From predictions to confidence intervals: an empirical study of conformal prediction methods for in-context learning

Zhe Huang[1], **Simone Rossi**[2], **Rui Yuan**[3], **Thomas Hannagan**[3]

[1]PSL University, [2]EURECOM, [3]Stellantis

Objective and Contributions

Uncertainty quantification using transformers' in-context learning (ICL) abilities is a promising area of research, but still underexplored [1].

Objectives:

- Leverage in-context learning for robust, scalable uncertainty quantification for noisy regression [3].
- Benchmark against reference methods for performance and computational efficiency.

Contributions:

- Introduce a novel **conformal prediction** method based on in-context learning.
- Demonstrate the effectiveness of our approach through extensive experiments, including exact oracle conformal predictors for linear self-attention.

In-context learning and conformal prediction

Full Conformal Prediction

Goal. Build set $\Gamma_\alpha(\mathbf{x}_{n+1})$ with coverage guarantee: $P(y_{n+1} \in \Gamma_\alpha(\mathbf{x}_{n+1})) \geq 1 - \alpha$

- **Augment** dataset $\mathcal{D}_n$ for $z \in \mathbb{R}$ (or $\{z \in \mathcal{Z}\} \subset \mathbb{R}$)

$$\mathcal{D}_{n+1}(z) = \mathcal{D}_n \cup \{(\mathbf{x}_{n+1}, z)\} \quad \text{with } \mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

- **Fit model** with ridge regression on $\mathcal{D}_{n+1}(z)$ for each z:

$$\widehat{\mathbf{w}}(z) = \arg\min_{\mathbf{w} \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2 + (z - \langle \mathbf{w}, \mathbf{x}_{n+1} \rangle)^2 + \lambda \|\mathbf{w}\|_2^2$$

- Compute **conformity scores** (residuals):

$$\widehat{R}_i(z) = \left|y_i - \langle \widehat{\mathbf{w}}(z), \mathbf{x}_i \rangle\right|, \quad \widehat{R}_{n+1}(z) = \left|z - \langle \widehat{\mathbf{w}}(z), \mathbf{x}_{n+1} \rangle\right|$$

- **Rank** the conformity scores and compute the p-value:

$$\widehat{\pi}(z) = 1 - \frac{1}{n+1} \operatorname{rank}\left(\widehat{R}_{n+1}(z)\right)$$

- **Build the prediction set** for $\mathbf{x}_{n+1}$:

$$\Gamma_\alpha(\mathbf{x}_{n+1}) = \{z \in \mathbb{R} \mid \widehat{\pi}(z) \geq \alpha\}$$

- **In-context learning for regression tasks.** Context $\mathcal{D}_n^{(\tau)} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i = \langle \mathbf{w}^{(\tau)}, \mathbf{x}_i \rangle + \varepsilon_i$, with $\mathbf{w}^{(\tau)} \sim \mathcal{N}(0, \gamma^2 \mathbf{I}_d)$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{x}_i \sim \mathcal{U}(-a, a)^d$. Predict unseen label y_{n+1} given $\mathbf{x}_{n+1}$ via single forward pass.
- **Tokenization scheme.** Tokens include both context and query:

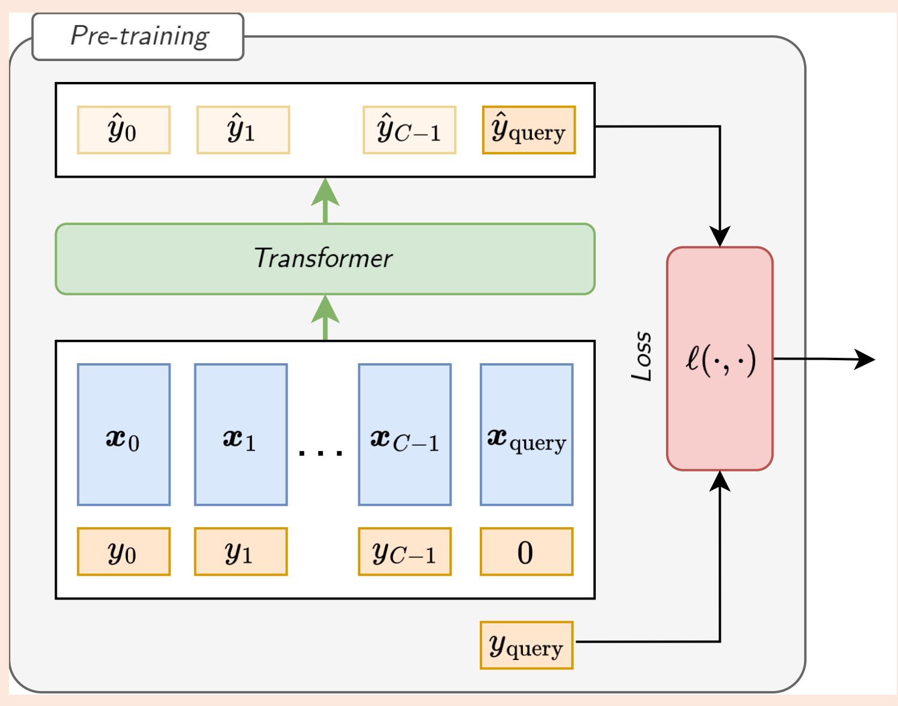
$$\mathbf{E}\left(\mathcal{D}_n^{(\tau)}, \mathbf{x}_{n+1}^{(\tau)}, z\right) = \begin{bmatrix} \mathbf{x}_1^{(\tau)} & \dots & \mathbf{x}_n^{(\tau)} & \mathbf{x}_{n+1}^{(\tau)} \\ y_1^{(\tau)} & \dots & y_n^{(\tau)} & z \end{bmatrix}$$

