

CSE/DSC 234

Data-Centric AI and AI Engineering

Arun Kumar

Topic 1: Creating Data and Evals

Outline

- ➔ ❖ Motivation for Data-Centric AI
 - ❖ Creating Data
 - ❖ Creating Evals

From Model-Centric to Data-Centric

- ❖ Transformers now dominate; DL architectures converged
- ❖ “Scaling laws”: data matters as much as parameters
- ❖ Yet dataset eng. remains “artisanal—painstaking and expensive”
- ❖ DCAI = systematically engineering data to build AI systems

AI system = Code + Data

Model-centric AI

How can you change the model (code) to improve performance?

Data-centric AI

How can you systematically change your data (inputs x or labels y) to improve performance?

[Andrew Ng Talk on Data-Centric AI 2021](#)

From Model-Centric to Data-Centric

- ❖ DCAI Competition: improve dataset, hold model fixed

DeepLearning.AI | LANDING AI

Data-Centric AI Competition

Join the data-centric AI movement!

Click here to enter the contest!



0:04 / 8:37

About the competition

A collaboration between DeepLearning.AI and Landing AI, the Data-Centric AI Competition aims to elevate data-centric approaches to improving the performance of machine learning models.

In most machine learning competitions, you are

Leaderboard

Rank	Submission	Accuracy
	Baseline	0.64421
	DeepLearning.AI (https://www.deeplearning.ai)	
1	iter3_002 Divakar Roy (https://www.linkedin.com/in/droyed/)	0.85826
2	baseline-cleaned-Naugmented Innotescus	0.85744

<https://https-deeplearning-ai.github.io/data-centric-comp/>

From Model-Centric to Data-Centric

- ❖ Two main levers:
 - ❖ *Better* data (refine quality)
 - ❖ *More* data (collect, augment, and/or synthesize)

Real-World AI Data Problems

- ❖ Label inconsistency; annotators disagree; e.g., “chipped” vs. “scratched”



Examples
of defects



Baseline system: 76.2% accuracy
Target: 90.0% accuracy

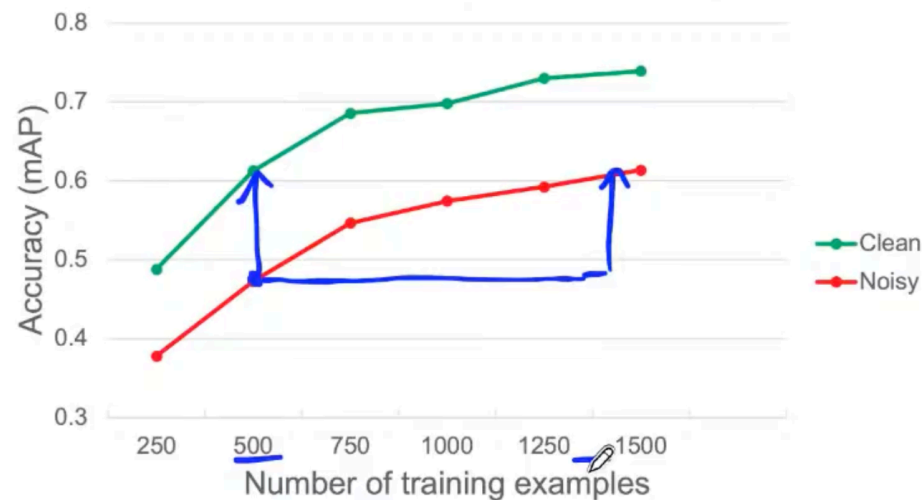
- ❖ Tweaking the model/hyperparameters -> No improvement
- ❖ Fixing labeling errors -> 16% lift!

Real-World AI Data Problems

- ❖ Label inconsistency; annotators disagree
 - ❖ ~3.3% label errors in top 10 ML benchmark test sets!
- ❖ Bias and fairness baked into datasets, not just models
- ❖ Many small-data domains; cannot rely on scale alone
- ❖ “Everyone wants to do the model work, not the data work”

