IN-CONTEXT FULL CONFORMAL PREDICTION: VALIDITY AND EFFICIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning enables pretrained models to adapt to new tasks without parameter updates, but rigorous uncertainty quantification for in-context predictions remains poorly understood. Conformal prediction provides finite-sample distribution-free coverage, yet full conformal prediction can be computationally expensive because it may require candidate-dependent model refitting. We develop an in-context full conformal procedure that replaces these refitting steps with forward evaluations of a fixed pretrained predictor. Under exchangeability and a natural context-order invariance condition, we establish finite-sample marginal coverage. We further characterize the statistical efficiency of the resulting prediction sets, showing that it depends jointly on prediction risk and prompt stability. As a concrete nonparametric instantiation, we show that standard softmax Transformers can approximate Gaussian local polynomial regression and achieve the minimax-optimal prediction rate over Hölder classes. Empirically, we show that replacing candidate-wise refitting with forward evaluations of a frozen in-context predictor substantially reduces online computation while maintaining comparable empirical coverage and set length. Together, these results establish in-context learning as a computationally efficient approach to full conformal prediction with pretrained predictors.

1 INTRODUCTION

Recent advances in large language models (LLMs) have revealed remarkable in-context learning (ICL) capabilities: given a small number of input–output demonstrations, a pretrained model can adapt to a new task without updating its parameters (Brown et al., 2020). A growing theoretical literature has begun to characterize how Transformers learn from in-context examples (Garg et al., 2022; Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023; Zhang et al., 2024). However, most existing analyses focus on predictive accuracy, leaving the uncertainty associated with in-context predictions largely uncharacterized. This raises a natural question: can in-context learning be equipped with rigorous uncertainty quantification?

Conformal prediction provides a natural framework for this purpose, producing prediction sets with finite-sample, distribution-free coverage (Vovk et al., 2005; Shafer & Vovk, 2008). Full conformal prediction, however, can be computationally expensive because the underlying predictor is typically refitted as the candidate response varies. ICL offers a natural alternative: repeated model fitting can be replaced by forward evaluations of a pretrained predictor under different prompts.

Recent work has explored conformal inference with in-context learning through empirical studies, Jackknife+-type procedures, and conformal-aware training for full conformal prediction (Huang et al., 2025; Mukherjee et al., 2025; Deng et al., 2025). In contrast, we develop a statistical theory for applying full conformal prediction to a pretrained in-context predictor without conformal-specific training, covering both finite-sample validity and statistical efficiency.

This shift from model refitting to prompt adaptation raises a new theoretical question: whether the resulting score construction still preserves the symmetry required by full conformal prediction. Beyond validity, it is also important to understand when the resulting prediction sets are statistically efficient, rather than merely coverage-valid. We develop a theory addressing both questions and then instantiate the prediction component using standard softmax Transformers.

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.

Contributions. Our main contributions are summarized as follows.

- We develop an in-context full conformal framework that replaces repeated model refitting with forward evaluations of a pretrained predictor, while retaining finite-sample validity.
- We establish oracle-approximation guarantees that characterize conformal-set efficiency through prediction risk and prompt stability.
- We construct standard softmax Transformers that approximate Gaussian local polynomial regression through softmax-weighted moment aggregation and normal-equation solving. We establish the minimax-optimal nonparametric prediction rate and prompt stability, yielding explicit conformal-set efficiency guarantees.
- Experiments show that ICL-FCP achieves near-nominal coverage, competitive prediction-set widths, and substantial computational speedups over classical full conformal prediction.

2 FULL CONFORMAL PREDICTION WITH IN-CONTEXT PREDICTORS

Full conformal prediction evaluates a candidate response z by augmenting the observed sample with $(\mathbf{x}_{n+1}, z)$, recomputing the corresponding nonconformity scores, and retaining candidates whose test score is sufficiently small relative to the calibration scores (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018). Its main computational cost arises from repeatedly refitting the predictive model as z varies. We replace these refitting steps with forward evaluations of a pretrained in-context predictor, while keeping its parameters fixed throughout conformal inference. A review of classical full conformal prediction is provided in Appendix A.

Let $D_n = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, and let $\mathbf{x}_{n+1}$ denote the test covariate. For a candidate response $z \in \mathbb{R}$, define $D^z_{n+1} = (D_n, (\mathbf{x}_{n+1}, z))$, with $y^z_i = y_i$ for $i \in [n]$ and $y^z_{n+1} = z$. For each $i \in [n+1]$, define the leave-one-out prompt $D^z_{-i} = ((\mathbf{x}_j, y^z_j) : j \in [n+1],\, j \neq i)$, with observations retaining their original order, and the corresponding prediction $\widehat{y}^z_i = f_\theta(D^z_{-i}, \mathbf{x}_i)$, where f_θ is the fixed pretrained in-context predictor. Thus, to predict the response of the ith observation, we use the remaining n observations as context. This symmetric construction will be essential for the conformal validity argument below.

Define the residual score $\widehat{S}^z_i = |y^z_i - \widehat{y}^z_i|$, and let $\widehat{q}^z_\theta$ be the empirical $(1-\alpha)(1+1/n)$-quantile of $\widehat{S}^z_1, \ldots, \widehat{S}^z_n$, with the convention that the quantile is $+\infty$ when its level exceeds 1. The resulting prediction set is defined as

$$\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) = \left\{ z \in \mathbb{R} : \widehat{S}^z_{n+1} \leq \widehat{q}^z_\theta \right\}. \tag{1}$$

The candidate response enters the two sides of (1) differently. For the test observation, $D^z_{-(n+1)} = D_n$, so $\widehat{y}^z_{n+1} = f_\theta(D_n, \mathbf{x}_{n+1})$ is independent of z, although the test residual $\widehat{S}^z_{n+1} = |z - \widehat{y}^z_{n+1}|$ depends on z directly. By contrast, for each $i \in [n]$, the candidate pair $(\mathbf{x}_{n+1}, z)$ appears in the prompt D^z_{-i}. Consequently, the calibration predictions $\widehat{y}^z_i$, and hence the cutoff $\widehat{q}^z_\theta$, may also vary with z.

Finite-sample validity. Conformal validity requires the score construction to respect the permutation symmetry of the augmented sample: permuting the observations should induce the same permutation of their nonconformity scores. In the construction above, each score is obtained by applying the same in-context predictor to the remaining observations. A natural sufficient condition is therefore that the predictor be insensitive to the ordering of its demonstrations.

Definition 1 (Context-order invariance)**.** Let $D_m = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, and for any permutation π of $[m]$, let $D^\pi_m = ((\mathbf{x}_{\pi(1)}, y_{\pi(1)}), \ldots, (\mathbf{x}_{\pi(m)}, y_{\pi(m)}))$. A predictor f_θ is *context-order invariant* if $f_\theta(D^\pi_m, \mathbf{x}) = f_\theta(D_m, \mathbf{x})$ for every D_m, query $\mathbf{x}$, and permutation π.

Much of the theoretical literature on ICL models the in-context demonstrations as i.i.d. samples from an underlying task (Garg et al., 2022; Ahn et al., 2023; Raventós et al., 2023; Akyürek et al., 2023; Bai et al., 2023; Zhang et al., 2024). Unlike sequential data such as natural language, the ordering of i.i.d. samples carries no statistical information. It is therefore natural to require the predictor to be

invariant to permutations of the demonstrations, which preserves the permutation symmetry needed for conformal validity.

We can now state the finite-sample validity guarantee.

Theorem 1 (Finite-sample validity)**.** *Suppose that the predictor f_θ is context-order invariant. If $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n+1}, y_{n+1})$ are exchangeable, then $\mathbb{P}\Big\{y_{n+1} \in \widehat{\Gamma}_\alpha(\mathbf{x}_{n+1})\Big\} \geq 1 - \alpha$.*

At $z = y_{n+1}$, context-order invariance makes the leave-one-out score vector permutation-equivariant, so the result follows from the standard conformal rank argument; see Appendix B. The guarantee is independent of predictive accuracy, which instead determines the efficiency of the prediction set studied in Section 3.

The Transformer class in Section 4.2 is context-order invariant by construction. Without positional encodings or order-dependent attention masks, self-attention is permutation-equivariant with respect to the context tokens (Lee et al., 2019); since the query token is kept fixed and the prediction is read from its final representation, permuting the demonstrations leaves the output unchanged.

Practical computation. For numerical evaluation, we restrict the candidate space to a finite grid $\mathcal{Z} = \{z_1, \ldots, z_G\}$. Following Lei (2019, Remark 5), we use G evenly spaced points over $[y_{\min} - 0.25\Delta,\ y_{\max} + 0.25\Delta]$, where $\Delta = y_{\max} - y_{\min}$. The enlarged range mitigates boundary truncation, while G controls the numerical resolution of the prediction set. Classical full conformal prediction requires G model refits, with total cost GC_{fit}, whereas the in-context construction requires $G(n+1)$ forward evaluations, with cost $G(n+1)C_{\text{fwd}}$. Here, C_{fit} denotes the cost of fitting the predictive model once and C_{fwd} the cost of one Transformer forward pass. The forward evaluations can be batched across candidate responses and leave-one-out contexts. The computational benefit of replacing candidate-wise refitting with frozen forward evaluations is demonstrated empirically in Table 1; see Appendix G.5 for detailed timing results.

3 Statistical efficiency of conformal sets

Theorem 1 guarantees finite-sample coverage without requiring the pretrained predictor to be accurate. We now study when the resulting conformal set is statistically efficient, namely, when its size and location are close to those of an oracle interval.

Statistical model and prediction risk. Let $P_{\mathcal{M}}$ be a distribution over measurable regression functions $m : [0,1]^d \to [-M, M]$. For each task, draw $m \sim P_{\mathcal{M}}$ and, conditionally on m, let $y_i = m(\mathbf{x}_i) + \varepsilon_i$, where $(\mathbf{x}_i, \varepsilon_i)_{i=1}^{n+1}$ are i.i.d. from $P_{\mathbf{x}} \otimes P_\varepsilon$. The noise distribution is supported on $[-1, 1]$, symmetric about zero, and has a density f_ε bounded by C_ε; write $\sigma^2 = \mathbb{E}(\varepsilon^2)$. The normalization to $[-1, 1]$ affects only constants. Pretraining uses T independent tasks from the same task distribution, while inference is performed on an independent task. Predictors take values in $[-M, M]$.

Let $\mathbb{E}_{\text{test}}$ and $\mathbb{P}_{\text{test}}$ denote expectation and probability over an independent inference task, with the predictor held fixed. Define $R(f) = \mathbb{E}_{\text{test}}[\{y_{n+1} - f(D_n, \mathbf{x}_{n+1})\}^2]$. By the independence and zero mean of ε_{n+1},

$$R(f) - \sigma^2 = \mathbb{E}_{\text{test}}[\{f(D_n, \mathbf{x}_{n+1}) - m(\mathbf{x}_{n+1})\}^2]. \tag{2}$$

Write $\widehat{f}_T := f_\theta$ for the predictor pretrained on T tasks. Its excess prediction risk is

$$r_{n,T} := R(\widehat{f}_T) - \sigma^2 = \mathbb{E}_{\text{test}}[\{\widehat{f}_T(D_n, \mathbf{x}_{n+1}) - m(\mathbf{x}_{n+1})\}^2]. \tag{3}$$

Thus, $r_{n,T}$ is random only through pretraining. We allow $T = T_n$ as $n \to \infty$. Unless otherwise stated, O_p is understood with respect to the randomness involved in the corresponding statement.

Oracle benchmark. For $Q_\tau(U) = \inf\{t : \mathbb{P}(U \leq t) \geq \tau\}$, let $q_\alpha = Q_{1-\alpha}(|\varepsilon|)$. If m were known, the natural oracle interval would be $\Gamma^*_\alpha(m, \mathbf{x}) = [m(\mathbf{x}) - q_\alpha, m(\mathbf{x}) + q_\alpha]$, which has conditional coverage $1 - \alpha$ and length $2q_\alpha$.

For a realized pretrained predictor $\widehat{f}_T$, let $\bar{q}_\alpha = Q_{1-\alpha}(|y - \widehat{f}_T(D_n, \mathbf{x})|)$ under $\mathbb{P}_{\text{test}}$. A related second-order oracle comparison was established by Lei et al. (2018, Theorem 5) for classical

regression estimators. Here we state the corresponding conditional result for a fixed pretrained predictor evaluated on an independent inference task.

Proposition 1 (Oracle-width comparison)**.** *Under the symmetric-noise model above, suppose the density h_ε of $|\varepsilon|$ is L_ε-Lipschitz on $[0,\infty)$, and $h_\varepsilon(t) \geq c_\alpha > 0$ whenever $|t - q_\alpha| < \eta_\alpha$ for some $0 < \eta_\alpha < q_\alpha$. For any realized predictor satisfying $L_\varepsilon r_{n,T} < 2c_\alpha\eta_\alpha$, we have $|2\bar{q}_\alpha - 2q_\alpha| \leq (L_\varepsilon/c_\alpha)r_{n,T}$.*

Proposition 1 shows that the oracle-width error is of order $r_{n,T}$, rather than $\sqrt{r_{n,T}}$. Under symmetric noise, the first-order perturbation vanishes, so prediction error has a second-order effect on the population residual quantile. The proof is given in Appendix D.1.

3.1 Prompt stability and oracle approximation

Since $y_{n+1} \in \mathcal{Y} := [-M-1, M+1]$ almost surely, we restrict the prediction set in (1) to $z \in \mathcal{Y}$ throughout the efficiency analysis, retaining the notation $\widehat{\Gamma}_\alpha$. This restriction preserves coverage.

At the true candidate $z = y_{n+1}$, write $\widehat{y}_i = \widehat{y}_i^{\,y_{n+1}}$ for $i \in [n]$. For distinct $i, j \in [n]$, let ε'_j be an independent copy of ε_j, set $y'_j = m(\mathbf{x}_j) + \varepsilon'_j$, and let $\widehat{y}_i^{(j)}$ denote the prediction obtained by replacing y_j with y'_j in the leave-one-out context for $\mathbf{x}_i$, with all other quantities fixed.

Assumption 1 (Prompt stability)**.** For $n \geq 2$, there exist deterministic sequences $\kappa_n \to 0$ and $\rho_n \to 0$ such that

$$\sup_{z\in\mathcal{Y}} \max_{i\in[n]} |\widehat{y}_i^{\,z} - \widehat{y}_i| = O_p(\kappa_n), \tag{4}$$

and

$$\frac{1}{n(n-1)} \sum_{\substack{i,j\in[n]\\ i\neq j}} \mathbb{E}_{\text{test}} |\widehat{y}_i - \widehat{y}_i^{(j)}| = O_p(\rho_n). \tag{5}$$

Here, $\mathbb{E}_{\text{test}}$ also averages over the independent replacement noises.

The first condition controls the effect of changing the candidate response on calibration predictions, analogous to classical full conformal stability (Lei et al., 2018), but with fixed model parameters. The second controls the dependence induced by overlapping leave-one-out prompts, analogous to algorithmic stability (Bousquet & Elisseeff, 2002; Li et al., 2023). The stability rates are understood along $T = T_n$ and may depend on this pretraining schedule.

Let Leb denote the Lebesgue measure and define $\widehat{\nu}_\alpha(\mathbf{x}) := \mathrm{Leb}(\widehat{\Gamma}_\alpha(\mathbf{x}))$. We can now quantify the discrepancy between the conformal set and the oracle benchmark.

Theorem 2 (Oracle approximation of conformal sets)**.** *Suppose $\widehat{f}_T$ is context-order invariant and takes values in $[-M, M]$, the density h_ε satisfies the conditions of Proposition 1, and Assumption 1 holds. If $r_{n,T} = O_p(a_{n,T})$ for some deterministic sequence $a_{n,T} \to 0$, then*

$$|\widehat{\nu}_\alpha(\mathbf{x}_{n+1}) - 2q_\alpha| = O_p\Big(\sqrt{n^{-1} + \rho_n} + a_{n,T} + \kappa_n\Big),$$

and

$$\mathrm{Leb}\Big\{\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \,\triangle\, \Gamma^*_\alpha(m, \mathbf{x}_{n+1})\Big\} = O_p\Big(\sqrt{n^{-1} + \rho_n} + \sqrt{a_{n,T}} + \kappa_n\Big),$$

where $A \triangle B = (A \cap B^c) \cup (A^c \cap B)$ denotes the symmetric difference of two sets A and B.

The first bound controls the deviation of the conformal-set length from the oracle length $2q_\alpha$, while the second measures the discrepancy between the conformal set and the oracle interval in symmetric difference. The term $\sqrt{n^{-1} + \rho_n}$ reflects the fluctuation of the empirical conformal cutoff under dependence among the leave-one-out scores, whereas κ_n captures sensitivity of the cutoff to the candidate response. Under symmetric noise, the excess prediction risk affects the population residual quantile only at second order, which yields the $a_{n,T}$ term in the length bound. By contrast, the discrepancy between the predictive center $\widehat{f}_T(D_n, \mathbf{x}_{n+1})$ and the oracle center $m(\mathbf{x}_{n+1})$ is of order $\sqrt{a_{n,T}}$, leading to the corresponding term in the symmetric-difference bound. The proof of Theorem 2 is given in Appendix D.3.

For linear smoothers, the two stability conditions reduce to bounds on the largest smoothing weight assigned to the candidate response and the average magnitude of the weights assigned to context responses, respectively. These bounds are preserved under one-Lipschitz post-processing, such as clipping. Examples and the Gaussian local-polynomial verification are given in Appendices C and C.1.

4 Nonparametric guarantees for softmax Transformers

The preceding section characterizes conformal-set efficiency in terms of prediction risk and two prompt-stability quantities. We now study the prediction-risk component under a nonparametric regression model. We construct a softmax Transformer that approximates Gaussian local polynomial regression. Combining this approximation with statistical learning bounds yields an excess-risk bound whose leading term attains the minimax-optimal nonparametric rate over Hölder classes. This result can be combined with Theorem 2 to control conformal-set efficiency whenever the required prompt-stability conditions also hold.

Recent work has established statistical guarantees for Transformer-based nonparametric regression in both supervised and in-context settings (Jiao et al., 2026; Kim et al., 2024; Ma et al., 2025; Liu et al., 2025; Shen et al., 2026; Kong et al., 2026). In particular, Kim et al. (2024) study minimax-optimal nonparametric ICL through learned basis representations, while Shen et al. (2026) connect attention-based prediction with kernel methods. Most closely related, Ching et al. (2026) show that pretrained Transformers can approximate local polynomial regression and attain the minimax-optimal rate over Hölder classes.

Our analysis differs from Ching et al. (2026) in architecture. Their construction uses linear attention with gradient-descent updates on kernel-weighted polynomial features, whereas ours uses standard softmax attention to aggregate Gaussian-weighted polynomial moments and positionwise ReLU networks to form and solve the resulting fixed-dimensional normal equations.

4.1 Nonparametric model and pretraining

We retain the data-generating model of Section 3 and assume that the task distribution $P_{\mathcal{M}}$ is supported on the Hölder class $\mathcal{H}^\beta([0,1]^d, M)$. For $\beta > 0$, let $p := \lceil \beta \rceil - 1$. We further assume that $P_{\mathbf{x}}$ has a Lebesgue density $f_{\mathbf{x}}$ satisfying $0 < c_{\mathbf{x}} \le f_{\mathbf{x}}(\mathbf{x}) \le C_{\mathbf{x}} < \infty$ for almost every $\mathbf{x} \in [0,1]^d$, a standard random-design condition for local polynomial regression (Fan, 2018; Györfi et al., 2002).

Definition 2 (Hölder class)**.** The class $\mathcal{H}^\beta([0,1]^d, M)$ consists of all $m \in C^p([0,1]^d)$ such that $\max_{|\boldsymbol{\alpha}|_1 \le p} \|\partial^{\boldsymbol{\alpha}} m\|_\infty \le M$, $\max_{|\boldsymbol{\alpha}|_1 = p} \sup_{\mathbf{x} \ne \mathbf{x}'} \frac{|\partial^{\boldsymbol{\alpha}} m(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} m(\mathbf{x}')|}{\|\mathbf{x} - \mathbf{x}'\|_2^{\beta - p}} \le M$. Here $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, $|\boldsymbol{\alpha}|_1 := \sum_{j=1}^d \alpha_j$, and $\partial^{\boldsymbol{\alpha}} := \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d}$.

For a bounded predictor class $\mathcal{F}$, define the pretraining empirical risk by $\widehat{R}_T(f) := T^{-1} \sum_{t=1}^T \{y_{n+1}^{(t)} - f(D_n^{(t)}, \mathbf{x}_{n+1}^{(t)})\}^2$, and let $\widehat{f}_T \in \arg\min_{f \in \mathcal{F}} \widehat{R}_T(f)$. We assume that a measurable minimizer is selected; our analysis concerns this statistical ERM target and does not require a particular optimization algorithm to attain it.

4.2 Transformer-induced hypothesis class

We next specify the hypothesis class $\mathcal{F}$ used for empirical risk minimization. Let $d_{\mathrm{model}} \ge d + 3$. For $D_n = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) \in ([0,1]^d \times \mathbb{R})^n$ and $\mathbf{x}_{n+1} \in [0,1]^d$, define

$$\mathrm{Embed}_{d_{\mathrm{model}}}(D_n, \mathbf{x}_{n+1}) := \begin{pmatrix} \mathbf{x}_1^\top & y_1 & 0 & 1 & \mathbf{0}_{d_{\mathrm{model}}-d-3}^\top \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_n^\top & y_n & 0 & 1 & \mathbf{0}_{d_{\mathrm{model}}-d-3}^\top \\ \mathbf{x}_{n+1}^\top & 0 & 1 & 1 & \mathbf{0}_{d_{\mathrm{model}}-d-3}^\top \end{pmatrix} \in \mathbb{R}^{(n+1) \times d_{\mathrm{model}}}. \tag{6}$$

The first n rows are context tokens and the last is the query token. The two coordinates following the response are a query indicator and a constant coordinate, respectively, while the remaining coordinates are auxiliary registers.

For each $\ell \in [L]$ and $r \in [H]$, let $\mathbf{Q}_{\ell,r}, \mathbf{K}_{\ell,r}, \mathbf{V}_{\ell,r} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$, and define

$$\text{Attn}_\ell(\mathbf{Z}) := \mathbf{Z} + \sum_{r=1}^{H} \text{softmax}_{\text{row}}\left(\frac{(\mathbf{Z}\mathbf{Q}_{\ell,r})(\mathbf{Z}\mathbf{K}_{\ell,r})^\top}{\sqrt{d_{\text{model}}}} \right) \mathbf{Z}\mathbf{V}_{\ell,r}, \tag{7}$$

where $\text{softmax}_{\text{row}}$ applies softmax to each row.

