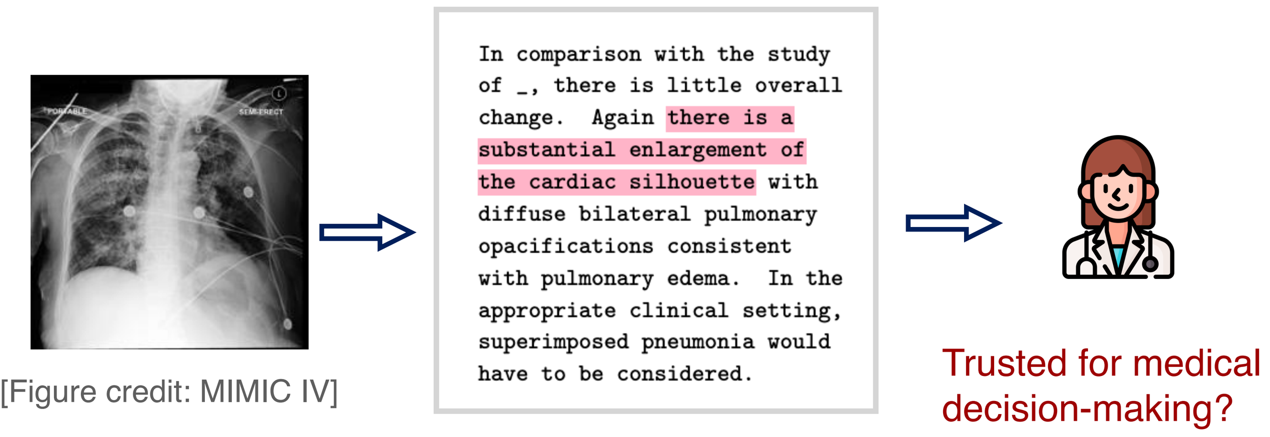
Conformal Alignment: Knowing When to Trust Foundation Models with Guarantees

Yu Gui* [1] Ying Jin* [2] Zhimei Ren* [3] * alphabetical ordering

[1]Department of Statistics, University of Chicago [2]Data Science Initiative, Harvard University [3]Department of Statistics and Data Science, University of Pennsylvania

LLM as "Radiologist"?

Shortage of Radiologist $\Longrightarrow$ use LLM?



[Figure credit: MIMIC IV]

X-ray scan LLM generated report

Question

Foundation model

$$f: \text{ Prompt } X \mapsto \text{ Output } Y$$

- How to safely use LLM outputs $Y = f(X)$?
- What guarantees are reasonable and how to achieve such guarantees?

Problem setup

- Available dataset with reference E_i:
$$\mathcal{D} = \mathcal{D}_{\mathtt{train}} \cup \mathcal{D}_{\mathtt{calib}} = \{(X_i, E_i)\}_{i\in[n]}$$
- Alignment function $\mathcal{A}: (f(X), E) \mapsto A$
- Test dataset $\mathcal{D}_{\mathtt{test}} = \{X_{n+j}\}_{j\in[m]}$
- An output is admissible if
$$A_i = \mathcal{A}(f(X_i), E_i) > c$$
- **Goal**: identify a subset $\mathcal{S} \subseteq [m]$ with "trustworthy" outputs, i.e. $A_{n+j} > c$

Goal: Selection with FDR control

Find a subset $\mathcal{S} \subseteq [m]$ such that

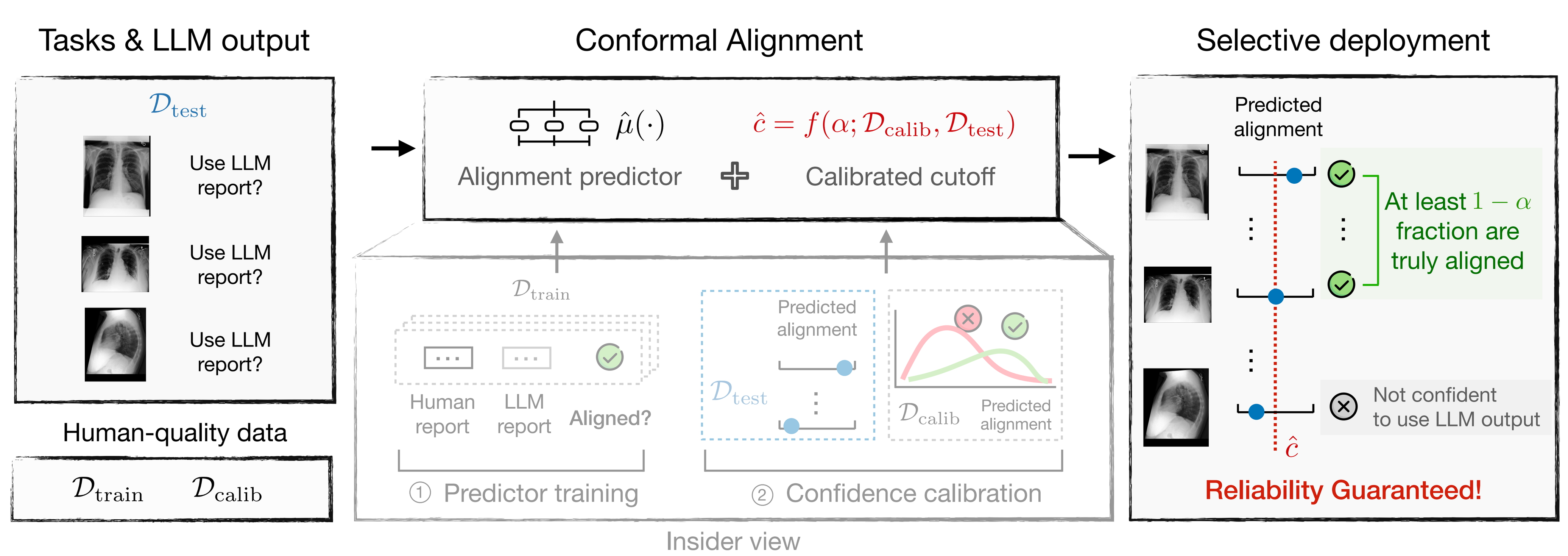
$$\mathrm{FDR}(\mathcal{S}) = \mathbb{E}\left[\frac{\sum_{j\in[m]} \mathbf{1}\{A_{n+j} \le c, j \in \mathcal{S}\}}{\max(|\mathcal{S}|, 1)}\right] \le \alpha$$

