Implicit Bias: Literature Overview

Felix Konrad

January 15, 2024

Abstract

The term "Implicit Bias" of an optimization method refers to the phenomenon that among the many parameter configurations with minimal training error, the optimization procedure seems to favour a set of parameters satisfying additional properties, possibly improving performance on unseen data.
In this document we aim to give an introduction to various optimization methods for neural networks and give an overview of "Implicit Bias" results thereof.
Note: This is a work in progress and thus not a complete overview of the literature.

Contents

1 Introduction

2 Optimization

2.1 Theory

2.1.1 The Karush-Kuhn-Tucker Theorem in $\mathbb{R}^d$

Theorem 2.1 (Karush-Kuhn-Tucker I, [Nes+18] Thm 3.1.26)**.** *Let*

$$f, g_i : \mathbb{R}^d \to \mathbb{R}, \qquad i = 1, \ldots, n$$

be convex differentiable functions. Furthermore, suppose Slater's condition

$$\exists x \in \mathbb{R}^d : g_i(x) < 0, \qquad i = 1, \ldots, n.$$

A point x_0 is optimal for the convex optimization problem

$$\begin{aligned} \min \;\; & f(x) \\ s.t. \;\; & g_i(x) \leq 0, \qquad i = 1, \ldots, n \end{aligned}$$

if and only if there exist non-negative $\lambda_i \in \mathbb{R}_{\geq 0}$ such that

$$\begin{aligned} \nabla f(x_0) - \sum_{i=1}^{n} \lambda_i \nabla g_i(x_0) &= 0 \\ \lambda_i \cdot g_i(x_0) &= 0 \qquad i = 1, \ldots, n \end{aligned}$$

Theorem 2.2 (Karush-Kuhn-Tucker II, [Nes+18] Thm 3.1.27)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function with bounded level sets and let $A \in \mathbb{R}^{m \times d}$ be a full-row-rank matrix. Then $x_0 \in \mathbb{R}^d$ is optimal for the convex optimization problem*

$$\begin{aligned} \min \;\; & f(x) \\ s.t. \;\; & Ax = b \end{aligned}$$

if and only if $Ax = b$ and there exists $\mu \in \mathbb{R}^m$ with

$$\nabla f(x_0) - A^T \mu = 0$$

2.2 Algorithms

2.2.1 Gradient Descent

Theorem 2.3 (Convergence of Gradient Descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and continuously differentiable with Lipschitz-continuous gradient, meaning for $x, y \in \mathbb{R}^d$ we have $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. Assume that there exists a minimum $f(x^*)$ of f. Choose a step size $\eta \in (0, \frac{1}{L}]$, then the iterates $(x_k)_{k\in\mathbb{N}}$ obtained by*

$$x_{k+1} := x_k - \eta \nabla f(x_k)$$

satisfy

$$f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\eta k}$$

In particular, $(f(x_k))_{k\in\mathbb{N}}$ converges to the minimum $f(x^*)$.

2.2.2 Stochastic Gradient Descent

Since it is common to have a large amount of data points in many deep learning tasks, it can be too memory consuming to compute the loss gradient with respect to *all* data points simultaneously. The idea, or at least one of the ideas, of stochastic gradient descent (SGD) is to randomly select a data point at each iteration step for which the gradient computation is performed. More precisely, given a dataset $(x_1, y_1), \ldots, (x_N, y_N)$, a model f_θ and a corresponding loss

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i),$$

in each step we sample $i_n \sim \mathrm{Uni}(\{1, \ldots, N\})$ and update

$$\theta_{n+1} = \theta_n - \eta_n \cdot \nabla_\theta \ell(f_{\theta_n}(x_{i_n}), y_{i_n}).$$

For this procedure, we are able to obtain similar convergence guarantees for SGD as we did for gradient descent (Theorem 2.3):

