Data Science for LLMs

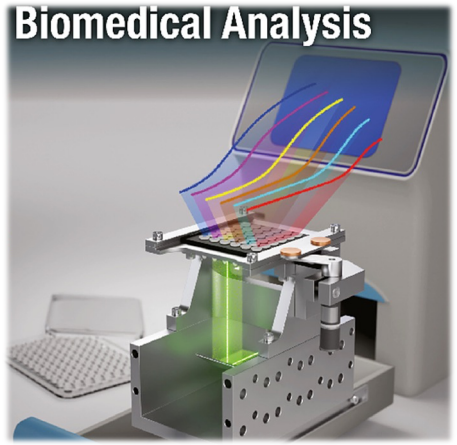
https://github.com/weAIDB/awesome-data-llm

Data: The Fuel of LLM Development

➢ LLMs have been widely used, which require **vast, high-quality** data, distributing across **diverse** domains

LLM Applications

General Domains



Specific Domains

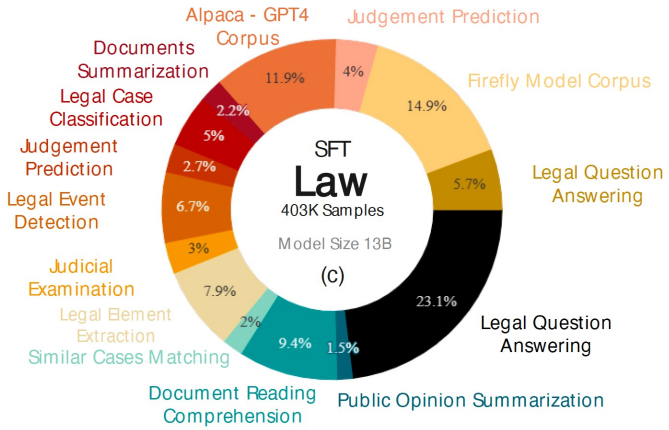
LLM Datasets

Qwen-2VL → *Qwen-2.5VL*

1.2 → **4 trillion** tokens

(with over **10,000** object categories for visual grounding)

DISC-Law-SFT



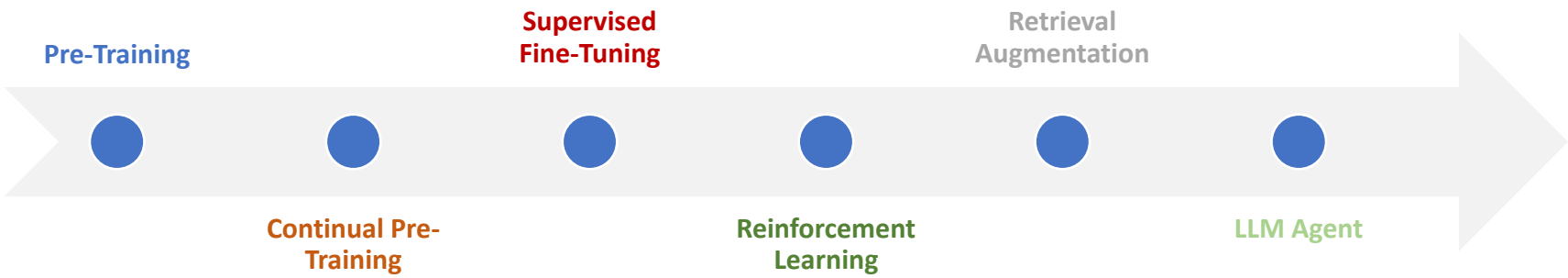
Data Size ↑

Data Quality ↑

A Survey of LLM × DATA. https://github.com/weAIDB/awsome-data-llm 🔥

Question to Ask:

What data is required for different LLM stages?



Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

1. Pre-Training

- 1T+ unlabeled samples

2. Continual Pre-Training

- 10M+ unlabeled samples

3. Supervised Finetuning

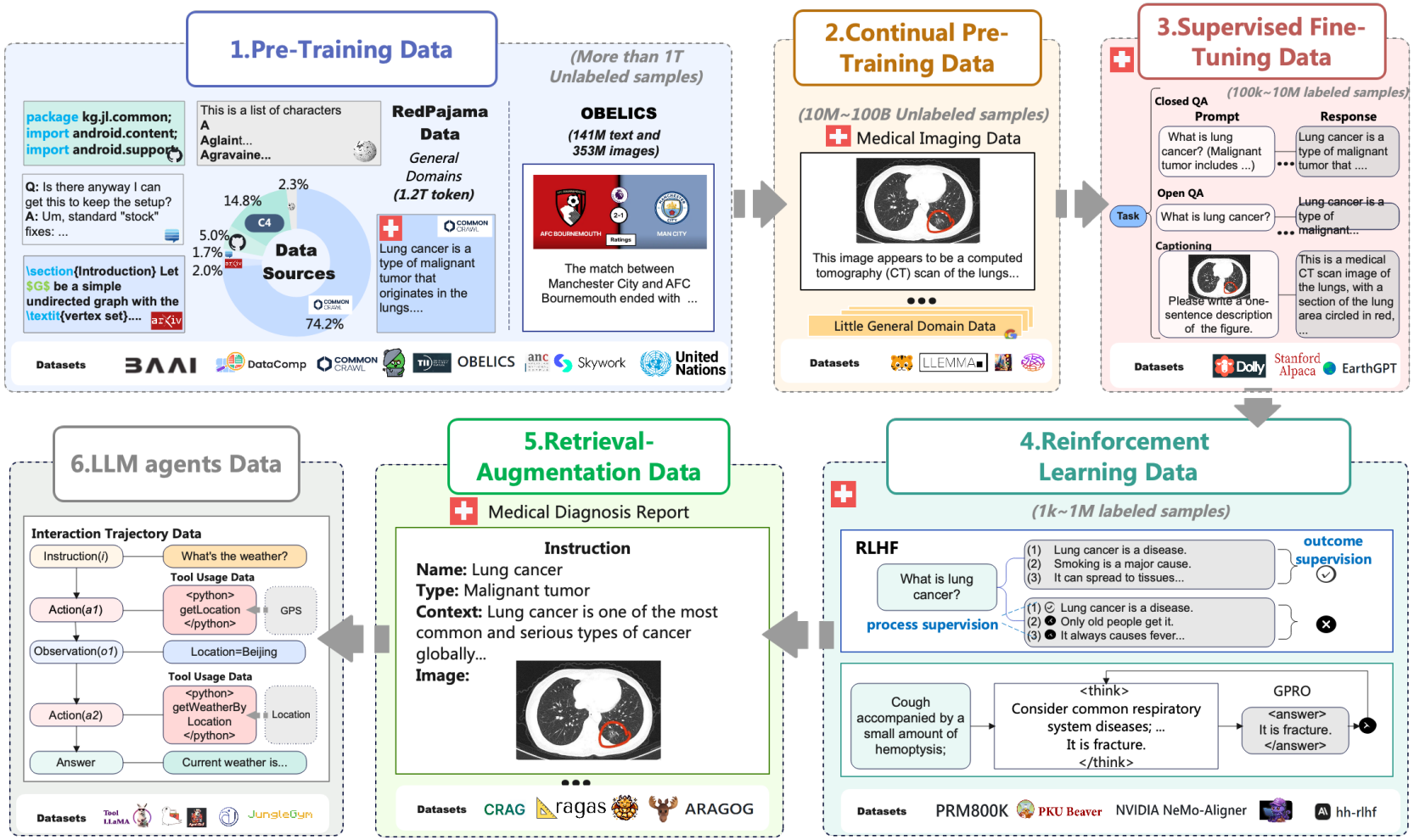
- 10k+ labeled samples

4. Reinforcement Learning

- 1k+ labeled samples

5. RAG

6. Agentic LLM

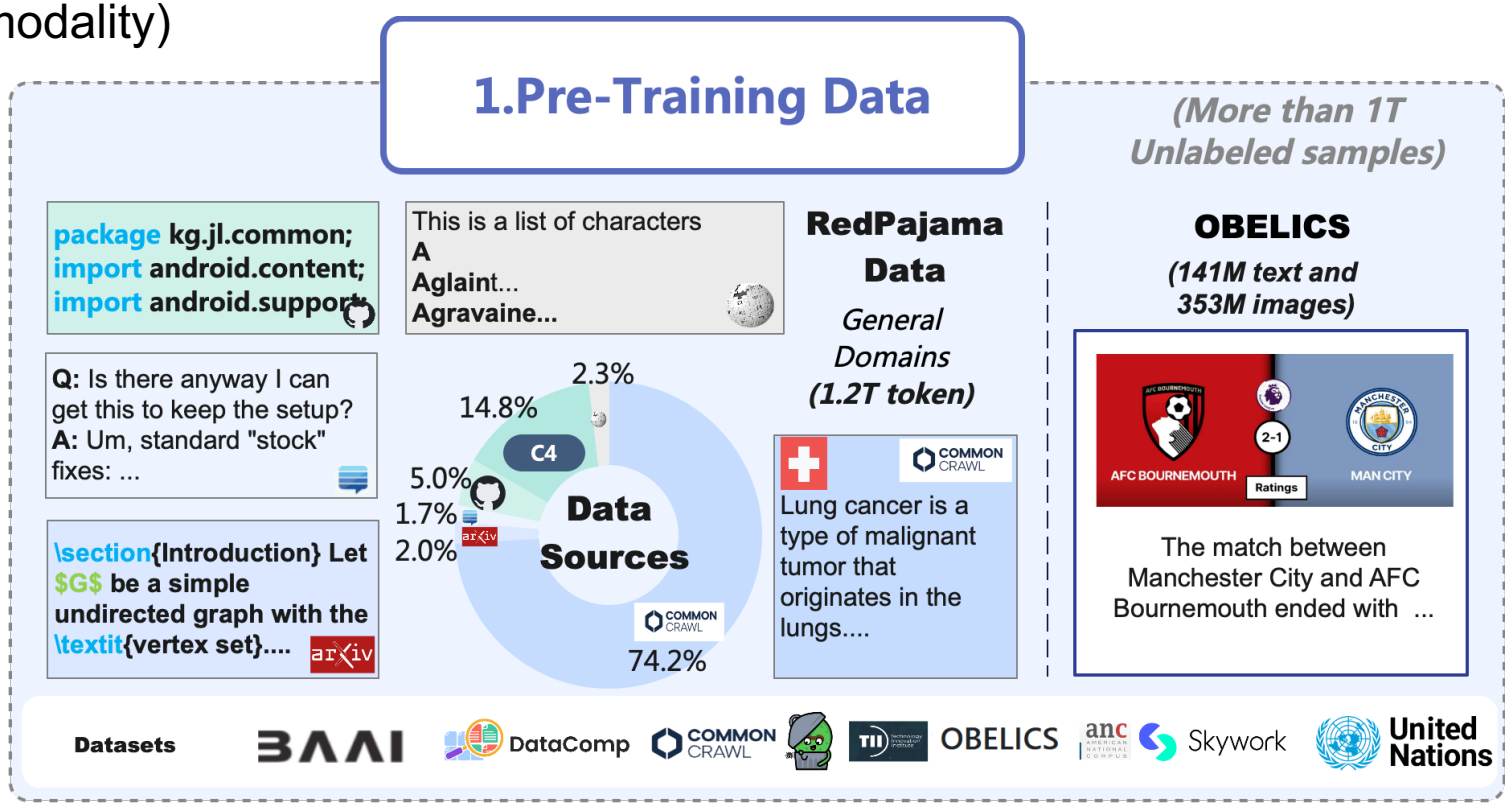


Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

1. Pre-Training (1T+ unlabeled samples)

- **Data Demand:**
  - Acquire broad language (and cross-modality) understanding;
  - Reduce the risk of overfitting.
- **Sources:**
  - Web crawls (e.g., HTML, WARC)
  - Open code repository
  - Books (text, EPUB)
  - Academic papers
  - Interleaved image-text

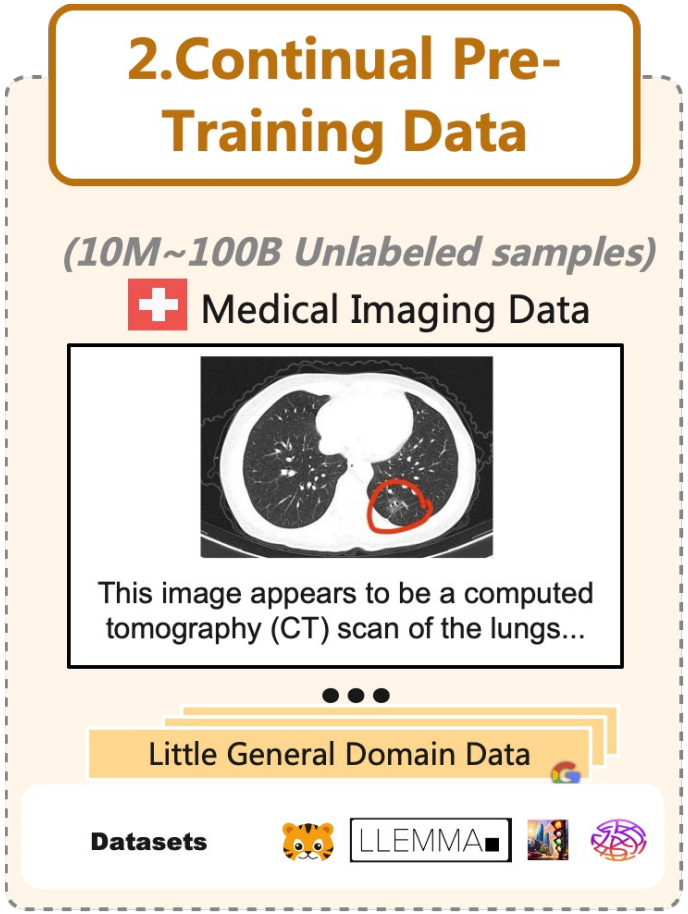


Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

2. Continual Pre-Training (10M+ unlabeled samples)

- **Data Demand:**
  - Fill knowledge gaps
  - Adapt the model to specific domains
- **Examples:**
  - BBT-FinCorpus: 300 GB of finance data
  - Medical-pt: 360,000 Chinese-English entries from medical encyclopedias

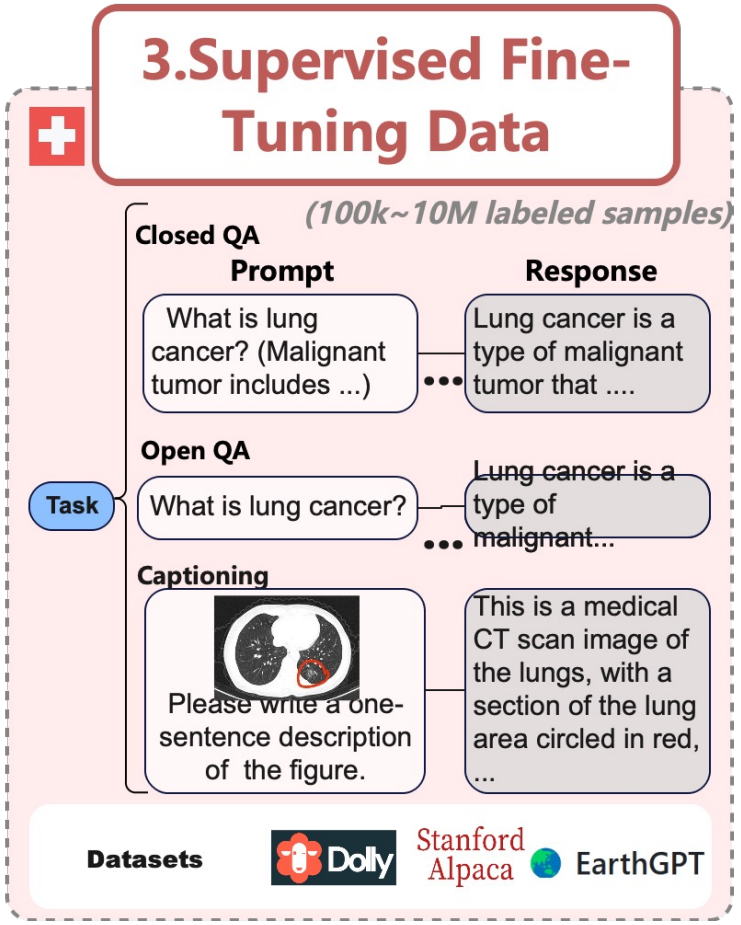


Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

3. Supervised Finetuning (10k+ labeled samples)

- **Data Demand:** Guide the model in learning a specific, narrower set of tasks
- **General Instruction Following**
  - Databricks-dolly-15K with 7 types of tasks, e.g., closed QA, summarization, information extraction
- **Specific Task Resolving**
  - DISC-Law-SFT: Legal information extraction (32k), legal judgment prediction (16k), legal event detection (27k), and legal question-answering (93k)

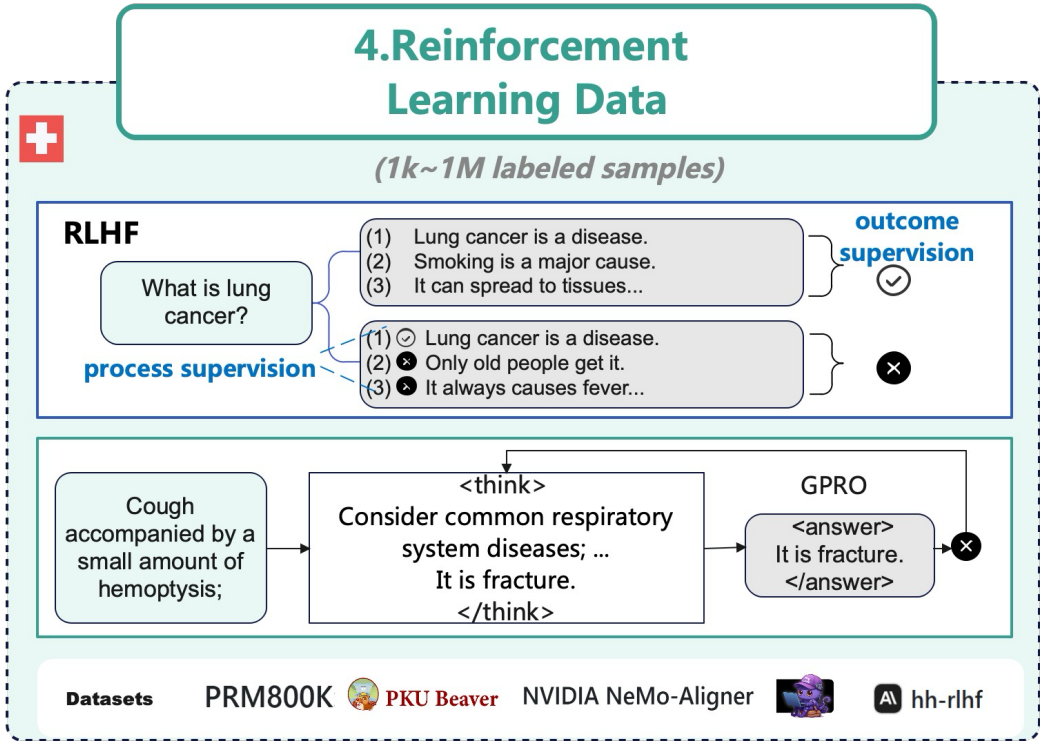


Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

4. Reinforcement Learning (1k+ labeled samples)

- **RLHF (Reinforcement Learning with Human Feedback)**
  - **Data:** Smaller than SFT's with more complex data annotations (compare and rank multiple candidate responses by *human preference*)
- **Reasoning-oriented Reinforcement Learning (RoRL)**
  - **Data:** Only feedback on whether the answer is correct or not (for long-term reasoning)

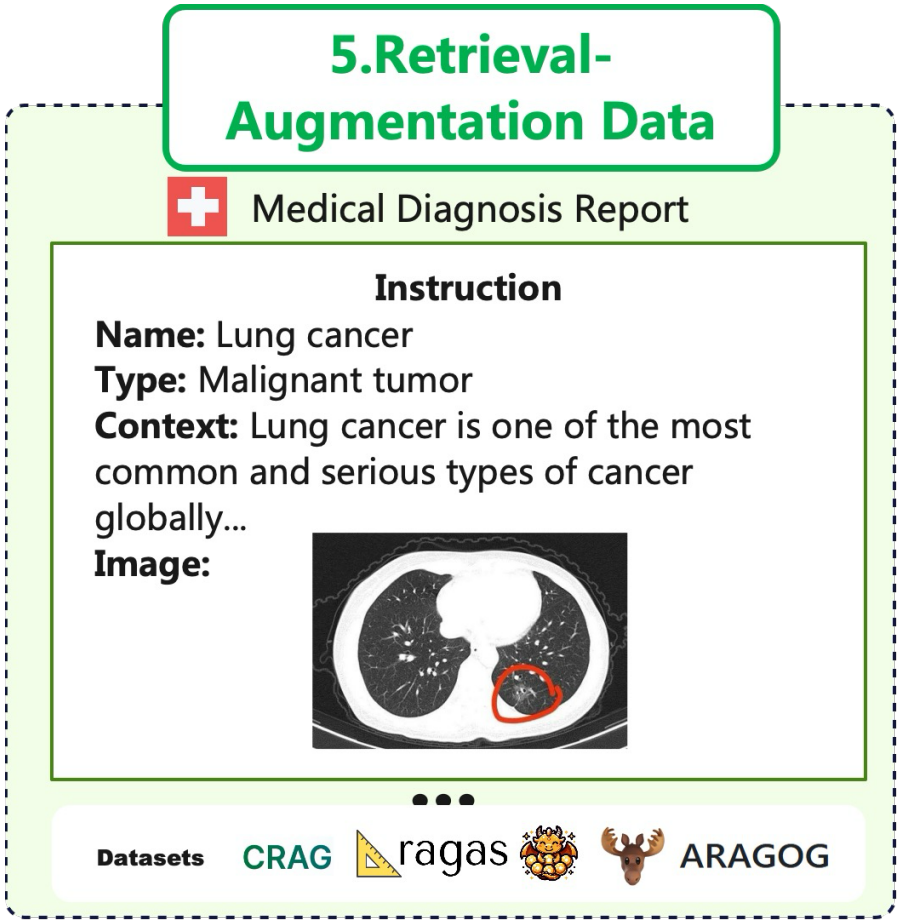


Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

5. RAG

- **Data Demand:** Strictly reviewed to ensure authenticity and validity, while dynamic data requires real-time updates
- **Examples**
  - **Medical:** Data from over 65,000 ICU patients and more than 200,000 patients treated in emergency department
  - **Legal:** 800+ national and local laws, regulations, and rules, aswell as 24,000 legal-related exam questions.
  - **User-personalized LLM**

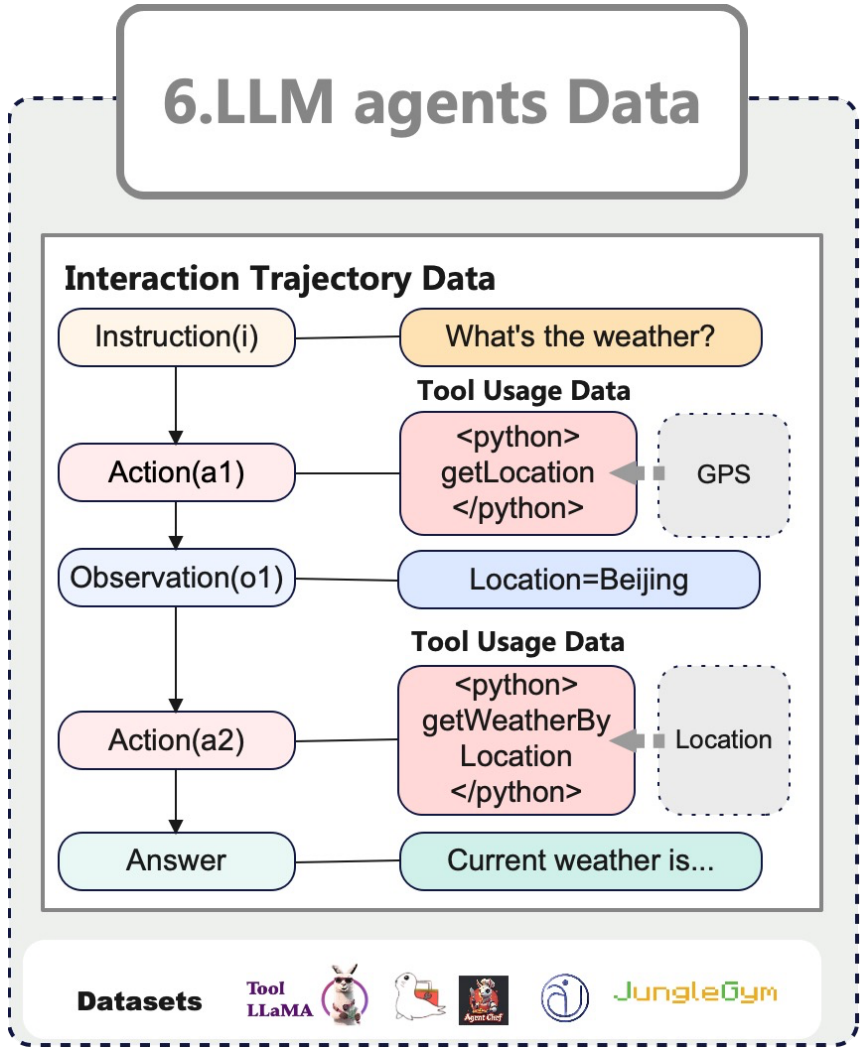


Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

6. Agent

- **Data Demand:** Require training data for advanced capabilities such as planning, tool orchestration, and multi-turn dialogue capability
- **Examples**
  - **Reasoning (UltraInteract):** Instruction as the root node; Both the correct actions and corresponding incorrect actions as nodes to construct a trajectory tree of human preference
  - **Tool (AutoTools):** Finetune on tool data (<python>code</python>)
  - **Dialogue (UltraChat):** Add an LLM to simulate user instructions and conversational content

