

# Building a flexible data platform for LLM training data

## Who am I?

---



**Manager of Technical Staff, Cohere**

**Lead - Data Acquisition**



**Co-founder, Toronto Modern Data Stack**

Prev: Data @ Shopify, Instacart, Super

# Today's talk

---

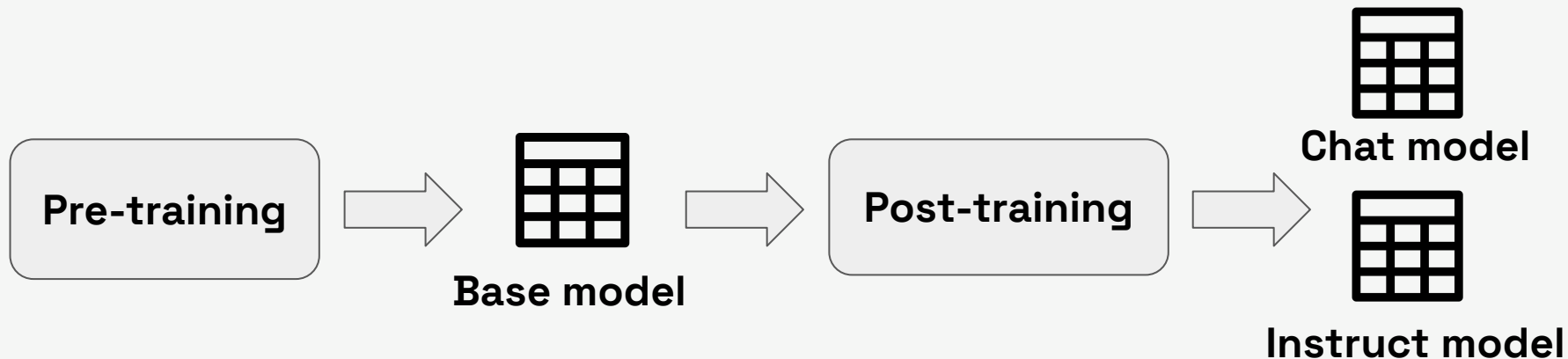
**Science of data for LLMs**

**Lifecycle of a pre-training dataset**

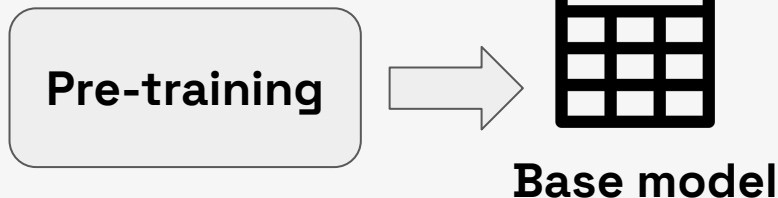
**Cohere's data platform**

# LLM training phases

---



# Pre-training



Self-supervised  
Billions of unlabelled docs  
Expensive

## Web pages

**Barack Hussein Obama II** ([/bəˈrɑːk huːˈseɪn ʊˈbɑːmə/](#) (listen)); born August 4, 1961)<sup>[1]</sup> is an [American politician](#) and [attorney](#). He was the 44th [president of the United States](#) from 2009 to 2017. He was the first [African-American](#) president in U.S. history. As a member of the [Democratic Party](#), he also served as member of the [Illinois Senate](#) from 1997 to 2004 and a [United States senator](#) from [Illinois](#) from 2005 to 2008.

## Code

```
class Bloom:
    def __init__(self, size: int = 8) -> None:
        self.bitarray = 0b0
        self.size = size

    def add(self, value: str) -> None:
        h = self.hash_(value)
        self.bitarray |= h

    def exists(self, value: str) -> bool:
        h = self.hash_(value)
        return (h & self.bitarray) == h

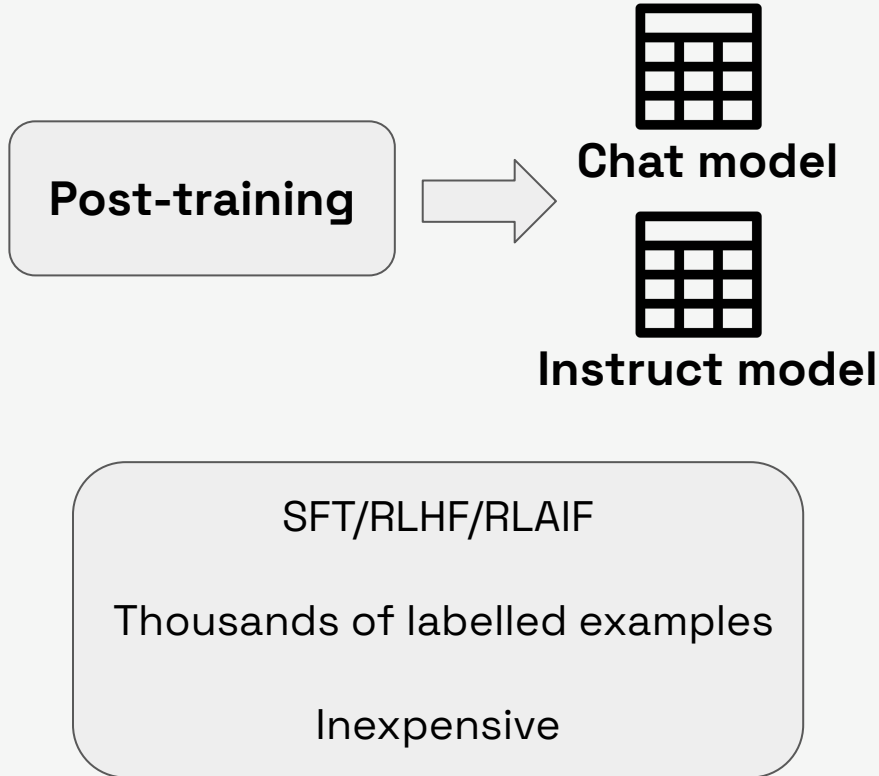
    def __contains__(self, other: str) -> bool:
        return self.exists(other)
```

## Academic

### Attention Is All You Need

<b>Ashish Vaswani*</b> Google Brain avaswani@google.com	<b>Noam Shazeer*</b> Google Brain noam@google.com	<b>Niki Parmar*</b> Google Research nikip@google.com	<b>Jakob Uszkoreit*</b> Google Research usz@google.com
<b>Llion Jones*</b> Google Research llion@google.com	<b>Aidan N. Gomez* †</b> University of Toronto aidan@cs.toronto.edu	<b>Lukasz Kaiser*</b> Google Brain lukaszkaizer@google.com	
<b>Illia Polosukhin* †</b> illia.polosukhin@gmail.com			

