

Addressing Data Sovereignty and Empowering Users by Promoting Trustworthy Low Resource AI  
Systems

Matthew Pon

A thesis

submitted in partial fulfillment of the  
requirements for the degree of

Master of Public Health

University of Washington

2024

Committee:

Carolin Spice

Tanya E. Karwaki

Program Authorized to Offer Degree:

Health Services

©Copyright 2024  
Matthew Pon

**Abstract**

Addressing Data Sovereignty and Empowering Users by Promoting Trustworthy Low Resource AI Systems

Matthew Pon

Chair of the Supervisory Committee:

Carolin Spice

Department of Health Systems and Population Health

In the last few years, advancements in artificial intelligence (AI) have dramatically transformed the digital world, with AI tools being integrated across a multitude of industries. The widespread adoption of Large Language Models (LLMs) has led to numerous benefits, such as improved data analysis, customer support, and plain language explanations. However, the proliferation of LLMs in digital services has also raised concerns related to cost, environmental impact, privacy, and algorithmic fairness. This research explores if a locally trained and run low-rank adaptations (LoRAs) can enable community-based organizations to create AI tools that can fine tune LLMs and address their specific needs while mitigating concerns around privacy, algorithmic fairness, cost, and environmental impact. Furthermore, this research provides guidelines for low-resource organizations to adopt this AI tool on local hardware.

## Key terms

**7B\_Vicuna:** 7 billion parameter Vicuna LLM with a LoRA applied

**7B\_Vicuna\_LoRA:** 7 billion parameter Vicuna LLM with a LoRA applied

**13B\_Vicuna:** 13 billion parameter Vicuna LLM with a LoRA applied

**13B\_Vicuna\_LoRA:** 13 billion parameter Vicuna LLM with a LoRA applied

**CPU:** Central processing unit (CPU) is the fundamental component of a computer system responsible for carrying out instructions and executing tasks.

**GPT:** Generative Pre-trained Transformer (GPT) is a type of large language model developed by the company OpenAI

**GPU:** A graphics processing unit (GPU) is specialized hardware designed to perform complex computations, such as those required by large language models, in parallel rather than sequentially, allowing for faster processing times.

**GUI:** Graphical User Interface (GUI) uses visual elements to more intuitively interact with digital systems.

**LLM:** Large language model (LLM) used to generate textual data and responses. They are often used as chat-bots and virtual assistants.

**LoRA:** Low-rank adaptation (LoRA) is an adapter that fine tunes a LLM through rank decomposition.

**ML:** Machine learning (ML) is a type of artificial intelligence that focuses on enabling systems to learn from and make predictions based on data without being explicitly programmed.

**RAM:** Random access memory (RAM) is a type of volatile memory found in computers, that temporarily stores data for the CPU before it is accessed or modified.

**SVD:** Singular Value Decomposition (SVD) is a mathematical technique used to reduce the dimensionality of data making it easier and more efficient to work with.

**VRAM:** Video random access memory (VRAM) refers to RAM on a GPU specifically designated for storing data specifically for the GPU.

## Table of Contents

1. Introduction .....	5
2. Methods .....	10
3. Results .....	13
4. Discussion .....	16
Appendix	
Setup Guide .....	19
Training Guide .....	22
References .....	25

# 1. Introduction

## *1.1 Background*

In the last few years artificial intelligence (AI) has been integrated in the digital world with AI tools being used in various fields and applications, spanning from science, technology, industry, and medical to digital AI powered virtual assistants.<sup>1,2</sup> Machine learning (ML) capabilities have been developed by researchers to aid in automation tasks like natural language processing, sentiment analysis, speech generation, and image classification.<sup>3,4,5</sup> Most recently, generative pre-trained transformers (GPT) have generated considerable attention for their ability to produce human-like textual interaction, the most recently notable being ChatGPT by OpenAI.<sup>6</sup> Generative AI, like ChatGPT, can create images, audio, and human-like text based on training data. The overall positive reception of machine powered chatbots by both businesses and consumers has prompted large technology companies like Microsoft, Apple, Google, and Amazon to rapidly integrate Large Language Model (LLM) powered chat assistants into their products.<sup>7,8</sup> LLMs have even become so ubiquitous in the digital world that other industries, such as healthcare, finance, entertainment, and government, have started integrating LLMs and ML to improve overall customer experience.<sup>9,10,11,12</sup> LLMs are poised to become increasingly important tools that we use to interact with and communicate through the digital world. However, it is essential to consider both the potential benefits and concerns posed by widespread adoption of AI tools.

LLMs and ML algorithms have the potential to deliver several benefits. One of these benefits is being able to process and analyze large amounts of data at once while minimizing noise.<sup>13</sup> Companies are leveraging this ability to analyze data and provide near real time recommendations to customers and small businesses.<sup>9</sup> Using this strategy, LLMs have been used to provide on-demand customer support where end users can ask questions in natural language and receive timely responses, reducing the need for human intervention. Another benefit, in the context of health and well-being, is improving mental health through LLM interaction.<sup>14</sup> When talking to a machine, people experience less social stigma since a machine is not capable of passing judgment. This allows users to communicate freely where they can experience positive emotional outcomes regardless of the conversation occurring between two humans or between a human and a LLM.<sup>15</sup> Finally, LLMs possibly have the potential to help tackle plain language explanations when users are faced with overly technical or legal definitions. According to the United States General Services Administration, plain language is a type of

communication that aims to correctly explain a concept that is easy to understand for most people, regardless of education or background.<sup>16</sup> In practice, writing in plain language could prove difficult given how perception of reality can be changed based on a wide variety of circumstances like culture, emotion, and society.<sup>17,18</sup> Having a LLM explain a concept and adapt responses to different individual users can help potentially mitigate perceptual differences.

However, the integration and operation of LLMs in digital services may raise concerns related to cost and environmental impact.<sup>19,20,21</sup> The cost of training LLMs are largely due to the necessity of having high-performance computing infrastructure to handle the complex algorithms.<sup>22</sup> In an interview, OpenAI's CEO Sam Altman said GPT-4 cost more than \$100 million to train.<sup>23,24,25</sup> According to the 2024 Artificial Intelligence Index Report by Stanford University's Institute for Human-Centered Artificial Intelligence, Google's Gemini Ultra cost \$191 million to train.<sup>26</sup> In addition to the high energy consumption, high-performance computing infrastructure can result in substantial carbon emissions.<sup>27,28</sup> As of May 2024, Meta has yet to release an official research paper about the training cost of LLaMA-3 but it is estimated from third party sources that Meta invested around \$14 billion in Nvidia H100 GPUs for training its LLaMa-3 LLM.<sup>29,30</sup> Nvidia's own product page reported specification for the H100, the GPUs that Meta is purchasing, states a maximum rated utilization of 350W of power, with one variant capable of up to 700W.<sup>31</sup> The concern comes when Meta reports purchasing 350,000 of the 350W capable cards.<sup>32</sup> This adds up to  $1.225 \times 10^5$  kW of power, which Meta was likely running for weeks.