Let $\text{ReLU}(t) := \max\{t, 0\}$ act entrywise, and define

$$\text{FF}_\ell(\mathbf{Z}) := \mathbf{Z} + \text{ReLU}\big(\mathbf{Z}\mathbf{W}_{\ell,1} + \mathbf{1}_{n+1}\mathbf{b}_{\ell,1}^\top\big)\, \mathbf{W}_{\ell,2} + \mathbf{1}_{n+1}\mathbf{b}_{\ell,2}^\top,$$

where $\mathbf{W}_{\ell,1} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$, $\mathbf{W}_{\ell,2} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$, $\mathbf{b}_{\ell,1} \in \mathbb{R}^{d_{\text{ff}}}$, and $\mathbf{b}_{\ell,2} \in \mathbb{R}^{d_{\text{model}}}$.

Writing $\mathcal{T}_\ell := \text{FF}_\ell \circ \text{Attn}_\ell$, let θ collect all trainable parameters across layers and heads, and define

$$f_\theta(D_n, \mathbf{x}_{n+1}) := \text{clip}_M\Big(\big[(\mathcal{T}_L \circ \cdots \circ \mathcal{T}_1)\big(\text{Embed}_{d_{\text{model}}}(D_n, \mathbf{x}_{n+1})\big)\big]_{n+1,d+1} \Big),$$

where $\text{clip}_M(t) := (-M) \vee (t \wedge M)$. We write $\|\theta\|_\infty$ for the largest absolute parameter value and $\|\theta\|_0$ for the number of nonzero parameters.

Definition 3 (Transformer-induced hypothesis class)**.** For $B, M > 0$ and $S \in \mathbb{N}$, define

$$\mathcal{F}(d_{\text{model}}, d_{\text{ff}}, H, L, B, S, M) := \{ f_\theta : \|\theta\|_\infty \le B,\ \|\theta\|_0 \le S \}. \tag{8}$$

The class $\mathcal{F}$ is context-order invariant by construction. Indeed, without positional encodings or order-dependent attention masks, each self-attention layer is permutation equivariant with respect to the context tokens, and the positionwise feed-forward layers preserve this equivariance. Since the query token is kept fixed and f_θ reads out only its final representation, permuting the context examples does not change the query output. Consequently, every $f_\theta \in \mathcal{F}$ satisfies the context-order invariance condition required by Theorem 1.

4.3 Gaussian local polynomial realization

We construct a Transformer comparator based on Gaussian local polynomial regression. Recall that $p = \lceil \beta \rceil - 1$. For $q \in \{p, 2p\}$, let $\mathcal{A}_q = \{\boldsymbol{\alpha} \in \mathbb{N}_0^d : |\boldsymbol{\alpha}|_1 \le q\}$ and $J_q = |\mathcal{A}_q| = \binom{d+q}{q}$. Define $\boldsymbol{\psi}_p(\mathbf{u}) = (\mathbf{u}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathcal{A}_p}$, with the constant monomial first, and let $\mathbf{e}_0 = (1, 0, \ldots, 0)^\top$.

For the Gaussian kernel $K(\mathbf{u}) = \exp(-\|\mathbf{u}\|_2^2/2)$, bandwidth $h \in (0, 1]$, and query $\mathbf{x}$, let $\mathbf{u}_i = (\mathbf{x}_i - \mathbf{x})/h$ and $a_h(\mathbf{x}) = 1 + \sum_{i=1}^n K(\mathbf{u}_i)$. Define

$$\mathbf{G}_h(D_n, \mathbf{x}) = a_h(\mathbf{x})^{-1} \sum_{i=1}^n K(\mathbf{u}_i) \boldsymbol{\psi}_p(\mathbf{u}_i) \boldsymbol{\psi}_p(\mathbf{u}_i)^\top,$$

$$\mathbf{b}_h(D_n, \mathbf{x}) = a_h(\mathbf{x})^{-1} \sum_{i=1}^n K(\mathbf{u}_i) \boldsymbol{\psi}_p(\mathbf{u}_i) y_i.$$

When $\mathbf{G}_h(D_n, \mathbf{x})$ is invertible, define the Gaussian local polynomial predictor by

$$f_{\text{LocPol},h}(D_n, \mathbf{x}) = \text{clip}_M\big(\mathbf{e}_0^\top \mathbf{G}_h(D_n, \mathbf{x})^{-1} \mathbf{b}_h(D_n, \mathbf{x})\big), \tag{9}$$

where $\text{clip}_M(t) = (-M) \vee (t \wedge M)$; otherwise, set $f_{\text{LocPol},h}(D_n, \mathbf{x}) = 0$. The common normalization $a_h(\mathbf{x})^{-1}$ cancels from the normal equations. The extra 1 accounts for the query token in the softmax construction, which has logit zero and contributes zero to the moment numerators.

We show that a softmax Transformer can approximate this predictor in a single forward pass. Set $W_0 := 2d + 4J_{2p} + 6J_p^2 + 8J_p + 4p + 40$.

Theorem 3 (Softmax Transformer approximation)**.** *Fix $0 < \lambda < \Lambda < \infty$. There exists a constant $C \ge 1$, depending only on $d, p, M, \lambda, \Lambda$, such that, for every $h \in (0, 1]$ and $\varepsilon \in (0, 1/2)$, there exist $L, S \in \mathbb{N}$ and a predictor $f_{h,\varepsilon} \in \mathcal{F}(W_0, 3W_0, 1, L, Ch^{-\max\{2,2p\}}, S, M)$ satisfying $L \vee S \le C \log(e/(h\varepsilon))$ and*

$$|f_{h,\varepsilon}(D_n, \mathbf{x}) - f_{\text{LocPol},h}(D_n, \mathbf{x})| \le \varepsilon \tag{10}$$

for every $n \in \mathbb{N}$, $D_n \in ([0,1]^d \times [-M-1, M+1])^n$, and $\mathbf{x} \in [0,1]^d$ satisfying $\lambda \mathbf{I}_{J_p} \preceq \mathbf{G}_h(D_n, \mathbf{x}) \preceq \Lambda \mathbf{I}_{J_p}$. Here, the same predictor $f_{h,\varepsilon}$ applies to every n.

Architecture. The explicit width satisfies $W_0 = O(d + J_{2p} + J_p^2)$, and the FFN hidden width is $3W_0$. The feed-forward layers first construct the polynomial features required by local polynomial regression. A softmax attention layer then assigns Gaussian-kernel weights to the context observations and aggregates the corresponding weighted polynomial moments. The subsequent feed-forward layers solve the resulting local normal equations and output the local polynomial prediction.

The proof is given in Appendix F. Under the random-design assumptions of Section 4.1, Lemma 4 shows that the required conditioning holds at an independent random query with probability at least $1 - C_1 \exp(-c_1 n h^d)$ whenever $nh^d \geq 1$. By decreasing λ and increasing Λ if necessary, we choose common conditioning constants for Lemmas 4 and 1. The resulting $f_{h,\varepsilon}$ is used in both the risk and stability analyses.

4.4 Excess risk and conformal efficiency

We now establish an excess-risk bound for pretrained Transformers using the preceding construction as a comparator.

Theorem 4 (Excess risk of softmax Transformers)**.** *Under the assumptions of Section 4.1, there exist constants $C_0, C \geq 1$, depending only on $d, \beta, M, c_{\mathbf{x}}, C_{\mathbf{x}}$, such that the following holds. For $n, T \in \mathbb{N}$, let $L_n = \lceil C_0 \log(en) \rceil$, $B_n = C_0 n^2$, and $\mathcal{F}_n = \mathcal{F}(W_0, 3W_0, 1, L_n, B_n, L_n, M)$. Any measurable empirical risk minimizer $\widehat{f}_T$ of $\widehat{R}_T$ over $\mathcal{F}_n$, based on T independent pretraining tasks, satisfies, for every $\gamma \geq 1$, with probability at least $1 - Ce^{-\gamma}$,*

$$r_{n,T} = R(\widehat{f}_T) - \sigma^2 \leq C \left[n^{-2\beta/(2\beta+d)} + \frac{\log^4(en) + \log(en)\log(eT) + \gamma}{T} \right]. \tag{11}$$

The first term in (11) is the minimax-optimal nonparametric regression rate over the Hölder class, matching earlier minimax-optimal ICL guarantees (Kim et al., 2024; Ching et al., 2026). Our result differs in that it is established for standard softmax attention, which is closer to the Transformer architecture commonly used in practice. The Gaussian kernel weights are realized directly through the softmax logits, while the resulting weighted moments are used to recover the local polynomial estimator. By contrast, Kim et al. (2024) use learned basis representations with linear attention, whereas Ching et al. (2026) implement local polynomial regression through linear attention and gradient-descent updates. The proof of Theorem 4 is provided in Appendix E.

Furthermore, the constructed Transformer in Theorem 3 satisfies the required stability conditions in Assumption 1 with explicit rates, which, together with the oracle-approximation theorem, yield explicit conformal-set efficiency guarantees; see Proposition 3 and Corollary 2 in Appendix C.

For the pretrained ERM, if $T \gtrsim n^{2\beta/(2\beta+d)} \log^4(en)$, Theorem 4 yields $r_{n,T} = O_p(n^{-2\beta/(2\beta+d)})$. Conformal efficiency additionally requires prompt stability.

Corollary 1 (Conformal efficiency under stable softmax Transformer ERM)**.** *Assume the conditions of Theorem 4 and the noise-regularity conditions of Theorem 2. If $\widehat{f}_T$ additionally satisfies $\rho_n = O(n^{-1})$, $\kappa_n = O(n^{-2\beta/(2\beta+d)})$, and $T \gtrsim n^{2\beta/(2\beta+d)} \log^4(en)$, then*

$$\left| \widehat{\nu}_\alpha(\mathbf{x}_{n+1}) - 2q_\alpha \right| = O_p\Big(n^{-1/2} + n^{-2\beta/(2\beta+d)} \Big) \tag{12}$$

and

$$\mathrm{Leb}\Big(\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \triangle \Gamma^*_\alpha(m, \mathbf{x}_{n+1}) \Big) = O_p\Big(n^{-1/2} + n^{-\beta/(2\beta+d)} \Big). \tag{13}$$

The excess-risk bound in Theorem 4 does not establish prompt stability for the empirical risk minimizer $\widehat{f}_T$. Existing Transformer stability analyses also require additional assumptions (Li et al., 2023; Mukherjee et al., 2025). We therefore state Corollary 1 under explicit prompt-stability conditions.

5 Experiments

Setup. We evaluate in-context full conformal prediction (ICL-FCP) on six regression families: sparse linear, generalized additive models (GAMs), Friedman-1, multivariate adaptive regression

splines (MARS)-style hinges, radial basis function (RBF) Gaussian processes, and neural teachers. Five independent 648,961-parameter Transformers are pretrained for 8,000 updates and then frozen. ICL-FCP uses the frozen predictor and leave-one-out scores defined in Section 2. Three baselines—multilayer perceptron (MLP), ExtraTrees, and RBF-kernel support vector regression (RBF-SVR)—instead refit a model for each candidate z, requiring G fits. All methods use the same conformal rank criterion $p(z) = (n+1)^{-1} \sum_{i=1}^{n+1} \mathbf{1}\{S_i^z \geq S_{n+1}^z\} > \alpha$, where S_i^z denotes the method-specific nonconformity score.

Validity and grid resolution. We assess finite-sample validity by comparing direct rank coverage with the attainable finite-rank reference. Across 24,000 evaluations on 9,000 independent tasks, all references lie within the Wilson 95% intervals (Figure 1a). To assess numerical accuracy, grid and direct membership disagree in only 28 cases (0.117%). Across 60 matched tasks, membership is unchanged for $G \in \{101, 201, 401, 801, 1601\}$, with $|L_{801} - L_{1601}| = 0.0022$ (Figure 1b).

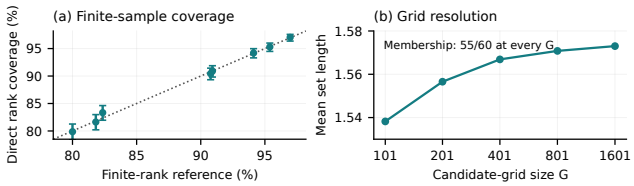


Figure 1: Finite-sample validity and numerical accuracy of Frozen ICL-FCP. (a) Direct rank coverage versus the attainable finite-rank reference, with Wilson 95% confidence intervals. (b) Mean prediction-set length under grid refinement on 60 matched tasks.

Prediction quality. We next examine whether ICL-FCP achieves competitive prediction-set efficiency while maintaining empirical coverage. Across 600 paired tasks, ICL-FCP achieves 90.33% coverage with mean set length 1.631, compared with 90.17% and 5.959 for MLP-FCP (5.986 after the endpoint follow-up of Table 6). Its 90th-percentile length is also substantially smaller (2.482 versus 20.381), while its mean set length is comparable to ExtraTrees-FCP and RBF-SVR-FCP (Table 1).

Table 1: Prediction quality on 600 paired tasks ($n = 32$, $\alpha = 0.1$, $G = 801$) and online latency on 60 matched tasks. Coverage includes paired-bootstrap 95% intervals; median and P90 denote set-length quantiles. ICL and MLP are timed on an RTX A6000 GPU, while ExtraTrees and RBF-SVR use CPU implementations on a Xeon w9-3495X. Time ratios are relative to Frozen ICL-FCP; cross-device comparisons are descriptive. RBF-SVR is the exact-rank ($\tau = 0$) record; the stabilized variant is the sensitivity check of Appendix G.3. The MLP length precedes the endpoint follow-up of Table 6 (5.986 afterwards).

Method	Coverage (%)	Mean length	Median	P90	Time (ms)	Speedup vs. **ICL**
Frozen ICL-FCP	**90.33** **[88.00, 92.50]**	**1.631**	**1.516**	**2.482**	**11.04**	**1.00**×
MLP-FCP	90.17 [87.83, 92.50]	5.959	1.812	20.381	788.58	71.44×
ExtraTrees-FCP	91.00 [88.67, 93.17]	1.590	1.507	2.422	5290.86	479.31×
Exact RBF-SVR-FCP	90.83 [88.50, 93.00]	1.690	1.566	2.564	55.31	5.01×

Per-family performance and online latency. Across six regression families, ICL-FCP achieves 86–93% empirical coverage with competitive set lengths (Table 2). Its online latency is 11.04 ms, yielding 71.4×, 479.3×, and 5.0× lower runtime than MLP-FCP, ExtraTrees-FCP, and RBF-SVR-FCP, respectively. The MLP comparison uses the same RTX A6000, whereas the CPU comparisons are descriptive. See Appendix G.5 for details.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing, language polishing, and experimental design.

Table 2: Per-family coverage, mean set length, and median online latency on 100 paired tasks per family. Coverage is descriptive; speedup ratios use the corresponding Frozen ICL-FCP baseline. ExtraTrees and RBF-SVR use archived CPU records; RBF-SVR is the exact-rank ($\tau = 0$) record.

Family	Method	Cov. (%)	Mean len.	Time (ms)	vs. **ICL**
Sparse linear	**ICL-FCP**	**90.0**	**1.280**	**10.97**	**1.00**×
	MLP-FCP	88.0	2.632	769.82	70.18×
	ExtraTrees	89.0	1.278	5,246.46	478.29×
	RBF-SVR	85.0	1.305	55.24	5.04×
GAM	**ICL-FCP**	**90.0**	**2.240**	**11.07**	**1.00**×
	MLP-FCP	90.0	7.743	1,532.35	138.42×
	ExtraTrees	93.0	2.036	5,318.13	480.41×
	RBF-SVR	95.0	2.235	51.70	4.67×
Friedman-1	**ICL-FCP**	**93.0**	**1.578**	**11.07**	**1.00**×
	MLP-FCP	87.0	6.655	785.50	70.94×
	ExtraTrees	93.0	1.910	5,112.34	461.71×
	RBF-SVR	93.0	1.713	55.43	5.01×
MARS-style	**ICL-FCP**	**86.0**	**1.569**	**10.86**	**1.00**×
	MLP-FCP	91.0	5.100	794.24	73.14×
	ExtraTrees	89.0	1.380	5,440.69	501.02×
	RBF-SVR	90.0	1.609	59.13	5.45×
RBF GP	**ICL-FCP**	**90.0**	**2.097**	**10.97**	**1.00**×
	MLP-FCP	88.0	9.914	767.93	70.00×
	ExtraTrees	89.0	1.922	5,042.08	459.58×
	RBF-SVR	90.0	2.187	53.58	4.88×
Neural teacher	**ICL-FCP**	**93.0**	**1.020**	**11.18**	**1.00**×
	MLP-FCP	97.0	3.707	1,159.35	103.72×
	ExtraTrees	93.0	1.012	5,382.73	481.57×
	RBF-SVR	92.0	1.091	91.50	8.19×

Reproducibility statement

The conformal procedure, statistical assumptions, and Transformer architecture are fully specified in the main text. Complete proofs of the theoretical results are provided in the appendices. Experimental details and implementation settings are provided in the experimental appendices and in the supplementary material.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. In *Forty-third International Conference on Machine Learning*, 2026.

Weicao Deng, Sangwoo Park, Min Li, and Osvaldo Simeone. Optimizing in-context learning for efficient full conformal prediction. *IEEE Signal Processing Letters*, 33:311–315, 2025.

Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Zhe Huang, Simone Rossi, Rui Yuan, and Thomas Hannagan. From predictions to confidence intervals: an empirical study of conformal prediction methods for in-context learning. *arXiv preprint arXiv:2504.15722*, 2025.

Yuling Jiao, Yanming Lai, Defeng Sun, Yang Wang, and Bokai Yan. Approximation bounds for transformer networks with application to regression. In *Forty-third International Conference on Machine Learning*, 2026.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. *Advances in Neural Information Processing Systems*, 37:106667–106713, 2024.

Insung Kong, Niklas Dexheimer, and Johannes Schmidt-Hieber. Generalization bounds for transformer-based next-token prediction in a language model. *arXiv preprint arXiv:2606.13280*, 2026.

Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A theoretical analysis of deep neural networks and parametric pdes. *Constructive Approximation*, 55(1):73–125, 2022.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106 (4):749–764, 2019.

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.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International conference on machine learning*, pp. 19565–19594. PMLR, 2023.

Chenrui Liu, Falong Tan, Chuanlong Xie, Yicheng Zeng, and Lixing Zhu. In-context learning as nonparametric conditional probability estimation: Risk bounds and optimality. *arXiv preprint arXiv:2508.08673*, 2025.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Tianyi Ma, Tengyao Wang, and Richard J Samworth. Optimal in-context adaptivity and distributional robustness of transformers. *arXiv preprint arXiv:2510.23254*, 2025.

Sohom Mukherjee, Ivane Antonov, Kai Günder, and Magnus Josef Maichle. Predictive inference is really free with in-context learning. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. In *International Conference on Learning Representations*, volume 2026, pp. 42067–42103, 2026.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Larry Wasserman. *All of nonparametric statistics*. Springer, 2006.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

A CLASSICAL FULL CONFORMAL PREDICTION

We briefly recall classical full conformal prediction in the regression setting. Let $D_n = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, where $\mathbf{x}_i \in [0,1]^d$ and $y_i \in \mathbb{R}$. Given a test covariate $\mathbf{x}_{n+1}$, the goal is to construct a prediction set $\Gamma_\alpha(\mathbf{x}_{n+1}) \subseteq \mathbb{R}$ with miscoverage level $\alpha \in (0,1)$ such that $\mathbb{P}\{y_{n+1} \in \Gamma_\alpha(\mathbf{x}_{n+1})\} \geq 1 - \alpha$.

For each candidate response $z \in \mathbb{R}$, define the augmented data set $D^z_{n+1} = (D_n, (\mathbf{x}_{n+1}, z))$, and let $y^z_i = y_i$ for $i \leq n$ and $y^z_{n+1} = z$. Given a nonconformity score $s((\mathbf{x}, y); D)$ that is symmetric with respect to permutations of the observations in D, full conformal prediction computes $S^z_i = s((\mathbf{x}_i, y^z_i); D^z_{n+1})$ for $i = 1, \ldots, n+1$. For example, if the score is based on the residual of a predictor $\widehat{f}_D$ fitted on D, then $S^z_i = |y^z_i - \widehat{f}_{D^z_{n+1}}(\mathbf{x}_i)|$. Thus, for each fixed candidate z, all $n+1$ nonconformity scores are evaluated using the same augmented data set D^z_{n+1}.

The full conformal prediction set retains those candidate values whose test score is not unusually large relative to the remaining scores:

$$\Gamma_\alpha(\mathbf{x}_{n+1}) = \left\{ z \in \mathbb{R} : S^z_{n+1} \leq \widehat{q}^z_{\text{full}} \right\}, \qquad \widehat{q}^z_{\text{full}} = \text{Quantile}\left(S^z_1, \ldots, S^z_n; (1-\alpha)\left(1 + \frac{1}{n}\right) \right).$$

Here, for $a_1, \ldots, a_n \in \mathbb{R}$, $\text{Quantile}(a_1, \ldots, a_n; \tau)$ denotes the $\lceil n\tau \rceil$th order statistic, with the convention $\text{Quantile}(a_1, \ldots, a_n; \tau) = +\infty$ when $\tau > 1$. Under exchangeability of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n+1}, y_{n+1})$ and permutation symmetry of the score construction, the standard conformal rank argument yields

$$\mathbb{P}\{y_{n+1} \in \Gamma_\alpha(\mathbf{x}_{n+1})\} \geq 1 - \alpha.$$

The main drawback of full conformal prediction is computational. Since D^z_{n+1} depends on the candidate value z, changing z generally requires refitting the predictor $\widehat{f}_{D^z_{n+1}}$ and recomputing the associated nonconformity scores. In regression, where z ranges over a continuum, this repeated refitting can make exact full conformal prediction computationally prohibitive.

B FINITE-SAMPLE VALIDITY

Proof of Theorem 1. For the true candidate $z = y_{n+1}$, we write $\widehat{S}_i(D_{n+1}) = \widehat{S}^{y_{n+1}}_i = |y_i - f_\theta(D_{-i}, \mathbf{x}_i)|$, where $D_{-i} = ((\mathbf{x}_j, y_j) : j \in [n+1],\, j \neq i)$.

