estimation variability without relearning the full task. Because each Stage 2 fit involves a small adaptation module on few samples, repeated fitting is less expensive, and a batched implementation trains many such estimators concurrently over the Stage 1 adapted backbone. Together, these strategies make resampling-based inference practical at foundation-model scale.

The framework accommodates multiple inference methods. We consider the ensemble subsampling method (ESM) (Meng & Li, 2025) and nonparametric bootstrap. Across vision and text tasks and diverse pretrained architectures, the framework provides well-calibrated confidence intervals while reducing training computation by up to 92% at the same total sample size. We further demonstrate its use in a biological application to gene-expression prediction from H&E images.

2 Related Work

Parameter-efficient adaptation. A large literature studies parameter-efficient adaptation of pretrained models, including adapter-based, prompt- and prefix-based, and low-rank adaptation methods (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Hu et al., 2022; Wang et al., 2025). LoRA, for example, freezes the pretrained weights and learns low-rank updates to selected weight matrices, substantially reducing the number of trainable parameters and the memory required for adaptation (Hu et al., 2022). These developments make adaptation of increasingly large pretrained models practical, but primarily target downstream prediction rather than frequentist statistical inference.

Uncertainty quantification for neural networks. Our inferential target differs from several major strands of uncertainty quantification for modern neural networks. Deep ensembles train multiple neural networks independently from different initializations and aggregate their predictive distributions to quantify predictive uncertainty (Lakshminarayanan et al., 2017), while calibration methods assess whether reported confidence scores reflect empirical frequencies (Guo et al., 2017; Minderer et al., 2021). Bayesian approaches instead characterize uncertainty through posterior distributions over model parameters and predictions, commonly summarized by the posterior predictive distribution $p(y_{\text{new}} \mid x, \mathcal{D}) = \int p(y_{\text{new}} \mid x, \phi) p(\phi \mid \mathcal{D})\, d\phi$ (Gal & Ghahramani, 2016). Recent Bayesian PEFT methods restrict this posterior treatment to low-dimensional adaptation parameters (Yang et al., 2024; Wang et al., 2024). Conformal methods construct prediction sets for future outcomes with coverage $P\{Y_{\text{new}} \in \widehat{C}_\alpha(X_{\text{new}})\} \geq 1 - \alpha$ under exchangeability (Lei et al., 2018; Angelopoulos & Bates, 2023). Their inferential object is the future response Y_{new}, whereas our target is the population conditional mean $\mu_0(x) = \mathbb{E}(Y \mid X = x)$ at a fixed input x, for which we seek a confidence interval satisfying $P\{\mu_0(x) \in C_\alpha^\mu(x)\} \to 1 - \alpha$. This is often harder than predictive inference. For example, in binary classification, an informative prediction set needs calibrated probability estimation (Guo et al., 2017), whereas a confidence interval for $\mu_0(x)$ must shrink as data accumulate, and its validity requires the estimation bias to be negligible relative to the estimation variability, beyond calibration.

Frequentist inference for neural networks. Several approaches have been developed for frequentist inference on neural-network estimators, including specially constructed estimators (Liu et al., 2022), objective perturbation (Kallus & McInerney, 2022), and likelihood-ratio methods (Sluijterman et al., 2025). These approaches provide principled inferential guarantees but rely on analytical structure specific to the estimator or training objective. A different route is provided by HulC, which constructs confidence sets from estimators trained on independent data splits, with validity governed by their median bias (Kuchibhotla et al., 2024). In parallel, much of the literature on frequentist inference for neural-network regression relies on resampling-based methods, which require relatively little model-specific analytical structure and can therefore be applied more flexibly to complex neural estimators. Ensemble-subsampling approaches train estimators on multiple subsamples and exploit their dependence structure for inference (Fei & Li, 2024; Meng & Li, 2025; Wu & Politis, 2025), while nonparametric bootstrap methods refit neural estimators on resampled datasets (Arie & Gorfine, 2024). However, these methods have primarily been developed with conventional DNNs, with computational savings obtained through parallelization or smaller subsamples rather than batched multi-estimator training.

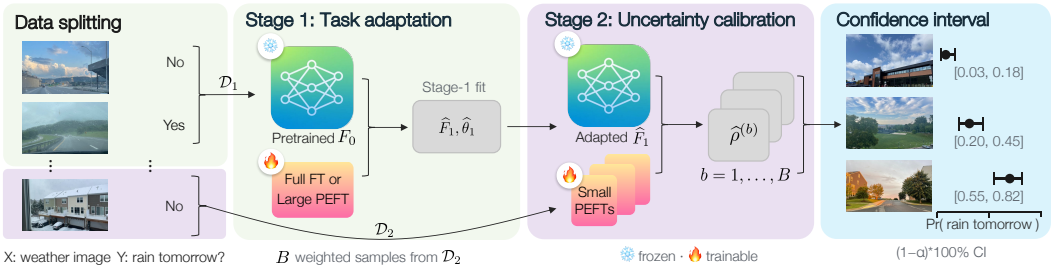


Figure 1: Overview of HIFI. Stage 1 performs a single task-adaptation fit on $\mathcal{D}_1$, yielding an adapted backbone $\widehat{F}_1$ and predictor $\widehat{\theta}_1$. Holding these components fixed, Stage 2 fits B estimator-specific residual PEFT modules using different sampling weights on the independent sample $\mathcal{D}_2$. The residual estimates are then combined to obtain a point estimate and pointwise confidence interval for $\mu_0(x)$.

Efficient ensemble and multi-adapter training. A separate line of work develops efficient mechanisms for training multiple predictors or parameter-efficient adapters over a shared backbone. BatchEnsemble shares model parameters across ensemble members and updates multiple members concurrently (Wen et al., 2020), while LoRA-based ensemble methods use a common pretrained backbone with member-specific low-rank adapters for predictive uncertainty and calibration (Wang et al., 2023; Mühlematter et al., 2026). More recently, Li et al. (2026) develop systems support for concurrently training multiple LoRA jobs over a shared backbone. These methods primarily target efficient ensembling, predictive uncertainty, or multi-adapter training. Our work instead brings these computational ideas into resampling-based frequentist inference by using batched PEFT training.

3 HIFI Framework

Let $Z = (X, Y) \sim P$, where X is a potentially high-dimensional or unstructured input and Y is a scalar response. The inferential target is the conditional mean $\mu_0(x) = \mathbb{E}(Y \mid X = x)$. We work on the link scale $\theta_0(x) = g\{\mu_0(x)\}$, where g is a known monotone link function, including the identity, logit, and log links for continuous, binary, and count outcomes, respectively. Pointwise inference is performed for $\theta_0(x)$ and transformed to the response scale through g^{-1}.

As illustrated in Figure 1, we separate task adaptation from statistical inference through a two-stage sample-splitting strategy. We split the labeled data into two independent samples, $\mathcal{D}_1 = \{Z_i^{(1)}\}_{i=1}^{n_1}$ and $\mathcal{D}_2 = \{Z_i^{(2)}\}_{i=1}^{n_2}$. The first sample is used for a single Stage-1 task adaptation, while the second is used by resampling-based procedures to construct Stage-2 estimators for statistical inference.

3.1 Stage-1 Task Adaptation

Let F_0 denote a pretrained foundation model. Stage 1 applies an adaptation procedure $\mathcal{A}_1$ to F_0 using $\mathcal{D}_1$, yielding $(\widehat{F}_1, \widehat{\theta}_1) = \mathcal{A}_1(F_0, \mathcal{D}_1, U_1)$, where $\widehat{F}_1$ is the adapted backbone, $\widehat{\theta}_1$ is the corresponding predictor with a task-specific prediction head, and U_1 collects training randomness. Stage 1 may use full fine-tuning or a higher-capacity PEFT parameterization.

To specify the Stage-1 optimization, let $\ell_1(y, t)$ denote a loss, whose conditional population risk is minimized at $t = \theta_0(x)$. The training objective is

$$\widehat{\mathcal{L}}_1(\theta) = \frac{1}{n_1} \sum_{i=1}^{n_1} \ell_1 \left(Y_i^{(1)}, \theta(X_i^{(1)}) \right), \tag{1}$$

where θ ranges over predictors induced by the adaptation parameterization and prediction head.

3.2 Stage 2: Uncertainty Calibration

Conditional on the Stage-1 fit, define the residual signal as $\rho_0(x) = \theta_0(x) - \widehat{\theta}_1(x)$. Stage 2 estimates it using B sampled or weighted versions of $\mathcal{D}_2$, represented by sampling-weight vectors $W_1, \ldots, W_B$, with $W_b = (W_{b1}, \ldots, W_{bn_2})$. The sampling weights depend on the inference method $\mathfrak{I}_\alpha$.

For each b, the Stage-1 adapted backbone $\widehat{F}_1$ and predictor $\widehat{\theta}_1$ are held fixed and shared across all Stage 2 fits. Starting from $\widehat{F}_1$, the bth fit learns estimator-specific PEFT parameters and a residual prediction head, yielding $\widehat{\rho}_b = \mathcal{A}_2(\widehat{F}_1, \mathcal{D}_2, W_b; \widehat{\theta}_1, U_{2,b})$, where $\mathcal{A}_2$ denotes the Stage-2 adaptation procedure and $U_{2,b}$ collects training randomness.

More specifically, let $\ell_2(y,t)$ denote the Stage-2 loss, which may be the same as or different from the Stage-1 loss ℓ_1, and whose conditional population risk is minimized at $t = \theta_0(x)$. Holding $\widehat{\theta}_1$ fixed, the bth training objective is

$$\widehat{\mathcal{L}}_{2,b}(\rho) = \frac{1}{\sum_{i=1}^{n_2} W_{bi}} \sum_{i=1}^{n_2} W_{bi}\, \ell_2\Big(Y_i^{(2)}, \widehat{\theta}_1(X_i^{(2)}) + \rho(X_i^{(2)})\Big). \tag{2}$$

Let $C_\alpha^\rho(x)$ denote a $(1-\alpha)$ confidence interval for $\rho_0(x)$. An inference method $\mathfrak{I}_\alpha$ constructs it from the sampling weights and Stage-2 residual estimates as $C_\alpha^\rho(x) = \mathfrak{I}_\alpha(\{W_b, \widehat{\rho}_b(x)\}_{b=1}^B)$. Since $\theta_0(x) = \widehat{\theta}_1(x) + \rho_0(x)$ conditional on Stage 1, the corresponding intervals on the link and response scales are $C_\alpha^\theta(x) = \widehat{\theta}_1(x) + C_\alpha^\rho(x)$ and $C_\alpha^\mu(x) = g^{-1}\{C_\alpha^\theta(x)\}$, respectively. This formulation is method-agnostic. In this paper, we use ESM and bootstrap.

3.3 HIFI Instantiation with LoRA and ESM

We illustrate HIFI using LoRA with ESM. The corresponding bootstrap instantiation is given in Appendix B.1. In Stage 1, for each selected pretrained weight matrix $M_{\ell,0}$, LoRA learns

$$\widehat{M}_{\ell,1} = M_{\ell,0} + \frac{\lambda_1}{q_1} \widehat{L}_{\ell,1} \widehat{R}_{\ell,1},$$

where q_1 and λ_1 are the LoRA rank and scaling parameter, respectively. The LoRA factors and prediction head are optimized according to equation 1, yielding the adapted backbone $\widehat{F}_1$ and predictor $\widehat{\theta}_1$.

For Stage 2, ESM independently draws B size-r subsets $I_1, \ldots, I_B$ of $\{1, \ldots, n_2\}$, with $W_{bi} = \mathbf{1}\{i \in I_b\}$. Holding the Stage 1 adapted matrices fixed, the bth fit uses

$$\widehat{M}_{\ell,2,b} = \widehat{M}_{\ell,1} + \frac{\lambda_2}{q_2} \widehat{L}_{\ell,2,b} \widehat{R}_{\ell,2,b}, \qquad b = 1, \ldots, B,$$

where q_2 and λ_2 are the Stage 2 LoRA rank and scaling parameter, respectively. Each fit learns new LoRA factors and a residual prediction head according to equation 2, yielding $\widehat{\rho}_b$.

ESM averages the residual estimates as $\widehat{\rho}_{\mathrm{ESM}}(x) = B^{-1} \sum_{b=1}^B \widehat{\rho}_b(x)$, yielding the link-scale point estimate $\widehat{\theta}_{\mathrm{ESM}}(x) = \widehat{\theta}_1(x) + \widehat{\rho}_{\mathrm{ESM}}(x)$. Let $\overline{W}_i = B^{-1} \sum_{b=1}^B W_{bi}$. The infinitesimal jackknife variance estimator is

$$\widehat{s}^2_{\mathrm{ESM}}(x) = \frac{n_2(n_2-1)}{(n_2-r)^2} \sum_{i=1}^{n_2} \left[\frac{1}{B} \sum_{b=1}^B (W_{bi} - \overline{W}_i)\{\widehat{\rho}_b(x) - \widehat{\rho}_{\mathrm{ESM}}(x)\} \right]^2.$$

The resulting pointwise interval is $C_\alpha^{\theta,\mathrm{ESM}}(x) = \widehat{\theta}_{\mathrm{ESM}}(x) \pm z_{1-\alpha/2} \widehat{s}_{\mathrm{ESM}}(x)$, with the response-scale interval obtained through g^{-1} and point estimate $\widehat{\mu}_{\mathrm{ESM}}(x) = g^{-1}\{\widehat{\theta}_{\mathrm{ESM}}(x)\}$.

3.4 Efficient Batched Stage-2 Training

The resampling-based Stage 2 procedures above require fitting B estimators, which can be computationally expensive when performed sequentially even under PEFT. To reduce this overhead, we stack the estimator-specific LoRA parameters and prediction heads along an additional ensemble dimension and execute their forward and backward computations using batched tensor operations on the GPU. The frozen backbone weights are shared across fits. For a group of B_{chunk} fits, each fit retains its own resampling weights, minibatch, loss, and optimizer state, while gradient updates are confined to its estimator-specific parameter slice. This batching strategy improves GPU utilization by replacing separate training loops with larger batched operations and is algebraically equivalent to fitting the same estimators separately, up to numerical differences.

The B fits are processed in groups of at most B_{chunk} estimators, with group size determined by GPU memory. Predictions are collected across groups before constructing confidence intervals. Although we use ESM and bootstrap, the same construction applies to other resampling-based procedures.

4 Theoretical Motivation

We use an idealized sparse ReLU-DNN setting to motivate the statistical and computational rationale of HIFI. Building on approximation–estimation risk decompositions (Schmidt-Hieber, 2020; Meng & Li, 2025), Appendix A derives risk bounds under explicit compatibility and learning assumptions. We then describe additional conditions under which conditional ESM validity transfers to unconditional coverage. This analysis provides theoretical motivation for the two-stage design.

Consider the γ-Hölder, d-dimensional special case of the bounded sparse ReLU-DNN setting in Meng & Li (2025). Let $\widehat{\theta}^{\text{single}}$ and $\widehat{\theta}^{\text{HIFI}}$ denote the direct-resampling and HIFI estimators, respectively, and let $\nu = 2\gamma/(2\gamma + d)$. Let r be the Stage 2 subsample size, and suppose the direct-resampling subsample size and Stage 1 sample size are $c_0 r$ and $c_1 r$, respectively, for fixed $c_0, c_1 \geq 1$. Applying Proposition A.3 in Appendix A.1 together with the corresponding DNN excess-risk rate gives

$$\begin{aligned} \mathbb{E}\|\widehat{\theta}^{\text{single}} - \theta_0\|_2^2 &\lesssim (c_0 r)^{-\nu} L_{c_0 r} \log^2(c_0 r), \\ \mathbb{E}\|\widehat{\theta}^{\text{HIFI}} - \theta_0\|_2^2 &\lesssim (c_1 r)^{-\nu} L_{c_1 r} \log^2(c_1 r) + r^{-\nu} L_r \log^2 r. \end{aligned} \tag{3}$$

Here L_s denotes the network depth for a fit on s observations.

