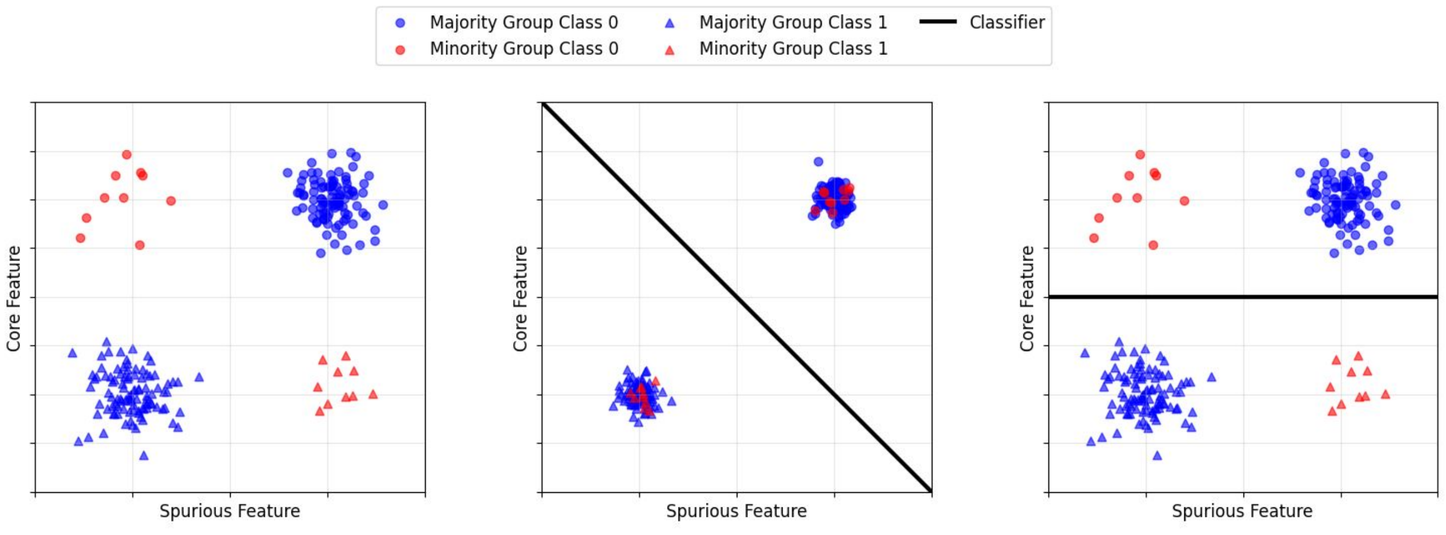
On the Unreasonable Effectiveness of Last-layer Retraining

John C. Hill, Tyler LaBonte, Xinchen Zhang, Vidya Muthukumar
Georgia Institute of Technology

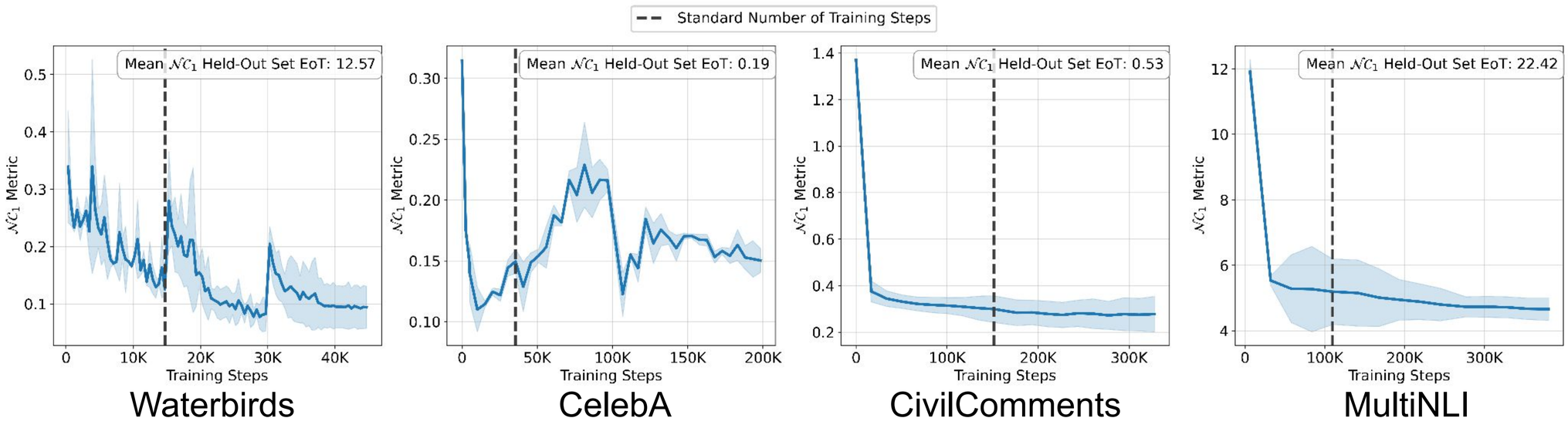
Abstract: The success of last-layer retraining (LLR) methods is determined by their ability to perform explicit (DFR) or implicit (CB-LLR/AFR) group balancing on the held-out set.



Dataset with Majority and Minority Groups | Collapsed Features with Biased Classifier | Held-out Features with Optimal Classifier

Question: How do LLR methods achieve high WGA without group annotations?