We first show that $(\widehat{S}_1, \ldots, \widehat{S}_{n+1})$ is exchangeable. Let π be any permutation of $[n+1]$. In the permuted sample $D^\pi_{n+1} = ((\mathbf{x}_{\pi(1)}, y_{\pi(1)}), \ldots, (\mathbf{x}_{\pi(n+1)}, y_{\pi(n+1)}))$, the leave-one-out context associated with $(\mathbf{x}_{\pi(i)}, y_{\pi(i)})$ is a permutation of $D_{-\pi(i)}$. Hence, by the context-order invariance of f_θ, we have $\widehat{S}_i(D^\pi_{n+1}) = \widehat{S}_{\pi(i)}(D_{n+1})$ for every $i \in [n+1]$. Thus, the score map is permutation-equivariant. Since the observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n+1}, y_{n+1})$ are exchangeable, it follows that the scores $\widehat{S}^{y_{n+1}}_1, \ldots, \widehat{S}^{y_{n+1}}_{n+1}$ are exchangeable as well.

Let $k = \lceil (n+1)(1-\alpha) \rceil$. If $k = n+1$, then $\widehat{q}^{y_{n+1}}_\theta = +\infty$ by convention, and the result is immediate. Suppose $k \leq n$. For each $i \in [n+1]$, let Q_{-i} be the kth order statistic of $\{\widehat{S}^{y_{n+1}}_j : j \neq i\}$. For any $n+1$ real numbers, at least k indices satisfy $\widehat{S}_i \leq Q_{-i}$; indeed, this is true for the indices corresponding to the k smallest values. Therefore,

$$\sum_{i=1}^{n+1} \mathbf{1}\{\widehat{S}^{y_{n+1}}_i \leq Q_{-i}\} \geq k \qquad \text{almost surely.}$$

By exchangeability, the events $\{\widehat{S}^{y_{n+1}}_i \leq Q_{-i}\}$ have the same probability for all $i \in [n+1]$. Consequently,

$$\mathbb{P}\{\widehat{S}^{y_{n+1}}_{n+1} \leq Q_{-(n+1)}\} \geq \frac{k}{n+1} \geq 1 - \alpha.$$

Since $Q_{-(n+1)} = \widehat{q}^{y_{n+1}}_\theta$, the event $\{\widehat{S}^{y_{n+1}}_{n+1} \leq Q_{-(n+1)}\}$ is exactly $\{y_{n+1} \in \widehat{\Gamma}_\alpha(\mathbf{x}_{n+1})\}$. This proves the result. □

C LINEAR-SMOOTHER STABILITY CRITERIA AND EXAMPLES

Verification of prompt stability. We verify Assumption 1 using classical linear smoothers. For fixed covariates and response-independent tuning parameters, a linear smoother has the form $\widehat{m}(\mathbf{x}) = \sum_j w_j(\mathbf{x}) y_j$, where the weights do not depend on the responses. Examples include k-nearest-neighbor regression, Nadaraya–Watson kernel regression, local polynomial regression, regression and smoothing splines, and squared-loss RKHS regression; see, for example, Wasserman (2006). We also allow a fixed one-Lipschitz post-processing map, such as clipping. The following proposition reduces prompt stability to bounds on the corresponding smoothing weights.

Proposition 2 (Linear-smoother criterion for prompt stability)**.** *Suppose that, for every $i \in [n]$ and $z \in \mathcal{Y}$,*

$$\widehat{y}_i^z = \phi\left(\sum_{j\in[n+1]\setminus\{i\}} w_{ij} y_j^z\right),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a fixed one-Lipschitz map, including the identity and clip_M. The possibly random weights w_{ij} may depend on the covariates, but not on the responses. If, for deterministic sequences $\kappa_n, \rho_n \to 0$, $\max_{i\in[n]} |w_{i,n+1}| = O_p(\kappa_n)$ and

$$\frac{1}{n(n-1)} \sum_{\substack{i,j\in[n]\\ i\neq j}} \mathbb{E}_{\mathrm{test}}|w_{ij}| = O_p(\rho_n),$$

then the two bounds in Assumption 1 hold with rates κ_n and ρ_n, respectively.

Proof. Because ϕ is one-Lipschitz and $y_j^z = y_j^{y_{n+1}}$ for every $j \in [n]$, $|\widehat{y}_i^z - \widehat{y}_i| \le |w_{i,n+1}|\,|z - y_{n+1}|$. Since $z, y_{n+1} \in \mathcal{Y} = [-M-1, M+1]$,

$$\sup_{z\in\mathcal{Y}} \max_{i\in[n]} |\widehat{y}_i^z - \widehat{y}_i| \le 2(M+1) \max_{i\in[n]} |w_{i,n+1}| = O_p(\kappa_n),$$

which proves part (i).

For distinct $i, j \in [n]$, response-independence of the weights and the one-Lipschitz property of ϕ give $|\widehat{y}_i - \widehat{y}_i^{(j)}| \le |w_{ij}|\,|y_j - y_j'| \le 2|w_{ij}|$, since $y_j - y_j' = \varepsilon_j - \varepsilon_j'$ and $\varepsilon_j, \varepsilon_j' \in [-1, 1]$. Therefore,

$$\frac{1}{n(n-1)} \sum_{\substack{i,j\in[n]\\ i\neq j}} \mathbb{E}_{\mathrm{test}}|\widehat{y}_i - \widehat{y}_i^{(j)}| \le \frac{2}{n(n-1)} \sum_{\substack{i,j\in[n]\\ i\neq j}} \mathbb{E}_{\mathrm{test}}|w_{ij}| = O_p(\rho_n),$$

which proves part (ii). □

Example 1 (Gaussian kernel and ridge regression)**.** Two standard examples illustrate the criterion.

1. *Gaussian kernel regression.* Let $K(\mathbf{u}) = \exp(-\|\mathbf{u}\|_2^2/2)$ and consider the Nadaraya–Watson predictor
$$\widehat{y}_i^z = \frac{\sum_{j\in[n+1]\setminus\{i\}} K((\mathbf{x}_j - \mathbf{x}_i)/h) y_j^z}{\sum_{\ell\in[n+1]\setminus\{i\}} K((\mathbf{x}_\ell - \mathbf{x}_i)/h)}.$$
Its weights are $w_{ij} = K((\mathbf{x}_j - \mathbf{x}_i)/h) / \sum_{\ell\neq i} K((\mathbf{x}_\ell - \mathbf{x}_i)/h)$. Under the design conditions of Section 4.1, if $nh^d/\log n \to \infty$, a conditional Bernstein bound followed by a union bound over $i \in [n]$ implies that the denominators are uniformly bounded below by cnh^d with probability tending to one, for some $c > 0$. Since $0 < K \le 1$, $\max_i w_{i,n+1} = O_p((nh^d)^{-1})$, so one may take $\kappa_n = (nh^d)^{-1}$.

   Moreover, $w_{ij} \ge 0$ and $\sum_{j\in[n+1]\setminus\{i\}} w_{ij} = 1$. Hence $(n(n-1))^{-1} \sum_{i\neq j,\ i,j\in[n]} \mathbb{E}_{\mathrm{test}}|w_{ij}| \le (n-1)^{-1}$, so one may take $\rho_n = n^{-1}$.

2. *Ridge regression.* For a fixed $\lambda > 0$, let X_{-i} have rows $\mathbf{x}_j^\top$ and let $\mathbf{y}_{-i}^z = (y_j^z)_{j\neq i}$, with $j \in [n+1] \setminus \{i\}$ in their original order. Consider the normalized ridge predictor
$$\widehat{y}_i^z = \mathbf{x}_i^\top \left(\frac{1}{n} X_{-i}^\top X_{-i} + \lambda I\right)^{-1} \frac{1}{n} X_{-i}^\top \mathbf{y}_{-i}^z.$$

Its weights are $w_{ij} = n^{-1}\mathbf{x}_i^\top(n^{-1}X_{-i}^\top X_{-i} + \lambda I)^{-1}\mathbf{x}_j$. Since $\|(n^{-1}X_{-i}^\top X_{-i} + \lambda I)^{-1}\|_{\mathrm{op}} \le \lambda^{-1}$ and $\|\mathbf{x}_i\|_2, \|\mathbf{x}_j\|_2 \le \sqrt{d}$, we have $|w_{ij}| \le d/(n\lambda)$. Thus one may take $\kappa_n = \rho_n = n^{-1}$.

The same rates hold after applying clip_M, since clipping is one-Lipschitz.

C.1 Gaussian local-polynomial stability

Lemma 1 (Uniform conditioning of the Gaussian local Gram matrix)**.** *Let $D_n = ((\mathbf{x}_i, y_i))_{i=1}^n$, where $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d. on $[0,1]^d$ with Lebesgue density $f_{\mathbf{x}}$ satisfying $0 < c_{\mathbf{x}} \le f_{\mathbf{x}} \le C_{\mathbf{x}} < \infty$ a.e. Fix $p \in \mathbb{N}_0$, let $\mathcal{A}_p := \{\boldsymbol{\alpha} \in \mathbb{N}_0^d : |\boldsymbol{\alpha}|_1 \le p\}$, $J_p := |\mathcal{A}_p|$, and $\boldsymbol{\psi}_p(\mathbf{u}) := (\mathbf{u}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathcal{A}_p}$. Let $K(\mathbf{u}) := \exp(-\|\mathbf{u}\|_2^2/2)$, and for $h \in (0,1]$ write $\mathbf{u}_i(\mathbf{x}) := (\mathbf{x}_i - \mathbf{x})/h$. Recall that $\mathbf{G}_h(D_n, \mathbf{x})$, defined in Section 4.3, depends only on the covariates.*

There exist constants $c, C, \lambda, \Lambda > 0$, depending only on $d, p, c_{\mathbf{x}}, C_{\mathbf{x}}$, such that on an event $\mathcal{U}_{n,h}$, simultaneously for all $\mathbf{x} \in [0,1]^d$,

$$\lambda \mathbf{I}_{J_p} \preceq \mathbf{G}_h(D_n, \mathbf{x}) \preceq \Lambda \mathbf{I}_{J_p}, \qquad \frac{1}{nh^d}\sum_{i=1}^n K(\mathbf{u}_i(\mathbf{x}))\{1 + \|\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))\|_2\} \le C.$$

If $nh^d \ge C\log(e/h)$, then $\mathbb{P}(\mathcal{U}_{n,h}) \ge 1 - C\exp(-cnh^d)$. For arbitrary n and $h \in (0,1]$, the weaker bound $\mathbb{P}(\mathcal{U}_{n,h}) \ge 1 - Ch^{-d(d+1)}\exp(-cnh^d)$ also holds, possibly vacuously.

If $|y_i| \le M + 1$, then on $\mathcal{U}_{n,h}$, $\sup_{\mathbf{x}} \|\mathbf{b}_h(D_n, \mathbf{x})\|_2 \le (M+1)\sqrt{\Lambda}$, where $\mathbf{b}_h(D_n, \mathbf{x}) := \{1 + \sum_i K(\mathbf{u}_i(\mathbf{x}))\}^{-1}\sum_i K(\mathbf{u}_i(\mathbf{x}))\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))y_i$.

Proof. Define $\widehat{\mathbf{A}}_h(\mathbf{x}) := (nh^d)^{-1}\sum_i K(\mathbf{u}_i(\mathbf{x}))\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))^\top$ and $\widehat{s}_h(\mathbf{x}) := (nh^d)^{-1}\{1 + \sum_i K(\mathbf{u}_i(\mathbf{x}))\}$. Then $\mathbf{G}_h(D_n, \mathbf{x}) = \widehat{\mathbf{A}}_h(\mathbf{x})/\widehat{s}_h(\mathbf{x})$.

Let $\mathbf{A}_h(\mathbf{x}) := \mathbb{E}\widehat{\mathbf{A}}_h(\mathbf{x})$ and $s_h(\mathbf{x}) := h^{-d}\mathbb{E}K((\mathbf{x}_1 - \mathbf{x})/h)$. With $\Omega_{\mathbf{x},h} := \{\mathbf{z} : \mathbf{x} + h\mathbf{z} \in [0,1]^d\}$, a change of variables gives

$$\mathbf{A}_h(\mathbf{x}) = \int_{\Omega_{\mathbf{x},h}} K(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})^\top f_{\mathbf{x}}(\mathbf{x} + h\mathbf{z})\, d\mathbf{z},$$
$$s_h(\mathbf{x}) = \int_{\Omega_{\mathbf{x},h}} K(\mathbf{z}) f_{\mathbf{x}}(\mathbf{x} + h\mathbf{z})\, d\mathbf{z}.$$

For $\boldsymbol{\tau} \in \{-1,1\}^d$, let $Q_{\boldsymbol{\tau}} := \prod_{j=1}^d I_{\tau_j}$, where $I_1 = [0, 1/2]$ and $I_{-1} = [-1/2, 0]$. For every $\mathbf{x}$, one can choose $\boldsymbol{\tau}$ so that $Q_{\boldsymbol{\tau}} \subseteq \Omega_{\mathbf{x},h}$, since $h \le 1$. The matrix $\mathbf{M}_{\boldsymbol{\tau}} := \int_{Q_{\boldsymbol{\tau}}} K(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})^\top d\mathbf{z}$ is positive definite: for every nonzero $\mathbf{v}$, $\mathbf{v}^\top\boldsymbol{\psi}_p(\mathbf{z})$ is a nonzero polynomial and hence cannot vanish a.e. on a rectangle with nonempty interior. Thus $a_- := c_{\mathbf{x}}\min_{\boldsymbol{\tau}}\lambda_{\min}(\mathbf{M}_{\boldsymbol{\tau}}) > 0$. Moreover, $a_+ := C_{\mathbf{x}}\lambda_{\max}\{\int_{\mathbb{R}^d} K(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})^\top d\mathbf{z}\} < \infty$. Hence $a_-\mathbf{I}_{J_p} \preceq \mathbf{A}_h(\mathbf{x}) \preceq a_+\mathbf{I}_{J_p}$ uniformly in $\mathbf{x}, h$. The same argument gives constants $0 < s_- \le s_+ < \infty$ such that $s_- \le s_h(\mathbf{x}) \le s_+$ uniformly.

It remains to control the empirical quantities uniformly in $\mathbf{x}$. For $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathcal{A}_p$, let $\varphi_{\boldsymbol{\alpha},\boldsymbol{\gamma}}(\mathbf{z}) := K(\mathbf{z})\mathbf{z}^{\boldsymbol{\alpha}+\boldsymbol{\gamma}}$, $\varphi_0 := K$, and $\varphi_*(\mathbf{z}) := K(\mathbf{z})\{1 + \sum_{\boldsymbol{\alpha}\in\mathcal{A}_p}|\mathbf{z}^{\boldsymbol{\alpha}}|\}$. These finitely many functions are bounded, square integrable, and globally Lipschitz.

For any such φ, set $\widehat{T}_{\varphi,h}(\mathbf{x}) := n^{-1}\sum_i h^{-d}\varphi((\mathbf{x}_i - \mathbf{x})/h)$ and $T_{\varphi,h}(\mathbf{x}) := \mathbb{E}\widehat{T}_{\varphi,h}(\mathbf{x})$. The summands are bounded by Ch^{-d} and have variance at most Ch^{-d}, so Bernstein's inequality gives, for fixed $t \in (0,1]$,

$$\mathbb{P}\Big(|\widehat{T}_{\varphi,h}(\mathbf{x}) - T_{\varphi,h}(\mathbf{x})| > t\Big) \le 2\exp(-cnh^d t^2).$$

Moreover, $|\widehat{T}_{\varphi,h}(\mathbf{x}) - \widehat{T}_{\varphi,h}(\mathbf{x}')| + |T_{\varphi,h}(\mathbf{x}) - T_{\varphi,h}(\mathbf{x}')| \le Ch^{-d-1}\|\mathbf{x} - \mathbf{x}'\|_2$.

Set $t_* := \min\{1, a_-/(8J_p), s_-/8\}$, and let $\mathcal{N}_h$ be an r_h-net of $[0,1]^d$ with $r_h := t_* h^{d+1}/(4C)$. Then $|\mathcal{N}_h| \le Ch^{-d(d+1)}$. Applying Bernstein's inequality at threshold $t_*/2$, followed by a

union bound over $\mathcal{N}_h$ and the fixed collection of functions, gives, with probability at least $1 - Ch^{-d(d+1)} \exp(-cnh^d)$,

$$\sup_{\mathbf{x}\in[0,1]^d} \max_{\varphi} |\widehat{T}_{\varphi,h}(\mathbf{x}) - T_{\varphi,h}(\mathbf{x})| \le t_*.$$

Here the Lipschitz bound extends the estimate from the net to all $\mathbf{x}$. In particular, $\sup_{\mathbf{x}} \|\widehat{\mathbf{A}}_h(\mathbf{x}) - \mathbf{A}_h(\mathbf{x})\|_{\mathrm{op}} \le J_p t_* \le a_-/8$, and $\sup_{\mathbf{x}} (nh^d)^{-1} \sum_i K(\mathbf{u}_i(\mathbf{x}))\{1 + \|\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))\|_2\} \le C$.

Weyl's inequality therefore yields $(a_-/2)\mathbf{I}_{J_p} \preceq \widehat{\mathbf{A}}_h(\mathbf{x}) \preceq 2a_+\mathbf{I}_{J_p}$. Similarly, with $\overline{s}_h(\mathbf{x}) := (nh^d)^{-1} \sum_i K(\mathbf{u}_i(\mathbf{x}))$, we have $s_-/2 \le \overline{s}_h(\mathbf{x}) \le 2s_+$. If $nh^d \ge 1$, then $s_-/2 \le \widehat{s}_h(\mathbf{x}) \le 2s_+ + 1$, and hence

$$\frac{a_-}{2(2s_+ + 1)} \mathbf{I}_{J_p} \preceq \mathbf{G}_h(D_n, \mathbf{x}) \preceq \frac{4a_+}{s_-} \mathbf{I}_{J_p}$$

uniformly in $\mathbf{x}$. Thus one may take $\lambda := a_-/\{2(2s_+ + 1)\}$ and $\Lambda := 4a_+/s_-$. The polynomial prefactor in the failure probability is absorbed into the exponential whenever $nh^d \ge C \log(e/h)$.

Finally, write $w_i(\mathbf{x}) := K(\mathbf{u}_i(\mathbf{x}))/\{1 + \sum_j K(\mathbf{u}_j(\mathbf{x}))\}$. For every unit vector $\mathbf{v}$, weighted Cauchy–Schwarz gives

$$|\mathbf{v}^\top \mathbf{b}_h(D_n, \mathbf{x})|^2 \le (M+1)^2 \mathbf{v}^\top \mathbf{G}_h(D_n, \mathbf{x}) \mathbf{v} \le (M+1)^2 \Lambda.$$

Taking the supremum over unit $\mathbf{v}$ proves $\|\mathbf{b}_h(D_n, \mathbf{x})\|_2 \le (M+1)\sqrt{\Lambda}$. □

Lemma 2 (Prompt stability of Gaussian local polynomial regression)**.** *Let $n \ge 2$, and suppose the inference data follow the regression model of Section 3, with covariates satisfying the assumptions of Lemma 1. Consider the clipped Gaussian local polynomial rule $f_{\mathrm{LocPol},h}$ in* (9)*. There exist constants $c, C > 0$ and a covariate-only event $\mathcal{V}_{n,h}$ such that, if $nh^d \ge C\{\log(en) + \log(e/h)\}$, then $\mathbb{P}(\mathcal{V}_{n,h}) \ge 1 - C(n+1)\exp(-cnh^d)$, and on $\mathcal{V}_{n,h}$ all $n+1$ leave-one-out prompt designs simultaneously satisfy*

$$\max_{i\in[n]} |w_{i,n+1,h}| \le \frac{C}{nh^d}, \qquad \frac{1}{n(n-1)} \sum_{\substack{i,j\in[n] \\ i\ne j}} |w_{ij,h}| \le \frac{C}{n}. \tag{14}$$

Here $w_{ij,h}$ is the coefficient of the jth visible response in the unclipped rule when the ith observation is used as the query. Consequently, the exact Gaussian local polynomial rule satisfies Assumption 1 with $\kappa_n = (nh^d)^{-1}$ and $\rho_n = n^{-1}$.

Proof. For a context of size n and query $\mathbf{x}$, the common normalization cancels from the normal equations. Hence, before clipping, the coefficient of y_j is

$$w_{j,h}(\mathbf{x}) = \mathbf{e}_0^\top \widehat{\mathbf{A}}_h(\mathbf{x})^{-1} \frac{K(\mathbf{u}_j(\mathbf{x}))\boldsymbol{\psi}_p(\mathbf{u}_j(\mathbf{x}))}{nh^d}.$$

On the event of Lemma 1, $\|\widehat{\mathbf{A}}_h(\mathbf{x})^{-1}\|_{\mathrm{op}} \le C$ uniformly in $\mathbf{x}$. Therefore $|w_{j,h}(\mathbf{x})| \le C(nh^d)^{-1} K(\mathbf{u}_j(\mathbf{x}))\|\boldsymbol{\psi}_p(\mathbf{u}_j(\mathbf{x}))\|_2$. Since a Gaussian times a fixed polynomial is uniformly bounded, $\max_j |w_{j,h}(\mathbf{x})| \le C(nh^d)^{-1}$, while the weighted-moment bound in Lemma 1 gives $\sum_j |w_{j,h}(\mathbf{x})| \le C$. Clipping is one-Lipschitz, so the same bounds control perturbations of the clipped prediction.

Apply Lemma 1 to each of the $n+1$ covariate subsamples obtained by leaving out one observation, and let $\mathcal{V}_{n,h}$ be the intersection of their good-design events. The bandwidth condition implies $nh^d \ge C \log(e/h)$, so a union bound gives a simultaneous failure probability at most $C(n+1)\exp(-cnh^d)$. On $\mathcal{V}_{n,h}$, (14) follows from the preceding individual and total-weight bounds.

For candidate-response perturbations, $|\widehat{y}_i^z - \widehat{y}_i| \le |w_{i,n+1,h}|\,|z - y_{n+1}|$, and $|z - y_{n+1}| \le 2(M+1)$; hence part (i) of Assumption 1 holds with $\kappa_n = (nh^d)^{-1}$. For distinct $i, j \in [n]$, replacing y_j by $y'_j = m(\mathbf{x}_j) + \varepsilon'_j$ gives $|y_j - y'_j| = |\varepsilon_j - \varepsilon'_j| \le 2$. Thus the second bound in (14) gives an average replace-one change of order n^{-1} on the simultaneous good event.

