databricks
mosaic research

Perplexed By Perplexity: Perplexity-Based Data Pruning With Small Reference Models

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L. Leavitt, Mansheej Paul

Outline

1. **Overview of LLM pretraining data**
2. How perplexity-based data pruning works
3. Results
4. Future directions

What Does A Data Researcher Do?

- Data quality is one of the largest factors for model quality
- To support the data diet of large models datasets are generally built from large scrapes of web data
- Once this large dataset is obtained we need to transform it to be of higher quality

What Is Perplexity?

- Measure of how surprised a model is on a sequence of text
- Proportional to the loss of the model on the sample

How Are Pretraining Datasets Curated?

- “Tried and true” Methods:
  - Rule based filtering (“bad” words, document length, etc.)
  - N-gram reference domain perplexity (similarity to “gold” domain, eg. Wikipedia)
- Pros:
  - Conceptually “simple” to implement
  - Very efficient and hence scalable
- Cons:
  - Potentially introduce bias
  - Rely on human heuristics of quality

How Are Pretraining Datasets Curated?

- Recent Methods:
  - Clustering statistics [Abbas et al.]
  - Domain learnability [Xie et al.]
  - Importance Resampling [Xie et al.]
  - And more!

Our Work Extends Previous Perplexity-Based Data Pruning Work

- Perplexity-based data pruning has been investigated previously [Marion et al.]
- We extend their work by:
  - Demonstrating that it works for small reference models
  - Investigating in new settings:
    a. Overtraining
    b. Data-constrained
    c. Downstream performance

Outline

1. Overview of LLM pretraining data
2. **How perplexity-based data pruning works**
3. Results
4. Future directions

What Are Important Properties Of A Data Pruning Method?

What Are Important Properties Of A Data Pruning Method?

1. Prune at the document level

What Are Important Properties Of A Data Pruning Method?

1. Prune at the document level
2. Efficient to run on large datasets

What Are Important Properties Of A Data Pruning Method?

1. Prune at the document level
2. Efficient to run on large datasets
3. Generally good data irrespective of domain (sometimes)

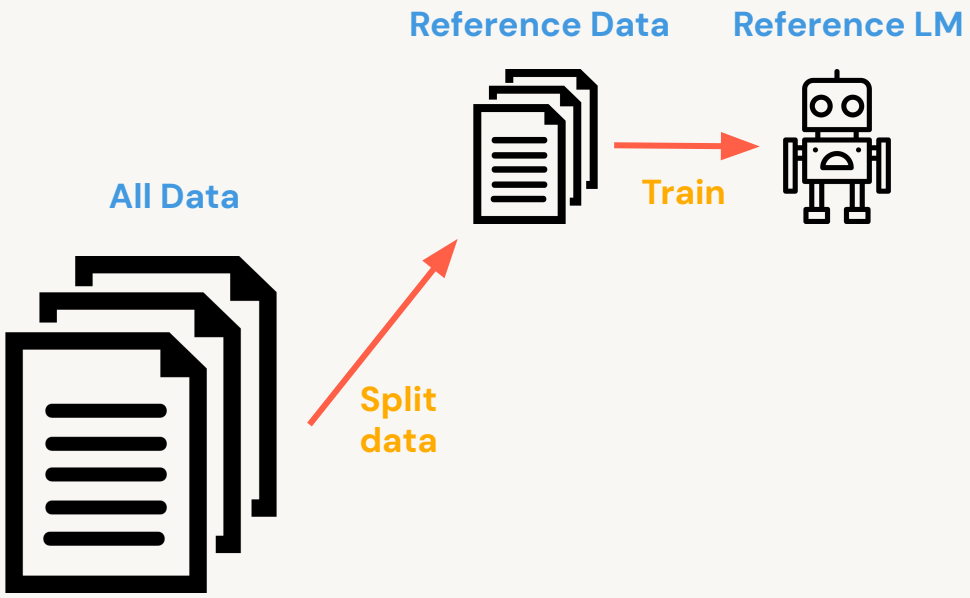
Strategy: Select Hard Examples To Train On

- How do we identify hard data points?
- Existing heuristics:
  - # Times example is forgotten [Toneva et al.], magnitude of gradient [Paul et al.], similarity to other datapoints [Abbas et al.], and many more
- One simple heuristic is the magnitude of the loss [Jiang et al.]
  - This is the heuristic we use in our work