The key compatibility requirement is that, conditional on Stage 1, the remaining Stage 2 class retains sufficient approximation capacity for the residual target (Assumption A.1 in Appendix A.1). Single-stage resampling refits all parameters on each subsample of size $c_0 r$, whereas HIFI fits the shared component once using $c_1 r$ observations and refits only the Stage 2 component using subsamples of size r. For comparable depth scaling, the two bounds have the same asymptotic rate order in r. Thus, HIFI reduces repeated fitting without changing the nonparametric rate order.

Valid inference. Write $h = (F, t)$ for a generic Stage 1 output and $\widehat{h}_1 = (\widehat{F}_1, \widehat{\theta}_1)$ for the fitted Stage 1 output. For fixed h, define the residual target $\rho_h = \theta_0 - t$. In addition to the excess-risk control in equation 3, valid inference requires the regularity conditions of the chosen resampling method. Conditional on a deterministic admissible h, Stage 2 targets the fixed residual function ρ_h. Suppose that, with h held fixed, the resulting Stage 2 DNN–ESM problem satisfies the conditions in Corollary 3.6 and Theorems 3.7–3.8 of Meng & Li (2025). Then the corresponding conditional inference is asymptotically valid.

To account for Stage 1 randomness, suppose there exist deterministic sets $\mathcal{H}_{n_2}$ and $\mathcal{T}_{n_2} \subseteq \mathcal{X}$ with $P(\widehat{h}_1 \in \mathcal{H}_{n_2}) \to 1$ and $P_X(\mathcal{T}_{n_2}) \to 1$, such that conditional ESM validity holds uniformly over $h \in \mathcal{H}_{n_2}$ and $x \in \mathcal{T}_{n_2}$. Then, for every deterministic sequence $x_{n_2} \in \mathcal{T}_{n_2}$, averaging over Stage 1 randomness yields $P\{\theta_0(x_{n_2}) \in C_\alpha^{\theta,\text{ESM}}(x_{n_2}; \widehat{h}_1)\} \to 1 - \alpha$. Thus, the HIFI–ESM interval attains asymptotically nominal unconditional coverage at any deterministic target sequence lying in the asymptotically full-probability target sets $\mathcal{T}_{n_2}$. Details are given in Appendix A.2.

Computational gains. Suppose training cost scales as the product of trainable parameters and sample size. With B resampling fits, single-stage resampling costs $\mathcal{C}_{\text{single}} = B p_0 c_0 r$, whereas HIFI costs $\mathcal{C}_{\text{HIFI}} = p_1 c_1 r + B p_2 r$, since the larger Stage 1 component is trained only once. For $p_1 = 0.8 p_0$, $p_2 = 0.2 p_0$, $c_0 = 4$, and $c_1 = 16$, corresponding to an 80/20 allocation of trainable parameters between Stages 1 and 2 and sample sizes $4r$ and $16r$ for the single-stage and Stage 1 fits, respectively, equation 3 gives the same excess-risk order under comparable depth scaling, while $\mathcal{C}_{\text{HIFI}}/\mathcal{C}_{\text{single}} = 3.2/B + 0.05$. At $B = 500$, HIFI therefore uses 5.64% of the training cost.

5 Synthetic Experiments

We evaluate HIFI across two modalities and six pretrained backbones. The vision task considers continuous-outcome regression, whereas the text task considers Bernoulli probability estimation. All methods use LoRA, with HIFI applying it in both stages. We evaluate ESM and bootstrap within HIFI and compare against the external comparators HulC (Kuchibhotla et al., 2024) and Bayesian-PEFT, adapted from the Bayesian neural-network framework of Sun et al. (2021).

Table 1: Statistical inference with HIFI and external baselines across pretrained vision models.

Model	Params	Method	Point Estimation		Statistical Inference		
			MAB	MAE	Cov. 90%	Cov. 95%	95% CI Length
ViT-B/32	87.5M	HIFI-ESM	0.018 (0.02)	0.041 (0.01)	0.908 (0.08)	0.937 (0.06)	0.282 (0.02)
		HIFI-Bootstrap	0.017 (0.02)	0.047 (0.01)	0.888 (0.07)	0.938 (0.05)	0.231 (0.02)
		HulC	0.028 (0.03)	0.040 (0.02)	0.867 (0.11)	0.914 (0.10)	0.206 (0.03)
		Bayesian-PEFT	0.054 (0.05)	0.060 (0.05)	0.702 (0.33)	0.764 (0.32)	0.183 (0.04)
DINOv2-B/14	86.6M	HIFI-ESM	0.011 (0.01)	0.044 (0.01)	0.933 (0.04)	0.955 (0.03)	0.319 (0.04)
		HIFI-Bootstrap	0.011 (0.01)	0.046 (0.01)	0.908 (0.04)	0.953 (0.03)	0.243 (0.03)
		HulC	0.021 (0.02)	0.073 (0.02)	0.931 (0.03)	0.965 (0.02)	0.498 (0.13)
		Bayesian-PEFT	0.029 (0.03)	0.049 (0.03)	0.984 (0.05)	0.993 (0.03)	0.484 (0.04)
Swin-B	87.8M	HIFI-ESM	0.013 (0.01)	0.035 (0.01)	0.925 (0.05)	0.951 (0.04)	0.236 (0.03)
		HIFI-Bootstrap	0.013 (0.01)	0.042 (0.01)	0.893 (0.06)	0.943 (0.04)	0.215 (0.03)
		HulC	0.028 (0.03)	0.041 (0.02)	0.860 (0.13)	0.906 (0.12)	0.207 (0.03)
		Bayesian-PEFT	0.033 (0.03)	0.057 (0.03)	0.881 (0.15)	0.925 (0.12)	0.281 (0.03)

Notes. MAB/MAE denote mean absolute bias/error; parentheses report target-wise SD; Bayesian-PEFT reports posterior credible intervals.

Unless otherwise stated, each experiment uses 100 replications with independently generated Stage 1 and Stage 2 samples, split into 8,000/500 and 1,000/200 training/validation observations, respectively. Training in both stages uses early stopping based on the corresponding validation set. For the external comparators, the Stage 1 and Stage 2 samples are pooled, so that each method uses all 9,700 observations. ESM uses 500 subsamples of size 501 without replacement, while bootstrap uses 400 resamples of size 1,000 with replacement. For evaluation, the vision task uses 500 target images and the text task uses 424 target texts, all held constant across replications. Implementation details for the inference methods and external comparators are provided in Appendix B, with model-specific architecture, adaptation, and optimization settings reported in Appendix D.

5.1 Task 1: Regression from Synthetic Images

We consider a vision-regression setting in which each input X is a 224×224 RGB image containing a single geometric object. The shape S is sampled uniformly from $\{\bigcirc, \square, \triangle\}$, and its geometric area fraction a is sampled from $\mathrm{Unif}(0.10, 0.32)$. Distinct foreground and background colors are sampled from a four-color palette, while object locations and the orientations of squares and triangles are randomized. Responses are generated as $Y = \beta(S)a + \varepsilon$, where $\varepsilon \sim N(0, 0.25^2)$, $\beta(\bigcirc) = 1$, $\beta(\square) = 2$, and $\beta(\triangle) = 3$. For a fixed target image x, let $S(x)$ and $a(x)$ denote its generating shape and geometric area fraction. By construction, $\mu_0(x) = \beta\{S(x)\}a(x)$, and our goal is to estimate $\mu_0(x)$ and construct a pointwise confidence interval. The model is trained only on the observed image–response pairs (X, Y) and is not given $S(X)$ or $a(X)$. Additional data-generation details are provided in Appendix C.1, with representative images shown in Figure 5.

We evaluate three pretrained vision backbones spanning different architectural and pretraining designs: ViT-B/32 (Dosovitskiy et al., 2021), a standard vision transformer; DINOv2 ViT-B/14 (Oquab et al., 2024), a self-supervised vision transformer; and Swin-B (Liu et al., 2021), a hierarchical shifted-window transformer. Across all three pretrained vision backbones, HIFI achieves low point-estimation error and stable near-nominal coverage. At the 95% level, HIFI-ESM achieves coverage of 0.937–0.955 and HIFI-Bootstrap 0.938–0.953 (Table 1), with HIFI-Bootstrap yielding the shorter 95% intervals across all three backbones. In contrast, HulC and Bayesian-PEFT exhibit less stable coverage across backbones, with 95% coverage ranging from 0.906 to 0.965 and 0.764 to 0.993, respectively. For 95% coverage, target-wise standard deviations remain within 0.03–0.06 for HIFI, compared with 0.02–0.12 for HulC and 0.03–0.32 for Bayesian-PEFT across backbones, while additional target-level diagnostics in Appendix E.1 show similarly stable behavior, indicating that HIFI provides robust inference across targets within each architecture.

5.2 Task 2: Probability Prediction from Text

We consider a controlled bilingual text task in which each input X describes when a passenger arrives relative to a scheduled flight, and $Y = 1$ indicates that the passenger catches the flight. Let $T(X) \in$

Table 2: Statistical inference with HIFI and external baselines across pretrained language models.

			Point Estimation		Statistical Inference		
Model	Params	Method	MAB	MAE	Cov. 90%	Cov. 95%	95% CI Length
Qwen3-Embedding-8B	7.6B	HIFI-ESM	0.014 (0.01)	0.060 (0.01)	0.899 (0.03)	0.928 (0.03)	0.374 (0.06)
		HIFI-Bootstrap	0.015 (0.01)	0.064 (0.02)	0.885 (0.04)	0.937 (0.03)	0.316 (0.07)
		HulC	0.075 (0.05)	0.097 (0.04)	0.849 (0.10)	0.889 (0.09)	0.418 (0.06)
		Bayesian-PEFT	0.085 (0.06)	0.102 (0.05)	0.830 (0.22)	0.890 (0.18)	0.426 (0.05)
XLM-R XXL	10.7B	HIFI-ESM	0.014 (0.01)	0.051 (0.01)	0.899 (0.04)	0.926 (0.03)	0.361 (0.08)
		HIFI-Bootstrap	0.016 (0.01)	0.061 (0.02)	0.878 (0.07)	0.930 (0.05)	0.311 (0.10)
		HulC	0.035 (0.03)	0.057 (0.02)	0.882 (0.07)	0.932 (0.06)	0.313 (0.06)
		Bayesian-PEFT	0.031 (0.02)	0.042 (0.02)	0.993 (0.03)	0.998 (0.01)	0.359 (0.05)
NV-Embed-v2	7.9B	HIFI-ESM	0.013 (0.01)	0.052 (0.01)	0.935 (0.02)	0.945 (0.02)	0.522 (0.08)
		HIFI-Bootstrap	0.010 (0.01)	0.048 (0.01)	0.952 (0.03)	0.970 (0.02)	0.321 (0.08)
		HulC	0.052 (0.04)	0.069 (0.03)	0.861 (0.11)	0.885 (0.10)	0.306 (0.07)
		Bayesian-PEFT	0.099 (0.07)	0.101 (0.07)	0.573 (0.45)	0.598 (0.45)	0.266 (0.02)

Notes. MAB/MAE denote mean absolute bias/error; parentheses report target-wise SD. Bayesian-PEFT reports posterior credible intervals.

$[0, 3.5]$ denote the arrival lead time encoded in the text, measured in hours. Larger values of $T(X)$ correspond to earlier arrivals and therefore higher success probabilities. Specifically, the conditional success probability is $\mu_0(X) = \text{logit}^{-1}\{2.8[T(X) - 1.05]\}$, and $Y \mid X \sim \text{Bernoulli}\{\mu_0(X)\}$. For a fixed target text x, our goal is to estimate $\mu_0(x)$ and construct a pointwise confidence interval.

For each observation, the arrival lead time T is generated using probability-stratified sampling, and the language is independently drawn as English or Chinese with equal probability. Conditional on T and the sampled language, the arrival-time information is expressed in natural language to form the text input X. Additional auxiliary flight information and response-irrelevant contextual information are included randomly to introduce nuisance textual variation. During training, the model observes only (X, Y) and is not given $T(X)$ or $\mu_0(X)$ explicitly. Representative examples are shown in Appendix Table 3, with complete data-generation details provided in Appendix C.2.

We evaluate three multi-billion-parameter language backbones with distinct designs: Qwen3-Embedding-8B (Zhang et al., 2025), a decoder-only embedding model; XLM-R XXL (Goyal et al., 2021), an encoder-only multilingual masked language model; and NV-Embed-v2 (Lee et al., 2025), a Mistral-based embedding model with latent-attention pooling.

The language experiments demonstrate that HIFI makes resampling-based inference feasible with multi-billion-parameter foundation models while maintaining reliable statistical performance. Across three heterogeneous language backbones, HIFI-ESM attains 95% coverage of 0.926–0.945, while HIFI-Bootstrap ranges from 0.930–0.970 (Table 2). By comparison, the external baselines show greater variation across architectures, with coverage of 0.885–0.932 for HulC and 0.598–0.998 for Bayesian-PEFT. HIFI also exhibits stable target-level calibration, with 95% coverage standard deviations of 0.02–0.05 across backbones, compared with 0.06–0.10 for HulC and 0.01–0.45 for Bayesian-PEFT. Additional analyses in Appendix E.2 show similarly stable behavior across targets.

5.3 Ablation Study 1: Effect of Stage-1 Task-Adaptation

We study the effect of adding Stage 1 task-adaptation data while holding the Stage 2 inference sample fixed. Using the bilingual flight simulation with Qwen3-Embedding-8B, we compare single-stage direct-resampling baselines based on ESM and bootstrap, which operate directly on the Stage 2-sized sample, with HIFI using $n_1 \in \{2{,}000, 4{,}500, 8{,}500\}$ additional Stage 1 observations. Except for the Stage 1 sample size, all simulation settings, evaluation targets, training settings, and inference settings are kept identical to those of the main experiment and Appendix Table 6.

Figure 2 shows two patterns. First, augmenting the single-stage procedures with a Stage 1 task-adaptation sample substantially improves estimation accuracy and produces much shorter confidence intervals. The single-stage baselines attain conservative coverage but exhibit substantially larger estimation bias. Second, within HIFI, increasing the Stage 1 sample size further improves performance. For both ESM and bootstrap, MAB decreases and 95% confidence intervals narrow steadily

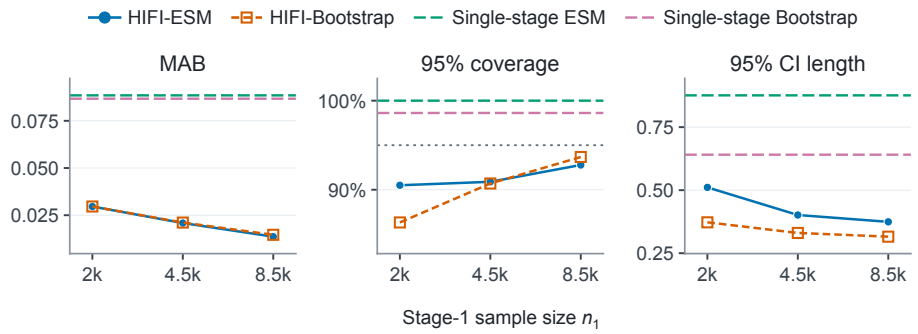


Figure 2: Effect of Stage 1 task adaptation on MAB, 95% coverage, and CI length.

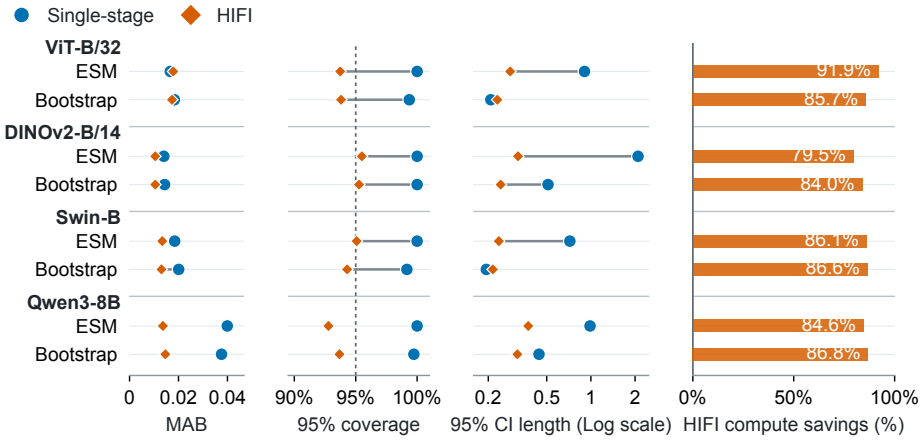


Figure 3: HIFI versus direct single-stage resampling under the same total data budget. HIFI training-compute savings measured in EFLOPs.

as n_1 increases, while coverage moves toward the nominal 95% level. At $n_1 = 8{,}500$, mean interval lengths are 51–57% smaller than those of the corresponding single-stage baselines, while empirical coverage reaches 92.8–93.7%, slightly below nominal. Appendix E.3 additionally reports the 90% results and target-wise analyses of estimation, coverage, and confidence-interval length, showing that the aggregate patterns are broadly consistent across targets.

