Early Time Classification with Accumulated Accuracy Gap Control

Liran Ringel, Regev Cohen, Daniel Freedman, Michael Elad, Yaniv Romano

Early Time Classification

Goal: Given a stream of data $X \in \mathcal{X}$, predict the label Y as early as possible, while maintaining accuracy comparable to that achieved by applying the classifier to the entire input.

Setting:

At timestep $t = 1$, we are exposed to
$$X^{\le 1} = X_1$$

At $t = 2$ we get additional context
$$X^{\le 2} = (X_1, X_2)$$

Until we get the entire stream
$$X = (X_1, X_2, X_3, \dots)$$

We want to reliably stop the inference ASAP, and output a prediction $\hat{Y}^{\text{early}}$

Question: What was the nationality of Ronald Fisher?
Options: **(1)** American **(2)** British **(3)** Canadian **(4)** Australian
Context: Sir Ronald Aylmer Fisher FRS (17 February 1890 – 29 July 1962) was a British polymath who was active as a mathematician, statistician, biologist, geneticist, and academic. For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". In genetics, his work used mathematics to combine Mendelian genetics and natural selection…
Answer: (2) British.

The Halting Problem

- We are handed:

  A pre-trained classifier: $\hat{f}(X^{\le t}) \in [0,1]^K$

  A confidence estimator: $\hat{\pi}(X^{\le t}) \in [0,1]$

  A holdout calibration set: $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$
- We stop the inference process when the model is confident enough in its prediction
- How? Find a confidence threshold $\underline{\hat{\lambda}}_t \in [0,1] \cup \{\infty\}$ for each timestep $t \in \{1, \dots, t_{\max}\}$, and stop the inference if
$$\hat{\pi}(X^{\le t}) \ge \underline{\hat{\lambda}}_t$$
  ⟵ Higher threshold ⟹ Later predictions
- The halt time is given by:
$$\hat{\tau}(X) = \tau_{\underline{\hat{\lambda}}}(X) = \min\{t : \hat{\pi}(X^{\le t}) \ge \underline{\hat{\lambda}}_t \text{ or } t = t_{\max}\}$$

STOP

Accuracy Gap

- Let $\hat{Y}^{\text{full}}$ and $\hat{Y}^{\text{early}}(\hat{\tau})$ be the predicted labels abtained by $\hat{f}(X)$ and $\hat{f}(X^{\le \hat{\tau}(X)})$, resp.
- We define a loss function:
$$L_{\text{gap}}\left(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}(\hat{\tau})\right) = \left(\mathbb{I}_{Y=\hat{Y}^{\text{full}}} - \mathbb{I}_{Y=\hat{Y}^{\text{early}}(\hat{\tau})}\right)_+$$