Our Method

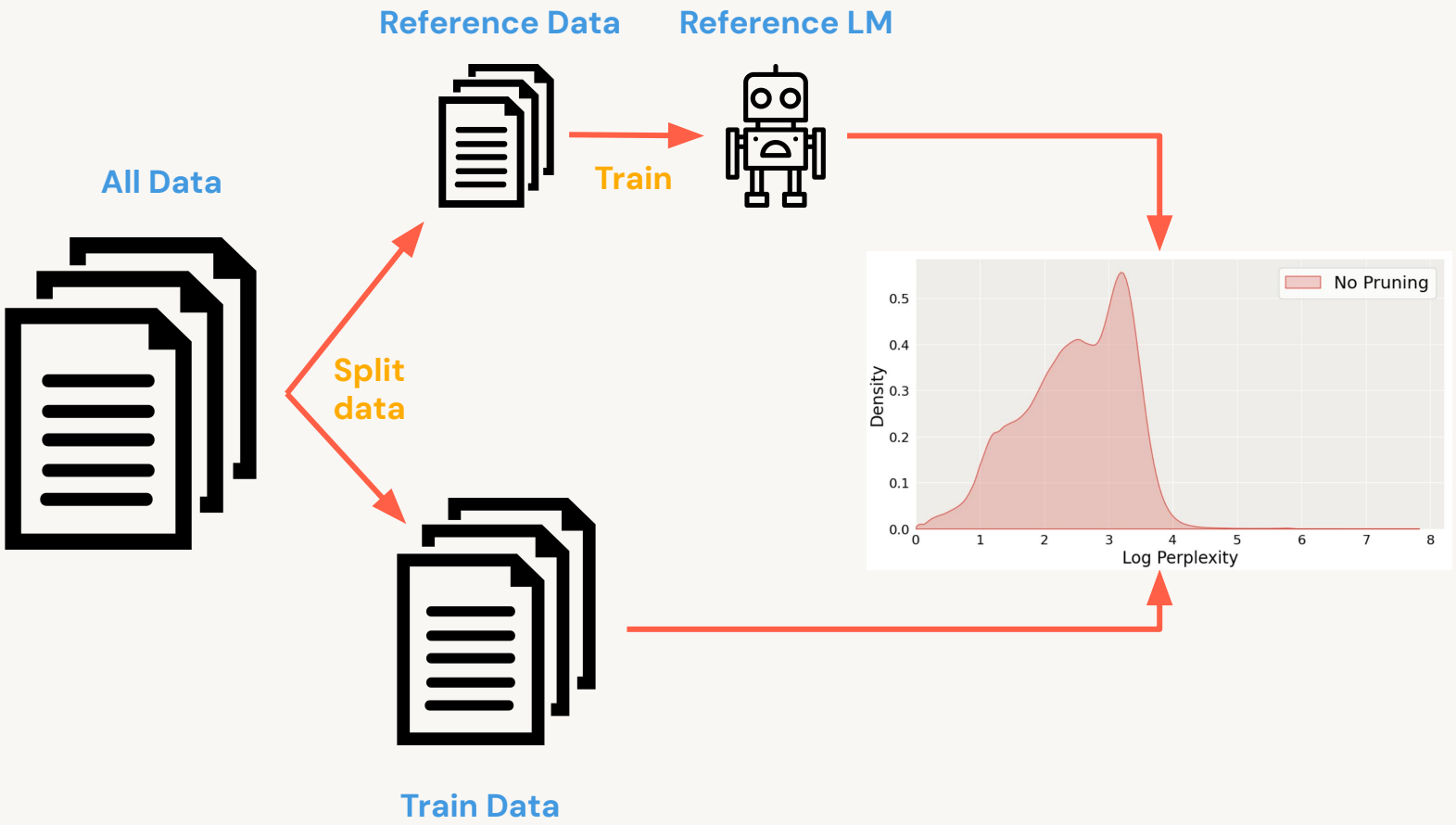
All Data

Our Method



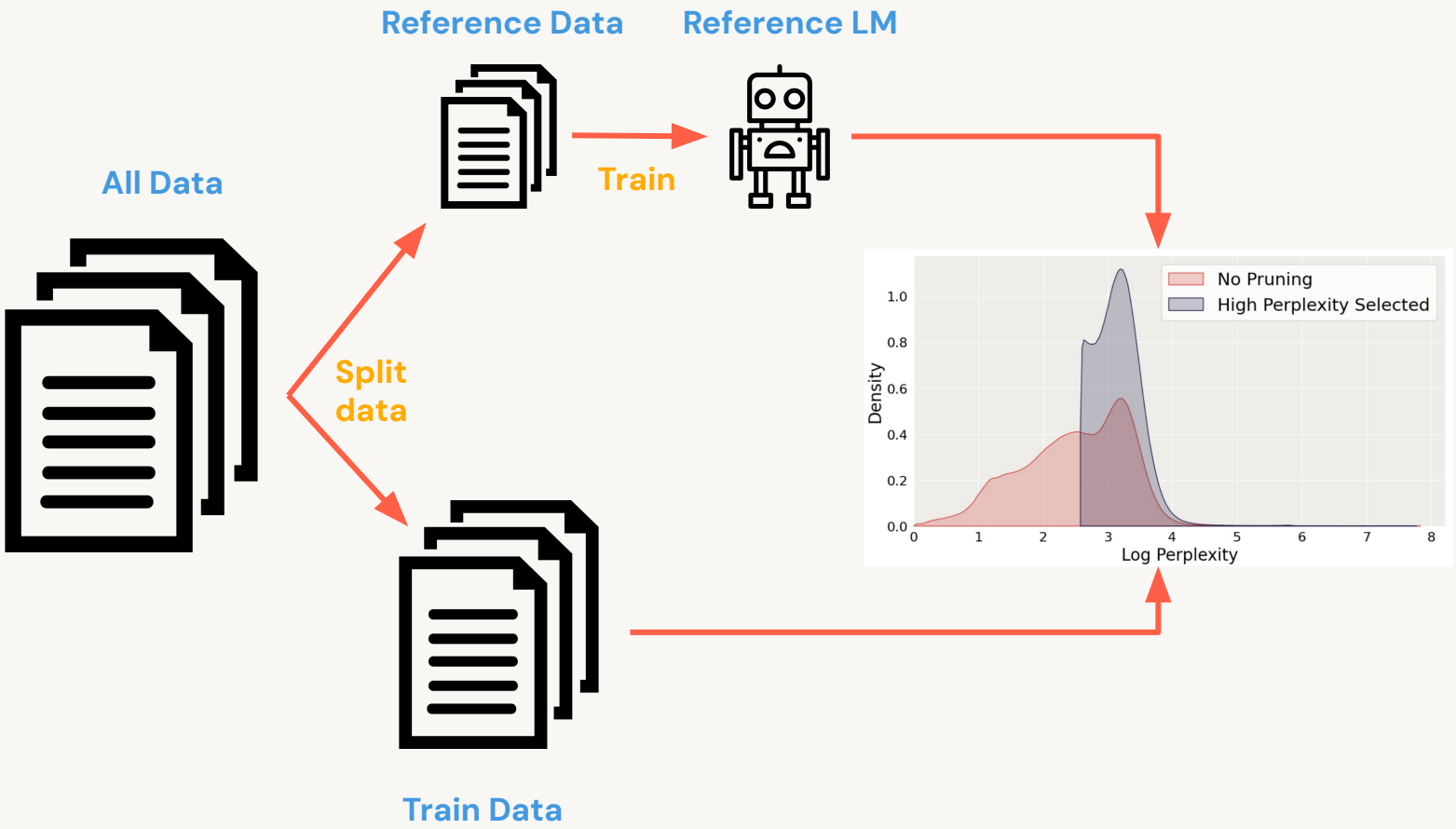
1. Create small split of data to train a reference LM

