A statistical theory of overfitting for imbalanced classification

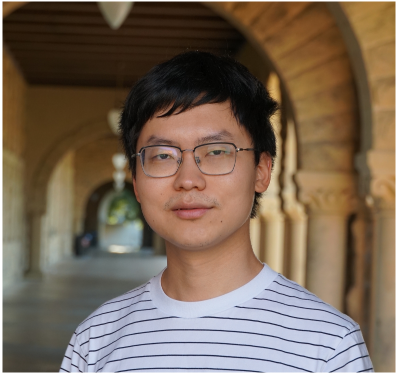
Jingyang Lyu* Kangjie Zhou† Yiqiao Zhong*

*Department of Statistics, University of Wisconsin–Madison
†Department of Statistics, Columbia University

November 7, 2025

Department of Statistics
SCHOOL OF COMPUTER, DATA & INFORMATION SCIENCES
UNIVERSITY OF WISCONSIN–MADISON

Collaborators



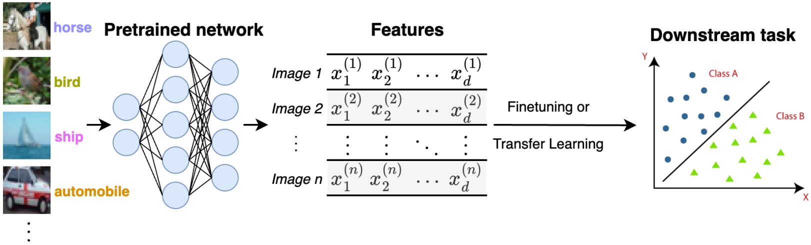
Kangjie Zhou, postdoc at Columbia U

Yiqiao Zhong, UW–Madison

Paper: https://arxiv.org/abs/2502.11323

Challenges in high dimensional imbalanced classification

Training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{x},y}$. Features $\boldsymbol{x}_i \in \mathbb{R}^d$. Binary labels $y_i \in \{\pm 1\}$.



Class imbalance. $\mathbb{P}(y_i = +1) \ll \mathbb{P}(y_i = -1)$. (WLOG, assume "+1" is minority)

Challenge 1: High-dimensional features from pretrained neural networks

Challenge 2: Class imbalance in downstream tasks

Challenges in high dimensional imbalanced classification

- A brief summary of **high dimensional** statistical theory:

	Low dimensions	**High dimensions**
Parameter estimation	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\\|\widehat{\boldsymbol{\beta}}\\|}, \frac{\boldsymbol{\beta}}{\\|\boldsymbol{\beta}\\|} \right\rangle \approx 1$	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\\|\widehat{\boldsymbol{\beta}}\\|}, \frac{\boldsymbol{\beta}}{\\|\boldsymbol{\beta}\\|} \right\rangle < 1$
Generalization	Train error $\approx$ Test error	Train error $<$ Test error

Table: Qualitative comparison for linear classification, $\boldsymbol{\beta}$ is the slope parameter vector.

  — Q: New angles for the (overfitting) effects of dimensionality?

- For **imbalanced** data, the minority has poor accuracy, classical theory and finite-sample correction fail in high dimensions, while the practice is heuristic-driven and ad hoc...

  — Q: How to quantify the impact of factors (imbalance ratio, SNR, dimension) on accuracy?

Empirical phenomenon: Simulation

Settings:

1. Generate a **(linearly) separable** training set from a Gaussian mixture model (GMM):

$$y_i = \begin{cases} +1, & \text{w.p. } \pi \quad \text{(minority)} \\ -1, & \text{w.p. } 1-\pi \quad \text{(majority)} \end{cases}, \qquad \boldsymbol{x}_i \,|\, y_i \sim \mathcal{N}(y_i \boldsymbol{\mu}, \mathbf{I}_d), \qquad i = 1, 2, \dots, n.$$

2. Train a **max-margin classifier (SVM)**: $\Longrightarrow \quad \widehat{\boldsymbol{\beta}}, \widehat{\beta}_0, \widehat{\kappa}$

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \beta_0\in\mathbb{R}, \kappa\in\mathbb{R}}{\text{maximize}} \quad & \kappa, \\ \text{subject to} \quad & y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \geq \kappa, \quad \forall\, 1 \leq i \leq n, \\ & \|\boldsymbol{\beta}\|_2 \leq 1. \end{aligned}$$

3. Visualize the distribution of **logit** $\widehat{f}(\boldsymbol{x}) = \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0$ on both training and testing set for each class $y = \pm 1$.