# Post-training (alignment)



## Instruction

**User:** Explain the concept of machine learning to a 10-year-old.

**Assistant:** Imagine you have a very smart robot friend named Robo...

## Conversational

**User:** Hi there!

**Assistant:** Hello! How can I assist you today?

**User:** I'd like to book a flight to Austin...

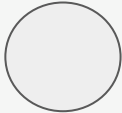
# Data volumes

---



**Pre-training**

**Billions of documents**  
**Trillions of tokens**  
**≈ 10-100M+ books**



**Post-training**

**Thousands of documents**  
**Millions of tokens**  
**≈ 10-100s of books**

# Pre-training

## Web pages

**Barack Hussein Obama II** ([/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/](#) (listen)); born August 4, 1961)<sup>[1]</sup> is an [American politician](#) and [attorney](#). He was the 44th [president of the United States](#) from 2009 to 2017. He was the first [African-American](#) president in U.S. history. As a member of the [Democratic Party](#), he also served as member of the [Illinois Senate](#) from 1997 to 2004 and a [United States senator](#) from [Illinois](#) from 2005 to 2008.

## Code

```
class Bloom:
    def __init__(self, size: int = 8) -> None:
        self.bitarray = 0b0
        self.size = size

    def add(self, value: str) -> None:
        h = self.hash_(value)
        self.bitarray |= h

    def exists(self, value: str) -> bool:
        h = self.hash_(value)
        return (h & self.bitarray) == h

    def __contains__(self, other: str) -> bool:
        return self.exists(other)
```

## Academic

### Attention Is All You Need

<b>Ashish Vaswani*</b> Google Brain avaswani@google.com	<b>Noam Shazeer*</b> Google Brain noam@google.com	<b>Niki Parmar*</b> Google Research nikip@google.com	<b>Jakob Uszkoreit*</b> Google Research usz@google.com
<b>Llion Jones*</b> Google Research llion@google.com	<b>Aidan N. Gomez* †</b> University of Toronto aidan@cs.toronto.edu	<b>Lukasz Kaiser*</b> Google Brain lukaszkaizer@google.com	
<b>Illia Polosukhin* †</b> illia.polosukhin@gmail.com			



# Post-training (alignment)

---

## Instruction

**User:** Explain the concept of machine learning to a 10-year-old.

**Assistant:** Imagine you have a very smart robot friend named Robo...

## Conversational

**User:** Hi there!


**Assistant:** Hello!  
How can I assist you today?

**User:** I'd like to book a flight to Austin...

# Data inputs into LLMs

---

## Pretraining:

**Barack Hussein Obama II** ([/bəˈrɑːk huːˈseɪn ʊˈbɑːmə/](#)  listen); born August 4, 1961)<sup>[1]</sup> is an [American politician](#) and [attorney](#). He was the 44th [president of the United States](#) from 2009 to 2017. He was the first [African-American](#) president in U.S. history. As a member of the [Democratic Party](#), he also served as member of the [Illinois Senate](#) from 1997 to 2004 and a [United States senator](#) from [Illinois](#) from 2005 to 2008.

## Post-training:

**User:** Explain the concept of machine learning to a 10-year-old.

**Assistant:** Imagine you have a very smart robot friend named Robo...

**User:** Hi there!

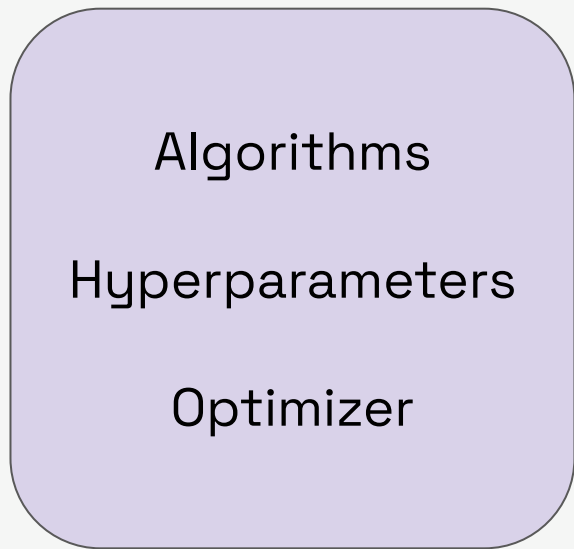
**Assistant:** Hello! How can I assist you today?