Our Method



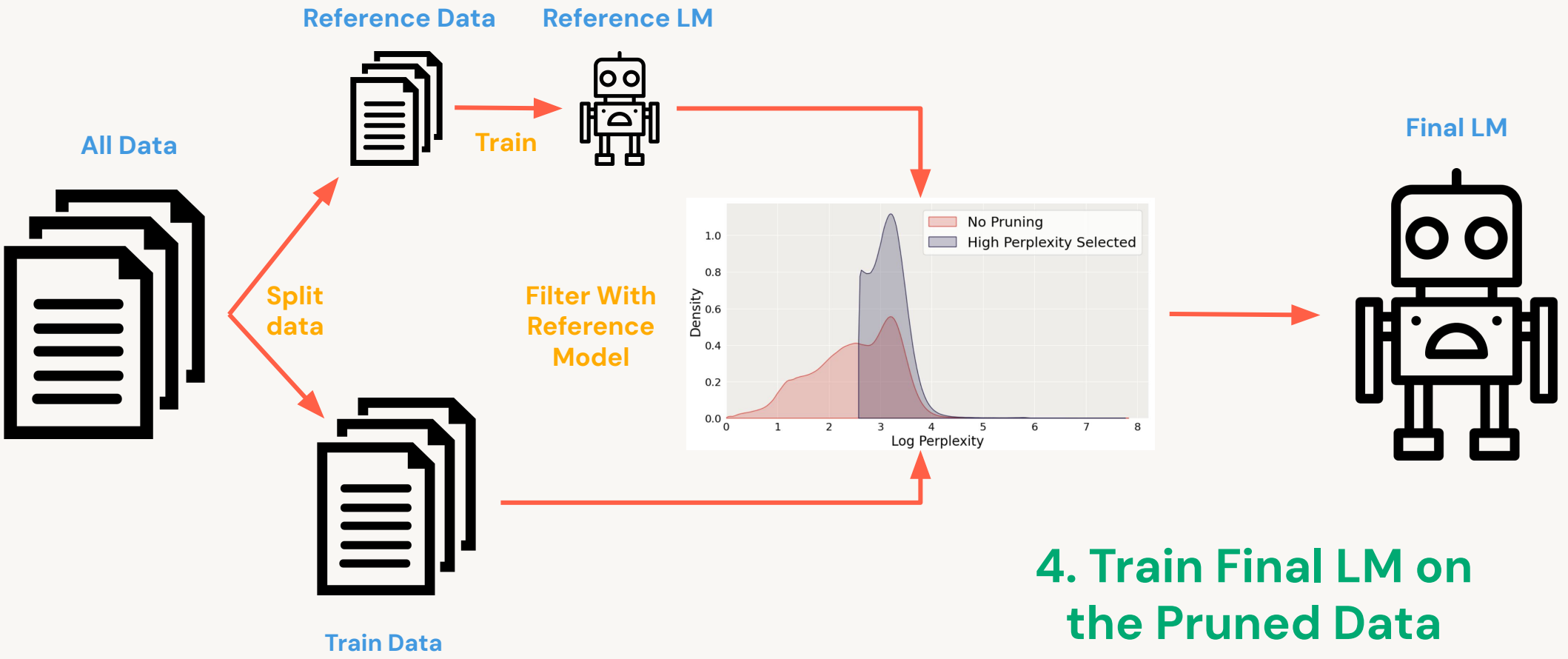
2. Evaluate perplexity of Reference LM on the remaining training data

Our Method



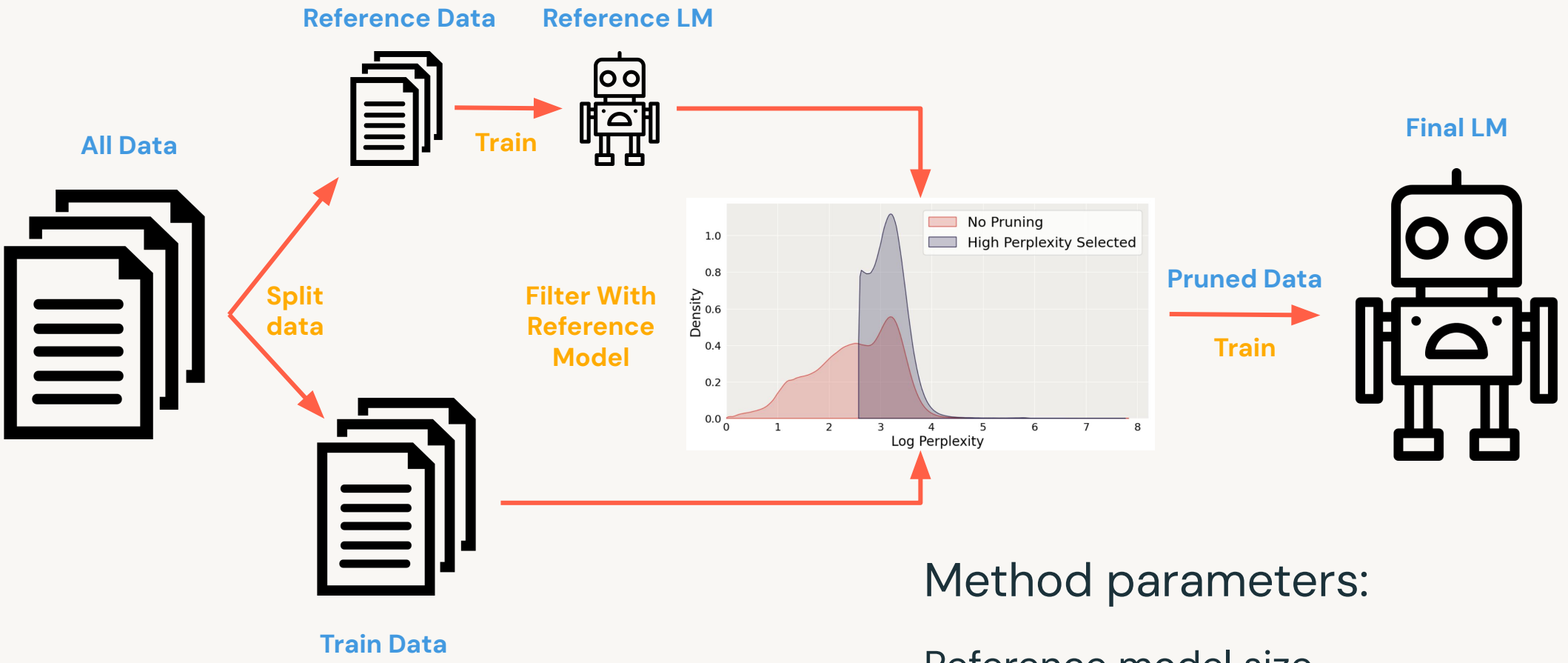
3. Select subset of train data based on each example's perplexity and selection settings