In addition, when interacting with LLMs it is important to consider the potential privacy implications and concerns around data collection and data sovereignty. When interacting with online services, any data uploaded or online activity may be logged and can be used in court.<sup>33</sup> Meta's transparency report showed that the US leads the world in government requests for user data. Meta reported, in their 2020 transparency report, that they complied with 88.07% of 236,928 requests the US government made for user data.<sup>34</sup> This is particularly concerning when we consider our online activity, including internet searches and browsing history, can be compiled to profile individuals, adjust credit scores, predict our habits, and allow companies to tailor interactions to influence our behaviors.<sup>35,36,37,38</sup> When communicating with LLMs people have a greater potential to share even more information than they would typically share with an average internet search. This means that chat bots or digital assistants, that exhibit human-like characteristics, are more likely to compel users to share personal information.<sup>39,40</sup> Moreover, the user data included in training data can be reverse engineered and

exposed.<sup>41</sup> AI assistants powered by human-like LLMs are poised to expose unprecedented levels of personal data from everyday digital interactions. There are no current commercial LLM solutions that offer complete transparency or ownership for the data you provide. Furthermore, even if de-identified, training data in LLMs can be leaked or reverse engineered from the model.<sup>43</sup>

“With web browsing and searches, you’re talking about some of the most sensitive, most personal, and most private details of Americans’ lives. Every thought that can come into people’s heads can be revealed in an internet search or a visit to a website. Their health history and medical concerns. Their political views. Their romantic lives and friendships. Their religious beliefs. Collecting this information is as close to reading minds as surveillance can get. It is digital mining of the personal lives of Americans.”

-US Senator Ron Wyden of Oregon.<sup>42</sup>

Finally, there is little control on the training data companies choose to include to train LLMs. Incomplete or bias training data can discriminate against populations and perspectives. AI has been known to exhibit algorithmic discrimination.<sup>43,44,45</sup> AI models can unfairly treat individuals or groups of individuals based on prejudiced or incomplete training data. Training bias can also impact LLM responses. Selecting against or preferentially including certain texts into the training dataset can shape responses. Various user reports show this can manifest as discrimination, toxicity, or contextual insensitivity.<sup>43,46,47,48</sup> This has the potential to further disadvantage vulnerable populations and reinforce inequality.

Unfortunately it is cost prohibitive, and environmentally irresponsible, to train custom LLMs that cater to every community. This means that smaller organizations with fewer resources are compelled to use commercial LLM solutions that could discriminate against the populations they serve. These commercial solutions also risk revealing sensitive information. A potential solution to this problem would be to locally train a low-rank adaptation (LoRA) using local hardware that never connects to the internet.

In context, LoRAs are essentially additional LLM fine tuning which can help prevent irrelevant or bias responses by allowing customization of a generative LLM to a specific task or community perspective. Since LLMs have fixed weights which can not change after being trained, LoRAs provide a way for users to customize a LLM. To accomplish this, a LoRA adds new trainable weights to an existing LLM. Then, LoRAs use compression so that all the new training weights are smaller than the

original LLM it was trained from. LoRAs are incredibly space efficient for the amount of data and guidance they provide. They achieve this data efficiency through truncated singular value decomposition (SVD) and the final size of LoRAs can be orders of magnitude smaller than the original weights.<sup>49</sup> SVD is a factorization technique from linear algebra that has been used to compress images and signals without much loss to the original data.<sup>50</sup> SVD is used to compress the new trained weights of the LoRA. The smaller LoRA is then reinjected into the LLM at the feed forward layer to influence output responses. The result is a very small and efficient way to influence a much larger ML model.

The approach of adding new trainable layers to a larger model has been used in automatic speech recognition. Fine tuning using a LoRA, the addition of trainable layers, has been used to help improve recognition for larger speech recognition models to low-resource, less represented, languages and communities.<sup>51,52</sup> This improved the recognition rate of languages outside of the original training data that the models were designed for. Fine tuning also added enough data to allow speech recognition in specialized fields with unique lexicons.<sup>53</sup> LoRAs allow for additional data to be incorporated into pre-trained models for automatic speech recognition or even LLMs.

Finally, LLMs are able to converse with users since they are trained to predict, then generate, the next sequence of text. LLMs learn to predict text based on large amounts of text that has been gathered and cleaned.<sup>54,55</sup> The large amount of text fed into the LLM means each response from the LLM represents a product of the available training text fed into it. Even if specific data is included in the original training dataset, a phenomenon called “catastrophic forgetting” can occur where large portions of the original dataset isn’t being represented in the final trained LLM.<sup>56</sup> This can manifest as averaged responses where specific data, even if included in the original dataset, isn’t being recalled or reproduced. LoRAs are a specific way to preserve the original data while further adding additional data.

## *1.2 Significance*

Locally training a LoRA allows communities to have a voice in how they are being represented during interactions with LLM tools. By including their data to further fine tune a LLM through a LoRA, communities can have their voices and perspectives included. Without a LoRA, LLM tools only present the views and perspectives for the dataset chosen by the companies training the LLMs. Locally training a LoRA has the potential to reduce algorithmic discrimination since LLMs are trained to be

broadly accepted rather than cater to specific communities with varying perspectives. Communities will be able to train LoRAs to add community specific factual data and influence the original LLM output in order to reduce discrimination. A LoRA shifts the focus from how major companies intend communities to interact with LLM tools to how communities prefer and need to use LLM tools by giving them a low cost customization method.

Locally training a LoRA also represents an option for communities to preserve data sovereignty. While commercially available LLMs require internet access and a large data center to run, a locally run small parameter LLM with a custom LoRA does not require any internet connection. All user interactions and data that a community provides to train the LoRA remains on hardware they physically control. This means a community is protected against sensitive data being leaked from data breaches or reverse engineering prompt attacks. Training their own LoRA gives communities the ability to completely control their data without relying on a third party.

The following explores if a LoRA can be locally trained on low cost commercially available hardware, reduce discrimination and bias present in a base LLM, and cater to a specific community's needs and perspectives. Will a locally trained LoRA allow communities to preserve their privacy, to retain sensitive information, and most importantly to interact with LLMs without having to resign to biases that are incumbent in commercial LLMs produced by large corporations?

## 2. Methods

Behind the scenes, LoRA training involves injecting new trainable layers after the feed-forward section of a transformer model. Then using SVD as a compression method, the new layers are compressed and saved. The smaller LoRA can be decompressed and injected back into the transformer model at any point in the future. Even if an order of magnitude less in size, see Table 1, the trained LoRA has a potential to influence output of a multi billion parameter LLM.<sup>48,52,53</sup>

Memory comparison of LLM and LoRA

	Size on Disk	Size in VRAM
7B Parameter LLM	3630 MB	4.4 GB
LoRA for 7B LLM	262 MB	0.4 GB
13B Parameter LLM	6760 MB	7.4 GB
LoRA for 13B LLM	409 MB	0.8 GB

Table 1: Size comparisons in disk storage and for working memory between LLM models and LoRAs.

The data for training the LoRA was gathered from Disabled American Veterans (DAV). DAV was chosen due to the investigator’s familiarity with the material and tone. To prepare the dataset, duplicated data was removed, special and non standard characters were replaced, phone numbers were dropped, website URLs were dropped, and any special formatting was converted to one standard style.<sup>46,47</sup>

Two LLMs were chosen for the training and evaluation. Two LLaMa 2 derivative models, a 7 billion parameter Vicuna model and a 13 billion parameter Vicuna model, were downloaded from Hugging Face via the oobabooga Text generation web UI. This webui was chosen because it represents the most accessible interface for an organization to get started with LoRA training. While there are several ways to train a LoRA in the webui, the most accessible way is to load the Vicuna models using Transformers and disabling any ExLlama kernel optimizations. This method does not require the user to install additional dependencies or enable special command line flags.

