Lecture 4.3 - Training Data for Larger LLMs

Generative AI Teaching Kit

This lecture

- Importance of Data Quality
- Dataset Curation
- Challenges in Data collection
- Data Augmentation
- Open vs. Proprietary Datasets

Importance of Data Quality

Data Quality – Diversity and Quality

When training large language models (LLMs) on millions, billions, or even trillions of tokens, the quality of data plays a crucial role in achieving high performance. A well-curated dataset directly impacts the model's accuracy, generalization, and fairness.

To build a powerful LLM, the dataset must balance **Diversity**, **Quality**, and **Quantity**:

1. Diversity: Preventing Bias & Enhancing Robustness

A diverse dataset reduces biases and ensures broad real-world applicability. It should include:

- Different domains (e.g., science, literature, code, news)
- Various languages and dialects
- Multiple writing styles (formal, informal, academic, conversational)

2. Quality: Minimizing Errors & Ensuring Reliability

High-quality data improves the model's ability to generate accurate and coherent responses. Key quality factors include:

- Clean and verified text (no spam, hallucinations, or misinformation)
- Correct grammar and semantics
- Factually reliable sources

3. Quantity: Enabling Generalization & Scalability

- A large volume of training data ensures the model generalizes well across topics.
- However, quantity alone is insufficient—low-quality or redundant data can degrade performance rather than improve it.

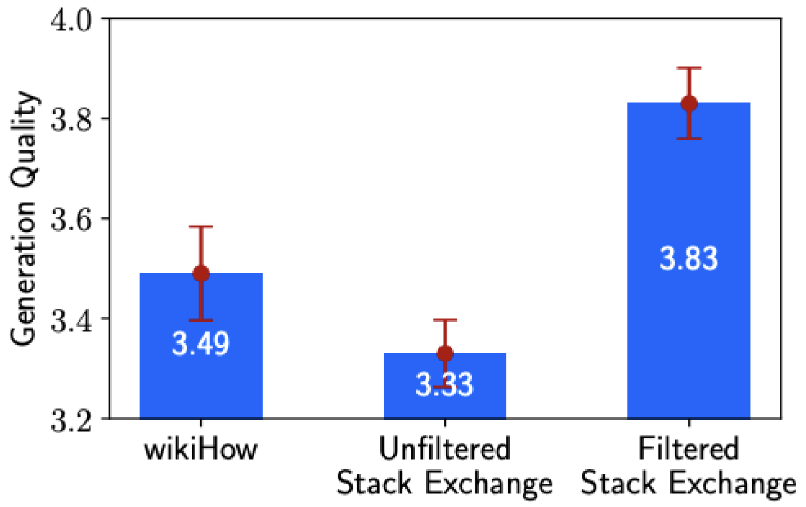


Figure 5: Performance of 7B models trained with 2,000 examples from different sources. **Filtered Stack Exchange** contains diverse prompts and high quality responses; **Unfiltered Stack Exchange** is diverse, but does not have any quality filters; **wikiHow** has high quality responses, but all of its prompts are "how to" questions.

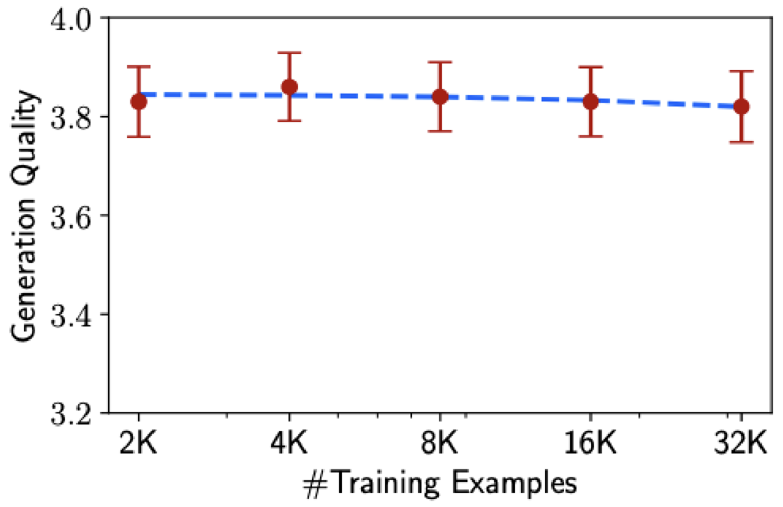


Figure 6: Performance of 7B models trained with exponentially increasing amounts of data, sampled from (quality-filtered) Stack Exchange. Despite an up to 16-fold increase in data size, performance as measured by ChatGPT plateaus.

Data Quality – **Diversity** and Quality

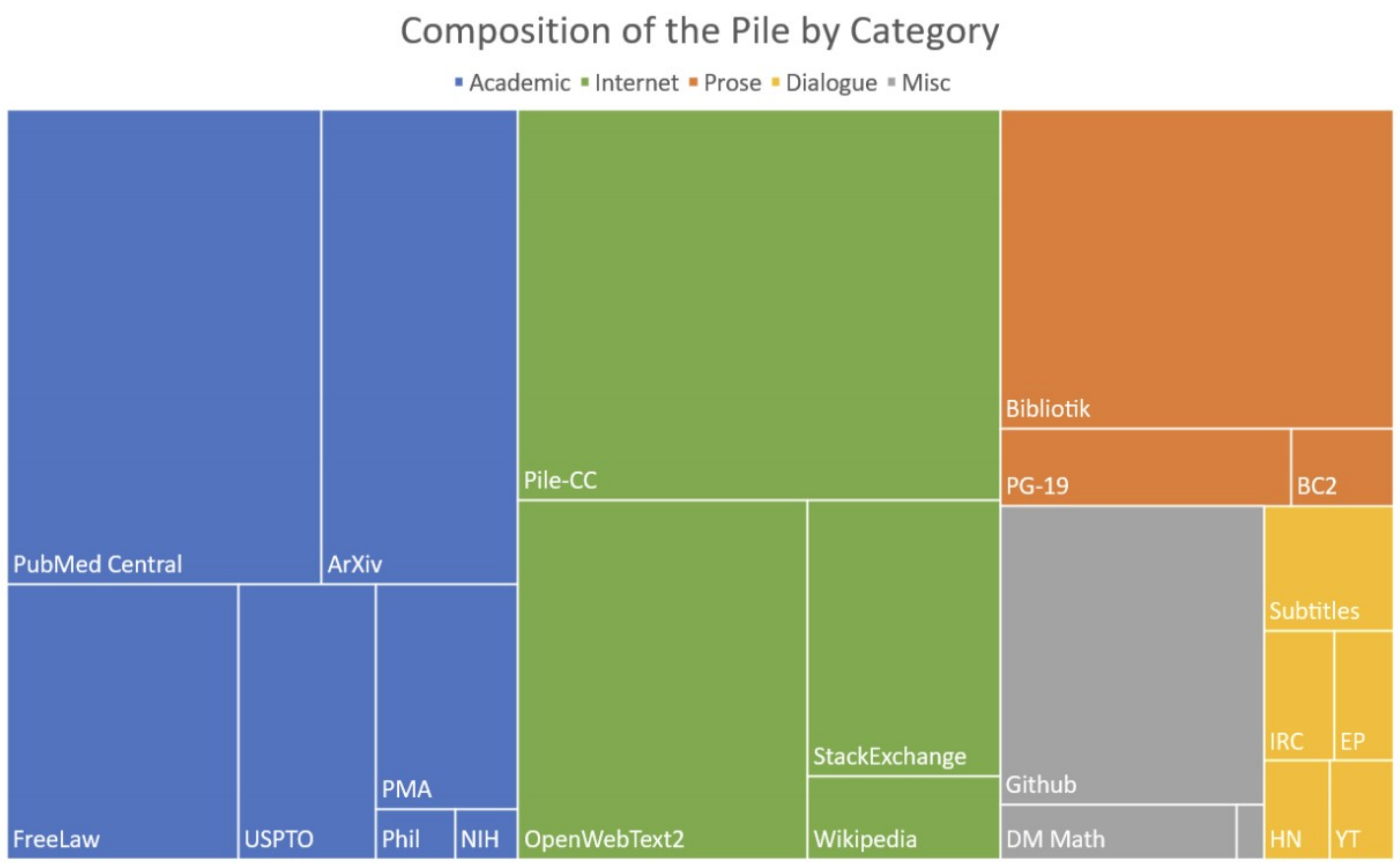
Diversity in LLM Training Data: *Ensuring Representation Across Languages, Dialects, Styles, and Domains*