Our Method



4. Train Final LM on the Pruned Data

Our Method



Method parameters:

Reference model size,
Selection criteria (high, medium, low),
Selection rate

Outline

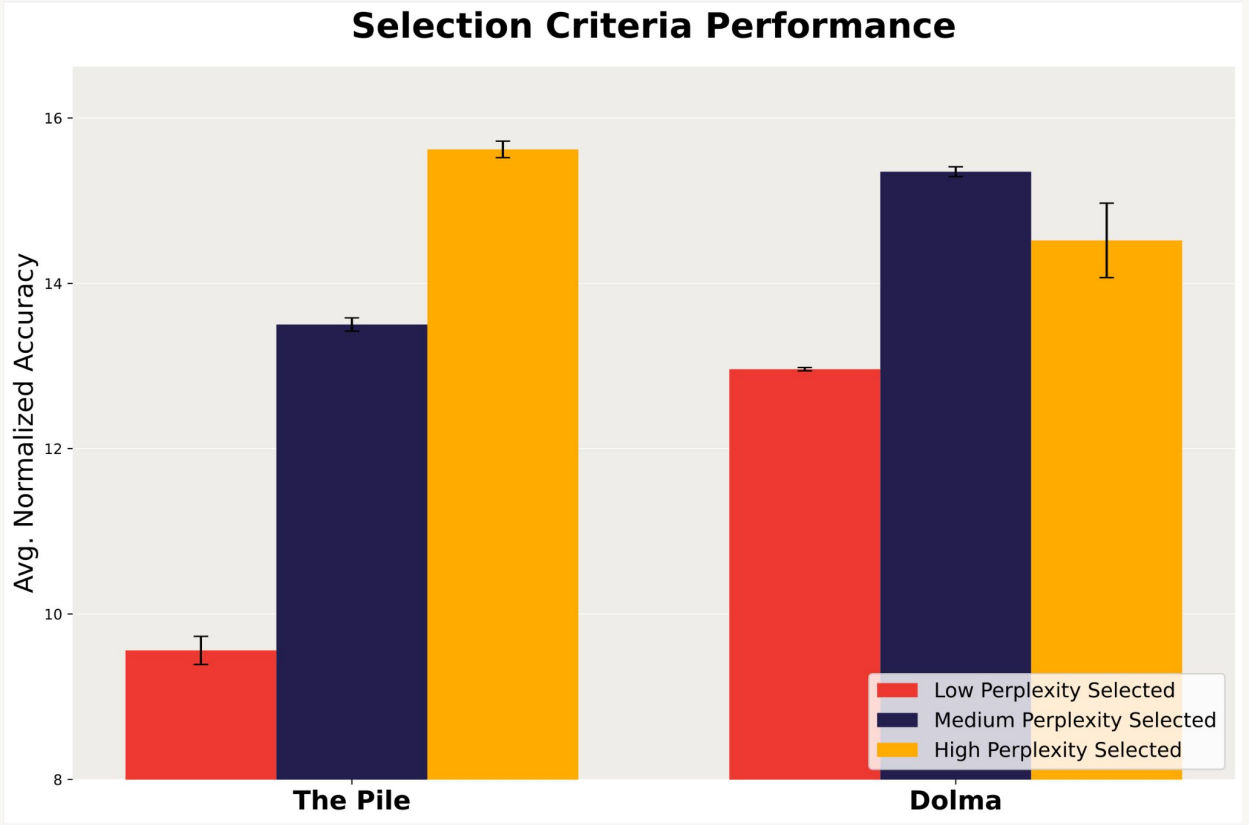
1. Overview of LLM pretraining data
2. How perplexity-based data pruning works
3. **Results**
4. Future directions

Experiment Setup

- Datasets
  - Train on the Pile [Gao et al.] and Dolma [Soldaini et al.]
    - The Pile is smaller and is composed of more diverse domains, Dolma is larger and primarily constructed general web data
- Models
  - 125 million parameter reference models and 1 billion & 3 billion parameter final models
- Evaluation
  - Evaluate on 33 different downstream tasks