**User:** I'd like to book a flight to Austin...

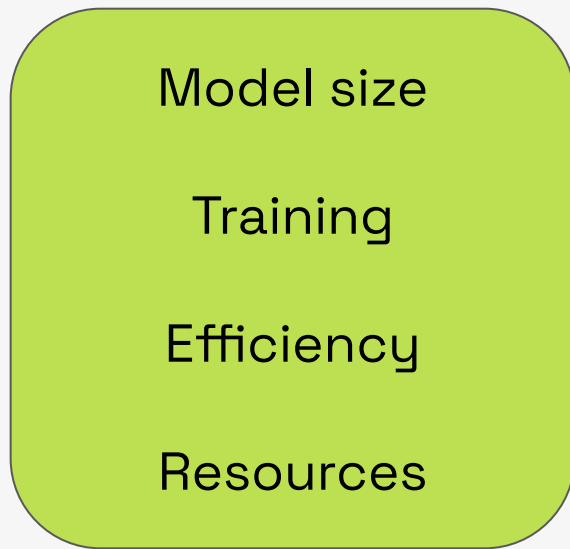
# Science of data for LLMs

# What determines the performance and capabilities of an LM?

---



**Model architecture**



**Compute**

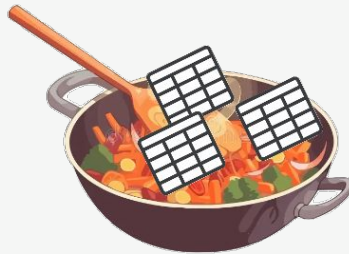


**Data**

# Why is data important?

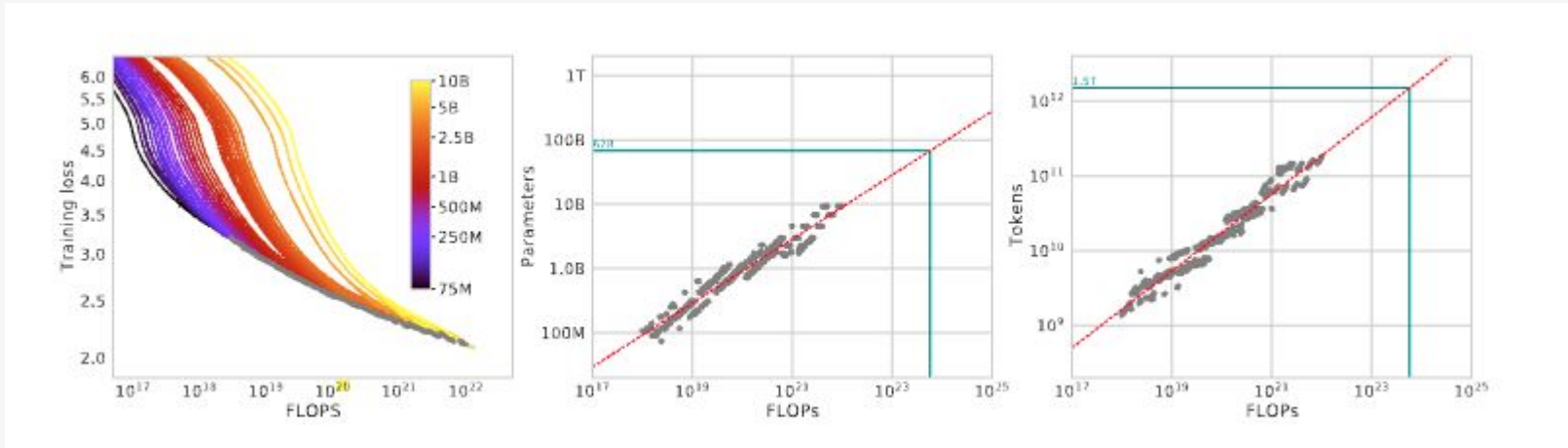
---

- Different model architectures trained on the same dataset will converge to the same point
- Performance is constrained by data more than anything else
- Data is the “secret sauce”



# Dataset size

- Model capabilities scale with data, parameters, and compute

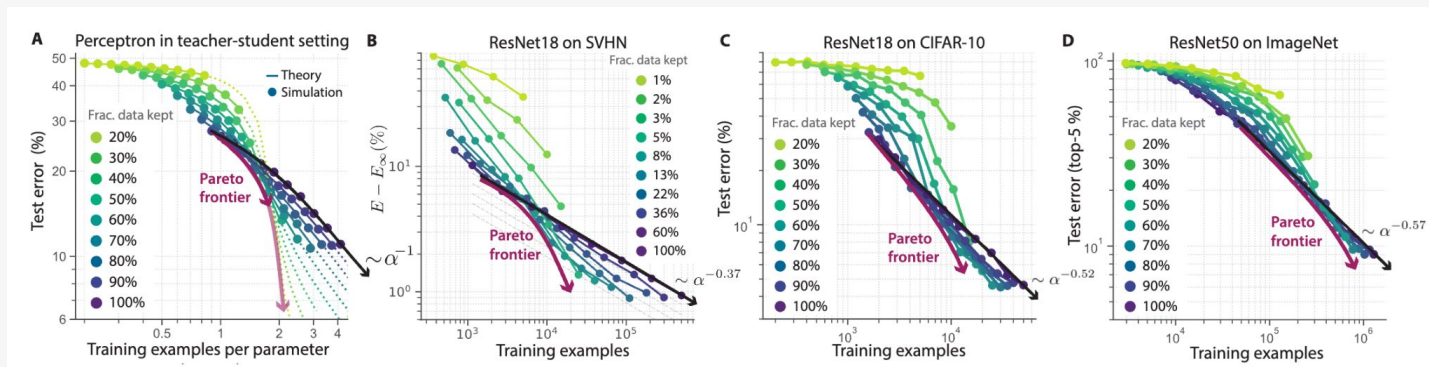


Training Compute-Optimal Large Language Models - Hoffmann et al. (2022)

- Recent models (LLaMa-2, Yi-34B, DBRX) have trained on even more tokens, seeing continued performance gains

# Data quality

- Clean, coherent, and relevant documents
- Faster convergence, better generalization, and higher overall ceiling
- Scaling laws may improve with higher quality data



Beyond neural scaling laws: beating power law scaling via data pruning - Sorscher et al. (2022)

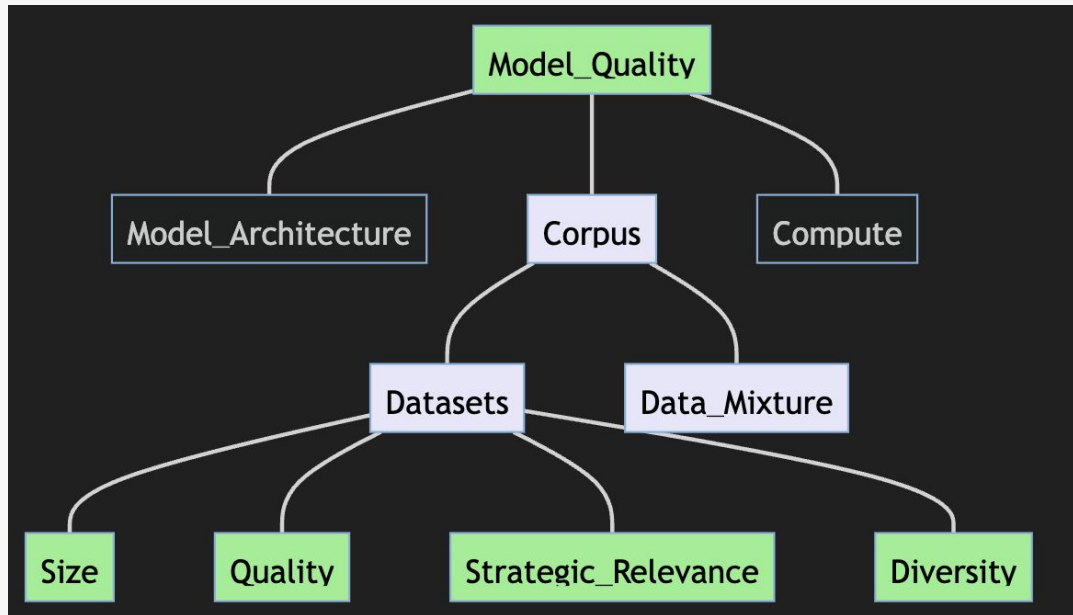
# Diversity

---

- Language encodes high dimensional information about the world
- The richness and diversity of language helps models learn robust and generalizable representations
- The *most* diverse dataset available is the internet (GPT-2+)
- Many types of diversity are important:
  - Linguistic
  - Domain
  - Task