Why Diversity Matters?
A diverse dataset is essential for training an LLM that can:
- Understand different cultures, linguistic nuances, and writing styles
- Reduce bias and prevent overfitting to dominant languages or topics
- Improve generalization across a wide range of applications

Key Aspects of Diversity
1. Languages & Dialects
- LLMs must learn from multiple languages to avoid being skewed toward high-resource languages like English.
- Dialectal variation (e.g., American vs. British English, Latin American vs. European Spanish) ensures natural communication in different regions.

2. Writing Styles
- Text should cover formal, informal, conversational, poetic, and technical styles to adapt to different use cases.

3. Domains & Knowledge Areas
- A balanced dataset includes news, scientific articles, literature, code, medical texts, legal documents, social media, and more.
- This ensures the model can perform well in general knowledge tasks and specialized applications.

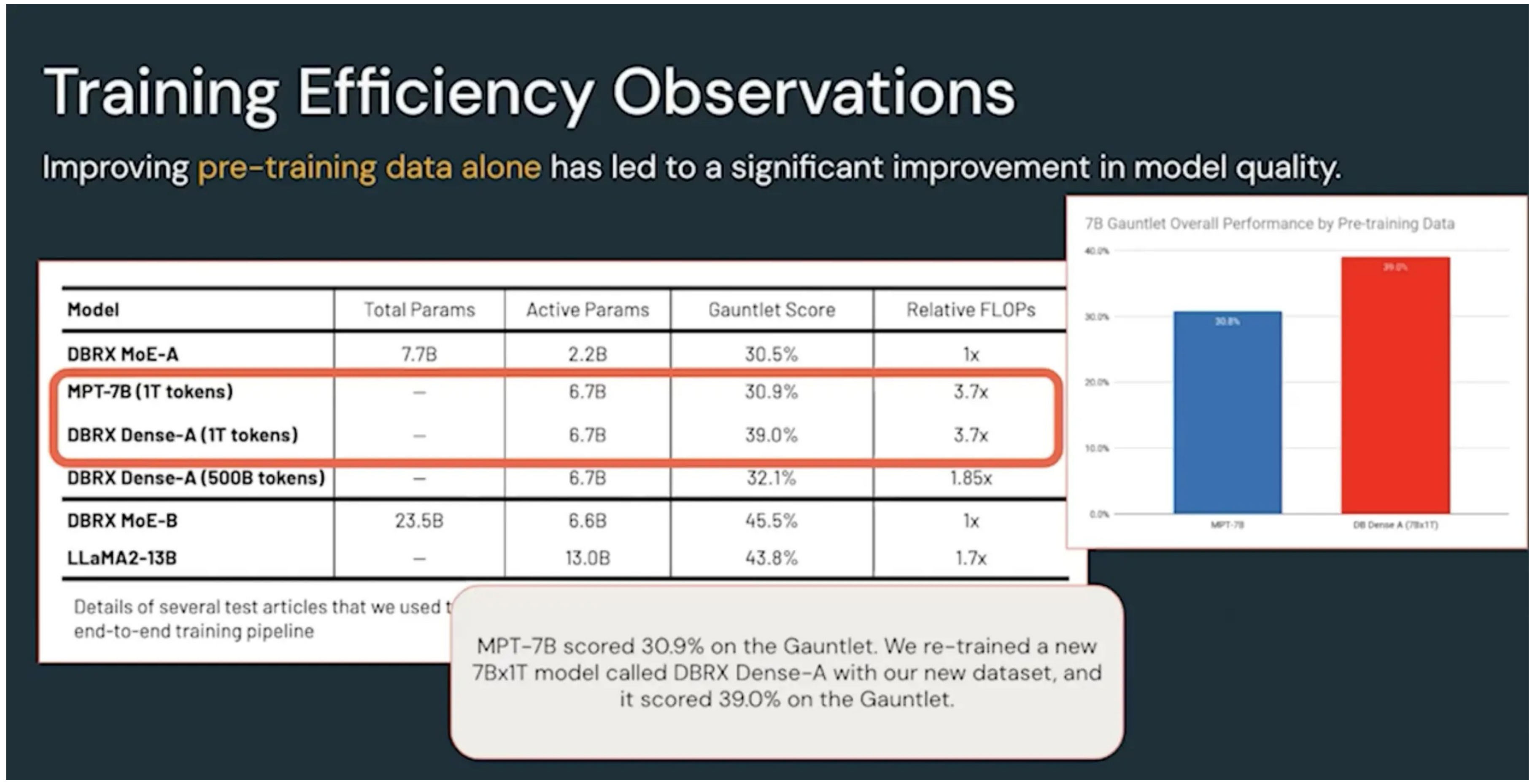


Data Quality – Diversity and **Quality**

Quality in LLM Training Data: Removing Noise, Filtering Low-Quality Content, and Prioritizing High-Value Sources

Why Does Quality Matter?

- Accuracy – Reduces hallucinations and misinformation
- Coherence – Ensures fluent, grammatically correct responses
- Reliability – Strengthens factual correctness and trustworthiness



Model	Total Params	Active Params	Gauntlet Score	Relative FLOPs
DBRX MoE-A	7.7B	2.2B	30.5%	1x
MPT-7B (1T tokens)	–	6.7B	30.9%	3.7x
DBRX Dense-A (1T tokens)	–	6.7B	39.0%	3.7x
DBRX Dense-A (500B tokens)	–	6.7B	32.1%	1.85x
DBRX MoE-B	23.5B	6.6B	45.5%	1x
LLaMA2-13B	–	13.0B	43.8%	1.7x

Key Aspects of Data Quality

1. Noise Removal: Filtering Out Unreliable Data

- Eliminate spam, duplicated content, and nonsensical text
- Avoid sources with misinformation, clickbait, and AI-generated text

2. Low-Quality Content: Identifying & Excluding Harmful Inputs

- Poor grammar, factual errors, or incomplete sentences degrade the model's performance.
- Remove overly biased, outdated, or offensive material.

3. High-Value Sources: Prioritizing Trusted & Structured Knowledge

- Books & Academic Papers – Provide well-researched, structured content
- Code Repositories – Enhance coding capabilities (e.g., GitHub, Stack Overflow)
- Peer-Reviewed Articles & Government Publications – Improve factual correctness

Data Quality – Toxicity

The Importance of Filtering
Filtering content is crucial for ensuring ethical AI development, improving user trust, and aligning models with safety standards. However, filtering presents a fundamental challenge: striking the right balance between removing harmful content and preserving diversity in training data.