5.4 Ablation Study 2: HIFI vs. Direct Resampling at Matched Sample Size

We further compare HIFI with direct single-stage resampling under a matched total sample size. The single-stage baseline pools the two samples (9,000 training and 700 validation observations) and applies ESM or bootstrap directly, using the same PEFT specification as HIFI Stage 2. We evaluate the three vision backbones and Qwen3-Embedding-8B with all other settings unchanged.

Figure 3 shows that the two-stage framework improves the statistical–computational trade-off at the same total sample size. Direct resampling is strongly conservative, with 95% coverage concentrated near 100% across model and inference methods, whereas HIFI tracks the nominal level much more closely, with coverage ranging from 92.8% to 95.5%. For ESM, HIFI also produces substantially shorter intervals across all four backbones. For bootstrap, intervals are shorter for DINOv2-B/14 and Qwen3-Embedding-8B and comparable for the other backbones. HIFI also yields lower MAB in seven of the eight comparisons with the largest improvements for Qwen3-Embedding-8B.

These statistical gains are accompanied by substantial computational savings: HIFI reduces training EFLOPs by 79.5–91.9% relative to direct resampling. Detailed training-compute calculations and wall-clock times are reported in Appendix E.5. The results are consistent with the two-stage design amortizing task adaptation in Stage 1 and using Stage 2 for uncertainty calibration. Additional 90% and target-wise results in Appendix E.4 show the same pattern.

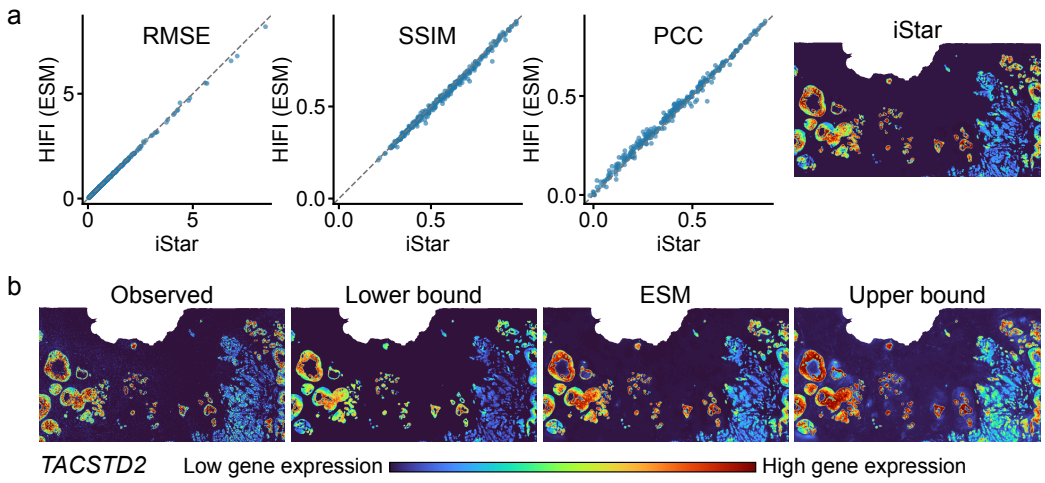


Figure 4: **Gene-expression inference on Xenium tissue Section 2. (a)** HIFI with ESM versus iStar across 313 genes by RMSE, SSIM, and PCC, with the iStar point estimate for *TACSTD2*. **(b)** HIFI point estimate and 95% pointwise confidence interval for *TACSTD2*.

6 Real Data Analysis

Spatial transcriptomics provides spatially resolved but inherently noisy measurements of gene expression. Estimating the conditional mean $\mu_0(x) = \mathbb{E}(Y \mid X = x)$ from local H&E morphology denoises these measurements and enriches the underlying signal (Zhang et al., 2024), and the resulting expression maps are routinely used in downstream analyses. Without uncertainty quantification, however, estimation error can be mistaken for biological structure and propagate into these analyses.

To demonstrate the practical value of HIFI, we use the Xenium human breast cancer dataset from Janesick et al. (2023), which contains two tissue sections. We train on Section 1 and evaluate on the held-out Section 2, using its H&E image to predict all 313 measured genes, with the observed expression used for evaluation. To facilitate a fair comparison with iStar (Zhang et al., 2024), we adopt the same preprocessing pipeline and HViT backbone, while using our HIFI framework for model fitting. Additional implementation details and experimental settings are provided in Appendix F.

HIFI retains point-prediction performance comparable to iStar while additionally enabling uncertainty quantification. As shown in Figure 4a, HIFI with ESM performs comparably to iStar across root mean squared error (RMSE), structural similarity index (SSIM), and Pearson correlation coefficient (PCC). Figure 4b illustrates the corresponding uncertainty quantification for *TACSTD2*. The confidence maps exhibit spatially heterogeneous uncertainty: some high-expression regions remain pronounced at the lower endpoint, whereas other regions show larger differences between the lower and upper endpoints. Appendix Figure 19 provides confidence maps for other genes. Bootstrap-based HIFI similarly yields point-prediction performance comparable to iStar (Appendix Figure 18) while providing corresponding confidence interval maps (Appendix Figure 20).

7 Discussion

We introduced HIFI, a two-stage framework that separates task adaptation from repeated uncertainty calibration for statistical inference with foundation models. By sharing a Stage 1 task-adapted representation across smaller Stage 2 PEFT fits, HIFI reduces the computational burden of resampling-based inference. Across vision and text tasks, HIFI achieves accurate point estimation and stable near-nominal coverage with substantially less training compute than direct resampling. The Xenium application further demonstrates its utility for uncertainty-aware biological research.

Several limitations remain. First, our experiments focus on LoRA, and extending HIFI to other PEFT parameterizations and resampling methods could broaden its applicability. Second, HIFI relies on PEFT to reduce estimation error, but PEFT effectiveness can be task-dependent, particularly for complex reasoning tasks. Repeated Stage 2 fitting can also remain costly for very large models, motivating extensions to more efficient and reasoning-oriented foundation models.

Reproducibility Statement

We provide detailed information to support reproduction of both the theoretical and empirical results. The main text specifies the HIFI framework, inference procedures, and experimental protocols, while the appendix provides the assumptions and complete proofs for the theoretical results. For all experiments, we document the data-generation or preprocessing procedures, model backbones, Stage 1 and Stage 2 data splits, PEFT configurations, optimization settings, resampling schemes, and evaluation protocols. The appendix also reports additional implementation details, hyperparameters, target-wise diagnostics, and supplementary results for the text, vision, and Xenium experiments.

AI Use Statement

In this work, we used generative AI tools to generate synthetic data sets, assist with language editing, organization, and formatting of author-written proofs, provide feedback on research methodology or experiments, implement methods, assist with translation, and interpret results. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, or clean and reformat datasets. Qualitative and thematic data analysis is not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Asaf Ben Arie and Malka Gorfine. Confidence intervals and simultaneous confidence bands based on deep learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PdbaruPVUY.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL https://arxiv.org/abs/2108.07258.

Axel Bücher and Ivan Kojadinovic. A note on conditional versus joint unconditional weak convergence in bootstrap consistency results. *Journal of Theoretical Probability*, 32(3):1145–1165, 2019.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691. PMLR, 2014.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Zhe Fei and Yi Li. U-learning for prediction inference via combinatory multi-subsampling: With applications to LASSO and neural networks. *arXiv preprint arXiv:2407.15301*, 2024. URL https://arxiv.org/abs/2407.15301.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pp. 29–33, 2021.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications*, 14(1):8353, 2023.

Nathan Kallus and James McInerney. The implicit delta method. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=etY_XXnPkoC.

Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. The HulC: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):586–622, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Kevin Li, Dibyadeep Saha, Avni Kanodia, and Fan Lai. tLoRA: Efficient multi-LoRA training with elastic shared super-models. *arXiv preprint arXiv:2602.07263*, 2026. doi: 10.48550/arXiv.2602.07263. URL https://arxiv.org/abs/2602.07263.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Ruiqi Liu, Ben Boukai, and Zuofeng Shang. Optimal nonparametric inference via deep neural network. *Journal of Mathematical Analysis and Applications*, 505(2):125561, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Xuran Meng and Yi Li. Inferring outcome means of exponential family distributions estimated by deep neural networks. *arXiv preprint arXiv:2504.09347*, 2025. URL https://arxiv.org/abs/2504.09347v2. Version 2, revised April 15, 2025.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Dominik J. Mühlematter, Michelle Halbheer, Alexander Becker, Dominik Narnhofer, Helge Aasen, Konrad Schindler, and Mehmet Ozgur Turkoglu. LoRA-ensemble: Efficient uncertainty modelling for self-attention networks. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=yhXXmOMpSQ.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875.

Laurens Sluijterman, Eric Cator, and Tom Heskes. Likelihood-ratio-based confidence intervals for neural networks. *Machine Learning*, 114:116, 2025. doi: 10.1007/s10994-024-06639-3.

Yan Sun, Wenjun Xiong, and Faming Liang. Sparse deep learning: A new framework immune to local traps and miscalibration. *Advances in Neural Information Processing Systems*, 34:22301–22312, 2021.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025. doi: 10.1007/s10462-025-11236-4. URL https://doi.org/10.1007/s10462-025-11236-4.

Xi Wang, Laurence Aitchison, and Maja Rudolph. LoRA ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023. doi: 10.48550/arXiv.2310.00035. URL https://arxiv.org/abs/2310.00035.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. BLoB: Bayesian low-rank adaptation by backpropagation for large language models. *Advances in Neural Information Processing Systems*, 37:67758–67794, 2024. doi: 10.52202/079017-2164.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklf1yrYDr.

Kejin Wu and Dimitris N Politis. Scalable subsampling inference for deep neural networks. *ACM/IMS Journal of Data Science*, 2(1):1–29, 2025.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9O6hbiDLU7.

Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle YY Lee, Kyung S Cho, Katalin Susztak, George X Xu, Michael D Feldman, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature biotechnology*, 42(9): 1372–1377, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A THEORETICAL ANALYSIS OF HIFI

A.1 APPROXIMATION AND EXCESS RISK

Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be independent i.i.d. samples of sizes n_1 and n_2 from the distribution of $Z = (X, Y)$, where r is the Stage 2 subsample size. The target is the link-scale regression function θ_0. Let $\mathcal{X}$ be the target domain, with $P_X(\mathcal{X}) = 1$, and hold the pretrained model fixed. Consider the full adaptation class

$$\mathcal{F}_{\mathbf{p}} = \{f(\cdot; w) : w \in \mathcal{W}_{\mathbf{p}}\}.$$

Here $\mathbf{p}$ denotes the adaptation specification, which may encode the number and allocation of trainable parameters, adaptation ranks, adapted layers, or other architectural complexity measures. Write $h = (F, t)$ for a Stage 1 output, where F is the adapted backbone and $t : \mathcal{X} \to \mathbb{R}$ is the corresponding link-scale prediction function. Write $\widehat{h}_1 = (\widehat{F}_1, \widehat{\theta}_1)$ for the fitted Stage 1 output. For fixed h, let $\mathcal{W}_{\mathbf{p}}(h) \subseteq \mathcal{W}_{\mathbf{p}}$ denote the adaptation parameters compatible with the fixed Stage 1 output, and define $\rho_h = \theta_0 - t$. The corresponding Stage 2 residual class is

$$\mathcal{R}_{\mathbf{p}}(h) = \{f(\cdot; w) - t : w \in \mathcal{W}_{\mathbf{p}}(h)\}.$$

Throughout, $\|f\|_2^2 = \int_{\mathcal{X}} |f(x)|^2 \, dP_X(x)$. The best approximation errors of the full and Stage 2 classes are

$$\mathcal{E}_{\mathrm{app}}(\mathbf{p}) := \inf_{w \in \mathcal{W}_{\mathbf{p}}} \|f(\cdot; w) - \theta_0\|_2^2, \qquad \mathcal{E}_{\mathrm{app},2}(\mathbf{p}; h) := \inf_{w \in \mathcal{W}_{\mathbf{p}}(h)} \|f(\cdot; w) - \theta_0\|_2^2.$$

Equivalently, $\mathcal{E}_{\mathrm{app},2}(\mathbf{p}; h) = \inf_{\rho \in \mathcal{R}_{\mathbf{p}}(h)} \|\rho - \rho_h\|_2^2$. Because $\mathcal{W}_{\mathbf{p}}(h) \subseteq \mathcal{W}_{\mathbf{p}}$, the Stage 2 class is a restricted subclass of the full adaptation class, and hence

$$\mathcal{E}_{\mathrm{app}}(\mathbf{p}) \le \mathcal{E}_{\mathrm{app},2}(\mathbf{p}; h) \qquad \text{for every } h.$$

We therefore define the approximation-capacity loss induced by the Stage 1 output as

$$d_1(h) := \mathcal{E}_{\mathrm{app},2}(\mathbf{p}; h) - \mathcal{E}_{\mathrm{app}}(\mathbf{p}) \ge 0.$$

Thus $d_1(h)$ measures how much approximation capacity is lost by fixing the Stage 1 output before Stage 2 fitting.

Let $\mathcal{E}(\mathbf{p}, m)$ denote a deterministic estimation-error bound for a learner with adaptation specification $\mathbf{p}$ fitted on m observations.

Assumption A.1. *There exists a constant $C_1 > 0$ such that*

$$\mathbb{E}\{d_1(\widehat{h}_1)\} \le C_1 \mathcal{E}(\mathbf{p}, n_1).$$

Here the expectation is over the Stage 1 randomness in $\widehat{h}_1$. Assumption A.1 requires the loss of Stage 2 approximation capacity induced by the learned Stage 1 output to be controlled by the Stage 1 estimation error. This is a natural compatibility condition for shared adaptation: when Stage 1 learns a useful shared backbone, fixing and reusing it should preserve most of the approximation capacity available to the full adaptation class.

Let $\widehat{\theta}^{\mathrm{single}}$ denote a generic single-stage fit based on a resample of size $c_0 r$, where $c_0 \ge 1$ is fixed, and let $\widehat{\rho}(\cdot; h)$ denote a generic Stage 2 residual fit based on a resample of size r. For fixed h, let P_h and E_h denote probability and expectation over $\mathcal{D}_2$ and the Stage 2 resampling mechanism, with h held fixed. The HIFI estimator and the corresponding integrated excess risks are

$$\widehat{\theta}^{\mathrm{HIFI}} = \widehat{\theta}_1 + \widehat{\rho}(\cdot; \widehat{h}_1), \qquad R_{\mathrm{single}} = \mathbb{E}\|\widehat{\theta}^{\mathrm{single}} - \theta_0\|_2^2, \qquad R_{\mathrm{HIFI}} = \mathbb{E}\|\widehat{\theta}^{\mathrm{HIFI}} - \theta_0\|_2^2.$$

Assumption A.2. *There exists a constant $C_{\mathrm{learn}} > 0$ such that*

$$R_{\mathrm{single}} \le C_{\mathrm{learn}}\{\mathcal{E}_{\mathrm{app}}(\mathbf{p}) + \mathcal{E}(\mathbf{p}, c_0 r)\}, \tag{4}$$

$$E_h\|\widehat{\rho}(\cdot; h) - \rho_h\|_2^2 \le C_{\mathrm{learn}}\{\mathcal{E}_{\mathrm{app},2}(\mathbf{p}; h) + \mathcal{E}(\mathbf{p}, r)\}, \qquad \textit{for all } h. \tag{5}$$

Motivated by oracle inequalities for neural-network regression (e.g., Schmidt-Hieber, 2020; Meng & Li, 2025), excess risk can be bounded by two sources of error: approximation error and estimation error. We use this structure to compare single-stage resampling with HIFI. For single-stage resampling, these terms are $\mathcal{E}_{\mathrm{app}}(\mathbf{p})$ and $\mathcal{E}(\mathbf{p}, c_0 r)$, respectively; for HIFI, conditional on the Stage 1 output h, they are $\mathcal{E}_{\mathrm{app},2}(\mathbf{p}; h)$ and $\mathcal{E}(\mathbf{p}, r)$.

