



SEPARATING THE TRUTH FROM THE MYTH

*Controlling Language Models by
Controlling Training Data*

Stella Biderman

biderman_stella@bah.com

ABOUT ME

Large Language Model Researcher @ Booz Allen Hamilton

- Primarily supporting the Laboratory for Physical Sciences (LPS)
- Also US Cyber Command, STRATCOM, and others

Focus areas:

- Training large language models (10B+ parameters, 100s of GPUs)
- Language model evaluation
- Data-centric interpretability and learning dynamics research
- Multilingual NLP and code language models

Executive Director of EleutherAI, an open source and open science LLM research non-profit

LANGUAGE MODELS AND THEIR DATA

Data-centric approaches to modifying the behavior of language models

As we get better at identifying explanations for behavior, the next stage is to try to leverage that knowledge to *control* behavior.

Unfortunately, most of the best work is being done in relatively academic environments and is often unaware of realistic operational constraints.

There's a lot of hype around everything in LLMs right now, and its hard to tell what's real.

My goal: to give an overview of emerging and important topics from the academic literature and how seriously they should be taken in operational settings.

DATA INTERVENTIONS FOR BIAS: THE PITCH

The Hope: if models learn representational biases from their data, we can edit the data to influence the biases that the models learn.

“The doctor asked ~~his~~ patient ...”

“The doctor asked **her** patient ...”

Applications:

- Bias mitigation for fairness purposes
- Deliberately controlling biases for operational simulations
- Strategic bias manipulation in influence campaigns
- Avoiding learning to reproduce undesirable ideologies

DATA INTERVENTIONS FOR BIAS: THE REALITY

BLUF: You should probably finetune instead of editing the training data, but you can achieve these goals easily

The good:

- Data interventions are doable, albeit finicky
- Finetuning is very good at accomplishing the same goals and isn't nearly as hard

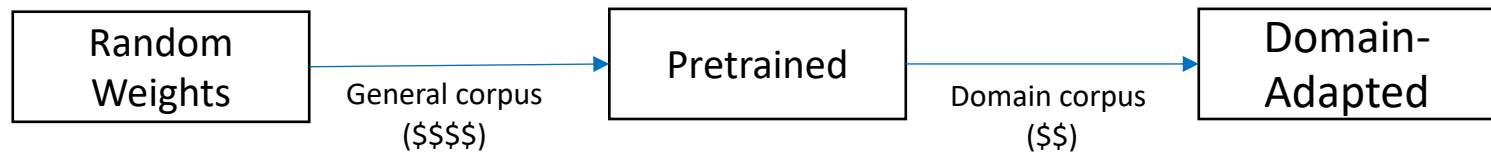
The bad:

- $P(\text{male} \mid \text{doctor}) \neq \%$ of doctors who are male in the training data
- In *Pythia* we show examples of how intervening on the last 10% of data doesn't necessarily move the needle

Recommendation: this is a promising area for research, but not ready as an off-the-shelf tool in your arsenal. Your needs are likely better met by finetuning

DOMAIN ADAPTATION: THE PITCH

The Hope: model trainers often have different goals from the government and may not train models on the right type of language for our use. We want to finetune models to adapt them to operate on language data that we care about.



Applications:

- New language acquisition
- Transferring capabilities across domains
- Teaching domain-specific jargon

Also known as “continual learning” “continual pretraining” or “cross-lingual finetuning”

DOMAIN ADAPTATION: THE REALITY

BLUF: This is cheap, easy, and effective

The good:

- Very little data and compute is necessary
- Most models are “mildly multilingual” which makes it easier to teach new languages
- Can be combined with instruction-finetuning effectively
- Works for style and formatting as well

The bad:

- [This line is intentionally left blank]

Recommendation: You should absolutely be doing this if you observe LLMs struggle to understand your domain-specific language

MEMORIZING FACTUAL INFORMATION: THE PITCH

The Hope: models can regurgitate long exact sequences from their training data. If we can encourage them to regurgitate *factual sequences* that could be a way to fight hallucination

Applications:

- Performance improvement
- Reliability and robustness
- Combatting hallucinations

MEMORIZING FACTUAL INFORMATION: THE REALITY

BLUF: This requires some bespoke work but does work with reasonable reliability.

The good:

- Models can learn the quality of the source data if given source data labels while they train and adjust how much they update accordingly
- Basic modifications like repeating data can substantially increase some types of memorization

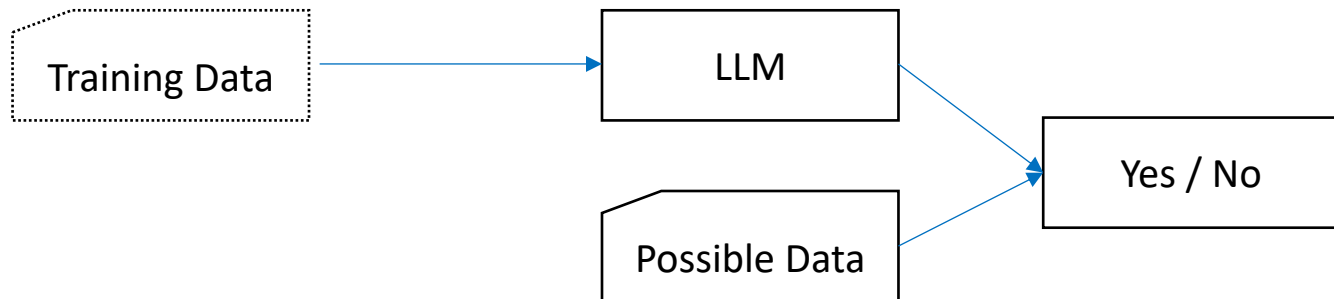
The bad:

- Only works *on average*, causing a model to memorize a specific fact is still hard.
- Works better for some types of memorization than for others

Recommendation: Not ready for non-expert users yet, but groups with LLM experts should look into this.

