Lecture 13: Data 1

Rayane Laabid & Weihua Shi

February 2026

1 Overview

Modern large language models are trained through multiple stages, each designed to serve a distinct role in the overall learning process. These stages are separated because different objectives, data regimes, and optimization signals are most effective at different points in training. Together, they transform a randomly initialized model into a capable, specialized, and human-aligned language assistant.

Pretraining Pretraining is the foundational stage in which the model acquires general linguistic competence and broad world knowledge. This stage is typically driven by self-supervised objectives such as next-token prediction and relies on large-scale, diverse, and mostly unlabeled text corpora. The outcome of pretraining is a general-purpose foundation model with strong representational capacity, but limited task awareness and alignment with human intent.

Mid-training Mid-training adapts and enhances the pretrained model by focusing on specific domains, skills, or reasoning capabilities. This stage often involves continued pretraining or task-oriented objectives on curated datasets, such as mathematics, code, scientific text, or multilingual corpora. The goal of mid-training is not alignment, but capability sharpening—improving performance and generalization in targeted areas.

Post-training Post-training aligns the model with human preferences, usability requirements, and safety constraints. Unlike earlier stages, it relies heavily on human-labeled data and explicit feedback signals, using techniques such as supervised fine-tuning and reinforcement learning from human or AI feedback. The result is a model that behaves helpfully, safely, and reliably in real-world interactions.

Summary of Training Stages

Stage	Main Purpose	Typical Data	Outcome
Pretraining	Learn general language and world knowledge	Large-scale, diverse, mostly unlabeled text (web, books, code, papers)	General-purpose foundation model
Mid-training	Specialize and enhance capabilities	Curated domain or task-specific datasets (math, code, science, multilingual data)	More capable and specialized model
Post-training	Align with human preferences and safety	Human-labeled data (instructions, rankings, feedback)	Usable, aligned assistant

In summary, pretraining builds general intelligence, mid-training sharpens specific capabilities, and post-training ensures alignment and usability for human interaction.

From Raw Web Data to Curated Corpora The effectiveness of pretraining critically depends on data quality. Early large-scale models demonstrated that raw web data alone is insufficient: without careful processing, it contains excessive noise, redundancy, and non-linguistic content. This insight motivated the development of increasingly sophisticated data construction pipelines.

A key historical milestone is the T5 work, which popularized the text-to-text paradigm. However, a major technical contribution of this line of research was the introduction of the C4 (Colossal Clean Crawled Corpus) dataset. C4 demonstrated that large-scale web data can only become useful for language modeling through aggressive, explicit, and value-laden filtering heuristics. This marked a shift from treating data as an afterthought to viewing data curation as a first-order design decision.

DataComp for Language Models (DCLM) DataComp for Language Models (DCLM) provides a standardized testbed for controlled experiments on pretraining data. It consists of a 240T-token corpus derived from Common Crawl, paired with reference pretraining recipes and a suite of 53 downstream evaluations. By fixing the evaluation protocol, DCLM enables systematic study of how data quality, scale, and filtering strategies affect model performance.

Extensive experiments within the DCLM framework show that **model-based filtering** is essential for assembling high-quality training datasets, outperforming purely heuristic or rule-based approaches.

DCLM-BASELINE DCLM-BASELINE is a high-quality subset of the DCLM corpus designed to emphasize efficiency. It enables a 7B-parameter model to achieve 64% 5-shot accuracy on MMLU using only 2.6T tokens, matching the performance of models such as Llama 3 8B while requiring significantly less compute. Compared to earlier benchmarks such as MAP-Neo, DCLM-BASELINE achieves state-of-the-art results for open-data models with approximately 40% lower computational cost.

From Ad-hoc Corpora to Designed Data Mixtures This evolution in data practices reflects a broader transition in language model training. Early models such as GPT-3 relied on large but relatively ad-hoc mixtures of web and curated datasets. In contrast, later efforts such as The Pile emphasized transparency, explicit data composition, and controllable weighting across sources. These works established that data composition, filtering, and deduplication are not secondary concerns, but central design choices that shape model behavior.

Building on this insight, RefinedWeb and FineWeb argue that sufficiently careful text extraction, rule-based filtering, and large-scale deduplication can enable web-only corpora to support strong language models without relying on proprietary or heavily curated sources.

DataComp-LM further reframes large-scale web corpora as the output of modular data processing pipelines, allowing researchers to isolate and study the contributions of heuristic cleaning, deduplication, and model-based filtering to downstream performance. Nemotron-CC extends this paradigm by incorporating ensemble quality scoring, synthetic data transformation, and legal constraints, highlighting that modern data pipelines optimize not only for scale and quality, but also for compliance and usability.

