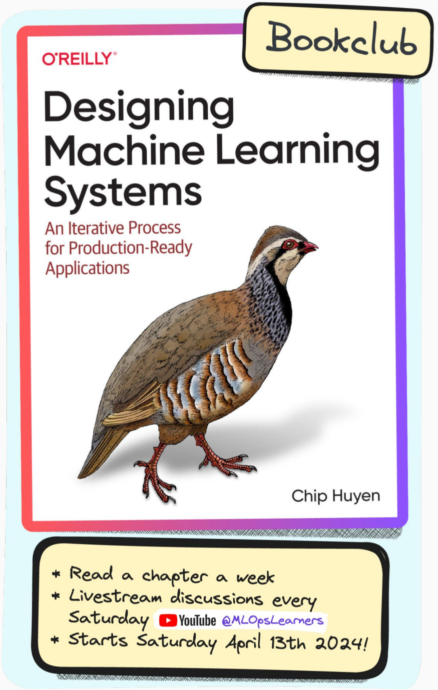


Chapter 4: Training Data

Group 2 (MLOpers)

Saturday, May 4, 2024

YouTube

https://www.youtube.com/watch?v=2xnYrib3tgI

MLOpers

Designing ML Systems Book Club

Chapter 4

TEAM MLOpers

Raksha V
Student
@ University of Michigan

Sebastien M
MLE
@Databook

Elias A, MD
Staff Software Engineer
@Gap

Introduction to Sampling in Machine Learning

Key focus:

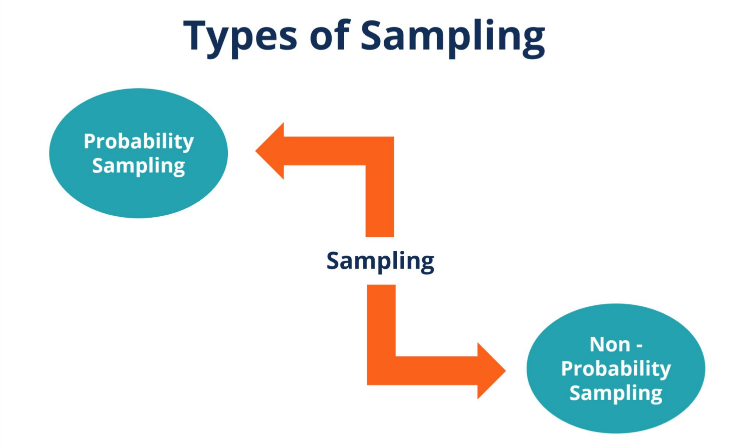
Sampling methods for creating training data.

Why Sampling is Necessary?

1. Limited access to all possible real-world data.
2. Feasibility issues due to resource constraints.
3. Quick experiments to test model hypotheses.

Types of Sampling

- Nonprobability Sampling
- Random Sampling



Nonprobability Sampling

Convenience Sampling : Samples selected based on availability.

Snowball Sampling : Future samples based on existing samples (e.g., scraping Twitter accounts).

Judgment Sampling: Samples selected by experts.

Quota Sampling: Samples based on predefined quotas without randomization (e.g., survey responses).

Limitations

- Not representative of real-world data.
- Riddled with selection biases.
- Commonly used despite limitations due to convenience.

Examples of Nonprobability Sampling

Language Modeling

Relies on easily collectible data sources like Wikipedia.

Sentiment Analysis

Utilizes biased data sources such as IMDB and Amazon reviews.

Self-Driving Cars

Data collection focuses on areas with favorable weather conditions.

Overview of Random Sampling Methods

Random Sampling Methods

Simple Random Sampling : All samples have equal probabilities of selection.

Stratified Sampling : Samples from different groups to ensure representation.

Weighted Sampling ; Assigns weights to samples to control selection probabilities.

Reservoir Sampling : Useful for streaming data, ensures equal probability for each sample.

Importance Sampling : Samples from one distribution based on another, useful in various ML tasks.

Reservoir Sampling

Definition: Ideal for data streams where the total size is unknown.