Accuracy on the entire input

Accuracy with early stopping

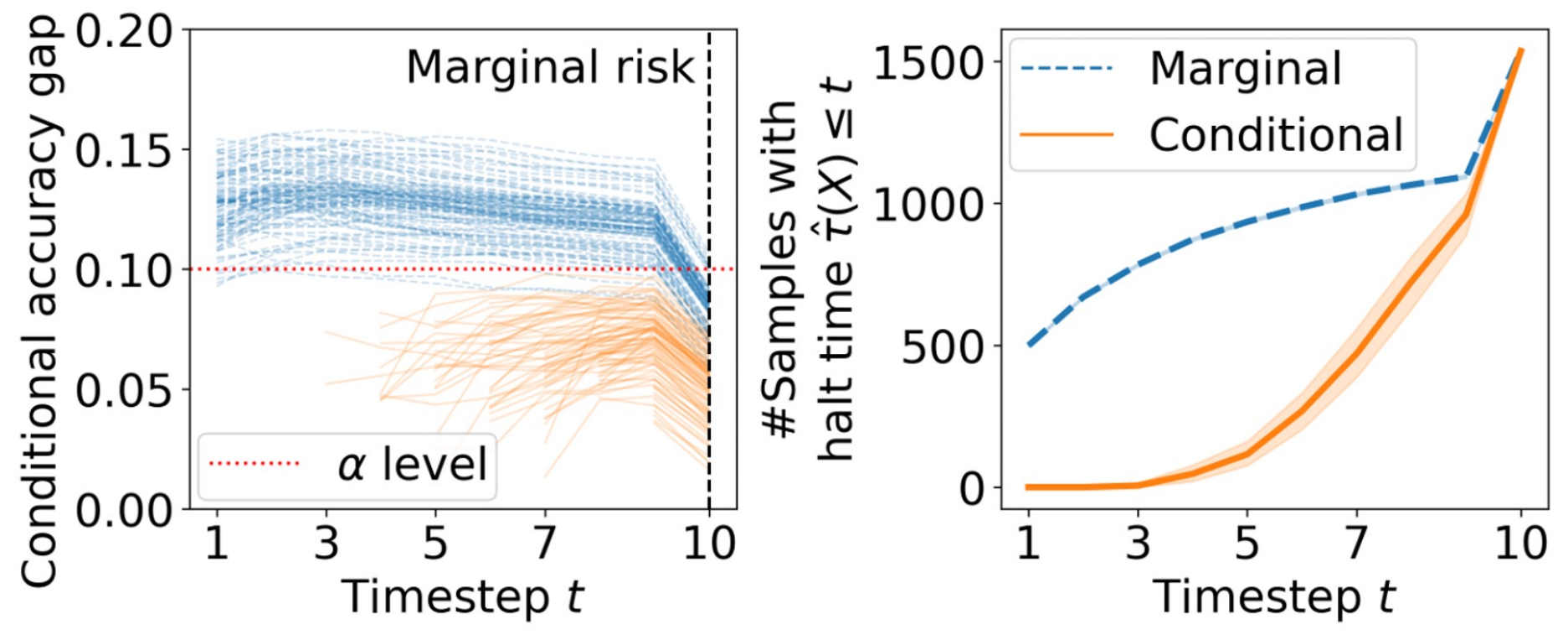
Accuracy Gap Control

- First, we offer a method that **marginally** controls the accuracy gap:
$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(R_{\text{gap}}^{\text{marginal}}\left(\tau_{\underline{\hat{\lambda}}}\right) \le \alpha\right) \ge 1 - \delta, \qquad (1)$$
  where
$$R_{\text{gap}}^{\text{marginal}}\left(\tau_{\underline{\hat{\lambda}}}\right) = \mathbb{E}_{P_{XY}}\left[L_{\text{gap}}\left(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}\left(\tau_{\underline{\hat{\lambda}}}\right)\right)\right]$$
  "With high probability, the proportion of samples where the decision to stop early results in an error is $\le \alpha$."
- Our main contribution: a framework that controls the accuracy gap **conditionally on the accumulated halt times**:
$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(R_{\text{gap}}^{\le t}(\hat{\tau}) \le \alpha \text{ for all } t \ge t_0\right) \ge 1 - \delta, \qquad (2)$$
  where
$$R_{\text{gap}}^{\le t}(\hat{\tau}) = \mathbb{E}_{P_{XY}}\left[L_{\text{gap}}\left(Y, \hat{Y}^{\text{full}}, \hat{Y}^{\text{early}}\right) \mid \hat{\tau}(X) \le t\right]$$
  is the accuracy gap of samples with halt time $\hat{\tau}(X) \le t$, and t_0 is the first timestep for which $\mathbb{P}(\hat{\tau}(\mathrm{X}) \le \mathrm{t}_0) > 0$.

  "The acc. gap is controlled across all accumulated halting times."