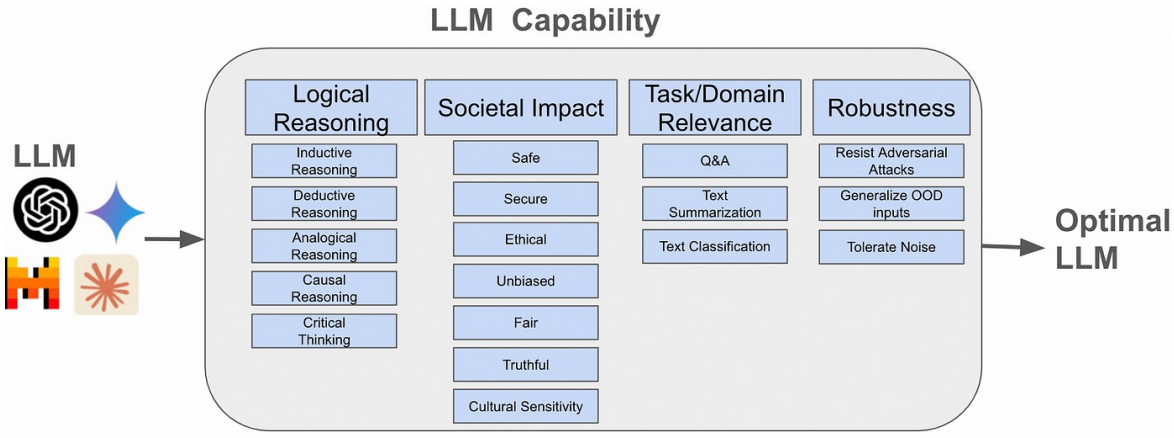


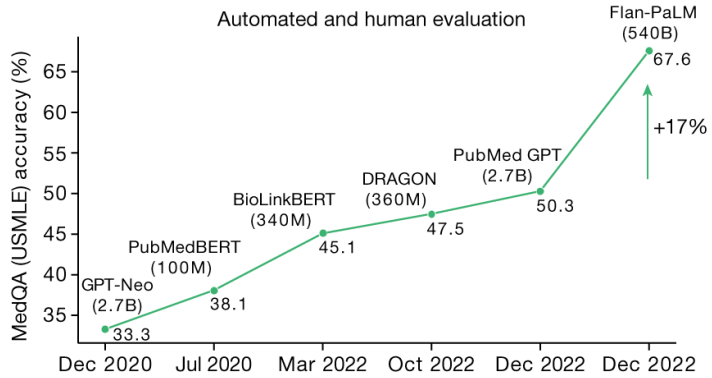
Data: The Fuel of LLM Development

➢ Different LLM stages have divese dataset requirements

7. Evaluation

- **Data:** Representative data samples that reflect different aspects of an LLM's capabilities.
- **General Domain**
  - MMMU: Perception, knowledge, and reasoning
- **Specific Domain**
  - HumanEval: 164 programming problems
  - MedQA: 61,097 medical exam questions from various regions





Q: How long does it take for newborn jaundice to go away?

Med-PaLM: Newborn jaundice is when a newborn baby's skin and eyes look yellow. It is caused by a buildup of a substance called bilirubin in the baby's blood. Bilirubin is a byproduct of the normal process of red blood cells breaking down. Newborn jaundice is common and typically harmless, but it can be a sign of a more serious condition. The jaundice typically goes away on its own within a few weeks. In some cases, it may last longer or require treatment. The time it takes for the jaundice to go away can vary depending on the cause and the severity of the jaundice. If the jaundice is severe or lasts longer than a few weeks, the doctor may recommend testing or treatment to determine the cause and prevent complications.

Question to Ask:

*What Makes a **Dataset** Truly "Good" for LLMs?*

Inclusiveness

Abundance

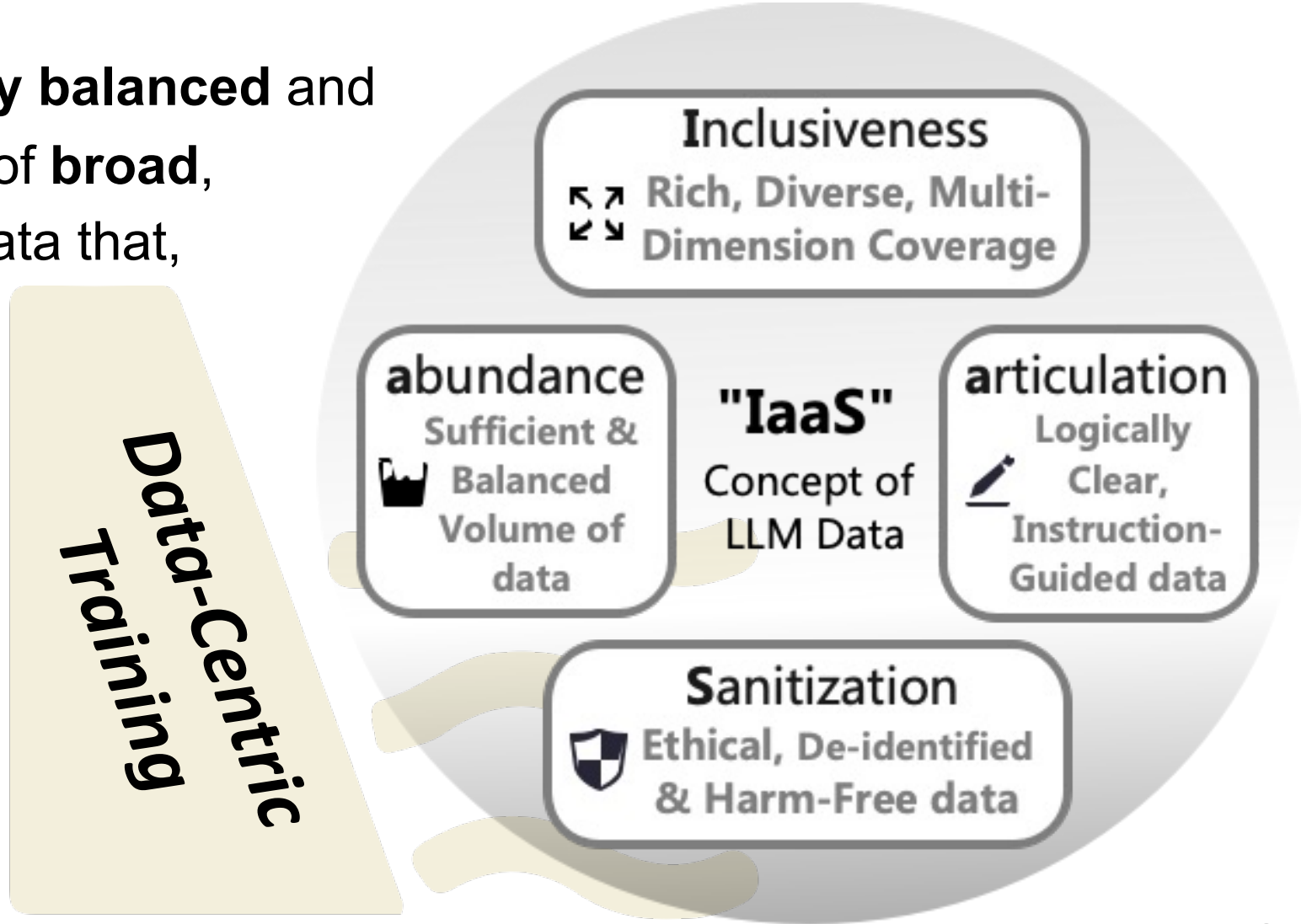
Articulation

Sanitization

The IaaS Concept of DATA4LLM

➢ The *IaaS* Concept

A good dataset is a **purposefully balanced** and **rigorously sanitized** collection of **broad**, **diverse**, and **well-articulated** data that, when used for training, yields **high-performing**, **safe**, and **resource-efficient** large language models.



The IaaS Concept of DATA4LLM

The ***IaaS*** Concept can be categorized into four keywords:

1. Inclusiveness

- Domains; Task Types; Sources; Languages;
- Expression Styles; Data Modalities

2. Abundance

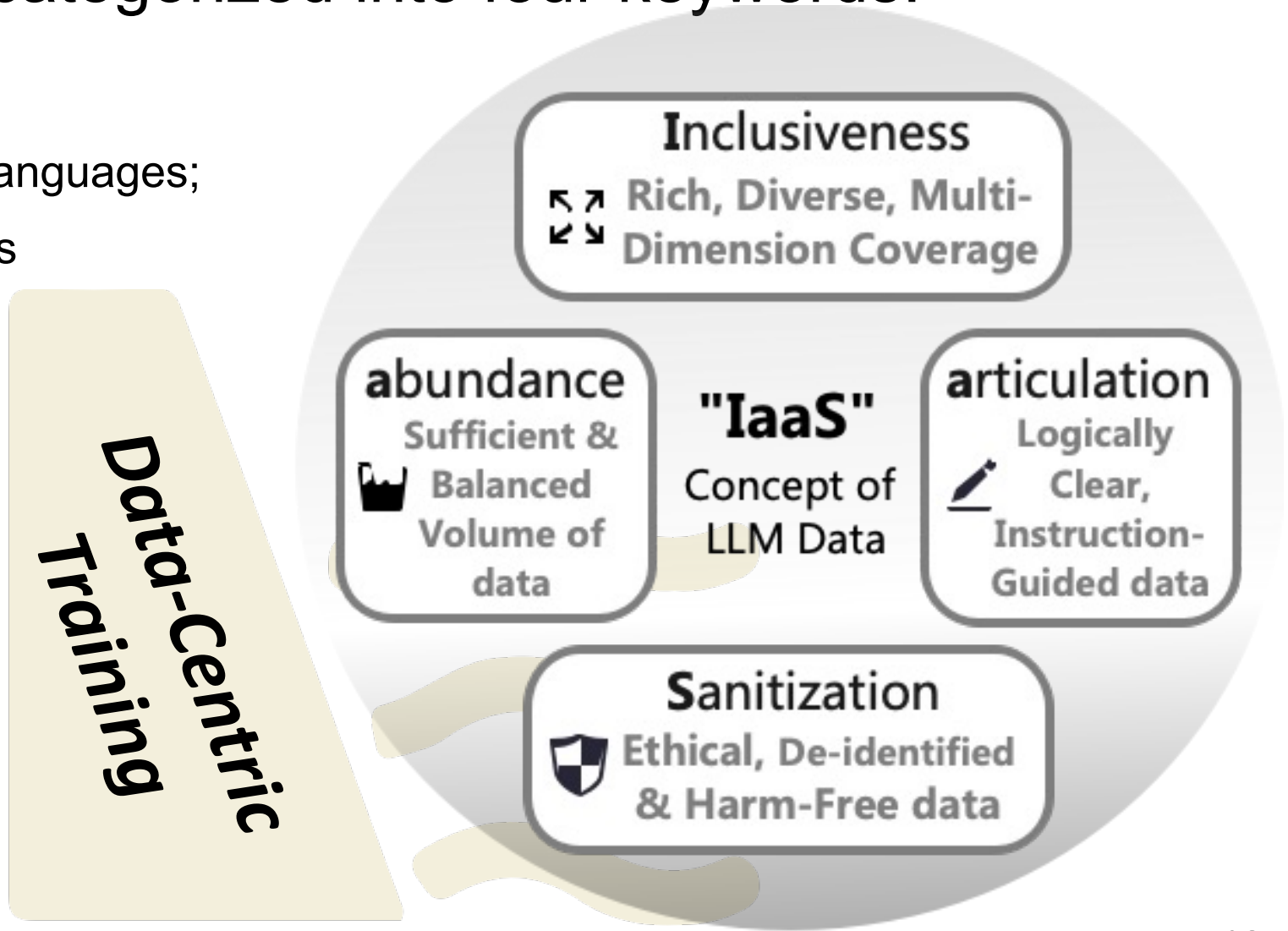
- Enhance domain-specific ability

3. Articulation

- Well-Formatted; Instructive;
- Step-by-step Reasoning

4. Sanitization

- Privacy; Ethical; Risk Removal



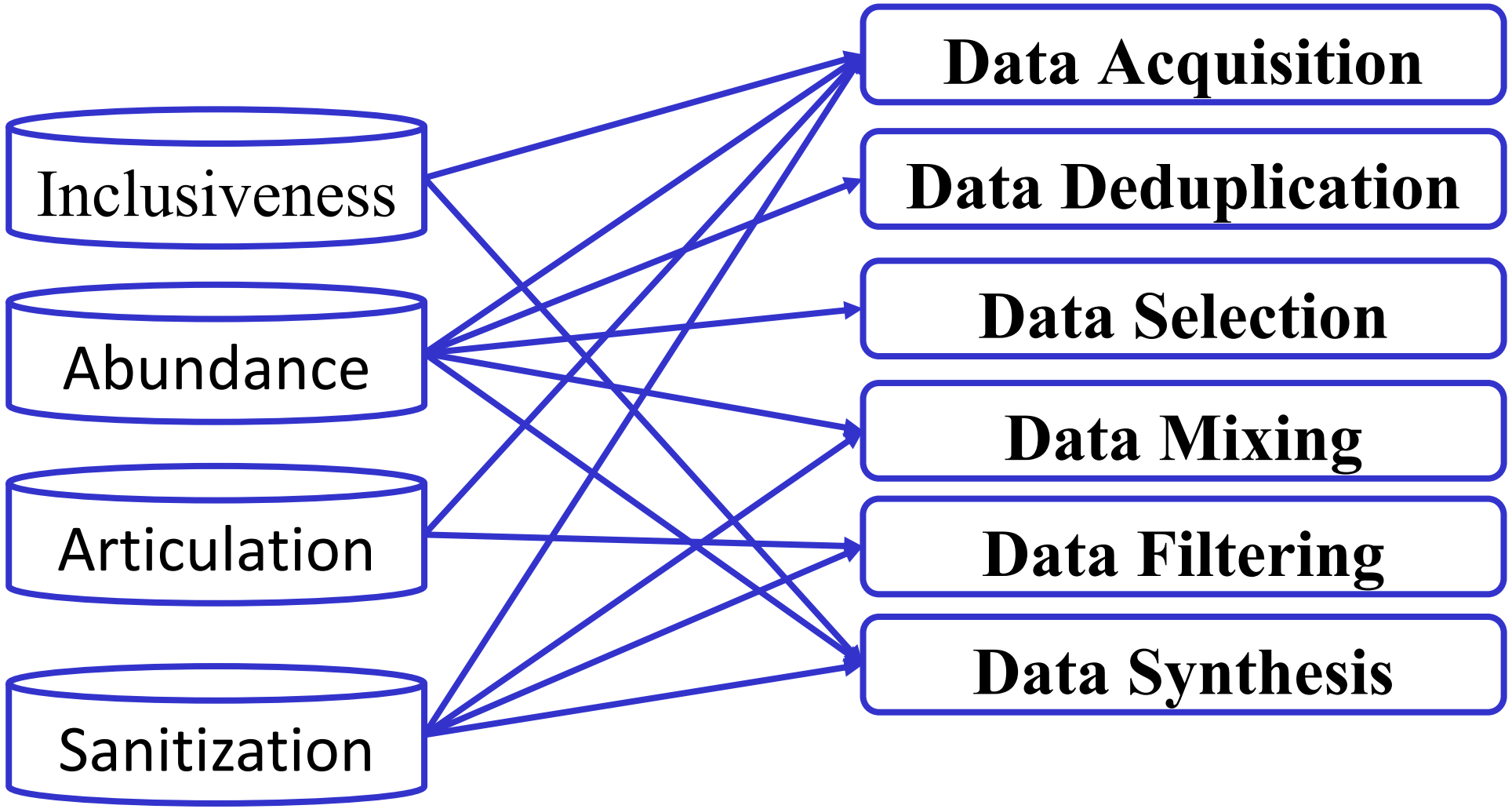
The IaaS Concept of DATA4LLM

➢ The ***IaaS*** Concept covers most LLM stages

Characters \ Stages	Pre-Training	Continual Pre-Training	Finetuning	RL	RAG	Agent	Evaluation
Inclusiveness	**Diverse Domains:** General Knowledge, Coding, STEM, Multilingual ...	**Specific Domain:** Finance, Medicine, ...	**Specific Domain:** QA, Summarization, Writing, Classification...	**Specific Domain:** **RLHF:** General, Medicine, Law... **RoRL:** Math, Coding, ...	**Specific Domain:** Finance, Medicine, Law...	**Planning; Tool; Multi-Turn Dialogue**	**Domain-Specific:** Legal Consultation, Code Synthesis and Translation...
Abundance	**Over 1 TB**; Tend to Increase Ratios like Math and Coding	**10M-100B**; Need General Data	**10k-10M**	**1k-1M**	---	**10k-10M**	**0.1k-100k**
Articulation	**Format:** Clear, Fluent, and Unambiguous Language; Well Formatted	**Format:** Clear, Fluent, and Unambiguous Language; Well Formatted	**Instruction-Response Pairs;** **Format:** Task-Oriented Response	**Step-By-Step Reasoning;** **Fine-Grained Annotation**	**Real-time;** **Extensive Info**	**Reasoning;** **Interaction Format**: Insturction+ Action+ Observation...	**Granular Evaluation;** **Free from Data Leakage**
Sanitization	Privacy Compliance; Toxicity-Free; Ethical Consistency; Risk Mitigation						