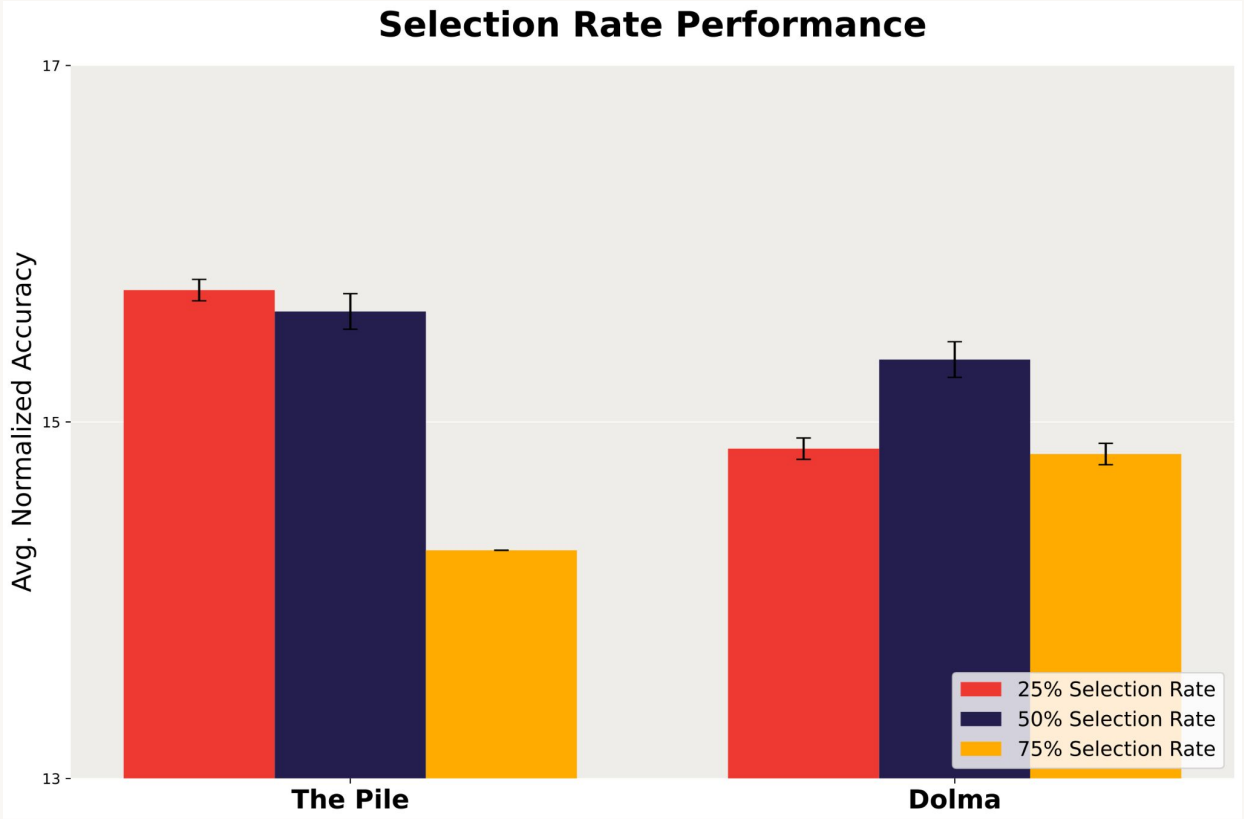
Optimal Selection Criteria Varies by Dataset

High perplexity samples are the best for the Pile but
medium perplexity samples are best for Dolma



Optimal Selection Rate Is More Robust Across Datasets

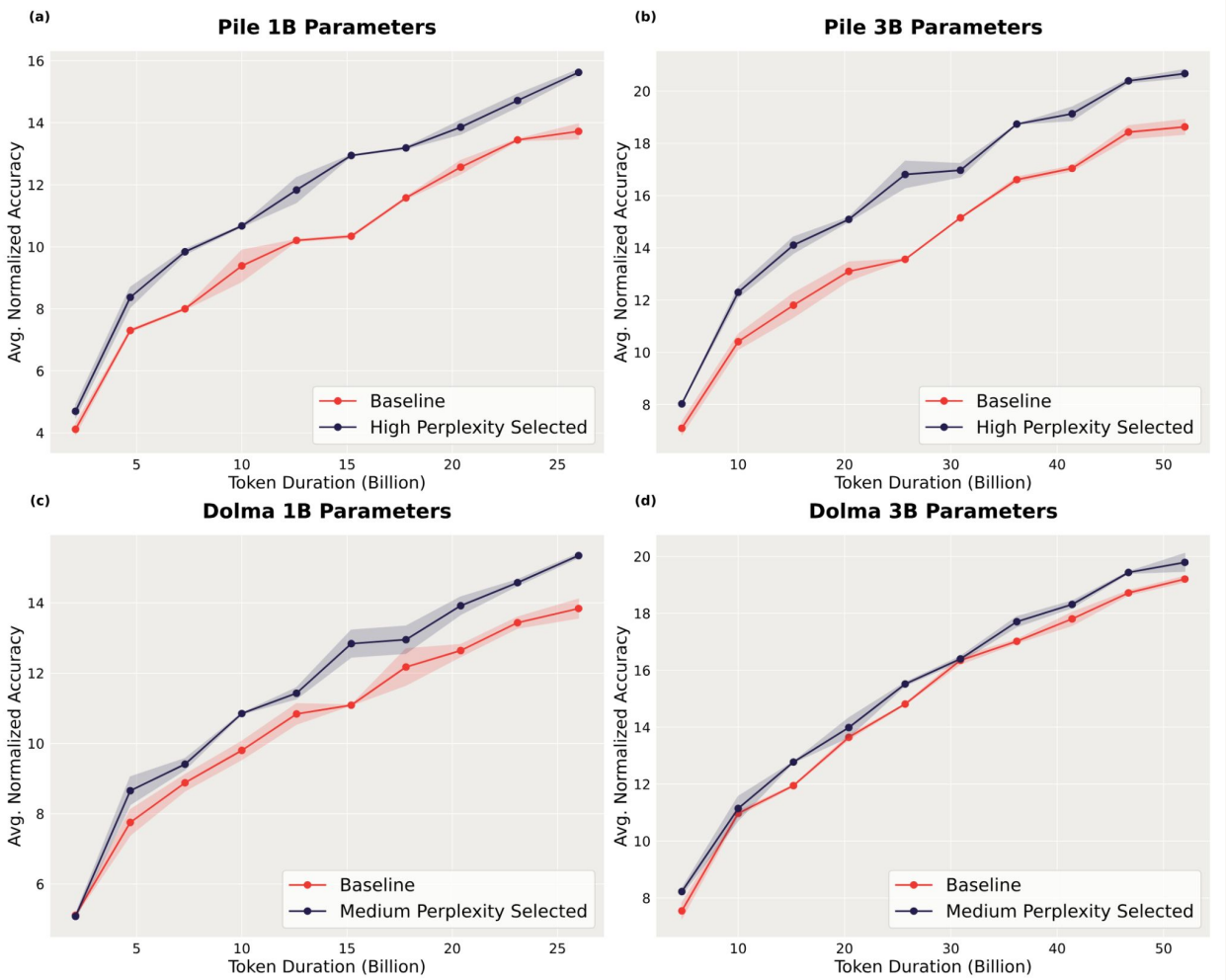
For both the Pile and Dolma, selecting 50% of samples is optimal



Perplexity Pruned Models Outperform Unpruned Models Through Training

Takeaways:

- Pruning improves the final model performance
- Pruning achieves the same accuracy with fewer pretraining steps
- Gains from pruning remain when increasing model size to 3 billion parameters