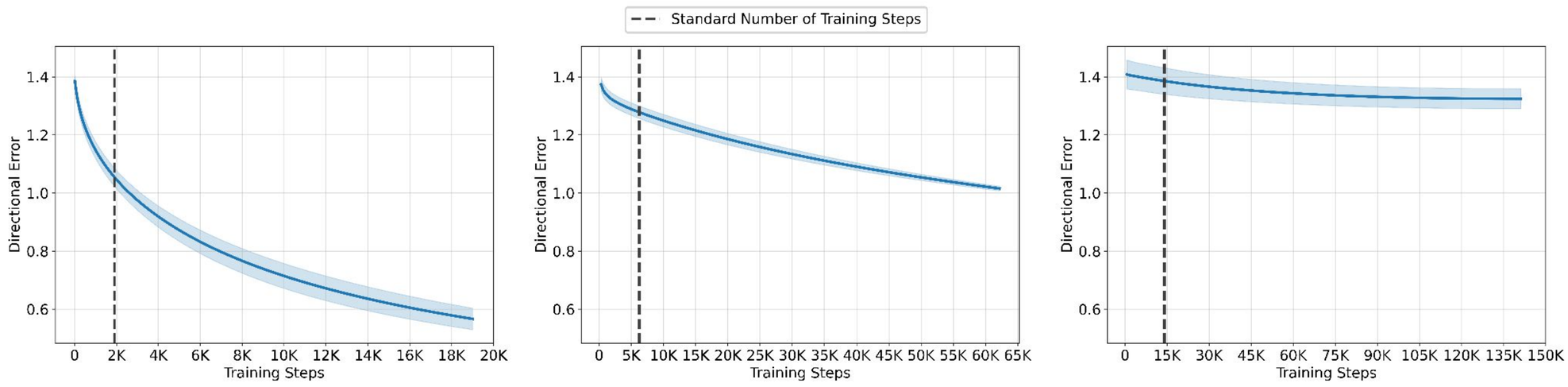
- ERM overfits to **spurious correlations**, harming minority group accuracy
- LLR on a group-balanced held-out set (DFR) is an effective solution to this problem
- Surprisingly, LLR can achieve similar success without group labels [1, 2]



Waterbirds | CelebA | CivilComments | MultiNLI

Hypothesis: Neural collapse (NC) and gradient descent (GD) implicit bias explain LLR

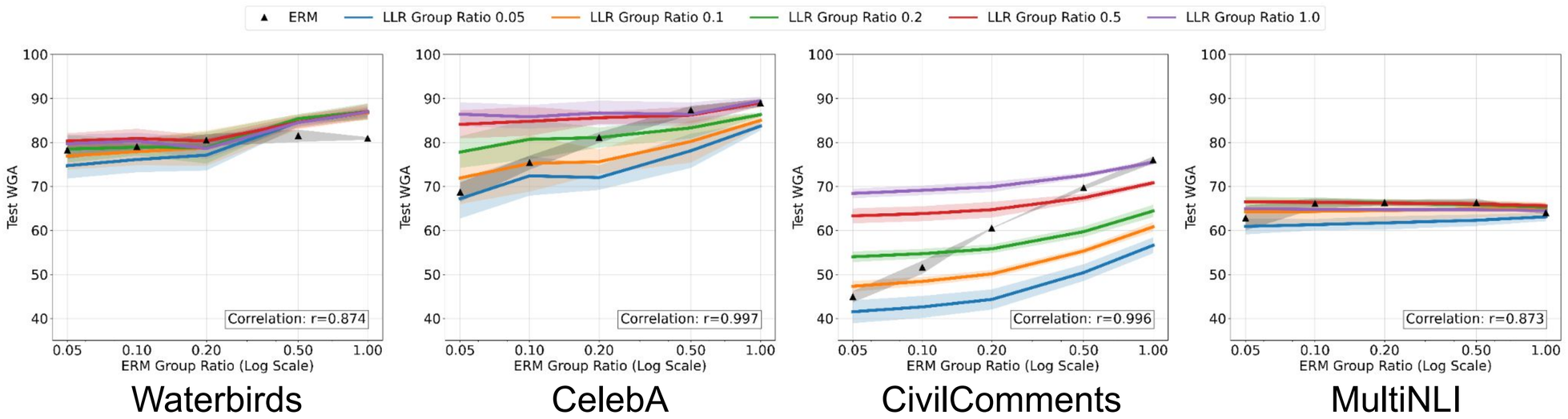
- **Idea 1:** NC occurs during training, causing ERM classifier to be dominated by the majority groups [3]
- **Idea 2:** Implicit bias of GD elicits a max-margin classifier on the held-out features during LLR [4, 5]
- **Hypothesis:** LLR avoids neural collapse (NC) through the held-out set, leading to the implicit bias of gradient descent (GD) benefitting robustness



Waterbirds | CelebA | CivilComments

Our Work: AFR and CB-LLR succeed primarily via *implicit* group-balancing

- Our empirical investigation **does not** support our hypothesis
- NC and GD implicit bias occur **too slowly** to play a significant role during normal training
- Instead, success of AFR [2] and CB-LLR [1] explained by **better group balance** in the held-out set
- LLR only improves over ERM when the held-out set is more balanced than the training set



Waterbirds | CelebA | CivilComments | MultiNLI

References

[1] LaBonte et al. "The group robustness is in the details: Revisiting finetuning under spurious correlations." NeurIPS 2024.
[2] Qiu et al. "Simple and Fast Group Robustness by Automatic Feature Reweighting". ICML 2023.
[3] Papyan et al. "Prevalence of neural collapse during the terminal phase of deep learning training." PNAS 2020.
[4] Soudry et al. " The implicit bias of gradient descent on separable data." JMLR 2018.
[5] Ji & Telgarsky. "The implicit bias of gradient descent on nonseparable data." COLT 2019.