고려대학교 글로벌인문학연구원 한자한문연구소 2025국제학술대회

# 글로벌한국학과 디지털인문학의 접점

# Beyond Bigger Data: How Dataset Quality Impacts Large Language Model (LLM) Performance


## Presenter: Dr. Prince Hamandawana

**Asst. Prof. Department of Software, Ajou University**


高麗大學校 Global人文學硏究院

Korea University Institute for Global Humanities Research and Collaboration

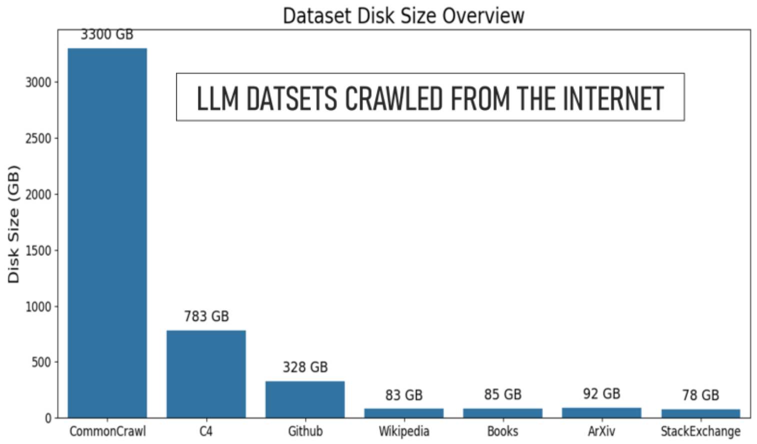
Dynamic Tomorrow

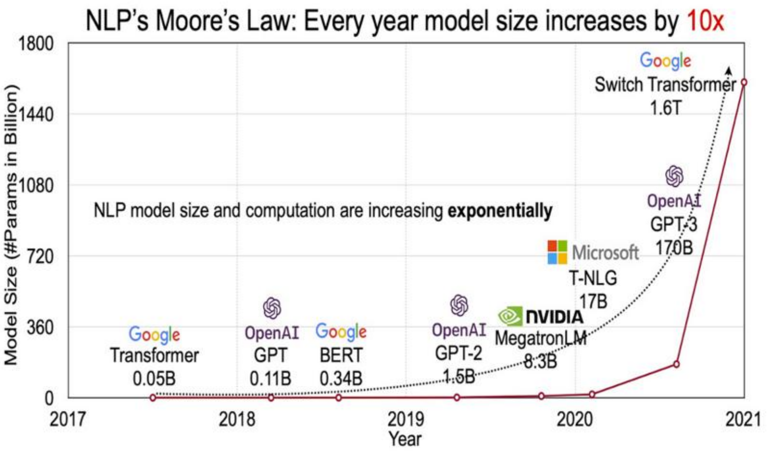
# Agenda

1. Introduction: The Role of Large Datasets in LLM Success
2. Challenges of Duplicated Datasets in LLMs
3. Proposed Solutions for Deduplication
4. Experimental Setup and Evaluation
5. Key Results from Deduplication Experiments
6. Conclusion: The Importance of Dataset Quality
7. Q&A Session

# 1. Introduction: The Role of Large Datasets in LLM Success

## Understanding the Foundation of LLMs




**Significance of Large Datasets**

Models like GPT-3 and BERT require **extensive training** data to capture the complexity of language and context. The scale directly influences their depth of understanding and fluency.

**Key Datasets in LLM Training**

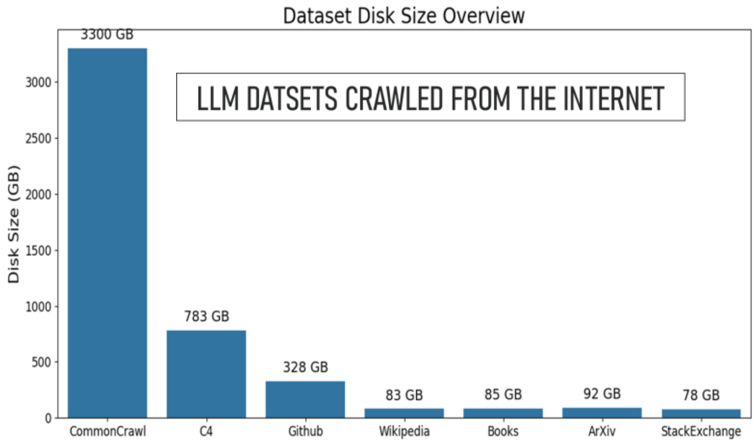
Datasets such as Wiki-40B and C4 are crucial for LLMs, as they include diverse linguistic styles and topics, encompassing **billions of words** and a range of genres.