# Putting it all together

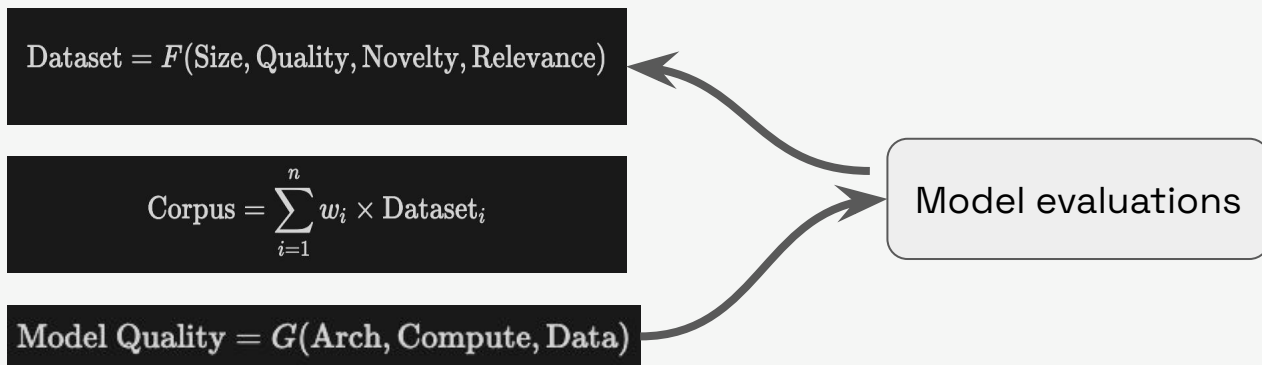


$$\text{Dataset} = F(\text{Size}, \text{Quality}, \text{Novelty}, \text{Relevance})$$

- Model quality does not monotonically improve with new data
- Small datasets may be more valuable than large datasets if they are diverse or high quality

## Evaluating the impact of a dataset

- Ablations and evals provide helpful feedback on corpus changes
- Can we predict the impact of a dataset on model performance...*before* training on it?



Kind of?

# Lifecycle of a pre-training dataset

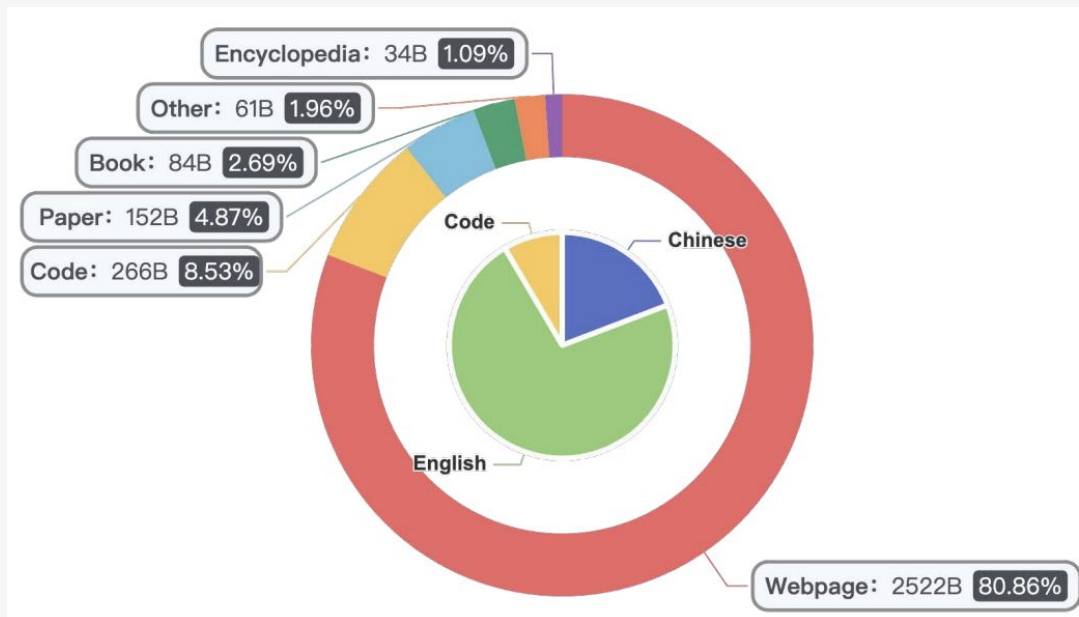
# What's in a corpus?

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

**GPT-3**

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

**LLaMa-1**



**Yi-34B**

# Data lifecycle

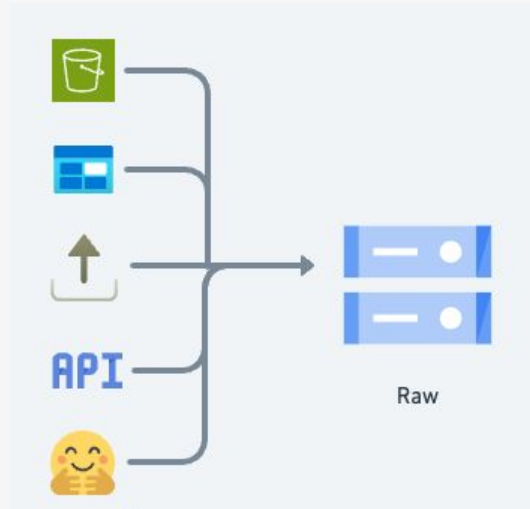
---



# Ingestion

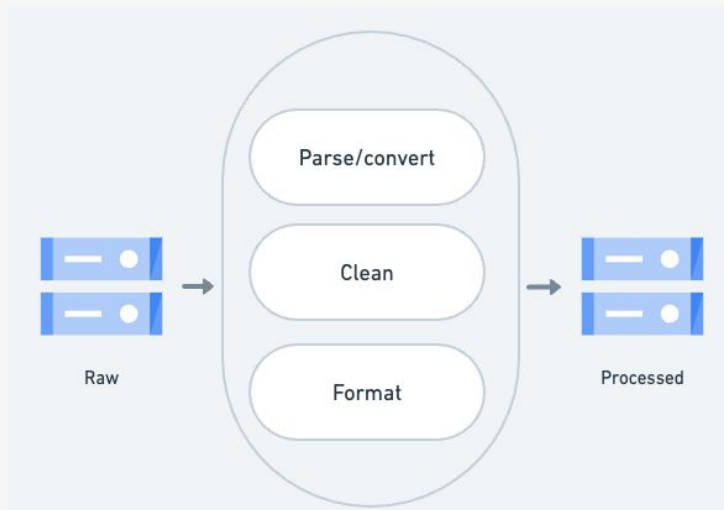
---

- Raw data is loaded into storage without any processing



# Transformation

- Parse out the purest possible version of a document
- Remove artifacts that not helpful for training
- Format documents into the desired structure