Methods for Detecting and Filtering Toxicity
To identify and mitigate harmful content, various toxicity detection models and filtering techniques are employed:

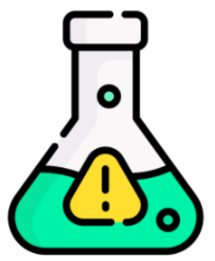
- Lexical and Heuristic Filtering: Identifies explicit offensive terms, hate speech, or discriminatory language.
- Contextual and ML-Based Approaches: Advanced models assess toxicity in context to reduce false positives from harmless discussions of sensitive topics.

The Trade-Off: Over-Filtering vs. Under-Filtering
Over-Filtering Risks

- Removing too much content can lead to censorship and reduce exposure to diverse viewpoints.
- Loss of domain-specific data from marginalized communities, affecting model inclusivity.

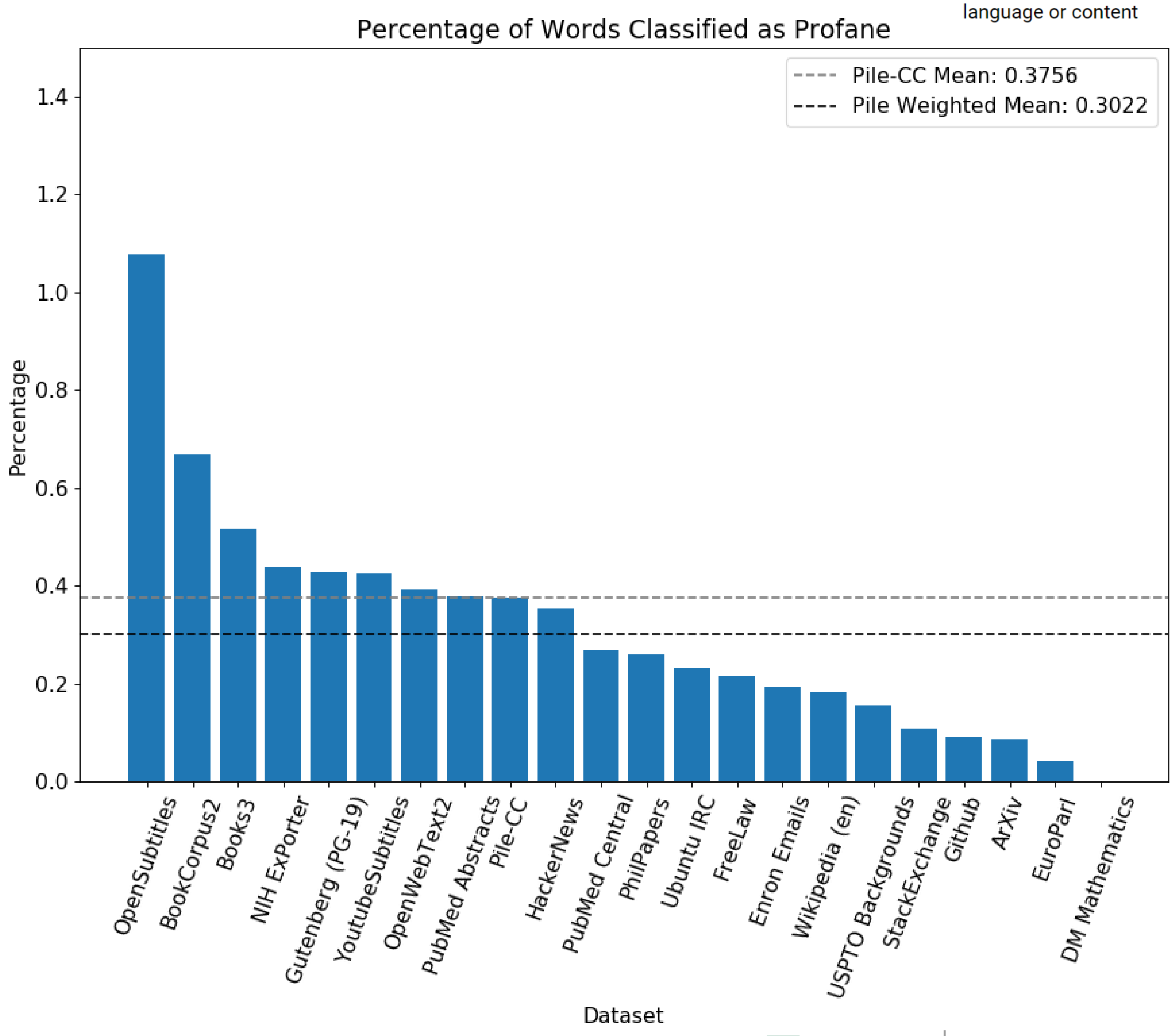
Under-Filtering Risks

- Retaining biased or toxic content can lead to models that amplify harmful stereotypes or generate unsafe outputs.
- Regulatory and ethical concerns arise when LLMs produce offensive, misleading, or discriminatory responses.



Toxicity

Harmful or discriminatory language or content



Dataset Curation

Preparing Data for Training

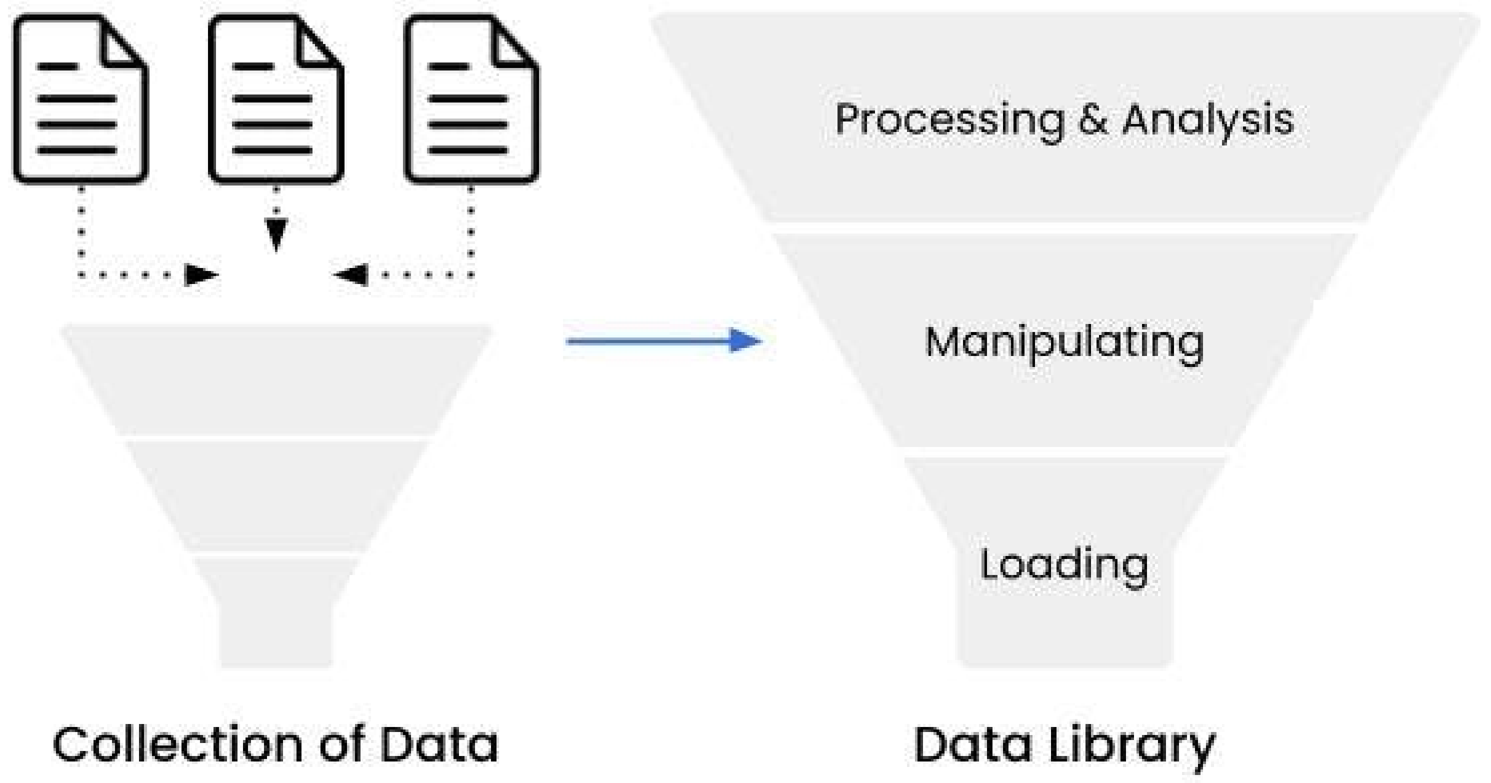
To train a high-performing language model, raw data must go through a structured curation pipeline consisting of four key stages:

1. **Sourcing** – Collecting text data from diverse and reliable sources.
2. **Cleaning** – Removing duplicates, filtering low-quality or harmful content, and standardizing formats.
3. **Tokenization** – Converting text into tokenized representations for efficient model processing.
4. **Storage** – Organizing the processed dataset in a format optimized for large-scale training.

Each step ensures that the data is high-quality, unbiased, and structured for optimal model performance.

Why Raw Internet Data Isn't Directly Usable

- **Noisy and Unstructured Format:** Internet data contains spam, misspellings, broken sentences, and low-quality text that can degrade model performance.
- **Duplicates and Redundancy:** Large-scale web crawls often collect duplicate or near-duplicate documents, leading to overfitting on repetitive data.
- **Bias and Toxicity:** Raw text can contain misinformation, hate speech, or biased narratives.
- **Tokenization Challenges:** Models process text as tokens, requiring specialized tokenization strategies (e.g., Byte Pair Encoding (BPE), SentencePiece).
- **Storage and Retrieval Considerations:** Training on trillions of tokens requires optimized storage formats (e.g., TFRecord, Arrow, LMDB) and efficient data loading pipelines.



Sourcing Data Mixes

Optimizing Data Mixes for LLM Training
The composition of training data plays a critical role in an LLM's ability t generalize across tasks and domains. The selection and proportion of different sources significantly impact performance.

Case Study: The Pile Dataset
The Pile (825 GiB) by EleutherAI is a well-curated dataset designed for large-scale LLMs. It includes 22 diverse sources such as academic papers, legal texts, and code repositories, enhancing cross-domain knowledge and adaptability.

Challenges in Data Mixing
- Diversity vs. Relevance: A broad mix improves generalization but may introduce noise.
- Bias Mitigation: Overrepresentation of specific domains can skew model behavior.
- Resource Constraints: Storing and processing large datasets demands high computational power.

Strategies for Optimized Data Mixing
- Data Mixing Laws: Predicts LLM performance across different data blends, enabling optimal selection pre-training.
- Efficient Online Data Mixing: Dynamically adjusts data proportions during training based on evolving model needs.

Careful curation and adaptive mixing strategies help balance diversity, quality, and efficiency in LLM training.

Table 1: Domain weights on The Pile. Baseline domain weights are computed from the default Pile dataset. DoReMi (280M) uses a 280M proxy model to optimize the domain weights.

Domain	Baseline	DoReMi (280M)	Difference	Domain	Baseline	DoReMi (280M)	Difference
Pile-CC	0.1121	0.6057	+0.4936	DM Mathematics	0.0198	0.0018	-0.0180
YoutubeSubtitles	0.0042	0.0502	+0.0460	Wikipedia (en)	0.0919	0.0699	-0.0220
PhilPapers	0.0027	0.0274	+0.0247	OpenWebText2	0.1247	0.1019	-0.0228
HackerNews	0.0075	0.0134	+0.0059	Github	0.0427	0.0179	-0.0248
Enron Emails	0.0030	0.0070	+0.0040	FreeLaw	0.0386	0.0043	-0.0343
EuroParl	0.0043	0.0062	+0.0019	USPTO Backgrounds	0.0420	0.0036	-0.0384
Ubuntu IRC	0.0074	0.0093	+0.0019	Books3	0.0676	0.0224	-0.0452
BookCorpus2	0.0044	0.0061	+0.0017	PubMed Abstracts	0.0845	0.0113	-0.0732
NIH ExPorter	0.0052	0.0063	+0.0011	StackExchange	0.0929	0.0153	-0.0776
OpenSubtitles	0.0124	0.0047	-0.0077	ArXiv	0.1052	0.0036	-0.1016
Gutenberg (PG-19)	0.0199	0.0072	-0.0127	PubMed Central	0.1071	0.0046	-0.1025

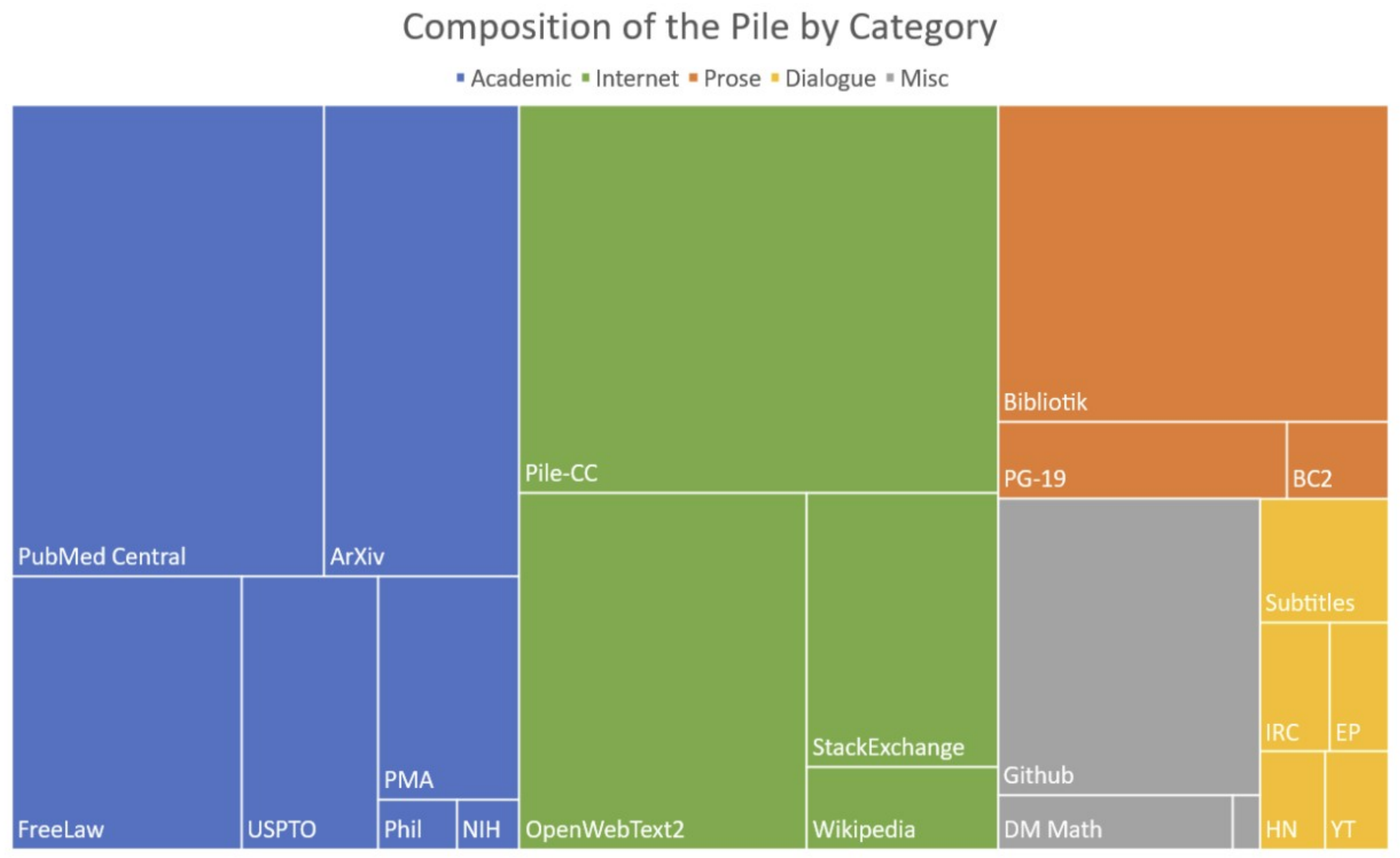


Figure 1: Treemap of Pile components by effective size.