while maximizing the selection power

$$\mathrm{Power}(\mathcal{S}) = \mathbb{E}\left[\frac{\sum_{j\in[m]} \mathbf{1}\{A_{n+j} > c, j \in \mathcal{S}\}}{\max(\sum_{j\in[m]} \mathbf{1}\{A_{n+j} > c\}, 1)}\right]$$

Conformal Alignment

1. $\mathcal{D}_{\mathtt{train}}$: fit a prediction model $g(X; f) \approx \mathcal{A}(f(X), E)$
2. $\mathcal{D}_{\mathtt{calib}}$: calculate $\widehat{A}_{n+j} = g(X_{n+j}; f)$ and conformal p-values
$$p_j = \frac{1 + \sum_{i\in\mathcal{D}_{\mathtt{calib}}} \mathbf{1}\{A_i \le c, \widehat{A}_i \ge \widehat{A}_{n+j}\}}{|\mathcal{D}_{\mathtt{calib}}| + 1}$$
3. Conformal Selection [Jin and Candés (2023)] via BH procedure: $\mathcal{S}_{\mathtt{CA}} = \{j \in [m] : p_j \le \alpha k^*/m\}$ with
$$k^* = \max\left\{k \in [m] : p_{(k)} \le \frac{\alpha k}{m}\right\}$$

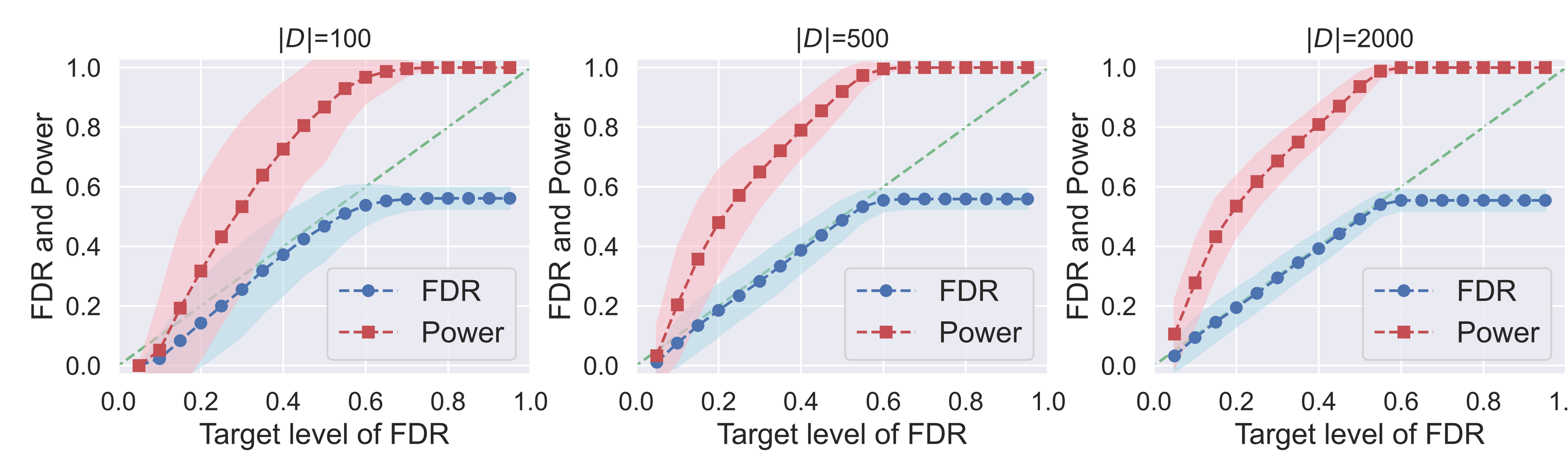


Theoretical Guanrantee

- (**FDR control**) Under *exchangeability* assumption,
$$\mathrm{FDR}(\mathcal{S}_{\mathtt{CA}}) \le \alpha$$
- (**Asymptotic power**) With $H(t) = \mathbb{P}(A \le c, g(X) \ge t)$ and some $t(\alpha)$,
$$\lim_{|\mathcal{D}_{\mathtt{calib}}|, m\to\infty} \mathrm{Power} = \mathbb{P}(H(g(X)) \le t(\alpha) \mid A > c)$$

Results with MIMIC-CXR

- X = X-ray scan, E = reports by human experts
- f : finetuned ViT
- $\mathcal{A}(f(X), E) = \mathbf{1}\{$CheXBert outputs $\ge$ 12 mathces$\}$
- g : `classifier`($A \sim$ `scores`)
  `scores` contain input uncertainty, output confidence, and self-evaluation scores as covariates [Kuhn et al (2023), Kadavath et al (2022), Lin et at (2024)] *(more details [1])*



References

[1] Gui, Y., Jin, Y., and Ren, Z. (2024). Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 34.