Simulation results

ELD = empirical (training) logit distribution, **TLD** = testing logit distribution

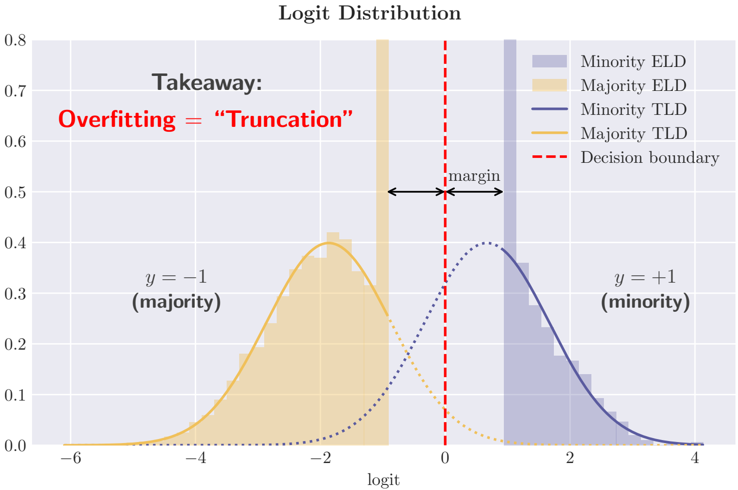


Figure: **Empirical (training) and testing logit distribution for binary Gaussian mixture model**

Real data experiments

Imbalanced classification for **tabular**, **image**, and **text** data

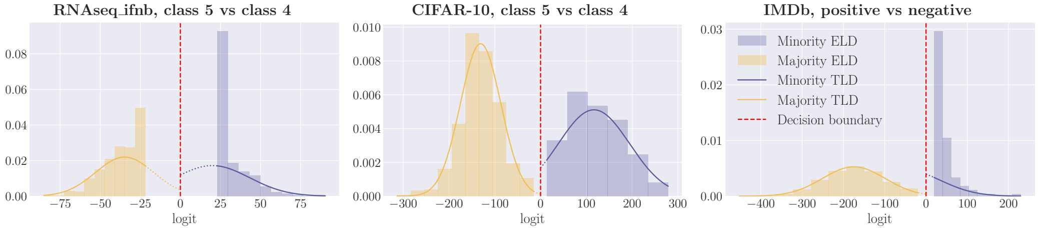


Figure: **ELD and TLD of logistic regression classifier (the last fully-connected layer) for real data**.

- **Left:** IFNB single-cell RNA-seq dataset.
- **Middle:** CIFAR-10 dataset preprocessed by pretrained ResNet-18 model for feature extraction.
- **Right:** IMDb movie review dataset preprocessed by BERT (110M) for feature extraction.