Theorem 2.4 (Convergence of SGD, [GG23] Thm 5.5)**.** *Let f be given by*

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x),$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and ∇f_i is L-Lipschitz. Assume a minimizer x^* of f exists. Let $x_0 \in \mathbb{R}^d$ and let $(x_n)_{n\in\mathbb{N}}$ be the iterates obtained by SGD with step sizes $(\eta_n)_{n\in\mathbb{N}}$. Then If $\eta_n = \eta < \frac{1}{2L}$, then for every $n \geq 1$ we have

$$\mathbb{E}\big[f(\bar{x}_n) - f(x^*)\big] \leq \frac{\|x_0 - x^*\|^2}{2\eta(1 - 2\eta L)n} + \frac{\eta}{1 - 2\eta L} \cdot \mathrm{Var}[\nabla f_i(x^*)],$$

where $\bar{x}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$.

Remark. Theorem 5.5 of [GG23] also includes a convergence Theorem for variable step sizes chosen as $\eta_n := \eta/\sqrt{n+1}$ with $\eta < (2L)^{-1}$.

This Theorem only tells us that $f(\bar{x}_n) \to f(x^*)$. It is not immediately clear whether $f(x_n) \to f(x^*)$ as well and whether the iterates $(x_n)_{n\in\mathbb{N}}$ converge to some minimizer (not necessarily x^*, as there may be multiple minimizers).
To guarantee convergence of the iterates $(x_n)_{n\in\mathbb{N}}$ we need stronger assumptions:

Theorem 2.5.

2.2.3 Mirror Descent

To motivate the Mirror Descent Algorithm, note that the iteration performed in Gradient Descent can be interpreted in the following sense: Let $\eta > 0$ be the step size and let f be the differentiable function we are trying to minimize. Given a current point x_k we find x_{k+1} by minimizing:

$$x_{k+1} := \arg\min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\eta} \|x - x_k\|^2 \right\}.$$

The idea is that for x near x_k, the first-order approximation of f given by $f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$ is quite accurate; Thus, instead of minimizing f (which was our original goal and is not doable in closed form) we minimize the local approximation. The addition of an error term $\frac{1}{2\eta}\|x - x_k\|^2$ prevents us from deviating from x_k too much. (Without this quadratic error term the minimization problem would also be ill-posed)
We can solve this minimization problem explicitly: Taking gradients yields

$$0 \overset{!}{=} \nabla_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\eta} \|x - x_k\|^2 \right\} = \nabla f(x_k) + \frac{1}{\eta}(x - x_k)$$

which rearranges to

$$x = x_k - \eta \nabla f(x_k),$$

which is exactly the iteration performed in Gradient Descent.
Mirror Descent now simply replaces the norm-penalty $\frac{1}{2\eta}\|x - x_k\|^2$ by a *Bregman Divergence*:

Definition 2.1 (Bregman Divergence)**.** Let $\Omega \subseteq \mathbb{R}^d$ be a closed convex set. Let $\psi : \Omega \to \mathbb{R}$ be a strictly convex differentiable function. The Bregman Divergence is defined as

$$D_\psi(x, y) := \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle, \qquad x, y \in \Omega.$$

A few properties immediately follow from Definition 2.1:

- Convexity of ψ implies $D_\psi \geq 0$
- *Strict* convextity of ψ implies $D_\psi(x, y) = 0$ if and only if $x = y$

Now we follow the same approach as before: Given a current point x_k we minimize

$$x_{k+1} := \arg\min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D_\psi(x, x_k) \right\}.$$

Taking gradients yields

$$\begin{aligned} 0 &\stackrel{!}{=} \nabla_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D_\psi(x, x_k) \right\} \\ &= \nabla f(x_k) + \frac{1}{\eta} \Big(\nabla \psi(x) - \nabla \psi(x_k) \Big) \end{aligned}$$

Since ψ is *strictly* convex and differentiable, its gradient is invertible and thus

$$x_{k+1} = (\nabla \psi)^{-1} \big[\nabla \psi(x_k) - \eta \nabla f(x_k) \big]$$

Example 2.1 (Examples of Bregman Divergences)**.** Here we list a few important examples of Bregman Divergences:

1. $\Omega = \mathbb{R}^d$ and $\psi(x) = \frac{1}{2}\|x\|^2$ gives $D_\psi(x, y) = \frac{1}{2}\|x - y\|^2$ and Mirror Descent corresponds to standard Gradient Descent

2. $\Omega = \{x \in \mathbb{R}^d : x_i \geq 0 \text{ and } \sum_{i=1}^d x_i = 1\}$ and $\psi(x) = \sum_{i=1}^d x_i \cdot \log x_i$ gives the *Kullback-Leibler divergence* D_ψ. Elements in Ω are interpreted as probability measures on a set with d elements.

3 Wide Neural Networks

3.1 Neural Tangent Kernel

Neural networks f_θ are, in general, *non-linear* functions of their parameters θ. This makes training difficult, as the loss landscape may be non-convex.
Question: How might we "linearize" a neural network?
One way to do so is to consider the first-order Taylor expansion

$$f_\theta^{lin}(x) := f_{\theta_0}(x) + \nabla_\theta f_{\theta_0}(x)^T \big(\theta - \theta_0\big),$$

where θ_0 is the parameter initialization. This model is a linear model (w.r.t. θ) and will thus have a convex loss landscape. To perform gradient descent on this model

Let ℓ be a loss function and let

$$\mathcal{L}(\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x^{(i)}), y^{(i)}),$$

where $f_\theta : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ is a neural network with L layers with parameters $\theta \in \mathbb{R}^p$. Suppose we initialize θ_0 and update the parameters via gradient flow:

$$\dot{\theta}_t = -\nabla_\theta \mathcal{L}(\theta_t) = \frac{1}{N}\sum_{i=1}^{N} \nabla_f \ell(f_{\theta_t}(x^{(i)}), y^{(i)}) \cdot \nabla_\theta f_{\theta_t}(x^{(i)}),$$

then the corresponding neural network f_{θ_t} evolves as

$$\begin{aligned}\dot{f}_{\theta_t}(x) &= \nabla_\theta f_{\theta_t}(x) \cdot \dot{\theta}_t \\ &= -\frac{1}{N}\sum_{i=1}^{N} \underbrace{\nabla_\theta f_{\theta_t}(x)^T \nabla_\theta f_{\theta_t}(x^{(i)})}_{=:\Theta(\theta_t)(x,x^{(i)})} \nabla_f \ell(f_{\theta_t}(x^{(i)}), y^{(i)})\end{aligned}$$

The matrix Θ is called the *Neural Tangent Kernel*. The name "kernel" is justified by the fact that since $\Theta \in \mathbb{R}^{n_L \times n_L}$ is positive semi-definite, we can view Θ as a kernel $(x, y) \mapsto \Theta(x, y) := x^T \Theta y$.
As this heuristic derivation shows, if we update our parameters via gradient flow, then the evolution of the neural network f_{θ_t} is described by a *kernel gradient descent.*
Understanding the convergence of the neural tangent kernel as the width goes to ∞ will help us understand gradient flow (and gradient descent) for wide neural networks. First, we take a look at what happens during initialization:

Theorem 3.1 (Infinite-Width Neural Networks are Gaussian Processes I, (reference))**.** *...*

3.2 Gradient Descent on Wide Networks

4 Implicit Bias

In this section, we will review selected papers on the topic of "Implicit Bias".

4.1 Linear Models

4.1.1 Implicit Bias of Gradient Descent for Linear Regression

In this section we are going to consider linear models $f_\omega : \mathbb{R}^d \to \mathbb{R}; f(x) := \omega^T x$ where $\omega \in \mathbb{R}^d$ denotes the parameters of the model. The loss function is given by the mean squared error

$$\ell(f_\omega, (x^{(i)}, y^{(i)})_{i=1,\dots,N}) = \sum_{i=1}^{N} \left[f_\omega(x^{(i)}) - y^{(i)}\right]^2.$$