3 Beijing Institute of Agricultural Machinery was founded in 1959. After being transformed into a scientific and technological enterprise in 2000, it has formed its own advantages in the field of research and development of facility agricultural engineering and industrial development. The level of modern agricultural equipment marked by greenhouse and animal husbandry engineering of facility agriculture is in the leading position in the field of agriculture in China, and has played a leading role. The technological supporting role and demonstration driving role of modern agricultural development have supported the process of agricultural industrialization in the capital and the whole country

4

5 About

6 |

7 News

8 |

9 Innovation Center

10 |

11 Major Business

12 |

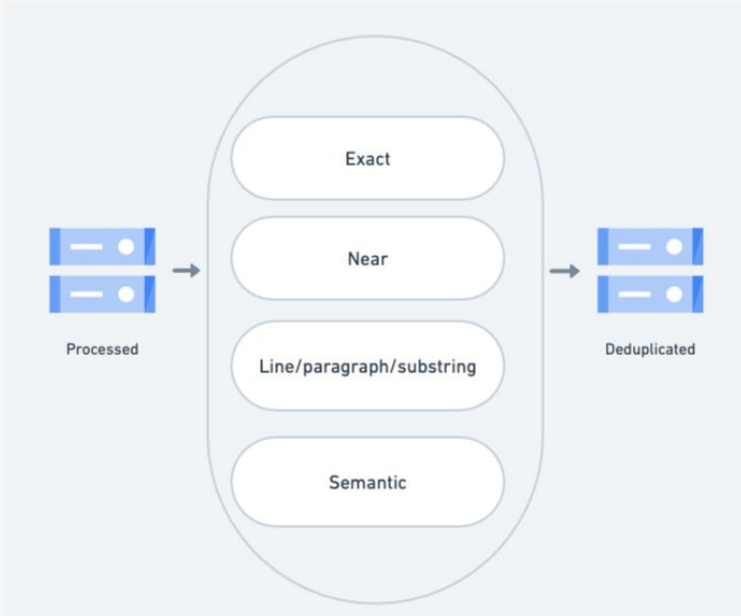
13 Product Services

14 |

15 Party building

1 Beijing Institute of Agricultural Machinery was founded in 1959. After being transformed into a scientific and technological enterprise in 2000, it has formed its own advantages in the field of research and development of facility agricultural engineering and industrial development. The level of modern agricultural equipment marked by greenhouse and animal husbandry engineering of facility agriculture is in the leading position in the field of agriculture in China, and has played a leading role. The technological supporting role and demonstration driving role of modern agricultural development have supported the process of agricultural industrialization in the capital and the whole country

# Deduplication



- Essential for web crawl data

- 1 Warning! PostcardPerfect.net has expired.
- 2 If this is your domain name you must renew it immediately before it is deleted and permanently removed from your account.
- 3 To renew this domain name visit <http://www.NameBright.com>

- 1 Warning! Lu-Tang.com has expired.
- 2 If this is your domain name you must renew it immediately before it is deleted and permanently removed from your account.
- 3 To renew this domain name visit <http://www.NameBright.com>

**35M documents!**



## Duplication in web crawls

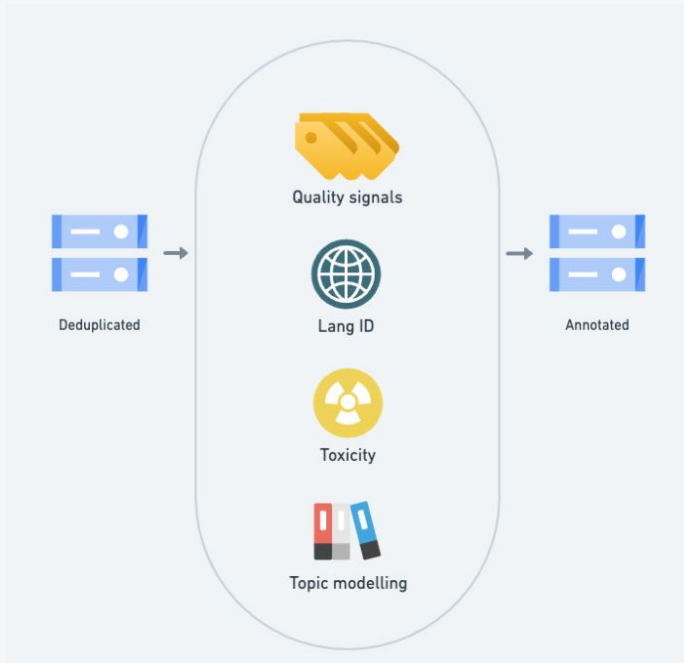
---

- 1 Warning! PostcardPerfect.net has expired.
- 2 If this is your domain name you must renew it immediately before it is deleted and permanently removed from your account.
- 3 To renew this domain name visit <http://www.NameBright.com>

- 1 Warning! Lu-Tang.com has expired.
- 2 If this is your domain name you must renew it immediately before it is deleted and permanently removed from your account.
- 3 To renew this domain name visit <http://www.NameBright.com>

# Annotation

- Label data with quality signals and classification metadata



- What is quality?
  - Ingestion/processing artifacts
  - Boilerplate/templated content
  - Coherence
  - “Low value” content