Real data experiments

Downstream task for large language models

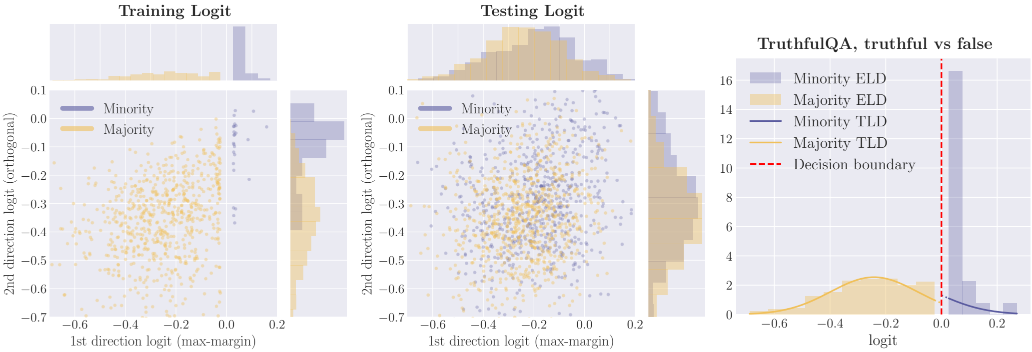


Figure: **ELD and TLD of Llama-3-8B-Instruct activation probing on TruthfulQA dataset. Left, Middle:** Scatter plot on training/testing data for activations (31th layer, 26th head) of truthful (minority) and false (majority) QA pairs after projection onto the top-2 directions. The marginal distributions are shown on the upper and right sides. **Right:** Marginal ELD and TLD on 1st direction.

Theoretical foundation

Theorem (Separable regime, simplified ver.)

Consider GMM with asymptotic regime $n/d \to \delta \in (0, \infty)$. When data is linearly separable w.h.p. ($\delta < \delta_c$), we have the following convergence on logit distribution:

(a) **(Testing logit)** *For a testing point $(\boldsymbol{x}_{\text{test}}, y_{\text{test}})$:*

$$y_{\text{test}} \left(\langle \boldsymbol{x}_{\text{test}}, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0 \right) \xrightarrow{w} \rho^* \|\boldsymbol{\mu}\| + G + Y\beta_0^*.$$

(b) **(Training logit)** *For a training point $(\boldsymbol{x}_i, y_i)$, there is a* **distortion effect** *on the distribution due to dependence between $(\boldsymbol{x}_i, y_i)$ and the classifier function $\widehat{f}$:*

$$y_i \left(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0 \right) \xrightarrow{W_2} \max\left\{ \kappa^*, \rho^* \|\boldsymbol{\mu}\| + G + Y\beta_0^* \right\}.$$

Extension: Non-separable regime

Proximal operator: $\text{prox}_{\lambda\ell}(x) = \arg\min_{t\in\mathbb{R}} \left\{\ell(t) + \frac{1}{2\lambda}(t-x)^2\right\}$, where ℓ is the loss function (e.g., logistic loss)

Overfitting appears as **nonlinear shrinkage** governed by $\text{prox}_{\lambda\ell}$ (Moreau-envelope gradient):

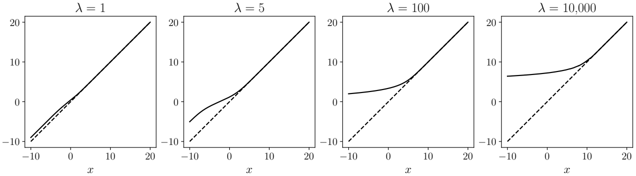


Figure: Plots of proximal operator $x \mapsto \text{prox}_{\lambda\ell}(x)$ where λ represents the strength of overfitting.

	limiting ELD (ν_*)	cause for overfitting (ξ^*)
separable data	$\max\{\kappa^*, \texttt{LOGITS}\}$	$R^*\sqrt{1-\rho^{*2}}\xi^* = (\kappa^* - \texttt{LOGITS})_+$
non-separable data	$\text{prox}_{\lambda^*\ell}(\texttt{LOGITS})$	$R^*\sqrt{1-\rho^{*2}}\xi^* = -\lambda^*\nabla e_{\lambda^*\ell}(\texttt{LOGITS})$
limiting TLD (ν_*^{test})	$\texttt{LOGITS}$	

$\texttt{LOGITS} := \rho^* \|\boldsymbol{\mu}\|_2 R^* + R^* G + Y\beta_0^*$ $\quad$ ($R^* := 1$ in separable case)

Table: Comparison of logit distributions on separable and non-separable data.