DARTMOUTH ENGINEERING NVIDIA

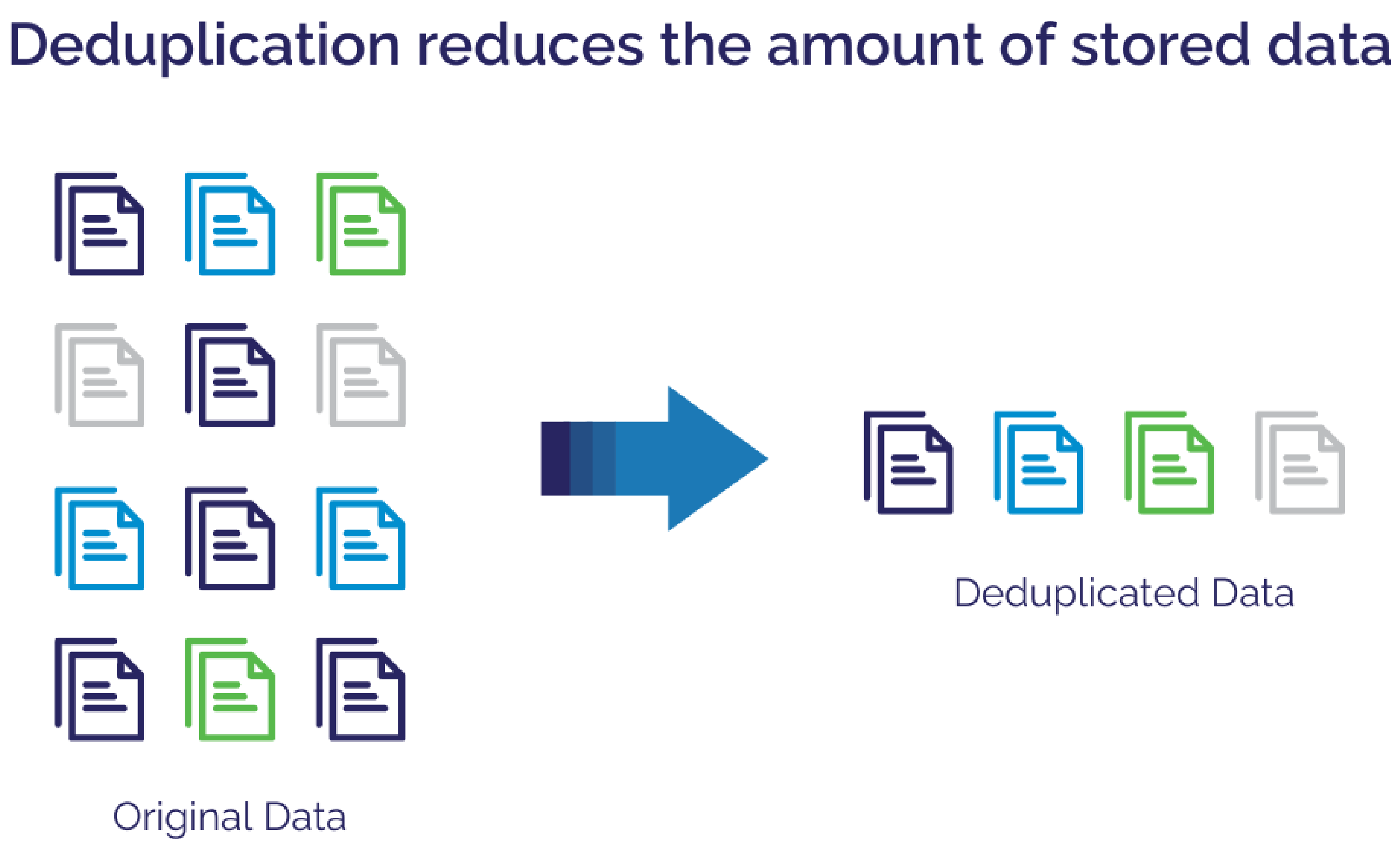
Deduplication and Cleaning

The Problem: Redundant and Noisy Data
Redundant data, such as repeated articles or copied content, can lead to overfitting and inefficient training. Additionally, raw text often contains noise, formatting issues, and irrelevant content, reducing model quality.

Deduplication Techniques
- Exact Matching: Identifies and removes identical text duplicates.
- Near-Duplicate Detection: Uses algorithms like SimHash and MinHash to detect similar but not identical content (e.g., paraphrased or slightly altered versions).

Data Cleaning Methods
- Noise Removal: Eliminates broken sentences, encoding errors, and malformed text.
- Formatting Standardization: Ensures consistent punctuation, spacing, and structure.
- Content Filtering: Removes irrelevant or low-quality text (e.g., spam, excessive boilerplate content).

Effective deduplication and cleaning improve efficiency, generalization, and overall LLM performance by ensuring high-quality, diverse, and non-redundant training data.