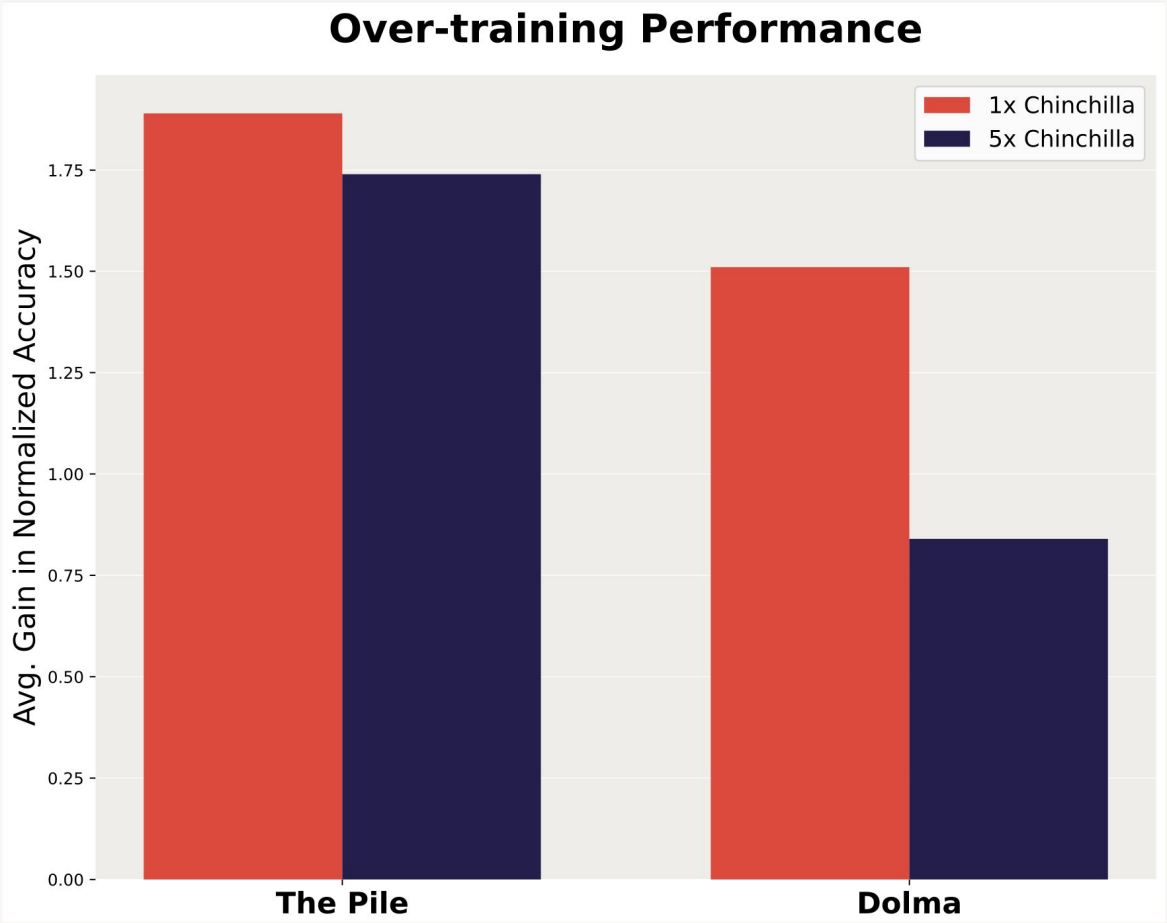


How Does Perplexity Pruning Work When Over-training?

- Chinchilla optimal to train for #tokens = 20 * #params
- Recent trend to train past chinchilla (i.e. more tokens)
  - DBRX, Llama, StarCoder2 [Gadre et. al.]
- Does pruning still work for over-trained models?
  - Investigate by training 1B models for 5x longer than chinchilla optimal

Perplexity Pruning Is Still An Improvement When Over-Training

Gain from pruning is the same for the Pile but slightly decreases on Dolma

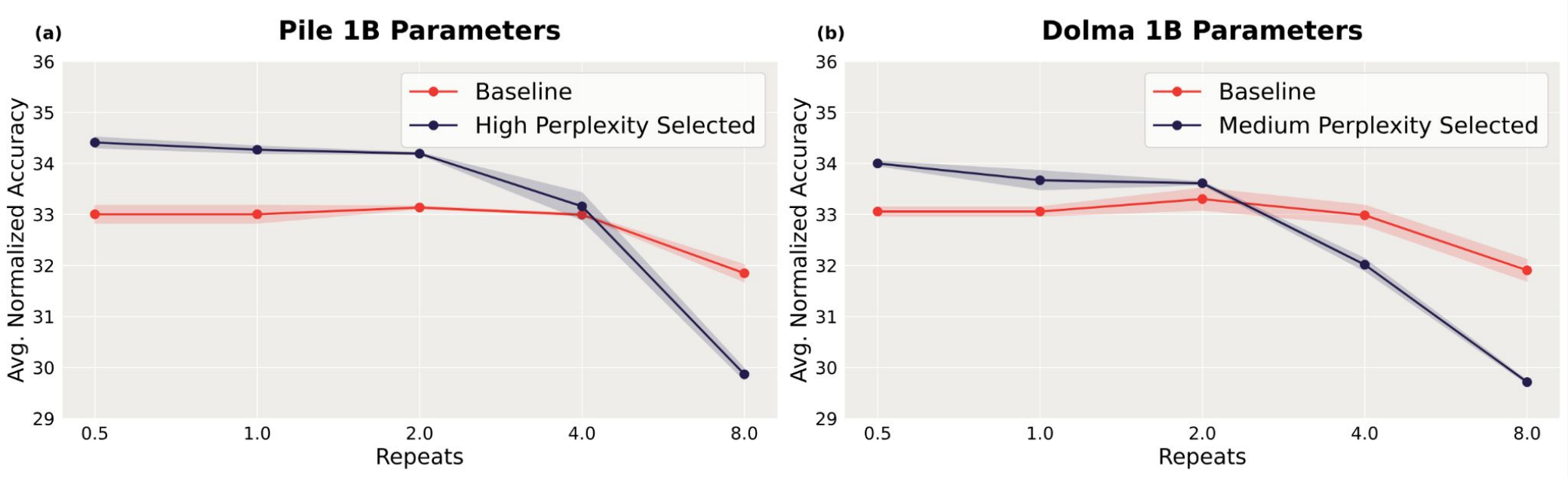


How Does Perplexity Pruning Work When Data-Constrained?