After enlarging the constant in the bandwidth condition, $C(n+1)\exp(-cnh^d) = O(n^{-1})$. On the complement, clipping bounds every prediction difference by $2M$. Averaging over the inference law therefore gives part (ii) of Assumption 1 with $\rho_n = n^{-1}$. □

C.2 Softmax Transformer stability and conformal efficiency

We transfer the stability of Gaussian local polynomial regression to the constructed softmax Transformer.

Proposition 3 (Prompt stability of the softmax Transformer)**.** *Let $f_{h,\varepsilon}$ be the predictor constructed in Theorem 3 with the common conditioning constants fixed in Section 4. If $nh^d \geq C\{\log(en) + \log(e/h)\}$, then*

$$\sup_{z\in\mathcal{Y}} \max_{i\in[n]} |\widehat{y}_i^z - \widehat{y}_i| = O_p\big((nh^d)^{-1} + \varepsilon\big)\,, \tag{15}$$

and

$$\frac{1}{n(n-1)} \sum_{i\neq j} \mathbb{E}_{\text{test}} |\widehat{y}_i - \widehat{y}_i^{(j)}| = O(n^{-1} + \varepsilon), \tag{16}$$

where the predictions are formed using $f_{h,\varepsilon}$.

Proof. Let $f_h := f_{\text{LocPol},h}$, and let $\mathcal{V}_{n,h}$ be the simultaneous good-design event in Lemma 2. Write $D_{-i} := D_{-i}^{y_{n+1}}$, $\widehat{y}_i^z := f_{h,\varepsilon}(D_{-i}^z, \mathbf{x}_i)$, and $\widehat{y}_i := f_{h,\varepsilon}(D_{-i}, \mathbf{x}_i)$. Define $\widehat{y}_i^{(j)}$ by replacing y_j with its independent copy in D_{-i}.

Since $\mathcal{V}_{n,h}$ depends only on the covariates, it is unchanged by these response perturbations. On $\mathcal{V}_{n,h}$, uniform approximation gives

$$|\widehat{y}_i^z - \widehat{y}_i| \leq 2\varepsilon + |f_h(D_{-i}^z, \mathbf{x}_i) - f_h(D_{-i}, \mathbf{x}_i)|.$$

Lemma 2 bounds the last term uniformly by $C(nh^d)^{-1}$. Since $\mathbb{P}(\mathcal{V}_{n,h}^c) \to 0$, this proves (15).

Similarly, on $\mathcal{V}_{n,h}$,

$$|\widehat{y}_i - \widehat{y}_i^{(j)}| \leq 2\varepsilon + |f_h(D_{-i}, \mathbf{x}_i) - f_h(D_{-i}^{(j)}, \mathbf{x}_i)|.$$

Averaging over $i \neq j$ and applying the same lemma gives $O(n^{-1}+\varepsilon)$. On $\mathcal{V}_{n,h}^c$, clipping bounds each difference by $2M$. The bandwidth condition, with a sufficiently large constant, ensures $\mathbb{P}(\mathcal{V}_{n,h}^c) = O(n^{-1})$. Taking expectations proves (16). □

Corollary 2 (Conformal efficiency of the softmax Transformer)**.** *Assume the model of Section 4.1 and the noise-regularity conditions of Theorem 2. Let $h_n = n^{-1/(2\beta+d)}$, $\tau_n = (en)^{-1}$, and $f_n := f_{h_n,\tau_n}$. Define $r_n := R(f_n) - \sigma^2$, and let κ_n, ρ_n denote the two stability rates of f_n. Then $r_n, \kappa_n = O(n^{-2\beta/(2\beta+d)})$ and $\rho_n = O(n^{-1})$. The conformal set $\widehat{\Gamma}_\alpha$ and its length $\widehat{\nu}_\alpha$, formed using f_n, satisfy*

$$|\widehat{\nu}_\alpha(\mathbf{x}_{n+1}) - 2q_\alpha| = O_p\Big(n^{-1/2} + n^{-2\beta/(2\beta+d)}\Big)$$

and

$$\text{Leb}\Big(\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \triangle \Gamma_\alpha^*(m, \mathbf{x}_{n+1})\Big) = O_p\Big(n^{-1/2} + n^{-\beta/(2\beta+d)}\Big)\,.$$

Proof. At the chosen bandwidth, $h_n^{2\beta} = (nh_n^d)^{-1} = n^{-2\beta/(2\beta+d)}$. Lemma 5 and Theorem 3 yield $r_n = O(n^{-2\beta/(2\beta+d)})$. Proposition 3 gives $\kappa_n = O(n^{-2\beta/(2\beta+d)})$ and $\rho_n = O(n^{-1})$. Applying the oracle-approximation argument of Theorem 2 to the fixed predictor f_n proves the claim. □

D Proofs for conformal-set efficiency

This appendix proves Proposition 1, establishes the leave-one-out concentration bound used for dependent calibration scores, and proves Theorem 2.

Throughout this appendix, for $t \geq 0$ and $u \in \mathbb{R}$, write $g_t(u) := \mathbb{P}(|\varepsilon - u| \leq t)$ and $H(t) := g_t(0) = \mathbb{P}(|\varepsilon| \leq t)$.

D.1 Proof of the oracle-width comparison

Proof of Proposition 1. Fix the realized predictor and write $\delta = \widehat{f}_T(D_n, \mathbf{x}) - m(\mathbf{x})$. Let $\bar{H}(t) = \mathbb{P}_{\text{test}}(|\varepsilon - \delta| \le t)$. Since the query noise is independent of $(m, D_n, \mathbf{x}, \widehat{f}_T)$, it is independent of δ. By symmetry, ε has density $v \mapsto h_\varepsilon(|v|)/2$, so $g_t(u) = \frac{1}{2}\int_{u-t}^{u+t} h_\varepsilon(|v|)\,dv$ and $\bar{H}(t) = \mathbb{E}_{\text{test}}[g_t(\delta)]$.

For almost every u,

$$|g_t'(u)| \le \frac{L_\varepsilon}{2}\big||t+u| - |t-u|\big| \le L_\varepsilon |u|.$$

Hence $|g_t(u) - g_t(0)| \le (L_\varepsilon/2)u^2$, and therefore

$$\sup_{t \ge 0} |\bar{H}(t) - H(t)| \le \frac{L_\varepsilon}{2}\mathbb{E}_{\text{test}}(\delta^2) = \frac{L_\varepsilon}{2} r_{n,T}.$$

Since H is continuous, $H(q_\alpha) = 1 - \alpha$. For any $b \in (L_\varepsilon r_{n,T}/(2c_\alpha), \eta_\alpha)$, the lower bound on h_ε and the mean-value theorem give $H(q_\alpha - b) \le 1 - \alpha - c_\alpha b$ and $H(q_\alpha + b) \ge 1 - \alpha + c_\alpha b$. The preceding uniform bound therefore implies $\bar{H}(q_\alpha - b) < 1 - \alpha < \bar{H}(q_\alpha + b)$, so $q_\alpha - b < \bar{q}_\alpha \le q_\alpha + b$. Letting $b \downarrow L_\varepsilon r_{n,T}/(2c_\alpha)$ yields

$$|\bar{q}_\alpha - q_\alpha| \le \frac{L_\varepsilon}{2c_\alpha} r_{n,T},$$

and multiplying by 2 proves the claim. □

D.2 Concentration of true-candidate leave-one-out scores

The next lemma controls the empirical fluctuation of the leave-one-out scores at the true candidate. If these scores were independent, the usual variance calculation would give an $n^{-1/2}$ rate. Here they are dependent because their leave-one-out prompts overlap. The idea is to replace one context response by an independent copy, thereby decoupling a pair of scores, and then use the average replace-one stability condition to control the resulting pairwise covariances. This yields the additional dependence term ρ_n.

At the true candidate $z = y_{n+1}$, write $D_{-i} := D_{-i}^{y_{n+1}}$ and $\delta_i := \widehat{y}_i - m(\mathbf{x}_i)$ for $i \in [n]$. Then $\widehat{S}_i := \widehat{S}_i^{y_{n+1}} = |\varepsilon_i - \delta_i|$, where δ_i is independent of ε_i.

Lemma 3 (Pointwise leave-one-out concentration)**.** *Suppose that part (ii) of Assumption 1 holds. Then, for any deterministic threshold $t \ge 0$, possibly depending on n and T,*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{|\varepsilon_i - \delta_i| \le t\} - g_t(\delta_i)\right] = O_p\Big(\sqrt{n^{-1} + \rho_n}\Big). \tag{17}$$

Proof. Fix $t \ge 0$ and write $Z_i(t) := \mathbf{1}\{|\varepsilon_i - \delta_i| \le t\} - g_t(\delta_i)$. We first show that each $Z_i(t)$ is centered. We then control $\mathrm{Cov}_{\text{test}}\{Z_i(t), Z_j(t)\}$ by replacing the ith context response in the prompt for the jth prediction with an independent copy. The variance bound then follows from Assumption 1. All expectations, covariances, and variances below are taken under the test-side law, with the convention from Section 3.1 that auxiliary replacement noises are also integrated out whenever they appear.

Conditional on the task function, the covariates, and all inference noises except ε_i, both t and δ_i are fixed, while ε_i retains distribution P_ε. Hence $\mathbb{E}_{\text{test}} Z_i(t) = 0$.

Fix $i \ne j$, let $\delta_j^{(i)} := \widehat{y}_j^{(i)} - m(\mathbf{x}_j)$, and define $Z_j^{(i)}(t) := \mathbf{1}\{|\varepsilon_j - \delta_j^{(i)}| \le t\} - g_t(\delta_j^{(i)})$. Because the response y_i in the prompt for $\widehat{y}_j$ has been replaced by the independent copy y_i', $Z_j^{(i)}(t)$ does not depend on ε_i. Therefore, averaging first over ε_i gives $\mathbb{E}_{\text{test}}\{Z_i(t)Z_j^{(i)}(t)\} = 0$. Since $\mathbb{E}_{\text{test}} Z_i(t) = \mathbb{E}_{\text{test}} Z_j(t) = 0$ and $|Z_i(t)| \le 1$,

$$|\mathrm{Cov}_{\text{test}}\{Z_i(t), Z_j(t)\}| \le \mathbb{E}_{\text{test}}|Z_j(t) - Z_j^{(i)}(t)|.$$

For any $a, b \in \mathbb{R}$, the symmetric difference of $[a-t, a+t]$ and $[b-t, b+t]$ has Lebesgue measure at most $2|a-b|$. Since $f_\varepsilon \le C_\varepsilon$, $|g_t(a) - g_t(b)| \le 2C_\varepsilon|a-b|$, and likewise

$$\mathbb{E}\,|\mathbf{1}\{|\varepsilon - a| \le t\} - \mathbf{1}\{|\varepsilon - b| \le t\}| \le 2C_\varepsilon|a-b|.$$

Moreover, both δ_j and $\delta_j^{(i)}$ are independent of ε_j, because the corresponding leave-one-out prompts omit y_j. Conditioning on these two shifts therefore yields

$$|\mathrm{Cov}_{\mathrm{test}}\{Z_i(t), Z_j(t)\}| \le 4C_\varepsilon \mathbb{E}_{\mathrm{test}}|\widehat{y}_j - \widehat{y}_j^{(i)}|.$$

Relabeling the ordered pairs (i, j) and applying part (ii) of Assumption 1 gives

$$\frac{1}{n(n-1)} \sum_{i \ne j} |\mathrm{Cov}_{\mathrm{test}}\{Z_i(t), Z_j(t)\}| = O_p(\rho_n). \tag{18}$$

Let $A_n := n^{-1} \sum_{i=1}^n Z_i(t)$. Since $|Z_i(t)| \le 1$,

$$\begin{aligned}\mathrm{Var}_{\mathrm{test}}(A_n) &\le \frac{1}{n} + \frac{1}{n^2} \sum_{i \ne j} |\mathrm{Cov}_{\mathrm{test}}\{Z_i(t), Z_j(t)\}| \\ &= O_p(n^{-1} + \rho_n),\end{aligned}$$

where the remaining randomness is due to pretraining. Thus $\mathrm{Var}_{\mathrm{test}}(A_n)/(n^{-1} + \rho_n) = O_p(1)$. For every $\eta > 0$, there is $C_\eta < \infty$ such that this ratio is at most C_η, except on an event of probability at most η, for all sufficiently large n along $T = T_n$. Conditional Chebyshev's inequality and averaging over pretraining then give, for every $A > 0$,

$$\mathbb{P}\Big\{|A_n| > A\sqrt{n^{-1} + \rho_n}\Big\} \le \eta + \frac{C_\eta}{A^2}.$$

Letting first $A \to \infty$ and then $\eta \downarrow 0$ proves (17). □

D.3 Proof of the oracle-approximation theorem

Proof of Theorem 2. The proof has three steps. First, we show that the empirical cutoff at the true candidate $z = y_{n+1}$ is close to the noise quantile q_α. This follows by combining the leave-one-out concentration lemma with the second-order perturbation bound for the residual distribution. Second, uniform candidate stability transfers this cutoff approximation from the true candidate to all $z \in \mathcal{Y}$. Finally, the resulting uniform cutoff bound yields a deterministic sandwich for the conformal set, from which the width and symmetric-difference bounds follow.

Throughout this proof, the predictor f_θ in (1) is the realized pretrained predictor $\widehat{f}_T$, and we write $\widehat{q}^z := \widehat{q}^z_\theta$ for brevity. As specified in Section 3.1, $\widehat{\Gamma}_\alpha$ denotes the prediction set in (1) restricted to $z \in \mathcal{Y}$.

Recall from Section 3.1 that, at the true candidate $z = y_{n+1}$, we write $D_{-i} = D_{-i}^{y_{n+1}}$, $\widehat{y}_i = \widehat{f}_T(D_{-i}, \mathbf{x}_i)$, $\delta_i = \widehat{y}_i - m(\mathbf{x}_i)$, and $\widehat{S}_i = |\varepsilon_i - \delta_i|$ for $i \in [n]$. Recall also that $q_\alpha = Q_{1-\alpha}(|\varepsilon|)$, $\mathcal{Y} = [-M-1, M+1]$, and $\Gamma^*_\alpha(m, \mathbf{x}) = [m(\mathbf{x}) - q_\alpha, m(\mathbf{x}) + q_\alpha]$.

Recall g_t and H, and define $\widehat{H}_n(t) := n^{-1} \sum_{i=1}^n \mathbf{1}\{\widehat{S}_i \le t\}$. The argument in the proof of Proposition 1 gives

$$\sup_{t \ge 0} |g_t(u) - H(t)| \le \frac{L_\varepsilon}{2} u^2, \qquad u \in \mathbb{R}. \tag{19}$$

We first control the average prediction error in the true-candidate leave-one-out prompts. Conditional on the realized predictor $\widehat{f}_T$, exchangeability of the inference observations and context-order invariance imply, for each $i \in [n]$,

$$\widehat{f}_T(D_{-i}, \mathbf{x}_i) - m(\mathbf{x}_i) \overset{d}{=} \widehat{f}_T(D_n, \mathbf{x}_{n+1}) - m(\mathbf{x}_{n+1})$$

under $\mathbb{P}_{\mathrm{test}}$. Hence, $\mathbb{E}_{\mathrm{test}}(n^{-1} \sum_{i=1}^n \delta_i^2) = r_{n,T}$.

Since $r_{n,T} = O_p(a_{n,T})$, for every $\eta > 0$ there exists $K_\eta < \infty$ such that $\mathbb{P}\{r_{n,T} > K_\eta a_{n,T}\} \le \eta$ for all sufficiently large n. On the event $\{r_{n,T} \le K_\eta a_{n,T}\}$, Markov's inequality under $\mathbb{P}_{\text{test}}$ gives, for every $A > 0$,

$$\mathbb{P}_{\text{test}}\left\{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2 > A a_{n,T}\right\} \le \frac{K_\eta}{A}.$$

Letting first $A \to \infty$ and then $\eta \downarrow 0$ yields

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i^2 = O_p(a_{n,T}). \tag{20}$$

We next control the cutoff at the true candidate. Let $k_n := \lceil (n+1)(1-\alpha) \rceil$ and $p_n := k_n/n$. For all sufficiently large n, $k_n \le n$, $p_n = 1 - \alpha + O(n^{-1})$, and $\widehat{q}^{\,y_{n+1}}$ is the p_n-quantile of $\widehat{H}_n$. Set $s_n := \sqrt{n^{-1} + \rho_n}$.

For any threshold t to which Lemma 3 applies,

$$\begin{aligned}\widehat{H}_n(t) - H(t) = \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}\{\widehat{S}_i \le t\} - g_t(\delta_i)\Big] \\ + \frac{1}{n}\sum_{i=1}^{n}\left[g_t(\delta_i) - H(t)\right].\end{aligned}$$

The first term is $O_p(s_n)$ by Lemma 3, while (19) and (20) give

$$\left|\frac{1}{n}\sum_{i=1}^{n}\{g_t(\delta_i) - H(t)\}\right| \le \frac{L_\varepsilon}{2n}\sum_{i=1}^{n}\delta_i^2 = O_p(a_{n,T}).$$

Therefore,

$$\widehat{H}_n(t) - H(t) = O_p(s_n + a_{n,T}). \tag{21}$$

Fix $\eta > 0$. For a sufficiently large constant $C_\eta > 0$, let $b_{n,T} := C_\eta(s_n + a_{n,T})$ and $t_\pm := q_\alpha \pm b_{n,T}$. Since $s_n + a_{n,T} \to 0$, for all sufficiently large n, $0 < t_- < q_\alpha < t_+$ and $b_{n,T} < \eta_\alpha$. The local lower bound on h_ε therefore gives $H(t_-) \le 1 - \alpha - c_\alpha b_{n,T}$ and $H(t_+) \ge 1 - \alpha + c_\alpha b_{n,T}$.

The bounds leading to (21) have constants independent of the admissible threshold. Applying them at t_- and t_+, and using $p_n = 1 - \alpha + O(n^{-1})$ and $n^{-1} = O(s_n)$, we may choose C_η sufficiently large so that

$$\mathbb{P}\{\widehat{H}_n(t_-) < p_n \le \widehat{H}_n(t_+)\} \ge 1 - \eta$$

for all sufficiently large n. Consequently, with probability at least $1 - \eta$, $t_- < \widehat{q}^{\,y_{n+1}} \le t_+$, and therefore

$$|\widehat{q}^{\,y_{n+1}} - q_\alpha| = O_p(s_n + a_{n,T}). \tag{22}$$

We now extend the cutoff bound uniformly over $z \in \mathcal{Y}$. For $i \in [n]$, since $y_i^z = y_i$, the reverse triangle inequality gives $|\widehat{S}_i^z - \widehat{S}_i| \le |\widehat{y}_i^z - \widehat{y}_i|$. Since an order statistic is 1-Lipschitz with respect to the maximum coordinate distance, part (i) of Assumption 1 implies

$$\sup_{z\in\mathcal{Y}}|\widehat{q}^{\,z} - \widehat{q}^{\,y_{n+1}}| = O_p(\kappa_n).$$

Combining this with (22), define

$$e_n := \sup_{z\in\mathcal{Y}}|\widehat{q}^{\,z} - q_\alpha| = O_p(s_n + a_{n,T} + \kappa_n). \tag{23}$$

In particular, $e_n = o_p(1)$.

It remains to translate the cutoff bound into a set approximation. Let $\widehat{\mu}_T := \widehat{f}_T(D_n, \mathbf{x}_{n+1})$. Since $D^z_{-(n+1)} = D_n$, $\widehat{S}^z_{n+1} = |z - \widehat{\mu}_T|$. On the event $e_n < q_\alpha$, we have $q_\alpha - e_n \le \widehat{q}^{\,z} \le q_\alpha + e_n$

uniformly over $z \in \mathcal{Y}$. Moreover, since $\widehat{\mu}_T \in [-M, M]$, $q_\alpha \le 1$, and $\mathcal{Y} = [-M-1, M+1]$, the inner interval below is contained in $\mathcal{Y}$. Thus

$$\begin{aligned}[\widehat{\mu}_T - q_\alpha + e_n,\, \widehat{\mu}_T + q_\alpha - e_n] &\subseteq \widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \\ &\subseteq [\widehat{\mu}_T - q_\alpha - e_n,\, \widehat{\mu}_T + q_\alpha + e_n] \cap \mathcal{Y}.\end{aligned} \tag{24}$$

It follows that

$$|\widehat{\nu}_\alpha(\mathbf{x}_{n+1}) - 2q_\alpha| \le 2e_n,$$

which, together with (23), proves the first assertion.

For the second assertion, let $I_T := [\widehat{\mu}_T - q_\alpha, \widehat{\mu}_T + q_\alpha]$. The sandwich (24) gives

$$\mathrm{Leb}\Big\{\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \triangle I_T\Big\} \le 4e_n.$$

Since I_T and $\Gamma^*_\alpha(m, \mathbf{x}_{n+1})$ have the same radius q_α,

$$\mathrm{Leb}\{I_T \triangle \Gamma^*_\alpha(m, \mathbf{x}_{n+1})\} \le 2|\widehat{\mu}_T - m(\mathbf{x}_{n+1})|.$$

By (3), $\mathbb{E}_{\mathrm{test}}|\widehat{\mu}_T - m(\mathbf{x}_{n+1})|^2 = r_{n,T}$. Hence $r_{n,T} = O_p(a_{n,T})$ and the same conditional Markov argument as above imply

$$|\widehat{\mu}_T - m(\mathbf{x}_{n+1})| = O_p(\sqrt{a_{n,T}}).$$

Therefore, by the triangle inequality for symmetric differences,

$$\begin{aligned}&\mathrm{Leb}\Big\{\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \triangle \Gamma^*_\alpha(m, \mathbf{x}_{n+1})\Big\} \\ &\qquad \le 4e_n + 2|\widehat{\mu}_T - m(\mathbf{x}_{n+1})|.\end{aligned}$$

Using (23) and $a_{n,T} \le \sqrt{a_{n,T}}$ for all sufficiently large n yields

$$\mathrm{Leb}\Big\{\widehat{\Gamma}_\alpha(\mathbf{x}_{n+1}) \triangle \Gamma^*_\alpha(m, \mathbf{x}_{n+1})\Big\} = O_p\big(s_n + \sqrt{a_{n,T}} + \kappa_n\big)\,.$$

Recalling that $s_n = \sqrt{n^{-1} + \rho_n}$ completes the proof. □

E STATISTICAL LEARNING PROOFS

For $h \in (0, 1]$, define

$$\widehat{\mathbf{A}}_h(\mathbf{x}) := \frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i(\mathbf{x}))\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))\boldsymbol{\psi}_p(\mathbf{u}_i(\mathbf{x}))^\top$$

and

$$\widehat{s}_h(\mathbf{x}) := \frac{1 + \sum_{i=1}^n K(\mathbf{u}_i(\mathbf{x}))}{nh^d}.$$

Then $\mathbf{G}_h(D_n, \mathbf{x}) = \widehat{\mathbf{A}}_h(\mathbf{x})/\widehat{s}_h(\mathbf{x})$. The following result is a Gaussian-kernel, random-query counterpart of the uniform local-design analysis for compactly supported kernels in Ching et al. (2026). The Gaussian kernel is used here because its normalized exponential weights match standard softmax attention. Moreover, the population-risk analysis below requires conditioning only at an independent random query, rather than uniformly over all query points; this removes the spatial covering factor in the probability bound. Since the kernel and the scope of the result differ, we give a self-contained proof.