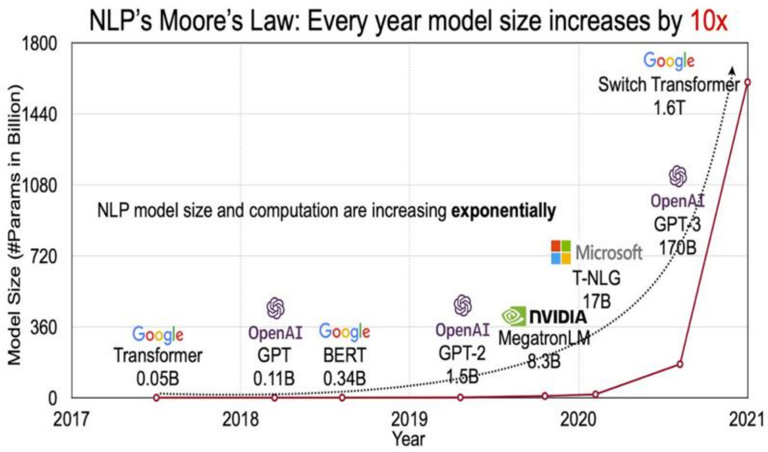
**Quality as a Determinant of Performance**

The quality of data directly correlates with the **model's predictive accuracy and generalization** abilities; low-quality datasets can severely hinder these aspects, leading to biased outputs.

# 1. Introduction: The Role of Large Datasets in LLM Success

## Understanding the Foundation of LLMs




These datasets often contain repeated text sequences

- **Duplicates / Near duplicates**

their depth of understanding and fluency.

a range of genres.

quality datasets can severely hinder these aspects, leading to biased outputs.

# 2. Challenges of Duplicated Datasets in LLMs

## Understanding the Pitfalls of Data Redundancy

**1. Memorization and Its Consequences:** When LLMs encounter duplicated data, they risk memorizing this content instead of genuinely understanding it, which may lead to outputting verbatim text rather than generating novel responses.

Case 1:23-cv-11195 Document 1-68 Filed 12/27/23 Page 4 of 127

ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES 3

**Output from GPT-4:**

many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capi-

**Actual text from NYTimes:**

many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House.

**Figure 1:** Examples of GPT-4 outputs The New York Times's copyrighted articles verbatim.

- **Legal Ramifications:** Memorization due to sample duplication leads to legal risks
- Models repeat exact training text, verbatim
- E.g., New York Times sued OpenAI for allegedly using its articles to train GPT-4

