CS336: Language Modeling from Scratch

Lecture 13: Data 1

Stanford University - Spring 2025

Abstract

Abstract: This lecture provides a comprehensive examination of data in language model training, arguing that data is the most important factor in getting language models right. We explore the evolution of training datasets from BERT's books and Wikipedia to modern multi-trillion token collections, examining data sources, processing techniques, filtering strategies, and copyright considerations. The lecture covers pre-training, mid-training, and post-training data paradigms, with detailed analysis of landmark datasets including Common Crawl, The Pile, C4, and recent advances in model-based filtering approaches.

Enhanced Summary by:
GitHub: HtmMhmd | LinkedIn: Hatem Mohamed

Contents

Contents

1 The Primacy of Data

1.1 Data as the Most Important Factor

Core Thesis: Data is the most important thing in getting language models right, surpassing even scaling laws in practical impact.

Evidence - Corporate Disclosure Patterns:

- Open-weight models (Llama 3, DeepSeek) fully disclose architecture details
- Papers extensively discuss training procedures and optimizations
- Data details remain deliberately vague and undisclosed
- Llama 3 paper: "We create our data set from a variety of data sources containing knowledge until the end of 2023"

Reasons for Data Secrecy:

- **Competitive dynamics:** Data curation represents significant competitive advantage
- **Legal liability:** Avoiding copyright infringement lawsuits
- **Trade secrets:** Processing techniques as intellectual property

Data Insight

Unlike architecture (single small team defines it), data work is highly parallelizable and scalable. Companies can easily hire hundreds of people working on different data aspects: multilinguality, code, images, domain-specific content. This scalability makes data curation a primary differentiator in model development.

1.2 Data Work Continuity

Historical Context: Before foundation models, data importance was clearly recognized because supervised learning required annotation.

Modern Reality: Even with reduced annotation requirements, data work involves extensive curation and cleaning - "the long tail of problems."

Scalability Advantage: Data processing can leverage large teams working on specialized domains, unlike architecture design which requires small, focused teams.

2 Training Stages and Data Types

2.1 Three-Stage Training Paradigm

Pre-training:

- Train on large amounts of raw data, usually from the web
- Focus on broad knowledge acquisition and language understanding
- Typical scale: Trillions of tokens

- Low-quality, high-volume approach

Mid-training:

- Curated smaller set of high-quality documents
- Target particular capabilities: math, code, long context
- Bridge between pre-training and post-training
- Typical scale: Billions to hundreds of billions of tokens

Post-training:

- Fine-tune on instruction following data
- Chat data and reinforcement learning
- Safety and alignment considerations
- Typical scale: Millions to billions of tokens

Boundary Blurring: In practice, lines between stages are blurry, with recent models having more stages and hybrid approaches.

2.2 Quality-Volume Trade-off

General Pattern: Start with large amounts of low-quality data, progressively train on smaller amounts of higher-quality data.

Terminology:

- **Base model:** Checkpoint after pre-training and mid-training
- **Instruct model:** After post-training and alignment

3 Historical Evolution of Datasets

3.1 BERT Era (2018): Books and Wikipedia

BooksCorpus:

- Source: Smashwords self-publishing platform (established 2008)
- Collection method: Scraped 7,000 books priced at zero
- Legal status: Violated terms of service, later taken down
- Historical significance: Demonstrated importance of book data
- 2015 context: "Wild West" era, AI copyright not prominent concern

Wikipedia Characteristics:

- **Content policy:** No original thought, everything from citations
- **Notability requirement:** Multiple sources must have covered topic

- **Editorial bias:** Small number of people contribute majority of content
- **Coverage limitations:** No recipes, limited opinion content, tail topics underrepresented
- **Quality reputation:** Generally reliable but subject to manipulation

Data Poisoning Vulnerability: Carlini's work demonstrated that malicious edits can be injected before periodic dumps, reaching training data before rollback policies take effect.