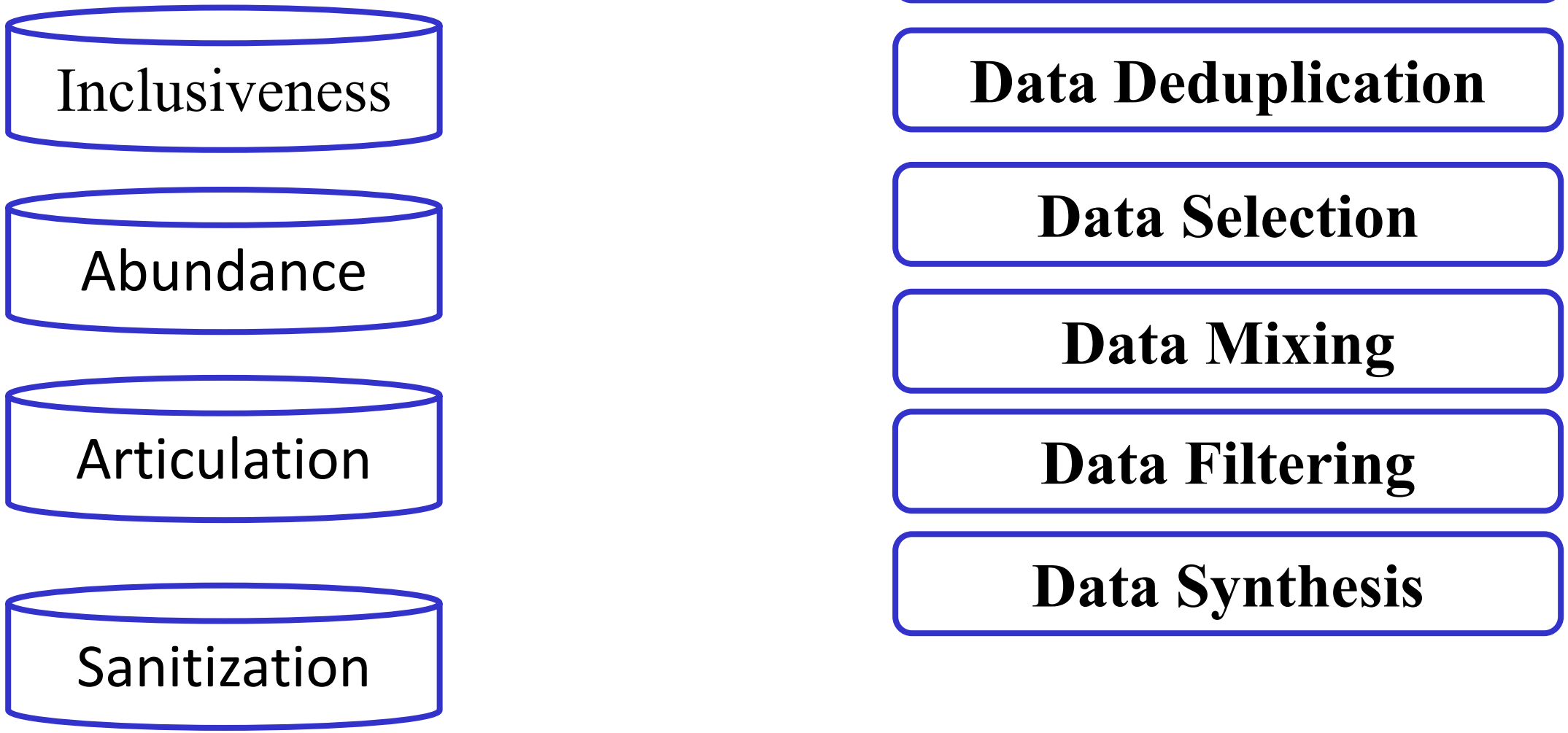
The IaaS Concept of DATA4LLM

- For an ***IaaS***-compliant dataset, use the data processing techniques



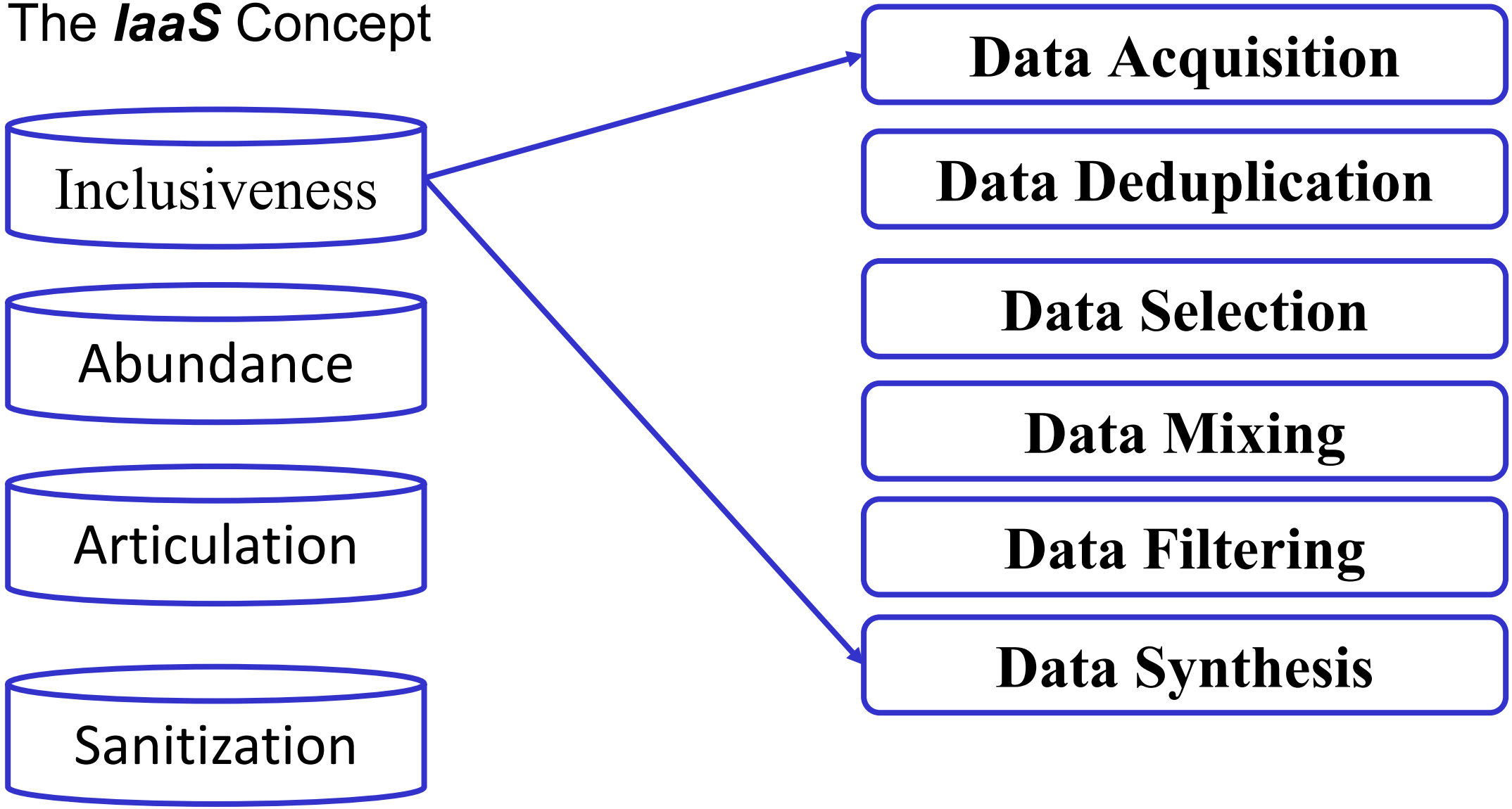
The IaaS Concept of DATA4LLM

- The ***IaaS*** Concept



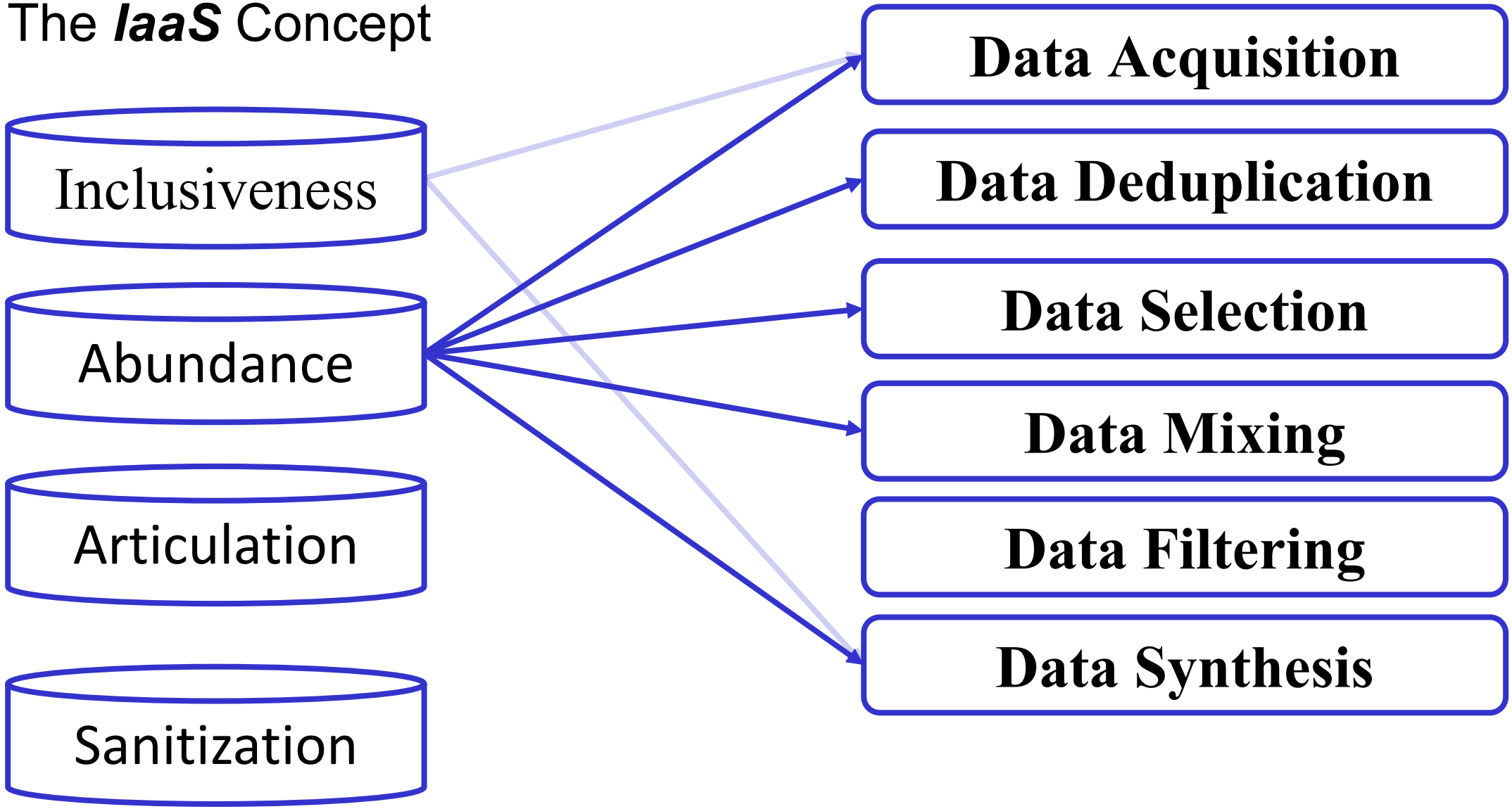
The IaaS Concept of DATA4LLM

- The ***IaaS*** Concept



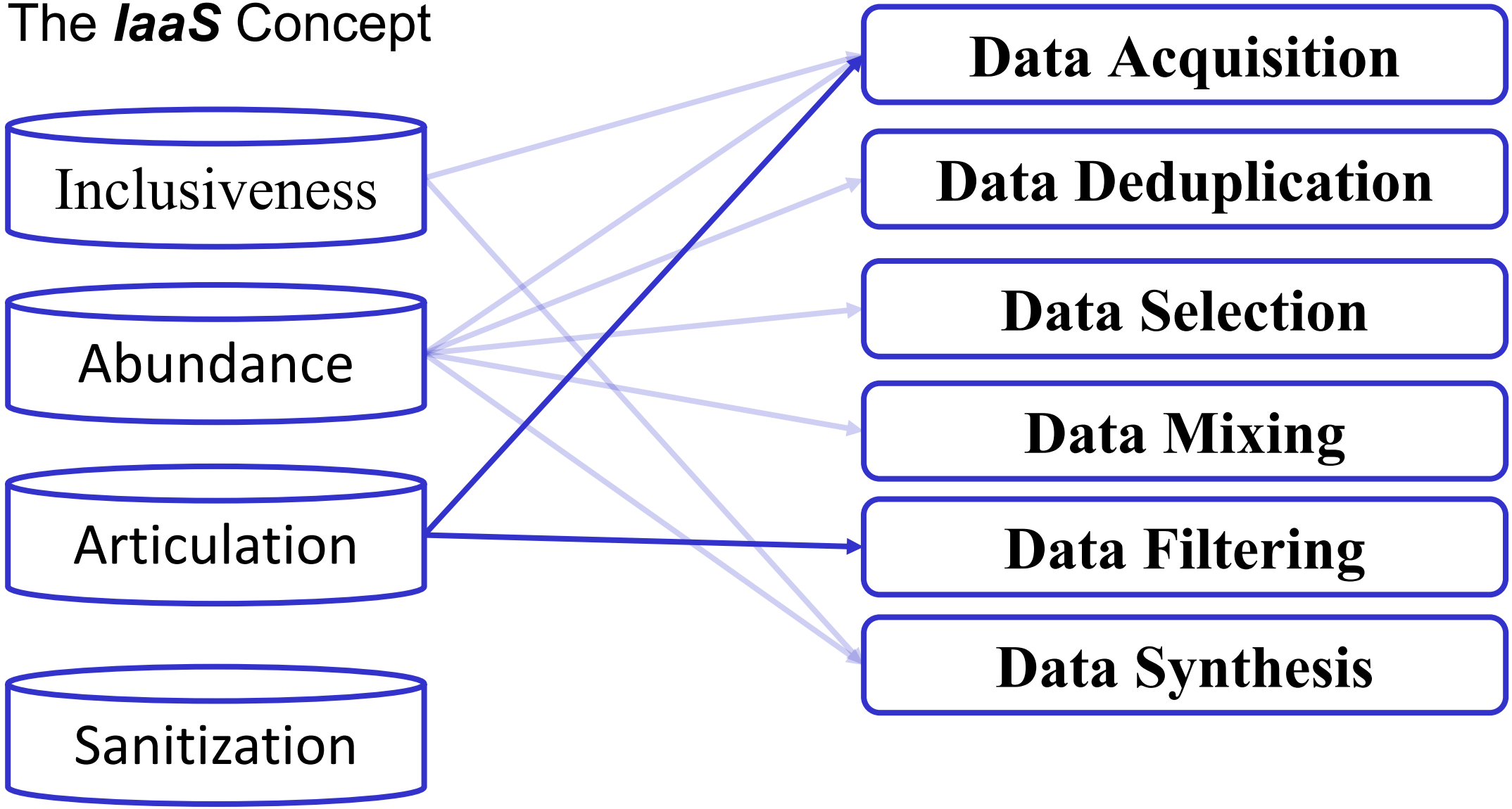
The IaaS Concept of DATA4LLM

- The ***IaaS*** Concept



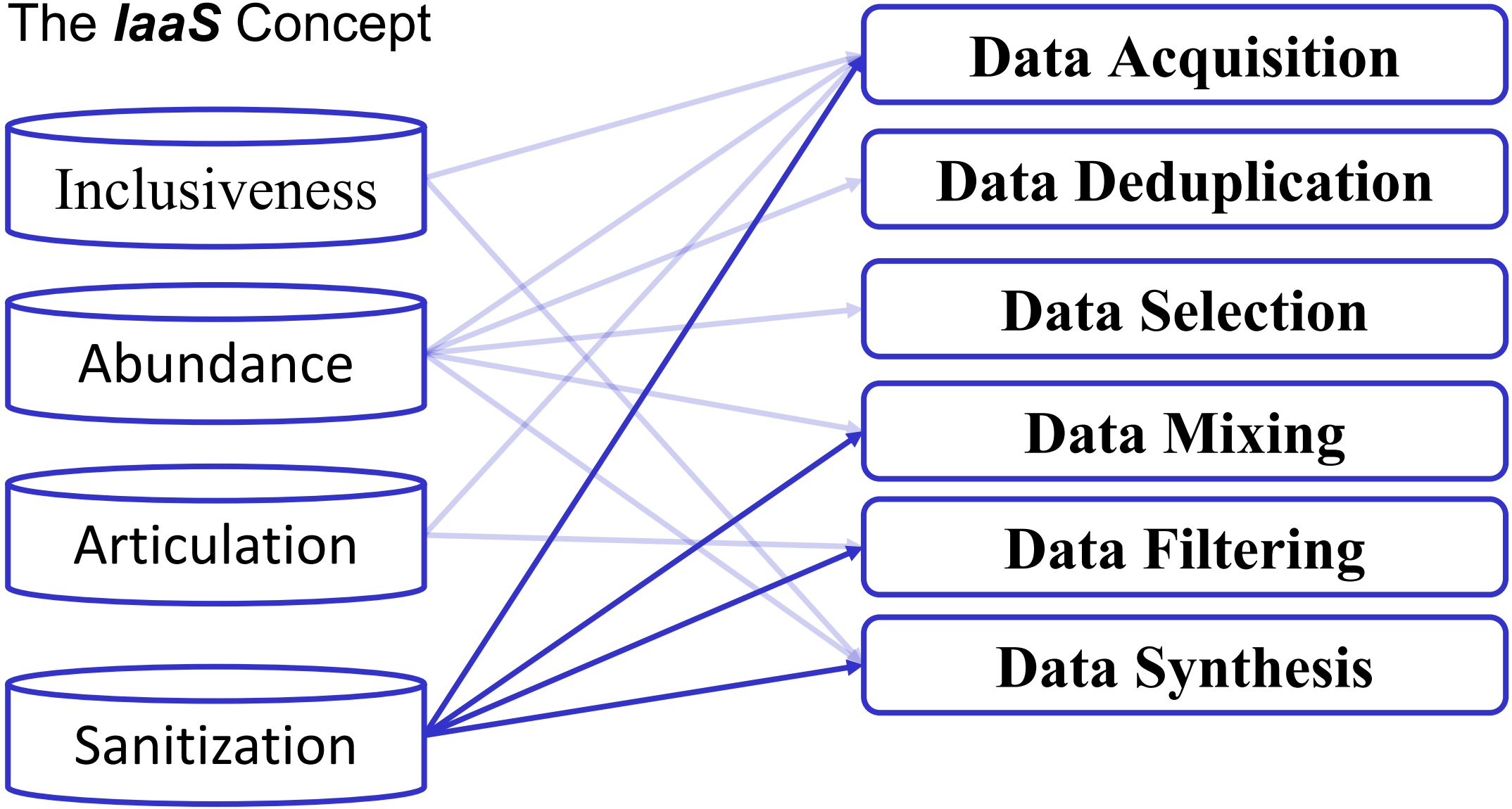
The IaaS Concept of DATA4LLM

- The ***IaaS*** Concept



The IaaS Concept of DATA4LLM

- The ***IaaS*** Concept



The IaaS Concept of DATA4LLM

➢ Required Techniques under ***IaaS***

Stage		Pre-training / Incremental Pre-training	Supervised Fine-Tuning	Reinforcement Learning	Inference	RAG
Data Processing	Acquisition	✓	✓	✓	N/A	✓
	De-duplication	✓	✓	N/A	N/A	N/A
	Filtering	✓	✓	N/A	N/A	×
	Selection	✓	✓	N/A	N/A	N/A
	Mixing	✓	✓	×	N/A	×
	Synthesis	✓	✓	✓	N/A	✓
Data Storage	Distribution	Distributed File System Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)
	Transmission	Caching Data Placement Parallelized Pipeline Data/Operator Offloading (CPUs)	Parallelized Pipeline Data/Operator Offloading (CPUs)	Parallelized Pipeline Data/Operator Offloading (CPUs)	×	N/A
	Fault Tolerance	✓	✓	✓	×	×
	KV Cache	N/A	N/A	N/A	Cache Space Management KV Indexing KV Placement KV Shrinking	KV Placement KV Shrinking
Data Serving	Selection	Sample-Scoring-Based Model-State-Based	Model-State-Based Experience-Based	N/A	×	SLM-Based Filtering LLM-Based Filtering Metric-Based Re-ranking LLM-Based Re-ranking
	Compression	N/A	N/A	N/A	✓	✓
	Packing	✓	✓	✓	×	×
	Provenance	×	×	×	✓	N/A

The IaaS Concept of DATA4LLM

➢ Required Techniques under ***IaaS***

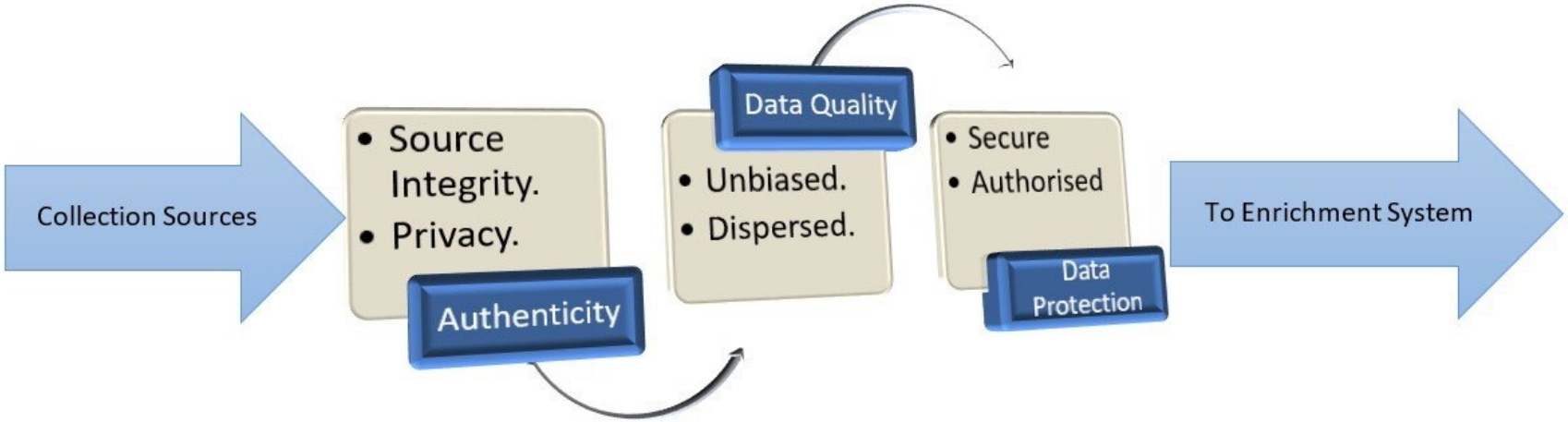
Stage		Pre-training / Incremental Pre-training	Supervised Fine-Tuning	Reinforcement Learning	Inference	RAG
Data Processing	Acquisition	✓	✓	✓	N/A	✓
	De-duplication	✓	✓	N/A	N/A	N/A
	Filtering	✓	✓	N/A	N/A	×
	Selection	✓	✓	N/A	N/A	N/A
	Mixing	✓	✓	×	N/A	×
	Synthesis	✓	✓	✓	N/A	✓
Data Storage	Distribution	Distributed File System Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)	Model Offload (GPUs, CPUs)
	Transmission	Caching Data Placement Parallelized Pipeline Data/Operator Offloading (CPUs)	Parallelized Pipeline Data/Operator Offloading (CPUs)	Parallelized Pipeline Data/Operator Offloading (CPUs)	×	N/A
	Fault Tolerance	✓	✓	✓	×	×
	KV Cache	N/A	N/A	N/A	Cache Space Management KV Indexing KV Placement KV Shrinking	KV Placement KV Shrinking
Data Serving	Selection	Sample-Scoring-Based Model-State-Based	Model-State-Based Experience-Based	N/A	×	SLM-Based Filtering LLM-Based Filtering Metric-Based Re-ranking LLM-Based Re-ranking
	Compression	N/A	N/A	N/A	✓	✓
	Packing	✓	✓	✓	×	×
	Provenance	×	×	×	✓	N/A

Data Acquisition —— Data Sources

- **Motivation: Pretraining needs abundant data, ensuring greater coverage, stronger abilities and less hallucination**
  - **Challenge: Where to collect the massive data?**
    - Public Data: Publicly available on the Internet with free access
      - Examples: Webpages (CommonCrawl, C4), digital books, code repositories (GitHub), etc.

Data Acquisition —— Data Sources

- **Motivation: Pretraining needs abundant data, ensuring greater coverage, stronger abilities and less hallucination**
  - **Challenge: Where to collect the massive data?**
    - Private Data: Privately owned by companies or organizations with limited access
      - Examples: Internal documents, event logs, subscriber-only content.
      - Constraints:
        - Prone to contain sensitive content, requiring inspecting and cleaning (e.g., anonymization)
        - Restricted use due to regional privacy laws (GDPR)



Data Acquisition —— Techniques

Method	Objective	Solution	Tools
Website Crawling	HTML Textual Content Extraction	Rule-based	Trafilatura [73]
		Rule-based	BET [144]
		ML-based	Dragnet [313]
	Automate Browser Interactions	HTML parsing	Beautiful Soup [6]
		Control web driver	Selenium [19]
		Wrap high-level API	Playwright [30]
		DevTools protocol	Puppeteer [31]
Layout-based	Content Extraction from Handwritten or Non-text Data	Model pipeline	PaddleOCR
		Model pipeline	MinerU [392]
		Multimodal LLM	GOT2.0 [407]
		Multimodal LLM	Fox [257]
Entity recognition & linking	New Sample Derivation	Bi-Transformer	ReFinED [68]
	Translation Consistency	Seq2seq Framework using References	AACTRANS [215]
	Text-Image Integration	Multimodal LLM	UMIE [367]

Data Acquisition —— Website Crawling

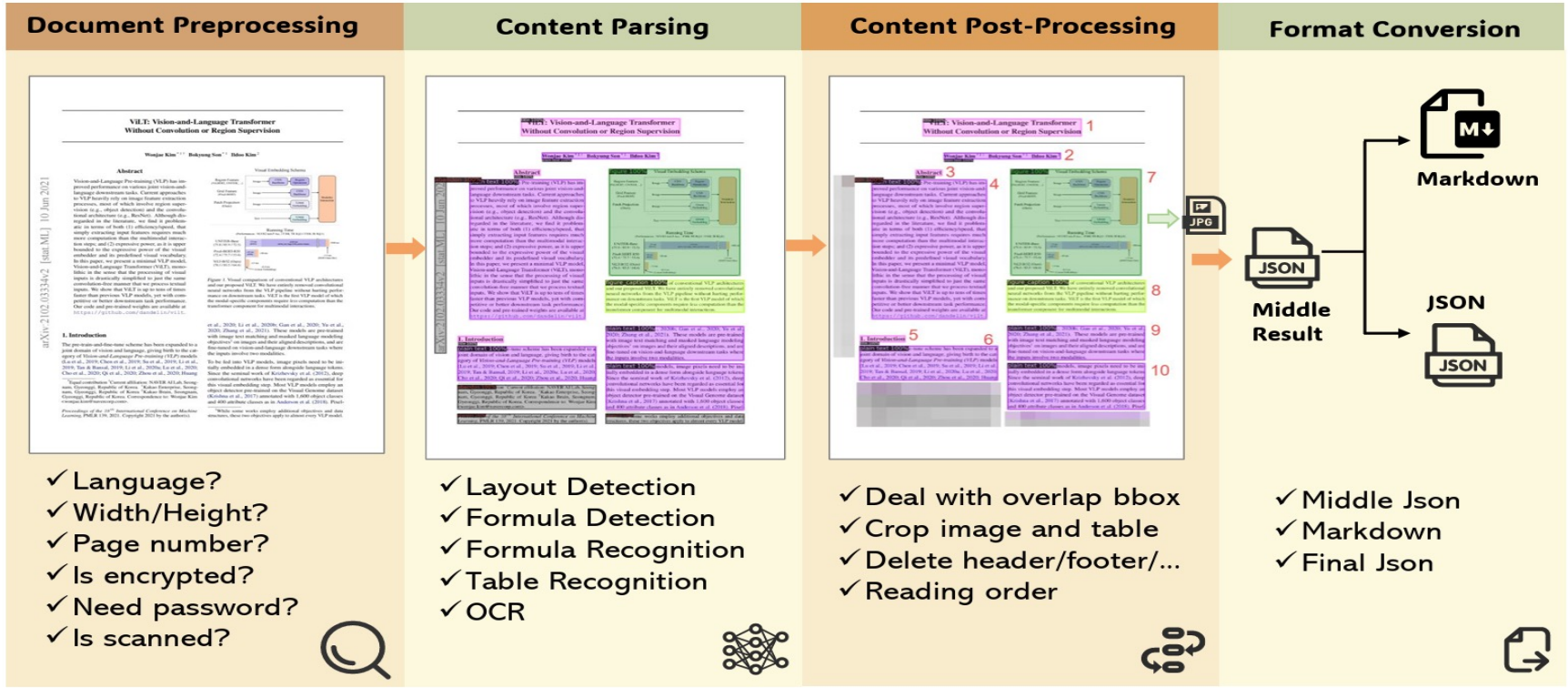
- **Motivation: Pretraining needs abundant data, ensuring greater coverage, stronger abilities and less hallucination**
  - **Challenge: How to identify and extract informative content inside the raw HTML data?**
    - Rule-based Crawling
      - Analyze and extract HTML pages through tags (head, body, footer, navbar, etc.), or
      - Find the elements with the largest region.
    - ML-based Crawling: Analyze whether a DOM node contains textual content using a HTML tag classifier
      - Available features include text density and word frequencies in “id” and “class” attributes.

$$P(y=1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}$$

Barbaresi, A. (2021, August). Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations (pp. 122-131).