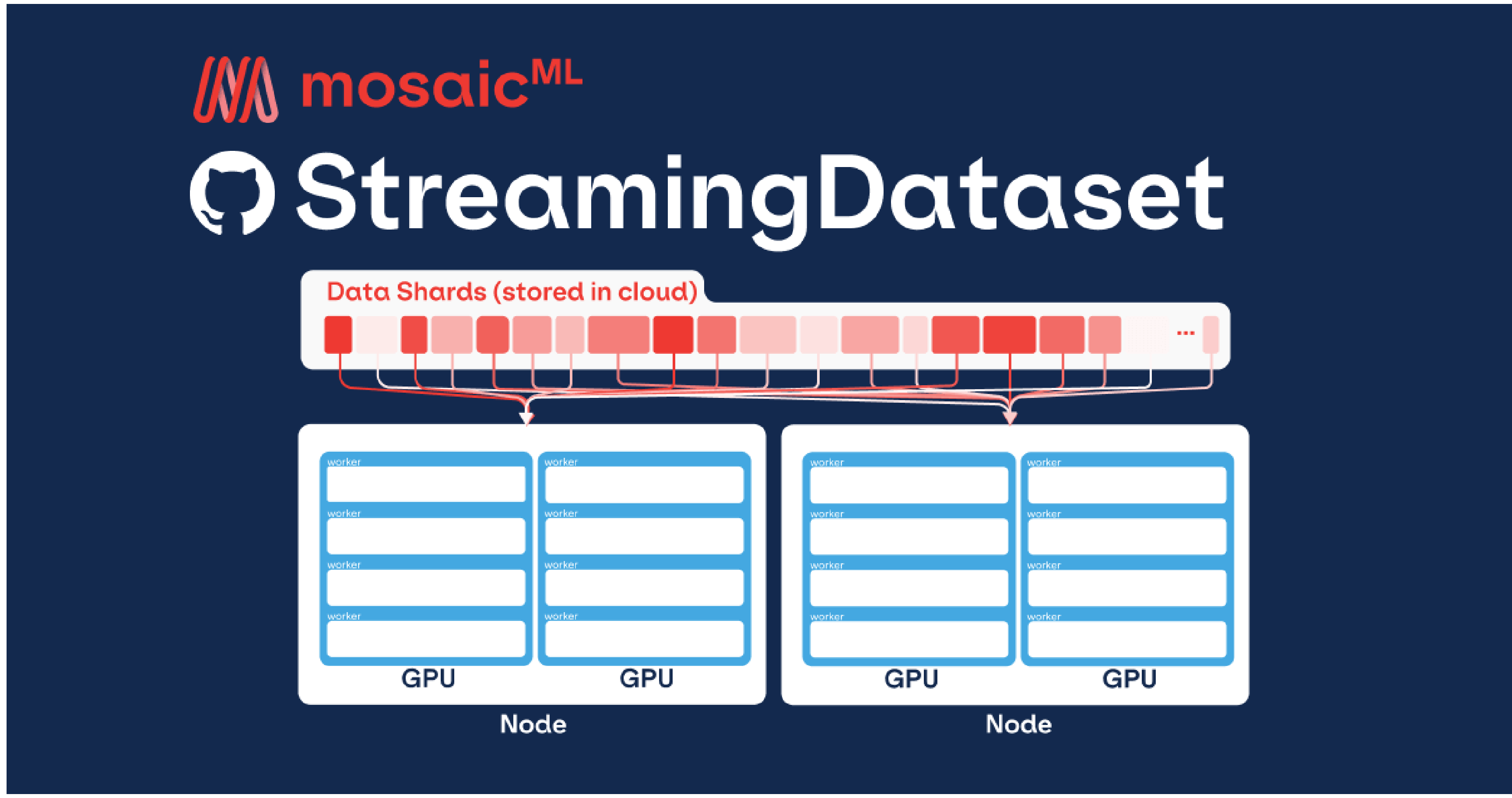


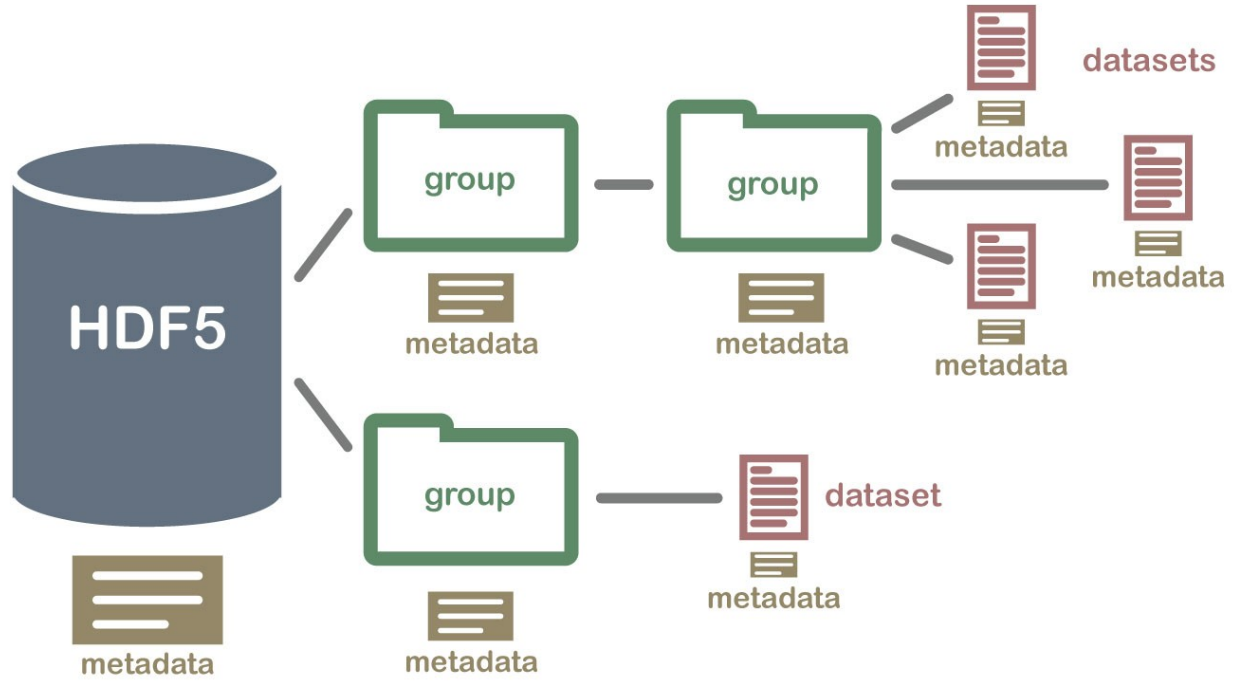
Tokenization and Storage of Datasets

Tokenization: Converting Text into Model-Ready Input
Raw text must be broken into tokens for efficient processing by LLMs. Tokenization balances efficiency and granularity, affecting model size, speed, and performance.

Common Tokenization Methods:
Byte Pair Encoding (BPE): Merges frequent character sequences into subwords.
WordPiece: Similar to BPE but optimized for linguistic structure (used in BERT).
SentencePiece: Works without predefined word boundaries, useful for multilingual models.

Storage: Handling Large-Scale Datasets
With terabytes of text, efficient storage is essential for scalability and fast retrieval.
Key Storage Formats:
TFRecord (TensorFlow format): Optimized for streaming large datasets in ML pipelines.
HDF5: Supports hierarchical, structured data storage for easy access.
Sharded Datasets: Splits data across multiple files or nodes to improve parallel processing.





ENGINEERING | NVIDIA.

Challenges in Data Collection

Copyright and Ethical Considerations of Datasets

Legal Risks: Copyrighted Material in Training Data
Training LLMs on copyrighted content without permission poses significant legal risks:

- Lawsuits & Regulatory Scrutiny: AI companies have faced litigation for unauthorized use of books, news articles, and other protected works.
- Unclear Legal Precedents: Many jurisdictions lack clear rulings on whether scraping and using copyrighted text for AI training constitutes infringement.

Fair Use & Licensing Considerations
Fair Use (U.S.): Some AI training may qualify under fair use, but this depends on factors like purpose, transformation, and market impact.
<u>Licensed & Open Data:</u>

- Creative Commons (CC-BY, CC0) allows reuse with attribution or without restrictions.
- Public Domain & Government Publications (e.g., U.S. federal documents) are free to use.
- Private Agreements: Some organizations license proprietary datasets explicitly for AI training.

Balancing Ethics and Compliance
AI developers must ensure responsible data use by favoring licensed, open-source, and ethically sourced datasets, reducing legal exposure while promoting fair AI development.

Ensuring Safe and Ethical Data in LLM Training

The Need for Filtering Harmful Content
LLMs trained on unfiltered web-scale data risk learning and generating harmful, misleading, or biased outputs. Careful dataset curation helps prevent:

- Hate speech and toxicity that can reinforce discrimination.
- Misinformation that undermines factual reliability.
- Personally Identifiable Information (PII) that raises privacy concerns.

Techniques for Filtering Harmful Data

- Hate Speech Detection: Automated classifiers (e.g., Perspective API, Jigsaw) flag toxic content.
- Misinformation Filtering: Prioritizing fact-checked sources and removing unreliable data.
- PII Redaction: Detecting and removing sensitive personal data (e.g., names, addresses, phone numbers).