## Quality signals

---

Linguistic

Content

Repetition

**Text heuristics**

Perplexity

Classifiers

Importance  
resampling

**ML/NLP heuristics**

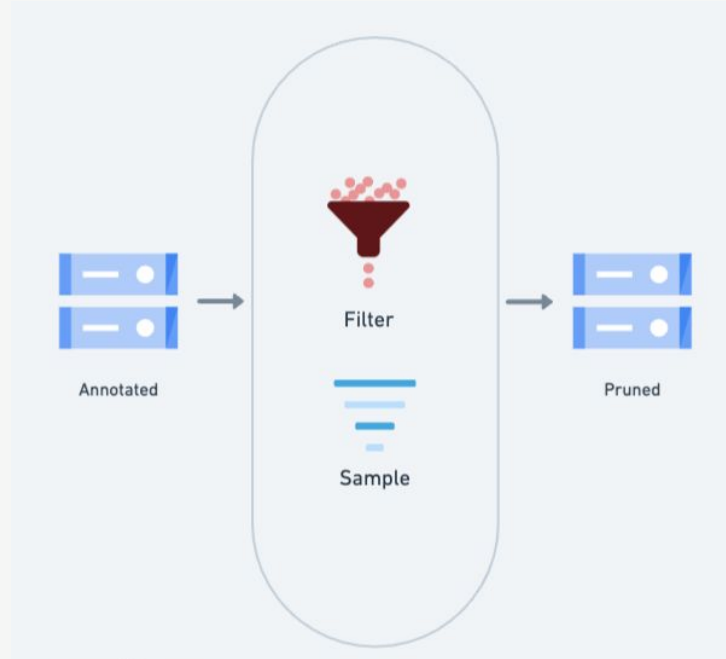
Prompt  
engineering

Finetunes

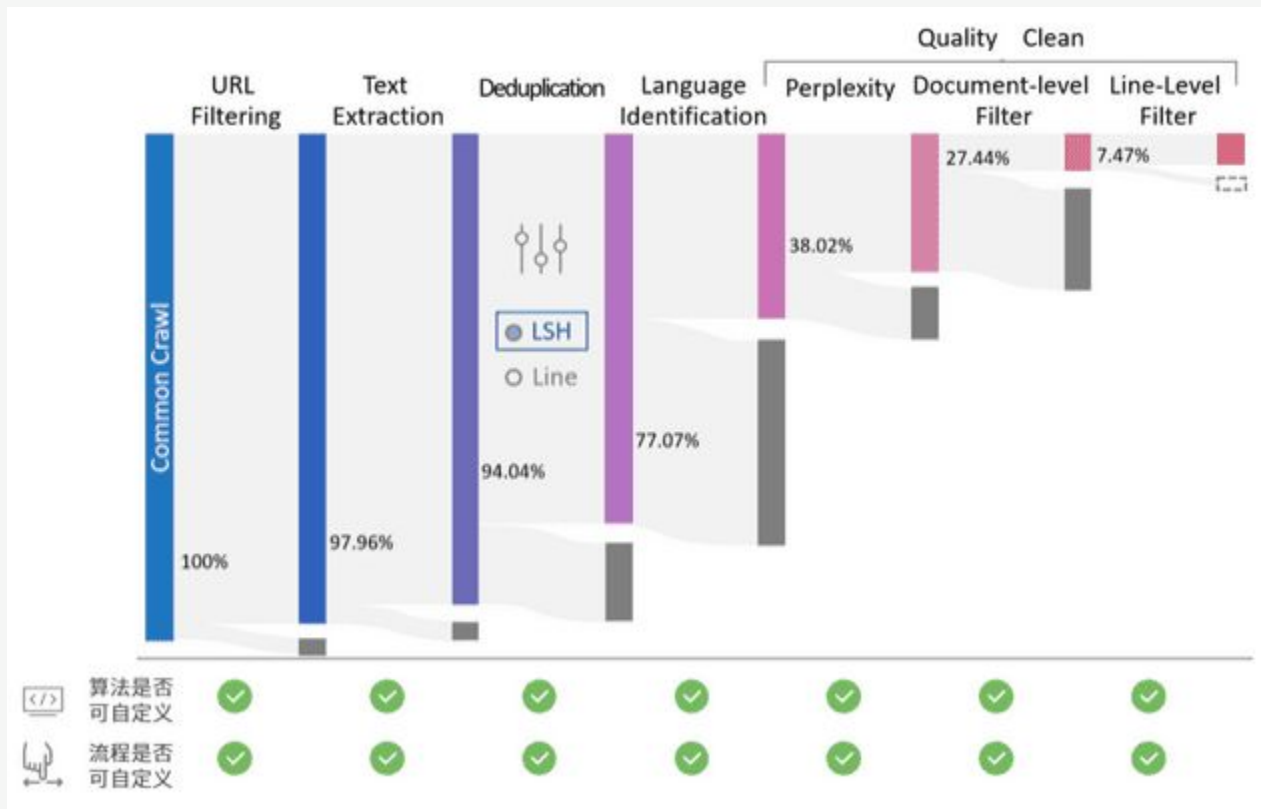
**LLM evaluators**

# Pruning

- Combination of filtering and sampling at document level



# Case study - DeepSeek 70B



## Challenges

---

- Working with web crawl data
- Source data may be multiple petabytes
- Finding the right signals for pruning
- Processing PDFs

# Cohere data platform

## Different needs than “traditional” platforms

---

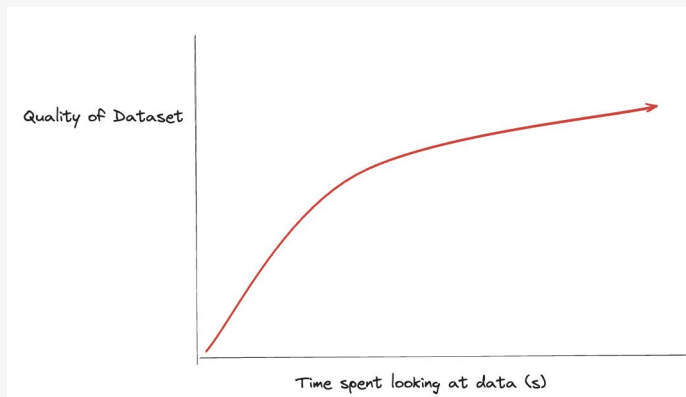
- Primarily unpartitioned, unstructured text and multimodal data
- Pipelines run semi-frequently, sometimes only once
  - Data acquisition jobs may run continuously
- Reliance on UDFs for text processing
- GPU acceleration sometimes needed