Mask $z = 0$ at training.

- **Linear Self-Attention (LSA) $\mathbf{f}_{\text{LSA}}$.** Alternative to softmax attention for ICL. Theoretically justified by [3, 2].
- **Pre-training.** Minimize loss w.r.t. LSA parameters $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\mathbf{f}_{\text{LSA}}(\boldsymbol{\theta}, \mathbf{E})_{[d+1, n+1]} - y_{n+1}\right)^2\right]$$

with expectation over $p(\mathbf{w}, \mathbf{x}, \varepsilon)$.



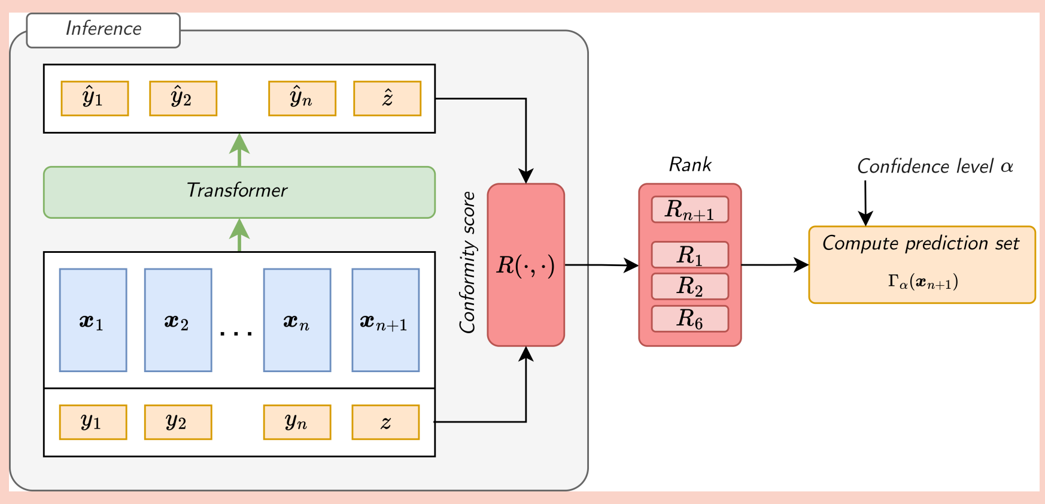
Bridging ICL and CP

Note: Pre-trained transformer converges to optimal Bayes-risk predictor ($\lambda = \gamma^2/\sigma^2$)

- **Predictive residuals via attention.** Use transformer outputs and avoid re-training for each z and predicting $\langle \mathbf{w}, \mathbf{x}_i \rangle$ for all i. Replace optimization with forward pass through the transformer.

$$\widehat{\mathbf{y}}(z) = \mathbf{f}_{\text{LSA}}(\boldsymbol{\theta}, \mathbf{E}(\mathcal{D}_n, \mathbf{x}_{n+1}, z))_{[d+1,:]}$$