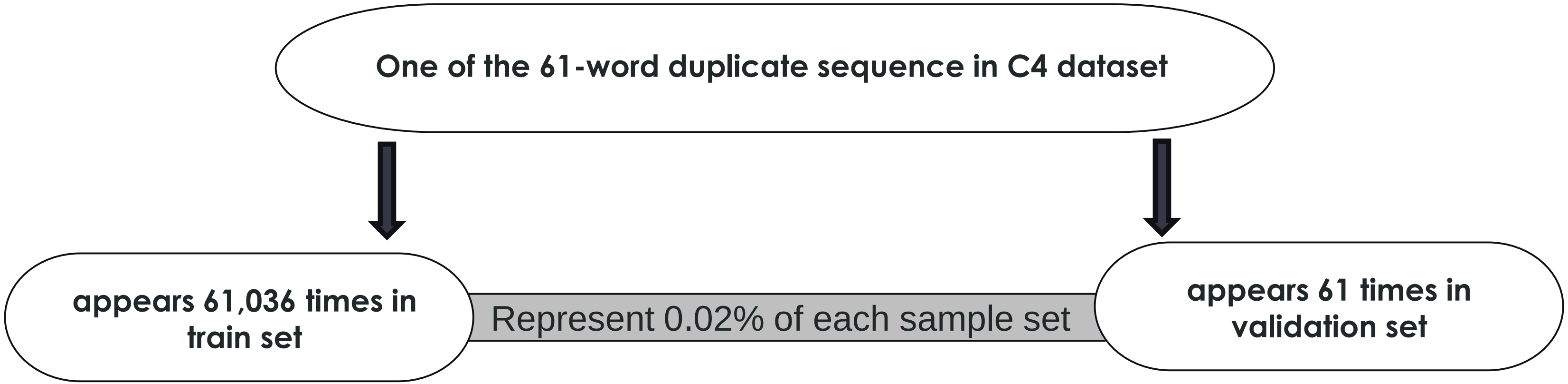
**SOURCE: STANFORD AI LAB**

# 2. Challenges of Duplicated Datasets in LLMs

## Understanding the Pitfalls of Data Redundancy

**2. Train/test data overlap :** Some test examples appear in training data, causing overestimation of model accuracy.

- **EXAMPLE:** Colossal Cleaned Common Crawl **(C4)** Dataset



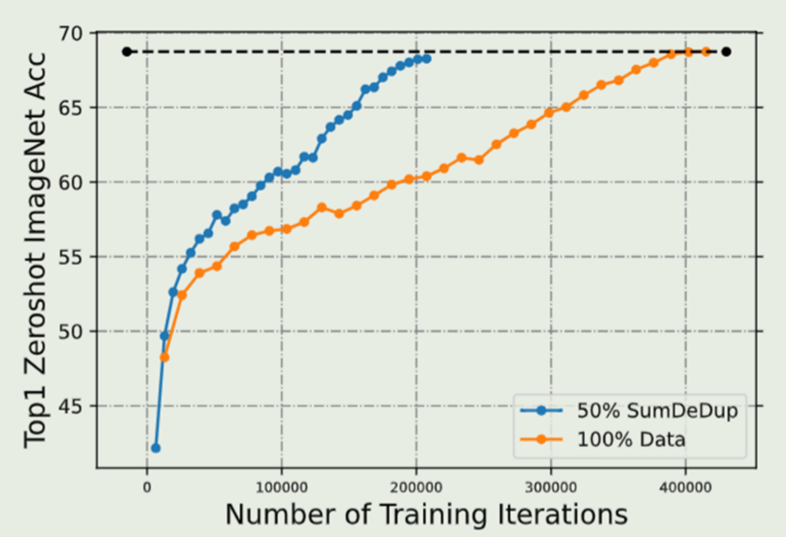

- This overlap results in inflated model accuracy estimates.
- Biases model selection toward models that overfit the training data.

# 2. Challenges of Duplicated Datasets in LLMs

## Understanding the Pitfalls of Data Redundancy

**3. Increased Training Costs:** Repeated data wastes resources and time

• More epochs required before model convergence




SOURCE: ABAS ET AL, 2023

# 2. Challenges of Duplicated Datasets in LLMs

## Understanding the Pitfalls of Data Redundancy

**4. Deduplicating exabyte scale datasets is complex.**

- Datasets are huge (hundreds of gigabytes to terabytes).
- Duplicates can be exact or near duplicates with small differences (e.g., dates, names).
- Naive duplicate detection (comparing every pair) is computationally expensive.
- Naive duplicate detection time complexity $O(n^2)$

Example of Identifying word sequences of a given length that repeat above a set threshold in Wiki40b dataset.

Command

```
cargo run self-similar --data-file data/wiki40b.test --length-threshold 100 --cache-dir /tmp/cache --num-threads 8
```

Output

```
Duplicates found: 3,374,227
```

# 2. Challenges of Duplicated Datasets in LLMs

## Understanding the Pitfalls of Data Redundancy

**4. Deduplicating exabyte scale datasets is complex.**

• Datasets are huge (hundreds of gigabytes to terabytes).

• Duplicates can be exact or near duplicates with small differences (e.g., dates, names).

• While removing duplicates from training data, we need to keep the test/validation sets clean.

So, there is a need for efficient duplicate sample data removal !!!

Command

```
cargo run self-similar --data-file data/wiki40b.test --length-threshold 100 --cache-dir /tmp/cache --num-threads 8
```

Output

```
Duplicates found: 3,374,227
```

# 3. Proposed Solutions for Deduplication

## Innovative Approaches to Clean Data

**1. Utilizing Exact Substring Matching**
By employing suffix arrays for exact matching, we can identify repeated sequences efficiently, significantly reducing redundancy in datasets.