- Results so far assume abundance of data such that no repetitions are required post-pruning
- Muennighoff et al. showed that repeated data is “worth less” than fresh data after multiple repetitions
- Many use cases (eg. frontier LLMs, domain specific models) require multiple repetitions
- Experiment with requiring {0.5, 1, 2, 4, 8} repetitions

Perplexity Pruning Is Still An Improvement When Data-Constrained

For up to 2 passes through the data perplexity pruning outperforms no pruning



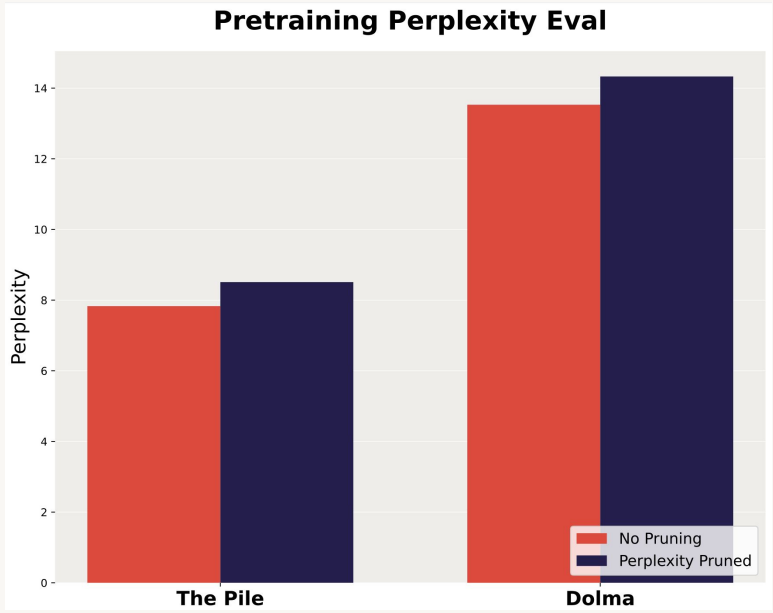
Is Pretraining Perplexity A Sound Evaluation Metric?

- Want models to perform well on “real world” tasks
- The field previously used a model’s perplexity on the test set of the pretraining data to approximate downstream performance
- Can a model with worse pretraining perplexity achieve better downstream performance? [Liu et al.]

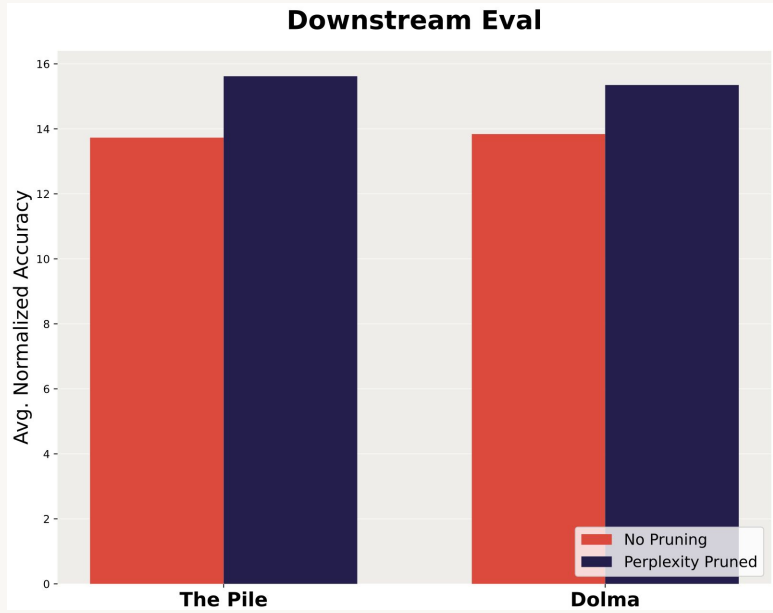
Pretraining Perplexity Is A Poor Metric For Data Pruning

Unpruned models have better pretraining perplexity but worse downstream task performance