Lemma 4 (Gaussian local-design control at a random query)**.** *Let $\mathbf{x}_1, \dots, \mathbf{x}_{n+1}$ be i.i.d. random vectors in $[0,1]^d$ with Lebesgue density $f_{\mathbf{x}}$ satisfying $0 < c_{\mathbf{x}} \le f_{\mathbf{x}} \le C_{\mathbf{x}} < \infty$ Lebesgue-a.e. on $[0,1]^d$. Fix $\beta > 0$, set $p = \lceil \beta \rceil - 1$ and $J_p = \binom{d+p}{p}$, and let $K(\mathbf{u}) = \exp(-\|\mathbf{u}\|_2^2/2)$. There exist constants $0 < \lambda < \Lambda < \infty$ and $C, c > 0$, depending only on $d, \beta, c_{\mathbf{x}}, C_{\mathbf{x}}$, such that, for every $h \in (0, 1]$ with $nh^d \ge 1$, there is an event $\mathcal{E}_h$, depending only on $(\mathbf{x}_1, \dots, \mathbf{x}_{n+1})$, for which*

$$\mathbb{P}(\mathcal{E}_h^c) \le C \exp(-cnh^d). \tag{25}$$

On $\mathcal{E}_h$, with $\mathbf{u}_i = (\mathbf{x}_i - \mathbf{x}_{n+1})/h$,

$$\lambda \mathbf{I}_{J_p} \preceq \mathbf{G}_h(D_n, \mathbf{x}_{n+1}) \preceq \Lambda \mathbf{I}_{J_p}, \qquad \|\widehat{\mathbf{A}}_h(\mathbf{x}_{n+1})^{-1}\|_{\mathrm{op}} \le C, \tag{26}$$

and

$$\frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i)(1 + \|\mathbf{u}_i\|_2^{p+\beta}) \le C, \qquad \frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i)^2 \|\boldsymbol{\psi}_p(\mathbf{u}_i)\|_2^2 \le C. \tag{27}$$

Proof. Fix $\mathbf{x} \in [0,1]^d$, and set

$$\mathbf{A}_h(\mathbf{x}) := \mathbb{E}\{\widehat{\mathbf{A}}_h(\mathbf{x})\}, \qquad s_h(\mathbf{x}) := h^{-d}\mathbb{E}K((\mathbf{x}_1 - \mathbf{x})/h).$$

Writing $\mathbf{t}$ for the original covariate and applying the change of variables $\mathbf{z} = (\mathbf{t} - \mathbf{x})/h$, we obtain

$$\mathbf{A}_h(\mathbf{x}) = \int_{\Omega_{\mathbf{x},h}} K(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})^\top f_{\mathbf{x}}(\mathbf{x} + h\mathbf{z})\, d\mathbf{z},$$

$$s_h(\mathbf{x}) = \int_{\Omega_{\mathbf{x},h}} K(\mathbf{z}) f_{\mathbf{x}}(\mathbf{x} + h\mathbf{z})\, d\mathbf{z},$$

where $\Omega_{\mathbf{x},h} = \prod_{j=1}^d [-x_j/h, (1-x_j)/h]$. For $\boldsymbol{\tau} \in \{-1,1\}^d$, let $Q_{\boldsymbol{\tau}} := \prod_{j=1}^d I_{\tau_j}$, where $I_1 = [0, 1/2]$ and $I_{-1} = [-1/2, 0]$. For every $\mathbf{x}$, one can choose $\boldsymbol{\tau}$ such that $Q_{\boldsymbol{\tau}} \subseteq \Omega_{\mathbf{x},h}$. Therefore,

$$c_* := c_{\mathbf{x}} e^{-d/8} \min_{\boldsymbol{\tau}\in\{-1,1\}^d} \inf_{\|\mathbf{v}\|_2=1} \int_{Q_{\boldsymbol{\tau}}} \{\mathbf{v}^\top \boldsymbol{\psi}_p(\mathbf{z})\}^2\, d\mathbf{z} > 0.$$

Indeed, for every nonzero $\mathbf{v}$, the function $\mathbf{z} \mapsto \mathbf{v}^\top \boldsymbol{\psi}_p(\mathbf{z})$ is a nonzero polynomial and cannot vanish Lebesgue-a.e. on a rectangle with nonempty interior; positivity of c_* then follows from continuity, compactness of the unit sphere, and the finiteness of $\{-1,1\}^d$. Consequently, $\mathbf{A}_h(\mathbf{x}) \succeq c_* \mathbf{I}_{J_p}$. Also,

$$\sup_{\mathbf{x},h} \|\mathbf{A}_h(\mathbf{x})\|_{\mathrm{op}} \le C_{\mathbf{x}} \int_{\mathbb{R}^d} K(\mathbf{z})\|\boldsymbol{\psi}_p(\mathbf{z})\|_2^2\, d\mathbf{z} < \infty.$$

The same argument without the polynomial factor bounds $s_h(\mathbf{x})$ uniformly above and away from zero. Thus, for some $0 < c_0 < C_0 < \infty$,

$$c_0 \mathbf{I}_{J_p} \preceq \mathbf{A}_h(\mathbf{x}) \preceq C_0 \mathbf{I}_{J_p}, \qquad c_0 \le s_h(\mathbf{x}) \le C_0 \tag{28}$$

uniformly over $\mathbf{x} \in [0,1]^d$ and $h \in (0,1]$.

For every fixed real $q \ge 0$, the functions $K(\mathbf{z})(1 + \|\mathbf{z}\|_2^q)$, $K(\mathbf{z})^2\|\boldsymbol{\psi}_p(\mathbf{z})\|_2^2$, and all entries of $K(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})\boldsymbol{\psi}_p(\mathbf{z})^\top$ are bounded and square integrable. Consequently, if ϕ is any one of these finitely many scalar functions, then, uniformly in $\mathbf{x}$ and $h \in (0,1]$,

$$\sup_{\mathbf{z}} |\phi(\mathbf{z})| \le C, \qquad \mathbb{E}\big[\phi((\mathbf{x}_1 - \mathbf{x})/h)^2\big] \le Ch^d.$$

Scalar Bernstein inequalities, applied at fixed thresholds and followed by a union bound over this fixed collection, therefore show that, conditionally on $\mathbf{x}_{n+1} = \mathbf{x}$, outside an event of probability at most $C\exp(-cnh^d)$,

$$\|\widehat{\mathbf{A}}_h(\mathbf{x}) - \mathbf{A}_h(\mathbf{x})\|_{\mathrm{op}} \le \frac{c_0}{2}, \qquad \left|\widehat{s}_h(\mathbf{x}) - \frac{1}{nh^d} - s_h(\mathbf{x})\right| \le \frac{c_0}{2},$$

and both bounds in (27) hold. Here the corresponding population weighted moments are uniformly bounded by the same change-of-variables argument.

Since $nh^d \ge 1$, (28) now gives

$$\frac{c_0}{2}\mathbf{I}_{J_p} \preceq \widehat{\mathbf{A}}_h(\mathbf{x}) \preceq 2C_0 \mathbf{I}_{J_p}, \qquad \frac{c_0}{2} \le \widehat{s}_h(\mathbf{x}) \le 2C_0 + 1.$$

It follows that

$$\frac{c_0}{2(2C_0+1)}\mathbf{I}_{J_p} \preceq \mathbf{G}_h(D_n, \mathbf{x}) \preceq \frac{4C_0}{c_0}\mathbf{I}_{J_p}, \qquad \|\widehat{\mathbf{A}}_h(\mathbf{x})^{-1}\|_{\mathrm{op}} \le \frac{2}{c_0}.$$

The conditional probability bound is uniform in $\mathbf{x}$. Evaluating the event at $\mathbf{x} = \mathbf{x}_{n+1}$ and averaging over $\mathbf{x}_{n+1}$ proves (25) and the remaining conclusions. □

Lemma 5 (Risk of Gaussian local polynomial regression)**.** *Under the nonparametric model of Section 4.1, let $p := \lceil \beta \rceil - 1$. For every $h \in (0,1]$ satisfying $nh^d \ge 1$, the clipped Gaussian local polynomial estimator satisfies*

$$R(f_{\mathrm{LocPol},h}) - \sigma^2 \le C\left\{h^{2\beta} + \frac{1}{nh^d} + \exp(-cnh^d)\right\}, \tag{29}$$

where $C, c > 0$ depend only on $d, \beta, M, c_{\mathbf{x}}, C_{\mathbf{x}}$.

Proof. Condition on a task regression function $m \in \mathcal{H}^\beta([0,1]^d, M)$ and on $\mathbf{x}_{n+1} = \mathbf{x} \in [0,1]^d$. For a multi-index $\boldsymbol{\alpha}$, write $\boldsymbol{\alpha}! := \prod_{j=1}^d \alpha_j!$, and define

$$\mathbf{q}_{m,\mathbf{x}} := \left(\frac{h^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} \partial^{\boldsymbol{\alpha}} m(\mathbf{x}) \right)_{\boldsymbol{\alpha} \in \mathcal{A}_p}.$$

Then $\mathbf{e}_0^\top \mathbf{q}_{m,\mathbf{x}} = m(\mathbf{x})$. The Taylor remainder property of $\mathcal{H}^\beta([0,1]^d, M)$ yields, for $\mathbf{u}_i = (\mathbf{x}_i - \mathbf{x})/h$,

$$\mathrm{rem}_i := m(\mathbf{x}_i) - \mathbf{q}_{m,\mathbf{x}}^\top \boldsymbol{\psi}_p(\mathbf{u}_i), \qquad |\mathrm{rem}_i| \le C h^\beta \|\mathbf{u}_i\|_2^\beta. \tag{30}$$

On the event $\mathcal{E}_h$ from Lemma 4, $\widehat{\mathbf{A}}_h(\mathbf{x})$ is invertible. Because the common normalization in $\mathbf{G}_h$ and $\mathbf{b}_h$ cancels, the untruncated local polynomial coefficient vector is

$$\widehat{\mathbf{q}} = \widehat{\mathbf{A}}_h(\mathbf{x})^{-1} \frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i) \boldsymbol{\psi}_p(\mathbf{u}_i) y_i.$$

Substituting $y_i = m(\mathbf{x}_i) + \varepsilon_i$ and (30) gives, on $\mathcal{E}_h$,

$$\widehat{\mathbf{q}} - \mathbf{q}_{m,\mathbf{x}} = \widehat{\mathbf{A}}_h(\mathbf{x})^{-1} (\mathbf{r}_h + \boldsymbol{\zeta}_h), \tag{31}$$

where

$$\mathbf{r}_h := \frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i) \boldsymbol{\psi}_p(\mathbf{u}_i) \mathrm{rem}_i, \qquad \boldsymbol{\zeta}_h := \frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i) \boldsymbol{\psi}_p(\mathbf{u}_i) \varepsilon_i.$$

Since J_p is fixed and $\|\boldsymbol{\psi}_p(\mathbf{u})\|_2 \le C(1 + \|\mathbf{u}\|_2^p)$, (30) and (27) imply that, on $\mathcal{E}_h$,

$$\|\mathbf{r}_h\|_2 \le C h^\beta \frac{1}{nh^d} \sum_{i=1}^n K(\mathbf{u}_i)(1 + \|\mathbf{u}_i\|_2^{p+\beta}) \le C h^\beta.$$

Conditionally on $(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x})$, independence, $\mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0$, and the independence of ε_i and $\mathbf{x}_i$ give

$$\mathbb{E}_\varepsilon \|\boldsymbol{\zeta}_h\|_2^2 = \frac{\sigma^2}{n^2 h^{2d}} \sum_{i=1}^n K(\mathbf{u}_i)^2 \|\boldsymbol{\psi}_p(\mathbf{u}_i)\|_2^2 \le \frac{C}{nh^d}$$

on $\mathcal{E}_h$. Therefore, (31) and (26) yield, on $\mathcal{E}_h$,

$$\mathbb{E}_\varepsilon \left[\left| \mathbf{e}_0^\top \widehat{\mathbf{q}} - m(\mathbf{x}) \right|^2 \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x} \right] \le C \left(h^{2\beta} + \frac{1}{nh^d} \right).$$

Because $m(\mathbf{x}) \in [-M, M]$, clipping at $[-M, M]$ cannot increase the squared error. On $\mathcal{E}_h^c$, both $f_{\mathrm{LocPol},h}(D_n, \mathbf{x})$ and $m(\mathbf{x})$ lie in $[-M, M]$, so their squared difference is at most $4M^2$. Consequently, Lemma 4 gives

$$\mathbb{E}\left[\{ f_{\mathrm{LocPol},h}(D_n, \mathbf{x}_{n+1}) - m(\mathbf{x}_{n+1}) \}^2 \,\middle|\, m \right] \le C \left\{ h^{2\beta} + \frac{1}{nh^d} + \exp(-cnh^d) \right\}.$$

The bound is uniform over $m \in \mathcal{H}^\beta([0,1]^d, M)$. Averaging over m and using $R(f) - \sigma^2 = \mathbb{E}\{f(D_n, \mathbf{x}_{n+1}) - m(\mathbf{x}_{n+1})\}^2$ proves (29). □

Lemma 6 (Covering number of sparse softmax transformers)**.** *Fix $d_{\mathrm{model}}, d_{\mathrm{ff}}, H \in \mathbb{N}$ and $M > 0$. There exists $C > 0$, depending only on these fixed quantities, such that, for every $L, S \in \mathbb{N}$, $B \ge 1$, and $\delta \in (0, 1]$,*

$$\log \mathcal{N}\left(\mathcal{F}(d_{\mathrm{model}}, d_{\mathrm{ff}}, H, L, B, S, M), \delta, \|\cdot\|_\infty \right) \le CS \left\{ L^2 \log(eB) + \log\left(\frac{e}{\delta}\right) + \log(eL) \right\}. \tag{32}$$

For each fixed prompt length n, the supremum norm is taken over all admissible prompts with $\mathbf{x}_i \in [0,1]^d$ and $|y_i| \le M + 1$. The bound is uniform in n.

Proof. Write $\|\cdot\|_{\max}$ for the entrywise maximum norm. The row-wise softmax satisfies

$$\|\operatorname{softmax}(\mathbf{a}) - \operatorname{softmax}(\mathbf{b})\|_1 \le 2\|\mathbf{a}-\mathbf{b}\|_\infty. \tag{33}$$

Since each softmax row is a probability vector and the dimensions and number of heads are fixed, if $\|\theta\|_\infty \le B$, then the maximal hidden state R_ℓ after ℓ blocks satisfies

$$R_\ell + 1 \le CB^3(R_{\ell-1}+1), \qquad R_0 \le M+1.$$

Hence

$$R_\ell + 1 \le (CB^3)^\ell (M+2), \qquad \log(R_\ell+1) \le C(\ell+1)\log(eB). \tag{34}$$

Consider two parameter vectors θ, θ' with $\|\theta\|_\infty \vee \|\theta'\|_\infty \le B$. Let $\mathbf{Z}_0 = \mathbf{Z}_0'$ be the common embedded input and, for $\ell \in [L]$, let $\mathbf{Z}_\ell, \mathbf{Z}_\ell'$ denote the outputs after ℓ blocks. Set

$$\rho_\ell := \|\mathbf{Z}_\ell\|_{\max} \vee \|\mathbf{Z}_\ell'\|_{\max}, \qquad d_\ell := \|\mathbf{Z}_\ell - \mathbf{Z}_\ell'\|_{\max}, \qquad \eta := \|\theta - \theta'\|_\infty.$$

For the ℓth attention sublayer,

$$\|\mathbf{Z}_{\ell-1}\mathbf{Q} - \mathbf{Z}_{\ell-1}'\mathbf{Q}'\|_{\max} \le C\{Bd_{\ell-1} + (\rho_{\ell-1}+1)\eta\},$$

and the same bound holds for the key and value projections. Consequently, the corresponding logit matrices satisfy

$$\|\mathbf{S}_\ell - \mathbf{S}_\ell'\|_{\max} \le C\left\{B^2(\rho_{\ell-1}+1)d_{\ell-1} + B(\rho_{\ell-1}+1)^2\eta\right\}.$$

Using (33) and the fact that softmax rows are probability vectors gives

$$\|\operatorname{Attn}_\ell(\mathbf{Z}_{\ell-1}) - \operatorname{Attn}_\ell'(\mathbf{Z}_{\ell-1}')\|_{\max} \le CB^3(\rho_{\ell-1}+1)^3(d_{\ell-1}+\eta).$$

The one-Lipschitz property of ReLU and a direct expansion of the two FFN matrix products therefore yield

$$d_\ell \le CB^6(\rho_{\ell-1}+1)^4(d_{\ell-1}+\eta), \qquad d_0 = 0. \tag{35}$$

Combining (34) and (35) and iterating over $\ell \in [L]$ gives

$$d_L \le \eta \exp\{CL^2\log(eB)\}.$$

Coordinate projection and clipping are one-Lipschitz, so

$$\sup_{(D_n,\mathbf{x})} |f_\theta(D_n,\mathbf{x}) - f_{\theta'}(D_n,\mathbf{x})| \le \Delta_{L,B}\|\theta-\theta'\|_\infty, \qquad \log\Delta_{L,B} \le CL^2\log(eB), \tag{36}$$

where $\Delta_{L,B} \ge 1$. By (33), the constant $\Delta_{L,B}$ is independent of the prompt length n.

Let P_L denote the total number of scalar trainable parameter slots. Since $d_{\text{model}}, d_{\text{ff}}$, and H are fixed, $P_L \le CL$. Set $s = S \wedge P_L$. Discretizing each active coordinate in $[-B, B]$ with mesh $\delta/\Delta_{L,B}$ and summing over supports gives

$$\mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty) \le \left(\frac{eP_L}{s}\right)^s \left(1 + \frac{2B\Delta_{L,B}}{\delta}\right)^s.$$

Taking logarithms and using $P_L \le CL$ and (36) proves (32). □

Lemma 7 (High-probability ERM oracle inequality)**.** *Let $\{(D_n^{(t)}, \mathbf{x}_{n+1}^{(t)}, y_{n+1}^{(t)})\}_{t=1}^T$ be independent copies of $(D_n, \mathbf{x}_{n+1}, y_{n+1})$, with $|y_{n+1}| \le M+1$. Let $\mathcal{F}$ be a class of functions bounded by M, and suppose that $\widehat{f}_T \in \arg\min_{f\in\mathcal{F}} \widehat{R}_T(f)$. Define $f_0(D_n, \mathbf{x}_{n+1}) := \mathbb{E}(y_{n+1} \mid D_n, \mathbf{x}_{n+1})$. Then, for every deterministic $f^\star \in \mathcal{F}$ and $\gamma \ge 1$, with probability at least $1 - Ce^{-\gamma}$,*

$$R(\widehat{f}_T) - R(f_0) \le 3\{R(f^\star) - R(f_0)\} + C\frac{\log\mathcal{N}(\mathcal{F}, T^{-1}, \|\cdot\|_\infty) + \gamma}{T}. \tag{37}$$

The constant $C > 0$ depends only on M; no assumption that $f_0 \in \mathcal{F}$ is required.

Proof. For $f \in \mathcal{F}$, define the excess loss $g_f := \{y_{n+1} - f(D_n, \mathbf{x}_{n+1})\}^2 - \{y_{n+1} - f_0(D_n, \mathbf{x}_{n+1})\}^2$. Since $\mathbb{E}\{y_{n+1} - f_0(D_n, \mathbf{x}_{n+1}) \mid D_n, \mathbf{x}_{n+1}\} = 0$, we have $\mathbb{E}g_f = R(f) - R(f_0) =$

$\mathbb{E}\{f(D_n, \mathbf{x}_{n+1}) - f_0(D_n, \mathbf{x}_{n+1})\}^2$. Moreover, $g_f = (f_0 - f)(2y_{n+1} - f - f_0)$, where f and f_0 are evaluated at $(D_n, \mathbf{x}_{n+1})$. Since $|y_{n+1}| \vee |f_0| \le M + 1$ and $|f| \le M$,

$$|g_f| \le C, \qquad \mathbb{E}g_f^2 \le C\{R(f) - R(f_0)\}.$$

Let $N_T := \mathcal{N}(\mathcal{F}, T^{-1}, \|\cdot\|_\infty)$. If $N_T = \infty$, the claim is immediate. Otherwise, let $\mathcal{F}_T$ be a T^{-1}-net of $\mathcal{F}$ with $|\mathcal{F}_T| = N_T$. Bernstein's inequality gives, for every fixed $f \in \mathcal{F}_T$ and $u \ge 1$, with probability at least $1 - 2e^{-u}$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} g_f(D_n^{(t)}, \mathbf{x}_{n+1}^{(t)}, y_{n+1}^{(t)}) - \mathbb{E}g_f \right| \le C\sqrt{\frac{\{R(f) - R(f_0)\}u}{T}} + C\frac{u}{T}.$$

Using $C\sqrt{ab} \le a/4 + Cb$, taking $u = \log N_T + \gamma$, and applying a union bound over $\mathcal{F}_T$, we obtain, with probability at least $1 - Ce^{-\gamma}$,

$$\left| \{\widehat{R}_T(f) - \widehat{R}_T(f_0)\} - \{R(f) - R(f_0)\} \right| \le \frac{1}{4}\{R(f) - R(f_0)\} + C\frac{\log N_T + \gamma}{T} \tag{38}$$

simultaneously for all $f \in \mathcal{F}_T$.