LoRA Training and Evaluation Flowchart

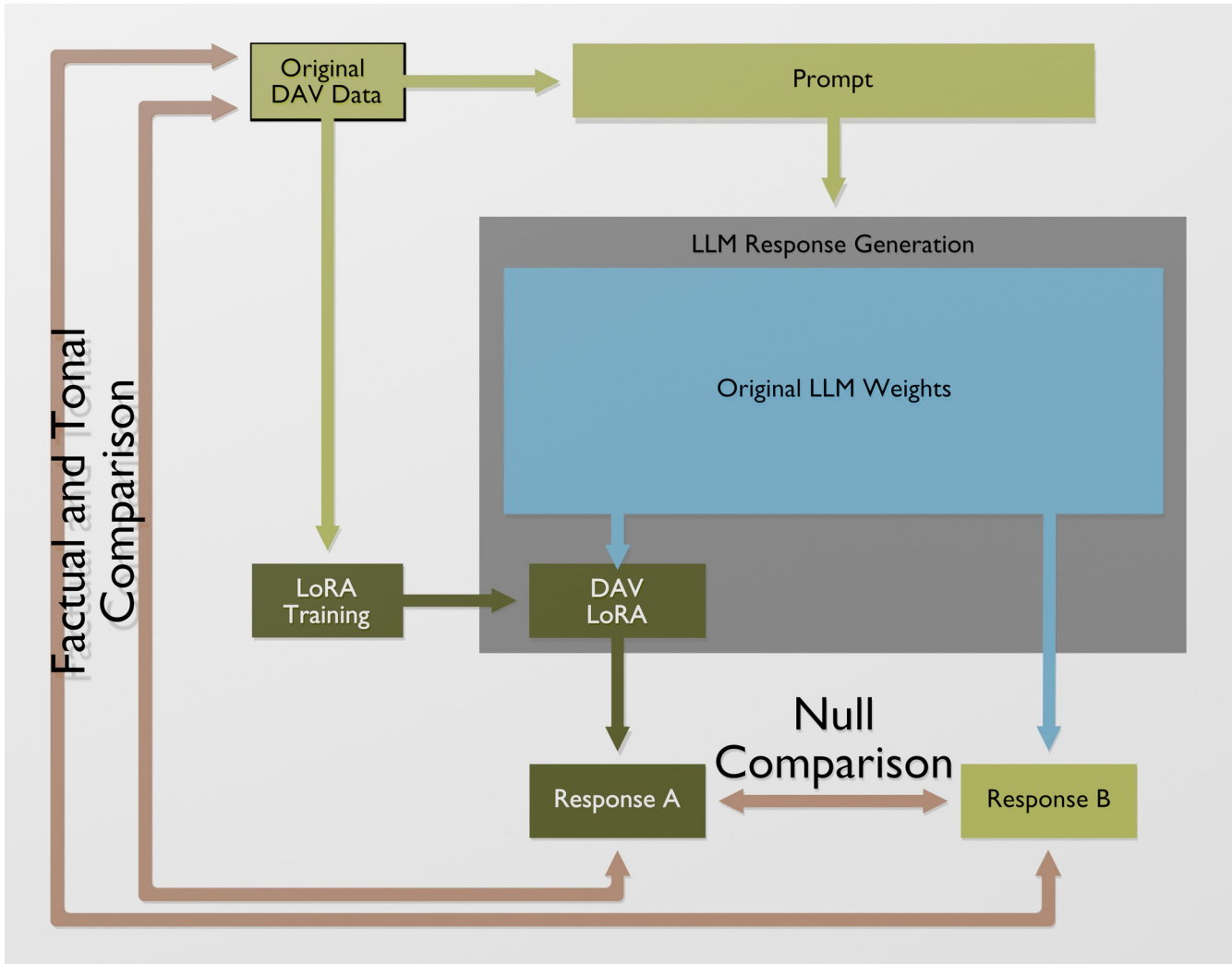


Figure 1: Overview showing the flow of data during training and points of comparison for evaluation.

To start training, the cleaned raw text file was loaded into the webui. Raw text was chosen since other methods require special formatting rules which decrease accessibility. The learning rate schedule was changed from default to Constant with a learning rate of  $3e-4$ . LoRA dropout was changed to 0.5 and Stop at loss point was disabled. The training was set to run for 1000 Epochs and checkpoints were saved every 5000 steps for points of comparison. This training strategy produces checkpoints every 5000 steps during training which can be used to check for underfitting and overfitting. Changing the LR scheduler to constant means there is a linear relationship between the number of training steps and the training loss. Functionally this allows the user to evaluate each incremental checkpoint and choose the one that provides the best results.

When training and checkpoint evaluation was complete, a LoRA at epoch 429.71 was chosen for the 7B Vicuna LLM and a LoRA at epoch 457.14 was chosen for the 13B Vicuna LLM. Both training runs took approximately 11 hours each to complete. The final testing LoRAs were chosen with slight overfitting since initial evaluation showed better adherence to data validity and tonal consistency. Using total system load reported from sensor data in HWiNFO and total training time, training both models consumed a total of 7kWh.

Testing methodology involved asking a battery of prompts and performing a binary comparison to address if the responses were factually and tonally consistent compared to the original training data. For a point of comparison, the same prompts were asked against the base Vicuna models as a control. Did the LoRA produce factually valid responses and did the response remain tonally consistent with the training data? Additionally, did the responses generated by the LoRA differ from the control responses?

Each LoRA and LLM was asked a battery of 50 questions taken directly from the training data to represent every major section from the dataset. All responses were recorded from a fresh conversation with no previous token history, which is important since continuing within the same conversation would include previous conversation history and would influence the generated responses. In addition, the same Seed was set for response generation since random Seed generation would produce different results given the same prompt.

Using a modified key informant interview technique the LLM models and LoRAs were asked about the training data.<sup>57</sup> Factually valid responses were coded as pass or fail, a passable response was recorded if the LLM's response contained the correct answer per the training material and additionally did not include factually incorrect information. Tonally correct responses had to adhere to the tone set by the DAV training data and failed if the response showed any tonal bias, such as toxicity, that was not consistent with the training data.

### 3. Results

Running the results through a Chi-squared test, (See Table 2 and Table 3) the results showed a statistically significant difference, with an alpha of 0.05, between correct and incorrect responses recorded off the base 7B parameter Vicuna compared to the responses recorded off the 7B parameter Vicuna + LoRA ( $p < 0.05$ ). However when moving to the 13B parameter Vicuna model and LoRA, the responses failed to show statistical significance ( $p > 0.05$ ). When comparing the tonally consistent responses to those that showed bias, the 7B LLM and LoRA were the only responses that showed statistical significance ( $p < 0.05$ ). The 13B model and LoRA again failed to show statistical significance in tonal responses ( $p > 0.05$ ).