**2. Employing Approximate Matching Techniques**
MinHash and locality-sensitive hashing can be leveraged to cluster near-duplicate documents, which is especially useful for dealing with slightly varied content across the web.

**3. Scalability Challenges**
Deduplicating data at the scale of exabytes requires not only sophisticated algorithms but also an investment in computing resources to effectively implement those solutions.

# 3. Proposed Solutions for Deduplication

## Innovative Approaches to Clean Data

1. **Utilizing Exact Substring Matching to improve Efficiency of naïve all pairs matching (quadratic time O(n2);**
    - Concatenate samples into long sequences of text segments, 50+ tokens.
    - Then use a Suffix Array data structure to find repeated substrings efficiently.
    - Suffix Array sorts all suffixes of the dataset text, enabling fast detection of repeated sequences.
    - Runs in linear time O(n) relative to dataset size, feasible for large datasets

**Let's see how this works !!!**

# 3. Proposed Solutions for Deduplication

## Innovative Approaches to Clean Data

### 1. Utilizing Exact Substring Matching with Suffix Arrays

- A **Suffix** is a substring at the end of a given string
- A **Suffix Array** stores the starting positions of all suffixes of a string, sorted in alphabetical order
- Lastly, we use the **Longest Common Prefix Array**, which tracks duplicates between 2 adjacent suffixes

Suffix Index

**Example suffixes**

| | | | | | | |
|---|---|---|---|---|---|---|
| [0] | B | A | N | A | N | A |
| [1] | | A | N | A | N | A |
| [2] | | | N | A | N | A |
| [3] | | | | A | N | A |
| [4] | | | | | N | A |
| [5] | | | | | | A |

**LCP**

| |
|---|
| 0 |
| 1 |
| 3 |
| 0 |
| 0 |
| 2 |

Starting position of suffix

**Example suffix Array**

| | | | | | | |
|---|---|---|---|---|---|---|
| [5] | A | | | | | |
| [3] | A | N | A | | | |
| [1] | A | N | A | N | A | |
| [0] | B | A | N | A | N | A |
| [4] | N | A | | | | |
| [2] | N | A | N | A | | |

# 3. Proposed Solutions for Deduplication

## Innovative Approaches to Clean Data

**1. Utilizing Exact Substring Matching to improve Efficiency of naïve all pairs matching (quadratic time O(n2);**

Starting position of suffix

**Example suffix Array**

| LCP | Starting position of suffix | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | [5] | A | | | | | |
| 1 | [3] | A | N | A | | | |
| 3 | [1] | A | N | A | N | A | |
| 0 | [0] | B | A | N | A | N | A |
| 0 | [4] | N | A | | | | |
| 2 | [2] | N | A | N | A | | |

**Total # of substrings from string banana**

- b
- a
- n
- a
- n
- a
- ba
- an
- na
- an
- na
- ban
- ana
- nan
- ana
- bana
- anan
- nana
- banan
- anana
- banana

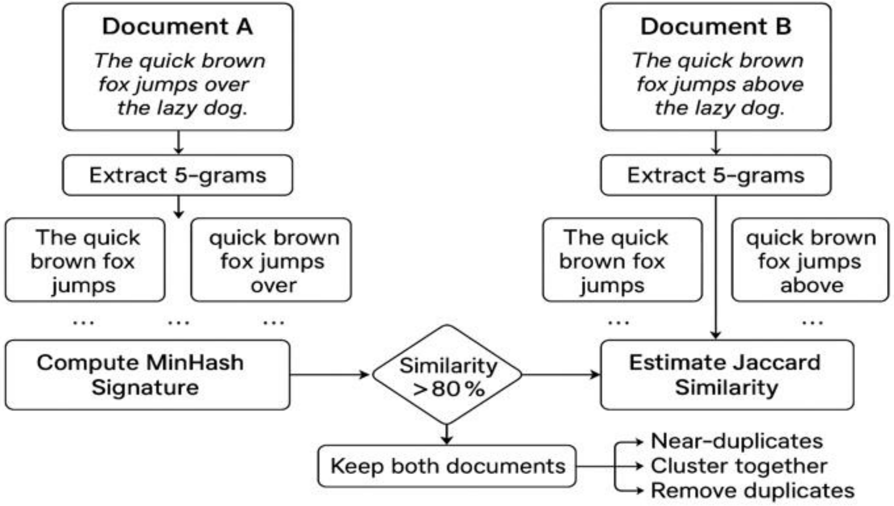
$$\text{Number of unique substrings} = \underbrace{\frac{n(n+1)}{2}}_{\text{Total \# of Strings = 21}} - \underbrace{\sum_{i=1}^{n} \text{LCP}(i)}_{\text{Total exact matches / duplicate strings = 6}}$$