Data Acquisition —— Layout Analysis

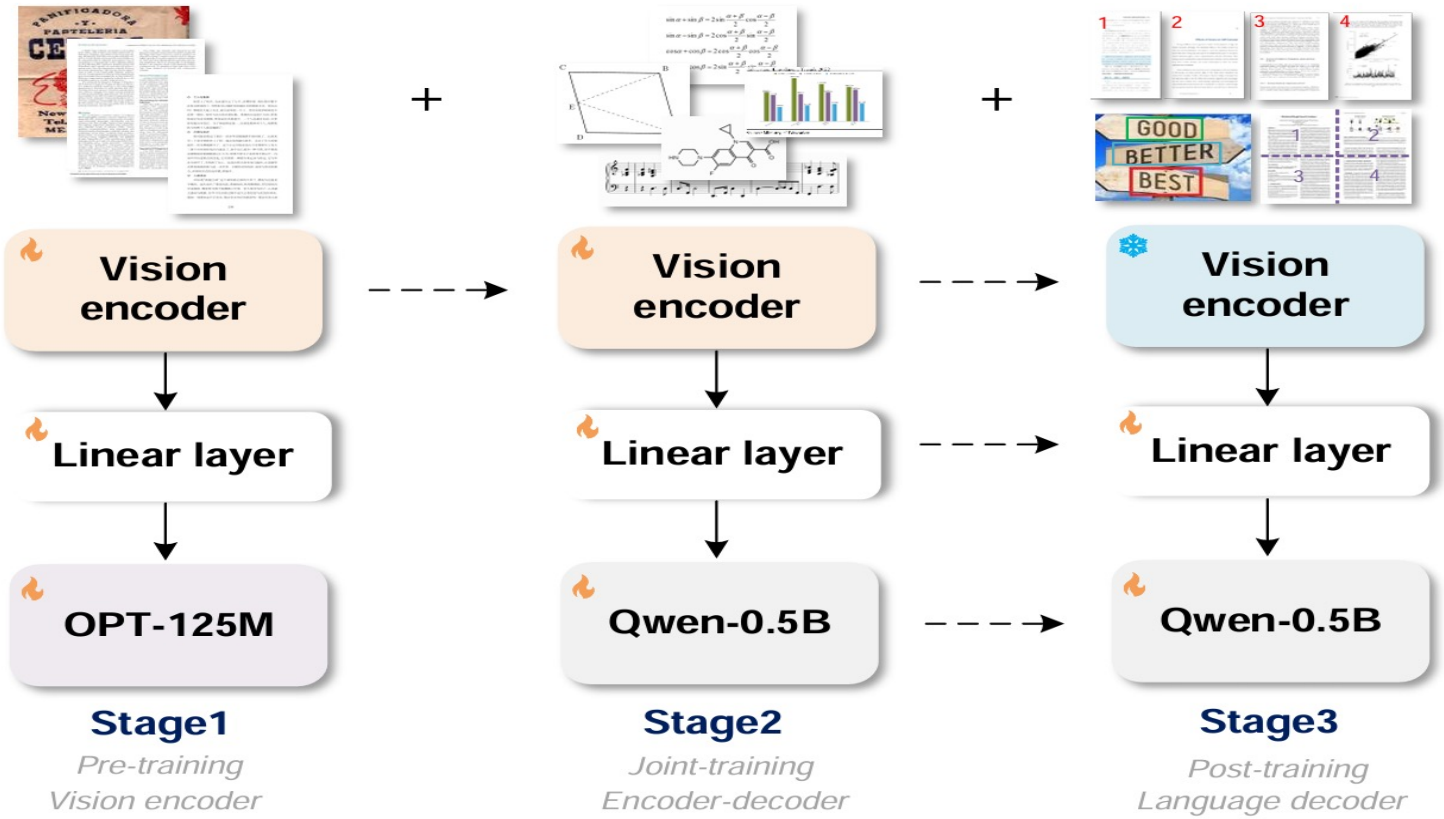
- Challenge: How to handle various types of documents with different elements.
  - Model Pipeline: Converts raw data (e.g., scanned books) into machine-readable formats in a pipeline manner, which consist of multiple small models.



Bin Wang, et al. MinerU: An Open-Source Solution for Precise Document Content Extraction [J]. arXiv preprint arXiv:2409.18839, 2024

Data Acquisition —— Layout Analysis

- Challenge: How to handle various types of documents with different elements.
  - Multimodal LLM: Adopts multimodal LLMs for end-to-end text acquisition.

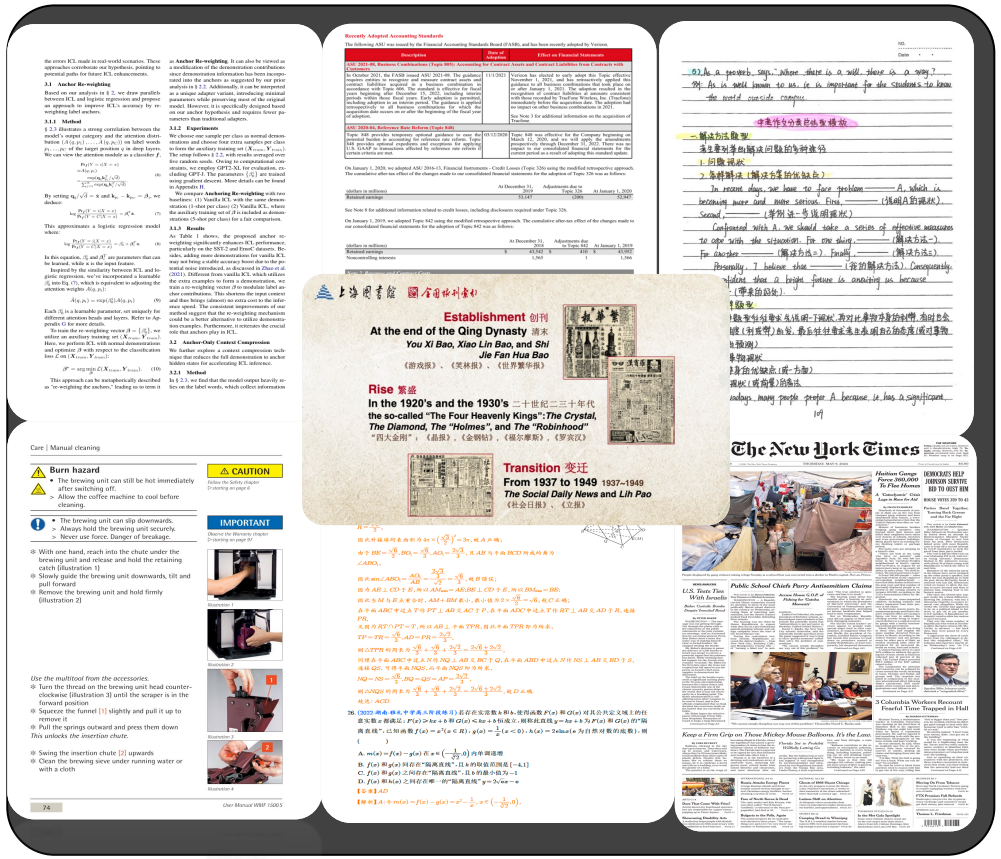


Haoran Wei, et al. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model[J]. arXiv preprint arXiv:2409.01704, 2024

Data Acquisition —— Layout Analysis

- Challenge: How to handle various types of documents with different elements.

多样性文档展示



布局检测挑战

- 文档类型多样，而可用数据集单一
- 文档元素尺度多样
- 多模态框架（LayoutLMv3等）实时性差

Data Acquisition —— Layout Analysis

多样性元素展示

20220607 项目第一次模拟测试卷
理科数学参考答案及评分标准

一、选择题：本大题共 12 个小题，每小题 5 分，共 60 分，在每小题给出的四个选项中，只有一项是符合题目要求的.

题号	1	2	3	4	5	6	7	8	9	10	11	12
答案	C	C	D	B	A	D	B	B	D	B	B	D

二、填空题：本大题共 4 小题，每小题 5 分，满分 20 分.

13. $x^2-\frac{y^2}{3}=1$　　14. $\frac{7}{5}$　　15. $\frac{8}{15}$　　16. 674

三、解答题：共 70 分. 解答应写出文字说明、证明过程或演算步骤. 第 17 题-21 题为必考题，每个试题考生都必须作答. 第 22 题、23 题为选考题，考生根据要求作答.

17.【解析】（1）因为点 A 运动的路程为 $\frac{2\pi}{3}$，所以 $\angle AOx=\frac{2\pi}{3}$，　……2 分

因为 $r_A=1, r_B=2$，所以 $\angle BOx=\frac{\pi}{3}$，则 $\angle AOB=\frac{\pi}{3}$，　……4 分

由余弦定理知 $AB^2=OA^2+OB^2-2OA\cdot OB\cdot\cos\angle AOB$，

得 $AB^2=1+4-2\times1\times2\times\frac{1}{2}=3$，所以 $|AB|=\sqrt{3}$.　……6 分

（2）设 $\angle BOx=\theta$，则 $\angle AOx=2\theta$，

所以 $A(\cos2\theta,\sin2\theta)$，$B(2\cos\theta,2\sin\theta)$，　……8 分

则 $x_1+y_2=\cos2\theta+2\sin\theta=-2\sin^2\theta+2\sin\theta+1=-2(\sin\theta-\frac{1}{2})^2+\frac{3}{2}$，　……10 分

所以当 $\sin\theta=\frac{1}{2}$ 时，x_1+y_2 取得最大值 $\frac{3}{2}$.　……12 分

18.【解析】（1）连接 BD 交 EC 于点 F，

由题意知，$PD\perp$ 平面 $ABCD$，所以 $PD\perp EC$，

又因为 $EC\perp PB$，$PD\cap PE=P$，

所以 $EC\perp$ 平面 PBD，则 $EC\perp BD$，　……2 分

因为 $PD=AB=2BC=2$，E 为斜边 AB 的中点，

所以 $BE=BC=1$，则 $\angle EBD=\angle CBD$，　……4 分

因为 $CD//AB$，所以 $\angle EBD=\angle CDB$，

则 $\angle CDB=\angle CBD$，所以 $CD=BC=1$；　……6 分

（2）连接 AD，因为 $AB=2, BC=1$，AB 为斜边，

所以 $\angle ABC=60°$，因为 $DC=BC=1$，

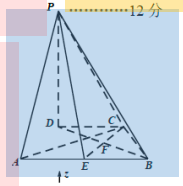
所以 $AD=1$，$\angle DAB=60°$，取 AE 的中点为 M，

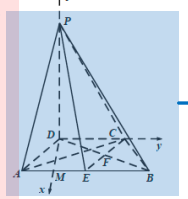
以 DM 为 x 轴，DC 为 y 轴，DP 为 z 轴建立空间直角坐标系，

则 $E(\frac{\sqrt{3}}{2},\frac{1}{2},0)$，$C(0,1,0)$，$P(0,0,2)$，

则平面 PDC 的法向量为 $\vec{n_1}=(1,0,0)$，　……8 分

— 高三理科数学（模拟一）答案第1页—





文本/标题

20220607 项目第一次模拟测试卷

二、填空题：本大题共 4 小题，每小题 5 分，满分 20 分

17.【解析】（1）因为点 A 运动的路程为 $\frac{2\pi}{3}$，所以 $\angle AOx=\frac{2\pi}{3}$，

表格元素

题号	1	2	3	4	5	6	7	8	9	10	11	12
答案	C	C	D	B	A	D	B	B	D	B	B	D

公式元素

$x^2-\frac{y^2}{3}=1$　　$AB^2=OA^2+OB^2-2OA\cdot OB\cdot\cos\angle AOB$

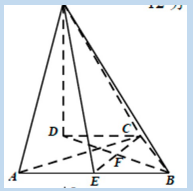
带格式文本

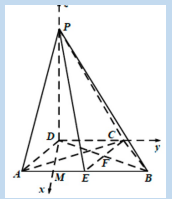
13. $x^2-\frac{y^2}{3}=1$　　14. $\frac{7}{5}$　　15. $\frac{8}{15}$　　16. 674

……2 分

……4 分

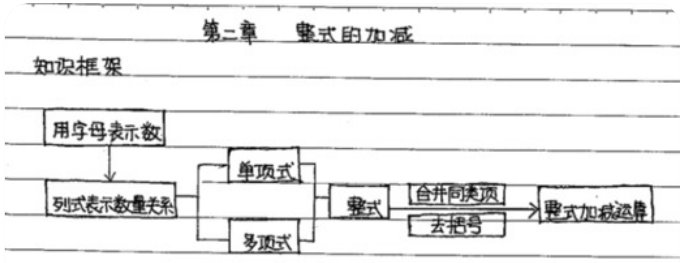
图片





Data Acquisition —— Layout Analysis

多样性文字展示




偽滿文獻

乾隆開欽命朝考卷


5

从表单满天飞到
数字化创新型队伍建设

铁福来如何打造数字化新型管理模式

铁福来装备制造集团股份有限公司成立于2003年，位于河南省平顶山市宝丰县产业集聚区，占地面积共9.2万平方米，注册资本金10296万元，是一家专业从事煤矿防突钻探高端装备生产制造的国家高新技术企业。

造业转型，从表单满天飞到数字化创新型管理队伍，近几年在数字化新型管理模式的推动下，铁福来持续创新，在煤矿防突钻探装备研发制造的实力居于全国领先行列。

2015年铁福来初始钉钉，发现钉钉的免费


下出死力戰必能破之偽朝穆提婆曰至尊計已成王子
齊主乃夜斬五龍門而出欲奔突厥從官多散
郎叩馬諫乃回向鄴時唯高阿那肱等十餘騎從
襄城王彦道繼至得數十人與俱
穆提婆西奔周軍陸令萱自殺家屬皆誅沒周主以
宜州刺史下詔諭齊羣臣曰若妙盡
命官榮爵賞各有加隆或我之將卒逃逸彼朝
從蕩滌自是齊臣降者相繼
初齊高祖爲魏丞相以唐邕典外兵曹太
兵曹皆以善書計工簿帳受委任及齊受禪諸司

図A-2～図A-4に、運動・スポーツに関する家族との関わり（する・みる・話す）と子どもの運動・スポーツ実施状況（実施頻度群）とのクロス集計結果を示した。ここでは、運動・スポーツに関する家族との関わりについて「よくしている」および「時々している」と回答したものを「している」とし、「ほとんどしていない」および「まったくしていない」と回答したものを「していない」として、子どもの運動・スポーツ実施頻度群との関係を分析した。

図A-2に、家族との運動・スポーツ、運動あそび実施状況と子どもの運動・スポーツ実施頻度群との関連について示した。家族と運動・スポーツ、運動あそびを実

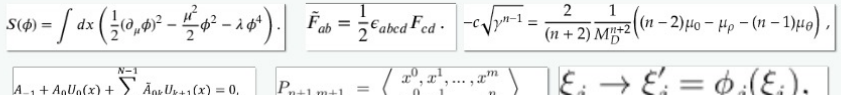
文字识别挑战

- 不同语言
- 不同尺度
- 不同清晰度
- 不同字体
- 不同排版
- ……

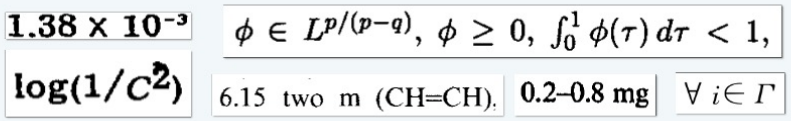
Data Acquisition —— Layout Analysis

多样性公式展示

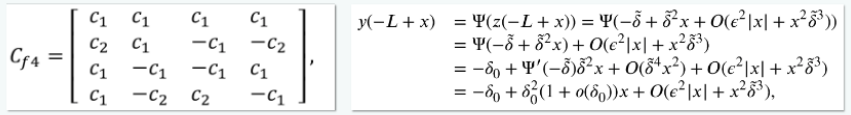
Simple Printed Expressions (SPE)



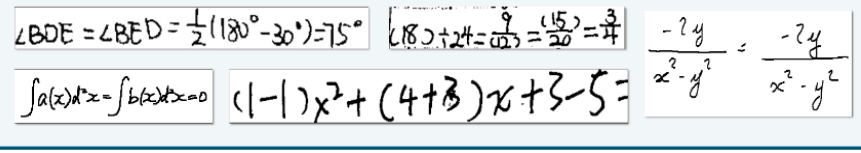
Screen Captured Expressions (SCE)



Complex Printed Expressions (CPE)



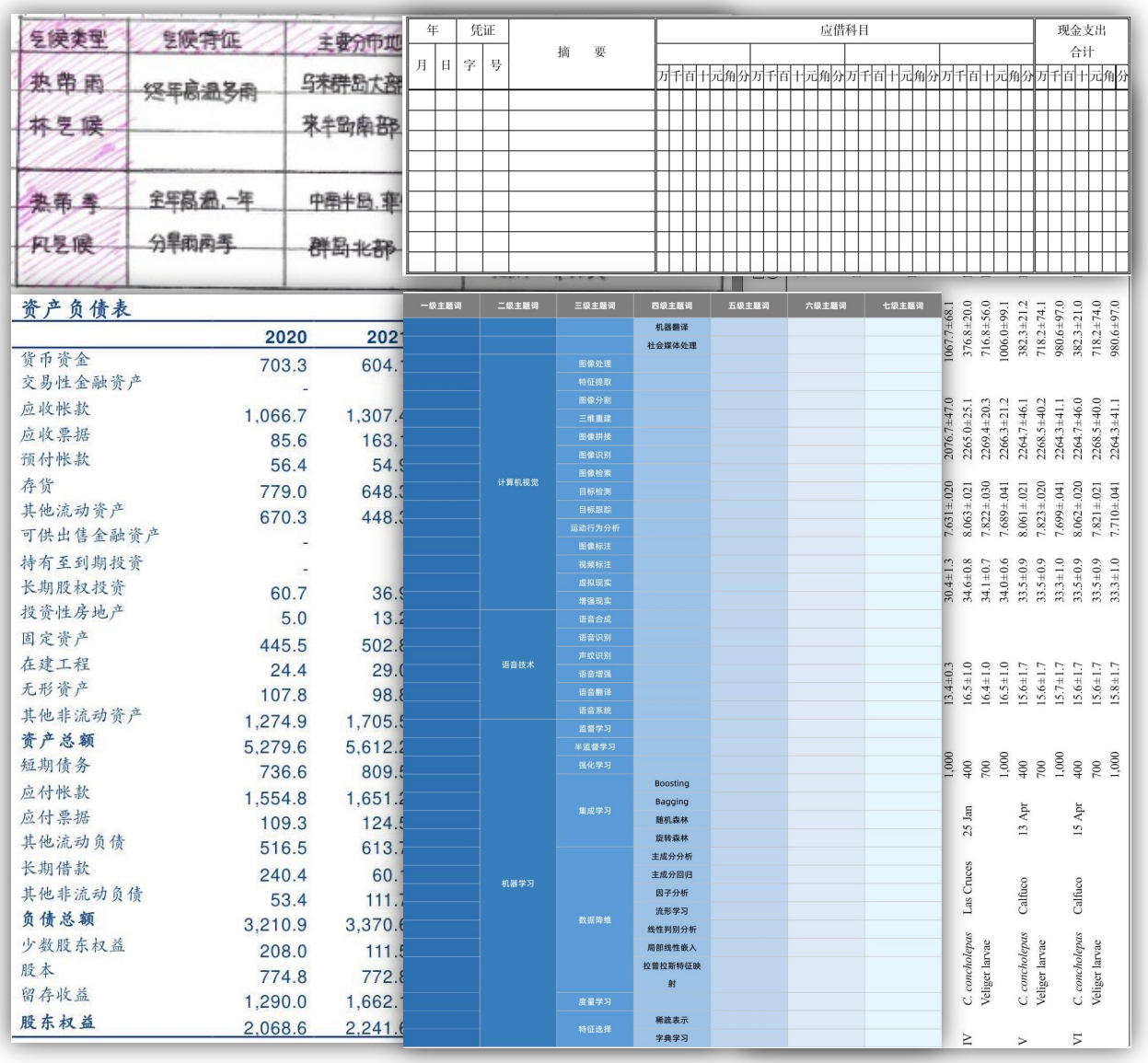
Handwritten Expressions (HWE)



公式识别挑战

- 简短公式 Vs 复杂长公式/多行公式
- 打印体公式 Vs 手写体公式/扫描件公式
- 公式识别多关注简短公式，忽略复杂长公式。

Data Acquisition —— Layout Analysis

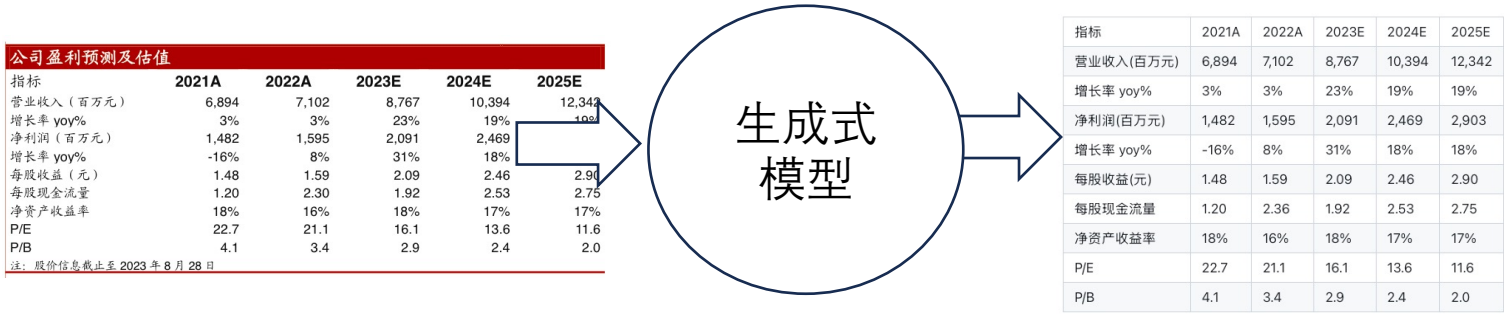


表格识别挑战

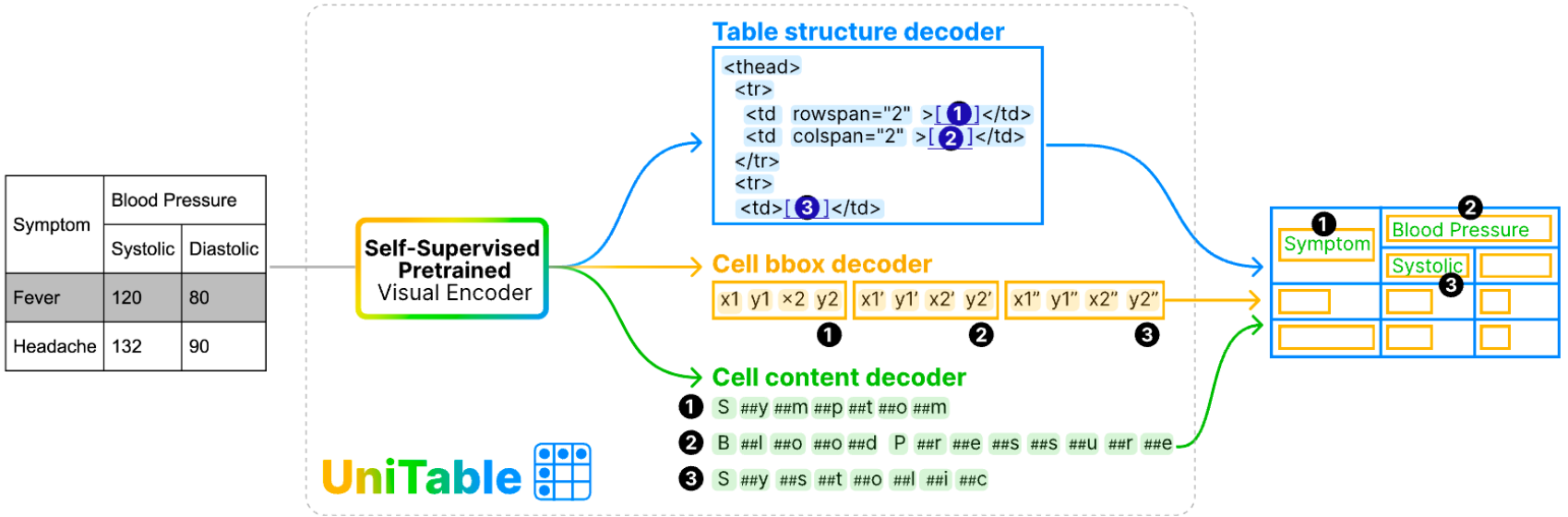
- 不同风格（有线、无线、三线...）
- 不同旋转角度
- 合并单元格、复杂长表格
- 空单元格表格
- 超长文字表格
- ……

Data Acquisition —— Layout Analysis

端到端模型（生成式）

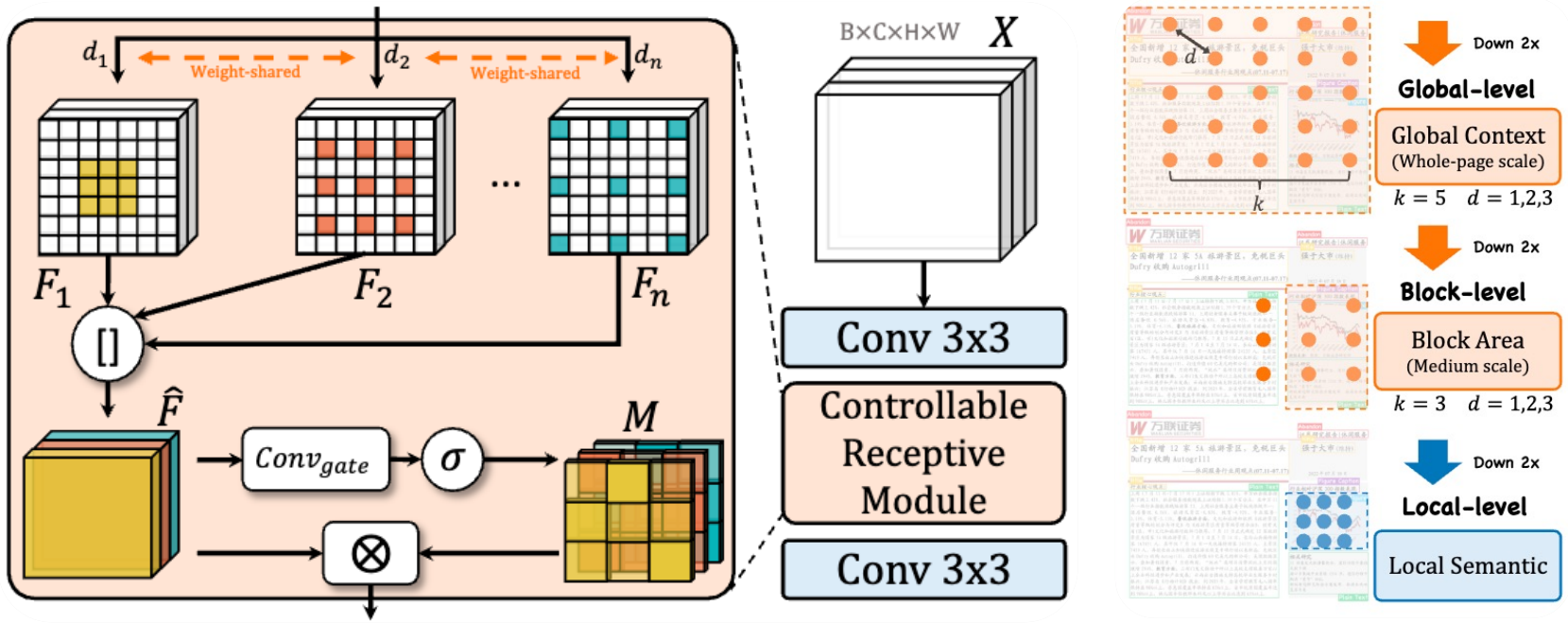


流水线方法（检测、识别、重构）



Data Acquisition —— Layout Analysis

- Challenge: Existing dataset for layout analysis has limited layout types and volume.
  - Generate diverse document images by searching for the best match between candidate elements (Candidates) and idle blocks (Mesh).