Quality > Quantity

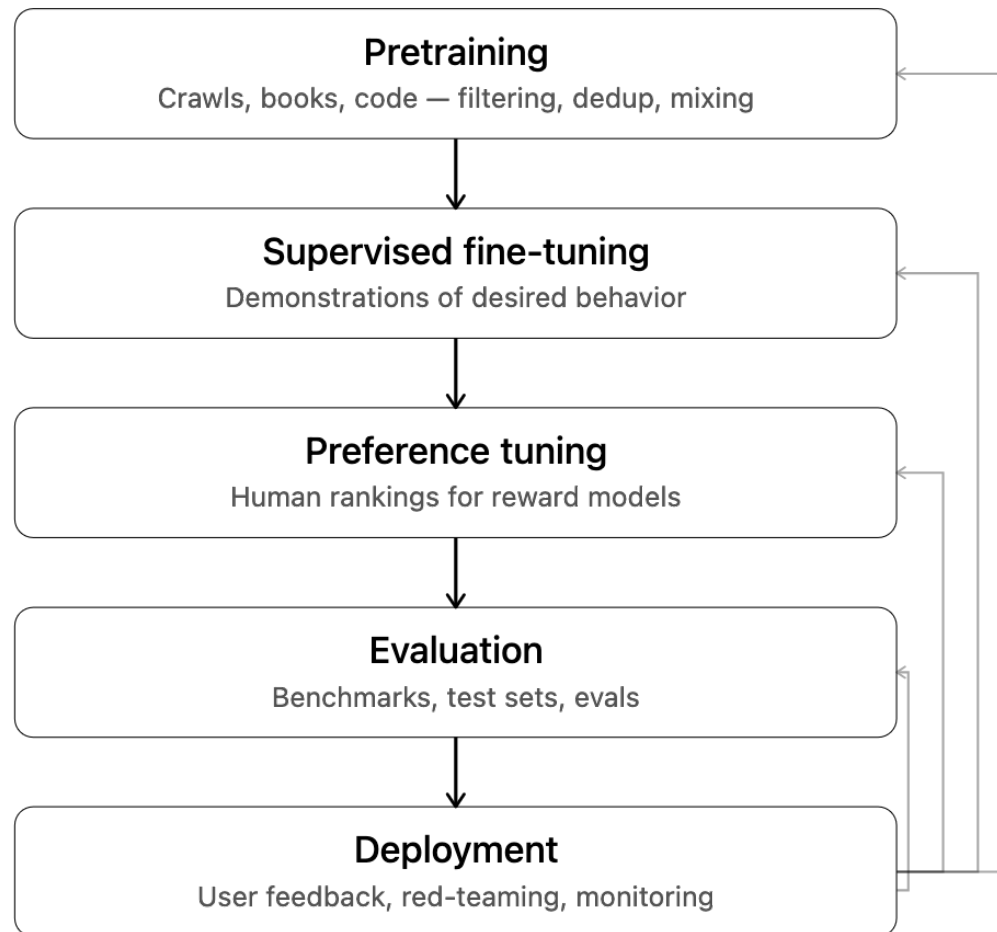
- ❖ Core insight: curated subset beats full noisy dataset.
- ❖ But “quality” depends on what you're optimizing for.
- ❖ Phi-1: 1.3B code model matched larger models via “textbook-quality” data.
- ❖ FineWeb-Edu: 1.3T from 15T -> better benchmarks by filtering hard.



Note: Big data problems where there's a long tail of rare events in the input (web search, self-driving cars, recommender systems) are also small data problems.

Data in LLM Development Lifecycle

- ❖ “Creating data” is continuous engineering, not a one-time step



Outline

- ❖ Motivation for Data-Centric AI
- ➔ ❖ Creating Data
- ❖ Creating Evals

Outline

- ❖ Motivation for Data-Centric AI
- ❖ Creating Data
- ➔ ❖ Human Feedback and Annotation
 - ❖ Data Curation and Structure
 - ❖ Data Synthesis
 - ❖ Multimodal Data
- ❖ Creating Evals

Human Feedback as Training Signal

- ❖ Classical ML: labels -> fixed ground “truth”
- ❖ LLMs: feedback is typically subjective, comparative, ongoing
 - ❖ Shift: from “annotate once” to continuous feedback loops

Demonstrations

"Here's how to answer this"

Prompt:

Summarize this paper

Desired response:

The paper presents a novel method for...

→ Used for SFT

Comparisons

"A is better than B"

Response A:

Clear, concise answer

Response B:

Vague, rambling answer

A > B

→ Used for preference

Corrections

"This is wrong, fix it"

Model output:

Paris is in Germany

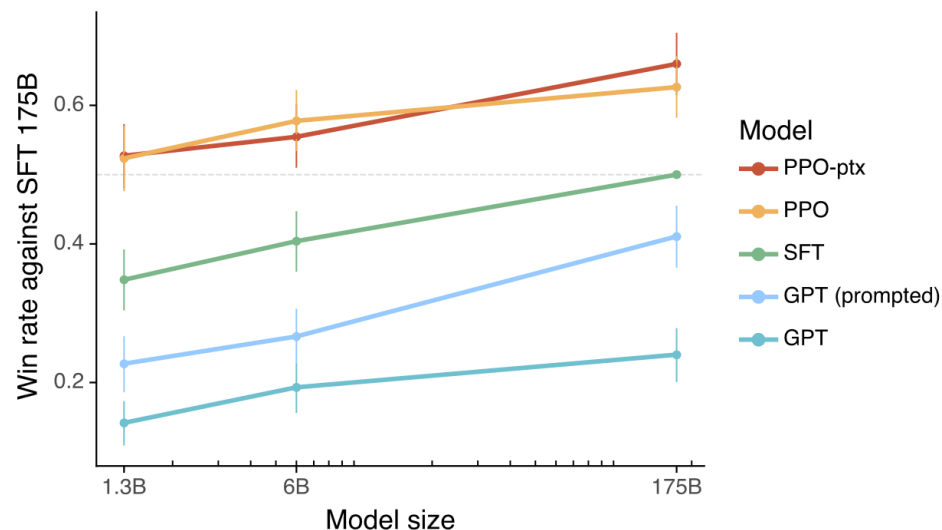
Corrected:

Paris is in France

→ Used for editorial

Pivot Point in AI: InstructGPT

- ❖ Showed how to make LLMs follow task-agnostic “instructions”
- ❖ First at-scale success of Reinforcement Learning with Human Feedback (RLHF) paradigm
- ❖ 40 screened contractors, detailed guidelines, calibration rounds
- ❖ 1.3B InstructGPT preferred over 175B GPT-3!
- ❖ 100x fewer params! Quality human feedback >> raw scale



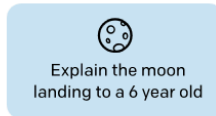
Pivot Point in AI: InstructGPT

Supervised Fine-Tuning (SFT)

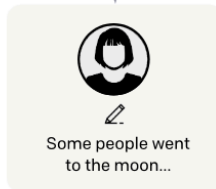
Step 1

Collect demonstration data, and train a supervised policy.

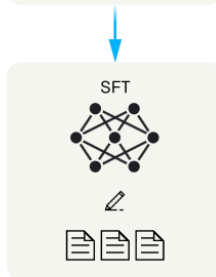
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.

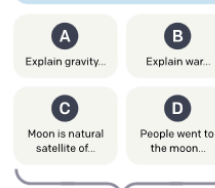
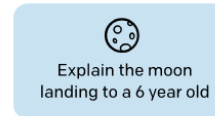


Reward Model (RM)

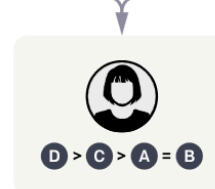
Step 2

Collect comparison data, and train a reward model.

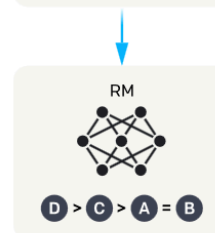
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



RL via Proximal Policy Optimization (PPO)

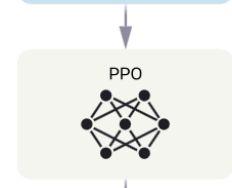
Step 3

Optimize a policy against the reward model using reinforcement learning.

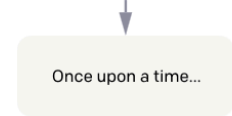
A new prompt is sampled from the dataset.



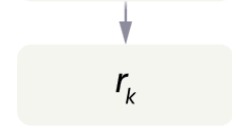
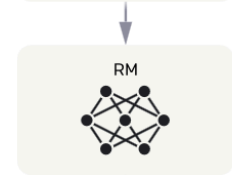
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Pivot Point in AI: InstructGPT

Submit Skip Page 3 / 11 Total time: 05:39

Instruction

Summarize the following news article:

====
(article)
====

Output A

summary1

Rating (1 = worst, 7 = best)

1 2 3 4 5 6 7

Fails to follow the correct instruction / task ? ☐ Yes ☐ No

Inappropriate for customer assistant ? ☐ Yes ☐ No

Contains sexual content ☐ Yes ☐ No

Contains violent content ☐ Yes ☐ No

Encourages or fails to discourage violence/abuse/terrorism/self-harm ☐ Yes ☐ No

Denigrates a protected class ☐ Yes ☐ No

Gives harmful advice ? ☐ Yes ☐ No

Expresses moral judgment ☐ Yes ☐ No

Notes

(Optional) notes

(a)

Ranking outputs

To be ranked

Rank 1 (best)

A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

Rank 2

Rank 3

Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

Rank 4

Rank 5 (worst)

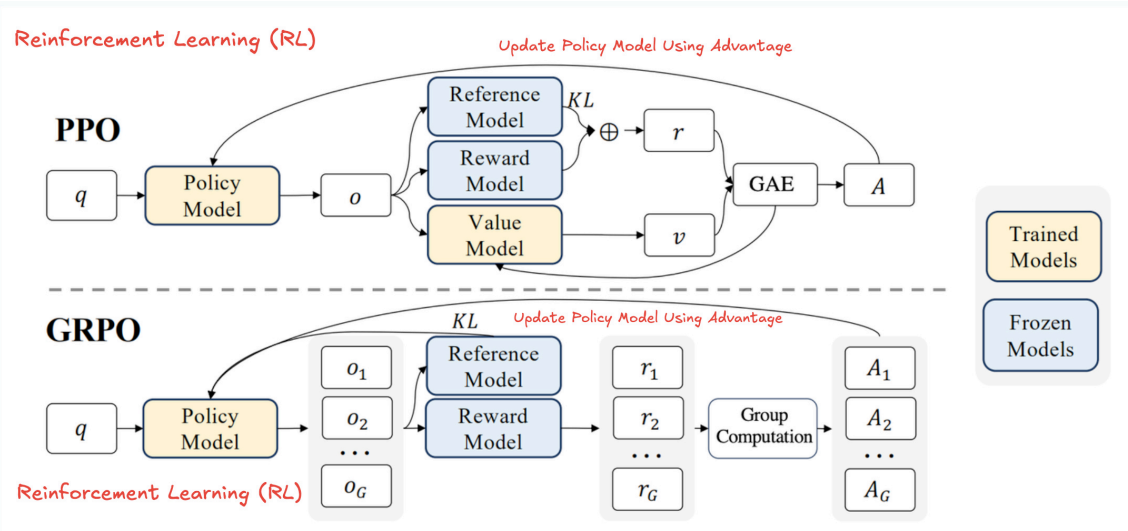
(b)