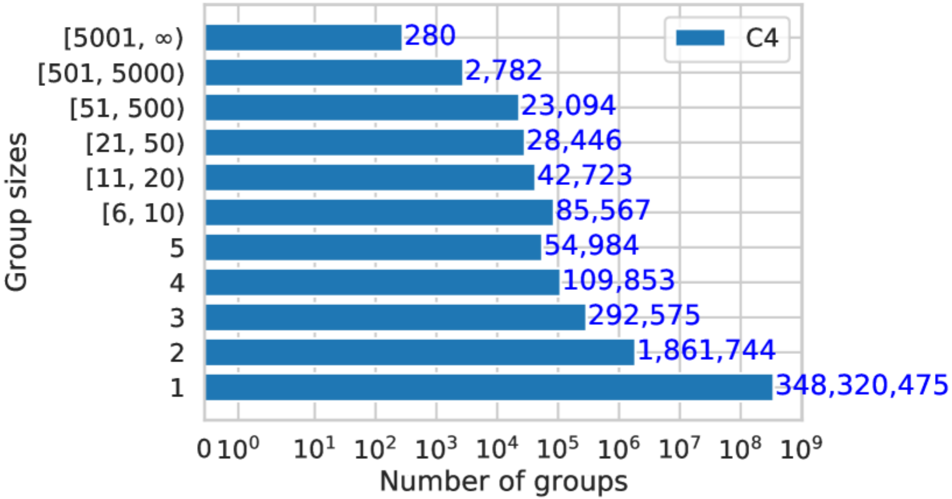
# 3. Proposed Solutions for Deduplication

## Innovative Approaches to Clean Data

### 2. Employing Approximate Matching Techniques (NearDup) with MinHash

- Represents each document by sets of 5-word sequences (5-grams).
- Uses MinHash to estimate similarity between documents without full comparisons.
- Employs **Jaccard Similarity**: Documents with similarity above 80% are considered near duplicates.
- Clusters duplicates and removes redundant documents.
- Handles cases where documents differ slightly (e.g., templated web pages).