Proposition A.3. *Under Assumptions A.1 and A.2,*

$$R_{\text{single}} \le C_{\text{learn}}\{\mathcal{E}_{\text{app}}(\mathbf{p}) + \mathcal{E}(\mathbf{p}, c_0 r)\},$$
$$R_{\text{HIFI}} \le C_{\text{learn}}\{\mathcal{E}_{\text{app}}(\mathbf{p}) + C_1\mathcal{E}(\mathbf{p}, n_1) + \mathcal{E}(\mathbf{p}, r)\}. \tag{6}$$

Proof. By the definition of $d_1(h)$,

$$\mathcal{E}_{\text{app},2}(\mathbf{p}; h) = \mathcal{E}_{\text{app}}(\mathbf{p}) + d_1(h).$$

Therefore, Assumption A.1 gives

$$\mathbb{E}\{\mathcal{E}_{\text{app},2}(\mathbf{p}; \widehat{h}_1)\} \le \mathcal{E}_{\text{app}}(\mathbf{p}) + C_1\mathcal{E}(\mathbf{p}, n_1).$$

Since

$$\widehat{\theta}^{\text{HIFI}} - \theta_0 = \widehat{\rho}(\cdot; \widehat{h}_1) - \rho_{\widehat{h}_1},$$

the tower property and Assumption A.2 yield

$$\begin{aligned} R_{\text{HIFI}} &= \mathbb{E}\left[E_{\widehat{h}_1}\left\|\widehat{\rho}(\cdot; \widehat{h}_1) - \rho_{\widehat{h}_1}\right\|_2^2\right] \\ &\le C_{\text{learn}}\left[\mathbb{E}\{\mathcal{E}_{\text{app},2}(\mathbf{p}; \widehat{h}_1)\} + \mathcal{E}(\mathbf{p}, r)\right] \\ &\le C_{\text{learn}}\{\mathcal{E}_{\text{app}}(\mathbf{p}) + C_1\mathcal{E}(\mathbf{p}, n_1) + \mathcal{E}(\mathbf{p}, r)\}. \end{aligned}$$

The single-stage bound follows directly from equation 4. □

The bounds reveal the key distinction between single-stage resampling and HIFI. Single-stage resampling fits the full adaptation class $\mathcal{F}_{\mathbf{p}}$ on each resample, leading to the estimation term $\mathcal{E}(\mathbf{p}, c_0 r)$. HIFI instead pays the Stage 1 estimation cost $\mathcal{E}(\mathbf{p}, n_1)$ once and, conditional on the Stage 1 output, resampling fits only the restricted Stage 2 class, leading to $\mathcal{E}(\mathbf{p}, r)$. Both procedures retain the same full-class approximation term $\mathcal{E}_{\text{app}}(\mathbf{p})$. Therefore, when the two procedures attain risk bounds of the same order, across B resampling fits HIFI incurs the Stage 1 estimation term only once and the Stage 2 estimation term B times, rather than refitting the full adaptation class $\mathcal{F}_{\mathbf{p}}$ B times as in single-stage resampling.

A.2 Inferential validity

Valid inference requires, in addition to the excess-risk bounds, the regularity conditions of the chosen resampling method.

HIFI–ESM validity Let $\mathcal{H}_{n_2}$ be deterministic sets of Stage 1 outputs h for which the conditional Stage 2 DNN–ESM problem for the residual target ρ_h satisfies the conditions in Corollary 3.6 and Theorems 3.7–3.8 of Meng & Li (2025). We further assume that there exist common deterministic target sets $\mathcal{T}_{n_2} \subseteq \mathcal{X}$ satisfying

$$P_X(\mathcal{T}_{n_2}) \ge 1 - \delta_{n_2}, \qquad \delta_{n_2} > 0, \qquad \delta_{n_2} \to 0,$$

on which the resulting conditional normal approximation centered at the true residual target ρ_h and variance consistency hold uniformly over $h \in \mathcal{H}_{n_2}$.

To account for the randomness of the Stage 1 fit, assume that

$$P(\widehat{h}_1 \in \mathcal{H}_{n_2}) \to 1.$$

Under this uniform conditional formulation, Slutsky's theorem gives, for every deterministic sequence $x_{n_2} \in \mathcal{T}_{n_2}$,

$$\sup_{h \in \mathcal{H}_{n_2}} \left|P_h\left\{\theta_0(x_{n_2}) \in C_\alpha^{\theta,\text{ESM}}(x_{n_2}; h)\right\} - (1-\alpha)\right| \to 0. \tag{7}$$

On the event $\widehat{h}_1 \in \mathcal{H}_{n_2}$, the conditional coverage error at $\widehat{h}_1$ is bounded by the supremum in equation 7. Hence,

$$P_{\widehat{h}_1}\left\{\theta_0(x_{n_2}) \in C_\alpha^{\theta,\text{ESM}}(x_{n_2}; \widehat{h}_1)\right\} \to_p 1 - \alpha.$$

Unconditional validity follows from the following lemma.

Lemma A.4. *Suppose that, for a deterministic target sequence x_{n_2} and a sequence of confidence intervals $C_\alpha(x_{n_2}; \widehat{h}_1)$,*

$$P_{\widehat{h}_1}\left\{\theta_0(x_{n_2}) \in C_\alpha(x_{n_2}; \widehat{h}_1)\right\} \to_p 1-\alpha.$$

Then

$$P\left\{\theta_0(x_{n_2}) \in C_\alpha(x_{n_2}; \widehat{h}_1)\right\} \to 1-\alpha.$$

Proof. Let

$$q_{n_2}(\widehat{h}_1) := P_{\widehat{h}_1}\left\{\theta_0(x_{n_2}) \in C_\alpha(x_{n_2}; \widehat{h}_1)\right\}.$$

By assumption, $q_{n_2}(\widehat{h}_1) \to_p 1-\alpha$. Since $0 \le q_{n_2}(\widehat{h}_1) \le 1$, convergence in probability to the constant $1-\alpha$ implies convergence in L_1. Hence,

$$\mathbb{E}\{q_{n_2}(\widehat{h}_1)\} \to 1-\alpha.$$

By the tower property,

$$\mathbb{E}\{q_{n_2}(\widehat{h}_1)\} = P\left\{\theta_0(x_{n_2}) \in C_\alpha(x_{n_2}; \widehat{h}_1)\right\},$$

which proves the result. □

Applying Lemma A.4 gives the HIFI–ESM coverage statement in the main text.

HIFI–Bootstrap validity For bootstrap inference, assume that, for each deterministic admissible Stage 1 output h, the conditional sampling law of the normalized Stage 2 estimation error centered at $\rho_h(x_{n_2})$ and the bootstrap law of the correspondingly normalized error centered at the original Stage 2 estimate converge in probability to the same distribution, with a continuous, strictly increasing cdf and symmetry about zero. Assume that the number of conditionally independent bootstrap replicates tends to infinity. The bootstrap quantile argument of Bücher & Kojadinovic (2019, Section 4, Lemma 4.2), together with the symmetry assumption, then yields asymptotically correct conditional coverage for percentile intervals. Under the analogous uniform conditional coverage requirement over $\mathcal{H}_{n_2}$ and $\mathcal{T}_{n_2}$, evaluating at $\widehat{h}_1$ and applying Lemma A.4 gives unconditional validity.

B Implementation Details

The implementation of HIFI with ESM is described in Section 3.3. Here we provide the corresponding bootstrap construction and implementation details for the external comparators.

B.1 Bootstrap

We use the LoRA construction in Section 3.3, holding the Stage 1 fit fixed. For Stage 2, we independently draw B bootstrap samples, each containing n_2 observations sampled with replacement from $\mathcal{D}_2$. The corresponding weights satisfy

$$(W_{b1}, \ldots, W_{bn_2}) \overset{\text{iid}}{\sim} \text{Multinomial}\left(n_2; \frac{1}{n_2}, \ldots, \frac{1}{n_2}\right), \qquad b = 1, \ldots, B,$$

where W_{bi} counts the occurrences of observation i in sample b. Each fit optimizes new Stage 2 LoRA factors and a prediction head according to equation 2, yielding $\widehat{\rho}_b$.

We average the B residual estimators,

$$\widehat{\rho}_{\text{Boot}}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\rho}_b(x), \qquad \widehat{\theta}_{\text{Boot}}(x) = \widehat{\theta}_1(x) + \widehat{\rho}_{\text{Boot}}(x).$$

Let $\widehat{Q}^{\rho}_p(x)$ denote the empirical p-quantile of $\{\widehat{\rho}_b(x)\}_{b=1}^B$. The pointwise percentile interval for $\theta_0(x)$ is

$$C^{\theta,\text{Boot}}_\alpha(x) = \widehat{\theta}_1(x) + \left[\widehat{Q}^{\rho}_{\alpha/2}(x), \widehat{Q}^{\rho}_{1-\alpha/2}(x)\right].$$

The point estimate and interval are transformed to the response scale through g^{-1}.

B.2 External Comparators

Bayesian-PEFT We adapt the Bayesian neural-network framework of Sun et al. (2021) to parameter-efficient fine-tuning. The pretrained backbone is held fixed, and approximate posterior sampling is performed over the LoRA parameters and prediction head using stochastic-gradient Hamiltonian Monte Carlo (Chen et al., 2014). The model is fitted directly to the observed responses. Text experiments use a Bernoulli likelihood, whereas vision experiments use a Gaussian likelihood with an estimated global noise scale. For each target, the point estimate is the mean of the retained conditional-mean predictions, and empirical equal-tailed quantiles form approximate posterior credible intervals. Model-specific sampling and prior settings are reported in Tables 5 and 7.

HulC We implement HulC (Kuchibhotla et al., 2024) as a direct comparator. The labeled data are partitioned into disjoint subsets, and an independent LoRA estimator is fitted to each subset from the pretrained backbone. For a fixed target x, let $\widehat{\theta}_b^{\mathrm{HulC}}(x)$, $b = 1, \ldots, B_\alpha$, denote the split-specific link-scale estimates. The link-scale interval is

$$C_\alpha^{\theta,\mathrm{HulC}}(x) = \left[\min_{1 \le b \le B_\alpha} \widehat{\theta}_b^{\mathrm{HulC}}(x), \max_{1 \le b \le B_\alpha} \widehat{\theta}_b^{\mathrm{HulC}}(x) \right],$$

and the response-scale interval is obtained through g^{-1}. We use the median of the split-specific estimates as the point estimate and the $\Delta = 0$ construction, with the number of disjoint fits chosen for the desired nominal confidence level.

C Synthetic Experiment Details

C.1 Synthetic Vision Data

We provide additional data-generation details for the synthetic image-regression task in Section 5.1. Each object is defined on a unit-square canvas, and a denotes its geometric area relative to the canvas. Triangles are equilateral, while squares and triangles are rotated by an angle sampled uniformly from $[0, 2\pi)$. Conditional on shape, area, and orientation, the object center is sampled uniformly over locations that keep the entire object inside the canvas. Foreground and background colors are sampled uniformly from the 12 ordered distinct pairs formed by red $(255, 0, 0)$, green $(0, 255, 0)$, yellow $(255, 255, 0)$, and blue $(0, 0, 255)$, independently of shape and area. For each fixed target image x_j, evaluation uses the conditional-mean target $\beta(S_j)a_j$. A panel of 500 target images is generated once and held fixed across all 100 replications. Representative examples are shown in Figure 5.

C.2 Bilingual Text Data Generation

We provide the data generation details for the bilingual simulation described in Section 5.2.

Probability-stratified sampling. The binary outcome probability is determined by the arrival lead time through

$$\mu(T) = \mathrm{logit}^{-1}\{2.8(T - 1.05)\}, \qquad T \in [0, 3.5].$$

Thus, the attainable probability range is

$$[\mu_{\min}, \mu_{\max}] = [\mu(0), \mu(3.5)] = [\mathrm{logit}^{-1}(-2.94), \mathrm{logit}^{-1}(6.86)].$$

To place equal sampling probability across regions of the success-probability range and avoid concentrating observations in only part of this range, we partition $[\mu_{\min}, \mu_{\max}]$ into 20 equal-width probability strata. For each observation, we independently draw a stratum index

$$J \sim \mathrm{Uniform}\{1, \ldots, 20\}.$$

Let

$$a_j = \mu_{\min} + \frac{j}{20}(\mu_{\max} - \mu_{\min}), \qquad j = 0, \ldots, 20,$$

denote the stratum boundaries, so that stratum j corresponds to $(a_{j-1}, a_j]$. Conditional on $J = j$, we sample

$$p^\star \sim \mathrm{Uniform}(a_{j-1}, a_j),$$

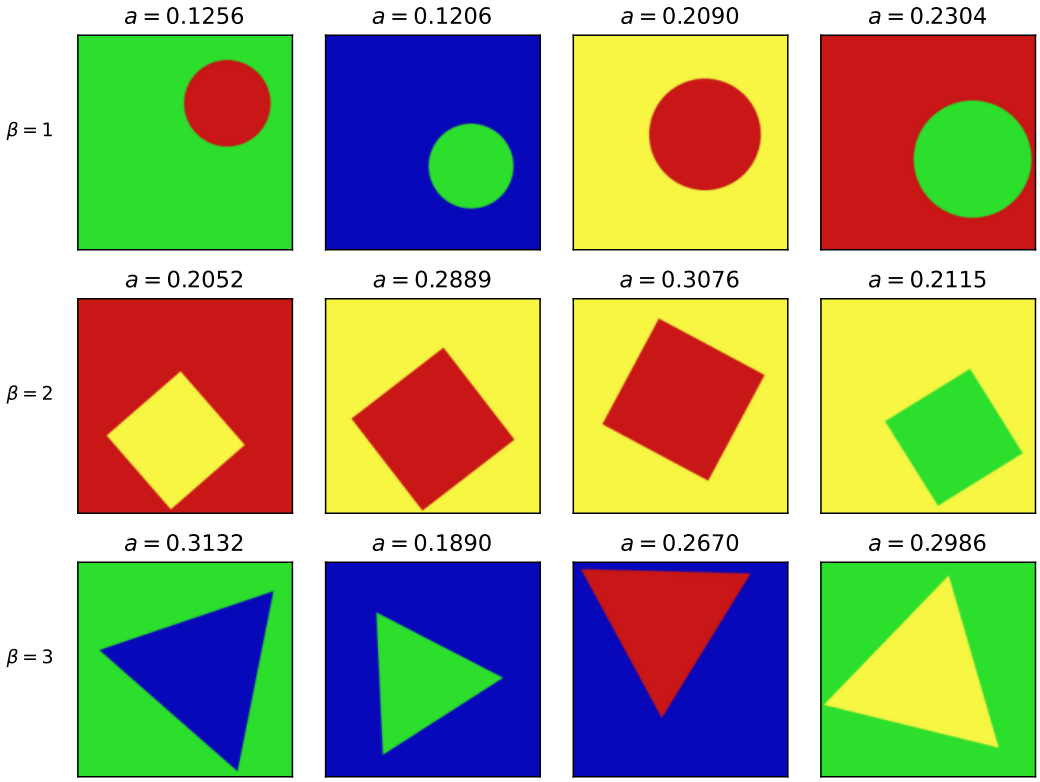


Figure 5: Examples of the synthetic vision data. Rows correspond to circles, squares, and equilateral triangles, respectively. Foreground and background colors, object positions, and polygon orientations vary across images.

and invert the probability model to obtain the corresponding arrival lead time, rounded to the nearest minute:

$$T = \frac{\text{round}\left\{60\left[1.05 + \frac{\text{logit}(p^\star)}{2.8}\right]\right\}}{60}.$$

Text construction. For each observation, independently of T, the language is drawn as English or Chinese with equal probability. Conditional on T and the sampled language, we generate a natural-language description of when the passenger arrives relative to the scheduled departure time. Additional textual variation is introduced through auxiliary flight information, such as the airline, flight number, gate, terminal, seat, or boarding group, with probability 0.50, and through response-irrelevant contextual information with probability 0.20. These components are generated independently of the arrival lead time. The sampled lead time is expressed in the resulting text input X; by construction, $T(X) = T$.