Figure 12: Screenshots of our labeling interface. (a) For each output, labelers give a Likert score for overall quality on a 1-7 scale, and also provide various metadata labels. (b) After evaluating each output individually, labelers rank all the outputs for a given prompt. Ties are encouraged in cases where two outputs seem to be of similar quality.

Beyond RLHF: Post-Training Toolkit

- ❖ Direct Preference Optimization (DPO): skip reward model, optimize directly on preferences; much simpler than PPO
- ❖ Constitutional AI (Anthropic) / RLAIIF: model self-critiques via principles; replaces human labelers with AI feedback
- ❖ RLVR: verifiable rewards for math/code (no human needed)
- ❖ Group Relative Policy Optimization (GRPO); skip critic model; sample N completions per prompt, normalize rewards within group; ~50% memory reduction vs PPO
- ❖ Typically modern post-training = SFT -> Preference FT -> RLVR
- ❖ Each stage has its own data requirements

Beyond RLHF: Post-Training Toolkit



Will revisit these 3 methods in Topic 4: Model Customization

Direct Preference Optimization (DPO)

x: "write me a poem about the history of jazz"



y_w

y_l

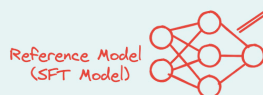
preference data



Policy Model



final LM

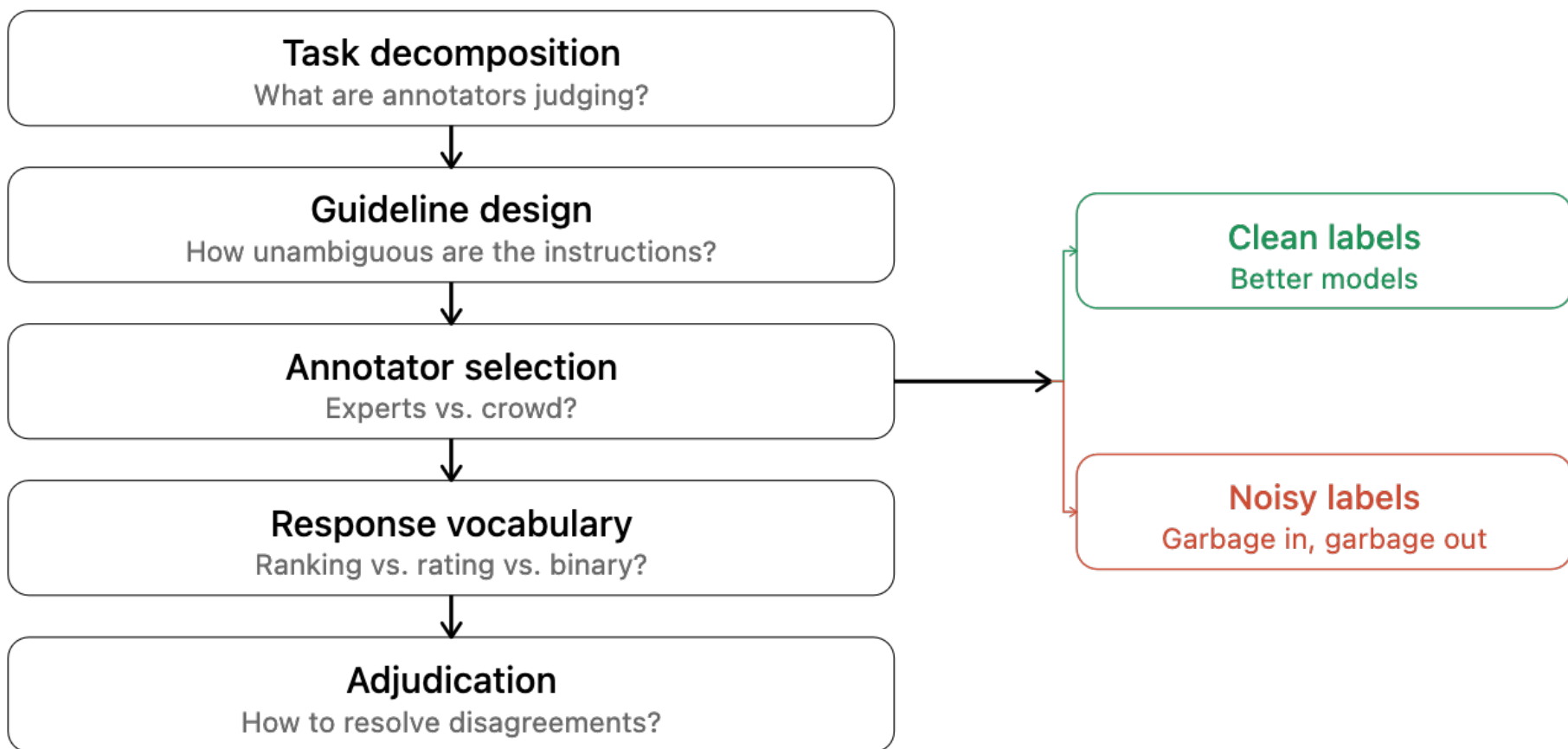


Reference Model (SFT Model)

maximum likelihood

Annotation is a Design Problem

❖ InstructGPT' success was partly annotation engineering



Measuring Annotation Quality

- ❖ If annotators disagree, your ground “truth” is not entirely “true”
- ❖ But high agreement $\leftrightarrow$ Validity; can be consistently wrong!
- ❖ Metrics (increasing generality):

Cohen's κ

Exactly 2 raters
No missing data
Nominal (ordinal via weighted κ)

Least general

Fleiss' κ

N raters (interchangeable)
No missing data
Nominal only

Middle ground

Krippendorff's α

Any number of raters
Missing data OK
Any scale (nominal, ordinal, interval, ratio)

Most general

Interpretation guidelines

Landis & Koch (1977)

For Cohen's & Fleiss' κ

< 0.00	No agreement
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Krippendorff

For α (stricter)

> 0.80	Reliable
0.67–0.80	Tentative conclusions