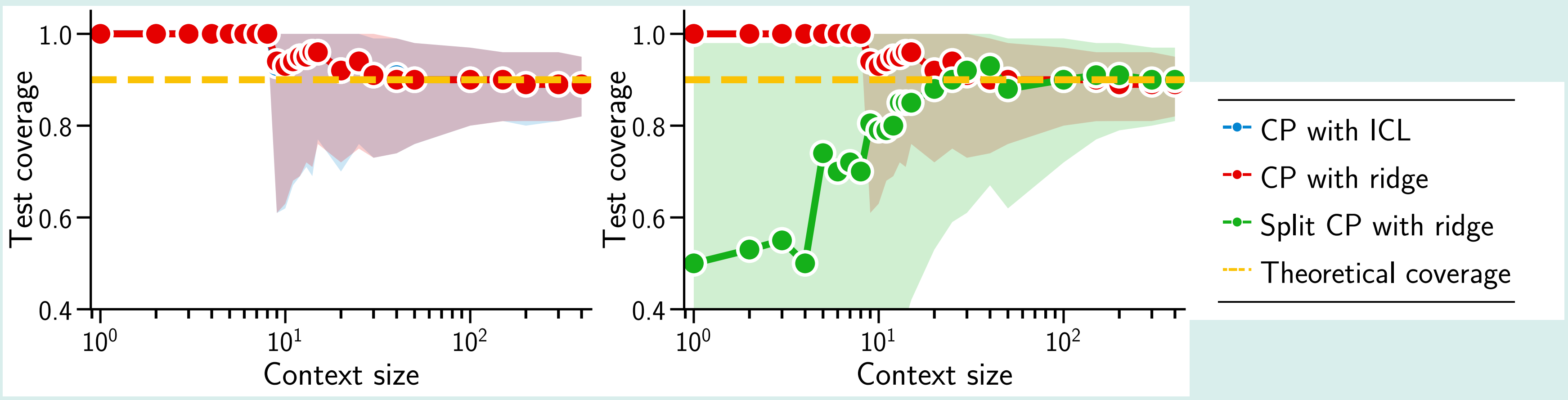
- **Directly** compute the residuals $\widehat{R}_i(z) = \left|y_i - \widehat{\mathbf{y}}(z)_{[i]}\right|$, $\widehat{R}_{n+1}(z) = \left|z - \widehat{\mathbf{y}}(z)_{[n+1]}\right|$
- **Compute** $\widehat{\pi}(z)$ and collect all z with $\widehat{\pi}(z) \geq \alpha$.
- For efficient computation, select $z \in \mathcal{Z}$ and parallelize over z and $\mathbf{x}_{n+1}$ as single batched forward pass.



Take a picture to download the full paper

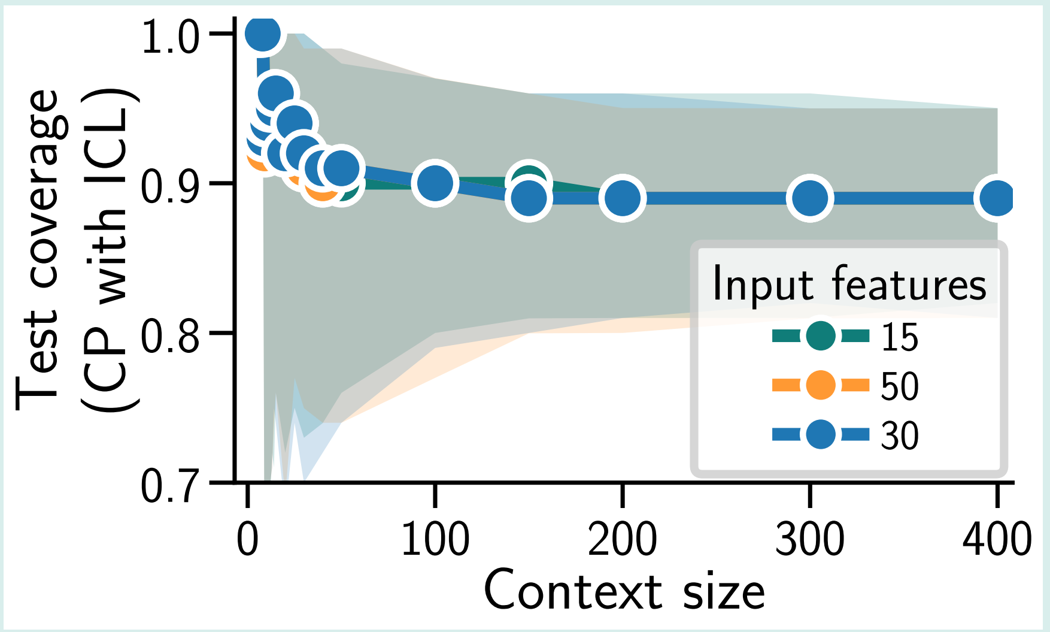
Do ICL and CP work together?