Process : Initialize a reservoir with the first k samples.

For each subsequent sample, replace it with a randomly

selected existing sample in the reservoir with decreasing probability.

Advantages

Ensures a representative sample for streaming data.

Efficient in memory usage.

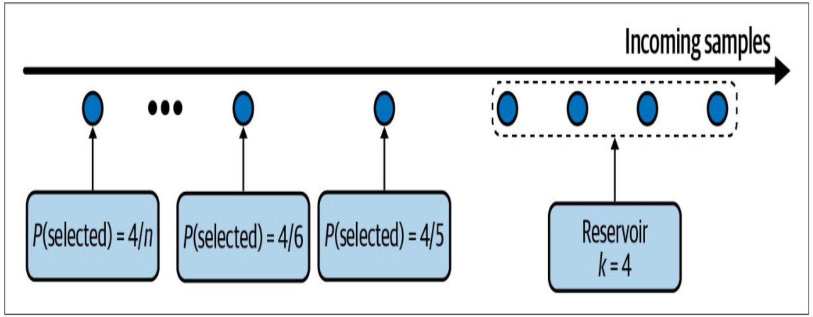


Figure 4-2. A visualization of how reservoir sampling works

Image Source : DSML Textbook

Importance Sampling

Definition : Used when sampling directly from the desired distribution is challenging.

Process : Sample from an easier or more convenient distribution (proposal distribution).

Adjust sampling by re-weighting the samples according to how probable they are in the target distribution versus the proposal distribution.

Advantages

Facilitates estimation of properties from a different distribution than the one sampled from.

Useful in scenarios like reinforcement learning where real-world trials are expensive or impractical.

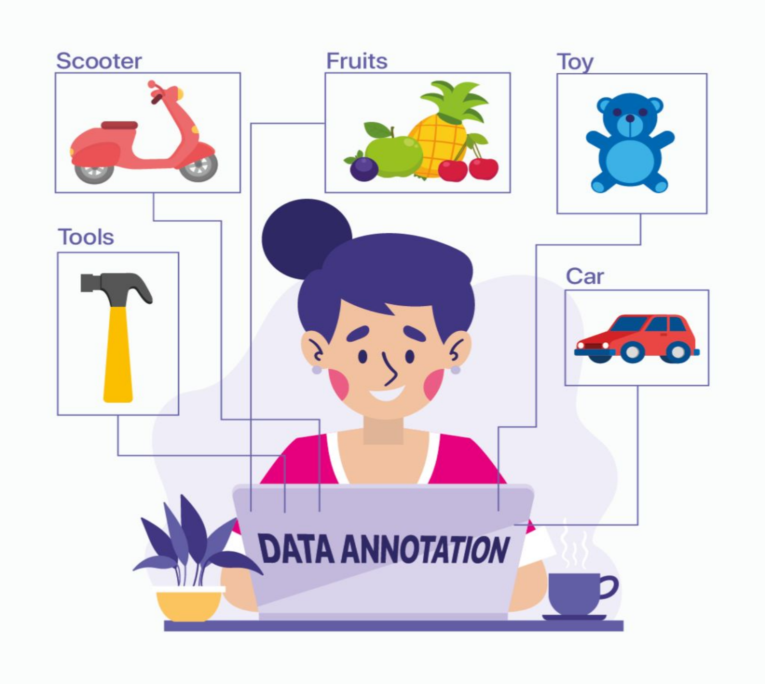
Introduction to Data Labeling in Machine Learning

Importance of Labeling : Crucial for supervised learning

Current Trends : Despite the rise of unsupervised learnin most production models are supervised and rely heavily on labeled data.

Methods of Labeling :

1. Hand Labels
2. Natural Labels



Challenges of Acquiring Labels

Hand-labeling Issues:

Cost: Expensive, especially when specialized expertise is needed (e.g., radiologists for X-rays).

Privacy Concerns: Risk with sensitive data; cannot easily outsource without compromising confidentiality.

Time Consumption: Slow process, e.g., transcribing speech can take 400 times the duration of the recording.