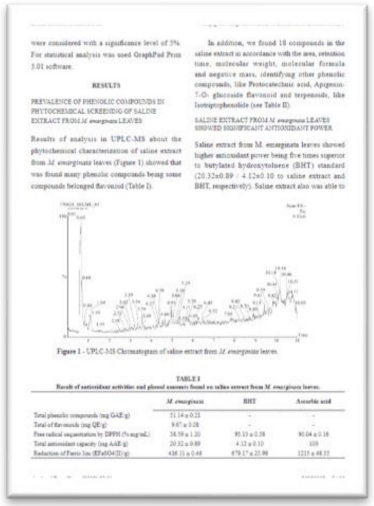


- ✓ 基于YOLO框架，精度高、速度快
- ✓ 基于多样性文档预训练，鲁棒性强
- ✓ Global-local模型优化，适配多尺度元素。

Zhiyuan Zhao, et al. DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception[J]. arXiv preprint arXiv:2410.12628, 2024

Data Acquisition —— Layout Analysis

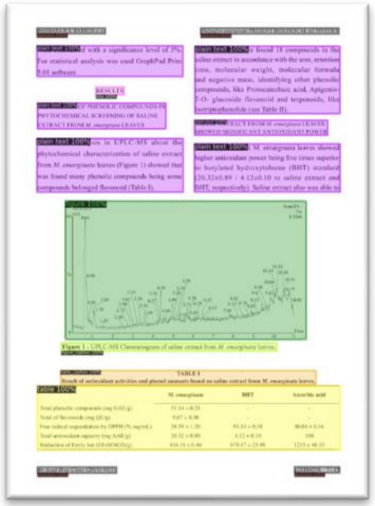
Original Document



Document:
- Paper
- Textbook
- Financial report
- PPT …

Language:
- English
- Chinese

Detection

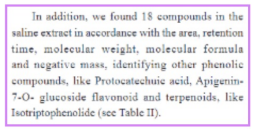


Layout Detection
- unilm LayoutLMv3

Formula Detection
- ultralytics Yolov8

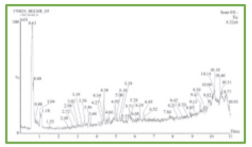
Recognition & Post-process

Plain texts → OCR



{"**text**": "In addition, we…",
"**poly**": [136.0, 781.0, …],
"**category_id**": 1}

Figures → Crop & Save



{"**poly**": [236.0, 11.0, …],
"**category_id**": 3}

Formula → Formula Recognition

Inline formula

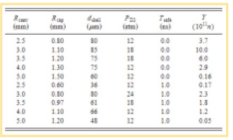
$n \to \infty$

Isolated formula

$$Q_C = 1 - \sum_{n=0}^{\infty} F(n, \gamma_1) P_0(n, \gamma_C),$$

{"**latex**": "n \\to \\infty",
"**poly**": [155.0, 631.0, …],
"**category_id**": 13}

Tables → Crop & Save



{"**poly**": [356.0, 90.0, …],
"**category_id**": 3}

Others (Header, Footer …) → Discard

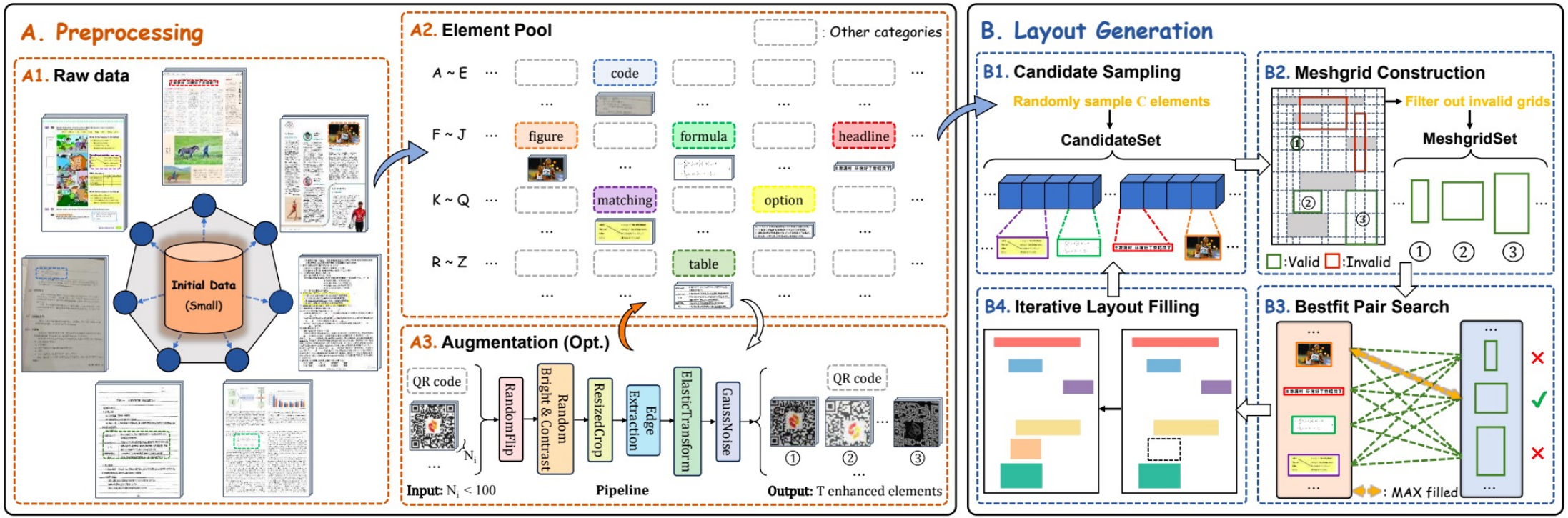
Discard

OCR
- PaddleOCR

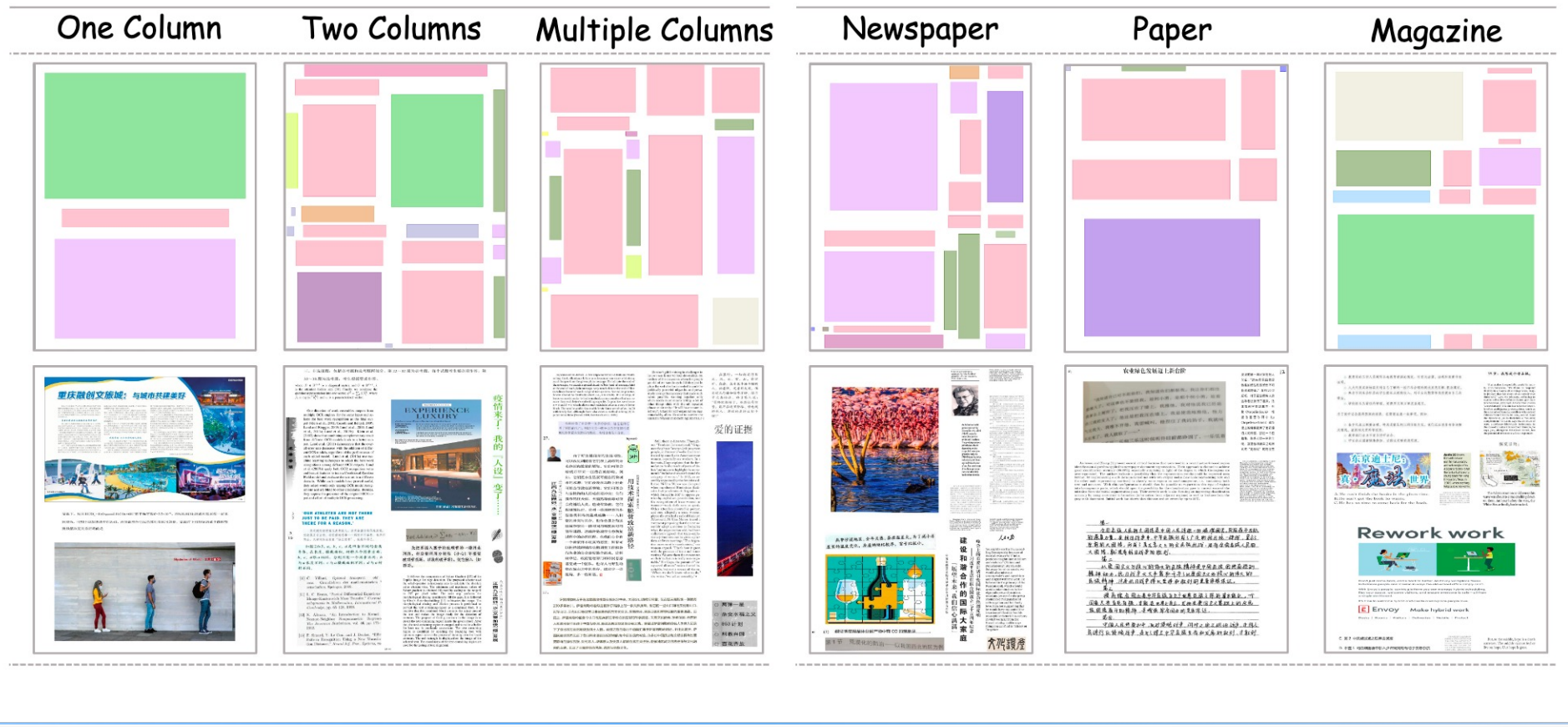
Formula Recognition
- UniMERNet

Data Acquisition —— Layout Analysis

1. Bestfit算法 合成多样性数据



Data Acquisition —— Layout Analysis



Data Acquisition —— Layout Analysis

2. （公式识别） UniMERNet网络结构

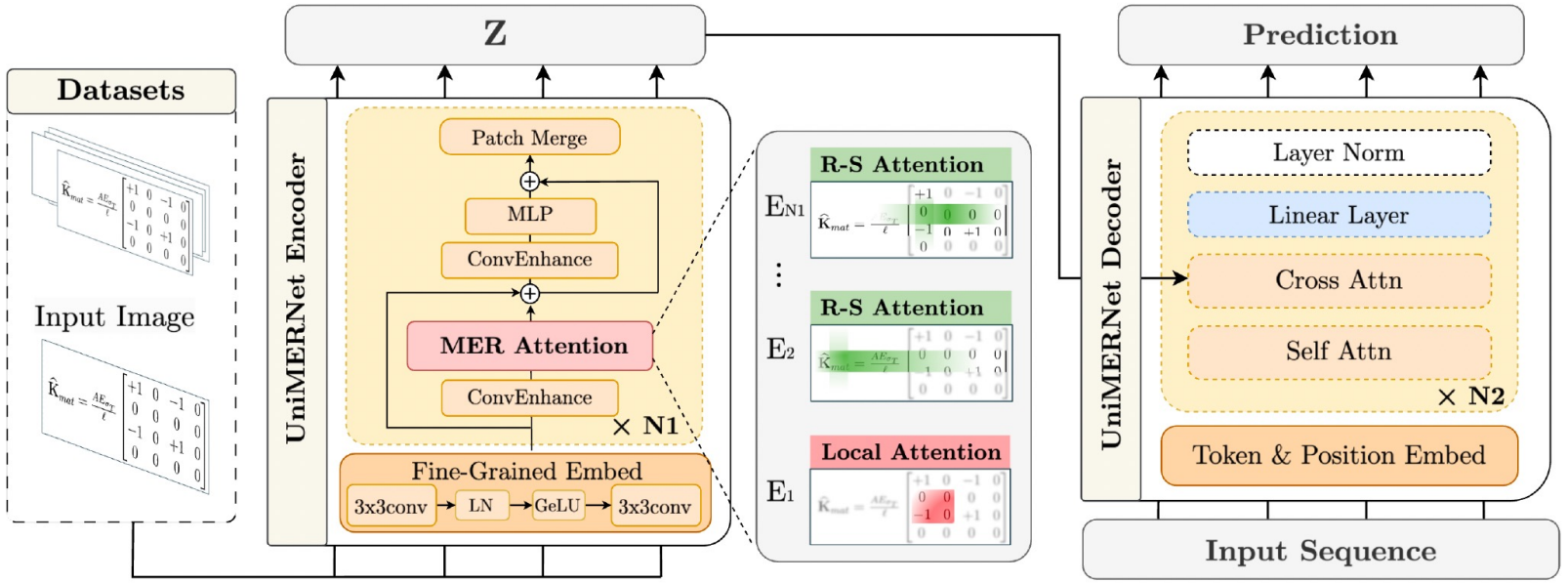
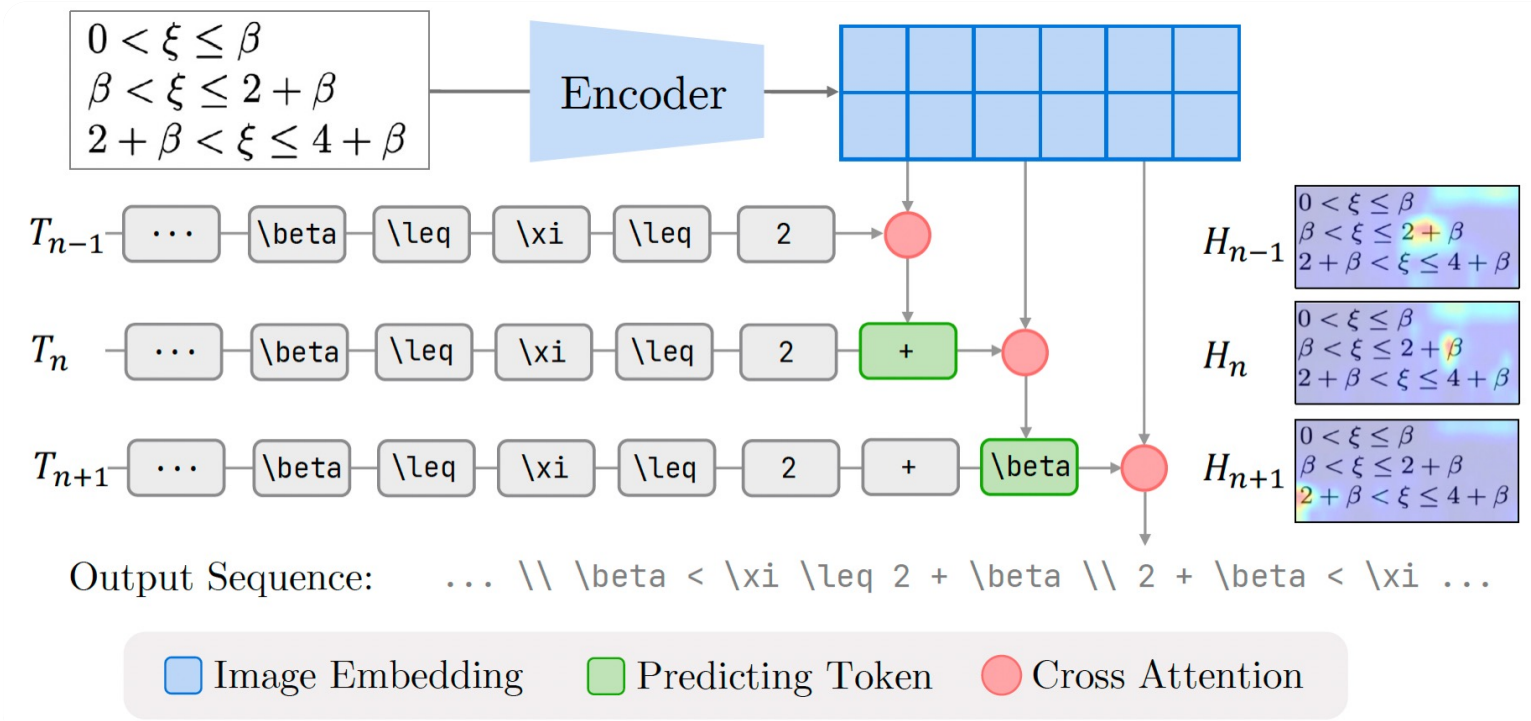


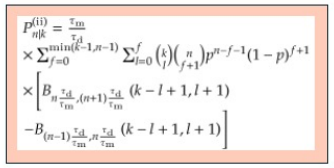
Figure 4: Overview of the UniMERNet architecture. The encoder-decoder framework with specialized MER blocks.

Data Acquisition —— Layout Analysis

2. （公式识别） UniMERNet网络结构



Data Acquisition —— Layout Analysis



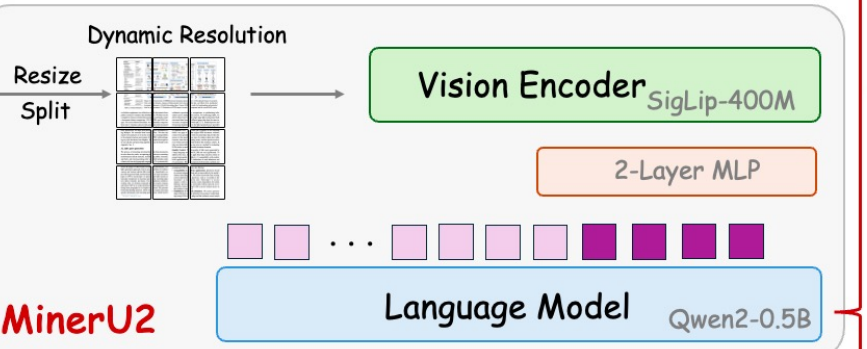
PROMPT

Document Parsing: Parse the entire document page.

Formula Recognition: Recognize formulas into LaTeX format.

Table Recognition: Recognize tables into OTSL format.

General OCR: OCR for general scenarios.



Training

OCR Related Data

SigLip-400M

2-Layer MLP

Qwen2-0.5B

Stage1 Pre-training

High Quality Data
Dynamic Resolution

SigLip-400M

2-Layer MLP

Qwen2-0.5B

Stage2 Join-training

Markdown format with layout:

Velho to Rio Comprido, with \(772~\mathrm{m}\) length each roughly. The tunnel cross section is \(81~\mathrm{m}^2\) in both bores. A concrete wall separates the north- bound traffic from the south- bound traffic. Each gallery has three lanes. CO concentrations and the \(1\textrm{h}\) average traffic flow are currently monitored. The tunnel ventilation system is composed of 78 fans placed near the roof (Rebouças 2006). The daily traffic volume ranges from 180,000 to 190,000 vehicles day- 1. ...

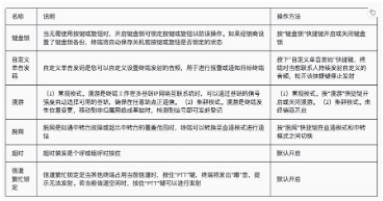
LaTeX format:

\begin{array}{rl} & P_{n|k}^{(ii)} = \frac{\tau_{\mathrm{m}}}{\tau_{\mathrm{d}}} \\ & \times \sum_{f = 0}^{\min (k - 1,n - 1)}\sum_{l = 0}^{f}\binom{k}{l}\binom{n}{f + 1}p^{n - f - 1}(1 - p)^{f + 1} \\ & \times \left[B_{n\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}},(n + 1)\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}}}(k - l + 1,l + 1)\right. \\ & \left. - B_{(n - 1)\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}},n\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}}}(k - l + 1,l + 1)\right] \end{array}

$$\begin{array}{rl} & P_{n|k}^{(ii)} = \frac{\tau_{\mathrm{m}}}{\tau_{\mathrm{d}}} \\ & \times \sum_{f=0}^{\min(k-1,n-1)} \sum_{l=0}^{f} \binom{k}{l}\binom{n}{f+1} p^{n-f-1}(1-p)^{f+1} \\ & \times \left[B_{n\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}},(n+1)\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}}}(k-l+1,l+1)\right. \\ & \left. - B_{(n-1)\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}},n\frac{\tau_{\mathrm{d}}}{\tau_{\mathrm{m}}}}(k-l+1,l+1)\right] \end{array}$$

OTSL & HTML format:

<table><tr><td>名称</td><td>说明</td><td>操作方法</td></tr><tr><td>键盘锁</td><td>当无需使用按键或旋钮时，开启键盘锁可锁定按键或旋钮以防误操作。如果经销商设置了键盘锁备份，终端将自动保存关机前按键或旋钮是否锁定的状态</td><td>...

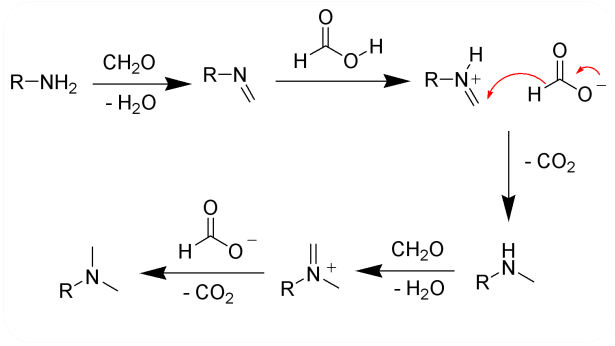


Text with on format:

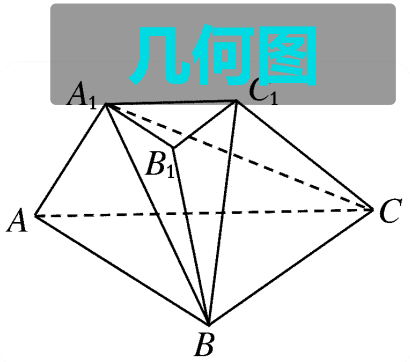
Rule # 12.31 Food and Beverages.
No food or beverages may be brought into a New York- Penn League stadium by fans except with the prior consent of either the club's general manager or the League President.
Thanks for your cooperation.