< 0.67	Revisit guidelines

Measuring Annotation Quality

❖ Example: Annotation Table

Item	Annotator A	Annotator B	Annotator C
1	+	+	+
2	+	+	-
3	-	-	-
4	+	-	+
5	-	-	
6	-	+	-

Cohen's Kappa

	A	B	C
1	+	+	+
2	+	+	-
3	-	-	-
4	+	-	+
5	-	-	
6	-	+	-

1. Cohen's kappa

2 raters only

Pick one pair: A and B on items 1–6 (ignore C entirely). Build the 2x2 contingency matrix:

		Rater B		
		+	-	
Rater A	+	2	1	3
	-	1	2	3
Col total		3	3	6

Green diagonal = agreement cells. Items 1,2 → both +. Items 3,5 → both -. Items 4,6 → disagree.

Step 1: Observed agreement from diagonal: $P_o = (2 + 2) / 6 = 4/6 = 0.667$

Step 2: Expected by chance — A says + on 3/6 items, B says + on 3/6 items.
 $P(\text{both } +) = 0.5 \times 0.5 = 0.25$, $P(\text{both } -) = 0.5 \times 0.5 = 0.25 \rightarrow P_e = 0.50$

Step 3: $\kappa = (P_o - P_e) / (1 - P_e) = (0.667 - 0.50) / (1 - 0.50) = 0.333$

Limitation: only works for exactly 2 raters. Can't use annotator C. Can't handle missing data.

Fleiss' Kappa: Step 1

2. Fleiss' kappa

N raters, all must rate all items

Use all 3 annotators — but item 5 must be dropped (C didn't rate it). Items 1–4, 6 only.

	A	B	C
1	+	+	+
2	+	+	-
3	-	-	-
4	+	-	+
5	-	-	
6	-	+	-

Step 1: Build the items × categories count matrix (n = 3 raters per item):

Item	c ₊	c ₋	
1	3	0	All agree +
2	2	1	C disagrees
3	0	3	All agree -
4	2	1	B disagrees
6	1	2	B disagrees
Total	8	7	15 ratings total

Fleiss' Kappa: Step 2

	A	B	C
1	+	+	+
2	+	+	-
3	-	-	-
4	+	-	+
5	-	-	
6	-	+	-

Step 2: Per-item agreement: $P_i = (\sum c_k^2 - n) / n(n-1)$, where c_k = count for category k , $n = 3$ raters.

Item	$c_+^2 + c_-^2$	$- n$	$\div n(n-1)$	P_i
1	$9 + 0 = 9$	6	$\div 6$	1.000
2	$4 + 1 = 5$	2	$\div 6$	0.333
3	$0 + 9 = 9$	6	$\div 6$	1.000
4	$4 + 1 = 5$	2	$\div 6$	0.333
6	$1 + 4 = 5$	2	$\div 6$	0.333

Step 3: $P_o = \text{mean of } P_i = (1 + .333 + 1 + .333 + .333) / 5 = 0.60$

Category proportions: $p_+ = 8/15 = 0.533$, $p_- = 7/15 = 0.467$

$P_e = p_+^2 + p_-^2 = 0.284 + 0.218 = 0.502$

Step 4:

$$\kappa = (P_o - P_e) / (1 - P_e) = (0.60 - 0.502) / (1 - 0.502) = 0.197$$

Limitation: all raters must rate all items. Item 5 dropped — data wasted. Nominal scale only.