The Importance of Conditional Accuracy Gap Control

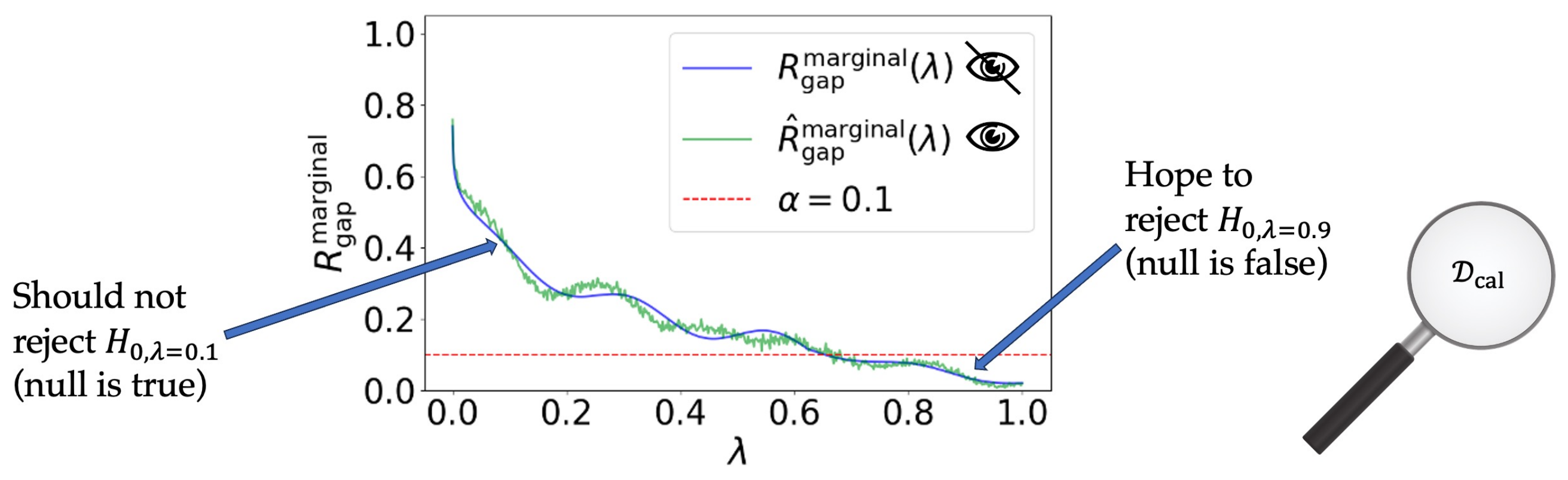
- Task: multiple choice question answering (QuALITY dataset [1])
- Model: Vicuna-13B [2]
- Target accuracy gap: $\alpha = 10\%$
- 100 different splits of calibration and test data



While the marginal method (blue) controls the acc. gap on average among all samples (halt time ≤ 10), the conditional method (orange) controls the accuracy gap also for early halt times

Marginal Method with a Single Threshold

- Want to rigorously tune a hyperparameter λ using the holdout set to achieve marginal accuracy gap control.
- We formalize the null hypothesis [3,4]:
$$H_{0,\lambda}: R_{\text{gap}}^{\text{marginal}}(\tau_\lambda) > \alpha$$
- Based on L_{gap} distribution, we calculate a p-value to test this hypothesis
- **Procedure**: run from the largest λ to the smallest, and return that last λ whose hypothesis was rejected (corresponding p-value is $\le \delta$)