The Challenge: Openness vs. Content Safety

- Strict filtering improves safety but may limit model diversity and robustness.
- Looser constraints preserve broad knowledge but increase risks of harmful outputs.
- Best practice: Use a layered approach, combining automated filtering, human review, and reinforcement learning to balance inclusivity with responsibility.

Data Augmentation

Dataset Size Requirements

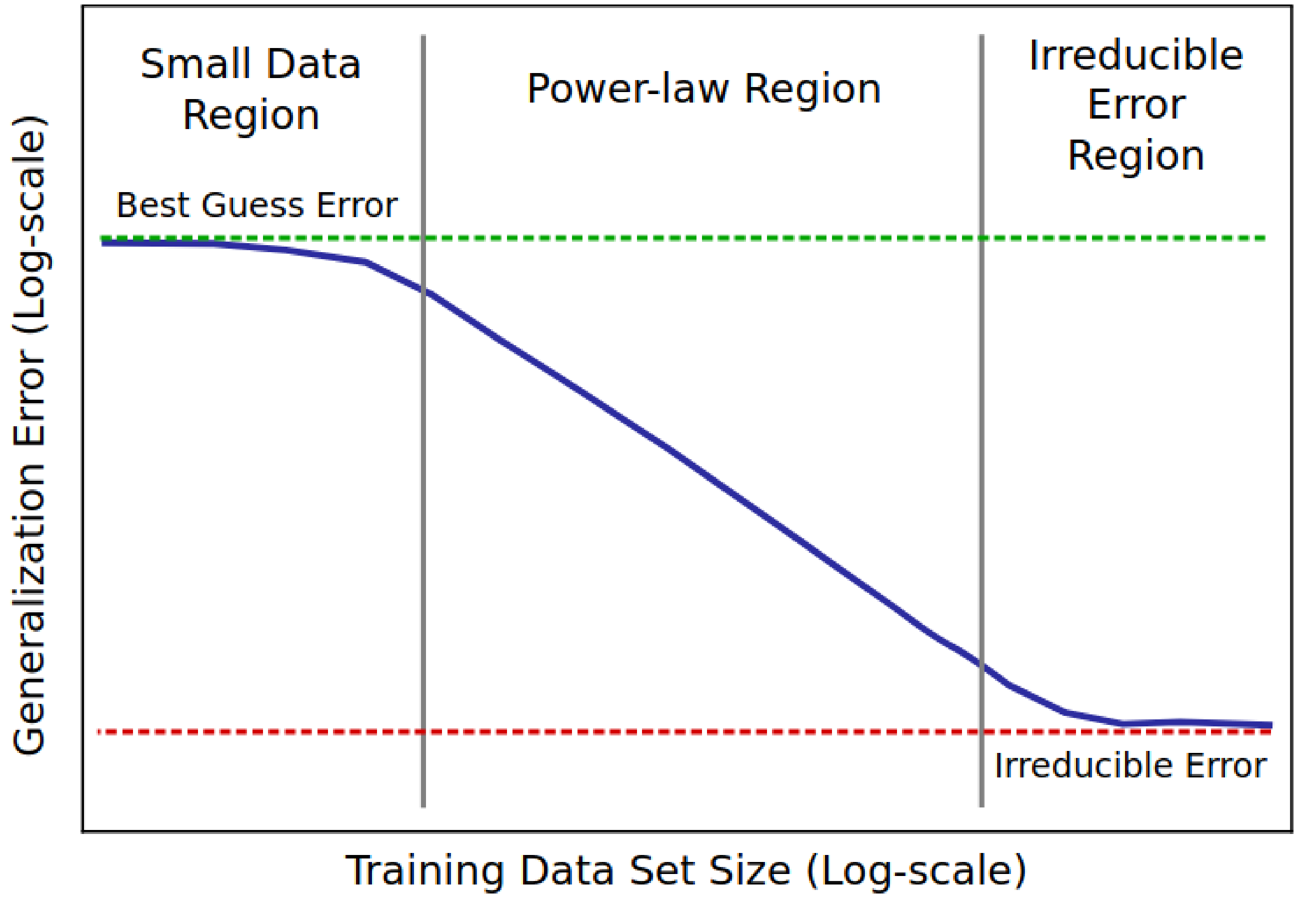
The Need for Massive Training Data
Larger LLMs require trillions of tokens for effective pretraining. However, sourcing enough high-quality text poses challenges:
The internet has finite high-quality data, requiring careful selection.
Expanding dataset size while maintaining quality and diversity is difficult.
Scaling up data must align with compute constraints and efficiency.

Challenges in Sourcing Large-Scale Data

- Limited Availability of High-Quality Text: Most web data is noisy or redundant.
- Ethical & Legal Constraints: Copyrighted materials cannot be freely used.
- Diminishing Returns: As models grow, more data is needed to maintain performance gains.

Scaling Laws and Dataset Size

- Empirical scaling laws show that model performance improves predictably with data size.
- Optimal data-to-parameter ratios must be maintained to prevent underfitting or wasted compute.
- Synthetic data generation and augmentation can help bridge gaps when real-world data is insufficient.



Synthetic Data Generation

Using AI-Generated Text for Data Augmentation

To overcome data limitations, researchers augment real-world datasets with AI-generated text from models like GPT. This approach helps:

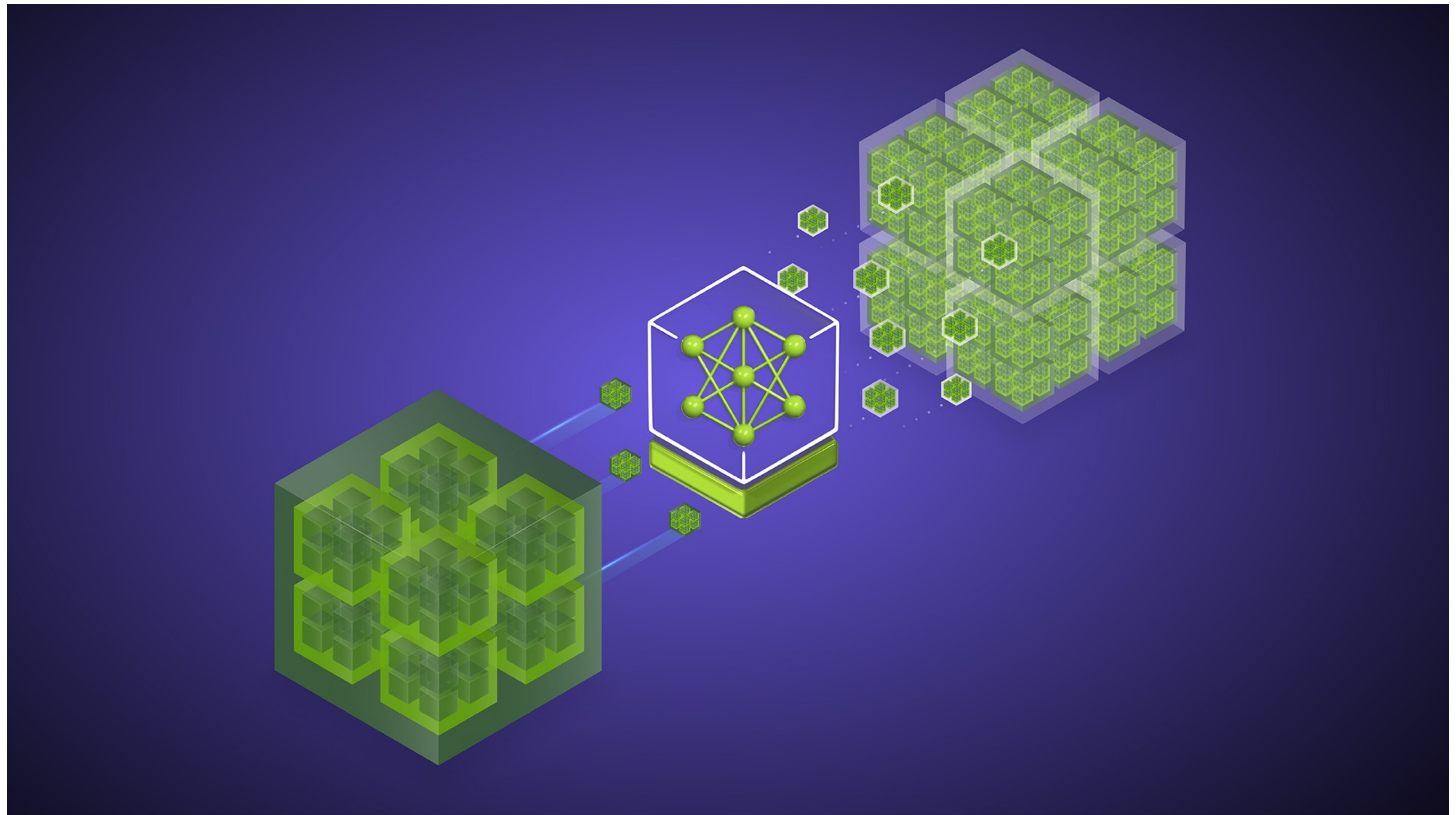
- Expand dataset size when high-quality human-written text is scarce.
- Fill gaps in underrepresented languages, topics, or styles.
- Reduce reliance on copyrighted or sensitive material by generating alternative text.