Cohen's vs. Fleiss' Kappa

- ❖ Cohen's Kappa can be misleading if pairs are cherrypicked!
- ❖ Fleiss' Kappa pools all annotators and is more robust
- ❖ Multi-rater metrics more common in practice

	A	B	C
1	+	+	+
2	+	+	-
3	-	-	-
4	+	-	+
5	-	-	
6	-	+	-

Pair	Items used	P_o	P_e	κ
A vs B	1-6	0.667	0.500	0.333
A vs C	1-4, 6	0.800	0.480	0.615
B vs C	1-4, 6	0.400	0.480	-0.154

Fair

Substantial

Worse than chance

The same data yields κ from -0.15 to 0.62 depending on which pair you pick. Rater B is the outlier: B dissents on items 4 and 6 where A and C agree, making B *worse than random* when paired with C.

Disagreement is Also a Signal

- ❖ Traditional: Disagreement = Error -> Fix it!
- ❖ Modern: Disagreement = Genuine ambiguity in data/semantics?
- ❖ Examples: sarcasm ($\alpha < 0.35$), toxicity, hate speech
- ❖ **Recourse for ambiguous semantics:**
 - ❖ Soft labels: keep probability distribution instead of majority vote (e.g., 60% toxic / 40% not)
 - ❖ Annotator-specific models: learn each rater's calibration separately (e.g., CrowdLayer, learning from crowds)
 - ❖ Report full distributions: don't collapse to a single label; downstream model trains on the disagreement itself

Label Errors are Ubiquitous in AI!

- ❖ Northcutt et al., NeurIPS 2021: labelerrors.com
- ❖ Errors found in *test* sets of 10 popular ML benchmarks!
- ❖ $\geq 3.3\%$ average; ImageNet validation $\geq 6\%$!
- ❖ Model rankings change on corrected test sets

Label Errors in ML Test Sets Find issues in your data Fix issues in your data More audits of famous datasets GitHub About

Dataset: ImageNet Label: All classes with noise

All classes with noise for ImageNet

great white shark tiger shark hammerhead shark
electric ray cock hen house finch jay magpie
American dipper kite bald eagle fire salamander
smooth newt newt American bullfrog tree frog
tailed frog loggerhead sea turtle
leatherback sea turtle mud turtle terrapin
box turtle banded gecko green iguana
Carolina anole desert grassland whiptail lizard
agama alligator lizard Gila monster
European green lizard chameleon Nile crocodile
worm snake ring-necked snake
eastern hog-nosed snake smooth green snake
water snake vine snake night snake
boa constrictor African rock python Indian cobra
green mamba sea snake Saharan horned viper
eastern diamondback rattlesnake harvestman
yellow garden spider barn spider
European garden spider southern black widow
tarantula wolf spider tick black grouse ptarmigan
ruffed grouse peacock partridge lorikeet coucal
bee eater hornbill hummingbird toucan duck
goose tusker platypus wallaby koala wombat
jellyfish sea anemone brain coral flatworm
nematode conch snail slug sea slug chiton
Dungeness crab rock crab fiddler crab
red king crab spiny lobster crayfish white stork
black stork spoonbill little blue heron great egret

ImageNet given label: siamang
Cleanlab guessed: baboon
MTurk consensus: baboon
ID: 00047520

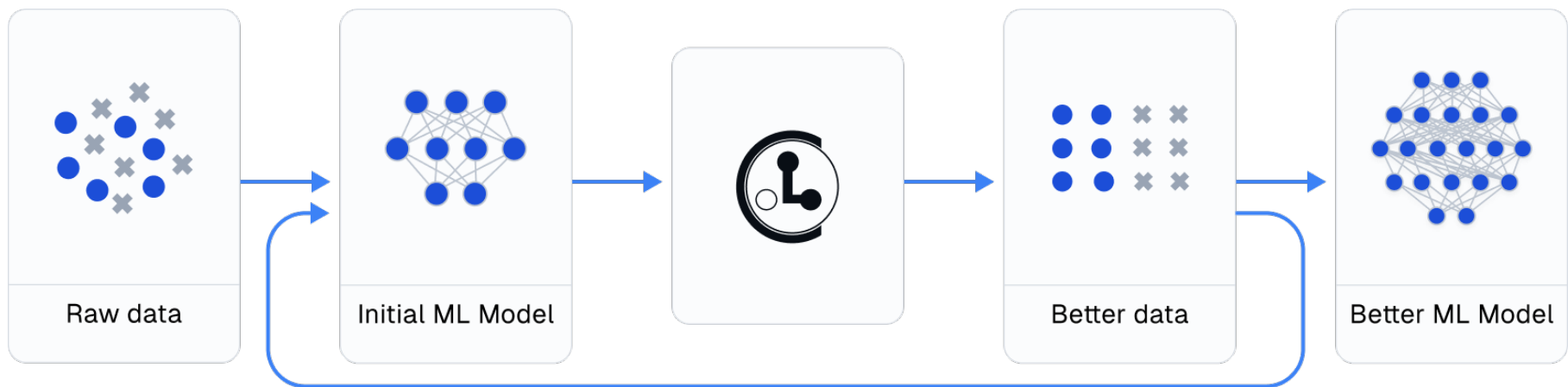
ImageNet given label: red panda
Cleanlab guessed: meerkat
MTurk consensus: meerkat
ID: 00022365