3.2 GPT-2 Era (2019): WebText

WebText Innovation:

- **Insight:** Use Reddit karma points as quality signal
- **Method:** Extract links from Reddit posts with >3 karma points
- **Scale:** 1 million pages, 40 gigabytes of text
- **Diversity:** Web-wide coverage filtered through social curation

Open Replication: OpenWebText created as public reproduction, widely used in research.

4 Common Crawl Foundation

4.1 Common Crawl Infrastructure

Establishment and Scale:

- Founded 2007, monthly web crawls for 17 years
- 100 different crawls with varying coverage
- Recent crawl: 2.7 billion pages added monthly
- Cost-effective: Less than two weeks on AWS machines

Crawling Methodology:

- **Seed URLs:** Hundreds of millions of starting points
- **Frontier approach:** BFS traversal of web with queue management
- **Politeness:** Respect robots.txt, avoid server overload
- **Dynamic handling:** URL normalization, duplicate detection

Output Formats:

- **WARC files:** Raw HTTP responses (HTML)
- **WET files:** Converted text (lossy HTML-to-text process)
- **Processing choice:** Raw WARC allows custom text extraction

4.2 HTML-to-Text Conversion Impact

Significant Performance Difference: DataComp-LM showed 4-point accuracy difference between raw Common Crawl WET files and properly processed text using Trafilatura.

Tool Comparison: Different HTML extraction tools produce materially different training outcomes.

4.3 Coverage and Limitations

Deliberate Incompleteness: Common Crawl is not comprehensive by design - prioritizes being "gentle and polite" over complete coverage.

Sparsity Example: Not all Wikipedia articles appear in Common Crawl despite being publicly accessible.

Robot.txt Compliance: Many frontier model providers now run their own crawlers because Common Crawl coverage is insufficient for their needs.

5 Early Filtering Approaches

5.1 CCNet (Meta)

Multilingual Focus: Generic procedure for extracting high-quality subsets from Common Crawl with emphasis on multiple languages.

Processing Pipeline:

1. Deduplication removal
2. Language identification using linear classifiers
3. Quality filtering based on 5-gram model similarity to Wikipedia
4. Wikipedia as high-quality surrogate

Limitations: Only captures content similar to Wikipedia, missing valuable non-Wikipedia-like content.

5.2 C4 (Google)

Colossal Clean Crawled Corpus: Introduced alongside T5 model, represents rule-based filtering approach.

Heuristic-Based Filtering:

- Keep lines ending in punctuation
- Remove pages with fewer than three sentences
- Remove content containing "bad words"
- Remove code (brace filtering)
- English-only filtering

Scale Reduction: 1.4 trillion tokens $\rightarrow$ filtered subset, demonstrating massive quality vs. quantity trade-offs.

Complementary Nature to CCNet:

- **Advantage:** Captures well-formed sentences that don't resemble Wikipedia
- **Disadvantage:** Includes spammy but grammatically correct content

6 GPT-3 Era Innovations

6.1 GPT-3 Dataset Composition

Multi-Source Approach:

- Common Crawl (processed)
- WebText2 (expanded Reddit-based filtering)
- Books1 and Books2 (mysterious book corpora)
- Wikipedia
- Total: 400 billion tokens (small by modern standards)

Quality Classification Innovation: Trained classifier to distinguish high-quality sources (WebText, Wikipedia, books) from general Common Crawl content.

6.2 The Pile (EleutherAI)

Community-Driven Curation: Decentralized Discord-based volunteer effort reacting to GPT-3's closed nature.

22 High-Quality Domains:

- Common Crawl and OpenWebText
- Stack Exchange (programming Q&A)
- Wikipedia and arXiv
- PubMed Central (NIH-mandated open access papers)
- Enron emails (only available corporate email dataset)
- Project Gutenberg (copyright-cleared books)
- Books3 (shadow library content, later removed)

Technical Improvements: Used WARC files over WET files and specialized text extraction tools.