For arbitrary $f \in \mathcal{F}$, choose $\widetilde{f} \in \mathcal{F}_T$ with $\|f - \widetilde{f}\|_\infty \le T^{-1}$. Since the squared loss is C-Lipschitz in its prediction on the relevant bounded range, $|\widehat{R}_T(f) - \widehat{R}_T(\widetilde{f})| + |R(f) - R(\widetilde{f})| \le C/T$. Combining this with (38) and enlarging C gives, for every $f \in \mathcal{F}$,

$$\left| \{\widehat{R}_T(f) - \widehat{R}_T(f_0)\} - \{R(f) - R(f_0)\} \right| \le \frac{1}{2}\{R(f) - R(f_0)\} + C\frac{\log \mathcal{N}(\mathcal{F}, T^{-1}, \|\cdot\|_\infty) + \gamma}{T}. \tag{39}$$

Since $\widehat{f}_T$ minimizes $\widehat{R}_T$ over $\mathcal{F}$, $\widehat{R}_T(\widehat{f}_T) - \widehat{R}_T(f_0) \le \widehat{R}_T(f^\star) - \widehat{R}_T(f_0)$. Applying the lower side of (39) to $\widehat{f}_T$ and the upper side to $f^\star$ gives

$$\frac{1}{2}\{R(\widehat{f}_T) - R(f_0)\} \le \frac{3}{2}\{R(f^\star) - R(f_0)\} + C\frac{\log \mathcal{N}(\mathcal{F}, T^{-1}, \|\cdot\|_\infty) + \gamma}{T}.$$

Multiplying by two and enlarging C proves (37). □

Proof of Theorem 4. Set $h_n = n^{-1/(2\beta+d)}$ and $\tau_n = (en)^{-1}$.

Theorem 3, with the common conditioning constants fixed in Section 4, gives a deterministic comparator $f_n := f_{h_n,\tau_n} \in \mathcal{F}_n$ for sufficiently large C_0. Indeed, $\log\{e/(h_n\tau_n)\} \le C\log(en)$ and, since $p < \beta$ and $d \ge 1$, $Ch_n^{-\max\{2,2p\}} = Cn^{\max\{2,2p\}/(2\beta+d)} \le Cn^2 \le C_0 n^2$. Identity blocks can be added if necessary to attain depth L_n without increasing the sparsity.

We first bound the risk of f_n. On the event $\mathcal{E}_{h_n}$ from Lemma 4, Theorem 3 gives $|f_n(D_n, \mathbf{x}_{n+1}) - f_{\mathrm{LocPol},h_n}(D_n, \mathbf{x}_{n+1})| \le \tau_n$. On $\mathcal{E}_{h_n}^c$, both predictors are clipped to $[-M, M]$. Hence

$$\mathbb{E}\,|f_n(D_n, \mathbf{x}_{n+1}) - f_{\mathrm{LocPol},h_n}(D_n, \mathbf{x}_{n+1})| \le \tau_n + 2M\mathbb{P}(\mathcal{E}_{h_n}^c) \le \frac{C}{n},$$

since $nh_n^d = n^{2\beta/(2\beta+d)}$ and $\exp(-cnh_n^d) \le C/n$.

As $|y_{n+1}| \le M + 1$, the squared loss is $(4M + 2)$-Lipschitz in its prediction on $[-M, M]$. Thus $|R(f_n) - R(f_{\mathrm{LocPol},h_n})| \le C/n$. Lemma 5 and $h_n^{2\beta} = (nh_n^d)^{-1} = n^{-2\beta/(2\beta+d)}$ yield $R(f_{\mathrm{LocPol},h_n}) - \sigma^2 \le Cn^{-2\beta/(2\beta+d)}$. Since $n^{-1} \le n^{-2\beta/(2\beta+d)}$, we obtain

$$R(f_n) - \sigma^2 \le Cn^{-2\beta/(2\beta+d)}. \tag{40}$$

We next control the pretraining error. Define $f_0(D_n, \mathbf{x}_{n+1}) := \mathbb{E}(y_{n+1} \mid D_n, \mathbf{x}_{n+1})$. By the regression model and the independence of ε_{n+1}, $R(f_0) = \sigma^2 + \mathbb{E}\,\mathrm{Var}(m(\mathbf{x}_{n+1}) \mid D_n, \mathbf{x}_{n+1}) \ge \sigma^2$.

Lemma 7, applied with $f^\star = f_n$, gives, with probability at least $1 - Ce^{-\gamma}$,

$$\begin{aligned} R(\widehat{f}_T) - \sigma^2 &\le 3\{R(f_n) - R(f_0)\} + R(f_0) - \sigma^2 \\ &\quad + C\frac{\log \mathcal{N}(\mathcal{F}_n, T^{-1}, \|\cdot\|_\infty) + \gamma}{T} \\ &\le 3\{R(f_n) - \sigma^2\} + C\frac{\log \mathcal{N}(\mathcal{F}_n, T^{-1}, \|\cdot\|_\infty) + \gamma}{T}, \end{aligned} \tag{41}$$

where the last inequality uses $R(f_0) \geq \sigma^2$.

Finally, Lemma 6 and the defining parameters of $\mathcal{F}_n$ give

$$\log \mathcal{N}(\mathcal{F}_n, T^{-1}, \|\cdot\|_\infty) \leq C\left\{\log^4(en) + \log(en)\log(eT)\right\}. \tag{42}$$

Substituting (40) and (42) into (41) proves (11). □

F SOFTMAX-TRANSFORMER REALIZATION

We first establish the auxiliary approximation ingredients and then prove Theorem 3.

ReLU arithmetic. We use ReLU networks to approximate the polynomial operations in feature construction, moment recovery, and the linear-system solver. Bandwidth-dependent rescalings are implemented separately. For a ReLU network Φ, let $W(\Phi)$, $L(\Phi)$, and $S(\Phi)$ denote its width, depth, and number of nonzero parameters, respectively.

Lemma 8 (Polynomial arithmetic with explicit width)**.** *Let $m, r \in \mathbb{N}$, let $k \geq 2$ be an integer, and let $\mathcal{K} \subset \mathbb{R}^m$ be contained in a compact box. Suppose that $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^r$ is a polynomial map of total degree at most k. There exists $C > 0$, depending only on $m, r, k, \mathbf{F}$, and the containing box, such that, for every $\zeta \in (0, 1/2)$, a ReLU network $\Phi_\zeta : \mathbb{R}^m \to \mathbb{R}^r$ satisfies $\sup_{\mathbf{z}\in\mathcal{K}} \|\Phi_\zeta(\mathbf{z}) - \mathbf{F}(\mathbf{z})\|_\infty \leq \zeta$, $W(\Phi_\zeta) \leq 2(m+r) + 2k + 26$, $L(\Phi_\zeta) \vee S(\Phi_\zeta) \leq C\log(e/\zeta)$, and $\|\theta(\Phi_\zeta)\|_\infty \leq C$. The input coordinates can be retained exactly throughout the construction. Retaining any fixed number q of additional passive coordinates increases the width by at most $2q$.*

Proof. Choose $B \geq 1$ such that $\mathcal{K} \subset [-B, B]^m$. Write $F_j(\mathbf{z}) = \sum_{\boldsymbol{\nu}\in\mathcal{I}_j} a_{j,\boldsymbol{\nu}} \mathbf{z}^{\boldsymbol{\nu}}$ and set $A_* := 1 + \max_{j\in[r]} \sum_{\boldsymbol{\nu}\in\mathcal{I}_j} |a_{j,\boldsymbol{\nu}}|$.

Polynomial approximation by deep ReLU networks is well studied; see, e.g., Lu et al. (2021). By Lemma B.6 of Ching et al. (2026), with $N = 2$, each monomial admits a ReLU approximator with error at most $\delta := \zeta/A_*$, width at most $2k + 26$, depth $O(\log(e/\delta))$, and parameter magnitudes bounded independently of δ. Its sparsity has the same logarithmic bound since its width is fixed.

Evaluate the finitely many monomials sequentially. Retain the m inputs and r partial sums exactly using signed identity channels, $t = \mathrm{ReLU}(t) - \mathrm{ReLU}(-t)$, which require $2(m + r)$ hidden units. Each monomial approximator uses at most $2k + 26$ additional units. Its output is added to the appropriate partial sum through an affine interface, after which the computational workspace is reused. Merging adjacent affine maps preserves the width and bounds the parameters independently of ζ; see Lemma B.2 of Ching et al. (2026). Thus $W(\Phi_\zeta) \leq 2(m + r) + 2k + 26$.

Since the inputs are retained exactly, the approximation error in coordinate j is at most $\sum_{\boldsymbol{\nu}\in\mathcal{I}_j} |a_{j,\boldsymbol{\nu}}|\delta \leq \zeta$. The number of monomials is fixed, so serial composition yields $L(\Phi_\zeta) \vee S(\Phi_\zeta) \leq C\log(e/\zeta)$ and $\|\theta(\Phi_\zeta)\|_\infty \leq C$. Finally, each additional passive coordinate can be retained using two signed identity channels. □

Implementation by residual FFNs. Let $W \in \mathbb{N}$. After zero padding, any ReLU layer $\mathbf{F}(\mathbf{z}) = \mathbf{B}\,\mathrm{ReLU}(\mathbf{A}\mathbf{z} + \mathbf{a}) + \mathbf{b}$ from $\mathbb{R}^W$ to $\mathbb{R}^W$, with hidden width at most W, is represented exactly by a residual FFN of hidden width $3W$. Indeed,

$$\mathbf{F}(\mathbf{z}) = \mathbf{z} + [\mathbf{B}, \quad -\mathbf{I}_W, \quad \mathbf{I}_W]\,\mathrm{ReLU}\left(\begin{bmatrix}\mathbf{A}\\ \mathbf{I}_W\\ -\mathbf{I}_W\end{bmatrix}\mathbf{z} + \begin{bmatrix}\mathbf{a}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}\right) + \mathbf{b}. \tag{43}$$

This follows from the fact $-\mathrm{ReLU}(\mathbf{z}) + \mathrm{ReLU}(-\mathbf{z}) = -\mathbf{z}$. Consequently, after padding and including the final affine output map, a width-W, depth-J ReLU network can be implemented positionwise using at most $J + 1$ Transformer blocks with model dimension W, FFN width $3W$, and identity residual attention sublayers. The identity blocks in (43) add $O(W)$ nonzero parameters per implemented layer. Thus, for the resulting residual implementation,

$$S_{\mathrm{res}} \leq CS(\Phi) + CW(J + 1).$$

The parameter-magnitude bound is preserved up to a fixed factor. For fixed W, these additional parameters preserve logarithmic depth and sparsity.

Raw-moment feature construction. Let q_j be the query indicator in (6), so $\mathbf{x}_{n+1} = \mathbf{x}$, $y_{n+1} = 0$, $q_i = 0$ for $i \in [n]$, and $q_{n+1} = 1$. Define $s_j = \|\mathbf{x}_j\|_2^2$ and $D_{\mathrm{raw}} = |\mathcal{A}_{2p}| + |\mathcal{A}_p| = \binom{d+2p}{2p} + \binom{d+p}{p}$. The required raw feature vector is

$$\mathbf{r}_j = \Big(\big((1-q_j)\mathbf{x}_j^{\boldsymbol{\gamma}}\big)_{\boldsymbol{\gamma}\in\mathcal{A}_{2p}}, \big((1-q_j)y_j\mathbf{x}_j^{\boldsymbol{\gamma}}\big)_{\boldsymbol{\gamma}\in\mathcal{A}_p} \Big) \in \mathbb{R}^{D_{\mathrm{raw}}}.$$

The two groups provide the raw moments needed for $\mathbf{G}_h$ and the response-weighted raw moments needed for $\mathbf{b}_h$, respectively; in particular, $\mathbf{r}_{n+1} = \mathbf{0}$.

Lemma 9 (Tokenwise polynomial feature preparation)**.** *For every $\delta \in (0, 1/2)$, after adding a fixed number of coordinates depending only on d, p, there is a stack of L_{feat} transformer blocks whose output at token j contains $\widetilde{s}_j, \widetilde{\mathbf{r}}_j$ satisfying*

$$\max_{j\in[n+1]} |\widetilde{s}_j - \|\mathbf{x}_j\|_2^2| \le \delta, \qquad \max_{j\in[n+1]} \|\widetilde{\mathbf{r}}_j - \mathbf{r}_j\|_\infty \le \delta. \tag{44}$$

All residual attention sublayers act as identity maps, and the FFN width is bounded by a constant. If S_{feat} and B_{feat} are the sparsity and maximum parameter magnitude, then $L_{\mathrm{feat}} \vee S_{\mathrm{feat}} \le C\log(e/\delta)$ and $B_{\mathrm{feat}} \le C$, where C depends only on d, p, M and is independent of n and δ.

Proof. On $\mathcal{K} = [0,1]^d \times [-M-1, M+1] \times [0,1]$, consider the fixed polynomial map

$$\mathbf{F}(\mathbf{x}, y, q) = \Big(\|\mathbf{x}\|_2^2, \big((1-q)\mathbf{x}^{\boldsymbol{\gamma}}\big)_{\boldsymbol{\gamma}\in\mathcal{A}_{2p}}, \big((1-q)y\mathbf{x}^{\boldsymbol{\gamma}}\big)_{\boldsymbol{\gamma}\in\mathcal{A}_p} \Big).$$

Lemma 8 gives a constant-width ReLU network Φ_δ with uniform error at most δ, depth and sparsity at most $C\log(e/\delta)$, and parameter magnitude at most C. Implement it positionwise using (43), retaining the input and constant coordinates through exact identity channels and setting the value matrices to zero. Every residual attention sublayer is therefore the identity map. At the end of the stack, the retained coordinates are returned to their prescribed token positions, and the moment-aggregation registers are exactly zero. Since the same map is shared across tokens, $(\widetilde{s}_j, \widetilde{\mathbf{r}}_j) = \Phi_\delta(\mathbf{x}_j, y_j, q_j)$ for $j \in [n+1]$. The uniform approximation bound proves (44), and parameter sharing makes the complexity independent of n. □

Gaussian-kernel moment aggregation.

Lemma 10 (Gaussian aggregation of raw polynomial features)**.** *Suppose an attention sublayer receives $\mathbf{x}_j, \widetilde{s}_j, \widetilde{\mathbf{r}}_j$ for $j \in [n+1]$, and that* (44) *holds. For every $h \in (0, 1]$, one square softmax attention head of the form* (7)*, using only $d+2$ active query–key coordinates and D_{raw} active value coordinates, can make the final query-token output contain a fresh-register vector $\widetilde{\mathbf{m}}$ satisfying*

$$\left\| \widetilde{\mathbf{m}} - \frac{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x})/h)\mathbf{r}_i}{1 + \sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x})/h)} \right\|_\infty \le C\delta h^{-2}.$$

The head has fixed sparsity and parameter magnitude at most Ch^{-2}, where C depends only on d, p, M; these bounds are uniform in n.

Proof. Write $d_m = d_{\mathrm{model}}$ and define the active parts of the query and key vectors by

$$\mathbf{q}_h(\mathbf{a}, \widetilde{s}) = \sqrt{d_m}\left(\frac{\mathbf{a}}{h}, -\frac{\widetilde{s}}{2h^2}, 1\right), \qquad \mathbf{k}_h(\mathbf{a}, \widetilde{s}) = \left(\frac{\mathbf{a}}{h}, 1, -\frac{\widetilde{s}}{2h^2}\right).$$

Pad these vectors with zeros to dimension d_m. Because each token contains $\mathbf{a}$, $\widetilde{s}$, and a constant coordinate, they are implemented by square matrices $\mathbf{Q}_h, \mathbf{K}_h$. With exact squared norms, the final query-to-j logit is $\ell_j = -\|\mathbf{x}_j - \mathbf{x}\|_2^2/(2h^2)$ for $j \in [n]$ and $\ell_{n+1} = 0$. Only this final query row is used below; the other attention rows are immaterial. The implemented logits satisfy $\max_j |\widetilde{\ell}_j - \ell_j| \le \delta h^{-2}$.

Let $\mathbf{p}$ and $\widetilde{\mathbf{p}}$ be the corresponding ideal and implemented attention-weight vectors. The softmax map satisfies

$$\|\operatorname{softmax}(\mathbf{a}) - \operatorname{softmax}(\mathbf{b})\|_1 \le 2\|\mathbf{a} - \mathbf{b}\|_\infty. \tag{45}$$

Indeed, if $\mathbf{p}$ is a probability vector, then $\|(\operatorname{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)\mathbf{u}\|_1 \le 2\|\mathbf{u}\|_\infty$; integrating the Jacobian along the segment between $\mathbf{a}$ and $\mathbf{b}$ proves (45). Hence $\|\widetilde{\mathbf{p}} - \mathbf{p}\|_1 \le 2\delta h^{-2}$.

Writing $Z_h = 1 + \sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x})/h)$, we have $p_i = K((\mathbf{x}_i - \mathbf{x})/h)/Z_h$ for $i \in [n]$ and $p_{n+1} = 1/Z_h$. Choose the square value map $\mathbf{V}_h$ to select $\widetilde{\mathbf{r}}_j$ and write the weighted sum into a fresh register. Since $\mathbf{r}_{n+1} = \mathbf{0}$,

$$\sum_{j=1}^{n+1} p_j \mathbf{r}_j = \frac{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x})/h)\mathbf{r}_i}{Z_h}.$$

Using $\max_j \|\widetilde{\mathbf{r}}_j - \mathbf{r}_j\|_\infty \le \delta$ and $\max_j \|\mathbf{r}_j\|_\infty \le M + 1$, we obtain

$$\left\| \widetilde{\mathbf{m}} - \sum_{j=1}^{n+1} p_j \mathbf{r}_j \right\|_\infty \le \delta + (M+1)\|\widetilde{\mathbf{p}} - \mathbf{p}\|_1 \le C\delta h^{-2}.$$

The query and key maps use coefficients of orders $1, h^{-1}, h^{-2}$, whereas the value map is a fixed coordinate selection. This proves the parameter bound. □

Local-polynomial coefficient recovery. For related ReLU approximation results on matrix inversion and linear-system solution, see Kutyniok et al. (2022). We provide a self-contained proof with the explicit width bound needed here.

Lemma 11 (Fixed-dimensional linear-system approximation)**.** *Fix $D \in \mathbb{N}$, $B_0 \ge 0$, and $0 < \lambda < \Lambda < \infty$. Set $W_{\mathrm{sol}} := 6D^2 + 4D + 32$. There exists $C > 0$, depending only on D, B_0, λ, Λ, such that, for every $\varepsilon \in (0, 1/2)$, a ReLU network $\mathcal{S}_\varepsilon : \mathbb{R}^{D^2+D} \to \mathbb{R}^D$ satisfies*

$$\sup_{\substack{\mathbf{G}=\mathbf{G}^\top,\ \lambda\mathbf{I}_D \preceq \mathbf{G} \preceq \Lambda\mathbf{I}_D \\ \|\mathbf{b}\|_2 \le B_0}} \left\| \mathcal{S}_\varepsilon(\mathbf{G}, \mathbf{b}) - \mathbf{G}^{-1}\mathbf{b} \right\|_2 \le \varepsilon. \tag{46}$$

Moreover, $W(\mathcal{S}_\varepsilon) \le W_{\mathrm{sol}}$, $L(\mathcal{S}_\varepsilon) \vee S(\mathcal{S}_\varepsilon) \le C\log(e/\varepsilon)$, and $\|\theta(\mathcal{S}_\varepsilon)\|_\infty \le C$.

Proof. Set $\rho = (\Lambda - \lambda)/(\Lambda + \lambda) \in (0, 1)$ and $\mathbf{X}_0 = 2(\lambda + \Lambda)^{-1}\mathbf{I}_D$. Then $\|\mathbf{I}_D - \mathbf{G}\mathbf{X}_0\|_{\mathrm{op}} \le \rho$. For the Newton–Schulz map $\mathcal{N}(\mathbf{G}, \mathbf{X}) = \mathbf{X}(2\mathbf{I}_D - \mathbf{G}\mathbf{X})$,

$$\mathbf{I}_D - \mathbf{G}\mathcal{N}(\mathbf{G}, \mathbf{X}) = (\mathbf{I}_D - \mathbf{G}\mathbf{X})^2. \tag{47}$$

Choose $q \in (\rho, 1)$, set $\vartheta = \rho/q$, and let $c_* = \min\{1/4, \vartheta(1-\vartheta)/(2\Lambda)\}$. By Lemma 8, for each $\xi \in (0, 1/2)$ there is a constant-width ReLU network $\mathcal{N}_\xi$ satisfying

$$\sup_{\substack{\|\mathbf{G}\|_{\mathrm{op}} \le \Lambda \\ \|\mathbf{X}\|_{\mathrm{op}} \le 2/\lambda}} \|\mathcal{N}_\xi(\mathbf{G}, \mathbf{X}) - \mathcal{N}(\mathbf{G}, \mathbf{X})\|_{\mathrm{op}} \le \xi. \tag{48}$$

Its depth and sparsity are at most $C\log(e/\xi)$, and its parameter magnitude is at most C. Here fixed-dimensional norm equivalence converts coordinatewise approximation into operator-norm approximation.

Similarly, for every $\eta \in (0, 1/2)$, a constant-width ReLU network $\mathcal{M}_\eta$ satisfies

$$\sup_{\substack{\|\mathbf{X}\|_{\mathrm{op}} \le 2/\lambda \\ \|\mathbf{b}\|_2 \le B_0}} \|\mathcal{M}_\eta(\mathbf{X}, \mathbf{b}) - \mathbf{X}\mathbf{b}\|_2 \le \eta, \tag{49}$$

with depth and sparsity at most $C\log(e/\eta)$ and parameter magnitude at most C.