ImageNet given label: Irish Wolfhound
Cleanlab guessed: Irish Water Spaniel
MTurk consensus: Irish Water Spaniel
ID: 00022071

Grappling with Label Errors

❖ Confident Learning:

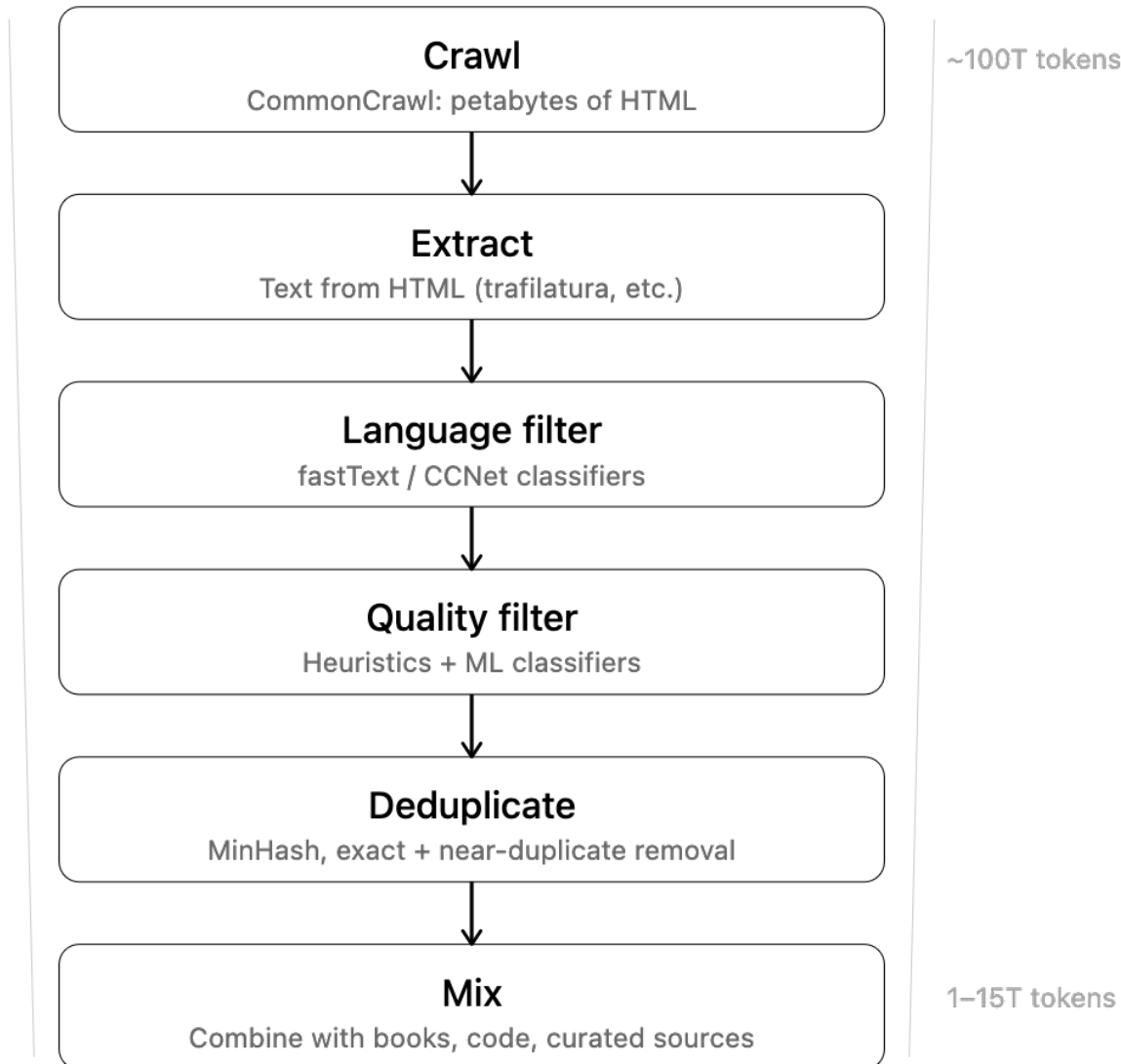
- ❖ Predicted probabilities + Given labels -> estimate joint distribution of noisy vs. true labels -> find errors
- ❖ Model-agnostic; no hyperparameters
- ❖ Open-sourced as cleanlab <https://github.com/cleanlab/cleanlab>



Outline

- ❖ Motivation for Data-Centric AI
- ❖ Creating Data
 - ❖ Human Feedback and Annotation
- ➔ ❖ Data Curation and Structure
 - ❖ Data Synthesis
 - ❖ Multimodal Data
- ❖ Creating Evals

Typical Data Pipeline of Pretraining



What Makes Text “High Quality”?