MEMBERSHIP INFERENCE ATTACKS: THE PITCH

The Hope: if you have blackbox access to a model and purported training data, you can check to see if the model was trained on the data



Applications:

- (Counter)intelligence operations related to LLMs
- Regulatory action against opaque model providers
- Data-sharing compliance
- Identifying data leakage
- Detecting data poisoning

MEMBERSHIP INFERENCE ATTACKS: THE REALITY

BLUF: very useful in many contexts but has some substantial caveats if *exact match* is the goal. Testing for “data that looks like this” is much easier than “this actual data”

The good:

- There is compelling evidence that you can infer training data
- Success beyond academic papers: many examples of it working in the real world

The bad:

- Recent research has shown that many papers with substantial effects are accidentally measuring *distribution shift*
- “True” membership inference is substantially harder and doesn’t really work (yet)

Recommendation: carefully evaluate how precise matching is required to achieve your mission goals and how good your guesses as to the model training data is likely to be

MEMBERSHIP INFERENCE ATTACKS: THE PITCH

The Hope: if you have blackbox access to a model and purported training data, you can check to see if the model was trained on the data

Applications:

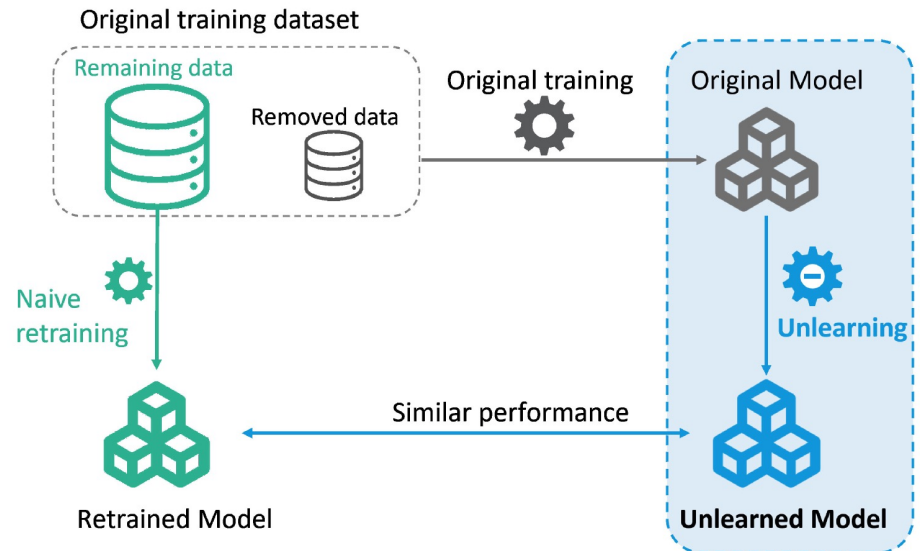
- (Counter)intelligence operations related to LLMs
- Regulatory action against opaque model providers
- Data-sharing compliance (unclear / context dependent)
- Identifying data leakage (context-dependent)
- Detecting data poisoning

MACHINE UNLEARNING

The Hope: if you wish you hadn't trained a model on certain data, you can fix that after the fact without incurring the costs of retraining from scratch.

Applications:

- Regulatory compliance
- Fixing mistakes
- Updating models as new information comes to light
- Counterfactual analysis



MACHINE UNLEARNING: THE REALITY

BLUF: it's not clear that there is a context where existing MU approaches can be relied on to accomplish our goals.

The good:

- [This line is intentionally left blank]

The bad:

- Most research focuses on finetuning because finetuning is cheap... but that also means you can just retrain the model.
- No MU technique has ever been seriously studied in the counterfactual pretraining setting
- Most papers create a proxy for “unlearning” which doesn't have counterfactual validity

Recommendation: Demand actual comparisons between models trained on the data and not trained on the data in realistic regimes. Whenever possible, don't train on data that contains undesirable information.

SOME OTHER THINGS YOU SHOULD KNOW

Anything that adds a multiplier to model training costs is a non-starter

- Differential privacy
- Adversarial robustness
- Verifiable computation

There is no compelling evidence that “safety finetuning” is effective at deterring a motivated adversarial user.

LLM detectors are bunk

LLMs are not programs

- If you download an adversary’s open source model, that isn’t malware
- Some *checkpoint formats* (pickle) can hide malware, but SafeTensors can’t

CLOSING THOUGHT

LLMs are not deployed in a vacuum. They are situated in an organizational workflow.

A good litmus test for if a LLM deployment is appropriate is whether you can justify the result without mentioning the LLM.

Contact: biderman_stella@bah.com