Data Evolution

The Pile represented a watershed moment in open-source data curation, demonstrating that community efforts could create datasets larger and more diverse than proprietary alternatives. Its 22-domain structure influenced subsequent dataset design patterns.

7 Specialized Data Sources

7.1 Code: GitHub and The Stack

GitHub Characteristics:

- 28 million public repositories
- Quality varies dramatically (random repos often disappointing)
- Includes non-code content (documentation, issues, commit history)
- Extensive deduplication required

The Stack Processing:

- GitHub Archive provides event snapshots via Google BigQuery
- Git cloned 137 million repositories
- Filtered for permissive licenses only
- Deduplication and cleaning
- Result: 3.1TB of processed code

Code Benefits Beyond Programming: Generally believed helpful for reasoning and other capabilities, though rigorous evidence limited.

7.2 Academic Content

arXiv: AI research benefits from preprint culture, unlike many other fields requiring journal publication.

PubMed Central: NIH mandate creates substantial open-access corpus of medical/biological research.

Semantic Scholar: AI2's academic search engine provides processed scholarly content at scale (40 million papers in recent datasets).

7.3 Q&A and Community Content

Stack Exchange Benefits:

- Question-answer format resembles instruction following
- Reputation and voting systems enable quality filtering
- Metadata (comments, votes) supports advanced filtering
- Blurs boundary between pre-training and post-training data

Reddit Data:

- Historical access through academic projects
- 2023 policy changes restricted access
- Submissions and comments provide conversational data

8 Modern Filtering Revolution

8.1 Transition to Model-Based Filtering

Historical Resistance: Early 2020s preference for manual rules to avoid model bias and protect marginalized content.

Paradigm Shift: Recognition that model-based filtering dramatically outperforms rule-based approaches on benchmarks.

Current Consensus: Model-based filtering now standard despite potential biases.

8.2 DCLM (DataComp for Language Models)

Competition Framework: Standardized infrastructure for comparing data processing algorithms.

Scale: DCLM-pool contains 240 trillion tokens from processed Common Crawl dumps.

DCLM-Baseline Creation:

- Aggressive filtering: 240T → 3.8T tokens (1.4
- FastText classifier trained on positive/negative examples
- Positive examples: OpenHermes (GPT-4 instruction data) + ELI5 (Reddit explanations)
- Negative examples: Random RefinedWeb samples

Performance Impact: 3

8.3 Nemotron-CC (NVIDIA)

Scale Motivation: DCLM-baseline's 3.8T tokens insufficient for large model training runs.

Token Preservation Focus: Optimize for retaining tokens while maintaining quality.

Advanced Techniques:

- **HTML Extraction:** jusText over Trafilatura for higher token retention
- **Educational Value:** Large model scores documents, distilled to efficient classifier
- **Ensemble Filtering:** Multiple classifiers with bucketed sampling
- **Content Rewriting:** Low-quality content rewritten by language models
- **Task Generation:** High-quality content augmented with synthetic tasks

Results: 6.3T tokens (nearly double DCLM-baseline) with superior benchmark performance.

9 Copyright and Legal Considerations

9.1 Copyright Fundamentals

Legal Foundation: US Copyright Act (1976) covers "original works of authorship fixed in tangible medium."

Key Principles:

- **Automatic protection:** No registration required for copyright

- **Expression vs. ideas:** Protects specific expression, not underlying concepts
- **Duration:** 75 years from publication
- **Scope expansion:** Copyright coverage has increased over time

Internet Reality: Most web content is copyrighted, making licensing or fair use appeal necessary.

9.2 Licensing Approaches

Commercial Licensing: Direct contracts with content creators (Google-Reddit, OpenAI-Stack Exchange).

Creative Commons: Bridges gap between public domain and full copyright protection.

Scale Challenge: Impossible to license "the internet" - random websites lack clear licensing mechanisms.