Data Acquisition —— Layout Analysis

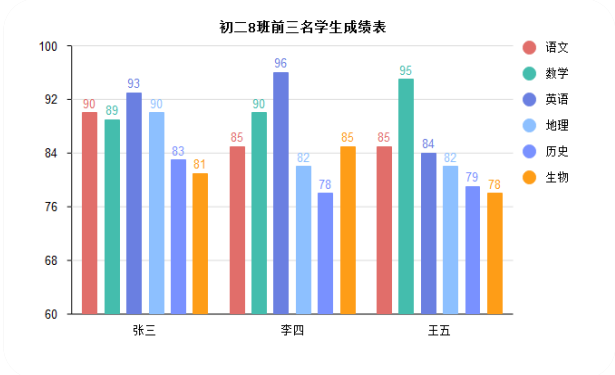
化学式



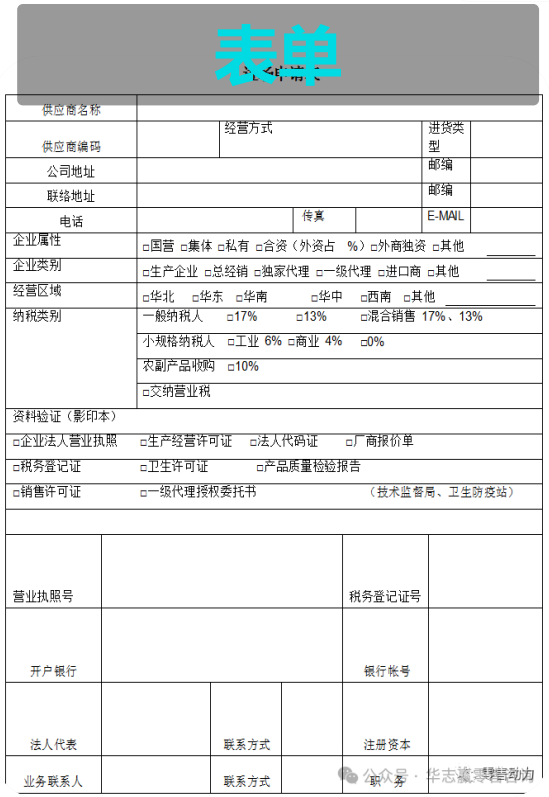
几何图



统计图表



表单



供应商名称					
供应商编码		经营方式		进货类型	
公司地址				邮编	
联络地址				邮编	
电话		传真		E-MAIL	
企业属性	□国营 □集体 □私有 □合资（外资占 %）□外商独资 □其他				
企业类别	□生产企业 □总经销 □独家代理 □一级代理 □进口商 □其他				
经营区域	□华北 □华东 □华南 □华中 □西南 □其他				
纳税类别	一般纳税人 □17% □13% □混合销售 17%、13%				
	小规格纳税人 □工业 6% □商业 4% □0%				
	农副产品收购 □10%				
	□交纳营业税				
资料验证（影印本）					
□企业法人营业执照 □生产经营许可证 □法人代码证 □厂商报价单					
□税务登记证 □卫生许可证 □产品质量检验报告					
□销售许可证 □一级代理授权委托书 （技术监督局、卫生防疫站）					
营业执照号				税务登记证号	
开户银行				银行帐号	
法人代表		联系方式		注册资本	
业务联系人		联系方式		职务	

代码段

3.单词大小写

```
str2 = "i love python"
print(str2.title())   # 单词首字母大写
print(str2.upper())   # 所有字母大写
print(str2.capitalize()) # 字符串首字母大写
```

```
I Love Python
I LOVE PYTHON
I love python
```

面对一个字符串，想将里面的单词首字母大写，只需要调用title()函数，而所有的字母大写只需要调用upper()函数，字符串首字母大写则是调用capitalize()函数即可。

4.字符串的拆分

```
str4 = "I love Python"
str4_1 = "I/love/Python"
str4_2 = "   I love Python   "
print(str4.split())  # 默认是按照空格进行拆分，返回的是列表
print(str4_1.split('/'))
print(str4_2.strip())  # 默认去除字符串左右两边的空格，返回的是字符串
print(type(str4_2.strip()))
```

```
['I', 'love', 'Python']
['I', 'love', 'Python']
I love Python
<class 'str'>
```

字符串的拆分可以直接利用split函数，进行实现，返回的是列表，而strip函数用于移除字符串头尾指定的字符（默认为空格或换行符。

Data Acquisition —— Layout Analysis

01

新一代文档解析架构探索

当前文档解析的两种主流方案：基于小模型的流程线方案（pipeline-based）及基于端到端大模型方案(vlm-based)均存在各自优势，是否有更易于拓展，高效推理且精准解析的方案仍需探索。

复杂场景、元素需要结合推理能力

复杂排版下的阅读顺序，复杂元素解析（复杂表格、表单及流程图）需要逻辑推理能力，仅依赖简单的识别能力不足以精准解析。

02

03

精度、速度、成本优化

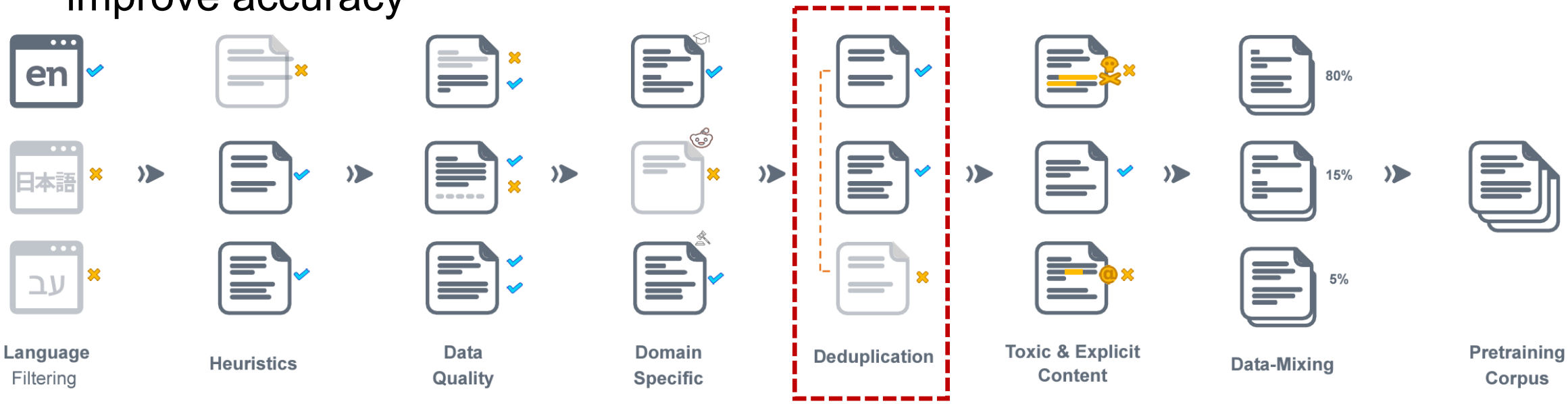
文档解析作为大模型时代的基础技术，需要质量足够高，推理足够快，成本足够小

Data Acquisition

- **Takeaways**
  - Data acquisition for large language models (LLMs) relies primarily on large-scale, unsupervised web scraping across diverse domains, contrasting with traditional ML's focus on labeled, domain-specific data.
  - Key data sources include public data (e.g., webpages, books, code repositories) and **private data** (e.g., internal logs, subscriber content), with the latter requiring strict privacy safeguards and secure handling.
  - Main acquisition methods are:
    - **Website crawling:** Use rule-based (e.g., Trafilatura, BET) or ML-based techniques (e.g., Dragnet) to extract clean text from HTML, aided by tools like Beautiful Soup, Selenium, and Playwright.
    - **Layout analysis:** Extract text from non-digital or structured documents using pipeline systems (e.g., PaddleOCR, MinerU) or end-to-end multimodal LLMs (e.g., GOT2.0, Fox), the latter offering higher versatility but lower efficiency.
    - **Entity recognition and linking**: Derive structured knowledge (e.g., triples) from text, using models like ReFinED and UMIE, though recent LLMs may implicitly learn such relationships, reducing the need for explicit linking.
  - Ensuring data quality, translation consistency (e.g., via AACTRANS), and multimodal integration (e.g., text-image alignment in UMIE) are critical for robust LLM pretraining.
  - Efficiency, scalability, and ethical compliance remain central challenges in large-scale data acquisition.

Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy



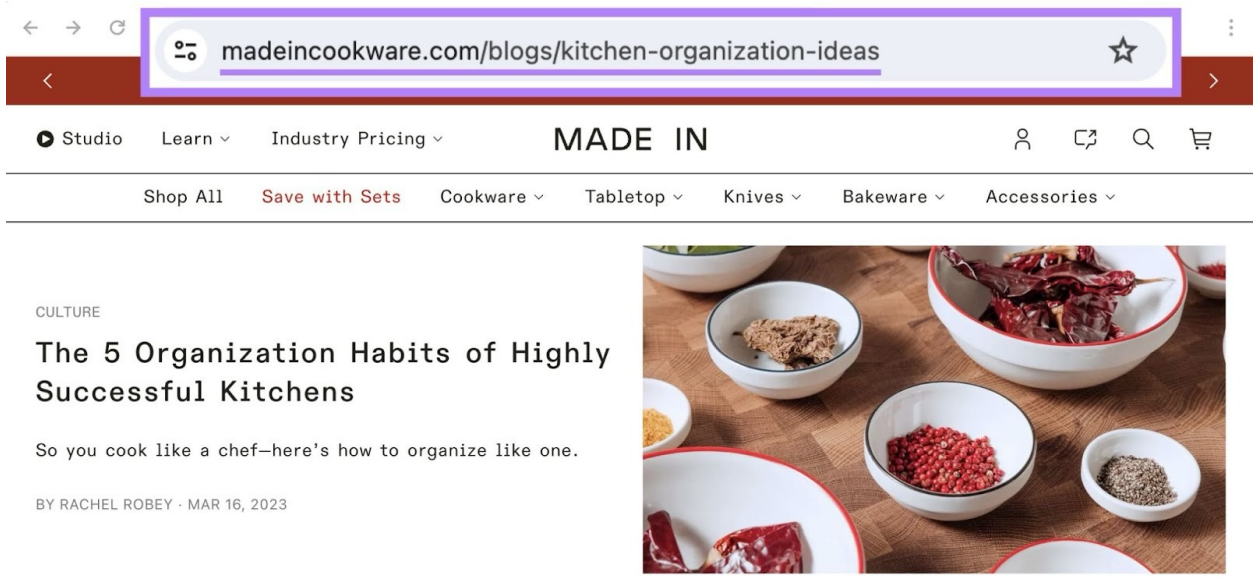
  - Utility Function: (1) Exactly match strings or documents or (2) Use approximate matching methods calculated according to similarity measures

Data Deduplication

Method	Objective	Modality	Work
Exact substring matching	Deduplicate samples with identical substrings	Text	MD5 [122] Suffix Array [299]
Hashing identification	Deduplicate samples with similar substrings	Text	SimHash [88] MinHash [81], [122], [299] MinHashLSH [347], [358] MinHash + Bloom Filter [207] DotHash [298]
Frequency analysis	Down-weighing samples with higher commonness	Text	SoftDeDup [167]
Embedding-based clustering	Deduplicate samples with identical topics but different formats	Text + Image	SemDeDup [46] SemDeDup + SSL Prototypes[385] FairDeDup [360]

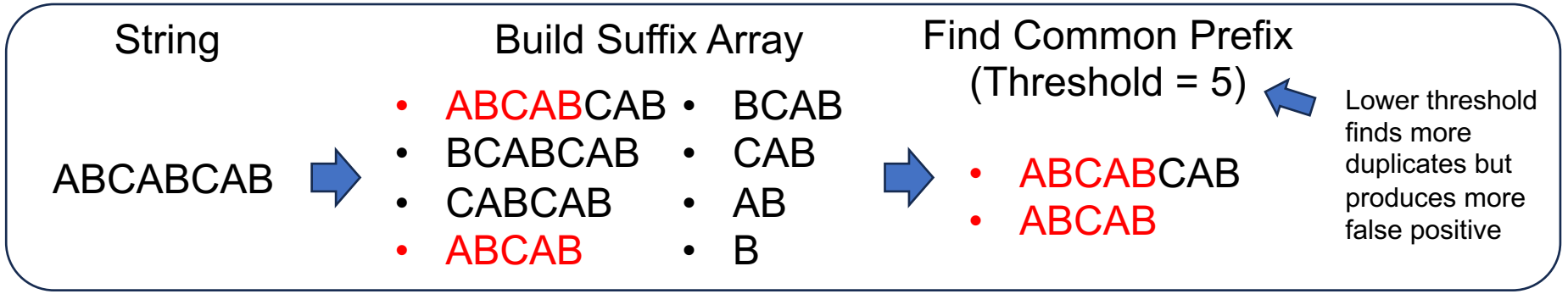
Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training performance
  - Exact Matching Techniques:
    - 1. URL Deduplication: Remove data that shares the same URL
      - Individual web page may appear in multiple datasets.
      - The samples sourced from the same URL will be identified as duplicates



Data Deduplication

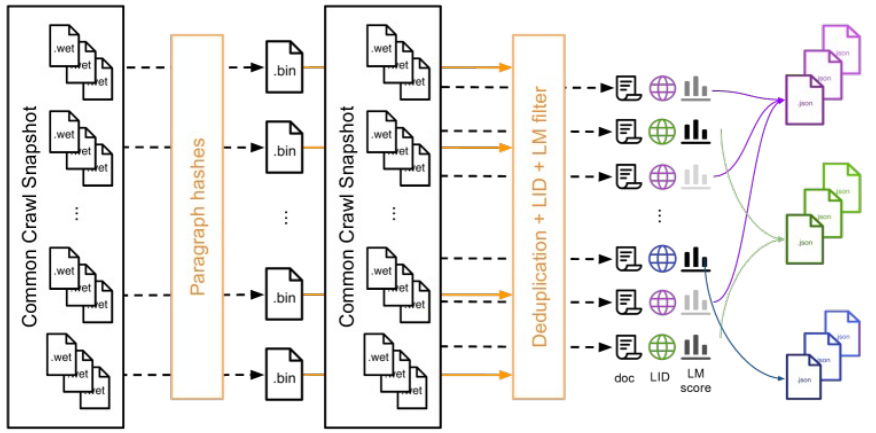
- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training performance
  - Exact Matching Techniques:
    - 2. Exact Substring Deduplication: Remove data that shares the same substring
      - Individual content may be referenced by multiple samples.
      - The samples with the same substring will be identified as duplicates.
    - Suffix Array: Find the duplicates through the common prefix of the suffice
      - **S1:** Combine all the samples into one single string
      - **S2:** Build the Suffix Array of the string
      - **S3:** Find duplicates through common prefix



Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating training data makes language models better.

Data Deduplication

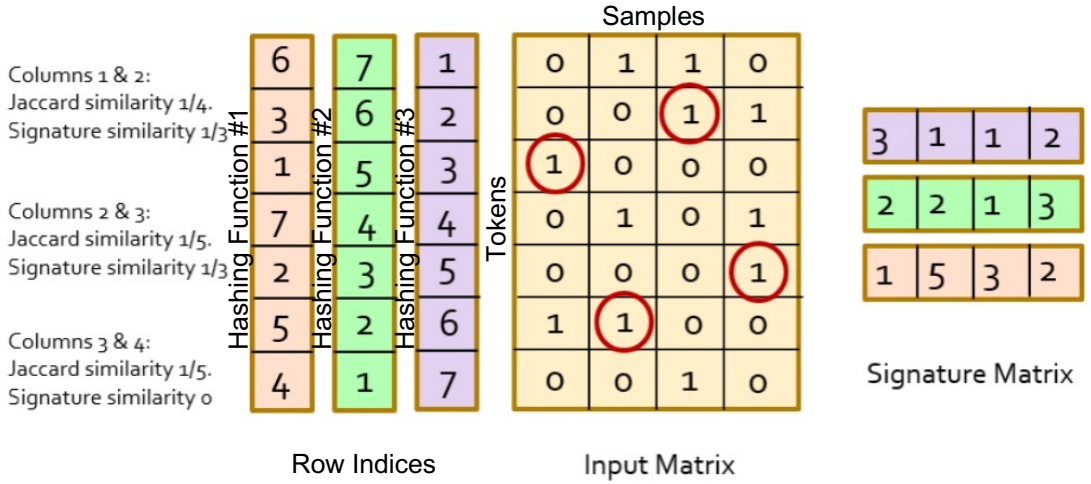
- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training performance
  - Exact Hashing Techniques:
    - 1. Hash Functions: Guarantee to find all exact matches
      - **(1) Initialize a Set for Hashes**
        - A set ~ The hashes of encountered text entries.
      - **(2) Hash Each Text Entry**
        - For each text entry, compute a simple hash (e.g., the sum of ASCII values of its characters).
      - **(3) Check for Duplicates**
        - If the hash of the current entry is already in the set, it is a duplicate and will be ignored.
        - If the hash is not in the set, add the hash to the set and keep the entry.



Efficient and Fast*, but *may find false positives due to hash collisions and remove non-matching documents

Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy
  - Approximate Hashing Techniques (MinHash):
    - **S1:** One-hot encode the samples by, e.g., n-grams, forming a matrix with row as token and column as sample
    - **S2:** Apply a series of hashing functions to row indices to shuffle the rows. For each sample:
      - After applying each hashing function, look for the lowest row index with value 1.
      - The sample signature is a vector of the lowest row indices.
      - Signature also works: Samples with similar set of tokens produce similar signatures.



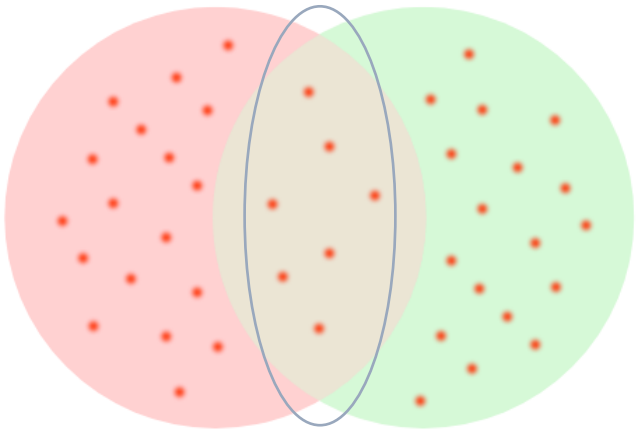
Manku, G. S., Jain, A., & Das Sarma, A. (2007, May). Detecting near-duplicates for web crawling. In Proceedings of the 16th international conference on World Wide Web.

Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy
  - Approximate Hashing Techniques (MinHash):
    - **S1:** One-hot encode the samples by, e.g., n-grams, forming a matrix with row as token and column as sample
    - **S2:** Apply a series of hashing functions to the row indices to obtain random row indices.
    - **S3:** Compute Jaccard Index between fingerprints.

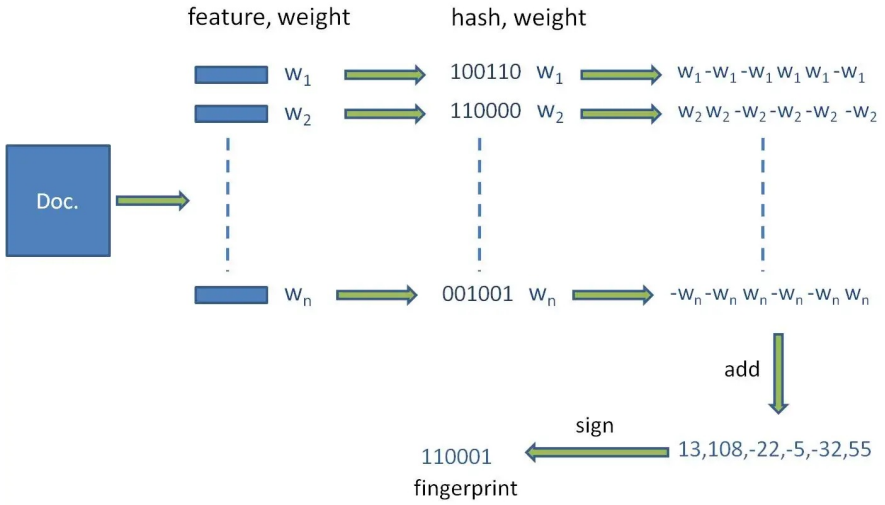
$$\text{Jaccard}(d_i, d_j) = |d_i \cap d_j| / |d_i \cup d_j|$$

      - Less computation by computing Jaccard Index on short fingerprints instead of long original data.
      - The more the hashing functions, the better the signature similarity approximates to the original one



Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy
  - Approximate Hashing Techniques (SimHash):
    - **S1:** Hash each token in the document into a fixed-dimension vector of $\{0, 1\}^d$, weighted by pre-defined or TF-IDF weight w that is positive for 1 and negative for 0.
    - **S2:** Add up weighted vector elements of each token to a single value, forming a new vector of the same dimension d
    - **S3:** Map the new vector to another vector of $\{0, 1\}^d$, which is the fingerprint of each sample
    - **S5:** Compute Hamming distance, the number of different elements between their vectors.



Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy
  - Approximate Hashing Techniques

维度	**SimHash**	**MinHash**
基本原理	将每个 token 哈希为固定维度的向量，加权累加后对每一位取符号（正为 1，负为 0）	对每个样本生成哈希签名，签名表示最早“命中”的 token 位置
最终表示	固定长度的二进制向量（指纹）	多个哈希函数下的最小值组成的签名向量
相似度计算方式	**汉明距离**（Hamming Distance）	**Jaccard 相似度**估计（签名重合程度）
适合的数据结构	高维稀疏向量（如 TF-IDF 文本）	集合（如 n-gram 集合、token 集合）
签名生成速度	一次遍历即可，速度快	需要多次哈希（多轮），速度稍慢
签名大小	较小，通常为 64 位或 128 位	可调大小（根据哈希函数个数）
鲁棒性	对文本顺序和轻微变动不太敏感	对 token 顺序不敏感，但对 token 改动较敏感
常见应用	文档去重（如搜索引擎）、海量文本相似检测	网页去重、大规模集合聚类、重复检测
相似性误差	会有一定信息损失，可能高估相似度	准确估计 Jaccard，但需要较多哈希函数才能稳定

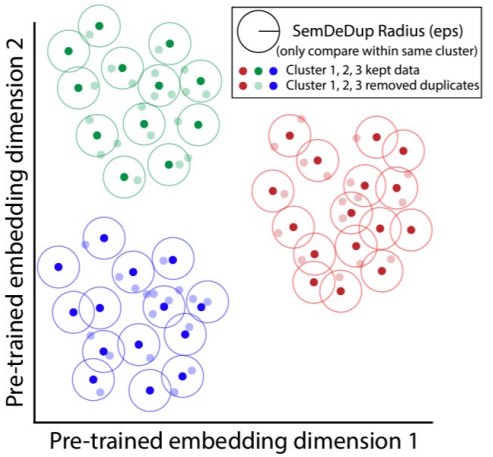
Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy
  - Approximate Frequency Techniques (SoftDeDup): Deduplicate by reweighting samples instead of pruning samples, preventing the loss of potentially valuable information
    - **S1:** Compute the frequency of each n-gram across all the samples
    - **S2:** Calculate the commonness of each sample by multiplying the frequencies of all the n-grams that appear in the document.
    - **S3:** Samples with higher commonness are more likely to be duplicates and thus be down-weighted.

He, N., Xiong, W., Liu, H., et al. SoftDedup: an Efficient Data Reweighting Method for Speeding Up Language Model Pre-training. ACL, 2024.

Data Deduplication

- **Motivation: Pretraining prefers to remove duplicates, ensuring greater coverage with less redundancy**
  - Data Deduplication: Remove duplicates to enhance training or sometimes improve accuracy
  - Embedding-Based Clustering Techniques:
    - Use pretrained models for semantic deduplication
    - **S1:** Embed each data point into a vector in the embedding space using existing LLM. The vector contains the tokens and the context between them, thus providing the semantic information.
    - **S2:** Cluster data points using k-means
    - **S3:** Within each cluster, pairwise cosine similarities between data points are calculated.
    - **S4:** For identified duplicates within a cluster, only the point with the lowest cosine similarity to the cluster centroid is kept, and the others are removed.