The Reality of Label Multiplicity and Data lineage

Diverse Annotator Input: Different annotators

often produce varied labeling outcomes.

Example: Entity recognition task with multiple annotators leading to conflicting labels.

Impact: Such variability can significantly

influence the training and performance of

It's good practice to keep track of the origin of each

of your data samples as well as its labels, a technique known as data lineage

ML models.

Image Source : DSML Book

Table 4-1. Entities identified by different annotators might be very different

Annotator	# entities	Annotation
1	3	[*Darth Sidious*], known simply as the Emperor, was a [*Dark Lord of the Sith*] who reigned over the galaxy as [*Galactic Emperor of the First Galactic Empire*].
2	6	[*Darth Sidious*], known simply as the [*Emperor*], was a [*Dark Lord*] of the [*Sith*] who reigned over the galaxy as [*Galactic Emperor*] of the [*First Galactic Empire*].
3	4	[*Darth Sidious*], known simply as the [*Emperor*], was a [*Dark Lord of the Sith*] who reigned over the galaxy as [*Galactic Emperor of the First Galactic Empire*].

Natural Labels - An Efficient Alternative

Definition and Examples:

Labels derived from system outputs or user interactions (e.g., Google Maps ETA predictions, stock price predictions).

Recommender systems where user clicks provide implicit feedback.

Advantages: Reduce the need for manual labeling, utilize real-time data feedback.

In summary, in addition to manual human labeling and Natural labels other emerging methods include self-supervised learning, semi-supervised learning,natural language supervision and automated labeling, often used in a blended approach to optimize the data labeling workflow.

Strategies to Address Labeling Challenges

Weak Supervision: Using programmatic heuristics to generate labels when hand-labeling isn't feasible.

Semi-Supervision: Combines a small amount of labeled data with a large amount of unlabeled data, using assumptions about the data structure.

Transfer Learning: Utilizing a model trained on a different but related task to reduce the need for extensive labeled data in the new task.

Active Learning: Selectively labeling data that the model deems most informative, improving efficiency and effectiveness.

Data Augmentation

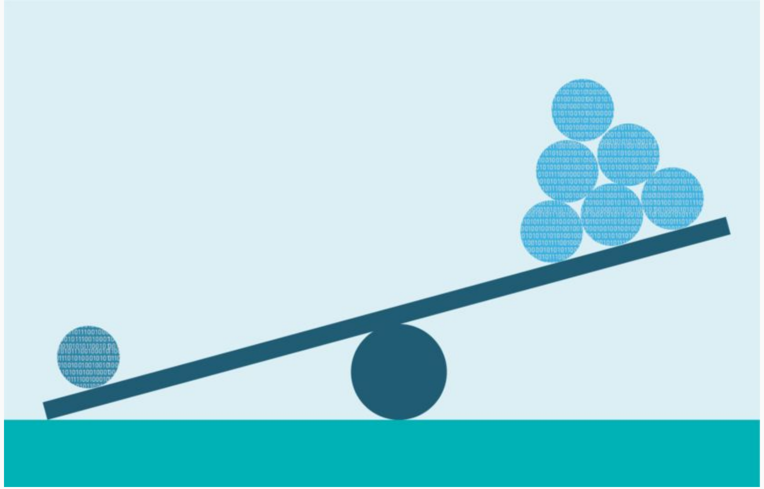
Perturbation:

- Rotation
- Zoom
- Crop

Data Synthesis:

- Template based
- Back translation
- Paraphrase

Class Imbalance



Loss functions:

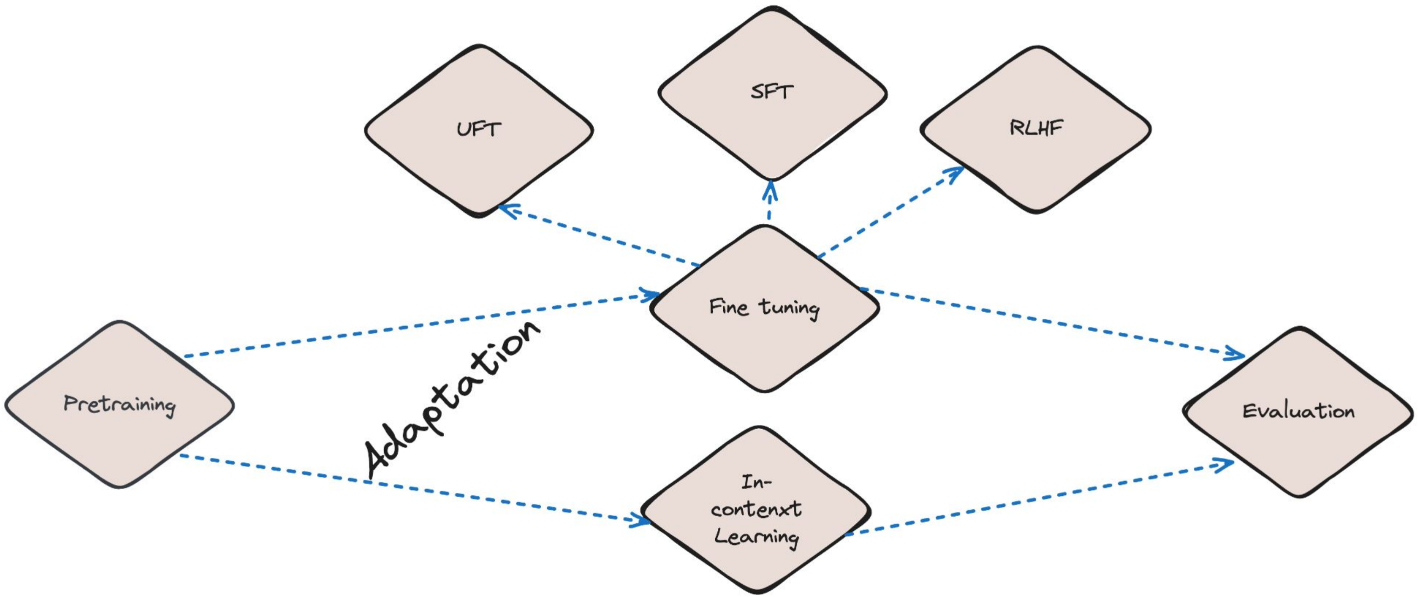
- Class-balanced Cross Entropy loss
- Focal loss

Metrics:

- F1 score
- ROC AUC
- PR AUC

The new era

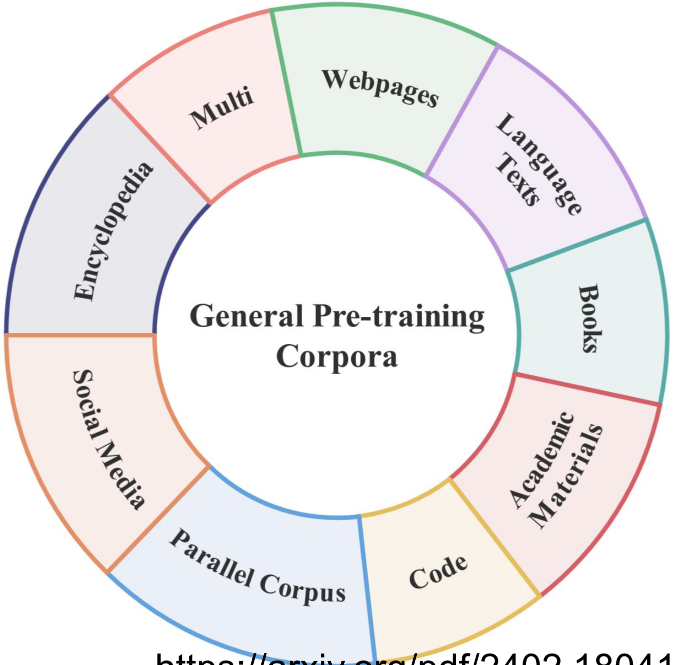
Stages in LLM training



Pretraining

- Pre Training corpora is massive amount of data
  - Compliance and copyright issues
  - Diversity - multicategory → better quality
  - Requires rigorous cleaning