- **Objectives**: Evaluate quality of predictive intervals combining ICL and CP.
- **Baseline**: CP with ridge (oracle), and split CP with ridge.



CP with ICL converges to theoretical coverage as context size increases, matching oracle performance. Split CP with ridge shows higher variance at small contexts.

Does CP with ICL scale?



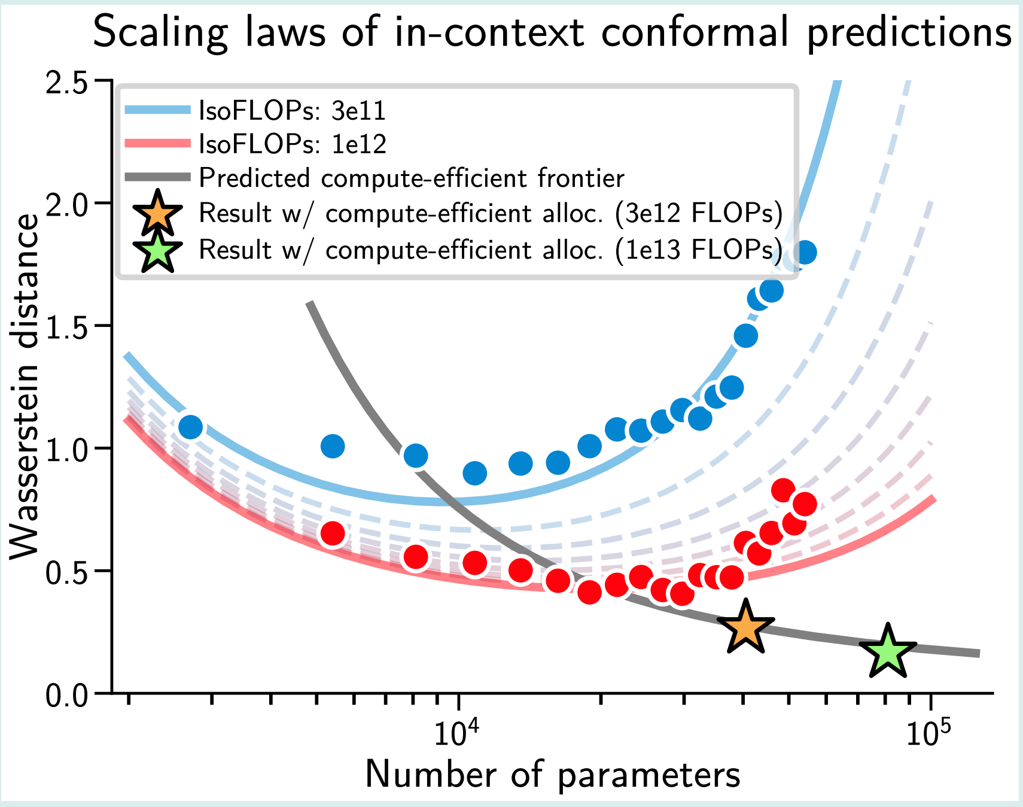
Coverage of CP with ICL converges to theoretical value independently of input dimension.

Method	Context: 100 points
CP with ICL	0.32s (0.29, 0.41)
CP with ridge	0.42s (0.40, 0.49)
Split CP with ridge	0.41s (0.37, 0.49)

CP with ICL is faster than (but comparable to) ridge-based methods at context size 100.

Scaling laws for conformal prediction with ICL

- **Question**: How does predictive interval quality scale with compute-contrained pre-training?
- **Method**: Predictive interval quality modeled as function of model size and data size, given compute budget.
- **How**: Train models with varying compute, model size, and data size, and evaluate predictive intervals (vs oracle).



Summary: (1) Can be used to predict the best model size and data size for a given compute budget. (2) Scaling follows Chinchilla laws but larger models yield better predictive intervals than increasing data alone.

Conclusions

CP with ICL provides reliable predictive intervals with guaranteed coverage, matches oracle performance, and reduces compute cost.

- **Discussion**:
  - Synthetic experiments show robust coverage and computational efficiency.
  - Method achieves similar performance to oracle predictors with a single forward pass.
  - Mechanistic interpretation of ICL enables fast and accurate uncertainty estimation.

Open questions

- Address limitations of simplified transformer architecture.
- Towards token-level uncertainty quantification for language models.

References

[1] F. Falck et al. (2024). "Is In-Context Learning in Large Language Models Bayesian? A Martingale Perspective".

[2] S. Garg et al. (2022). "What can transformers learn in-context? A case study of simple function classes".

[3] J. Von Oswald et al. (2023). "Transformers learn in-context by gradient descent".