Lower is better



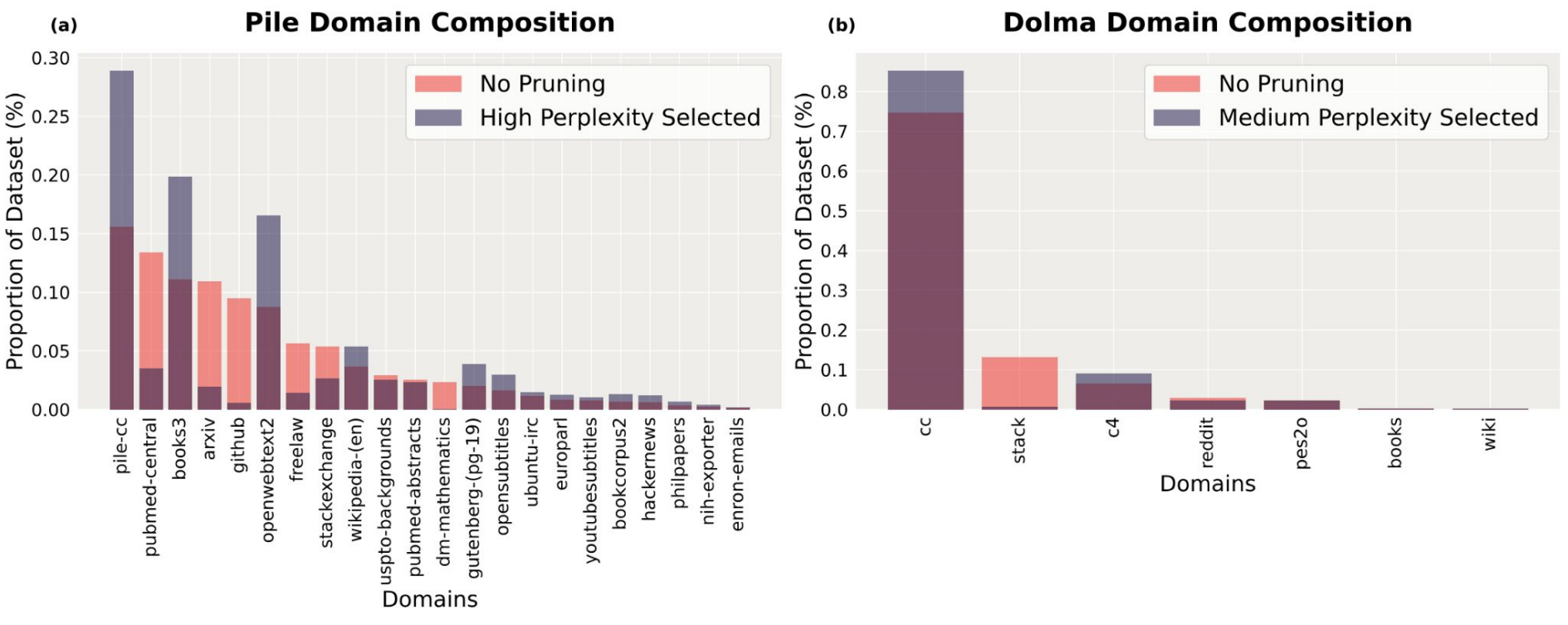
Higher is better



How Does Pruning Affect Domain Composition?

- The Pile and Dolma are composed of domains (sub datasets)
- Can interpret how perplexity-based data pruning works by looking at the domain composition before and after pruning

Pruning Upsamples General Web Domains And Downsamples Specialized Domains

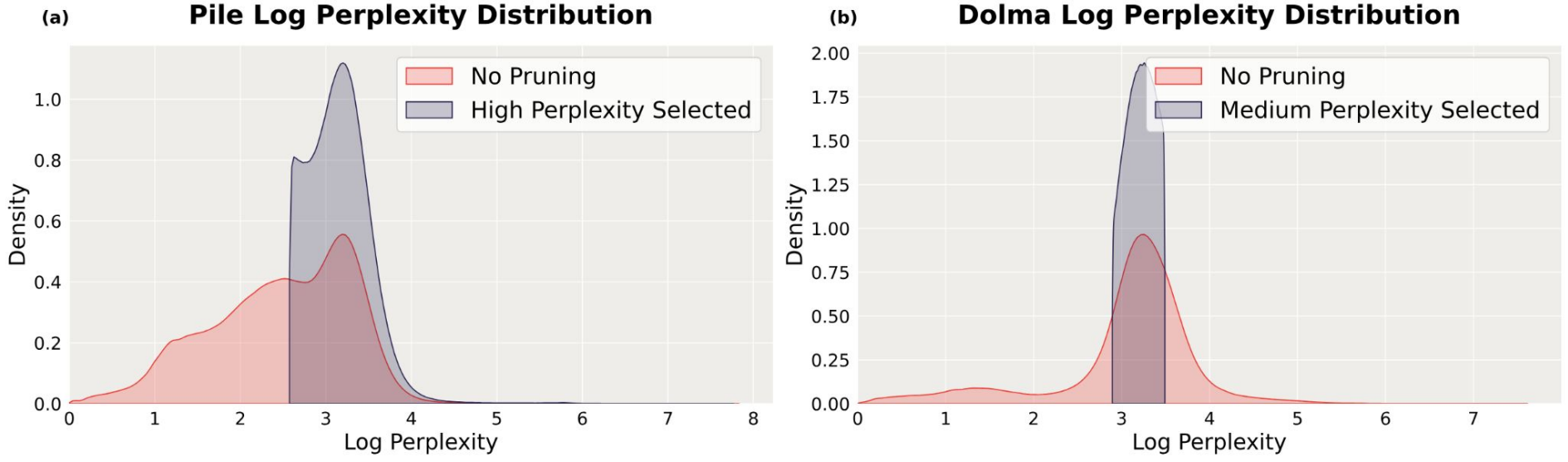


How Are Reference Model Perplexities Distributed?

- We can begin to understand the differences between datasets by inspecting the distribution of reference model perplexities
- Interpret prunings affect by examining the distribution post pruning

Reference Model Perplexity Is Distributed Similarly Across Datasets Post Pruning

Before pruning the Pile is skewed and multimodal while Dolma is symmetric and unimodal



Outline

1. Overview of LLM pretraining data
2. How perplexity-based data pruning works
3. Results
4. **Future directions**

Research Artifacts Release (coming soon)

Follow Up Directions

- Do perplexity-pruned models exhibit less bias?
- Can we define a curricula via reference model perplexity?
- Does training on the domain composition of the pruned dataset achieve the same performance?
- How much larger than the reference model can the final model be?

References

- Abbas, Amro, et al. "Semdedup: Data-efficient learning at web-scale through semantic deduplication." arXiv preprint arXiv:2303.09540 (2023).
- Xie, Sang Michael, et al. "Doremi: Optimizing data mixtures speeds up language model pretraining." Advances in Neural Information Processing Systems 36 (2024).
- Xie, Sang Michael, et al. "Data selection for language models via importance resampling." Advances in Neural Information Processing Systems 36 (2023): 34201-34227.
- Marion, Max, et al. "When less is more: Investigating data pruning for pretraining llms at scale." arXiv preprint arXiv:2309.04564 (2023).
- Muennighoff, Niklas, et al. "Scaling data-constrained language models." Advances in Neural Information Processing Systems 36 (2024).
- Toneva, Mariya, et al. "An empirical study of example forgetting during deep neural network learning." arXiv preprint arXiv:1812.05159 (2018).
- Paul, Mansheej, Surya Ganguli, and Gintare Karolina Dziugaite. "Deep learning on a data diet: Finding important examples early in training." Advances in neural information processing systems 34 (2021): 20596-20607.
- Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." arXiv preprint arXiv:2101.00027 (2020).
- Soldaini, Luca, et al. "Dolma: An open corpus of three trillion tokens for language model pretraining research." arXiv preprint arXiv:2402.00159 (2024).
- Liu, Hong, et al. "Same pre-training loss, better downstream: Implicit bias matters for language models." International Conference on Machine Learning. PMLR, 2023.

Questions?

Contact: ankner@mit.edu