Throughout this, we will assume a fixed dataset $x^{(1)}, \dots, x^{(N)} \in \mathbb{R}^d; y^{(1)}, \dots, y^{(N)} \in \mathbb{R}$ and we will abbreviate $\ell(\omega) := \ell(f_\omega, (x^{(i)}, y^{(i)})_{i=1,\dots,N})$.
A classical result regarding Gradient Descent for such models is that Gradient Descent (with suitable step size) not only converges to a global minimum, but in addition the iterates of Gradient Descent converge to a global minimum *closest to the initial parameters* $\omega^{(0)}$. We make this precise in the following Theorem:

Theorem 4.1 (Implicit Bias of Gradient Descent for Linear Models)**.** *Let* $X^T = (x^{(1)}, \dots, x^{(N)}) \in \mathbb{R}^{d\times N}, y = (y^{(1)}, \dots, y^{(N)}) \in \mathbb{R}^N$ *be any realizable dataset with* $d \geq N$*. Let* $\omega^{(0)} \in \mathbb{R}^d$ *and define*

$$\omega^{(k+1)} := \omega^{(k)} - \eta \nabla \ell(\omega^{(k)}).$$

Then for small enough $\eta > 0$ *the sequence* $(\omega^{(k)})_{k\in\mathbb{N}}$ *converges to a global minimum* ω^* *solving the following optimization problem*

$$\begin{aligned} \min_{\omega} \ & \|\omega - \omega^{(0)}\|^2 \\ s.t. \ & X\omega = y \end{aligned} \tag{MIN}$$

Proof. The proof is quite elegant and reveals a strategy that will be used repeatedly when proving "Implicit Bias" results; hence we will include it here:
Note that by the Karush-Kuhn-Tucker Theorem (Theorem 2.2) we can characterize an optimal solution ω^* to the optimization problem (MIN): A point $\omega^* \in \mathbb{R}^d$ is optimal for (MIN) if and only if it satisfies:

$$X\omega^* = y, \qquad \exists \lambda \in \mathbb{R}^N : (\omega^* - \omega^{(0)}) + X^T\lambda = 0.$$

Now we need to prove two things:

1. The iterates $(\omega^{(k)})_{k\in\mathbb{N}}$ converge to a point ω^* with 0 loss, i.e. $X\omega^* = y$: Note that f_ω is linear in ω and thus $\ell(\omega)$ is strictly convex and differentiable with Lipschitz gradient $\nabla_\omega \ell$. Hence, by Theorem 2.3 the convergence of Gradient Descent to a global minimum follows. Since $d \geq N$, there exists an exact interpolation of the data points, meaning $\ell(\omega^*) = 0$.

2. The iterates $(\omega^{(k)})_{k\in\mathbb{N}}$ satisfy the second KKT condition: Note that

$$\omega^{(k)} - \omega^{(0)} = \eta \sum_{l=0}^{k-1} \nabla \ell(\omega_l) = 2\eta \sum_{l=0}^{k-1} X^T (X\omega^{(l)} - y)$$

and hence for $\lambda = -2\eta \sum_{l=0}^{k-1} (X\omega^{(l)} - y)$ we see that the second KKT condition is satisfied. By continuity, the limiting parameters $\omega^* = \lim_{k\to\infty} \omega^{(k)}$ also satisfies this condition.

Combining these two claims gives the stated result. □

4.1.2 "Characterizing Implicit Bias in Terms of Optimization Geometry"

This section is based on the paper [Gun+18].
In Section 2.2.6 we have seen how mirror descent can be seen as a generalization of gradient descent. In Section 4.1.1 we have characterized the implicit bias of gradient descent for least squares regression. Now [Gun+18] generalize this theorem to mirror descent for *losses with a unique finite root*:

Definition 4.1 (Losses with a unique finite root)**.** We say ℓ is a loss function with a unique finite root if for all y and any sequence $(y_n)_{n\in\mathbb{N}}$ we have $\ell(y_n, y) \to \inf_{\hat{y}} \ell(\hat{y}, y)$ if and only if $y_n \to y$.