Let $A = (B_0 \vee 1)\vartheta/\lambda$, and let $R \in \mathbb{N}_0$ be minimal such that $Aq^{2^R} \le \varepsilon/2$. For $s = 0, \dots, R-1$, set $\xi_s = c_* q^{2^{s+1}}$, $\widetilde{\mathbf{X}}_0 = \mathbf{X}_0$, and $\widetilde{\mathbf{X}}_{s+1} = \mathcal{N}_{\xi_s}(\mathbf{G}, \widetilde{\mathbf{X}}_s)$. Propagating $(\mathbf{G}, \mathbf{b})$ through identity channels and composing the updates gives a ReLU network; define

$$\mathcal{S}_\varepsilon(\mathbf{G}, \mathbf{b}) = \mathcal{M}_{\varepsilon/2}(\widetilde{\mathbf{X}}_R, \mathbf{b}).$$

We claim that, for every $s = 0, \dots, R$,

$$\|\mathbf{I}_D - \mathbf{G}\widetilde{\mathbf{X}}_s\|_{\mathrm{op}} \le \vartheta q^{2^s}, \qquad \|\widetilde{\mathbf{X}}_s\|_{\mathrm{op}} \le \frac{2}{\lambda}. \tag{50}$$

For $s = 0$, these bounds follow from $\vartheta q = \rho$ and $\|\mathbf{X}_0\|_{\mathrm{op}} = 2/(\lambda + \Lambda) \le 2/\lambda$. If they hold at $s < R$, then (47) and (48) give

$$\begin{aligned}\|\mathbf{I}_D - \mathbf{G}\widetilde{\mathbf{X}}_{s+1}\|_{\mathrm{op}} &\le \|\mathbf{I}_D - \mathbf{G}\widetilde{\mathbf{X}}_s\|_{\mathrm{op}}^2 + \Lambda\xi_s \\ &\le \left\{\vartheta^2 + \frac{\vartheta(1-\vartheta)}{2}\right\} q^{2^{s+1}} \le \vartheta q^{2^{s+1}}.\end{aligned}$$

Moreover, $\|\widetilde{\mathbf{X}}_{s+1}\|_{\mathrm{op}} \le \lambda^{-1}\{1 + \|\mathbf{I}_D - \mathbf{G}\widetilde{\mathbf{X}}_{s+1}\|_{\mathrm{op}}\} \le 2/\lambda$, proving (50).

Consequently,

$$\|\widetilde{\mathbf{X}}_R\mathbf{b} - \mathbf{G}^{-1}\mathbf{b}\|_2 \le \frac{B_0\vartheta}{\lambda} q^{2^R} \le \frac{\varepsilon}{2}.$$

Together with (49), this proves (46).

If $R \ge 1$, minimality gives $Aq^{2^{R-1}} > \varepsilon/2$, and hence $2^R \le C\log(e/\varepsilon)$; this is immediate for $R = 0$. Furthermore,

$$\sum_{s=0}^{R-1} \log(e/\xi_s) \le C\sum_{s=0}^{R-1}(1 + 2^s) \le C\log(e/\varepsilon).$$

For the width count, reserve registers for $\mathbf{G}$, the current iterate, the next iterate, $\mathbf{b}$, and the final coefficient vector. This uses $3D^2 + 2D$ scalar registers. Both the Newton–Schulz update and the final matrix–vector product have degree at most three. The sequential implementation in Lemma 8 therefore has width at most

$$2(3D^2 + 2D) + 2 \cdot 3 + 26 = 6D^2 + 4D + 32 = W_{\mathrm{sol}}.$$

The next iterate is computed in a separate output bank; only after all its entries are available is it copied to the current-iterate bank. The next-iterate bank is then reset and reused. These copies, resets, and the initialization are affine operations. The width is independent of $B_0, \lambda, \Lambda, \varepsilon$, and the depth and sparsity are at most $C\log(e/\varepsilon)$. Lemma 8 also gives uniform parameter bounds for the update and final multiplication networks; the initialization and identity channels use only fixed parameters. Thus the maximum parameter magnitude is bounded independently of ε. □

Architecture of the construction. The blocks are ordered as L_{feat} raw-feature blocks, one Gaussian aggregation block, and L_{post} coefficient-recovery blocks. In the first and third stages, zero value maps make the residual attention sublayers identity maps, so all computations are tokenwise. In the middle block, softmax attention performs the only cross-token aggregation; setting $\mathbf{W}_{\ell,2} = \mathbf{0}$ and $\mathbf{b}_{\ell,2} = \mathbf{0}$ makes its residual FFN sublayer the identity.

Proof of Theorem 3. Write $D := J_p$, $D_{\mathrm{raw}} := J_{2p} + J_p$, $\mathbf{u}_i := (\mathbf{x}_i - \mathbf{x})/h$, $\mathbf{G}_h := \mathbf{G}_h(D_n, \mathbf{x})$, and $\mathbf{b}_h := \mathbf{b}_h(D_n, \mathbf{x})$. Choose a sufficiently small $c_0 \in (0, 1)$, depending only on $d, p, M, \lambda, \Lambda$, and set $\delta = c_0\varepsilon h^{2p+2}$ and $\eta = c_0\varepsilon$.

Stage 1: raw-moment feature construction. Apply Lemma 9. The first L_{feat} blocks preserve the input coordinates and produce $\widetilde{s}_j, \widetilde{\mathbf{r}}_j$ satisfying (44). Their residual attention sublayers are identity maps, and

$$L_{\mathrm{feat}} \vee S_{\mathrm{feat}} \le C\log(e/\delta), \qquad B_{\mathrm{feat}} \le C.$$

Stage 2: Gaussian-kernel moment aggregation. Apply Lemma 10. The middle attention sublayer writes into fresh query registers approximations to

$$\frac{\sum_{i=1}^n K(\mathbf{u}_i)\mathbf{x}_i^{\boldsymbol{\gamma}}}{a_h(\mathbf{x})}, \quad \boldsymbol{\gamma} \in \mathcal{A}_{2p}, \qquad \frac{\sum_{i=1}^n K(\mathbf{u}_i)y_i\mathbf{x}_i^{\boldsymbol{\gamma}}}{a_h(\mathbf{x})}, \quad \boldsymbol{\gamma} \in \mathcal{A}_p,$$

with coordinatewise error at most $C\delta h^{-2}$. Its residual FFN sublayer is the identity map. Since $\delta h^{-2} = c_0\varepsilon h^{2p} \le c_0$, the approximate raw moments remain in a fixed enlarged compact set.

Stage 3: local-polynomial coefficient recovery. For every multi-index $\boldsymbol{\eta}$,

$$(\mathbf{x}_i - \mathbf{x})^{\boldsymbol{\eta}} = \sum_{\boldsymbol{\gamma} \le \boldsymbol{\eta}} \binom{\boldsymbol{\eta}}{\boldsymbol{\gamma}} \mathbf{x}_i^{\boldsymbol{\gamma}}(-\mathbf{x})^{\boldsymbol{\eta}-\boldsymbol{\gamma}}.$$

Consequently, for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}_p$,

$$\{\mathbf{G}_h\}_{\boldsymbol{\alpha},\boldsymbol{\beta}} = h^{-|\boldsymbol{\alpha}|_1-|\boldsymbol{\beta}|_1} \sum_{\boldsymbol{\gamma}\le\boldsymbol{\alpha}+\boldsymbol{\beta}} \binom{\boldsymbol{\alpha}+\boldsymbol{\beta}}{\boldsymbol{\gamma}} (-\mathbf{x})^{\boldsymbol{\alpha}+\boldsymbol{\beta}-\boldsymbol{\gamma}} \frac{\sum_{i=1}^n K(\mathbf{u}_i)\mathbf{x}_i^{\boldsymbol{\gamma}}}{a_h(\mathbf{x})}, \tag{51}$$

$$\{\mathbf{b}_h\}_{\boldsymbol{\alpha}} = h^{-|\boldsymbol{\alpha}|_1} \sum_{\boldsymbol{\gamma}\le\boldsymbol{\alpha}} \binom{\boldsymbol{\alpha}}{\boldsymbol{\gamma}} (-\mathbf{x})^{\boldsymbol{\alpha}-\boldsymbol{\gamma}} \frac{\sum_{i=1}^n K(\mathbf{u}_i) y_i \mathbf{x}_i^{\boldsymbol{\gamma}}}{a_h(\mathbf{x})}. \tag{52}$$

Before the deterministic rescalings, these polynomial maps are Lipschitz on the fixed compact domain of the query and raw-moment coordinates. Thus the raw-moment error contributes at most $C\delta h^{-2p-2}$, because the largest rescaling in (51)–(52) is h^{-2p}. Lemma 8, implemented by residual FFNs as above, approximates the pre-rescaling polynomial maps to accuracy ηh^{2p}. These maps have coefficients and input domains independent of h, so their ReLU parameters are bounded by a constant independent of h, ε. Apply the subsequent rescalings in a separate positionwise block. On the full token register vector, let $\mathbf{A}_h$ act diagonally by $h^{-|\boldsymbol{\alpha}|_1-|\boldsymbol{\beta}|_1}$ on the Gram-moment registers, by $h^{-|\boldsymbol{\alpha}|_1}$ on the response-moment registers, and by the identity on all other coordinates. Thus $\|\mathbf{A}_h\|_{\max} \le h^{-2p}$, and the map $\mathbf{z} \mapsto \mathbf{A}_h\mathbf{z}$ is implemented exactly by one residual FFN via

$$\mathbf{z} + (\mathbf{A}_h - \mathbf{I})\{\mathrm{ReLU}(\mathbf{z}) - \mathrm{ReLU}(-\mathbf{z})\} = \mathbf{A}_h\mathbf{z}.$$

Its hidden width is $2W_0$, its parameter magnitude is at most Ch^{-2p}, and its sparsity is bounded by a constant. Keeping this block separate ensures that these rescaling coefficients are not multiplied into the parameters of the pre-rescaling polynomial networks. The resulting $\widetilde{\mathbf{G}}_h, \widetilde{\mathbf{b}}_h$ therefore satisfy

$$\|\widetilde{\mathbf{G}}_h - \mathbf{G}_h\|_{\mathrm{op}} + \|\widetilde{\mathbf{b}}_h - \mathbf{b}_h\|_2 \le C\left\{\delta h^{-2p-2} + \eta\right\} \le Cc_0\varepsilon. \tag{53}$$

Here and below, we use the equivalence of norms in fixed dimensions. Symmetrize $\widetilde{\mathbf{G}}_h$ by the fixed affine map $(\widetilde{\mathbf{G}}_h + \widetilde{\mathbf{G}}_h^\top)/2$; this does not increase its operator-norm error from the symmetric $\mathbf{G}_h$. The symmetrization uses at most one further residual FFN block with constant parameter magnitude and sparsity. All attention sublayers in this stage are identity maps, and $L_{\mathrm{cen}} \vee S_{\mathrm{cen}} \le C\log\{e/(\eta h^{2p})\}$, while $B_{\mathrm{cen}} \le Ch^{-2p}$.

It remains to solve the approximate normal equations. Set $w_i := K(\mathbf{u}_i)/a_h(\mathbf{x})$ for $i \in [n]$; then $\sum_{i=1}^n w_i \le 1$. For every unit vector $\mathbf{v} \in \mathbb{R}^D$, weighted Cauchy–Schwarz and the upper assumed conditioning bound $\mathbf{G}_h \preceq \Lambda\mathbf{I}_D$ give

$$\begin{aligned}|\mathbf{v}^\top\mathbf{b}_h|^2 &\le \left(\sum_{i=1}^n w_i y_i^2\right)\left(\sum_{i=1}^n w_i \left\{\mathbf{v}^\top\boldsymbol{\psi}_p(\mathbf{u}_i)\right\}^2\right)\\ &\le (M+1)^2\Lambda.\end{aligned}$$

Hence $\|\mathbf{b}_h\|_2 \le B_* := (M+1)\sqrt{\Lambda}$. After the exact symmetrization, set

$$e_G := \|\widetilde{\mathbf{G}}_h - \mathbf{G}_h\|_{\mathrm{op}}, \qquad e_b := \|\widetilde{\mathbf{b}}_h - \mathbf{b}_h\|_2.$$

By (53), decreasing c_0 if necessary ensures $e_G \le \lambda/2$ and $e_b \le 1$, uniformly over $h \in (0,1]$ and $\varepsilon \in (0, 1/2)$. For every unit vector $\mathbf{v} \in \mathbb{R}^D$,

$$\lambda - e_G \le \mathbf{v}^\top\widetilde{\mathbf{G}}_h\mathbf{v} \le \Lambda + e_G.$$

Consequently,

$$\frac{\lambda}{2}\mathbf{I}_D \preceq \widetilde{\mathbf{G}}_h \preceq 2\Lambda\mathbf{I}_D, \qquad \|\widetilde{\mathbf{b}}_h\|_2 \le B_* + 1.$$

In particular, both matrices are invertible and

$$\|\mathbf{G}_h^{-1}\|_{\mathrm{op}} \le \lambda^{-1}, \qquad \|\widetilde{\mathbf{G}}_h^{-1}\|_{\mathrm{op}} \le 2\lambda^{-1}.$$

Apply Lemma 11, with $\lambda/2$, 2Λ, and B_*+1, to approximate $\widetilde{\mathbf{G}}_h^{-1}\widetilde{\mathbf{b}}_h$ to accuracy $c_0\varepsilon$. Its residual-FFN implementation has identity residual attention sublayers and constant FFN width, with $L_{\mathrm{sol}} \vee S_{\mathrm{sol}} \le C\log(e/\varepsilon)$ and $B_{\mathrm{sol}} \le C$.

The identity

$$\widetilde{\mathbf{G}}_h^{-1}\widetilde{\mathbf{b}}_h - \mathbf{G}_h^{-1}\mathbf{b}_h = \widetilde{\mathbf{G}}_h^{-1}(\widetilde{\mathbf{b}}_h - \mathbf{b}_h) + \widetilde{\mathbf{G}}_h^{-1}(\mathbf{G}_h - \widetilde{\mathbf{G}}_h)\mathbf{G}_h^{-1}\mathbf{b}_h$$

and the spectral bounds yield

$$\begin{aligned}\left\|\widetilde{\mathbf{G}}_h^{-1}\widetilde{\mathbf{b}}_h - \mathbf{G}_h^{-1}\mathbf{b}_h\right\|_2 &\le \frac{2}{\lambda}\|\widetilde{\mathbf{b}}_h - \mathbf{b}_h\|_2 \\ &\quad + \frac{2B_*}{\lambda^2}\|\widetilde{\mathbf{G}}_h - \mathbf{G}_h\|_{\mathrm{op}} \\ &\le Cc_0\varepsilon.\end{aligned}$$

Together with the $c_0\varepsilon$-accuracy of the linear-system network, and after decreasing c_0 by a fixed factor, the first coordinate $\widetilde{\beta}_0$ of the solver output satisfies $|\widetilde{\beta}_0 - \mathbf{e}_0^\top \mathbf{G}_h^{-1}\mathbf{b}_h| \le \varepsilon$. A final positionwise affine map writes $\widetilde{\beta}_0$ to the $(d+1)$st coordinate of the query token, as required by the fixed readout. This map is implemented by at most one additional block with identity residual attention. The prescribed output clipping then returns $\mathrm{clip}_M(\widetilde{\beta}_0)$. Since clip_M is one-Lipschitz and $f_{\mathrm{LocPol},h}(D_n, \mathbf{x}) = \mathrm{clip}_M(\mathbf{e}_0^\top \mathbf{G}_h^{-1}\mathbf{b}_h)$, we obtain (10).

Explicit width and parameter count. Reserve disjoint scalar-register banks as follows:

Retained quantities	Number of registers
$\mathbf{x}_j, y_j, q_j, 1, \widetilde{s}_j$	$d+4$
$\widetilde{\mathbf{r}}_j, \widetilde{\mathbf{m}}_j$	$2D_{\mathrm{raw}}$
$\widetilde{\mathbf{G}}_h, \widetilde{\mathbf{X}}_s, \widetilde{\mathbf{X}}_{s+1}$	$3D^2$
$\widetilde{\mathbf{b}}_h, \widetilde{\boldsymbol{\beta}}$	$2D$

Thus the total register count is $R_0 := d + 4 + 2D_{\mathrm{raw}} + 3D^2 + 2D$. Unused banks are initialized to zero. Polynomial terms are evaluated sequentially and added to their output registers. The current iterate is preserved until the next iterate has been completely evaluated; an affine copy then replaces it, and the next-iterate bank is reset. No register bank is allocated separately for each iteration.

The feature maps have degree at most $\max\{2, 2p+1, p+2\}$, the pre-rescaling moment-conversion maps have degree at most $2p+1$, and the solver maps have degree at most three. Hence $k_0 := 2p+3$ bounds all these degrees. In the sequential construction of Lemma 8, the retained registers use at most $2R_0$ signed identity channels, while the single active monomial subnetwork has width at most $2k_0 + 26 = 4p + 32$. The hidden width of every ordinary ReLU computation is therefore at most

$$2R_0 + 4p + 32 = 2d + 4J_{2p} + 6J_p^2 + 8J_p + 4p + 40 = W_0.$$

The affine copies, resets, symmetrization, and final coordinate selection fit within the same width budget. At the interface with the aggregation block, the retained variables occupy their designated token coordinates and its output-moment bank is exactly zero.

After zero padding, take $d_{\mathrm{model}} = W_0$ and $d_{\mathrm{ff}} = 3W_0$. Equation (43) implements the preceding ReLU computations by residual FFNs of this width. The attention maps are square; only $d+2$ query–key coordinates and D_{raw} value coordinates are active in the single aggregation sublayer. All other residual attention sublayers are identity maps. Since $p+1 \le J_p$ and $J_p \ge 1$, the explicit choice satisfies $W_0 = O(d + J_{2p} + J_p^2)$.

Let $L_{\mathrm{out}} \le 1$ count the final readout-placement block, and set

$$L_{\mathrm{post}} = L_{\mathrm{cen}} + L_{\mathrm{sol}} + L_{\mathrm{out}}, \qquad L = L_{\mathrm{feat}} + 1 + L_{\mathrm{post}}.$$

All fixed affine interface operations are included in the corresponding stage counts. The number of monomials in each polynomial map depends only on d, p. Serial evaluation therefore changes only the constants in the depth and sparsity bounds. Residualization adds $O(W_0)$ nonzero parameters per implemented layer, and W_0 is independent of n, h, ε. Using $\delta = c_0\varepsilon h^{2p+2}$ and $\eta = c_0\varepsilon$, we obtain

$$\begin{aligned}L \vee \|\theta\|_0 &\le C\log(e/\delta) + C\log\left(\frac{e}{\eta h^{2p}}\right) + C\log(e/\varepsilon) + C \\ &\le C\log\left(\frac{e}{h\varepsilon}\right).\end{aligned}$$

The feature and solver networks have parameter magnitudes bounded by C, independently of h, ε. The attention maps have parameter magnitude at most Ch^{-2}, and the only bandwidth-dependent parameters in the recovery blocks are the separate affine rescalings, bounded by Ch^{-2p}. The residual-FFN implementation, affine register operations, and symmetrization introduce only fixed factors. Therefore, since $h \in (0, 1]$,

$$\|\theta\|_\infty \le C \max\{1, h^{-2}, h^{-2p}\} = Ch^{-\max\{2,2p\}}.$$

Let $\theta(h, \varepsilon)$ denote the constructed parameters and set $f_{h,\varepsilon} := f_{\theta(h,\varepsilon)}$. Taking $S := \|\theta(h, \varepsilon)\|_0$ yields $f_{h,\varepsilon} \in \mathcal{F}(W_0, 3W_0, 1, L, Ch^{-\max\{2,2p\}}, S, M)$. The parameters are independent of n.

The positionwise maps are shared across tokens, no positional encoding is used, and the middle softmax-weighted sum is permutation invariant in the context tokens. Permuting the first n tokens while fixing the query token therefore leaves the output unchanged. This proves context-order invariance and completes the proof. □

G Additional experimental details and results

This section provides the experimental protocol, additional validity and efficiency results, numerical diagnostics, and timing details supporting Section 5.

G.1 Experimental design and task prior

Task generation. Each task is generated from $y = f(x) + \sigma(x)\epsilon$, where $\epsilon \sim N(0, 1)$. The regression function, input distribution, and noise mechanism are sampled anew for each task and remain fixed within that task. Conditional on these choices, context and test observations are i.i.d. Table 3 specifies the six function families. Inputs are zero-padded to 16 dimensions when necessary. The Transformer receives neither the function-family identifier nor the true test response.

Table 3: Six regression-function families. Here $[a]_+ = \max(a, 0)$, and $U(a, b)$ denotes the uniform distribution on $[a, b]$.

Family (d)	Function and parameter distribution
Sparse linear (16)	$f(x) = A\sum_{j\in S}\beta_j x_j$. $\lvert S\rvert \sim \mathrm{Unif}\{2, \dots, 7\}$, $\beta_j \sim N(0, 3/16)$, $A \sim U(0.85, 1.15)$.
Generalized additive model (GAM) (16)	$f(x) = \frac{1}{1.8}\sum_{j=1}^{4} a_j b_j(v_j)$, $v_j = x_{s_j}$ (four active coordinates). $\lvert a_j\rvert \sim U(0.65, 1.25)$ with random signs. Bases: $b_1(v) = \sin(\pi v)$, $b_2(v) = \cos(1.5\pi v)$, $b_3(v) = 1.8(v^2 - 1/3)$, $b_4(v) = 1.4(v^3 - 0.6v)$.
Friedman-1 (10)	$f(x) = \frac{A}{6.8}\big[10\sin(\pi u_1 u_2) + 20(u_3 - 0.5)^2 + 10u_4 + 5u_5 - 14.4\big]$. $u_j = (x_{s_j} + 1)/2$ (five active coordinates), $A \sim U(0.9, 1.1)$.
MARS-style hinges (16)	$f(x) = \frac{2.1}{\sqrt{8}}\left[\sum_{k=1}^{4} a_k h_k(x) + \sum_{k=5}^{8} a_k h_k(x)\widetilde{h}_k(x)\right]$. $h_k(x) = [s_k(x_{j_k} - t_k)]_+$; $\widetilde{h}_k$ is an independent hinge. $a_k \sim N(0, 1)$, $t_k, \widetilde{t}_k \sim U(-0.65, 0.65)$. Coordinates are sampled with replacement.
RBF-kernel GP (8)	$f(X) \sim N(0, A^2(K + 10^{-4}I))$ on three active coordinates S. $K_{ij} = \exp(-\lVert x_{i,S} - x_{j,S}\rVert^2/(2\ell^2))$, $\ell \sim U(0.55, 1.15)$, $A \sim U(0.8, 1.2)$. The nugget is part of the generator.
Random neural-network teacher (16)	$f(x) = A\, w^\top \mathrm{GELU}\left(W_2\, \mathrm{GELU}(W_1 v + b_1) + b_2\right)$. Five active inputs v; two width-24 hidden layers with Gaussian error linear unit (GELU) activations. Independent zero-mean Gaussian entries of W_1, W_2, w have variances $1/5, 1/24, 1/24$. Bias entries: $b_1 \sim U(-0.35, 0.35)$, $b_2 \sim U(-0.20, 0.20)$; $A \sim U(1.1, 1.6)$.