9.3 Fair Use Analysis

Four Factor Test:

1. **Purpose and character:** Educational vs. commercial use, transformative nature
2. **Nature of work:** Factual content more likely fair use than fictional
3. **Amount used:** Snippets vs. complete works (problematic for language models)
4. **Market effect:** Whether use displaces original creator's market

Training-Specific Issues:

- Initial copying for training technically violates copyright
- Training arguably transformative (extracting patterns, not copying content)
- Models can memorize and reproduce training content
- Economic impact on content creators remains contentious

Copyright Challenge

Copyright complexity extends beyond verbatim memorization to include plots, characters, and semantic content. Even with minimal n-gram overlap, models reproducing copyrighted characters or storylines could face infringement claims. This makes legal compliance far more complex than simple deduplication.

10 Post-Training Data Evolution

10.1 Task-Based Instruction Data

SuperNaturalInstructions: Community effort creating 1,600+ standardized tasks in instruction format.

FLAN (2022): Converted traditional NLP benchmarks into instruction-following format.

Template Problem: Early instruction datasets suffered from artificial, repetitive prompt structures.

10.2 Synthetic Data Generation

Self-Instruct (Alpaca): Language models generate their own training examples.

Conversational Data:

- Vicuna used ShareGPT conversations (now deprecated)
- Models chatting with themselves
- Evol-instruct for increasing complexity

Web-to-QA Conversion: Extract question-answer pairs from quiz sites and educational content.

OpenHermes: Aggregation of multiple synthetic datasets, used in DCLM quality filtering.

10.3 Human-Annotated Instruction Data

Llama 2 Chat: Professional annotators created high-quality instruction data.

Quality vs. Quantity: Claims that small amounts of high-quality human annotation outperform large synthetic datasets.

Cost Optimization: Potential for even smaller annotation budgets combined with reinforcement learning.

11 Long Context and Specialized Capabilities

11.1 Long Context Extension

Computational Reality: Quadratic scaling of transformers requires staged approach to context length.

Mid-Training Integration: Add long context capabilities after basic model competence established.

Data Sources:

- **Books:** Natural long-range dependencies
- **Mathematical content:** Extended reasoning chains
- **Synthetic generation:** Artificially created long-dependency tasks

11.2 Capability-Specific Data

Mathematical Reasoning: Specialized datasets for improving mathematical problem-solving.

Code Generation: Programming-specific datasets beyond general code training.

Multimodal Extensions: Integration of image, video, and other modalities requires specialized processing.

12 Recent Developments

12.1 Data Access Restrictions

Platform Policy Changes (2023): Reddit, Stack Exchange, and other platforms restricted data access as AI training value became apparent.

Commercial Licensing Era: Transition from free academic access to paid commercial licensing.

Competitive Advantage: Platforms leveraging data ownership for AI development partnerships.

12.2 Llama-Nemotron Post-Training

Multi-Source Approach: Combination of public datasets and synthetically generated content.

Model Diversity: Generated data from multiple language models rather than single source.

Data Release: Actual dataset available for research and analysis.

13 Data Processing Best Practices

13.1 HTML-to-Text Conversion

Tool Selection Impact: Choice of extraction tool significantly affects training performance.

Information Preservation: Balance between content retention and noise reduction.

Structure Maintenance: Preserving document structure while removing formatting artifacts.

13.2 Deduplication Strategies

Multiple Levels:

- Exact duplicate removal
- Near-duplicate detection
- Cross-source deduplication
- Temporal duplicate handling

Quality vs. Quantity: More aggressive deduplication improves quality but reduces training data volume.

13.3 Quality Filtering Evolution

Rule-Based Era: Manual heuristics based on linguistic features.

Model-Based Era: Classifiers trained on positive/negative examples.

Hybrid Approaches: Combining multiple filtering techniques with ensemble methods.