The mean squared error loss function is a special case of a loss function with a unique finite root, as any L_p-loss ($p \geq 1$) or the Huber loss. However, common loss functions like the logistic loss of the exponential loss are not within this class. Such loss functions will be dealt with later in this chapter.
We now come to the main theorem regarding mirror descent for losses with a unique finite root, which is to be seen as a generalization of Theorem 4.1:

Theorem 4.2 (Implicit Bias of Mirror Descent for Losses with a unique finite root, [Gun+18] Thm 1)**.** *Let ℓ be a loss function with unique finite root, $(x_n, y_n)_{n=1,\dots,N}$ be a realizable dataset. Let ψ be a strictly convex potential. Consider the linear model $f_\omega(x) = \langle \omega, x\rangle$. Assume that the iterates $(\omega_n)_{n\in\mathbb{N}}$ obtained by mirror descent starting at ω_0 with step sizes $(\eta_n)_{n\in\mathbb{N}}$ satisfy*

$$\lim_{n\to\infty} \omega_n = \omega_\infty \text{ exsits and satisfies } \ell(f_{\omega_\infty}(x_n), y_n) = 0 \forall n.$$

Then ω_∞ is the solution to the following optimization problem:

$$\begin{aligned} \min_{\omega\in\mathbb{R}^d} \ & D_\psi(\omega, \omega_0) \\ s.t. \ & f_\omega(x_n) = y_n, \quad n = 1, \dots, N \end{aligned}$$

In particular, if we let ω_0 be a minimizer of ψ s.t. $\nabla\psi(\omega_0) = 0$, then ω_∞ is again a minimizer of ψ with $\langle \omega_\infty, x_n\rangle = y_n$.

Proof. The proof is completely analogous to the proof of Theorem 4.1. Note that since ψ is convex, the function $\omega \mapsto D_\psi(\omega, \omega_0)$ is convex as well and hence the minimization problem given above is a convex minimization problem. Applying the Karush-Kuhn-Tucker Theorem (Theorem 2.2) tells us that ω_∞ is a minimizer if and only if

$$\langle \omega_\infty, x_n \rangle = y_n \text{ for all } n \text{ and } \exists \lambda \in \mathbb{R}^N : \nabla\psi(\omega_\infty) - \nabla\psi(\omega_0) - \sum_{n=1}^{N} \lambda_n x_n = 0.$$

By assumption, the iterates of mirror descent converge to a point ω_∞ with 0 loss, hence the first condition is satisfied.
To verify the second condition, we will show that every iterate ω_n of mirror descent satisfies this condition, hence by convergence of $\omega_n \to \omega_\infty$ and by continuity of the condition it follows that ω_∞ satisfies the second condition as well. Recall that ω_{n+1} satisfies

$$\nabla\psi(\omega_n) = \nabla\psi(\omega_{n-1}) - \eta_n \nabla\mathcal{L}(\omega_{n-1})$$

and thus by a telescopic sum

$$\nabla\psi(\omega_n) - \nabla\psi(\omega_0) = -\sum_{k=1}^{n} \eta_k \nabla\mathcal{L}(\omega_k) = -\sum_{k=1}^{n} \eta_k \sum_{i=1}^{N} \ell'(\omega_k^T x_i, y_i) \cdot x_i \in \operatorname{span}(x_1, \ldots, x_N),$$

where ℓ' denotes the derivative of ℓ in the first component. Thus the second KKT condition holds for every ω_n. □

Remark. Note that unlike Theorem 4.1, in Theorem 4.2 we *assume* the convergence of mirror descent for given step sizes $(\eta_n)_{n\in\mathbb{N}}$, but no convergence guarantee is given for any sequence of step sizes. Hence the assumption of Theorem 4.2 are stronger than those of Theorem 4.1.