Extension: Multiclass

$\widehat{f}_k(\cdot)$ is the logit for label k

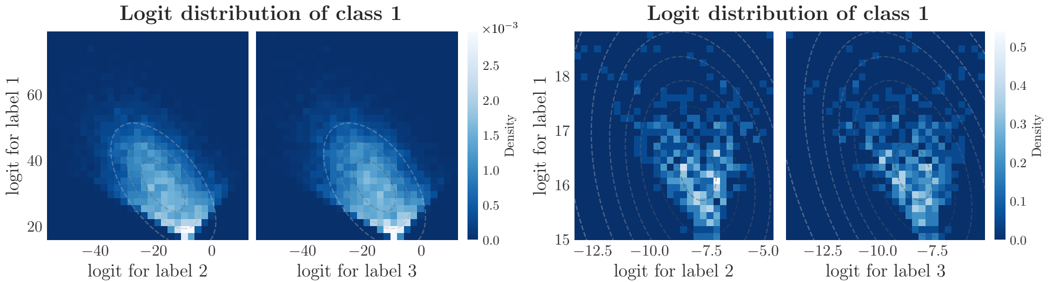


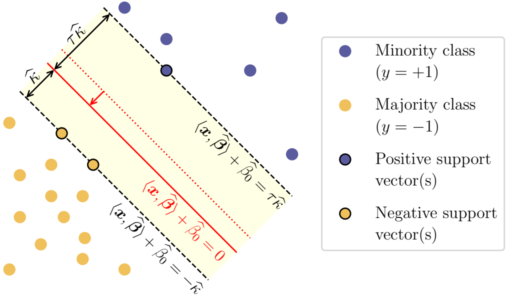
Figure: **Joint empirical logit distributions of multinomial logistic regression.** The heatmaps display empirical joint logits $\left(\widehat{f}_1(\boldsymbol{x}_i), \widehat{f}_k(\boldsymbol{x}_i)\right)$ for features $\boldsymbol{x}_i$ from class 1, where $k = 2, 3$. Overlaid Gaussian density contours (dashed curves) depict testing logit distributions. **Left:** 3-GMM simulation. **Right:** CIFAR-10 image features preprocessed by pretrained ResNet-18.

Rebalancing margin

Rebalancing margin is crucial in separable regime.

Consider **margin-rebalanced SVM**:

$$
\begin{aligned}
\underset{\boldsymbol{\beta}\in\mathbb{R}^d,\beta_0\in\mathbb{R},\kappa\in\mathbb{R}}{\text{maximize}} \quad & \kappa, \\
\text{subject to} \quad & y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \ge \tau\kappa, \quad \forall\, i : y_i = +1 \\
& y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \ge \ \kappa, \quad \forall\, i : y_i = -1 \\
& \|\boldsymbol{\beta}\|_2 \le 1.
\end{aligned}
$$



Margin ratio: $\tau > 0$.

- **Note:** $\widehat{\beta}$ does not depend on τ.
- Question: what is the optimal τ?

Setting 1: proportional regime

Gaussian mixture model with $n/d \to \delta \in (0, \infty)$

Proposition (Proportional regime)

Define $\tau^{\rm opt}$ *as the optimal margin ratio which minimizes the asymptotic* **balanced error**

$$\tau^{\rm opt} := \arg\min_{\tau} \mathrm{Err}_{\rm b}^* = \arg\min_{\tau} \left(\mathrm{Err}_+^* + \mathrm{Err}_-^*\right)/2.$$

(a) *When* $\tau = \tau^{\rm opt}$, *we have* $\beta_0^* = 0$ *and* $\mathrm{Err}_+^* = \mathrm{Err}_-^* = \mathrm{Err}_{\rm b}^*$. *In particular,*

$$\tau^{\rm opt} = \frac{g_1^{-1}\left(\dfrac{\rho^*}{2\pi \|\boldsymbol{\mu}\|_2\, \delta}\right) + \rho^* \|\boldsymbol{\mu}\|_2}{g_1^{-1}\left(\dfrac{\rho^*}{2(1-\pi) \|\boldsymbol{\mu}\|_2\, \delta}\right) + \rho^* \|\boldsymbol{\mu}\|_2}, \qquad \textit{where} \quad \begin{array}{c} g_1(t) = \mathbb{E}[(G+t)_+] \\ G \sim \mathcal{N}(0,1),\ (t)_+ = 0 \vee t \end{array}$$