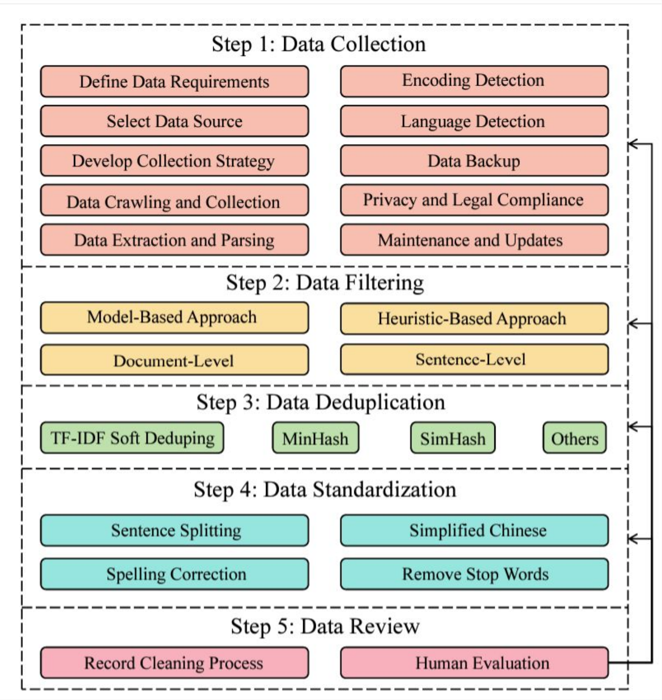
https://arxiv.org/pdf/2402.09668



https://arxiv.org/pdf/2402.18041

Pretraining

- Preprocessing steps
  - Data filtering: human vs classifier
  - Data dedup: affects performance and data memorization
  - Privacy reduction: heuristics-based



https://arxiv.org/pdf/2402.18041

Pretraining

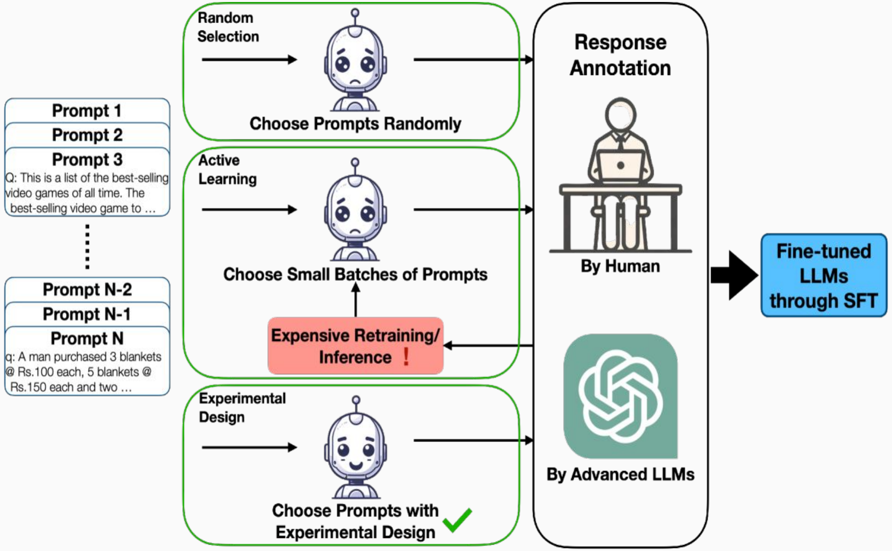
- Massive data is not sustainable
    - Rise of data efficient LLMs
        - Coverage sampling - in embedding space
        - Quality-score sampling -
            - Density sampling
            - Ask-LLM sampling
- Random sampling (ignores unbalanced distribution) → clusterclip sampling (cluster sampling followed by clipping overrepresented samples) (https://arxiv.org/html/2402.14526v1)