14 Future Directions and Challenges

14.1 Scaling Data Processing

Token Requirements: Modern models require tens of trillions of training tokens.

Quality-Quantity Balance: Tension between aggressive filtering and maintaining sufficient scale.

Multimodal Integration: Extending processing pipelines to handle diverse data types.

14.2 Legal and Ethical Considerations

Copyright Evolution: Legal frameworks adapting to AI training use cases.

Content Creator Rights: Balancing AI development with creator compensation.

Open Source Challenges: Tensions between transparency and legal compliance.

14.3 Data Quality Research

Evaluation Metrics: Beyond benchmark scores to capture data quality dimensions.

Domain-Specific Optimization: Tailoring data processing for specific model capabilities.

Synthetic Data Quality: Improving generated content to match or exceed human-created data.

15 Summary and Key Takeaways

Data Insight

Core Insights for Data Strategy:

1. Data curation is the most parallelizable and scalable aspect of language model development
2. Model-based filtering has largely replaced rule-based approaches for superior performance
3. The boundary between pre-training and post-training data continues to blur
4. Copyright compliance requires sophisticated legal and technical strategies
5. Quality-quantity trade-offs remain fundamental to data strategy decisions

Historical Evolution Pattern: From small, curated datasets (books + Wikipedia) to massive web-scale collections (trillions of tokens) with increasingly sophisticated filtering techniques.

Technical Learnings:

- HTML-to-text conversion choices significantly impact model performance
- Aggressive filtering (99+ removal) can improve results despite reduced scale
- Social signals (Reddit karma, Stack Exchange votes) provide effective quality indicators
- Multi-source ensemble approaches outperform single-source strategies

Legal Reality: Most training data involves copyrighted material, requiring either licensing agreements or fair use arguments. The legal landscape continues evolving as AI training use cases mature.

Future Trajectory: Data work will likely become even more important as model capabilities plateau and data quality becomes the primary differentiator. Organizations with superior data processing capabilities and unique data access will maintain competitive advantages.