Response generation. For the resulting text X, we set

$$\mu_0(X) = \text{logit}^{-1}\{2.8[T(X) - 1.05]\},$$

and independently draw

$$Y \mid X \sim \text{Bernoulli}\{\mu_0(X)\},$$

where $Y = 1$ indicates that the passenger catches the flight and $Y = 0$ otherwise. Table 3 provides representative realizations from the bilingual DGP, showing the generated text X, the encoded arrival lead time $T(X)$, the corresponding conditional mean $\mu_0(X)$, and the sampled outcome Y. The model is trained only on the observed pairs (X, Y) and is not given $T(X)$ or $\mu_0(X)$.

Table 3: Illustrative observations from the bilingual text DGP at representative arrival lead times. English translations of the Chinese inputs are included only for readability.

Language	Model input X	$T(X)$ (min)	$\mu_0(X)$	Y
English	Kate walked into the terminal 15 minutes before the scheduled departure.	15	0.0962	0
English	Zack arrived at the airport 49 minutes before his flight.	49	0.3422	1
English	His flight DL8517 was assigned to gate B17. Henry reached the gate 1 hour and 16 minutes ahead of takeoff.	76	0.6472	0
English	American listed the flight as AA2789 from Terminal A. Frank reached the airport with 1 hour and 50 minutes left until the scheduled flight.	110	0.8996	1
Chinese	航班预定起飞前16分钟，周强已经抵达机场。 *English translation: Zhou Qiang had already arrived at the airport 16 minutes before the scheduled departure.*	16	0.1004	0
Chinese	航站楼里人特别多。 她的航班是长荣BR6967。 距起飞还剩50分钟时，林雪走进了候机厅。 *English translation: The terminal was especially crowded. Her flight was EVA Air BR6967. With 50 minutes remaining before takeoff, Lin Xue entered the departure lounge.*	50	0.3528	0
Chinese	他的航班是东航MU5511。 航班预定起飞前1小时16分钟，郑磊已经抵达机场。 *English translation: His flight was China Eastern MU5511. Zheng Lei had already arrived at the airport 1 hour and 16 minutes before the scheduled departure.*	76	0.6472	1
Chinese	她的航班从一号航站楼出发。 查完行程后，吴丽在飞机起飞前1小时50分钟到达了机场。 *English translation: Her flight departed from Terminal 1. After checking the itinerary, Wu Li arrived at the airport 1 hour and 50 minutes before takeoff.*	110	0.8996	1

D Model Architecture and Training Settings

Across both modalities, the learned Stage 1 LoRA updates are merged into the backbone and frozen before Stage 2, which uses fresh LoRA adapters and a residual prediction head with a zero-initialized output layer. The remaining architecture, adaptation, optimization, and inference settings are reported below.

D.1 Vision Experiment

HIFI settings are summarized in Table 4, and settings for HulC and Bayesian-PEFT are reported in Table 5.

D.2 Text Experiment

HIFI settings are summarized in Table 6, and settings for HulC and Bayesian-PEFT are reported in Table 7.

D.3 Ablation Study 1

We use Qwen3-Embedding-8B with the architecture, LoRA configuration, and optimization hyperparameters reported in Table 6. We vary the total Stage 1 data budget over $\{2{,}000, 4{,}500, 8{,}500\}$, maintaining the same training-to-validation ratio as in the main experiment, while keeping the Stage 2 architecture and data budget fixed. The single-stage reference uses the same member architecture and optimization settings as HIFI Stage 2, but fits the response directly from the pretrained backbone with a standard randomly initialized prediction head.

D.4 Ablation Study 2

The HIFI configurations follow Tables 4 and 6 for the corresponding backbones. Single-stage learners use the corresponding HIFI Stage 2 architecture, LoRA configuration, and optimization

Table 4: Architecture, adaptation, and training settings for the vision experiments.

Setting	ViT-B/32	DINOv2-B/14	Swin-B
Pooling	CLS + mean patch (concatenated)	CLS + mean patch (concatenated)	Global average
Prediction head	Two-layer MLP; 64 hidden units; GELU		
Training loss	Mean squared error		
Stage 1			
Training / validation sample size	8,000 / 500		
Blocks with LoRA	All 12	Last 4	All 24
LoRA projections / dropout	Q, K, V, and O; dropout 0.05		
LoRA rank / scaling	Rank 12; scaling 24	Rank 4; scaling 8	Rank 12; scaling 24
Learning rate	5×10^{-5}	4×10^{-4}	5×10^{-5}
Training settings	AdamW; batch size 16; weight decay 10^{-4}; at most 80 epochs		
Stage 2			
Training / validation sample size	1,000 / 200		
Blocks with LoRA	Last 6	Last 2	Final stage (2 blocks)
LoRA configuration	Q and V projections; rank 8; scaling 16; dropout 0.05		
Learning rate	10^{-4}	5×10^{-5}	10^{-4}
Training settings	AdamW; batch size 8 per member; weight decay 10^{-4}		
Maximum epochs	40	25	25
Stage-2 inference			
ESM subsamples / size	500 / 501; without replacement		
Bootstrap resamples / size	400 / 1,000; with replacement		

hyperparameters. Each Single-stage learner uses a training pool of 9,000 observations and a validation pool of 700 observations. For single-stage ESM, we draw 500 subsamples of size 3,621 without replacement from the 9,000 training observations. For single-stage bootstrap, we draw 400 resamples of size 9000 with replacement from the same training pool. HIFI allocates these observations to Stage 1 (8,000 training and 500 validation) and Stage 2 (1,000 training and 200 validation).

E Additional Synthetic Experiment Results

E.1 Target-level Results for the Vision Task

Figure 6 reports empirical 95% coverage across individual evaluation targets, while Figure 7 summarizes the corresponding distribution of mean 95% interval lengths. HIFI-ESM and HIFI-Bootstrap exhibit stable target-wise calibration across all three vision backbones, with coverage concentrated near the nominal level for most targets and only a limited number showing noticeable under-coverage. In contrast, Bayesian-PEFT exhibits pronounced under-coverage for many targets under ViT-B/32, while HulC shows heavier lower tails in several settings. In terms of interval length, HIFI-Bootstrap generally produces shorter intervals than HIFI-ESM, while both methods maintain relatively stable interval lengths across targets. These results show that HIFI provides well-calibrated uncertainty quantification across evaluation targets while maintaining competitive interval efficiency.

For visualization, 50 targets are selected from the full target panel by stratified random sampling across shape classes and true-mean bins. Figure 8 shows the corresponding mean point estimates and mean endpoints of the 95% confidence intervals across replications. Across all three backbones, the HIFI estimators closely track the identity line over the range of true conditional means. The external baselines exhibit noticeable target-specific bias, with mean estimates for some targets deviating substantially from the identity line.

Table 5: Settings for the external vision comparators.

Setting	ViT-B/32	DINOv2-B/14	Swin-B
HulC			
Training / validation size		9,000 / 700	
Adapted blocks	Last 6	Last 2	Final stage (2 blocks)
LoRA configuration		Q and V; rank 8; scaling 16; dropout 0.05	
Learning rate	10^{-4}	5×10^{-5}	10^{-4}
Training		AdamW; batch size 8; weight decay 10^{-4}	
Maximum epochs	40	25	25
Disjoint fits (90% / 95%)		5 / 6	
Bayesian-PEFT			
Training size		9,700	
Adapted blocks	All 12	All 12	All 24
LoRA configuration		Q, K, V, and O; rank 12; scaling 24	
SGHMC learning rate	10^{-4}	5×10^{-5}	5×10^{-5}
SGHMC sampling		Batch size 8; momentum 0.9	
Posterior samples		600 burn-in updates; thinning 30; 300 draws	
Weight prior		Zero-mean Gaussian spike-and-slab mixture	
Spike / slab variance		10^{-6} / 10^{-2}	
Slab probability		10^{-7}	
Likelihood		Gaussian; observation SD σ estimated from data	
Observation-SD prior		$\sigma \sim \mathrm{HalfNormal}(1)$	

Notes. Both comparators use the backbone-specific pooling and prediction head reported in Table 4.

Table 6: Architecture, adaptation, and training settings for the text experiments.

Setting	Qwen3-Embedding-8B	XLM-R XXL	NV-Embed-v2
Pooling	Last token	CLS token	Native latent attention
Maximum sequence length	96 tokens	96 tokens	112 tokens
Input prompt	`Instruct: Given a passenger flight-status description, estimate the probability that the passenger makes the flight.` `Query: <text>`		
Prediction head		Two-layer MLP; 128 hidden units; GELU	
Training loss		Binary cross-entropy with logits	
LoRA configuration		Q and V projections; rank 16; scaling 32; dropout 0.05	
Stage 1			
Training / validation sample size		8,000 / 500	
Blocks with LoRA	All 36	All 48	All 32
Learning rate	7.5×10^{-5}	5×10^{-5}	3×10^{-5}
Training settings		AdamW; batch size 8; weight decay 10^{-4}; at most 20 epochs	
Stage 2			
Training / validation sample size		1,000 / 200	
Blocks with LoRA	Last 6	Last 4	Last 4
Learning rate	4×10^{-5}	10^{-4}	3×10^{-4}
Training settings		AdamW; batch size 8 per member; weight decay 0; at most 25 epochs	
Stage-2 inference			
ESM subsamples / size		500 / 501; without replacement	
Bootstrap resamples / size		400 / 1,000; with replacement	

Table 7: Settings for the external text comparators.

Setting	Qwen3-Embedding-8B	XLM-R XXL	NV-Embed-v2
HulC			
Training / validation size		9,000 / 700	
Adapted blocks	Last 6	Last 4	Last 4
LoRA configuration		Q and V; rank 16; scaling 32; dropout 0.05	
Learning rate	4×10^{-5}	2×10^{-4}	3×10^{-4}
Training		AdamW; batch size 8; weight decay 0	
Maximum epochs		25	
Disjoint fits (90% / 95%)		5 / 6	
Bayesian-PEFT			
Training size		9,700	
Adapted blocks	All 36	Last 12	Last 12
LoRA configuration		Q and V; rank 16; scaling 32	
Likelihood		Bernoulli with a logit link	
SGHMC learning rate	5×10^{-5}	5×10^{-3}	5×10^{-4}
SGHMC sampling		Batch size 8; momentum 0.9; temperature 1	
Burn-in updates	600	1,200	600
Posterior samples		300 draws; thinning interval 30	
Parameter prior	Gaussian spike-and-slab mixture	$\mathcal{N}(0,1)$	$\mathcal{N}(0,10^{-2})$: LoRA and head weights; $\mathcal{N}(0,1)$: head biases

Notes. Both comparators use the pooling, input prompt, sequence-length limit, and prediction-head architecture in Table 6.

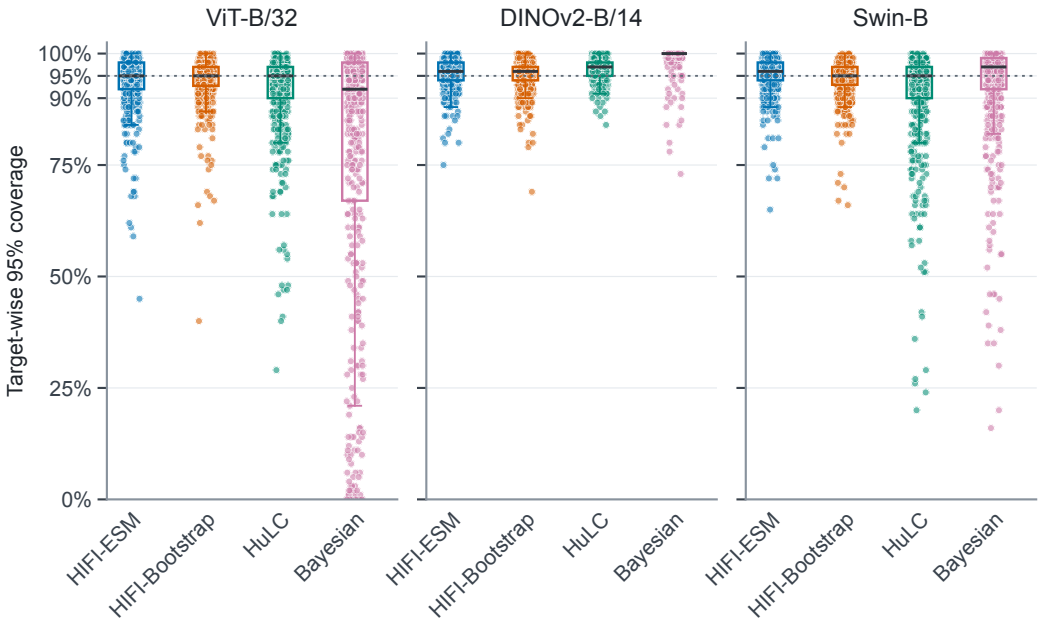


Figure 6: Target-wise empirical coverage of nominal 95% pointwise confidence intervals for the vision regression task. Each point represents one fixed evaluation target and gives the proportion of replications in which the corresponding interval contains $\mu_0(x)$. The dashed horizontal line marks the nominal 95% coverage level.

E.2 Target-level Results for the Text Task

Figure 9 reports empirical 95% coverage across individual evaluation targets, while Figure 10 summarizes the corresponding distribution of mean 95% interval lengths. HIFI-ESM and HIFI-

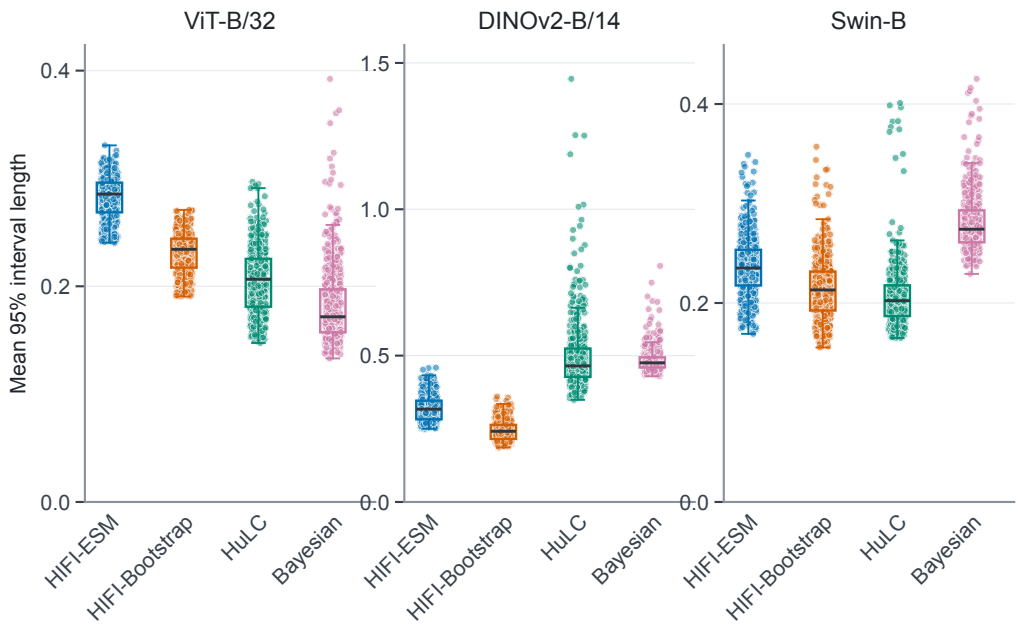


Figure 7: Target-wise mean lengths of nominal 95% pointwise confidence intervals for the vision regression task. Each point represents one evaluation target and reports the confidence-interval length averaged across replications. Boxplots summarize the distribution of mean interval lengths across targets.