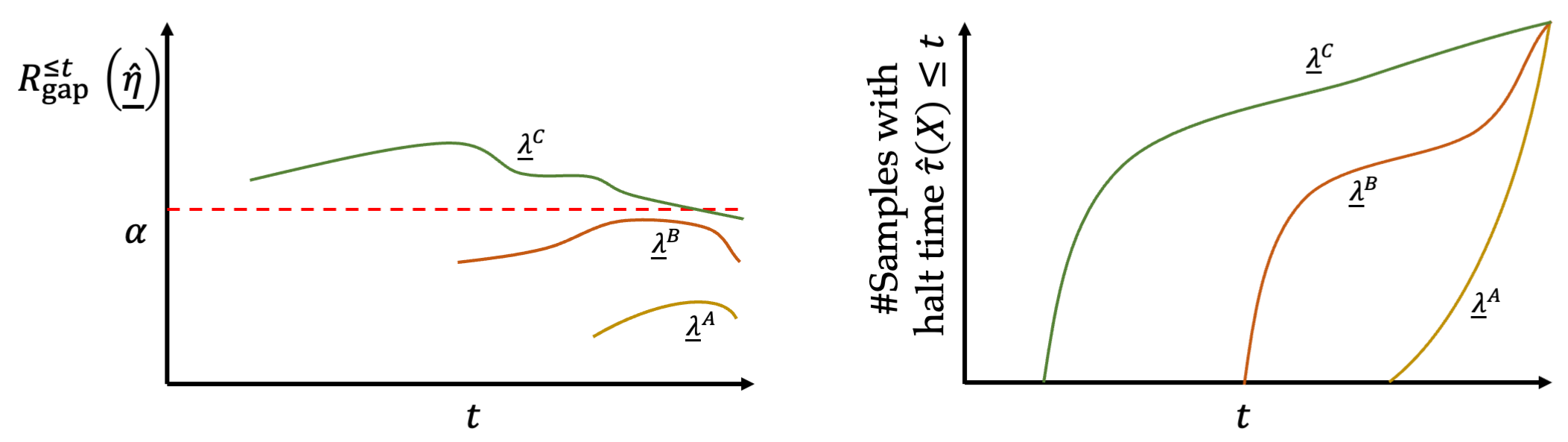
Should not reject $H_{0,\lambda=0.1}$ (null is true)

Hope to reject $H_{0,\lambda=0.9}$ (null is false)

λ scan order. **The predetermined order is crucial to prove that $\hat{\lambda}$ satisfies (1).**

Advancing: Conditional Method with Time-Dependent Thresholds

Want to achieve (2) while utilizing time-dependent thresh. to obtain earlier halt times



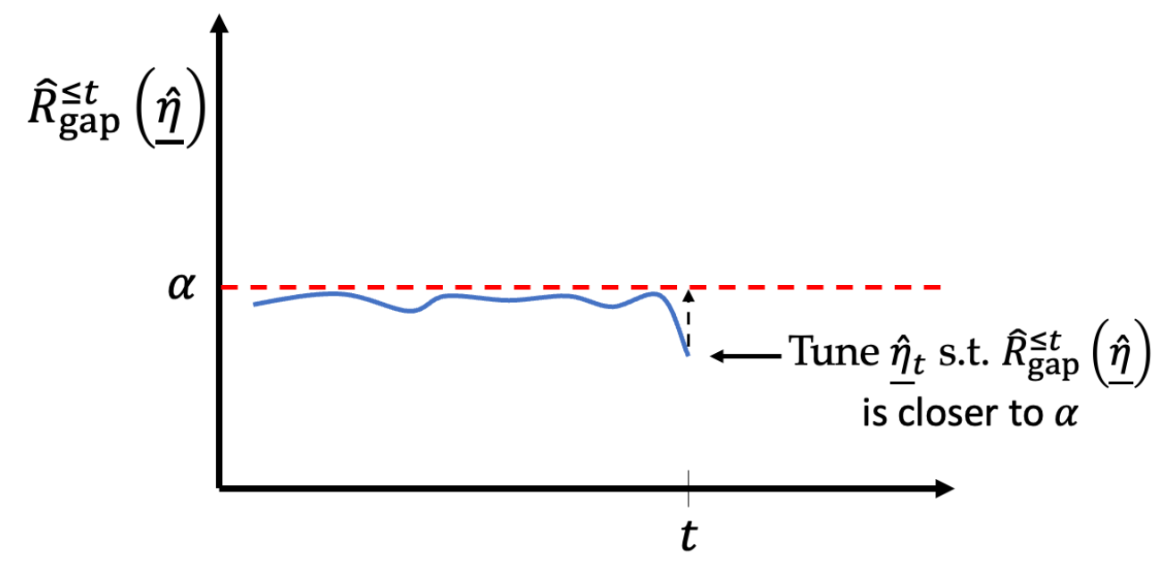
But this flexibility comes with challenges:

1. There are $(|\Lambda| + 1)^{t_{\max}}$ possible configurations for $\underline{\lambda}$, infeasible to check all options
2. Can overfit the calibration data, especially for early timesteps as sample size can be very small