16 Data Evolution Mindmap

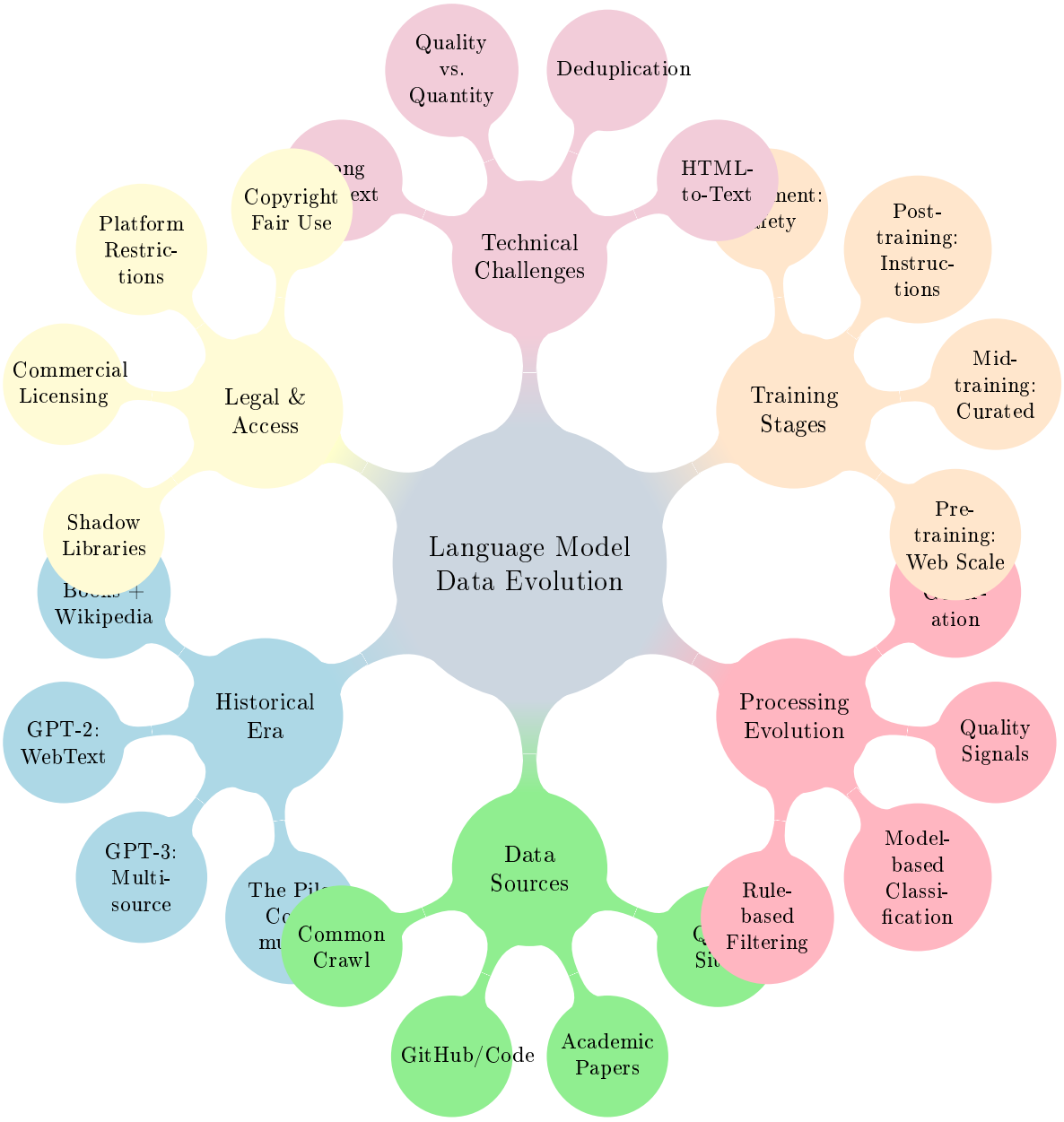

Figure 1: Comprehensive mindmap of language model data evolution, illustrating the progression from simple curated datasets to complex multi-stage processing pipelines with sophisticated filtering and legal considerations.

16.1 Mindmap Interpretation

This mindmap captures the complex evolution of language model training data across six interconnected dimensions:

Historical Era (Blue): The progression from BERT's simple books and Wikipedia foundation through GPT-2's WebText innovation to modern multi-source approaches, culminating in

community-driven efforts like The Pile that democratized high-quality dataset creation.

Data Sources (Green): The core repositories that power modern language models, from the foundational Common Crawl web archive through specialized sources like GitHub code repositories, academic paper collections, and community Q&A platforms that provide diverse, high-quality content.

Processing Evolution (Red): The methodological transformation from simple rule-based filtering through sophisticated model-based classification systems to modern approaches incorporating quality signals and synthetic data generation for optimal training corpus creation.

Training Stages (Orange): The multi-phase training paradigm that begins with massive web-scale pre-training, progresses through curated mid-training for specific capabilities, advances to instruction-following post-training, and concludes with alignment for safety and usability.

Technical Challenges (Purple): The persistent engineering problems that significantly impact model performance: HTML-to-text conversion quality, comprehensive deduplication across sources, quality-quantity optimization trade-offs, and long context extension requirements.

Legal & Access (Yellow): The complex legal and commercial landscape encompassing copyright fair use arguments, platform access restrictions, evolving commercial licensing models, and the controversial use of shadow library content that shapes data availability and compliance strategies.

The mindmap reveals how data work has evolved from simple curation to sophisticated engineering requiring legal expertise, technical innovation, and strategic resource allocation. The interconnections show how advances in one area (e.g., model-based filtering) enable progress in others (e.g., quality-quantity optimization), while challenges in one dimension (e.g., copyright restrictions) drive innovation in others (e.g., synthetic data generation).