(b) *When* $\tau = \tau^{\rm opt}$, *the testing error* $\mathrm{Err}_{\rm b}^*$ *is a* **decreasing** *function of* $\|\boldsymbol{\mu}\|_2$ *(signal strength),* δ *(aspect ratio) and* $\pi \in (0, 1/2)$ *(imbalance ratio).*

- When π is small, roughly speaking $\tau^{\rm opt} \asymp 1/\sqrt{\pi}$.

Setting 2: high imbalance

Sub-Gaussian mixture model with $\pi \to 0$, $\|\boldsymbol{\mu}\| \to \infty$, $\delta = n/d \to \infty$

Theorem (High imbalance regime, sub-Gaussian mixture model)

Consider $\pi \asymp d^{-a}$, $\|\boldsymbol{\mu}\|^2 \asymp d^b$, $n \asymp d^{c+1}$, for some $a, b, c > 0$, and $a - c < 1$ (i.e. $n\pi \to \infty$).

(a) **High signal** *(no need for margin rebalancing):* $\boxed{a - c < b}$. *If $1 \le \tau_d \ll d^{b/2}$, then*

$$\mathrm{Err}_+^* = o(1), \qquad \mathrm{Err}_-^* = o(1).$$

(b) **Moderate signal** *(margin rebalancing is crucial):* $\boxed{b < a - c < 2b}$. *If we choose $d^{a-b-c} \ll \tau_d \ll d^{(a-c)/2}$, then*

$$\mathrm{Err}_+^* = o(1), \qquad \mathrm{Err}_-^* = o(1).$$

However, if we naively choose $\tau_d \asymp 1$, then

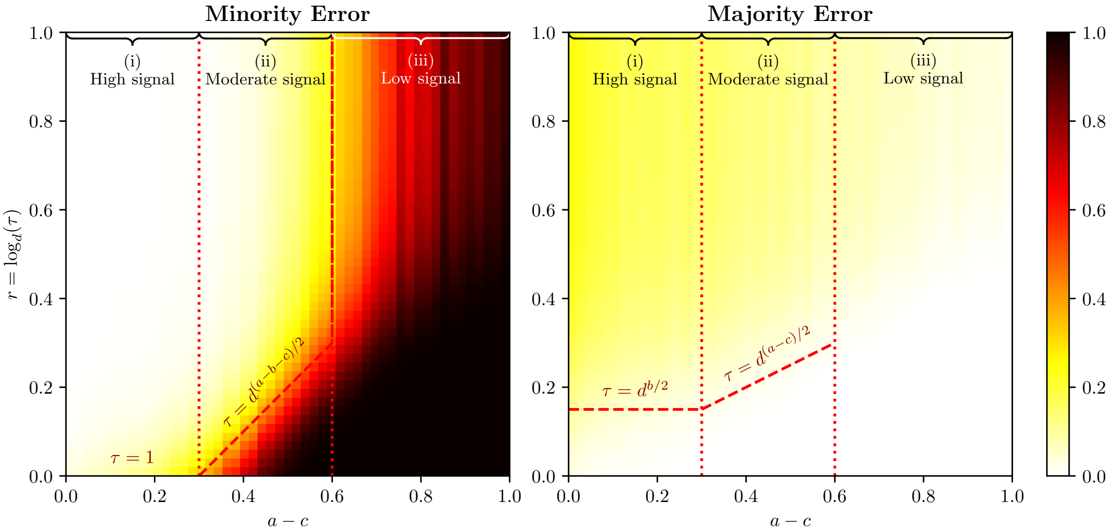
$$\mathrm{Err}_+^* = 1 - o(1), \qquad \mathrm{Err}_-^* = o(1).$$

(c) **Low signal** *(no better than random guess):* $\boxed{a - c > 2b}$. *For any τ_d, we have*

$$\mathrm{Err}_\mathrm{b}^* \ge \tfrac{1}{2} - o(1).$$

Simulation: $\tau = d^r$

$\pi \asymp d^{-a}$, $\|\boldsymbol{\mu}\| \asymp d^{b/2}$, $n \asymp d^{c+1}$ (fix $b = 0.3$, $c = 0.1$, $d = 2000$)



Simulation: Consequences for uncertainty quantification

Setup: 2-GMM, $n = 1,000$, $d = 500$, $\pi = 0.05$, $\|\boldsymbol{\mu}\| = 1$, train SVM with $\tau = \tau^{\mathrm{opt}}$.