Our solution is a two stage procedure: candidate screening followed by a rigorous testing procedure

Conditional Method – Candidate Screening

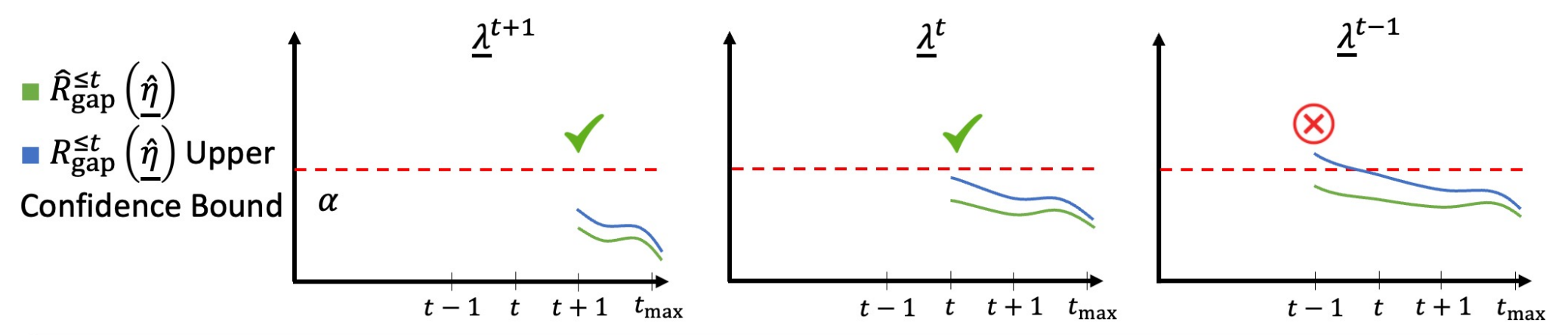
- Use a validation set to heuristically find a data-adaptive threshold vector $\underline{\hat{\eta}}$, with an eye towards early stopping with conditional risk control
- **Greedy method:** start from $\underline{\hat{\eta}}_1$ to $\underline{\hat{\eta}}_{t_{\max}}$. At the t-th step, set $\underline{\hat{\eta}}_t$ to the smallest value s.t. the empirical accumulated accuracy gap up to timestep t satisfies: $\hat{R}_{\text{gap}}^{\le t}\left(\underline{\hat{\eta}}\right) \le \alpha$



Tune $\underline{\hat{\eta}}_t$ s.t. $\hat{R}_{\text{gap}}^{\le t}\left(\underline{\hat{\eta}}\right)$ is closer to α

Conditional Method – Testing

- Form a valid stopping rule $\tau_{\underline{\hat{\lambda}}}$, given the heuristic threshold vector $\underline{\hat{\eta}}$
- **Approach:** a greedy method that utilizes a holdout calibration set, and gradually reveal the time-dependent thresholds from last to first
  - The first candidate is $\underline{\lambda}^{t_{\max}} = \left(\infty, \dots, \infty, \underline{\hat{\eta}}_{t_{\max}}\right)$
  - If $H_{0,\underline{\lambda}^{t_{\max}}}$ was rejected, test for $\underline{\lambda}^{t_{\max}-1} = \left(\infty, \dots, \infty, \underline{\hat{\eta}}_{t_{\max}-1}, \underline{\hat{\eta}}_{t_{\max}}\right)$
  - Continue this process until failing to reject or when all elements of $\underline{\hat{\eta}}$ are revealed



THM: Under the i.i.d assumption, the above procedure attains conditional accuracy gap control with:
✓ Finite number of samples
✓ Any (unknown) distribution
✓ Any predictive model $\hat{f}$

References

- [1] Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. QuALITY: Question answering with long input texts, yes! In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5336–5358, 2022.
- [2] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685, 2023.
- [3] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. "Learn then test: Calibrating predictive algorithms to achieve risk control." arXiv preprint arXiv:2110.01052, 2021.
- [4] Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Consistent accelerated inference via confident adaptive transformers. In Conference on Empirical Methods in Natural Language Processing, pages 4962–4979, 2021.