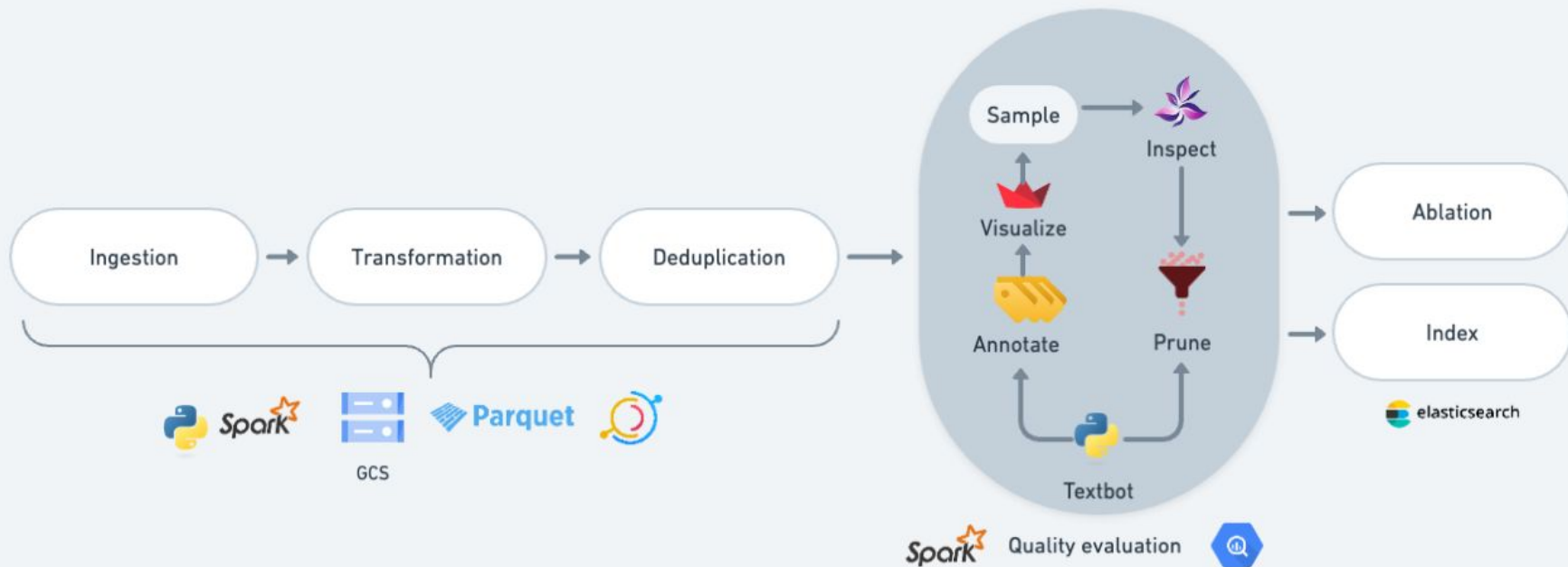
## Guiding principles

---

- Be as non-destructive as possible when processing data
- Create reusable components
- Simplify and accelerate data quality engineering
- Help people become “one with the data”



# Architecture



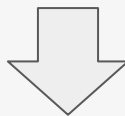
# Quality evaluation

---

- Textbot: Signal library and scalable compute engine

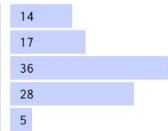
```
from textbot.analyzer import Analyzer

analyzer = Analyzer(signals=["doc_num_tokens", "doc_lang_id"])
analyzer.analyze(["What is the capital of Texas?"])
```



```
{
  "doc_num_tokens": 6,
  "doc_lang_id": "en",
}
```

# Document introspection

100 rows	ⓘ												
abc raw	▼ ⓘ												
abc content	▼ ⓘ												
123 document_len	▲ ⓘ												
ⓘ Total count	100												
ⓘ Unique values	~96												
ⓘ Range	69 .. 58,711												
 <table border="1"><thead><tr><th>Range</th><th>Count</th></tr></thead><tbody><tr><td>[-1 .. 130)</td><td>14</td></tr><tr><td>[130 .. 1,800)</td><td>17</td></tr><tr><td>[1,800 .. 9,000)</td><td>36</td></tr><tr><td>[9,000 .. 30,000)</td><td>28</td></tr><tr><td>[30,000 .. 80,000)</td><td>5</td></tr></tbody></table>		Range	Count	[-1 .. 130)	14	[130 .. 1,800)	17	[1,800 .. 9,000)	36	[9,000 .. 30,000)	28	[30,000 .. 80,000)	5
Range	Count												
[-1 .. 130)	14												
[130 .. 1,800)	17												
[1,800 .. 9,000)	36												
[9,000 .. 30,000)	28												
[30,000 .. 80,000)	5												
123 __index_level_0__	▼ ⓘ												
123 year	▼ ⓘ												

Filters | Group by | Sort

1 of 100

content

Supreme Court of Alaska.  
2  
3 No. 6878.  
4  
5 **\*\*In re the Disciplinary Matter Involving Michelle V. MINOR, Respondent Attorney.\*\***  
6 Feb. 10, 1983.  
7  
8 Before BURKE, C.J., and RABINOWITZ, MATTHEWS and COMPTON, JJ.  
9  
10 CONNOR, J., not participating.  
11  
12 Paul L. Davis and Lee Holen, Boyko & Davis, Anchorage, for respondent.  
13  
14 Richard J. Ray, Anchorage, for Alaska Bar Association.  
15  
16 OPINION  
17  
18 COMPTON, Justice.  
19  
20 The present case is before this court pursuant to Alaska Bar Rule II IS(j). At issue is whether the Disciplinary Board of the Alaska Bar Association properly concluded that Michelle Minor, an attorney licensed to practice law

Supreme Court of Alaska.  
No. 6878.  
\* \* \*

**OPINION.** Justice.

The present case is before this court pursuant to Alaska Bar Rule II IS(j). At issue is whether the Disciplinary Board of the Alaska Bar Association properly concluded that Michelle Minor, an attorney licensed to practice law in this state, was guilty of professional misconduct as defined by Alaska Bar Rule II IS(j).

The Disciplinary Board of the Alaska Bar Association (the "Board") presented this case to the court on October 10, 1983. The Board's complaint was captioned with Alaska Bar Rule II IS(j) as the authority for its jurisdiction. The Board's complaint charged that Michelle Minor, Respondent Attorney, had committed professional misconduct by failing to disclose to her client, Paul L. Davis and Lee Holen, Boyko & Davis, Anchorage, Alaska, the fact that she had been suspended from the practice of law by the State Bar of Michigan on October 28, 1982.

On November 16, 1983, after conducting a trial in the Supreme Court of Alaska, the court rendered its decision in this case. The court held that Michelle Minor had committed professional misconduct as defined by Alaska Bar Rule II IS(j) and suspended her from the practice of law for 12 months. The court also held that the Board's decision to suspend Michelle Minor was supported by the weight of the evidence. The court affirmed the Board's decision to suspend Michelle Minor.

The court's decision is supported by the following findings of fact:

1. Michelle Minor is a member of the State Bar of Michigan and is a member of the State Bar of Alaska. Michelle Minor is a member of the State Bar of Michigan and is a member of the State Bar of Alaska.