Response Results

	Correct	Incorrect	Tonally Consistent	Bias
Vicuna_7B	20	26	17	29
Vicuna_7B_LoRA	36	10	37	9
Vicuna_13B	35	11	25	21
Vicuna_13B_LoRA	40	6	33	13

Table 2: Coded and recorded responses from Base LLM models and LoRAs.

Reported p values

	Validity	Tone
7B_Vicuna versus 7B_Vicuna_LoRA	$P < 0.01$	$P < 0.01$
13B_Vicuna versus 13B_Vicuna_LoRA	$P = 0.28$	$P = 0.13$

Table 3: Reported p-values from Chi-squared test comparing responses from the base LLM models and their trained LoRAs.

The 7 billion parameter Vicuna model and its LoRA showed significant improvement to the generated responses when compared to the base model. In comparison to the base model, the LoRA produced more valid and tonally correct responses. The majority of invalid responses produced by both LoRAs involved asking about disability calculations and rates. While the Vicuna LLM is well suited for conversational style responses, it is not well suited for calculations or math. This highlights the importance of choosing a suitable LLM and understanding LLM limitations.

While the 13 billion parameter Vicuna base model compared to its LoRA did not produce statistically significant responses compared to the base 13B LLM, failing to reject the null, the 13B

Vicuna LoRA did produce the most detailed responses of any model. It is important to remember the p-value is only one measure for judging evidence and the totality of all data should be considered.<sup>58</sup> One example was the 13B LoRA responding correctly and then continuing to fully list Veteran’s Day deals at participating restaurants. Where the other three responses only stated broadly that there were Veteran’s Day deals. The 13B LoRA also was more likely to produce nuanced responses that took advantage of the extra 6B parameters. The responses from the 13B\_Vicuna\_LoRA produced more detail than those provided from base models as well as the 7B\_Vicuna\_LoRA. The responses it provided were also nearly identical with the training data and tonally correct (see Table 4 and Table 5 for an example). These examples showed that while the 13B\_Vicuna\_LoRA responses did not statistically significantly differ from the base LLM, the responses did factually and tonally differ.

A possible explanation for this phenomenon would be the consistent size of the training data relative to the increased size of the base LLM. The 13B Vicuna model with its significantly larger network, may have had a harder time finding this optimal set of weights, especially with the limited guidance signal available in the training data. There may have not been enough guidance material to properly support the extra 6B parameters in the 13B LLM. While it is possible that selecting a checkpoint from a higher epoch could improve the results, the selected LoRA already showed signs of overfitting. Perhaps a larger training dataset or increasing the target modules during training could produce more factually valid and tonally consistent responses for the 13B LoRA.

Both the trained LoRAs, and the base LLMs, struggled when prompted to do calculations. When asked to calculate disability ratings, all responses failed to produce correct arithmetic or follow basic algebra. However the responses did remain tonally consistent with how the calculations were supposed to be presented. Other LLMs may be more suited for calculations since Vicuna was trained for conversation and instruction following. This highlights the importance of choosing a LLM suited for the task.

The statistical significance from the 7B test shows that it’s possible that a locally trained and run LoRA can adequately represent the perspective and needs of a specific community. The responses from the 7B model, and to some degree the 13B model, remained factually relevant and didn’t show significant signs of bias or toxicity. Most importantly the training occurred on low cost consumer hardware and didn’t need external cloud based services. The training data never left the computer system which ensured complete data sovereignty.

Table 4: Prompt and response generated by base 13B Vicuna on DIC.

What is Dependency and Indemnity Compensation?
Survivors of deceased veterans may be eligible for certain educational benefits through the Department of Veterans Affairs (VA). These include the Survivors' and Dependents' Educational Assistance Program (DEA) and the Fry Scholarship program. Under DEA, dependents of veterans who died as a result of their service-connected disabilities or in the line of duty may receive up to 45 months of educational assistance.
-AI response generated by the 13B_Vicuna on training data from <a href="https://www.dav.org/">https://www.dav.org/</a>

Table 5: Prompt and response generated by 13B Vicuna with the trained LoRA on DIC

What is Dependency and Indemnity Compensation?
<p>Dependency and Indemnity Compensation (DIC) is a monthly benefit paid to the surviving spouse, children or parents of any veteran who passed away on or after Jan. 1, 1957. It may also be payable to certain survivors of veterans whose service-connected disabilities did not cause their deaths. The amount of DIC is equal to the amount paid to a surviving spouse with dependent children, which means that if the spouse has both dependents and non-dependents, the spouse will only receive the amount payable for one group, either dependents or non-dependents. Surviving parents can only receive DIC if they are the natural parents of the veteran or if they were formally adopted by the veteran.</p> <p>To be eligible for DIC, both of the following must be true for surviving spouses and children:          The veteran must have died while on active duty or on active or inactive duty for training.          The deceased veteran must have been permanently disabled or a purple heart recipient.</p> <p>For surviving parents, all of the following must be true:          The veteran must have been their child's biological, adoptive or foster parent.          The veteran must have died while on active duty or on active or inactive duty for training.          The deceased veteran must have been permanently disabled or a purple heart recipient.          Note: "Inactive training" refers to training periods when service members are on leave or exercises. They are still considered on active duty for training purposes.</p> <p>Eligibility for surviving children          To receive DIC, surviving children must meet all of the following requirements:          Not included in the surviving spouse's DIC.          Under age 18, or between ages 18 and 23 while attending school.          Children adopted out of the veteran's family may be eligible for DIC if all other eligibility criteria are met.</p> <p>Eligibility for surviving parents          To receive DIC, both of the following must be true for surviving parents:          Biological, adoptive or foster parent of the veteran or service member.          Income below a certain amount.</p>
-AI response generated by the 13B_Vicuna_LoRA on training data from <a href="https://www.dav.org/">https://www.dav.org/</a>

## 4. Discussion

### 4.1 Limitations and Considerations

While the 7B LoRA showed statistical significance and shows that low cost privacy focused AI tools are available, this analysis still only represents one community. To improve external validity, more locally trained LoRAs need to be tested on a greater variety of communities with differing viewpoints. Additionally while moving forward, community involvement could help increase access to training data and participation in the evaluation process. It will be important that communities understand that any data they provide will remain local in their control and never transmitted over the internet.

Additionally, while no data is transmitted over the internet, the physical security of the data at rest must be protected. This includes all the training data and user interactions with the LoRA. The computer and its hard drive must be secured against theft or surveillance just as if it were a paper folder containing the same information. Physical security practices and cybersecurity practices still need to be observed to ensure data privacy. A good guide to ensuring data is protected and secure is to review the HIPAA rules for hardware.

### 4.2 Conclusion

Training a LoRA on a community-specific dataset allows organizations to customize a LLM. This mitigates algorithmic discrimination and improves privacy, while still providing access to the benefits of LLM tools. Since LLMs are trained on datasets composed of large amounts of data from various sources that may contain biases and incomplete information, training a LoRA on a community-specific dataset can help organizations to customize a LLM to better address the needs and preferences of their target population. Organizations will have complete control over the training data and can ensure the dataset is relevant to their mission. Also since the data is never transmitted over the internet, locally training a LoRA allows organizations to have complete data sovereignty. Protecting communities against aggressive data collection and digital profiling can help improve community resilience.