Bootstrap exhibit relatively stable target-wise calibration across all three language backbones, with coverage concentrated near the nominal level for most targets. HulC shows under-coverage for a subset of targets, whereas Bayesian-PEFT exhibits near-complete coverage under XLM-R XXL but pronounced under-coverage for many targets under Qwen3-Embedding-8B and especially NV-Embed-v2. In terms of interval length, HIFI-Bootstrap generally produces shorter intervals than HIFI-ESM. Notably, the shorter intervals produced by Bayesian-PEFT under NV-Embed-v2 coincide with substantial target-wise under-coverage. These results show that HIFI provides well-calibrated uncertainty quantification across evaluation targets.

For visualization, 50 targets are selected from the evaluation panel by stratified random sampling across languages and true-probability bins. Figure 11 shows the corresponding mean point estimates and mean lower and upper endpoints of the 95% intervals across replications. Across all three backbones, the HIFI estimators closely track the identity line over the range of true conditional probabilities. The external baselines show more pronounced target-specific bias, with HulC deviating toward the center of the probability range for some targets and Bayesian-PEFT exhibiting particularly strong systematic deviations under NV-Embed-v2.

E.3 Ablation Study 1

We provide complementary results on the effect of Stage 1 task-adaptation. Figure 12 shows the same qualitative pattern at the 90% confidence level. Holding the Stage 2 inference sample fixed, increasing n_1 moves empirical coverage toward the nominal 90% level while reducing mean interval length for both ESM and bootstrap. The corresponding single-stage baselines attain conservative coverage, but with substantially larger intervals.

Figure 13 shows similar target-wise patterns. For ESM and bootstrap, HIFI coverage is below 95% at smaller n_1 and moves closer to the nominal level at $n_1 = 8{,}500$. The interval lengths decrease across most targets as n_1 increases, while the single-stage baselines produce substantially wider intervals.

Figure 14 further compares point estimates and mean 95% interval endpoints for 50 representative targets. The single-stage fits show larger deviations from the identity line and much wider intervals.

Under HIFI, the point estimates track the truth more closely as n_1 increases, while the corresponding intervals become progressively narrower. These target-level results are consistent with the aggregate findings in Figure 2.

E.4 Ablation Study 2

Figure 15 reports the corresponding results at the 90% confidence level. Consistent with the 95% results in Figure 3, direct single-stage resampling remains strongly conservative across backbones and methods, whereas HIFI produces coverage closer to the nominal level with substantially shorter intervals in most settings.

Figures 16 and 17 provide target-level results at the 95% confidence level. The single-stage procedures exhibit coverage concentrated near 100% across most targets, while HIFI coverage is concentrated closer to the nominal 95% level. HIFI produces substantially shorter intervals for ESM across all four backbones. For bootstrap, HIFI yields shorter intervals for DINOv2-B/14 and Qwen3-Embedding-8B, with comparable interval lengths for ViT-B/32 and Swin-B.

E.5 Computational Cost

Table 8 reports training EFLOPs and wall-clock time per replication for HIFI and the single-stage configurations. Vision experiments were run on A40 or RTX 3090 GPUs, while the language-model experiments used A100 80GB or B200 GPUs. All reported experiments use a single GPU.

Training EFLOPs are estimated analytically from the model architecture and executed workload, counting one multiply–add as two FLOPs. The calculation includes forward computation through the backbone, LoRA adapters, and prediction heads; backward computation for activation and trainable-parameter gradients; loss computation; and AdamW updates. For HIFI, the total includes the Stage 1 training cost once together with the cost of all Stage 2 ensemble or bootstrap members. For the single-stage configurations, it includes training all ensemble or bootstrap members directly on the responses using the combined training sample. We report mean training EFLOPs per replication.

F Additional Real-Data Analysis Details and Results

We provide additional details on the Xenium data analysis.

F.1 Data preprocessing and splitting

We use two consecutive sections from the same breast cancer specimen in the Xenium dataset of Janesick et al. (2023), following the cross-section prediction setting of iStar (Zhang et al., 2024). Section 1 (GSM7780153) provides the source data for model fitting, whereas Section 2 (GSM7780154) is held out entirely for evaluation. We retain transcripts with quality scores of at least 20 and use all 313 genes.

Following the iStar preprocessing pipeline, the registered H&E images are processed at $0.5\,\mu\mathrm{m}$ per image pixel. The native prediction grid consists of $8 \times 8\,\mu\mathrm{m}$ superpixels. Supervision is provided through pseudo-Visium spots with diameter $55\,\mu\mathrm{m}$ and center-to-center spacing $100\,\mu\mathrm{m}$, each associated with 37 native superpixels.

The 4,455 source-section spots are partitioned into spatial blocks, and the blocks are randomly assigned to the Stage 1 and Stage 2 training and validation sets. Predictions are produced natively at $8\,\mu\mathrm{m}$ resolution and aggregated to $16\,\mu\mathrm{m}$ blocks for reporting. Training and resampling operate at the pseudo-Visium-spot level. Table 9 summarizes the principal settings.

F.2 Model implementation and inference

Model implementation HIFI uses the pretrained hierarchical vision transformer backbone of iStar. Each superpixel combines pretrained HViT features with mean RGB intensities and spatial coordinates. RGB intensities are divided by 255, and coordinates are scaled to $[-1, 1]$ using each section's tissue bounding box.

Table 8: Training compute and wall-clock time for HIFI and single-stage configurations across tasks, backbones, and inference methods.

Backbone	Method	Procedure	GPU	Training EFLOPs/rep	Wall-clock time/rep
ViT-B/32	ESM	Single-stage	A40	0.2934	2 h 44 min
	ESM	HIFI	RTX 3090	0.0239	**15 min**
	Bootstrap	Single-stage	A40	0.4685	2 h 46 min
	Bootstrap	HIFI	RTX 3090	0.0668	**32 min**
DINOv2-B/14	ESM	Single-stage	A40	0.3174	1 h 48 min
	ESM	HIFI	RTX 3090	0.0649	**37 min**
	Bootstrap	Single-stage	A40	1.0715	6 h 04 min
	Bootstrap	HIFI	RTX 3090	0.1720	**1 h 01 min**
Swin-B	ESM	Single-stage	A40	0.2239	1 h 03 min
	ESM	HIFI	RTX 3090	0.0312	**24 min**
	Bootstrap	Single-stage	A40	0.4497	2 h 20 min
	Bootstrap	HIFI	RTX 3090	0.0604	**34 min**
Qwen3-Embedding-8B	ESM	Single-stage	A100 80GB	5.9344	7 h 37 min
	ESM	HIFI	A100 80GB	0.9394	**1 h 21 min**
	Bootstrap	Single-stage	A100 80GB	10.1817	13 h 00 min
	Bootstrap	HIFI	A100 80GB	1.3349	**1 h 51 min**
XLM-R XXL	ESM	HIFI	A100 80GB	1.4369	**2 h 00 min**
	Bootstrap	HIFI	A100 80GB	2.8516	**3 h 47 min**
NV-Embed-v2	ESM	HIFI	B200	1.8604	**55 min**
	Bootstrap	HIFI	B200	6.4052	**2 h 59 min**

Note. All experiments were run on a single GPU. Wall-clock time and EFLOPs are reported per replication.

Following iStar, the model is trained through weak supervision at the pseudo-Visium-spot level. For each gene g, let a_g and s_g denote the minimum and range, respectively, of its expression counts across the Stage 1 training spots. These quantities are used to normalize the targets in both stages and are kept fixed throughout model fitting and inference. Let $\widetilde{Y}_s \in \mathbb{R}^{313}$ denote the resulting normalized expression vector for spot s.

Let $\mathcal{M}_s$ denote the set of 37 native superpixels associated with spot s, and let $f_\theta(X_u) \in \mathbb{R}^{313}$ denote the normalized gene-expression prediction for superpixel u. The spot-level prediction is $\bar{f}_\theta(s) = |\mathcal{M}_s|^{-1} \sum_{u \in \mathcal{M}_s} f_\theta(X_u)$, and model fitting minimizes $\mathcal{L}(\theta) = (313|\mathcal{S}|)^{-1} \sum_{s \in \mathcal{S}} \|\widetilde{Y}_s - \bar{f}_\theta(s)\|_2^2$. Thus, spot-level measurements provide weak supervision while the model retains predictions on the finer superpixel grid.

We apply HIFI using the two-stage fitting and inference procedure, with model fitting and resampling performed at the pseudo-Visium-spot level. Model architecture, PEFT, optimization, and other implementation settings are summarized in Table 9.

Statistical inference Let $p_{ug}^{(b)}$ denote the final normalized prediction of member b for gene g at native superpixel u. Using the Stage 1 normalization parameters defined above, we convert this prediction to a native count contribution as $c_{ug}^{(b)} = (s_g p_{ug}^{(b)} + a_g)/37$, where the division by 37 accounts for the 37 superpixels contributing to each pseudo-Visium spot.

For each nonoverlapping 16 μm block j, let $\mathcal{A}_j$ contain its valid native superpixels. We first aggregate within each member and then compute the ensemble point estimate:

$$T_{jg}^{(b)} = \sum_{u \in \mathcal{A}_j} c_{ug}^{(b)}, \qquad \widehat{\mu}_{jg} = \frac{1}{B} \sum_{b=1}^{B} T_{jg}^{(b)}, \qquad B = 500. \tag{8}$$

These operations are performed separately for ESM and bootstrap. For ESM, we apply the ESM variance estimator to the aggregated member predictions and their source-spot inclusion indicators.

Table 9: HIFI settings for the Xenium experiment. Batch sizes and resampling sample sizes count pseudo-Visium spots.

Shared setting	Specification	
Input features	192 HViT-4K + 384 HViT-256 + 3 RGB + 2 coordinates	
MLP layer dimensions	$581 \to 256 \to 256 \to 256 \to 256 \to 313$	
Hidden activation	Leaky ReLU with negative slope 0.1	
Trainable MLP layers	All hidden layers and the output layer	
LoRA targets	QKV and attention output projections	
LoRA parameters	Rank 8; scaling 16; dropout 0	
Optimizer	Adam; learning rate 10^{-4}; weight decay 0	
Training batch size	8	
Stage-specific setting	**Stage 1**	**Stage 2**
Training/validation spots	2,495 / 624	1,069 / 267
HViT-256 adaptation	LoRA in all 12 blocks	Additional member-specific LoRA in the last 3 blocks
MLP initialization	Random initialization	Initialized from Stage 1 hidden-layer weights; final layer initialized to zero
MLP output activation	$\mathrm{ELU}_{0.01}(t) + 0.01$	Identity
Stage 2 resampling		
ESM	500 subsets of 532 out of 1,069 spots; without replacement	
Bootstrap	500 samples of 1,069 spots; with replacement	

Writing the resulting variance estimate as $\widehat{V}_{jg}^{\mathrm{ESM}}$, the nominal 95% interval is

$$\left[\widehat{\mu}_{jg} - z_{0.975}\sqrt{\widehat{V}_{jg}^{\mathrm{ESM}}},\ \widehat{\mu}_{jg} + z_{0.975}\sqrt{\widehat{V}_{jg}^{\mathrm{ESM}}}\right]. \tag{9}$$

For bootstrap, the interval endpoints are the empirical 2.5th and 97.5th percentiles of $\{T_{jg}^{(b)}\}_{b=1}^{B}$.

F.3 Evaluation

We rerun the official iStar implementation on all 4,455 source-section spots using its released image features. Each run uses the official median aggregation of five model states, each trained for 400 epochs. We perform five runs with different optimization randomness. The main figure 4 displays the spatial prediction from the first run.

All methods are evaluated on the same set of 107,052 registered $16\,\mu\mathrm{m}$ blocks containing at least one valid native superpixel. Observed and predicted expression are evaluated as block-total counts.

We evaluate point-prediction performance using root mean squared error (RMSE), Pearson correlation coefficient (PCC), and structural similarity index (SSIM). RMSE and PCC are computed directly from the observed and predicted block-total count maps. SSIM follows the released iStar implementation, using uniform 7×7 windows, sample covariance, and only windows fully contained in the registered support. For each gene, the SSIM dynamic range is defined by the observed count range and held fixed across all methods and iStar runs.

For each gene and metric, the iStar reference is the average over the five runs, $M_g^{\mathrm{iStar}} = 5^{-1}\sum_{r=1}^{5} M_g^{(r)}$. The scatter plots contain one point per gene and compare these reference values with the corresponding HIFI scores.

F.4 Additional results and interpretation

Figure 18 compares bootstrap with iStar across all 313 genes and shows comparable point-prediction performance.

Figure 19 presents additional genes with their ESM confidence interval maps, while Figure 20 provides three additional visual examples of bootstrap confidence interval maps. These spatial maps illustrate the uncertainty quantification provided by HIFI on a real scientific prediction task.

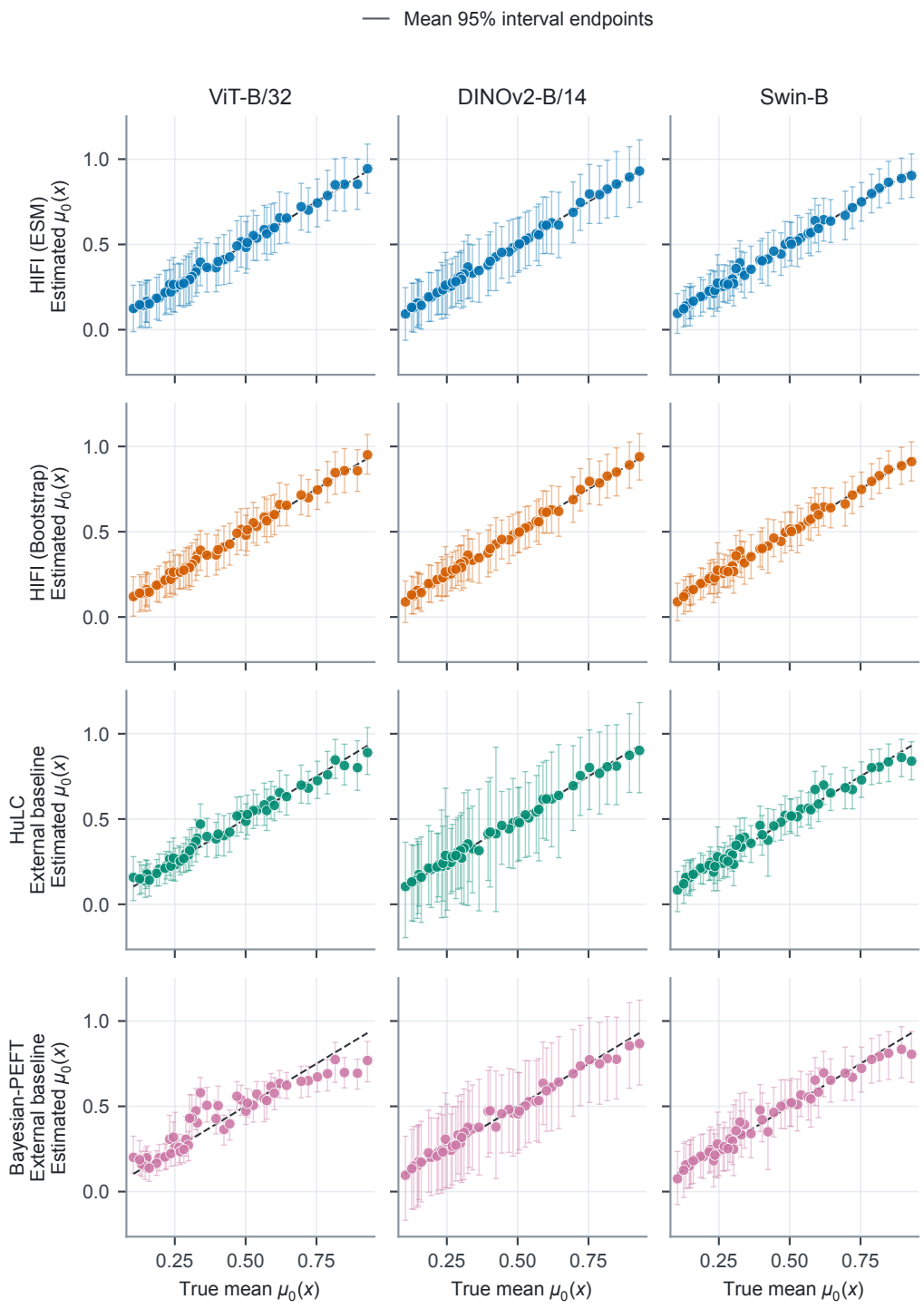


Figure 8: For visualization, 50 targets are selected from the full target panel by stratified random sampling across shape classes and true-mean bins. Columns correspond to the three vision backbones, and rows correspond to HIFI-ESM, HIFI-Bootstrap, HulC, and Bayesian-PEFT. Points show the mean estimated conditional mean for each target, and vertical segments show the mean lower and upper endpoints of the corresponding 95% intervals. The dashed diagonal line represents perfect agreement with the true conditional mean $\mu_0(x)$.