Benefits of Synthetic Data

Scalability: Enables training LLMs without solely depending on real-world text.
Customization: Tailors data to specific domains (e.g., scientific, legal, conversational).
Bias Reduction: Helps balance training data by supplementing underrepresented perspectives.

Risks and Challenges

- Model Collapse: Excessive use of AI-generated text leads to a feedback loop where models train on their own outputs, reducing originality and diversity.
- Quality Control: AI-generated text may contain errors, biases, or lack real-world grounding, requiring careful filtering.
- Detectability: Hard to distinguish synthetic vs. real data, complicating dataset validation.



New Data Sources

Expanding Data Sources Beyond Web Scraping
While web crawling provides a broad dataset, it has limitations in quality, structure, and diversity. To supplement training data, LLM developers explore alternative sources that offer high-quality, specialized, and underrepresented text.

Alternative Data Sources

OCR of Historical Documents

- Optical Character Recognition (OCR) extracts text from books, archives, handwritten manuscripts, and scanned PDFs.
- Expands LLM knowledge in history, literature, and cultural studies.
- Challenge: Requires error correction due to OCR inaccuracies.

Transcriptions of Spoken Language

- Automatic Speech Recognition (ASR) systems generate text from audio sources, capturing conversational, dialectal, and low-resource languages.
- Improves LLM dialogue capabilities and multilingual support.
- Challenge: Speech data requires heavy processing to clean disfluencies and improve text structure.

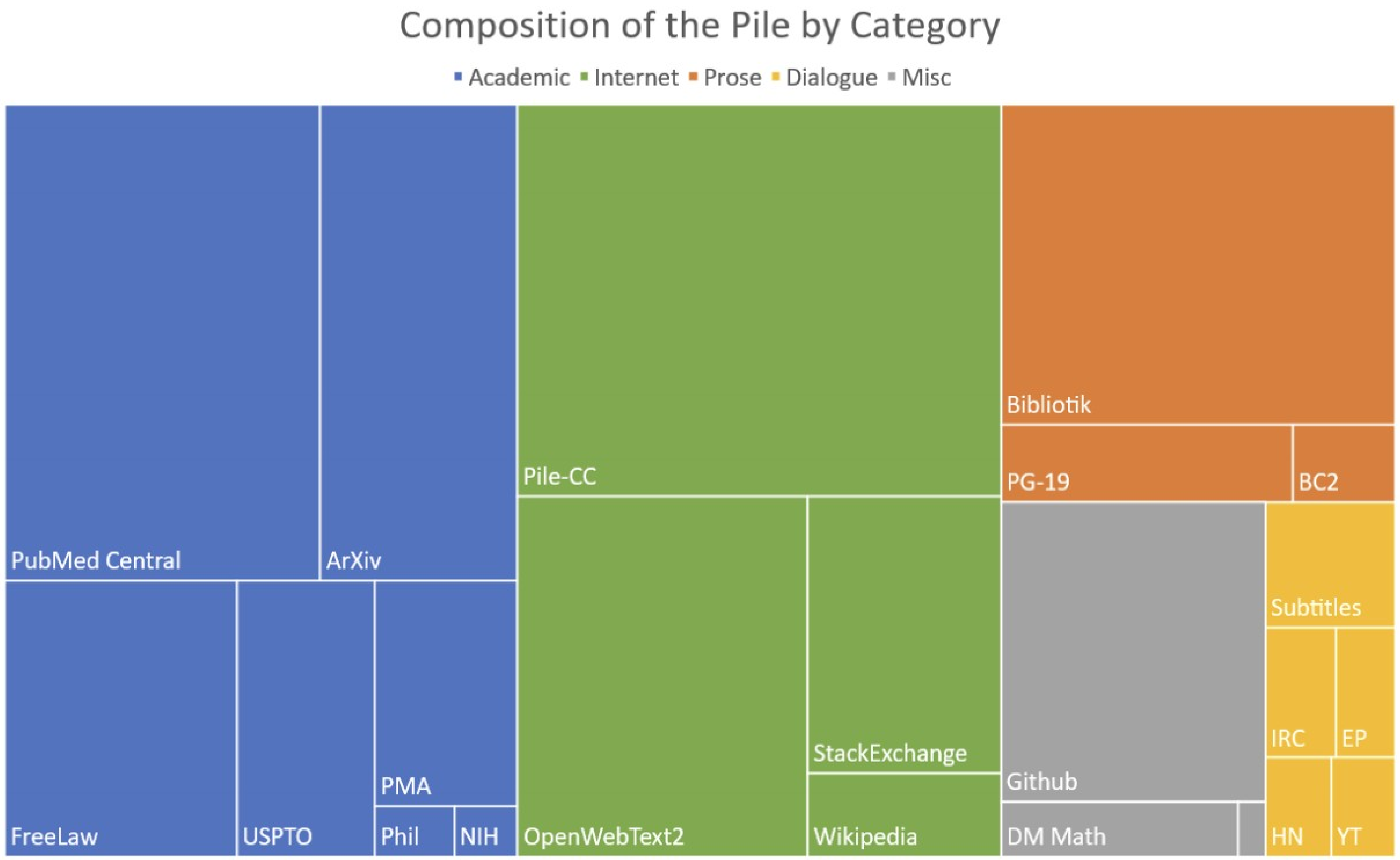
Why These Sources Matter

- More diverse than web data, improving model generalization.
- Higher quality and curated, reducing noise and misinformation.
- Access to specialized knowledge that's underrepresented online.

Wrap Up

Training Data for Larger LLMs

- Today, we explored the role of data quality and curation in training LLMs.
- We saw that curating, filtering, and tokenizing data is essential for model performance.
- We discussed the challenges of copyright, ethics, and content filtering, balancing openness with safety.
- We examined data augmentation as a way to expand training datasets while mitigating data scarcity.
- Finally, we considered the trade-offs of proprietary datasets, which provide advantages but raise ethical concerns.



Thank you!