Additionally, training a LoRA on community-specific data allows organizations with fewer resources to utilize LLMs. Training both LoRAs on personal hardware consumed 7kWh of energy which cost around 1 USD at local utility rates. While utility rates can vary, the cost of training a LoRA is still significantly lower than the \$191 million that Google spent training their Gemini LLM.<sup>26</sup> This creates an accessible entry point towards decentralized LLM and promotes end user accessibility, especially when utilized in combination with graphical user interfaces like oobabooga Text generation web UI.

Moreover, it's important to raise awareness for privacy since online LLM tools represent an unprecedented way to collect user data. The human like responses generated from LLMs are more likely to prompt users to overshare personal information than a traditional internet search.<sup>39</sup> In addition, user and community data can be obtained by performing reverse warrants. A reverse warrant occurs when law enforcement legally compel companies to release user information based on specific criteria, such as key words used in an internet search.<sup>33,59</sup> Normally warrants are limited to a specific detailed scope like a person or residence during a specific time, which is supported by probable cause. However in the case of reverse warrants the number of individuals who match the search criteria can remain undefined and can extend longitudinally.<sup>60</sup> This is particularly concerning for individuals or communities who wish to establish digital curtilage to protect against unreasonable search and seizures as guaranteed by the Fourth Amendment since cloud based LLMs are being integrated into virtual assistants like Apple's Siri or Amazon's Alexa or directly into operating systems like Microsoft's Copilot. Establishing digital curtilage and maintaining privacy can prove difficult when companies install digital LLM agents that can log our interactions and then comply with government requests for user data.

It's important to show communities that private, affordable, and easily accessible options for popular services are available. Using privacy focused applications allow communities to minimize the attack surface that companies use to collect user data, data that can be used to algorithmically discriminate against entire populations. This analysis covers how locally run LLMs and LoRAs are suitable alternatives to commercially AI tools, all while being low cost, easy to set up, and private.

## Appendix

## Setup Guide

The following guide will provide setup instructions on how to properly install and run local LLM. The oobabooga text-generation-webui was chosen based on it having an accessible and easy to use graphical user interface (GUI). First we will be choosing hardware, next installing the webui, downloading a LLM, and finally begin talking to the LLM.

### Recommended Hardware

Before starting, first decide on the hardware. Most modern CPUs will be able to run a LLM using this interface. However while the LLM will absolutely run on a computer with a CPU only, it is highly recommend to use a GPU for text generation and training. GPUs are purpose built with architecture to support parallel processing rather than sequential processing.<sup>61</sup> A GPU with 6GB of VRAM can be found on eBay for under \$90, as of April 2024, and will greatly speed up text generation and training times. The following table is included as aid in choosing a GPU with an appropriate amount of VRAM.

Memory comparison of LLM and LoRA

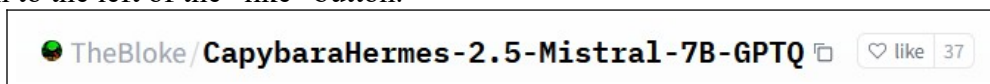
	Size on Disk	Size in VRAM
7B Parameter LLM	3630 MB	4.4 GB
LoRA for 7B LLM	262 MB	0.4 GB
13B Parameter LLM	6760 MB	7.4 GB
LoRA for 13B LLM	409 MB	0.8 GB

### Installing the webui

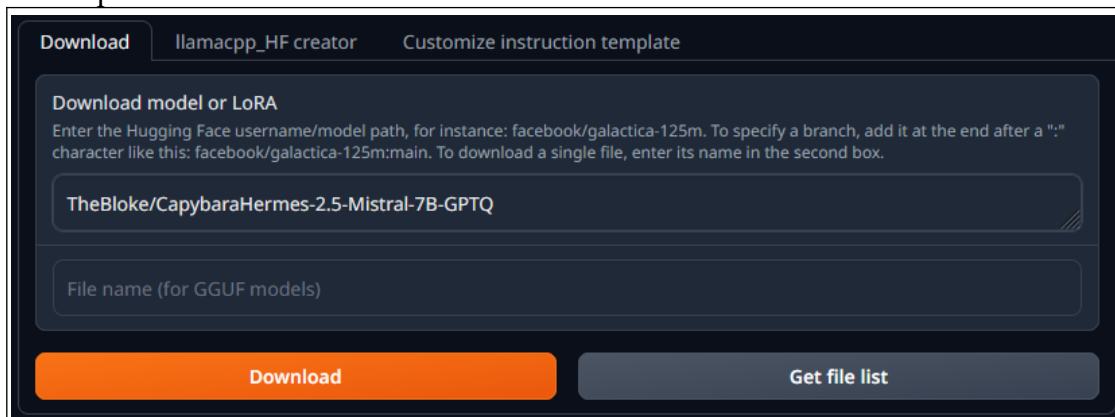
1. First to install the webui, clone the repository using Git or directly download the Source code (.zip) from the GitHub repository. This link provides the most up-to-date release
  - 1.1. <https://github.com/oobabooga/text-generation-webui/archive/refs/heads/main.zip>.
2. Next, unzip the folder to somewhere where there are no spaces in the folder path.
  - 2.1. A good example for a folder path would be “C:/text-generation-webui.”
  - 2.2. A bad example for a folder path containing spaces would be “C:/Program Files/text generation webui.”
3. Once the folder has been unzipped click on the appropriate starting file. Most users will need to click on the Windows script.
  - 3.1. For Linux click on start\_linux.sh
  - 3.2. For Windows click on start\_windows.bat
  - 3.3. For Mac click on start\_macos.sh
  - 3.4. For Windows subsystem for linux click on start\_wsl.bat.
4. A command terminal will open and automatically install the webui. When prompted, answer what GPU you plan to run the interface on. The CPU only option is also available here.
5. Finally, once the installation completes, a browser will open with the webui.

## Choosing a LLM

1. First identify what you want to use a LLM for. Some models are better suited for chat and instruction following while other models are more suited for coding.
  - 1.1. Another consideration would be language. Some LLMs are trained in different languages.
  - 1.2. Have an idea about desired model size. Table 1 is a great way to understand the relationship between parameter size and VRAM requirements. 6 GB of VRAM is roughly enough to run a 7B parameter LLM with its LoRA.
2. Next search for a suitable LLM on Hugging Face. This is a site that hosts the LLM models we can download and use with the webui.
  - 2.1. To help search for available LLMs a community curated leader board can be found here [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)
  - 2.2. There are also individuals that come highly recommend for their work in providing custom LLM weights. <https://huggingface.co/TheBloke>
3. Once a suitable LLM has been found, it can be downloaded in the webui by clicking on the square icon to the left of the “like” button.



4. Next in the webui, navigate to the “Model” tab at the top of the page. Look for the section with a bright orange Download button and paste the model card information that you copied from the last step here. It should look like this.



5. Finally clicking on the Download button will download the model onto the computer. Once downloaded no further internet connection is required. Everything required to run a local LLM has already been loaded onto the computer.

## Usage

1. To start using the model that was just downloaded, click on the model drop menu at the upper left corner on the same page. If it doesn't show up, the list can be refreshed using the blue icon on the right next to the drop menu. The proper settings to load the model will automatically be detected and loaded by the webui.