From Pretraining to Instruction-Following Models While pretraining and mid-training primarily focus on learning representations and capabilities, instruction-following and alignment emerge primarily during post-training. FLAN-style instruction tuning teaches the model what tasks are and how to perform them, whereas post-training instruction and preference optimization teach the model how to behave as a helpful, safe, and human-aligned assistant.

Earlier BERT-style datasets such as BooksCorpus are unsuitable for modern chat-oriented models not only due to legal and ethical concerns, but also because they lack conversational structure and behavioral supervision. Similarly, although Wikipedia is legally clean, high-quality, and well-structured, its openness and periodic snapshotting expose it to data poisoning risks, underscoring that no data source is entirely risk-free.

Finally, systems such as Alpaca and Vicuna illustrate complementary data construction paradigms that progressively transform a pretrained language model into an instruction-following and chat-capable assistant through synthetic instruction generation and supervised fine-tuning.

2 Abbreviations

Abbreviation	Full Name	Explanation (in the context of LLM data)
API	Application Programming Interface	A programmatic interface that allows automated, structured access to data from a service (e.g., Reddit API), as opposed to manual browsing.
Crawling	Web Crawling	Automated downloading of web pages by simulating a browser; often produces raw HTML and is used for large-scale data collection such as Common Crawl.
Dump	Data Dump	An official, static snapshot of a dataset released by a platform (e.g., Wikipedia dumps), useful for reproducibility and academic research.
HTML	HyperText Markup Language	The raw markup language of web pages; commonly produced by crawling and requires significant processing before training.
JSON	JavaScript Object Notation	A structured data format commonly returned by APIs, easier to parse and process than raw HTML.
CC	Common Crawl	A public corpus created by large-scale web crawling, frequently used as a source for LLM pretraining data.

Data Processing and Dataset Construction

Term	Meaning	Explanation
Raw data	Raw snapshot	Unprocessed text collected directly from the web or APIs; typically noisy and unsuitable for direct training.
Processed data	Processed text	Text that has undergone cleaning, filtering, deduplication, and normalization, making it suitable for training.
Curated	Curated dataset	A dataset that has been intentionally filtered and selected using heuristics or quality models to improve signal-to-noise ratio.
Aggregated	Aggregated dataset	A dataset formed by combining multiple sources with controlled mixing ratios (e.g., The Pile, Dolma).
Dedup	Deduplication	The process of removing duplicate or near-duplicate text to prevent memorization and data leakage.

Training Stages and Optimization Methods

Abbreviation	Full Name	Explanation
PT	Pretraining	Large-scale self-supervised training to learn general language structure and world knowledge.
MT	Mid-training	Additional training focused on improving specific capabilities (e.g., math, reasoning, domain knowledge).
FT	Fine-tuning	A general term for additional training after pretraining; often refers to post-training adjustments.
SFT	Supervised Fine-Tuning	Training on instruction–answer pairs to teach the model how to respond correctly and follow instructions.
DPO	Direct Preference Optimization	A preference-based optimization method that teaches the model which of multiple answers is preferred by humans.
RLHF	Reinforcement Learning from Human Feedback	A reinforcement learning approach using human preference data to align model behavior.
RLAIF	Reinforcement Learning from AI Feedback	A variant of RLHF where preferences are generated by a stronger model instead of humans.

Instruction, Reasoning, and Capability Terms

Abbreviation	Full Name	Explanation
IF	Instruction Following	The ability of a model to understand and execute tasks specified in natural language instructions.
FLAN	Fine-tuned Language Net	A multitask instruction-format dataset designed to improve task generalization rather than assistant behavior.
CoT	Chain of Thought	Explicit intermediate reasoning steps included in training data to improve multi-step reasoning.
Infilling	Infilling	A training objective where the model fills in missing text in the middle of a sequence rather than appending at the end.
Long context	Long-context modeling	The ability of a model to process very long input sequences, enabled by both architecture and data.
Tokens	Tokens	The basic units of text processed by the model; not equivalent to words.

Synthetic Data and Distillation

Term	Meaning	Explanation
Synthetic data	Synthetic data	Data generated by a model rather than humans, often used to scale training or control difficulty.
Distillation	Knowledge Distillation	Training a smaller or weaker model using outputs generated by a stronger teacher model.
Self-distillation	Self-distillation	A process in which a model generates data to further train itself, often to stabilize behavior or reasoning patterns.
Teacher model	Teacher model	A stronger model used to generate synthetic or preference data (e.g., GPT-4).
Student model	Student model	The model being trained using real, synthetic, or distilled data.