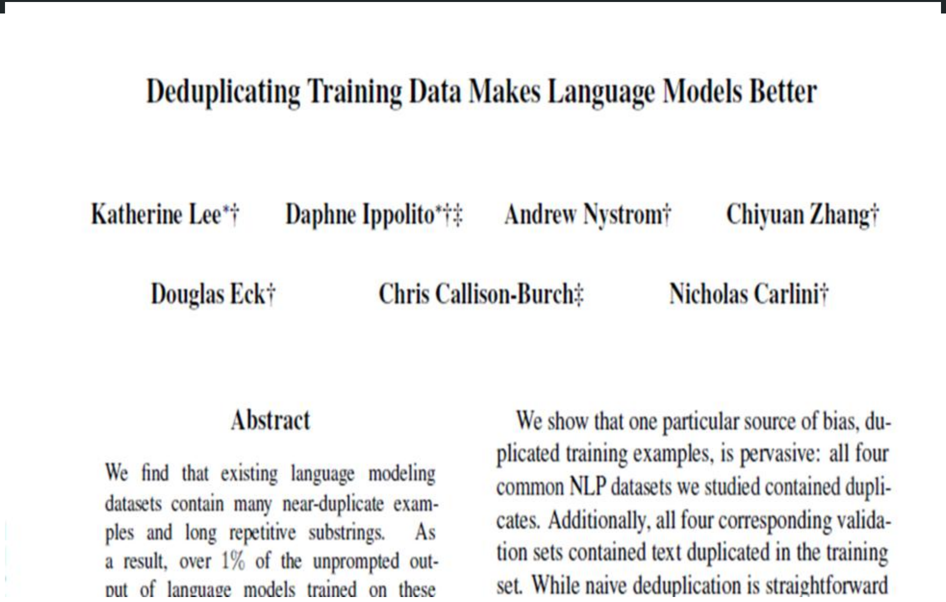


# 3. Proposed Solutions for Deduplication

## SOLUTION IMPLEMENTATION

## Google AI Research LAB



Deduplicating Training Data Makes Language Models Better

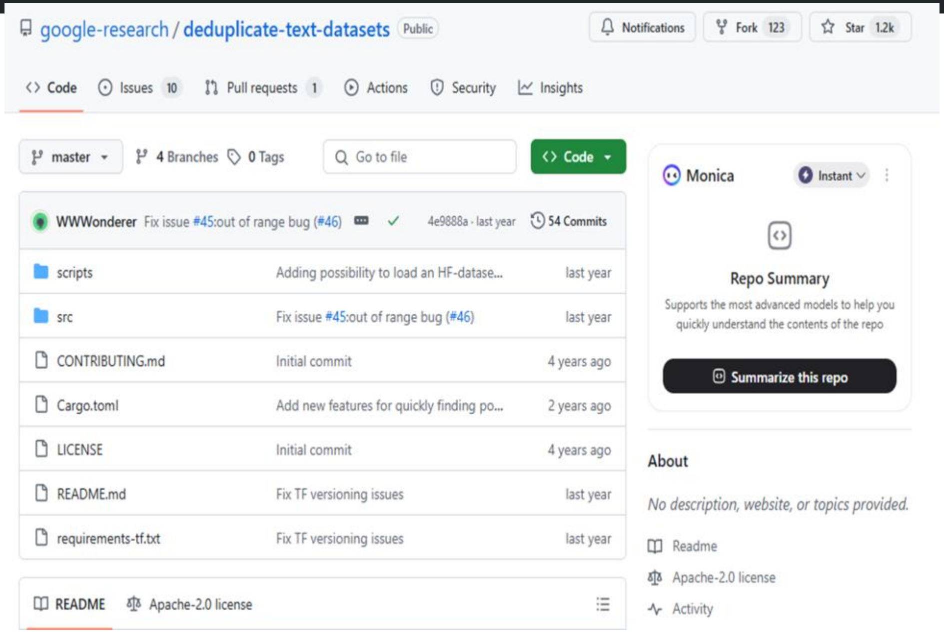
Katherine Lee*† Daphne Ippolito*†‡ Andrew Nystrom† Chiyuan Zhang†

Douglas Eck† Chris Callison-Burch‡ Nicholas Carlini†

Abstract

We find that existing language modeling datasets contain many near-duplicate examples and long repetitive substrings. As a result, over 1% of the unprompted output of language models trained on these

We show that one particular source of bias, duplicated training examples, is pervasive: all four common NLP datasets we studied contained duplicates. Additionally, all four corresponding validation sets contained text duplicated in the training set. While naive deduplication is straightforward



# 4. Experimental Setup and Evaluation

## Testing Deduplication Techniques

**Overview of Experimental Design**
Detailed application of deduplication techniques on datasets like C4 and RealNews is essential for understanding effectiveness across varying conditions.

**Contrasting Original and Deduplicated Data**
By comparing baseline datasets against deduplicated versions, we can derive insights into model performance, focusing on accuracy and generalizability metrics.

**Evaluating Outcomes of Deduplication**
Utilizing metrics such as perplexity and memorization rates, we assess the tangible impacts of deduplication on LLMs.

# 4. Experimental Setup and Evaluation

## Testing Deduplication Techniques

**Overview of Experimental Design**
Detailed application of deduplication techniques on 4 Datasets to understanding effectiveness across varying conditions.

**Contrasting Original and Deduplicated Data**
Compare baseline datasets against deduplicated versions to derive insights into model performance,

**Evaluating Outcomes of Deduplication**
Utilize metrics such as perplexity and memorization rates to assess the impacts of deduplication on LLMs.

| Aspect | Details |
|---|---|
| Datasets | - C4, RealNews, Wiki-40B, LM1B |
| Model | - Transformer-based- 1.5 billion parameters |
| Training Data | - Original data vs Deduplicated data |
| Evaluations | - Perplexity (text prediction ability)<br>- Memorization (copying training data) |

# 5. Key Results from Deduplication Experiments

## Insights on Model Performance Improvements

### Dataset Deduplication

- Near-duplicates found in all datasets, up to 13.6% of examples in RealNews.
- Exact Deduplication reduced dataset size by up to 19%.
- Reduced train-test overlap significantly, improving evaluation fairness.