Fine-tuning

- Further training of LLMs on curated dataset
- Types
  - Supervised fine-tuning
    - Instruction fine-tuning
  - RLHF - preference alignment
- Consideration
  - Parameter efficient fine-tuning (PEFT) over transfer learning (T5 and mT5) or FFT
    - Retains model in-context learning ability and less expensive
  - Reduce cost of human annotated with Self-play fine-tuning (https://arxiv.org/pdf/2401.01335)

Fine-tuning

- Active learning better than random sampling
- Active learning expensive due to model retraining and inference for every batch
- Experimental design low cost and better label-efficiency



https://arxiv.org/html/2401.06692v1

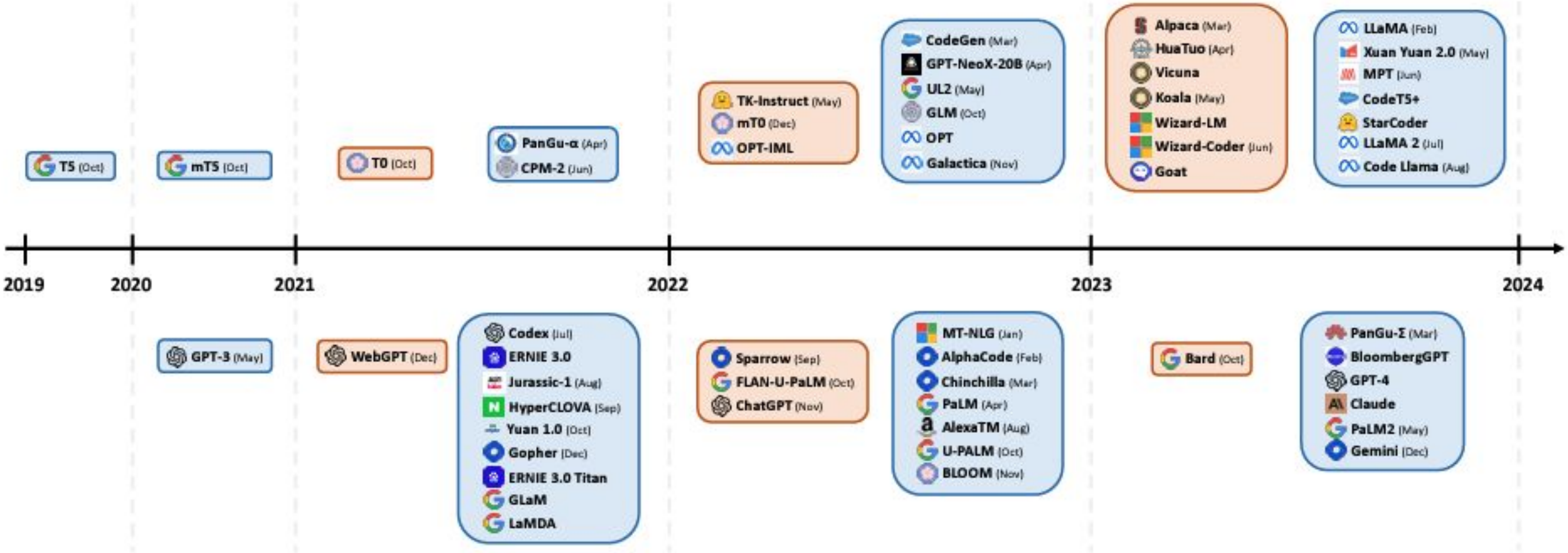
Bias

- Pre-training - bias and stereotypes in the massive corpora
- Overrepresentation of some training data (challenging class imbalance)
- Encoding bias – BERT associating disability with more negative sentiments (Hutchinson et al)

Privacy

- Preventing adversarial attacks with adversarial fine-tuning
  - Membership inference attacks
  - Training data extraction (https://arxiv.org/pdf/2012.07805)
- Training with differential privacy (DP-SGD)
- Data protection is not equivalent to privacy protection for natural language data
- Data sanitization mayn’t be enough, private data is context dependent

Where it is heading?



https://arxiv.org/pdf/2307.06435