Reliability diagrams: For each p (x-axis), calculate $\mathbb{P}(y = 1 \mid \widehat{p}(\boldsymbol{x}) = p)$ (y-axis) on test set.

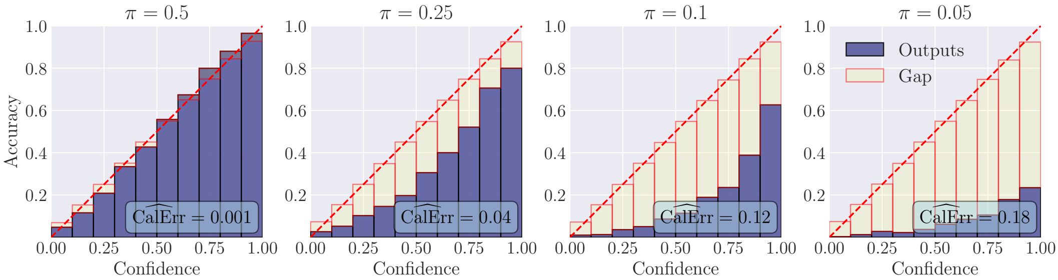


Figure: **Imbalance worsens calibration**.

Summary

Goal 1. Provide a new angle of **characterizing overfitting** for imbalanced classification.

	Low dimensions	**High dimensions**
Parameter estimation	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right\rangle \approx 1$	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right\rangle < 1$
Generalization	Train error $\approx$ Test error	Train error $<$ Test error
Distribution of logits	1D projection of $P_{\boldsymbol{x}}$	Skewed/distorted 1D projection of $P_{\boldsymbol{x}}$

Goal 2. Quantify the **adverse effects** of overfitting, esp. for the minority class.

Increasing dimension $(\delta = \frac{n}{d} \downarrow)$

More severe imbalance $(\pi \downarrow)$

Low signal strength $(\|\boldsymbol{\mu}\| \downarrow)$

$\Longrightarrow$ **More severe overfitting** $\Longrightarrow$

Test accuracy $\downarrow$

Poor calibrated

Thank you for listening.

ArXiv paper

GitHub page

References

- El Karoui, Noureddine, et al. "On robust regression with high-dimensional predictors." Proceedings of the National Academy of Sciences 110.36 (2013): 14557-14562.
- Donoho, David, and Andrea Montanari. "High dimensional robust m-estimation: Asymptotic variance via approximate message passing." Probability Theory and Related Fields 166 (2016): 935-969.
- Sur, Pragya, and Emmanuel J. Candès. "A modern maximum-likelihood theory for high-dimensional logistic regression." Proceedings of the National Academy of Sciences 116.29 (2019): 14516-14525.
- Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." Proceedings of the National Academy of Sciences 116.32 (2019): 15849-15854.
- Bartlett, Peter L., et al. "Benign overfitting in linear regression." Proceedings of the National Academy of Sciences 117.48 (2020): 30063-30070.

References

- Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
- Zhang, Hongyi, et al. "mixup: Beyond Empirical Risk Minimization." International Conference on Learning Representations. 2018.
- Cao, Kaidi, et al. "Learning imbalanced datasets with label-distribution-aware margin loss." Advances in neural information processing systems 32 (2019).
- Montanari, Andrea, and Kangjie Zhou. "Overparametrized linear dimensionality reductions: From projection pursuit to two-layer neural networks." arXiv preprint arXiv:2206.06526 (2022).