2. Michelle Minor was suspended from the practice of law by the State Bar of Michigan on October 28, 1982.

3. Michelle Minor was not disclosed to her client, Paul L. Davis and Lee Holen, Boyko & Davis, Anchorage, Alaska, that she had been suspended from the practice of law by the State Bar of Michigan on October 28, 1982.

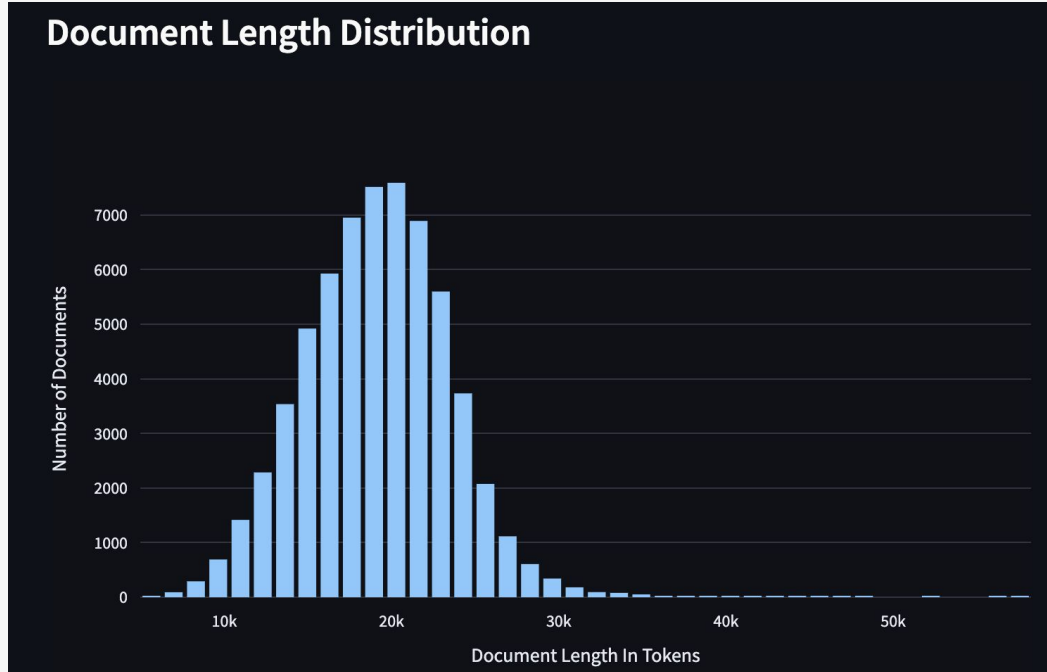
4. Michelle Minor's failure to disclose to her client that she had been suspended from the practice of law by the State Bar of Michigan on October 28, 1982, constituted professional misconduct as defined by Alaska Bar Rule II IS(j).

5. The Board's decision to suspend Michelle Minor was supported by the weight of the evidence.

6. The court affirmed the Board's decision to suspend Michelle Minor.

# Dataset-wide analytics

---



# Other infrastructure

- Data lineage tracking with Datahub
- Indexing and searching across data

The screenshot shows the Nautilus web interface for a Pretraining Keyword Search. The search query is "frequently used in transformation of eukaryotes are". The search results table has two columns: "\_index" and "content". The first result is at index 0, and the second is at index 3. The content of the second result is a paragraph of text describing genes and their functions.

**Nautilus**

SQL Query Executor [Pretraining Keyword Search](#) [Command Keyword Search](#) [Metadata](#) [Datasets](#)

## Pretraining Keyword Search

Search pretraining data:

frequently used in transformation of eukaryotes are

Exact string match?

Run pretraining search

Unique indices:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

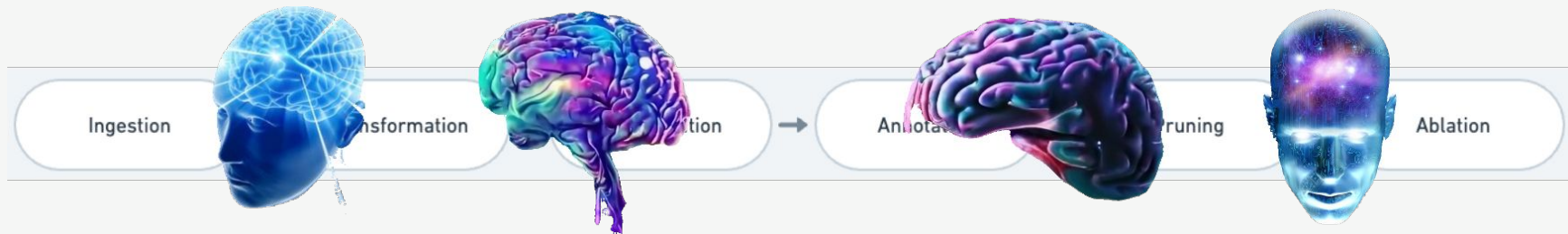
11 22 33 44 55 66 77 88 99 110 121 132 143 154 165 176 187 198 209 220 231 242 253 264 275 286 297 308 319 330 341 352 363 374 385 396 407 418 429 440 451 462 473 484 495 506 517 528 539 550 561 572 583 594 605 616 627 638 649 660 671 682 693 704 715 726 737 748 759 770 781 792 803 814 825 836 847 858 869 880 891 902 913 924 935 946 957 968 979 990

Search Results:

_index	content
0	
3	The genes frequently used in transformation of eukaryotes are neomycin phosphotransferase, which confers resistance to the aminoglycosides kanamycin and G418; histidinol dehydrogenase, which confers resistance to L-histidinol in a histidine lacking medium; hygromycin phosphotransferase, which confers resistance to hygromycin B; and sh ble gene from Streptococcus hindustanus, which confers resistance to bleomycin-philomycin antibiotics. These antibiotics are expensive to use in large quantities and the presence of these antibiotics is required to keep the resistant cell lines under constant selection to induce the propagation of the gene conferring antibiotic resistance.

## Lessons learned

- Treat training data like code (i.e. use peer review)
- Make inspecting, labelling, and visualizing data frictionless
- Create evals to help you make data-driven decisions
- Inject intelligence into your data processing pipelines



# Acknowledgements

---

- Thank you to my team at Cohere!
- Open source projects:
  - RedPajama
  - DataJuicer
  - Datatrove
  - Dolma
  - CCnet
- **@georgejrjr** for an excellent primer: “Datasets of models to come”







March 28, 2024

# Thank you!

Follow [@jtalms](https://twitter.com/jtalms) on



Jonathan Talmi