Input distributions. Each task selects one of three input distributions with equal probability: (i) independent uniform coordinates, $x_j \sim U(-1, 1)$; (ii) correlated coordinates, $x = \tanh(g)$, where $g_1 \sim N(0, 1)$ and $g_j = \rho g_{j-1} + \sqrt{1 - \rho^2}e_j$, with $e_j \sim N(0, 1)$ and $\rho \sim U(0.2, 0.7)$; or (iii) clustered inputs, $x = \mathrm{clip}(sc + 0.32e, -1, 1)$, where $c_j \sim U(-0.55, 0.55)$, s is a uniform random

sign, and e is a standard Gaussian vector. The parameters ρ and c are fixed within each task, whereas s and the Gaussian variables are resampled for each observation. The transformations are applied coordinatewise.

Noise and active coordinates. Each task samples $b \sim U(0.16, 0.28)$ and selects, with equal probability, either homoskedastic noise, $\sigma(x) = b$, or heteroskedastic noise, $\sigma(x) = b[0.65 + 0.35(x_1 + 1)]$. Errors are independently sampled for each observation. Active coordinates are sampled uniformly without replacement, except for MARS-style hinge terms, which allow repeated coordinates.

Training and evaluation mixtures. The pretraining probabilities for the six families, in the order of Table 3, are $(1, 1, 1, 1, 3, 1)/8$. Evaluation assigns equal weight to all six families.

G.2 Training and conformal construction

We describe the Transformer training procedure, the implementation of the conformal scores, the refitting baselines, and the numerical evaluation protocol.

Transformer training. We train five independent Transformers, each for 8,000 AdamW updates with 96 newly generated tasks per update and context sizes sampled from $\{16, 32, 64\}$. Each model has width 128, three cross-attention blocks, four attention heads, feed-forward width 512, and 648,961 parameters. We use a learning rate of 8×10^{-4}, weight decay of 10^{-4}, and gradient clipping at 1.0. The checkpoint from the final prespecified update is used without evaluation-based selection, and its parameters are frozen throughout conformal inference.

ICL score implementation. We implement the leave-one-out predictions $f_\theta(D^z_{-i}, \mathbf{x}_i)$ defined in Section 2. For each observation i in the augmented sample, the query token contains only $\mathbf{x}_i$, while the context tokens contain the pairs $(\mathbf{x}_j, y^z_j)$. A diagonal attention mask excludes observation i from its own context. Context tokens are encoded independently, and no positional encoding is used. This prevents the queried response y^z_i from entering its own prediction and makes the resulting score vector permutation-equivariant. The same frozen model computes all $\widehat{S}^z_i$ without parameter updates.

Refitting baselines. For each candidate z, a baseline algorithm $\mathcal{A}$ fits one predictor $\widehat{f}_z = \mathcal{A}(D^z_{n+1})$ and computes the residual scores $S^{\mathcal{A},z}_i = |y^z_i - \widehat{f}_z(\mathbf{x}_i)|$, $i \in [n+1]$. All scores for a given candidate use the same fitted predictor. Thus, an initial grid of G candidates requires G, rather than $G(n+1)$, model fits.

The MLP has two hidden layers of width 24 with Gaussian error linear unit (GELU) activations and an output transformation $20\tanh(r/20)$. For each candidate, it is initialized using the same fixed, data-independent seed 7301 and trained for 300 full-batch Adam steps with learning rate 0.003. The weight-only L2 penalty of 0.01 is selected by point-prediction mean squared error on 600 independent validation tasks, without using coverage, set length, or test tasks. Biases are not penalized. Candidate-wise fits are vectorized as independent optimization trajectories, without warm starts, parameter sharing across candidates, or cross-candidate score caching.

ExtraTrees uses 64 trees, a minimum leaf size of 2, all input features, and no bootstrap. RBF-SVR uses $C = 1$, $\gamma = 0.1$, $\epsilon = 0.2$, and solver tolerance 10^{-7}. Augmented observations are canonically sorted before fitting ExtraTrees and RBF-SVR. All methods use the same conformal rank criterion. The true test response is used only to evaluate coverage after constructing the prediction set. RBF-SVR rank sensitivity is examined in Appendix G.3.

Finite-grid construction. We approximate the prediction set in (1) using an initial grid of G equally spaced candidates over $[y_{\min} - 0.25\Delta,\ y_{\max} + 0.25\Delta]$, where $\Delta = \max(y_{\max} - y_{\min}, 0.001)$ and $y_{\min}, y_{\max}$ are the minimum and maximum observed responses. If an accepted grid point reaches either endpoint, we double the interval half-width, retaining the existing grid points, for up to four expansion rounds. Consecutive accepted grid points form closed intervals; disconnected intervals are retained, and isolated accepted points contribute zero length.

We distinguish two coverage measurements. *Grid coverage* evaluates membership in the returned union of intervals. *Direct rank coverage* evaluates the true test response using the conformal rank rule without grid approximation. The latter corresponds to the finite-sample validity result.

Uncertainty estimation. Prediction-quality intervals are obtained from 10,000 paired, stratified bootstrap draws, resampling tasks within each checkpoint–family cell. For timing comparisons, repeated measurements are first reduced to a median for each task, followed by bootstrap resampling within families. Coverage across sample sizes and miscoverage levels is reported with pointwise Wilson intervals. The held-out-family experiment uses 20,000 paired task-level bootstrap draws.

G.3 Validity and numerical accuracy

We assess whether the empirical coverage of ICL-FCP agrees with the finite-sample theory and whether finite-grid inversion accurately approximates the prediction sets. We further examine numerical effects that could influence the baseline comparisons.

Coverage across sample sizes and levels. To assess coverage beyond the primary setting, we evaluate ICL-FCP across $n \in \{16, 32, 64\}$ and multiple miscoverage levels. For each n, we sample 3,000 independent tasks and reuse them across levels, yielding 9,000 tasks and 24,000 task–level evaluations.

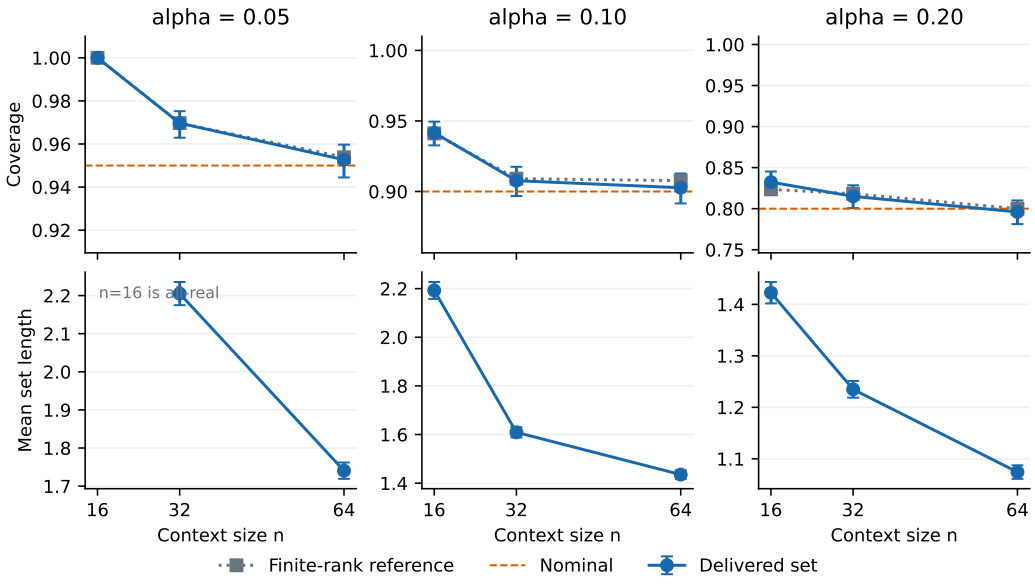


Figure 2: Returned-grid coverage and mean set length across context sizes and miscoverage levels. The reference is the attainable finite-rank coverage. At $n = 16, \alpha = 0.05$, the rank rule accepts every response, so this cell is shown analytically rather than simulated.

Table 4: Coverage across sample sizes and levels. Each simulated row uses 3,000 independent tasks. Direct coverage includes pointwise Wilson 95% intervals; disagreement counts differences between grid and direct membership.

n	α	Rank ref. (%)	Direct cov. (%)	Grid cov. (%)	Mean length	Disagree (%)
32	0.05	96.97	97.03 [96.36, 97.58]	96.97	2.205	0.07
64	0.05	95.38	95.30 [94.48, 96.00]	95.27	1.741	0.03
16	0.10	94.12	94.20 [93.31, 94.98]	94.17	2.193	0.03
32	0.10	90.91	90.90 [89.82, 91.88]	90.77	1.609	0.13
64	0.10	90.77	90.43 [89.33, 91.43]	90.27	1.435	0.17
16	0.20	82.35	83.33 [81.96, 84.62]	83.23	1.423	0.10
32	0.20	81.82	81.63 [80.21, 82.98]	81.50	1.235	0.13
64	0.20	80.00	79.87 [78.39, 81.26]	79.60	1.074	0.27

The attainable finite-rank reference lies within the Wilson 95% interval in every evaluated setting, with a maximum absolute deviation of 0.98 percentage points. Thus, the observed direct coverage is consistent with the finite-sample theory across the tested sample sizes and miscoverage levels. Grid and direct membership disagree in only 28 evaluations (0.117%), all corresponding to responses accepted by the direct rule but omitted by the grid (Figure 2 and Table 4).

Grid-resolution analysis. We assess whether the primary grid size $G = 801$ provides an accurate numerical approximation. Using the same 60 tasks, we compare $G \in \{101, 201, 401, 801, 1601\}$.

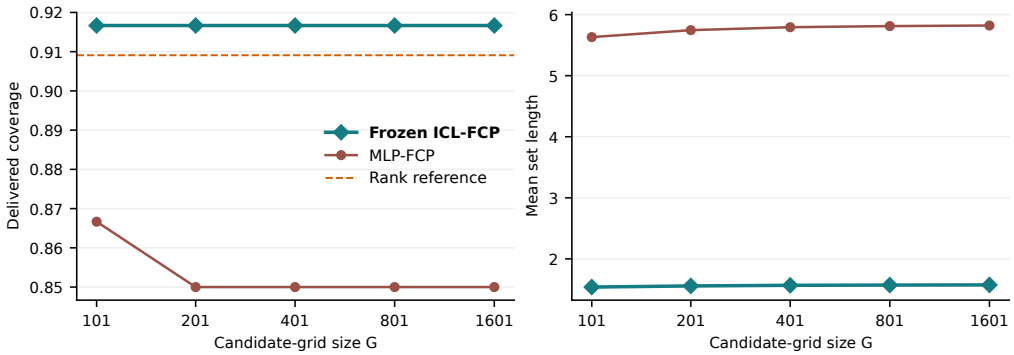


Figure 3: Returned coverage and mean set length on the same 60 tasks across five candidate-grid resolutions.

Table 5: Grid-resolution analysis. $|L - L_{1601}|$ compares mean set lengths with the finest grid.

Method	G	Coverage (%)	Mean length	$\lvert L - L_{1601}\rvert$	Mismatch	Flags
Frozen ICL-FCP	**101**	**91.7**	**1.538**	**0.0348**	**0**	**0**
	201	**91.7**	**1.557**	**0.0164**	**0**	**0**
	401	**91.7**	**1.567**	**0.0061**	**0**	**0**
	801	**91.7**	**1.571**	**0.0022**	**0**	**0**
	1601	**91.7**	**1.573**	**0.0000**	**0**	**0**
MLP-FCP	101	86.7	5.632	0.1897	0	1
	201	85.0	5.746	0.0756	1	1
	401	85.0	5.794	0.0278	1	1
	801	85.0	5.812	0.0094	1	1
	1601	85.0	5.822	0.0000	1	1

ICL-FCP yields identical realized-response membership decisions across all five resolutions, while the mean set lengths at $G = 801$ and $G = 1601$ differ by only 0.0022. These results indicate that $G = 801$ provides a stable numerical approximation on the evaluated tasks (Figure 3 and Table 5).

Endpoint and membership diagnostics. We check whether accepted grid endpoints lead to truncated prediction sets or affect the primary set-length comparisons.

Table 6: Numerical diagnostics on the 600 primary tasks. Initial flags indicate accepted grid endpoints before targeted re-evaluation. The Frozen ICL-FCP row is bold.

Method	Grid/rank disagree	Initial flags	Unresolved	Median length	P90 length
Frozen ICL-FCP	**1**	**0**	**0**	**1.516**	**2.482**
MLP-FCP	3	5	0	1.812	20.381
ExtraTrees-FCP	3	0	0	1.507	2.422
Exact RBF-SVR-FCP	11	0	0	1.566	2.564

The initial grid produces five endpoint flags for MLP-FCP. Re-evaluation at $G = 1601$ with the same expansion rule resolves all five, leaving no unresolved endpoints for any primary method. Replacing the five original MLP-FCP set lengths with their corrected values changes its pooled mean from 5.959 to 5.986, without affecting coverage. Thus, the endpoint corrections do not alter the observed set-length comparison (Table 6).

RBF-SVR rank sensitivity. Finally, we assess whether the RBF-SVR results are sensitive to numerical stabilization of the conformal rank comparison. We compare the primary exact rule ($\tau = 0$) with a validation-selected stabilized rule ($\tau = 10^{-4}$).

Table 7: RBF-SVR rank sensitivity on the same 600 quality tasks. The exact rule is the primary specification; the stabilized rule uses $\tau = 10^{-4}$. Timing uses separate 60-task CPU runs.

SVR rank rule	τ	Grid cov. (%)	Direct cov. (%)	Mean length	Time (ms)	Mismatch
Exact	0	90.83	91.33	1.690	55.31	11
Stabilized	10^{-4}	92.50	92.50	1.747	86.39	0

Under the exact rule, direct and grid coverage are 91.33% and 90.83%, respectively, with 11 membership disagreements. The stabilized rule yields 92.50% coverage for both evaluations. Although numerical stabilization changes some membership decisions, the qualitative set-length and timing conclusions remain unchanged (Table 7).

G.4 Prediction-set efficiency

We provide additional baseline comparisons and examine how prediction-set efficiency depends on optimization and the pretraining task distribution.

Additional baseline comparisons. The primary and per-family results appear in Tables 1 and 2. Table 8 supplements these results with paired coverage comparisons against the two nonparametric baselines. All comparisons use the same 600-task subset, with 20 tasks per checkpoint–family cell across five checkpoints and six families.

Table 8: Paired coverage comparisons with the two nonparametric baselines. Coverage differences are ICL minus baseline in percentage points; brackets are paired-bootstrap 95% intervals.

Baseline	**ICL** - baseline coverage (pp)
ExtraTrees-FCP	-0.67 [-2.83, 1.50]
Exact RBF-SVR-FCP	-0.50 [-2.50, 1.50]

MLP optimization-budget sensitivity. We examine whether the MLP-FCP efficiency gap is sensitive to its optimization budget. Table 9 reports the results.

Table 9: MLP optimization-budget sensitivity on an independent 36-task diagnostic at $G = 401$. Coverage brackets are Wilson 95% intervals.

Method	Adam steps	Coverage (%)	Mean / median length	Post-hoc fit MSE
Frozen ICL-FCP	**–**	**88.89** **[74.69, 95.59]**	**1.582 / 1.542**	**–**
MLP-FCP	100	91.67 [78.17, 97.13]	5.775 / 1.468	0.04911
MLP-FCP	300	91.67 [78.17, 97.13]	6.221 / 1.673	0.02144
MLP-FCP	600	94.44 [81.86, 98.46]	6.776 / 1.585	0.01771

Across 100, 300, and 600 Adam steps, lower training error does not produce shorter MLP-FCP sets, and the efficiency gap remains. The same six tasks per family are reused across budgets. This diagnostic uses one checkpoint. The MSE is computed post hoc from the fit augmented by the realized response and is never used to construct the prediction set.

Held-out function-family evaluation. We examine how excluding a function family from pretraining affects coverage and set length. Figure 4 and Table 10 compare two checkpoints on fresh MARS-style tasks.

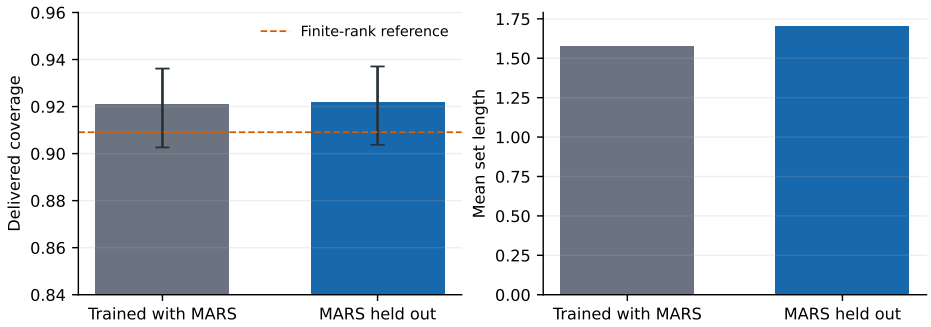


Figure 4: Paired evaluation on 1,000 fresh MARS-style tasks: the canonical seed-0 checkpoint and a seed-0 checkpoint pretrained without the MARS-style family.

Table 10: Held-out function-family results. Coverage intervals are Wilson 95% intervals.

Checkpoint	Direct coverage (%)	Delivered coverage (%)	Mean length	Mismatch
Trained with MARS	92.20 [90.37, 93.71]	92.10 [90.26, 93.62]	1.574	1
MARS held out	92.20 [90.37, 93.71]	92.20 [90.37, 93.71]	1.705	0

Direct rank coverage is 92.2% for both checkpoints. Holding MARS-style tasks out of pretraining increases mean set length from 1.574 to 1.705, a paired increase of 0.131 (95% interval: [0.115, 0.146]), or 8.29%. This illustrates the separation between marginal validity, which is controlled by the conformal rank, and prediction-set efficiency, which depends on the learned score.

G.5 COMPUTATIONAL EFFICIENCY

We evaluate the computational benefit of replacing candidate-wise refitting with frozen ICL inference. We first measure end-to-end online latency, then examine scaling with grid size and compare against a structured solver to clarify the scope of the speedup.

Timing protocol and main results. We benchmark ICL-FCP, MLP-FCP, and RBF-kernel ridge regression (RBF-KRR) on the same NVIDIA RTX A6000 using 60 fixed tasks, ten per function family. Inputs and the frozen checkpoint are resident. After five warm-ups per method and family, each task is measured three times, with Latin-rotated method order. We report the median of per-task medians, task-level interquartile ranges, and paired stratified bootstrap intervals for latency ratios.

Timing includes grid construction, adaptive expansion, candidate fitting or scoring, conformal ranking, and prediction-set construction. Offline pretraining, validation, loading, task generation and transfer, and coverage evaluation are excluded.

ICL-FCP takes 11.04 ms, compared with 788.58 ms for MLP-FCP, yielding a paired $71.4\times$ speedup (95% interval: [70.0, 136.1]). Because adaptive expansion is included, the methods evaluate different mean numbers of candidates (841.1 for ICL-FCP and 1468.5 for MLP-FCP). The reported speedup therefore measures complete online prediction-set construction, rather than per-candidate computational complexity.

Scaling with grid size. We examine whether the computational advantage persists as the number of candidate responses G increases. Table 11 compares ICL-FCP and MLP-FCP under the same GPU protocol.

Table 11: Online latency across grid sizes on 60 tasks using an NVIDIA RTX A6000. Each entry is the median task latency over three measured repeats.

Initial grid G	**ICL time (ms)**	MLP time (ms)	MLP/**ICL**
101	**3.107**	820.637	264.15×
201	**4.065**	826.055	203.22×
401	**6.305**	799.074	126.73×
801	**9.647**	775.316	80.37×
1601	**13.274**	830.983	62.60×

ICL-FCP latency increases with grid size, but it avoids the repeated optimization required by MLP-FCP. The latter remains dominated by candidate-wise fitting (Table 11). Thus, the computational gain comes from eliminating model refitting, not from avoiding candidate evaluation.

CPU baseline comparisons. We additionally compare ICL-FCP with the CPU implementations of ExtraTrees-FCP and exact RBF-SVR-FCP. Both are timed on an Intel Xeon w9-3495X with eight persistent candidate workers, using the same 60 task identities and online timing scope. Table 12 reports their latencies; the main-text ratios use the corresponding ICL-FCP medians from the RTX A6000 benchmark (11.04 ms overall and 10.86–11.18 ms by family).

The per-family RBF-SVR medians of Table 2 are taken from the same exact-rank record as Table 12. The stabilized rule ($\tau = 10^{-4}$) was timed once, in a separate multi-method CPU block, and measured the identical SVR work about $1.56\times$ slower (86.39 ms against 55.31 ms median); because it changes the acceptance threshold rather than the amount of fitting work, the main text reports the exact-rank record throughout and retains the stabilized variant only as the sensitivity check of Appendix G.3.

Table 12: Separate CPU implementation timings on 60 retained task identities.

Method	Median [IQR] (ms)	Device / parallelism
ExtraTrees-FCP	5290.86 [5126.66, 5406.74]	Xeon w9-3495X / 8 workers
Exact RBF-SVR-FCP	55.31 [51.10, 63.93]	Xeon w9-3495X / 8 workers

The measured CPU implementations have higher online latency than ICL-FCP, extending the observed implementation-level advantage beyond the MLP baseline. These cross-device comparisons are descriptive and do not isolate the effect of the algorithm from that of hardware or implementation.

Structured KRR comparison. Finally, we compare ICL-FCP with RBF-KRR to determine whether its computational advantage extends to methods with efficient refitting procedures. Unlike the iterative MLP baseline, RBF-KRR uses batched candidate-wise direct linear solves. We compare prediction quality on a separate 1,500-task subset and latency on the same 60 GPU timing tasks.

Table 13: Structured-solver comparison on an RTX A6000. Prediction quality uses 1,500 paired tasks, and latency uses 60 matched tasks. The final column reports method time divided by ICL-FCP time.

Method	Coverage (%)	Mean length	Time (ms)	Method/**ICL** time
MLP-FCP	90.87	6.063	788.58 [767.41, 1575.95]	71.44×
Frozen ICL-FCP	**90.67**	**1.624**	**11.04 [10.83, 11.16]**	**1.00**×
RBF-KRR-FCP	91.47	1.843	10.05 [8.70, 13.05]	0.91×

RBF-KRR is slightly faster than ICL-FCP (10.05 versus 11.04 ms), whereas ICL-FCP achieves a shorter mean prediction set (1.624 versus 1.843). This comparison identifies the scope of the computational advantage: frozen ICL substantially reduces the cost of generic candidate-wise nonlinear refitting, while structured linear-algebra methods can achieve comparable online latency.

The 1,500-task quality subset differs from the 600-task common subset in Tables 1 and 2; pooled prediction-quality results are therefore reported separately (Table 13).