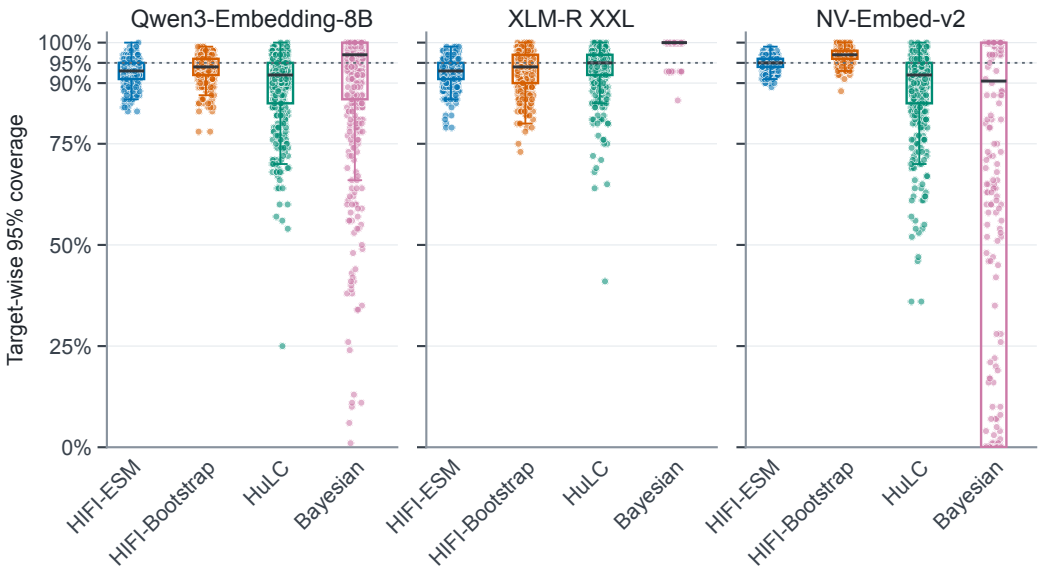


Figure 9: Target-wise empirical coverage of nominal 95% pointwise confidence intervals for the text probability-prediction task. Panels correspond to the three language backbones, and within each panel results are shown for HIFI-ESM, HIFI-Bootstrap, HulC, and Bayesian-PEFT. Each point represents one fixed evaluation target and gives the proportion of replications in which the corresponding interval contains the true conditional probability $\mu_0(x)$. Boxplots summarize the distribution across targets, and the dashed horizontal line marks the nominal 95% coverage level.

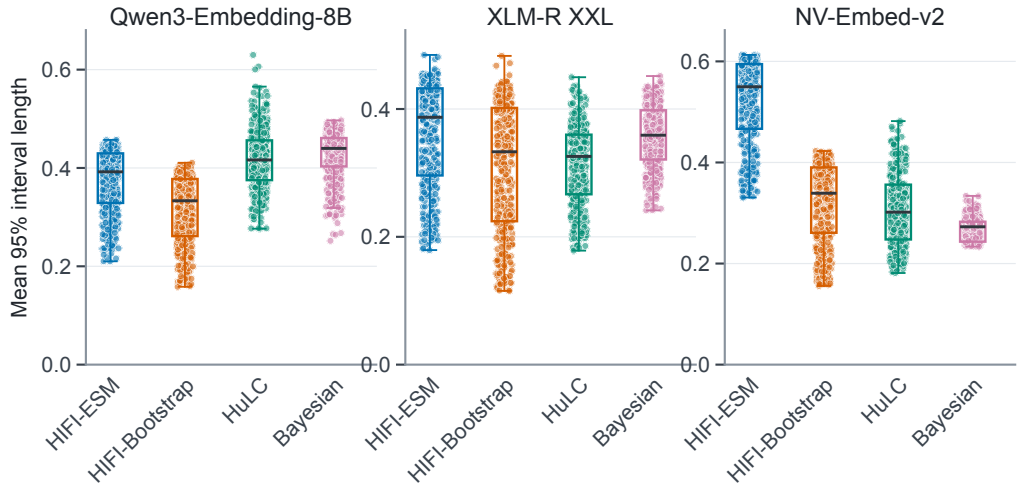


Figure 10: Target-wise mean lengths of nominal 95% pointwise confidence intervals for the text probability-prediction task. Each point represents one evaluation target and reports the confidence-interval length averaged across 100 replications. Boxplots summarize the distribution of mean interval lengths across targets.

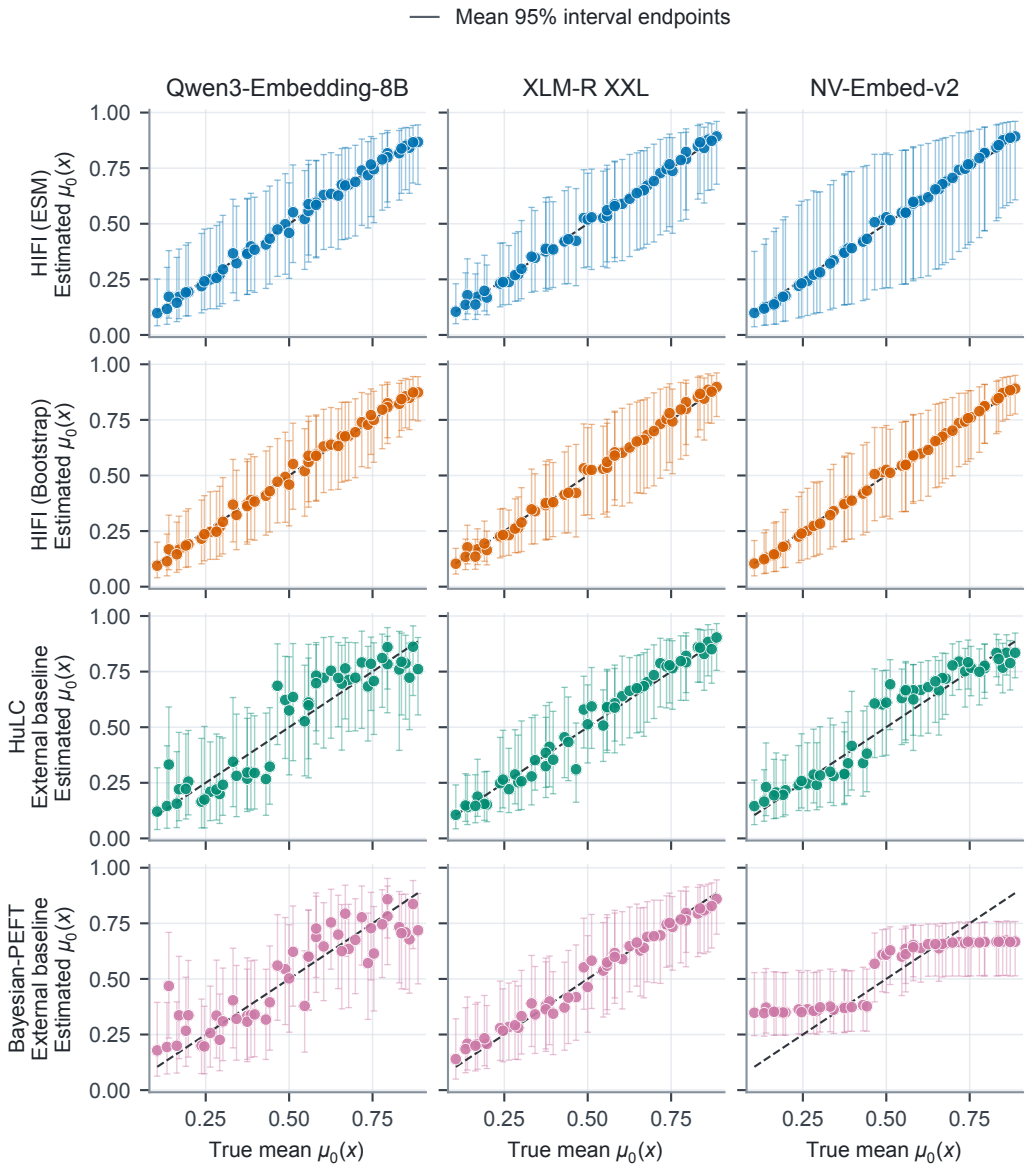


Figure 11: Target-level probability estimation and uncertainty for 50 evaluation targets selected by stratified random sampling across languages and true-probability bins. Columns correspond to the three language backbones, and rows correspond to HIFI-ESM, HIFI-Bootstrap, HulC, and Bayesian-PEFT. Points show the estimated conditional probability averaged across replications for each target, and vertical segments show the corresponding mean lower and upper endpoints of the 95% intervals. The dashed diagonal line represents perfect agreement with the true conditional probability $\mu_0(x)$.

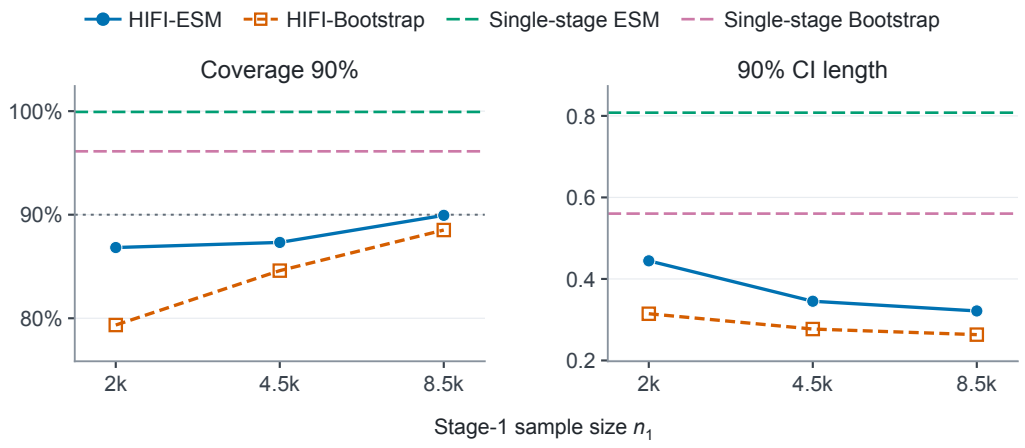


Figure 12: Effect of Stage 1 task-adaptation sample size on 90% coverage and CI length. The dotted line marks the nominal 90% coverage level.

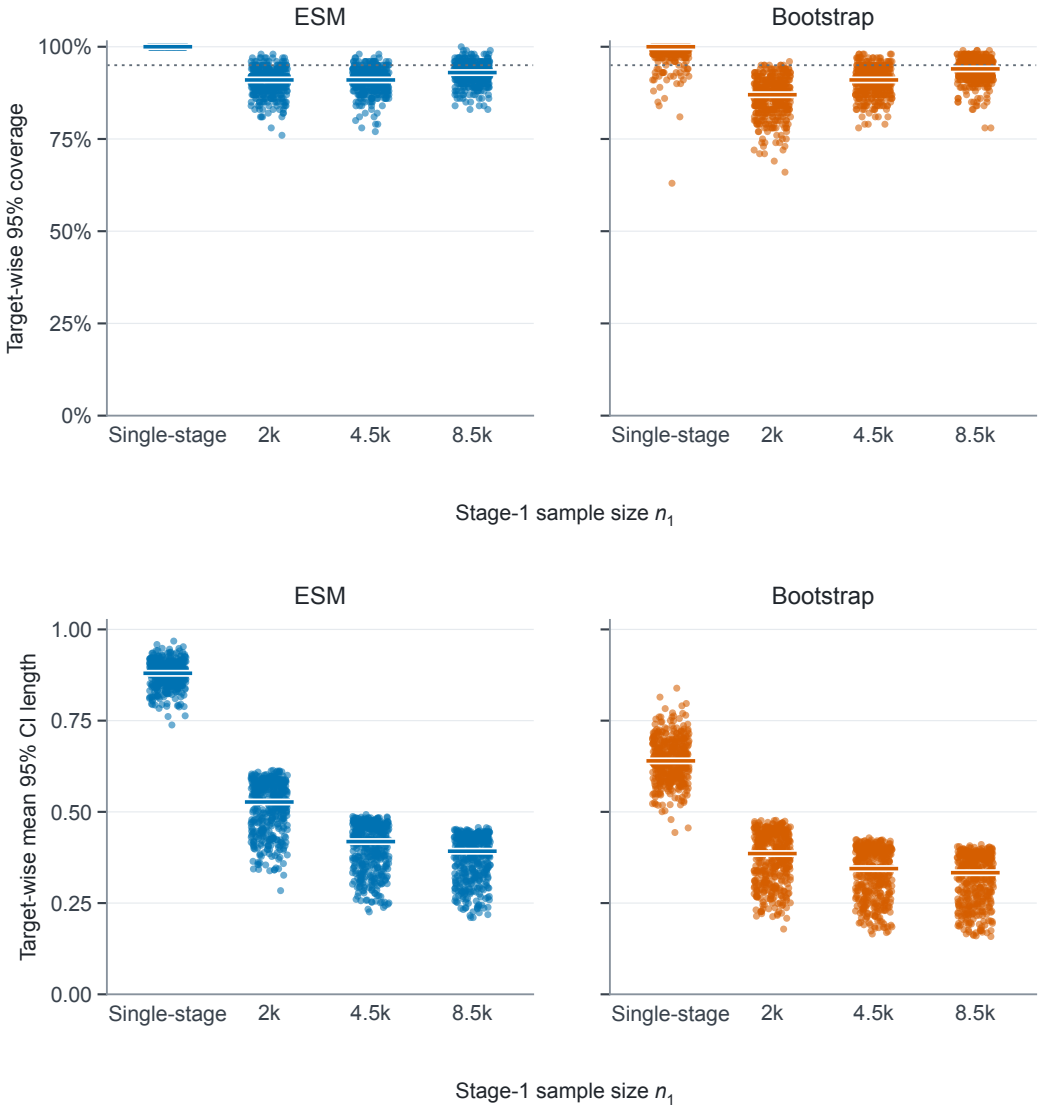


Figure 13: Target-wise 95% coverage (top) and mean 95% confidence-interval length (bottom) for ESM and bootstrap across the single-stage baseline and HIFI with $n_1 \in \{2{,}000, 4{,}500, 8{,}500\}$. Each point represents one of the 424 evaluation targets, with coverage and interval length computed across 100 replications.