**(1) Fraction of samples identified with MinHash (NearDedup)**

| Dataset | Train Duplicates (%) | Validation Duplicates (%) | Validation Overlap with Train (%) |
|---|---|---|---|
| **C4** | 3.04% | 1.59% | 4.60% |
| **RealNews** | **13.63%** | 1.25% | **14.35%** |
| **LM1B** | 4.86% | 0.07% | 4.92% |
| **Wiki40B** | 0.39% | 0.26% | 0.72% |

**(2) Fraction of samples identified with Exact Matching (Suffix Arrays)**

| Dataset | Train Duplicates (%) | Validation Duplicates (%) | Validation Overlap with Train (%) |
|---|---|---|---|
| **C4** | 7.18% | 0.75% | 1.38% |
| **RealNews** | **19.40%** | 2.61% | **3.37%** |
| **LM1B** | 0.76% | 0.02% | 0.02% |
| **Wiki40B** | 2.76% | 0.52% | 0.67% |

# 5. Key Results from Deduplication Experiments

## Insights on Model Performance Improvements

### Model Performance

1. **Models trained on deduplicated data:**

- Memorized text **10 times less** often.
- Achieved equal or better perplexity (sometimes improved by up to 10%).
- Deduplication reduced dataset size by up to 19%.
- Reduced train-test overlap significantly, improving evaluation fairness.

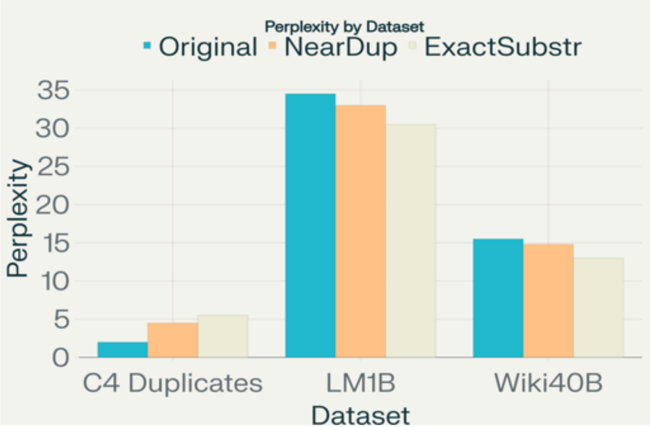
2. **Models trained on original data:**

- overfit duplicates, hurting generalization.

**LLM Generated Text: 100 000 sequences with no prompting**

| Model | 1 Epoch | 2 Epochs |
|---|---|---|
| XL-ORIGINAL | 1.926% | 1.571% |
| XL-NEARDUP | 0.189% | 0.264% |
| XL-EXACTSUBSTR | 0.138% | 0.168% |

*1% of tokens from the original model are exact duplicates from training data, reduced to just 0.1% with deduplicated training.*



# 6. Conclusion: The Importance of Dataset Quality

A Call for Commitment to Dataset Quality and Integrity

**Recap on Dataset Quality Findings**
The overarching narrative is clear; high-quality, deduplicated datasets are vital for optimizing LLM performance and training outcomes.

**Future Implications for LLM Development**
Moving forward, it is essential to prioritize dataset quality over sheer volume, as a cleaner dataset will yield more effective models suited for real-world applications.

**Encouragement for Best Practices**
Advocating for the widespread adoption of deduplication techniques among researchers and developers will fortify the integrity of datasets used in AI.

# References


1. Lee, Katherine et al. “Deduplicating Training Data Makes Language Models Better.” Annual Meeting of the Association for Computational Linguistics (2021).

2. Abbas, A., Tirumala, K., Simig, D., Ganguli, S., & Morcos, A.S. (2023). SemDeDup: Data-efficient learning at web-scale through semantic deduplication. ArXiv, abs/2303.09540.

3. google-research deduplicate-text-datasets. GitHub - google-research/deduplicate-text-datasets

# 7. Q&A Session

Engaging with the Audience

**Floor for Audience Questions on**
clarifications surrounding dataset challenges and deduplication techniques

**Practical Applications of Findings**
Real-world implementation strategies for deduplication techniques