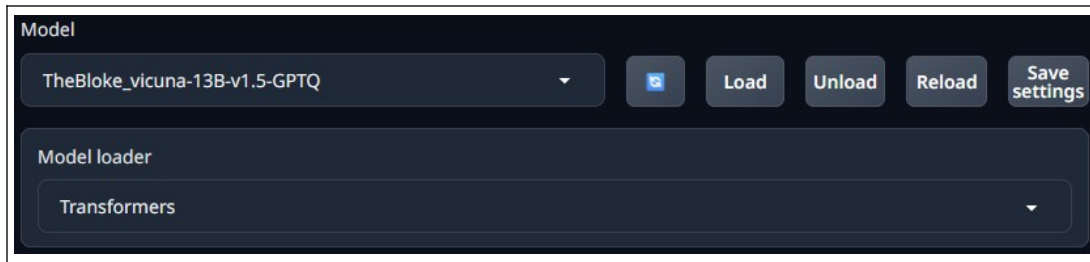
2. Next, clicking the Load button will proceed to load the model using its default recommended settings.
3. Finally, click back to the “Chat” tab to start a chat with the loaded LLM.
  - 3.1. Type a message and click “Generate” to start talking.



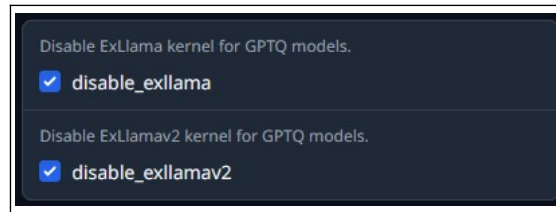
## Training Guide

The following guide will provide instructions for training a LoRA on the oobabooga Text Generation Web UI. The settings were chosen to provide maximum accessibility as well as provide some flexibility in checkpoint selection.

1. First follow the instructions for loading a LLM contained in the Setup Guide and stop before actually loading a LLM. Before clicking Load, several settings have to be changed for the webui to properly start training. Change the Model Loader to “Transformers”.

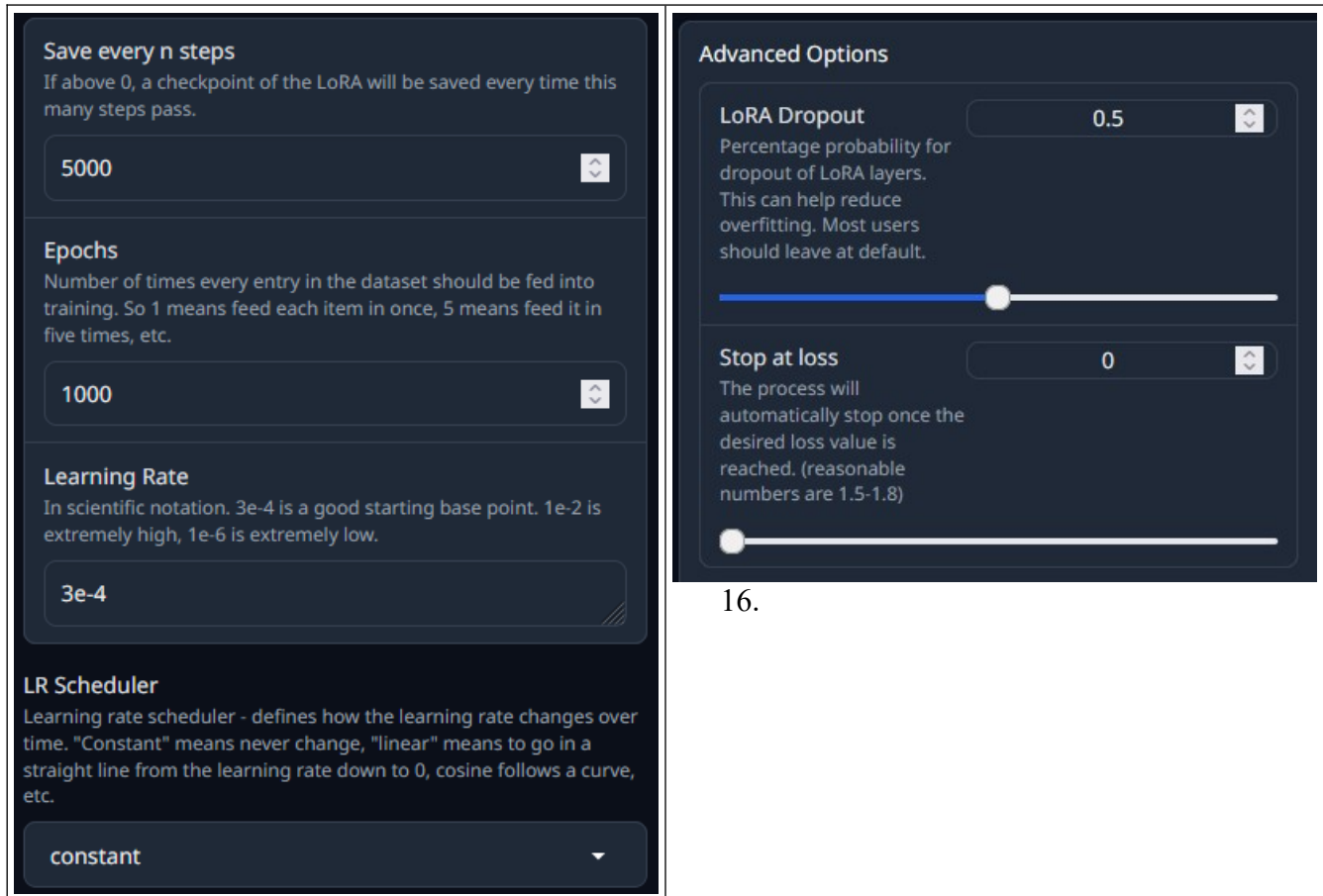


3. Next, disable exllama optimizations.



5. Now the LLM can be loaded for training. Click Load like before and head to the “Training” tab found at the very top.
6. Open a file browser and navigate to where you unzipped the webui installation.
  - 6.1. If following the Setup Guide that would be “C:/text-generation-webui”
7. Drop the cleaned training data into the “datasets” folder found in the “training” folder.
  - 7.1. The full path to this folder would be “C:/text-generation-webui/training/datasets”
    - a) Use as much available text as possible for better results.
    - b) The training data should be all one format and using standard characters in one .txt file.
    - c) The training data should not have duplicated or repeating entries, phone numbers, website URLs, or calculations.
    - d) Avoid using negatives. For the term “non-professional” instead consider using “casual.”
    - e) Avoid using multiple tones. Keep the training data tonally consistent.
8. After dropping the .txt file containing the cleaned training data, change from “Formatted Dataset” to “Raw text file” on the right hand of the screen. Clicking on the blue refresh button and using the drop down menu will reveal the .txt file.
9. Next create a name for the LoRA into the text field under “Name”
10. Enter the number 5000 under “Save every n steps”

11. Change “Epochs” to 1000 or lower
12. Select “Constant” under “LR Scheduler”
13. Click on Advanced Options at the bottom and change LoRA Drop out to 0.5
14. Ensure that “Stop at loss” is set to 0



17. The options should look like this when completed. Everything else can be left to default.
18. Click “Start LoRA Training” to start training and training progress will be displayed.
19. Once completed, the trained LoRAs will be in “C:/text-generation-webui/loras”. It’s recommended to save a copy of the LoRA before proceeding with checkpoint evaluation.
20. To test the LoRA, return to the “Model” tab and reload the base LLM. Next on the right side of the page, refresh the LoRA drop down list with the blue button to see available LoRAs.
  - 20.1. Selecting a LoRA and clicking “Apply LoRA” will load the very last checkpoint, which will likely show overfitting.
  - 20.2. However, selecting different checkpoints contained in “C:/text-generation-webui/loras” could possibly help. Open the folder containing the LoRA. Inside there are checkpoint

folders that contain the state of the LoRA at different stages of training. Copy the five files from different checkpoints into the main folder to check how it affects text generation.