Strictly Monotone Losses

Definition 4.2 (Strictly Monotone Losses)**.** We call ℓ a strictly monotone loss function if ℓ is bounded from below and for all y the map $\hat{y} \mapsto \ell(\hat{y}, y)$ is strictly monotonically decreasing (w.r.t. $\hat{y}$)
We further assume, without loss of generality, that $\inf_{\hat{y}} \ell(\hat{y}, y) = 0$ for all y and that $\ell(\hat{y}, y) \to 0$ as $\hat{y} \cdot y \to +\infty$.

4.1.3 "The Implicit Bias of Gradient Descent on Separable Data"

This section is based on [Sou+18], which deals with the implicit bias of gradient descent for linear support vector machines. In particular, given a dataset $x_1, \ldots, x_N \in \mathbb{R}^d$ with labels $y_1, \ldots, y_N \in \{-1, +1\}$ we consider the homogeneous linear model $f_\omega(x) := \omega^T x$ (no bias term). The goal of SVMs is to find ω

such that $\text{sgn}(w^T x_i) = y_i$ for all i. The hard-margin formulation of this problem reads

$$\begin{aligned} \min_{\omega} \quad & \frac{1}{2}\|\omega\|_2^2 \\ s.t. \quad & y_i \cdot \omega^T x_i \geq 1, \quad i = 1, \dots, N \end{aligned} \tag{H-SVM}$$

The main result of [Sou+18] states that for linearly separable datasets and a specific class of loss-functions (β-smooth, decreasing, with an exponential tail) the iterates of gradient descent $(\omega_n)_{n\in\mathbb{N}}$ roughly behave like $\omega^* \cdot \log t$, where ω^* is the solution to the hard-margin SVM formulation (H-SVM). In particular their limiting direction is given by that of ω^*, meaning

$$\lim_{n\to\infty} \frac{\omega_n}{\|\omega_n\|} = \frac{\omega^*}{\|\omega^*\|}.$$

We now give the relevant definitions, precisely state the main result and give an outline for its proof:

Definition 4.3 ([Sou+18] Def. 2)**.** A function $f(u)$ has a "tight exponential tail", if there exist positive constants c, a, μ_+, μ_-, L, l, such that

$$\begin{aligned} \forall u > L : & f(u) \leq c(1 + \exp(-\mu_+ u))e^{-au} \\ \forall u > l : & f(u) \geq c(1 - \exp(-\mu_- u))e^{-au} \end{aligned}$$

Both the exponential loss $\ell(u) := \exp(-u)$ as well as the logistic loss $\ell(u) := \log(1 + \exp(-u))$ have tight exponential tails.

Theorem 4.3. *Assume our dataset is linearly separable. Let the loss function ℓ satisfy:*

1. *ℓ is decreasing.*
2. *ℓ is β-smooth, meaning ℓ' is β-Lipschitz.*
3. *ℓ has a tight exponential tail.*

Then if $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$, we have that for any starting point ω_0, the gradient descent iterates $(\omega_n)_{n\in\mathbb{N}}$ satisfy

$$\omega_n = \omega^* \log n + \rho(n),$$

where ω^* is the solution to (H-SVM) and ρ grows at most as $\|\rho(n)\| = \mathcal{O}(\log\log n)$. In particular,

$$\lim_{n\to\infty} \frac{\omega_n}{\|\omega_n\|} = \frac{\omega^*}{\|\omega^*\|}.$$

4.2 Neural Networks

4.2.1 Linear Neural Networks

4.2.2 "Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks"

This section is based on the paper [JM20]

References

[Gun+18] Suriya Gunasekar et al. "Characterizing implicit bias in terms of optimization geometry". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1832–1841.

[Nes+18] Yurii Nesterov et al. *Lectures on convex optimization*. Vol. 137. Springer, 2018.

[Sou+18] Daniel Soudry et al. "The implicit bias of gradient descent on separable data". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2822–2878.

[JM20] Hui Jin and Guido Montúfar. "Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks". In: *arXiv preprint arXiv:2006.07356* (2020).

[GG23] Guillaume Garrigos and Robert M Gower. "Handbook of convergence theorems for (stochastic) gradient methods". In: *arXiv preprint arXiv:2301.11235* (2023).