❖ **Heuristic Rules:**

- ❖ Cheapest, fastest
- ❖ Min length, uppercase ratio, blocklists, perplexity via KenLM trained on Wikipedia
- ❖ Used by C4, Gopher; runs on CPU; billions of docs

❖ **Lightweight Classifiers:**

- ❖ Still fast, but learns patterns beyond hand-coded rules
- ❖ fastText trained on Wikipedia-like vs. random web
- ❖ Used by CCNet, FineWeb

What Makes Text “High Quality”?

❖ LLM-as-Judge:

- ❖ More expensive to bootstrap, but captures subtle quality signals
- ❖ Used by FineWeb-Edu; Llama-3-70B scores 500K pages for educational quality, distill into a small classifier, apply at scale; 15T → 1.3T tokens

❖ Human Annotation:

- ❖ Most expensive and likely slowest
- ❖ RedPajama-V2: 40+ quality signals per document, human-verified
- ❖ Highest fidelity but does not scale to full crawls
- ❖ Used for evaluation/calibration more than bulk filtering

Peer Instruction Activity

(Switch to Google Form)

Data Deduplication

- ❖ Duplicates => memorization, regurgitation, wasted compute!
- ❖ Methods to deduplicate at scale:
 - ❖ Exact hash
 - ❖ URL-based deduplication
 - ❖ Line-level
 - ❖ MinHash LSH (n-gram shingles, Jaccard ~ 0.8)
 - ❖ SemDeDup (embedding-based)
- ❖ Scale: 100B+ docs -> need LSH, Bloom filters
- ❖ SlimPajama: Global dedup. of RedPajama; better benchmarks

Data Mixing

- ❖ Diverse mixes of sources outperform uniform sourcing.
- ❖ LLM pretraining = mix of web, books, code, Wikipedia, papers.
 - ❖ Example: LLaMA: ~67% web, 4.5% Wikipedia, 4.5% books, 6.5% code.
- ❖ How to choose:
 - ❖ Manual tuning
 - ❖ DoReMi (proxy model)
 - ❖ Ablation

Open Datasets for Pretraining

- ❖ Popular large datasets:
 - ❖ C4 (2019, 175B)
 - ❖ The Pile (2020, 825GB)
 - ❖ RedPajama-V2 (2023, 30T)
 - ❖ RefinedWeb (2023)
 - ❖ FineWeb/FineWeb-Edu (2024, 15T/1.3T)
- ❖ Trend: more openness, better tooling, principled curation
- ❖ Document your data: Gebru et al., “Datasheets for Datasets”

Key Takeaways

1. AI in practice has shifted from model-centric to data-centric. Creating/curating data is now a first-class citizen.
2. Human feedback (demonstrations, comparisons, corrections, etc.) are a critical signal.
3. Label errors are ubiquitous in AI. Annotation quality is measurable (IAA).
4. Typical curation at scale = Filtering + Dedup. + Mixing.
5. Data quality > quantity.

Recommended Readings

- ❖ Andrew Ng, “MLOps: Model-Centric to Data-Centric AI” 2021
- ❖ Ouyang et al., InstructGPT, NeurIPS 2022
- ❖ Rafailov et al., DPO, NeurIPS 2023
- ❖ Northcutt et al., Confident Learning, JAIR 2021 + cleanlab.ai

Review Questions

1. Define data-centric AI. How does it differ from model-centric AI? Are there other prominent forms of “centricity” for AI?
2. What are the 3 main types of human feedback signals used in LLM post-training? Which type of training stage does each primarily feed?
3. Briefly explain one statistical difference between Cohen’s kappa and Fleiss’ kappa. Why is Krippendorff’s alpha preferable in practice?
4. Does a Fleiss’ kappa score of 1.0 mean the labels are perfect? Why or why not? Explain your rationale with an example.
5. InstructGPT at 1.3B was preferred over GPT-3 at 175B. What does this imply about where AI practitioners should invest effort?
6. When should annotator disagreement be treated as a signal rather than noise? What are two concrete strategies for preserving that signal?
7. When is LLM-as-Judge useful for quality filtering? When is it not? Why?
8. Why is mixing data sources useful for LLM pretraining?

Outline

- ❖ Motivation for Data-Centric AI
- ❖ Creating Data
 - ❖ Human Feedback and Annotation
 - ❖ Data Curation and Structure
-  ❖ Data Synthesis
 - ❖ Multimodal Data
- ❖ Creating Evals