- 1 Xu, Yongjun et al. "Artificial intelligence: A powerful paradigm for scientific research." *Innovation (Cambridge (Mass.))* vol. 2,4 100179. 28 Oct. 2021, doi:10.1016/j.xinn.2021.100179
- 2 Guan, Yanchu, et al. *Intelligent Virtual Assistants with LLM-Based Process Automation*. 2023. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2312.06677>
- 3 Chaubey, Pavan Kumar et al. "Sentiment Analysis of Image with Text Caption using Deep Learning Techniques." *Computational intelligence and neuroscience* vol. 2022 3612433. 25 Jun. 2022, doi:10.1155/2022/3612433
- 4 Jim, Jamin Rahman, et al. "Recent Advancements and Challenges of NLP-Based Sentiment Analysis: A State-of-the-Art Review." *Natural Language Processing Journal*, vol. 6, Mar. 2024, p. 100059. DOI.org (Crossref), <https://doi.org/10.1016/j.nlp.2024.100059>
- 5 Ramadan, Rabie, et al. "Can AI Generate Efficient and Accepted Speech?" *WAS Science Nature* vol. 6, 1 March 2024.
- 6 Ray, Partha Pratim. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope." *Internet of Things and Cyber-Physical Systems*, vol. 3, 2023, pp. 121–54. DOI.org (Crossref), <https://doi.org/10.1016/j.iotcps.2023.04.003>
- 7 Van Der Vlist, Fernando, et al. "Big AI: Cloud Infrastructure Dependence and the Industrialisation of Artificial Intelligence." *Big Data & Society*, vol. 11, no. 1, Mar. 2024, p. 20539517241232630. DOI.org (Crossref), <https://doi.org/10.1177/20539517241232630>
- 8 Rikap, Cecilia. *Same End by Different Means: Google, Amazon, Microsoft and Meta's Strategies to Organize Their Frontier AI Innovation Systems*. Working Paper, 2023–03, CITYPERC Working Paper, 2023. [www.econstor.eu](http://www.econstor.eu), <https://www.econstor.eu/handle/10419/280829>
- 9 Cao, Longbing. *AI in Finance: Challenges, Techniques and Opportunities*. 2021. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2107.09051>
- 10 Peng, Cheng, et al. "A Study of Generative Large Language Model for Medical Research and Healthcare." *Npj Digital Medicine*, vol. 6, no. 1, Nov. 2023, pp. 1–10. [www.nature.com](http://www.nature.com), <https://doi.org/10.1038/s41746-023-00958-w>
- 11 Yang, Xi, et al. *GatorTron: A Large Language Model for Clinical Natural Language Processing*. preprint, Health Informatics, 28 Feb. 2022. DOI.org (Crossref), <https://doi.org/10.1101/2022.02.27.22271257>
- 12 Department of Defense. 2023. Establishment of Chief Digital and Artificial Intelligence Officer Generative Artificial Intelligence and Large Language Models Task Force, Task Force Lima. Retrieved from <https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/>
- 13 Chen, Daoyuan, et al. *Data-Juicer: A One-Stop Data Processing System for Large Language Models*. 2023. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2309.02033>
- 14 Dosovitsky, Gilly, et al. "Artificial Intelligence Chatbot for Depression: Descriptive Study of Usage." *JMIR Formative Research*, vol. 4, no. 11, Nov. 2020, p. e17065. DOI.org (Crossref), <https://doi.org/10.2196/17065>
- 15 Ho, Annabell et al. "Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot." *The Journal of communication* vol. 68,4 (2018): 712-733. doi:10.1093/joc/jqy026
- 16 *Plainlanguage.gov | What Is Plain Language?* <https://www.plainlanguage.gov/about/definitions/>. Accessed 25 Apr. 2024
- 17 Engelmann, Jan B, and Marianna Pogosyan. "Emotion perception across cultures: the role of cognitive mechanisms." *Frontiers in psychology* vol. 4 118. 12 Mar. 2013, doi:10.3389/fpsyg.2013.00118

- <sup>18</sup> Kastanakis, Minas N., and Benjamin G. Voyer. “The Effect of Culture on Perception and Cognition: A Conceptual Framework.” *Journal of Business Research*, vol. 67, no. 4, Apr. 2014, pp. 425–33. DOI.org (Crossref), <https://doi.org/10.1016/j.jbusres.2013.03.028>
- <sup>19</sup> Strubell, Emma, et al. *Energy and Policy Considerations for Deep Learning in NLP*. 2019. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.1906.02243>
- <sup>20</sup> Wu, Carole-Jean, et al. *Sustainable AI: Environmental Implications, Challenges and Opportunities*. 2021. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2111.00364>.
- <sup>21</sup> Lacoste, Alexandre, et al. *Quantifying the Carbon Emissions of Machine Learning*. 2019. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.1910.09700>.
- <sup>22</sup> Ghosh, Debi Prasad. *Kumbhakarna: Six Months Pretrain and One Day Deployment -A Cost Analysis of Large Language Model Training on Cloud*. 2024. DOI.org (Datacite), <https://doi.org/10.13140/RG.2.2.26118.59202>
- <sup>23</sup> Knight, Will. “OpenAI’s CEO Says the Age of Giant AI Models Is Already Over.” *Wired*. [www.wired.com](http://www.wired.com), <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/>
- <sup>24</sup> Biderman, Stella, et al. *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. 2023. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2304.01373>
- <sup>25</sup> Sharir, Or, et al. *The Cost of Training NLP Models: A Concise Overview*. 2020. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2004.08900>
- <sup>26</sup> *AI Index Report 2024 – Artificial Intelligence Index*. <https://aiindex.stanford.edu/report/>. Accessed 22 May 2024
- <sup>27</sup> Faiz, Ahmad, et al. *LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models*. arXiv:2309.14393, arXiv, 19 Jan. 2024. [arXiv.org](http://arxiv.org/abs/2309.14393), <http://arxiv.org/abs/2309.14393>
- <sup>28</sup> Luccioni, Alexandra Sasha, et al. “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model.” *Journal of Machine Learning Research*, vol. 24, no. 253, 2023, pp. 1–15. [www.jmlr.org](http://www.jmlr.org), <http://jmlr.org/papers/v24/23-0069.html>
- <sup>29</sup> “Meta Intros Two GPU Training Clusters for Llama 3 | TechTarget.” *Enterprise AI*, <https://www.techtarget.com/searchenterpriseai/news/366573246/Meta-intros-two-GPU-training-clusters-for-Llama-3>. Accessed 25 Apr. 2024
- <sup>30</sup> Leswing, Kif. “Nvidia’s Top A.I. Chips Are Selling for More than \$40,000 on eBay.” *CNBC*, 14 Apr. 2023, <https://www.cnn.com/2023/04/14/nvidias-h100-ai-chips-selling-for-more-than-40000-on-ebay.html>
- <sup>31</sup> “NVIDIA H100 Tensor Core GPU.” *NVIDIA*, <https://www.nvidia.com/en-us/data-center/h100/>. Accessed 25 Apr. 2024
- <sup>32</sup> “Building Meta’s GenAI Infrastructure.” *Engineering at Meta*, 12 Mar. 2024, <https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/>
- <sup>33</sup> Tokson, Matthew. “Knowledge and Fourth Amendment Privacy.” *Northwestern University Law Review*, vol. 111, no. 1, Dec. 2016, pp. 139–204, <https://scholarlycommons.law.northwestern.edu/nulr/vol111/iss1/3>
- <sup>34</sup> Meta’s Transparency Reports: Government Requests for User Data. Accessed 5 Dec. 2023. <https://transparency.fb.com/reports/government-data-requests>
- <sup>35</sup> Onay, Ceylan, and Elif Öztürk. “A Review of Credit Scoring Research in the Age of Big Data.” *Journal of Financial Regulation and Compliance*, vol. 26, no. 3, July 2018, pp. 382–405. DOI.org (Crossref), <https://doi.org/10.1108/JFRC-06-2017-0054>
- <sup>36</sup> Aggarwal, Nikita. “The Norms of Algorithmic Credit Scoring.” *SSRN Electronic Journal*, 2020. DOI.org (Crossref), <https://doi.org/10.2139/ssrn.3569083>