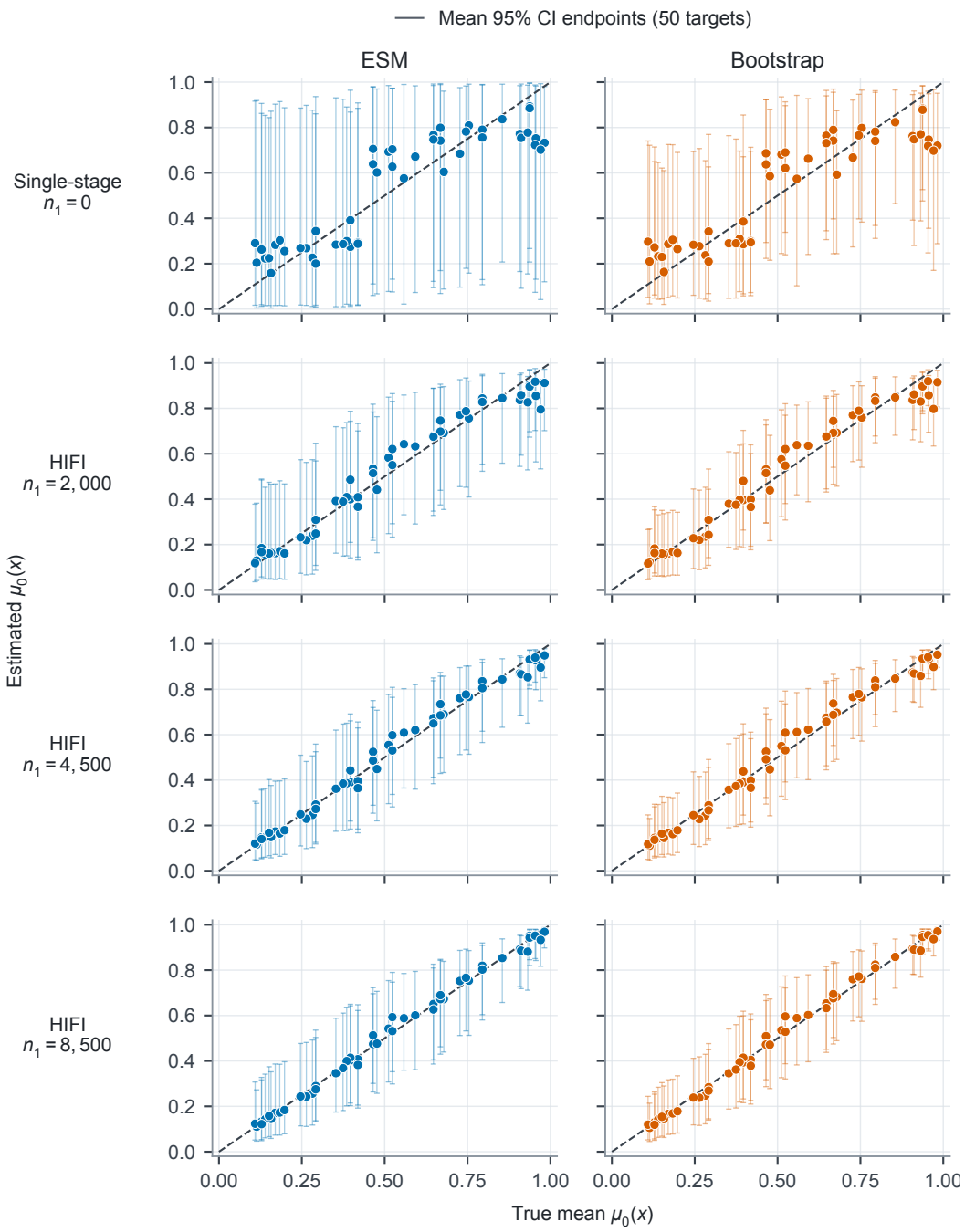


Figure 14: Target-wise point estimates and mean endpoints of 95% confidence intervals for 50 representative targets, selected by stratified random sampling across languages and true-probability bins. Columns correspond to ESM and bootstrap; rows show the single-stage baseline and HIFI with $n_1 \in \{2{,}000, 4{,}500, 8{,}500\}$. Points denote estimates averaged over 100 replications, and vertical segments connect the mean lower and upper confidence-interval endpoints. The dashed line is the identity line.

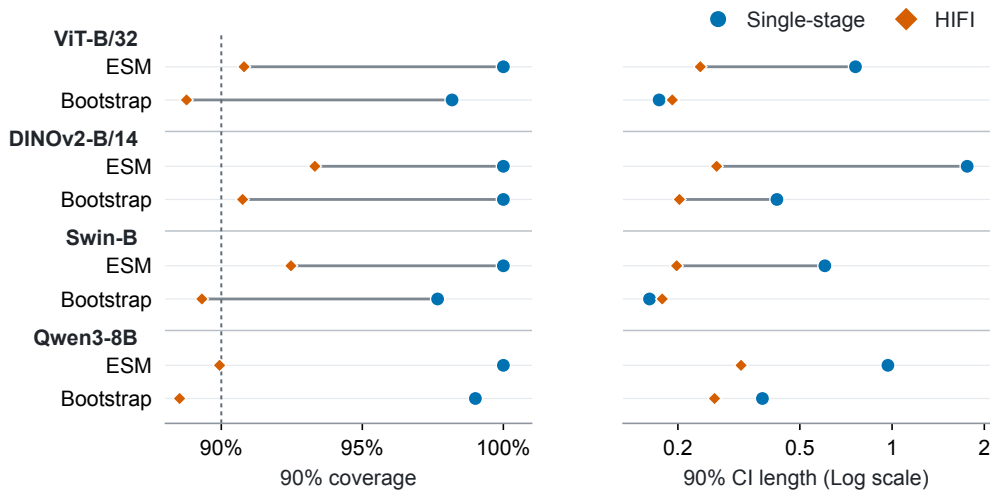


Figure 15: HIFI versus direct single-stage resampling under the same total data budget at the 90% confidence level.

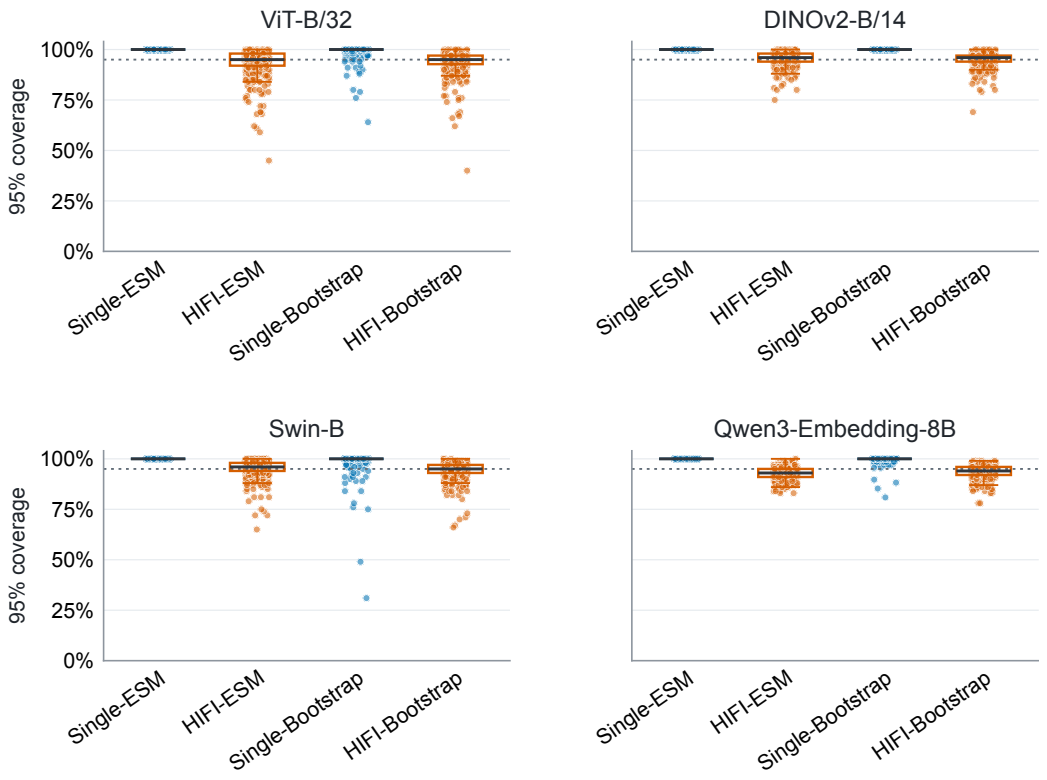


Figure 16: Target-wise empirical coverage of nominal 95% confidence intervals for HIFI and direct single-stage resampling across backbones and inference methods. Each point represents one fixed evaluation target.

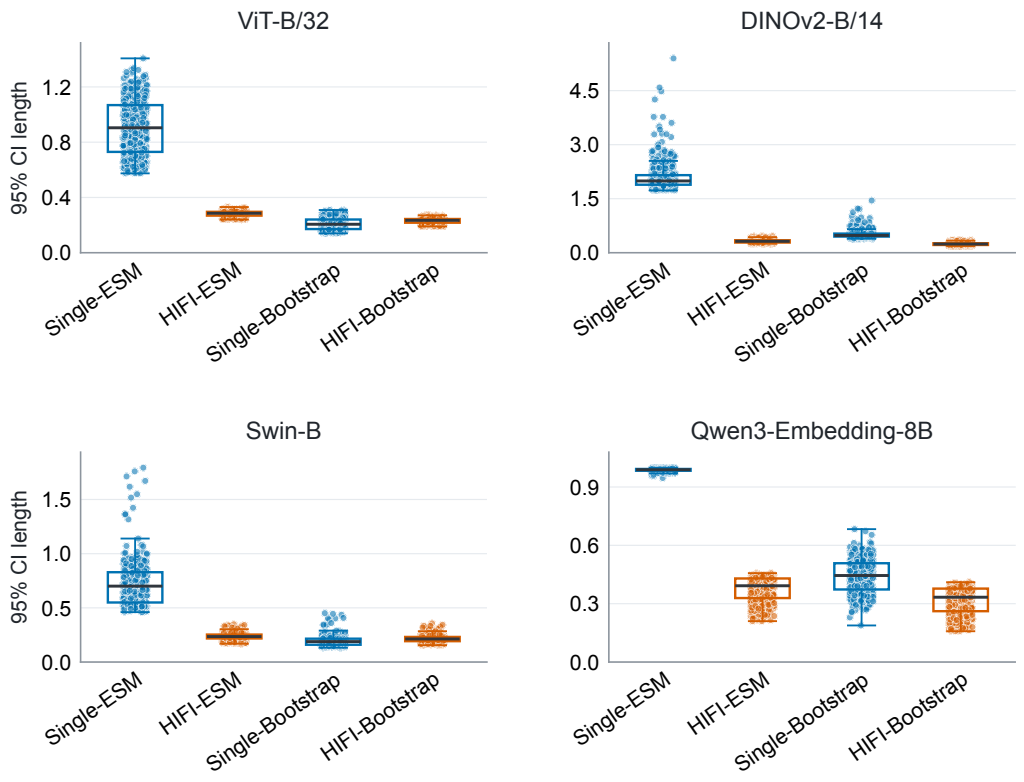


Figure 17: Target-wise lengths of nominal 95% confidence intervals for HIFI and direct single-stage resampling across backbones and inference methods.

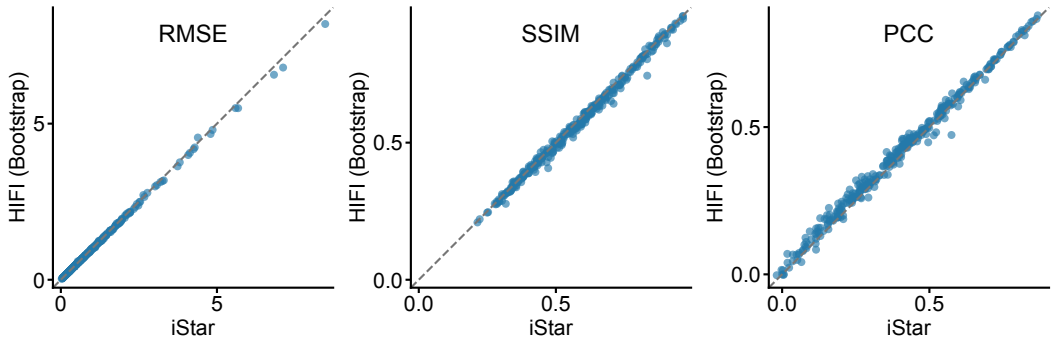


Figure 18: Point-prediction comparison between HIFI with bootstrap and iStar at 16 μm resolution. Each point represents one of the 313 genes, and dashed lines indicate equality. Lower RMSE and higher SSIM and PCC indicate better agreement with observed expression.

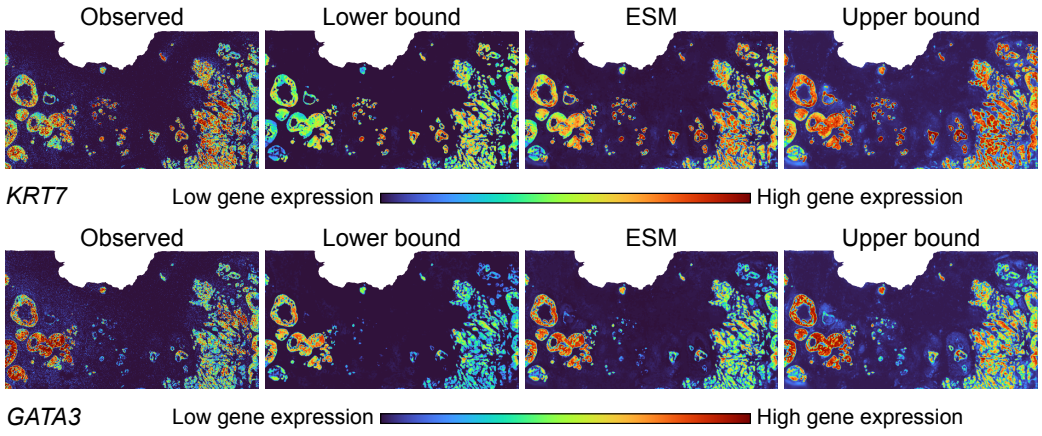


Figure 19: Additional ESM interval maps for *KRT7* and *GATA3*. Columns show observed expression, the lower endpoint, the point estimate, and the upper endpoint of the nominal 95% pointwise interval for mean expression. The corresponding *TACSTD2* example appears in Figure 4.

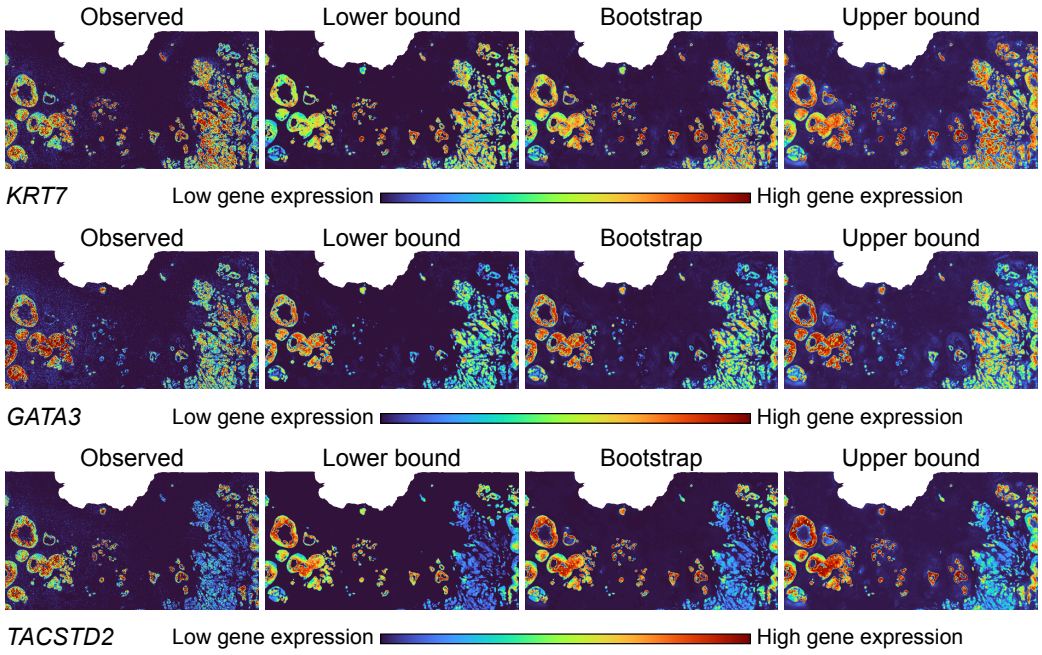


Figure 20: Bootstrap interval maps for *KRT7*, *GATA3*, and *TACSTD2*. Columns show observed expression, the lower percentile endpoint, the bootstrap mean, and the upper percentile endpoint. The endpoints correspond to nominal 95% pointwise intervals. Each gene uses the same display mapping as its ESM panels.