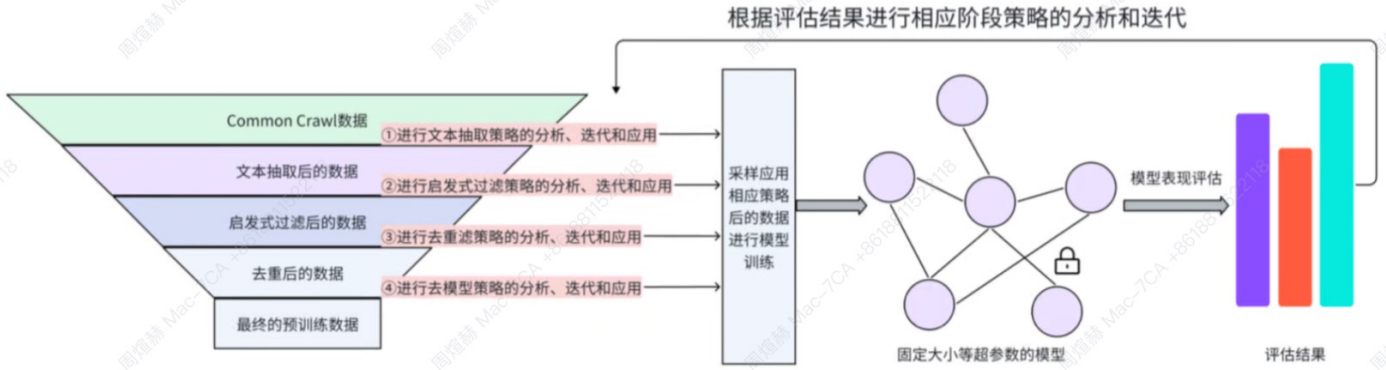
Abbas, A., Tirumala, K., Simig, D., Ganguli, S., & Morcos, A. S. (2023). Semdedup: Data-efficient learning at web-scale through semantic deduplication.

Data Deduplication

- **Takeaways**
  - Redundant data negatively impacts LLM performance by reducing generalization ability and increasing overfitting.
  - Deduplication improves training efficiency, prevents memorization of repeated patterns, and mitigates bias.
  - Challenges include efficiently **encoding semantic content** for comparison and **scalability** of deduplication methods for large datasets.
  - Future Directions: 1) Enhancing accuracy in detecting semantically similar but structurally different duplicates; 2) Developing fair deduplication strategies that preserve underrepresented groups.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.



	A	B	C	D	E	F	G	H
1	符号语义异常				长度异常			
2		head	mid	tail		head	mid	tail
3	句子以:: ,, '"?? .。开头	1.15	0.78	2.38	单行字符超过1000	1.89	0.84	1.36
4	多余的逆转义字符	0.31	0.49	2.01	连续空白行过多	0.00	0.00	0.00
5	中文中夹杂英文符号 [一-龥][!?.,:][一-龥]	2.41	2.39	11.01	等长文本连续出现50次以上	0.01	0.01	0.10
6	中文中夹杂英文符号(不含句号)	0.00	0.00	0.00	单行文本<5 连续出现5次以上	0.01	0.02	0.09
7	数字序号为句号 ^\d{1,3}。.+	0.08	0.11	0.33	空行过多，占比超过30%	0.00	0.00	0.00
8	中文文本中出现多余的空格	1.17	0.93	2.16	内容长度大于1000万	0.00	0.00	0.00
9	markdown中引用深度大于3	0.00	0.00	0.00	单条数据长度小于等于10	0.00	0.00	0.00
10	标题中含有多余的#	0.01	0.01	0.04	单条数据长度小于等于100	0.00	0.00	0.00
11	句子末尾出现冒号或左侧符号	0.35	0.43	0.57				
12	一行中只存在一个标点符号	0.01	0.01	0.00				
13	单个汉字换行连续出现三次及以上	0.00	0.00	0.00				
14								
15	内容语义异常				内容重复			
16		head	mid	tail		head	mid	tail
17	整行为“text”	0.00	0.00	0.00	相同的标点符号连续出现多次	0.79	0.99	4.19
18	去掉前后空白字符整行为数字	0.01	0.03	0.00	在行末尾出现连续六个中文句号	0.03	0.11	0.66
19	连续行序号异常	0.17	0.19	0.64	内容重复	0.23	0.40	0.26

Data Filtering

Category	Objective	Methods
Sample-level Filtering	Remove low-quality samples	Perplexity Measuring [383], [61], [288], [239], [238] Influence Assessment [254], [168] Clustering [45], [436] Model Scoring [411], [264], [345] Mixed Methods [285], [84], [126]
Content-level Filtering	Remove partial-noising samples	Privacy Anonymization [275], [268] Image & Video Filtering [437], [216], [390]

Data Filtering

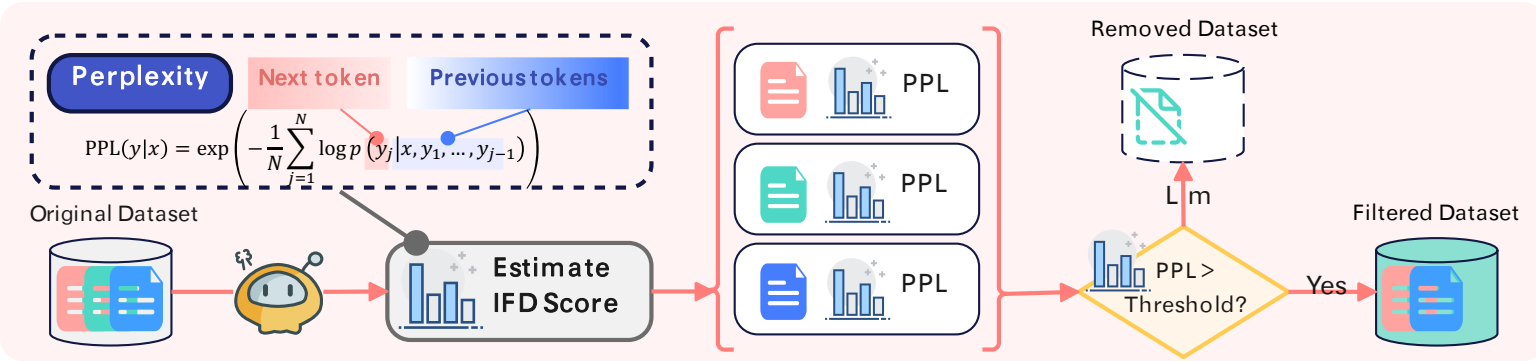
- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Perplexity Measuring: Filter by sample response generation difficulty.
    - Higher perplexity score indicates higher difficulty for model to generate sample response.
    - Filter samples by generating their responses using an individual model and assessing their perplexity scores.
    - Perplexity score is calculated by conditional probabilities: $\mathrm{PPL}(y|x) = \exp\left(-\frac{1}{N}\sum_{j=1}^{N}\log p\left(y_j|x, y_1, \ldots, y_{j-1}\right)\right)$
    - Given a sentence "I love machine learning", its perplexity score is calculated by

$$P(\mathrm{i}) = 0.2,\ P(\mathrm{love} \mid \mathrm{i}) = 0.1,\ P(\mathrm{machine} \mid \mathrm{i,love}) = 0.05,\ P(\mathrm{learning} \mid \mathrm{i,love,machine}) = 0.01$$

$$PPL(\mathrm{learning|i,love,machine}) = \exp\left(-\frac{1}{4}(\log(0.2) + \log(0.1) + \log(0.05) + \log(0.01))\right) = 17.78279$$

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Perplexity Measuring: Filter by sample response generation difficulty.
    - **S1:** Compute perplexity score with a surrogate model (smaller size for faster computing).
    - **S2:** Rank samples by perplexity values and select subsets based on the criteria (e.g., top-ranked domains or medium/high perplexity samples).
    - **S3:** Train larger models on the optimal subset.



Ankner, Z., Blakeney, C., Sreenivasan, K., Marion, M., Leavitt, M. L., & Paul, M. (2024). Perplexed by perplexity: Perplexity-based data pruning with small reference models.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Perplexity Measuring: Filter by sample response generation difficulty.
    - **Learning Percentage (LP):** Models learn easier samples first and harder ones later. Assess sample difficulty by measuring its perplexity drop ratio during training.
    - **S1:** Train a model to track sample perplexity across epochs.
    - **S2:** Calculate the learning percentage after the first epoch $LP(1)$.
    - **S3:** Rank samples and split them into three parts (hardest, medium, easiest).
    - **S4:** Train larger models on the hardest subset.

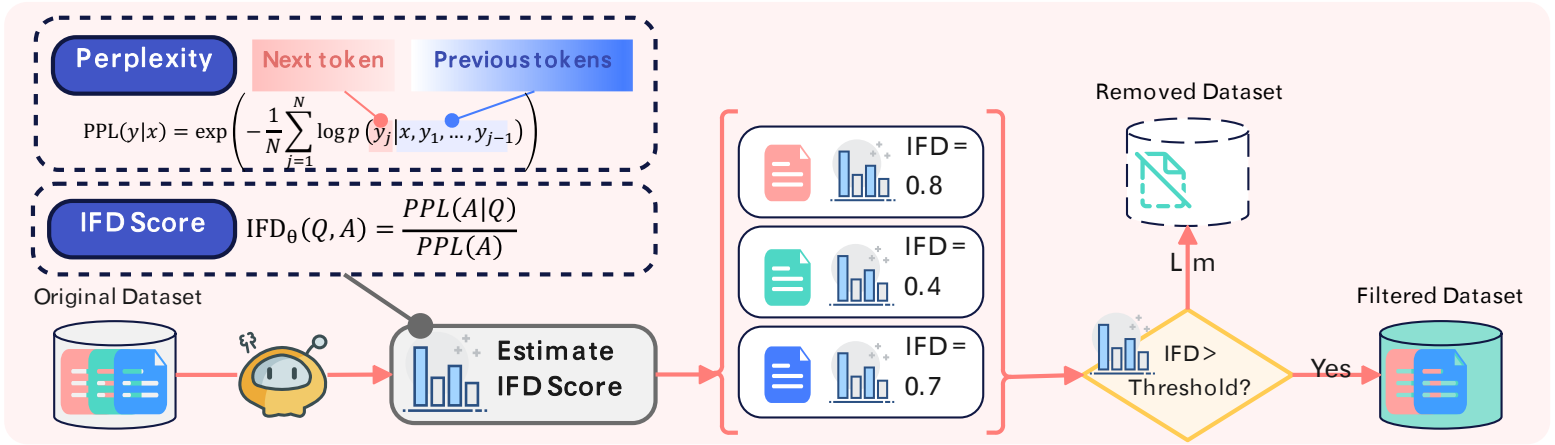
$$\mathcal{LP}(i) = \frac{\mathcal{P}_{i-1} - \mathcal{P}_i}{\mathcal{P}_0 - \mathcal{P}_n}$$

measures the perplexity drop ratio of a sample between the specific epoch i and the whole training procedure.

Mekala, D., Nguyen, A., & Shang, J. (2024). Smaller language models are capable of selecting instruction-tuning training data for larger language models.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Perplexity Measuring: Filter by sample response generation difficulty.
    - **Instruction-Following Difficulty (IFD):** Measures how much the instruction + input part of the sample would affect sample perplexity by comparing perplexity w/o the part.
    - **S1:** Compute IFD score using an existing model.
    - **S2:** Rank samples by IFD scores.
    - **S3:** Train model on the samples with higher IFD scores.



Li, M., Zhang, Y., Li, Z. et al.: From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning.
Li, M., Zhang, Y., He, S. et al.: Superfiltering: Weak-to-strong data filtering for fast instruction-tuning.

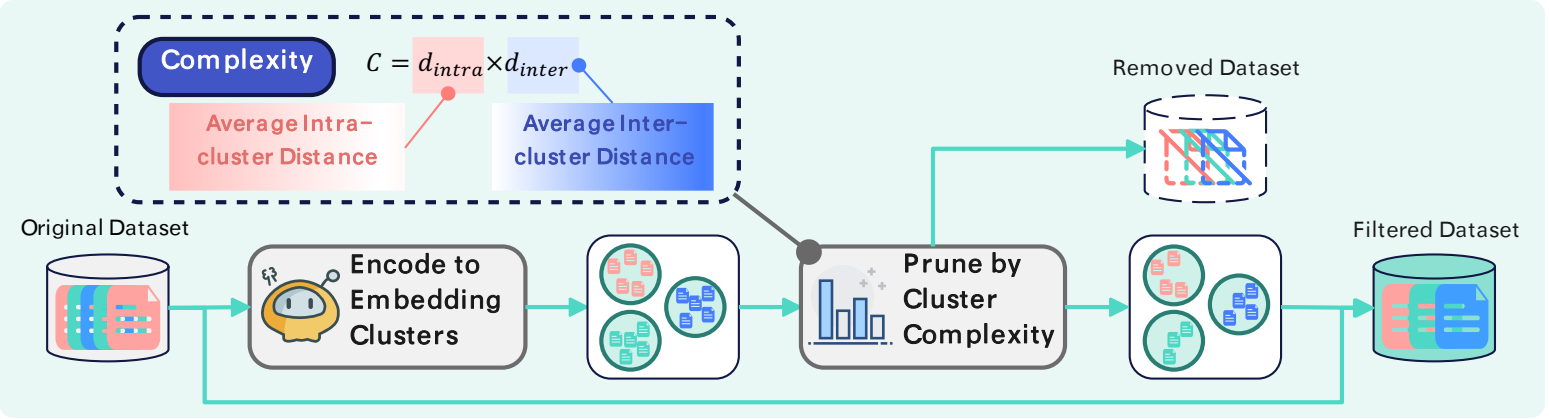
Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Influence Assessment: Filter by how much a sample would affect model parameters or model performance (e.g., accuracy, fairness).
    - **DEALRec:** Estimate the influence of a sample when removing/upweighting samples.
      - Estimate model parameter change through $\widehat{\theta_{-s}} - \hat{\theta} \approx \frac{1}{n} H_{\hat{\theta}}^{-1} \nabla_{\theta \mathcal{L}}(s, \hat{\theta})$, and extend it to
      - Estimate model performance change through $I_{\text{remove, loss}}(s, \mathcal{D}) = \frac{1}{n} \sum_i \frac{1}{n} \nabla_{\theta \mathcal{L}}(s_i, \hat{\theta})^{\top} H_{\hat{\theta}}^{-1} \nabla_{\theta \mathcal{L}}(s, \hat{\theta})$
    - **SHED:** Assess the influence of a sample on model performance using Shapley value.
      - Iteratively removing n samples to measure their contribution to model performance by comparing model performance w/o this subset.

Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., & Chua, T. S. (2024, July). Data-efficient Fine-tuning for LLM-based Recommendation. In Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval (pp. 365-374).
He, Y., Wang, Z., Shen, Z., Sun, G., Dai, Y., Wu, Y., ... & Li, A. (2024). Shed: Shapley-based automated dataset refinement for instruction fine-tuning. arXiv preprint arXiv:2405.00705.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Clustering: Groups similar samples together, allowing selection within clusters to reduce redundancy and across clusters to increase diversity.
    - **S1:** Encode samples into embeddings and cluster similar samples using cosine similarity.

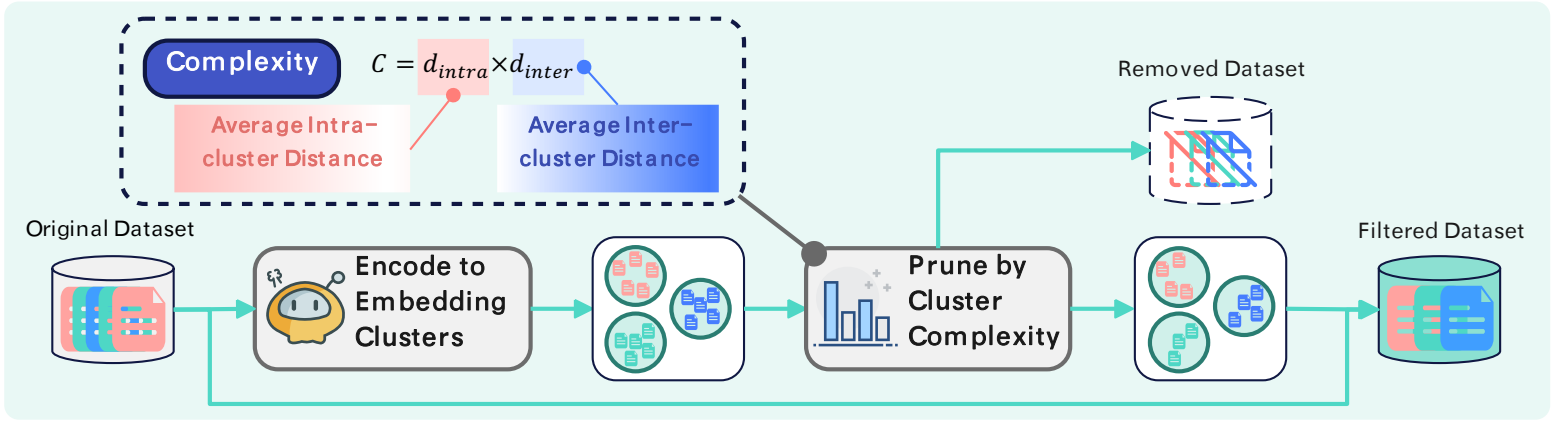


Abbas, A., Rusak, E., Tirumala, K., Brendel, W., Chaudhuri, K., & Morcos, A. S. (2024). Effective pruning of web-scale datasets based on complexity of concept clusters.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Clustering: Groups similar samples together, allowing selection within clusters to reduce redundancy and across clusters to increase diversity.
    - **S1:** Encode samples into embeddings and cluster similar samples using cosine similarity.
    - **S2:** Calculate cluster complexity based on intra-cluster and inter-cluster distances:

$$C = d_{intra} \times d_{inter}$$

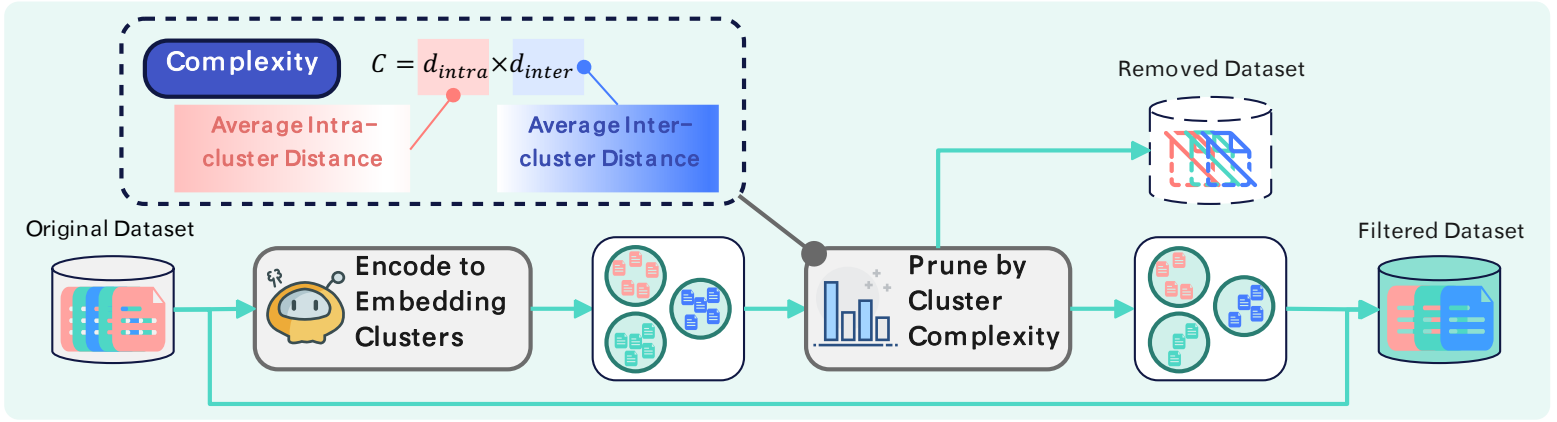


Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Clustering: Groups similar samples together, allowing selection within clusters to reduce redundancy and across clusters to increase diversity.
    - **S1:** Encode samples into embeddings and cluster similar samples using cosine similarity.
    - **S2:** Calculate cluster complexity based on intra-cluster and inter-cluster distances:

$$C = d_{intra} \times d_{inter}$$

    - **S3:** Resample clusters probabilistically to balance diversity and quality.



Abbas, A., Rusak, E., Tirumala, K., Brendel, W., Chaudhuri, K., & Morcos, A. S. (2024). Effective pruning of web-scale datasets based on complexity of concept clusters.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Model Scoring: Use LLMs to evaluate sample quality explicitly or implicitly through prompt engineering or human-labeled data.
    - **S1:** Prompt GPT-3.5-turbo to compare pairs of samples along four quality criteria (writing style, fact & trivia amount, educational value, and the expertise required to understand)

Wettig, A., Gupta, A., Malik, S., & Chen, D. (2024). Qurating: Selecting high-quality data for training language models.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Model Scoring: Use LLMs to evaluate sample quality explicitly or implicitly through prompt engineering or human-labeled data.
    - **S1:** Prompt GPT-3.5-turbo to compare pairs of samples along four quality criteria (writing style, fact & trivia amount, educational value, and the expertise required to understand)
    - **S2:** Record binary confidence $p_{B \succ A} \in [0,1]$ and translate it into probabilistic sample quality rating $p_{B \succ A} = \sigma(s_B - s_A)$ through Bradley-Terry model.

Wettig, A., Gupta, A., Malik, S., & Chen, D. (2024). Qurating: Selecting high-quality data for training language models.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Model Scoring: Use LLMs to evaluate sample quality explicitly or implicitly through prompt engineering or human-labeled data.
    - **S1:** Prompt GPT-3.5-turbo to compare pairs of samples along four quality criteria (writing style, fact & trivia amount, educational value, and the expertise required to understand)
    - **S2:** Record binary confidence $p_{B \succ A} \in [0,1]$ and translate it into probabilistic sample quality rating $p_{B \succ A} = \sigma(s_B - s_A)$ through Bradley-Terry model.
    - **S3:** Train a rating model on these quality ratings.

Wettig, A., Gupta, A., Malik, S., & Chen, D. (2024). Qurating: Selecting high-quality data for training language models.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Model Scoring: Use LLMs to evaluate sample quality explicitly or implicitly through prompt engineering or human-labeled data.
    - **S1:** Prompt GPT-3.5-turbo to compare pairs of samples along four quality criteria (writing style, fact & trivia amount, educational value, and the expertise required to understand)
    - **S2:** Record binary confidence $p_{B \succ A} \in [0,1]$ and translate it into probabilistic sample quality rating $p_{B \succ A} = \sigma(s_B - s_A)$ through Bradley-Terry model.
    - **S3:** Train a rating model on these quality ratings.
    - **S4:** Use the rating model predict quality ratings for new samples on each criterion.

Wettig, A., Gupta, A., Malik, S., & Chen, D. (2024). Qurating: Selecting high-quality data for training language models.

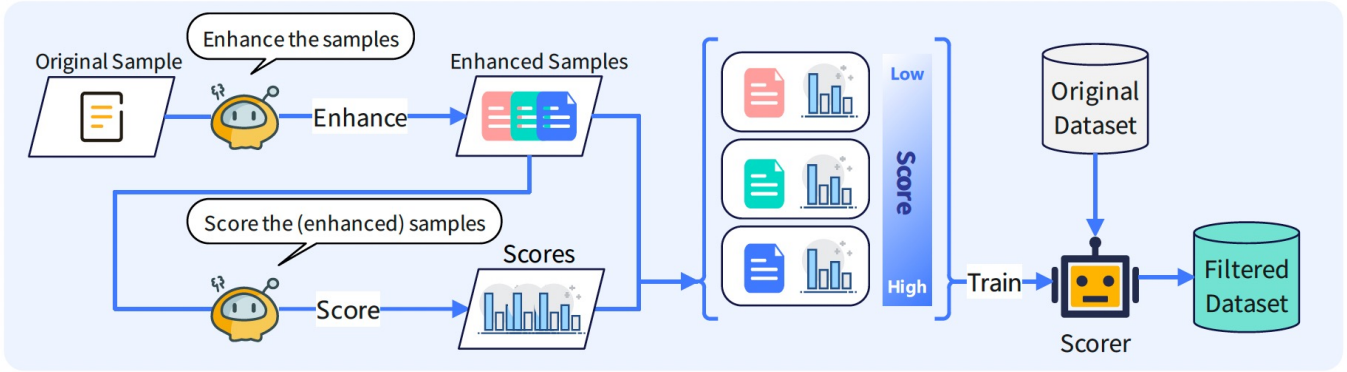
Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Model Scoring: Use LLMs to evaluate sample quality explicitly or implicitly through prompt engineering or human-labeled data.
    - **S1:** Prompt GPT-3.5-turbo to compare pairs of samples along four quality criteria (writing style, fact & trivia amount, educational value, and the expertise required to understand)
    - **S2:** Record binary confidence $p_{B \succ A} \in [0,1]$ and translate it into probabilistic sample quality rating $p_{B \succ A} = \sigma(s_B - s_A)$ through Bradley-Terry model.
    - **S3:** Train a rating model on these quality ratings.
    - **S4:** Use the rating model predict quality ratings for new samples on each criterion.
    - **S5:** Resample new samples by the predicted ratings.