- <sup>37</sup> Duhigg, Charles. “How Companies Learn Your Secrets.” *The New York Times*, 16 Feb. 2012. *NYTimes.com*, <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>
- <sup>38</sup> *AI Credit Scoring: The Future of Credit Risk Assessment*. <https://www.datrics.ai/articles/the-essentials-of-ai-based-credit-scoring>. Accessed 1 Apr. 2024
- <sup>39</sup> Ischen, Carolin, et al. “Privacy Concerns in Chatbot Interactions.” *Chatbot Research and Design*, edited by Asbjørn Følstad et al., vol. 11970, Springer International Publishing, 2020, pp. 34–48. *DOI.org (Crossref)*, [https://doi.org/10.1007/978-3-030-39540-7\\_3](https://doi.org/10.1007/978-3-030-39540-7_3)
- <sup>40</sup> Brown, Hannah, et al. “What Does It Mean for a Language Model to Preserve Privacy?” *2022 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2022, pp. 2280–92. *DOI.org (Crossref)*, <https://doi.org/10.1145/3531146.3534642>
- <sup>41</sup> Li, Haoran, et al. *Privacy in Large Language Models: Attacks, Defenses and Future Directions*. 2023. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2310.10383>
- <sup>42</sup> *Wyden Opposes Warrantless Government Surveillance of Americans’ Internet Browsing History | U.S. Senator Ron Wyden of Oregon*. <https://www.wyden.senate.gov/news/press-releases/wyden-opposes-warrantless-government-surveillance-of-americans-internet-browsing-history->. Accessed 5 Feb. 2024
- <sup>43</sup> Wang, Boxin, et al. *DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models*. 2023. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2306.11698>
- <sup>44</sup> O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition, Crown, 2016
- <sup>45</sup> “Blueprint for an AI Bill of Rights | OSTP.” *The White House*, <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>. Accessed 1 Apr. 2024
- <sup>46</sup> Thompson, Andrew. “Google’s Sentiment Analyzer Thinks Being Gay Is Bad.” *Vice*, 25 Oct. 2017, <https://www.vice.com/en/article/j5jnmj8/google-artificial-intelligence-bias>
- <sup>47</sup> Grant, Nico. “Google Chatbot’s A.I. Images Put People of Color in Nazi-Era Uniforms.” *The New York Times*, 22 Feb. 2024. *NYTimes.com*, <https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html>
- <sup>48</sup> Robertson, Adi. “Google Apologizes for ‘Missing the Mark’ after Gemini Generated Racially Diverse Nazis.” *The Verge*, 21 Feb. 2024, <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>
- <sup>49</sup> Hu, Edward J., et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2106.09685>
- <sup>50</sup> Wang, Huiwen, et al. *Enhancing the SVD Compression*. 2021. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2112.03715>
- <sup>51</sup>P, Sudhakar, et al. “Query-by-Example Spoken Term Detection for Zero-Resource Languages Using Heuristic Search.” *ACM Transactions on Asian and Low-Resource Language Information Processing*, July 2023, p. 3609505. *DOI.org (Crossref)*, <https://doi.org/10.1145/3609505>
- <sup>52</sup>Wang, Pidong, et al. “Source Language Adaptation Approaches for Resource-Poor Machine Translation.” *Computational Linguistics*, vol. 42, no. 2, June 2016, pp. 277–306. *DOI.org (Crossref)*, [https://doi.org/10.1162/COLI\\_a\\_00248](https://doi.org/10.1162/COLI_a_00248)
- <sup>53</sup>Amin, Nadine Amr Mahmoud. *Low-Resource Automatic Speech Recognition Domain Adaptation: A Case-Study in Aviation Maintenance*. 2023, p. 2199294 Bytes. *DOI.org (Datacite)*, <https://doi.org/10.25394/PGS.23750598.V1>
- <sup>54</sup> Dodge, Jesse, et al. *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*. 2021. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2104.08758>
- <sup>55</sup> Lee, Katherine, et al. “Deduplicating Training Data Makes Language Models Better.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022, pp. 8424–45.

- DOI.org (Crossref), <https://doi.org/10.18653/v1/2022.acl-long.577>
- <sup>56</sup> Pfeiffer, Jonas, et al. *AdapterFusion: Non-Destructive Task Composition for Transfer Learning*. 2020. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2005.00247>
- <sup>57</sup> UCLA Center for Health Policy Research. (n.d.) Section 4: Key informant interviews. Retrieved from [https://healthpolicy.ucla.edu/programs/health-data/trainings/Documents/tw\\_cba23.pdf](https://healthpolicy.ucla.edu/programs/health-data/trainings/Documents/tw_cba23.pdf)
- <sup>58</sup> Gibson, Eric W. “The Role of  $p$ -Values in Judging the Strength of Evidence and Realistic Replication Expectations.” *Statistics in Biopharmaceutical Research*, vol. 13, no. 1, Jan. 2021, pp. 6–18. DOI.org (Crossref), <https://doi.org/10.1080/19466315.2020.1724560>
- <sup>59</sup> Rathi, Mohit. *Rethinking Reverse Location Search Warrants*, 111 *J. Crim. L. & Criminology* 805 (2021). <https://scholarlycommons.law.northwestern.edu/jclc/vol111/iss3/5>
- <sup>60</sup> Granick, Chad Marlow, Jennifer Stisa. “Celebrating An Important Victory In The Ongoing Fight Against Reverse Warrants | ACLU.” *American Civil Liberties Union*, 29 Jan. 2024, <https://www.aclu.org/news/privacy-technology/fight-against-reverse-warrants-victory>
- <sup>61</sup> “About CUDA.” *NVIDIA Developer*, <https://developer.nvidia.com/about-cuda>. Accessed 26 April 2024