Wettig, A., Gupta, A., Malik, S., & Chen, D. (2024). Qurating: Selecting high-quality data for training language models.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove low-quality or noisy samples and ensure diversity in the selected subset.
  - Model Scoring: Use LLMs to evaluate sample quality explicitly or implicitly through prompt engineering or human-labeled data.
    - **S1:** Prompt ChatGPT to evolve the samples along instruction complexity and response quality, and to score these evolved samples
    - **S2:** Train two scorers (complexity and quality) on the evolved samples with their scores.
    - **S3:** Use both scorers to score new samples. Multiply both scores to form the final score
    - **S4:** Resample samples by the final score



Liu, W., Zeng, W., He, K., Jiang, Y., & He, J. (2023). What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove noise or sensitive information within samples.
    - Clean text by removing invalid characters, unnecessary texts, or harmful content (e.g., bias ranging from gender and racial stereotypes to cultural and socioeconomic prejudices).
      - SEAL: Train a selector based on a safety-aligned model (e.g., Merlinite-7b) using bi-level optimization:
        - Minimize the safety loss on the safe dataset
        - Minimize the fine-tuning loss on the filtered dataset during training to make the selector prioritize safe and high-quality samples in selecting data

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), 1-21.
Shen, H., Chen, P. Y., Das, P., & Chen, T. (2024). Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection.

Data Filtering

- **Motivation: Pretraining prefers to remove low-quality or noisy samples, simplifying training while retaining performance**
  - Data Filtering: Remove noise or sensitive information within samples.
    - Clean by removing invalid characters, unnecessary texts, harmful content.
    - Detect and anonymize private/sensitive information while preserving semantics.
      - Use NER models (spaCy, Flair, etc.) to tag personally identifiable information (e.g., names, addresses) and replace tagged entities with placeholders (e.g., `[NAME]`, `[LOCATION]`), or
      - **DeID-GPT:** Prompt LLM to redact PII within the given text:

Please de-identify the following clinical notes by replacing any terms that could be a name, an address, a date, or an ID with the term '[redacted]'. [content]

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023, May). Analyzing leakage of personally identifiable information in language models. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 346-363). IEEE.
Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., ... & Li, X. (2023). Deid-gpt: Zero-shot medical text de-identification by gpt-4.

Data Filtering

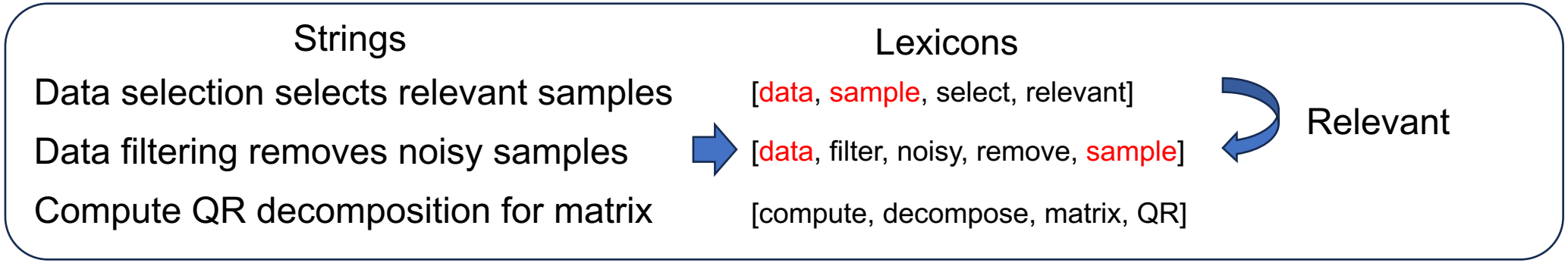
- **Takeaways**
  - Data filtering aims to remove low-quality or sensitive samples from training datasets, reducing computational overhead while maintaining or improving model performance.
  - Challenges include balancing data reduction with model performance and ensuring diversity while filtering out redundant or noisy samples.
  - **Sample-level filtering** focuses on removing entire low-quality samples based on metrics like perplexity, influence assessment, clustering, entropy, and model scoring.
  - **Content-level filtering** targets partial noise or sensitive content within samples rather than removing entire entries.
  - Future Directions include improving efficiency for massive datasets, enhancing accuracy in detecting low-quality or noisy text and developing better algorithms for locating potential sensitive information.

Data Selection

Method	Stage	Evaluation Metric
Similarity	Pre-training, Fine-tuning	Cosine Similarity [423] Bag-of-Words Similarity [421] Lexicon Set Overlap [321] Bayes-based Selection [80]
Optimization	Fine-tuning	Linear Search [130] Gradient-Influence Search [417] Kernel-Density Regularization [269]
Model	Pre-training	Logits-based LM-Score [465]

Data Selection

- **Motivation: Adapt LLMs to specific domains (e.g., medical or legal).**
  - Data Selection: Select subsets of already well-cleaned data samples to align LLMs with target tasks while maintaining generalization.
  - Lexicon Set Overlap: Measures relevance of a dataset to a specific domain using overlap between lexicons.
    - Break strings into lexicon sets, and select samples with large lexicon intersection to the target task

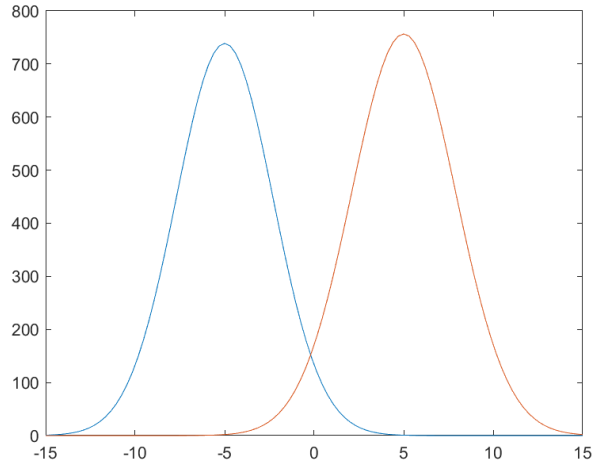


Qin, R., Xia, J., Jia, Z., Jiang, M., Abbasi, A., Zhou, P., ... & Shi, Y. (2024, June). Enabling on-device large language model personalization with self-supervised data selection and synthesis. In Proceedings of the 61st ACM/IEEE Design Automation Conference (pp. 1-6).

Data Selection

- **Motivation: Adapt LLMs to specific domains (e.g., medical or legal).**
  - Data Selection: Select subsets of already well-cleaned data samples to align LLMs with target tasks while maintaining generalization.
  - Bag-of-Words Similarity: Utilize bag-of-words to compute feature distributions for raw and target data.
    - **S1:** Represent raw and target data as bag-of-words features.
    - **S2:** Estimate importance weights $w_i = \frac{\widehat{p_{\text{feat}}}(z_i)}{\widehat{q_{\text{feat}}}(z_i)}$.
    - **S3:** Resample raw data with probability $\frac{w_i}{\sum_{i=1}^{N} w_i}$ to match the target distribution.

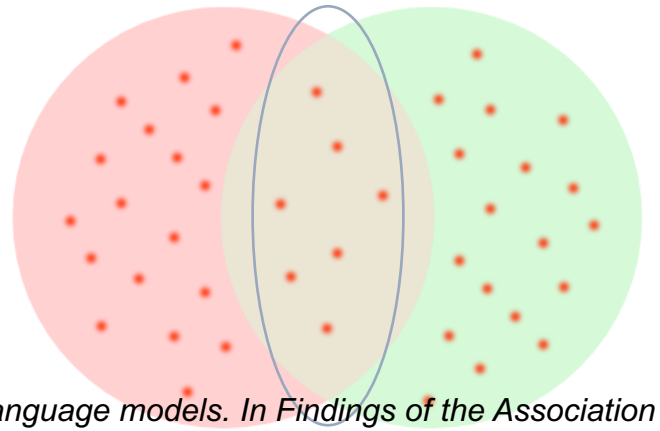
"Data selection selects relevant samples" → p(data) = 1/5, p(sample) = 1/5, p(select) = 2/5, p(relevant) = 1/5 →



Xie, S. M., Santurkar, S., Ma, T., & Liang, P. S. (2023). Data selection for language models via importance resampling. Advances in Neural Information Processing Systems, 36, 34201-34227.

Data Selection

- **Motivation: Adapt LLMs to specific domains (e.g., medical or legal).**
  - Data Selection: Select subsets of already well-cleaned data samples to align LLMs with target tasks while maintaining generalization.
  - Cosine Similarity: Compare embeddings of task-specific labeled data and unlabeled data.
    - **S1:** Encode both labeled and unlabeled data into embeddings
    - **S2:** Measure similarity between embeddings using cosine similarity.
    - **S3:** Select unlabeled samples that align closely with the task's embedding distribution.



Xie, Y., Aggarwal, K., & Ahmad, A. (2024, August). Efficient continual pre-training for building domain specific large language models. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 10184-10201).

Data Selection

- **Motivation: Adapt LLMs to specific domains (e.g., medical or legal).**
  - Data Selection: Select subsets of already well-cleaned data samples to align LLMs with target tasks while maintaining generalization.
  - Optimization-based Selection: Select subsets towards reducing model loss and improving model performance on the target tasks.
    - **Approach 1:** Minimizes model loss on target tasks using linear datamodels.
      - **S1:** Use a linear datamodel $\tau_{\theta_x}(1_S)$ to estimate how training samples affect model loss.
      - **S2:** Adjust characteristic vector 1_S to reflect the subset and estimate parameters θ_x via regression.
      - **S3:** Select the subset S of size k that minimizes the loss $\widehat{L_{D_{\text{targ}}}}(S)$.

Engstrom, L., Feldmann, A., & Madry, A. (2024). Dsdm: Model-aware dataset selection with datamodels.

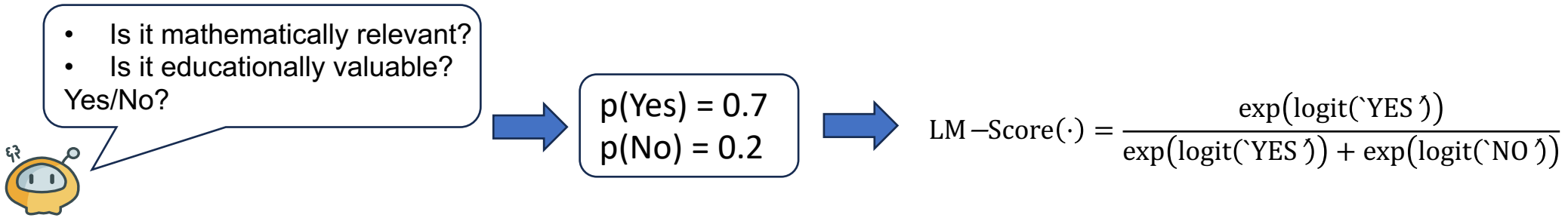
Data Selection

- **Motivation: Adapt LLMs to specific domains (e.g., medical or legal).**
  - Data Selection: Select subsets of already well-cleaned data samples to align LLMs with target tasks while maintaining generalization.
  - Optimization-based Selection: Select subsets towards reducing model loss and improving model performance on the target tasks.
    - **Approach 2:** Identifies impactful data by analyzing gradient similarities.
      - **S1:** Fine-tune the model on a random subset using LoRA.
      - **S2:** Compute Adam LoRA gradients for each sample and project them into lower-dimensional features.
      - **S3:** For downstream tasks, calculate gradient features of validation samples.
      - **S4:** Estimate influence using cosine similarity and select top-scoring samples.

Xia, M., Malladi, S., Gururangan, S., Arora, S., & Chen, D. (2024). Less: Selecting influential data for targeted instruction tuning.

Data Selection

- **Motivation: Adapt LLMs to specific domains (e.g., medical or legal).**
  - Data Selection: Select subsets of already well-cleaned data samples to align LLMs with target tasks while maintaining generalization.
  - Model-Based: Leverages LLMs themselves to evaluate and select high-quality samples.
    - S1: Prompt the LLM to assess relevance and educational value of each sample.
    - S2: Extract logits for responses ("Yes"/"No") and compute LM-Score
    - S3: Calculate composite score and select samples with highest scores.



Zhang, Y., Luo, Y., Yuan, Y., & Yao, A. C. (2024). Autonomous data selection with language models for mathematical texts. In ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models.

Data Selection

- **Takeaways**
  - Data selection involves choosing subsets of already cleaned data to adapt large language models (LLMs) to specific domains, such as medical or legal fields.
  - Three main types of filtering are discussed: Similarity-Based, Optimization-Based and Model-Based selections:
    - Similarity-based methods select samples with similar feature to the target task.
    - Optimization-based methods select samples that improve model performance on the target task.
    - Model-based methods prompt LLM models to select relevant samples to the target task.
  - Challenges include computational efficiency and robust generalization across tasks.
  - Future Directions include improving efficiency for massive datasets, enhancing accuracy in extracting domain patterns and developing better algorithms that generalizes well with the incoming data.

Data Mixing

Taxonomy	Stage	Methods	Traits
Before Training (Human Experience)	Pre-training	Multi-Source Data Adjusting [139], [347]	Intuitive and easy to implement, suitable for rapid experimentation.
		Entropy-Based Mixing [152]	Low computation cost with quality quantification by entropy.
Before Training (Model-Based Optimization)	Pre-training	Linear Regression Model [263]	Only 10% of DoReMi's [420] computational resources are required. Simultaneously train hundreds of small models to accelerate optimization.
	Pre-training	Bivariate Data Mixing Law [152]	Avoid iterative training of proxy models (low computational costs). Show relation between loss and training steps.
	Continual Pre-training	Chinchilla Scaling Law [323]	Support knowledge transferring to new domains (↓ over 95% training costs).
	Pre-training	Exponential Functions [439]	Support datasets without explicit domain division.
	Continual Pre-training	Power-law Function [160]	Compared to single-objective optimization like [323] [160] ensures that domain performance improvement does not compromise general capabilities.
	Pre-training	Classification Model [251]	Reverse engineering for finding the suitable data recipe of LLMs.
During Training (Bilevel Optimization)	Pre-training	Calculate domain contribution by gradient inner products[135]	Requires a proxy model, performances well in OOD datasets.
	Fine-tuning	Dynamically adjust weights by gradient alignment values [302]	Multiple applications like multilingual training, instruction following, large-scale data reweighting
During Training (Distributionally Robust Optimization)	Pre-training	Group DRO [420]	For pre-training, smooth adjusting to prevent abrupt weight changes
	Fine-tuning	Task-level DRO [278]	For fine tuning, quick response to task difficulty changes

Data Mixing For LLM training

- **Challenge: How to optimize dataset ratios for better performance and training efficiency.**
  - Before Training(Human Experience): Empirically set and experiment different ratios of datasets based on various factors (e.g., complexity and diversity of the datasets) that likely improve LLMs' abilities.
    - e.g., Two-phase training, which focuses diversity in data in phase 1, high quality data such as math and code in phase 2, explores their effect with 5 blends each.

Category	Domain	Blend1	Blend2	Blend3	Blend4	Blend5
Web Crawl	-	65.0	65.0	58.0	59.0	70.0
High Quality	Math	1.9	1.9	1.9	2.9	1.9
	Wiki	0.1	0.1	0.1	0.1	0.1
	Code	15.0	8.0	15.0	20.0	13.0
Medium Quality	Books	5.5	9.0	9.0	5.5	4.5
	Papers	3.5	5.0	5.0	3.5	1.9
	CC_{dv}	4.0	6.0	6.0	4.0	3.6
Multilingual	-	5.0	5.0	5.0	5.0	5.0

Table 2: Phase-1 Blends (in %)

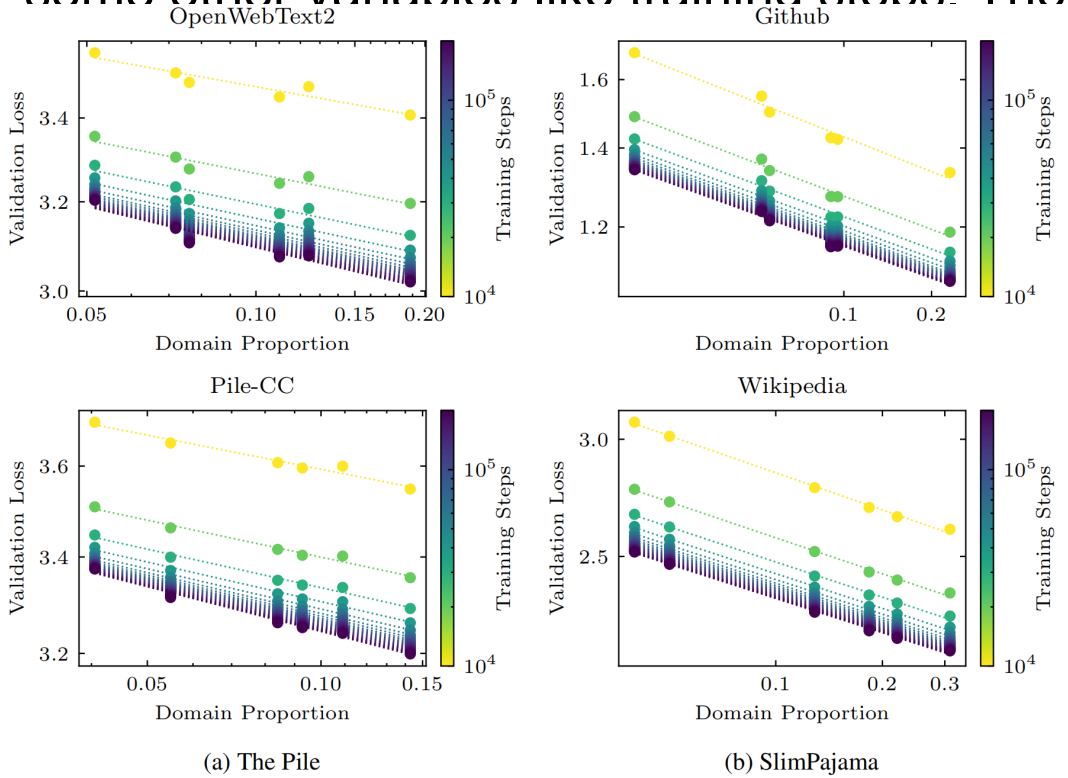
Category	Domain	Blend1	Blend2	Blend3	Blend4	Blend5
Web Crawl	-	31.0	35.0	31.0	40.0	35.0
High Quality	Math	24.0	24.0	24.0	24.0	29.0
	Wiki	1.0	1.0	1.0	1.0	1.0
	Code	20.0	25.0	29.0	20.0	20.0
Medium Quality	Books	8.0	4.0	4.0	4.0	4.0
	Papers	4.0	2.0	2.0	2.0	2.0
	CC_{dv}	7.0	4.0	4.0	4.0	4.0
Multilingual	-	3.7	3.7	3.7	3.7	3.7
Task Data	-	1.3	1.3	1.3	1.3	1.3

Table 3: Phase-2 Blends (in %)

Feng, S., Prabhumoye, S., Kong, K. et al.: Maximize your data's potential: Enhancing llm accuracy with two-phase pretraining. arXiv preprint arXiv:2412.15285 (2024)

Data Mixing For LLM training

- **Challenge: How to optimize dataset ratios for better performance and training efficiency.**
  - Before Training(Model-Based Optimazation): Model the relationship that depicts the relation between (i) the distribution of each domain or datasets, (ii) validation loss, and (iii) some other variables like training steps. Then find the optimal ratio based on the models.



(a) The Pile (b) SlimPajama

e.g., Based on the observation of data from experiments for scaling behaviour, Bivariate Data Mixing Law depicts the relation among domain's proportion, training steps and validation loss.

Ge, C., Ma, Z., Chen, D. et al.: Bimix: A bivariate data mixing law for language model pretraining (2025)

Data Mixing For LLM training

- **Challenge: How to optimize dataset ratios for better performance and training efficiency.**
  - During Training (Bilevel Optimization): Use a closed-loop optimization technique for two nested optimazation problem that ensures model parameters are optimized.
    - e.g., By measuring t**he contribution of a domain to other domains**, which is calculated by the dot product of the gradient of the domain and the sum of gradients from all other domains and the amount of gradients, dynamically adjust the domain by the contributions.

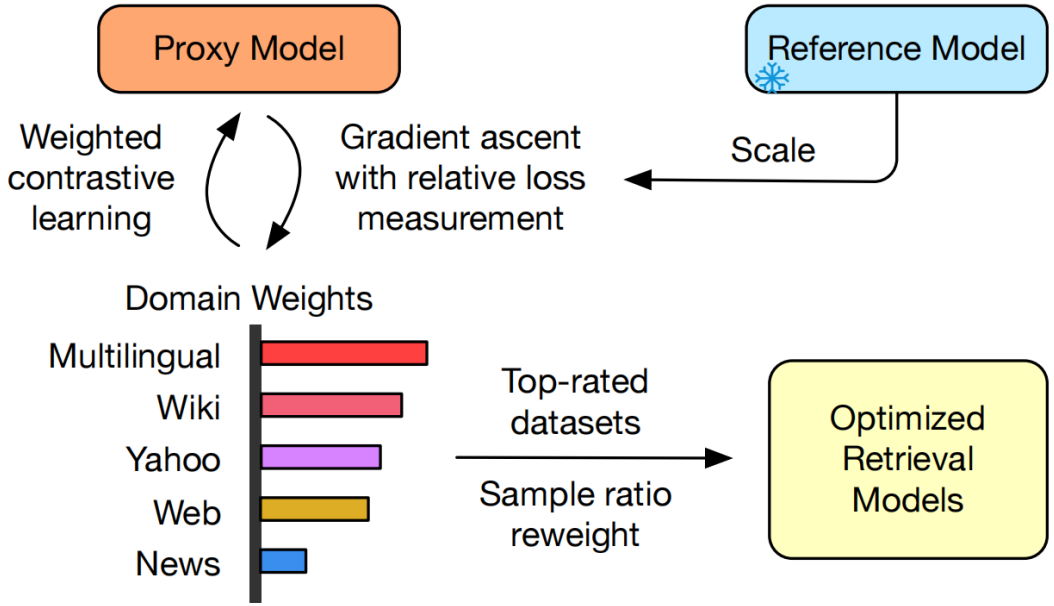
      Minimize the maximum loss across all domains
    - Train a larger model using the optimized domain ratios

Fan, S., Pagliardini, M., Jaggi, M.: Doge: Domain reweighting with generalization estimation. arXiv preprint arXiv:2310.15393 (2023)

Data Mixing For LLM training

- **Challenge: How to optimize dataset ratios for better performance and training efficiency.**
  - During Training (Distributionally Robust Optimization): Adopt Distributionally Robust Optimization (DRO) for a robust data mixing strategy (which can be sub-optimal but with low uncertainty)

e.g., Optimize By DRO using a small proxy model, which first trains a reference model, we have the validation loss of each domain for reference, then trains proxy model, by measuring the **loss difference to reference**, which **indicates the improvement potential**, dynamically adjust the domain weights, and tilt to domains with larger loss difference.



Ma, G., Ma, Y., Wu, X. et al.: Task-level distributionally robust optimization for large language model-based dense retrieval. arXiv preprint arXiv:2408.10613(2024)

Takeaways

- Human experience mixing takes least amount of cost to get a better data ratio for training by using the ratios given by these works, by experimenting like this still needs quite amount of training.
- For now, almost all mixing methods need a small proxy model for experiments then scale up to larger one.
- Model-based methods provide other variables like training step that help us see how these variables affect LLM performance with domain ratios.
- Optimize methods like Bilevel Optimization and DRO performs very well but find difficulties for applying on larger models (Largest model with 34B from recent works), as it’s done during training.

Data Synthesis and Distillation For LLM training

Stage	Category	Methods
Distillation	Reasoning Augmentation	Cot [353]
		Prompt with Tools [496]
	Data Augmentation	Prompt with Multi-Agent [467]
Pre-Training	Data Augmentation	Distillation + Fine Tuning + Prompt [481]
		Prompt [99], [98], [282], [93],[344], [92]
SFT	Data Augmentation	Prompt [233], [179], [260], [290]
	Reasoning Augmentation	Prompt [178], [173], [346]
		Human Label [253]
		Automated Label [399]
		High Quality Reasoning Data [442], [230]
RL	Prompts Optimization	Prompt [401]
	Human Feedback	RLHF [71]
		RLHF By LLM [476]
RAG	Privacy Protection	Prompt [450]

Data Synthesis and Distillation For LLM training

- **Motivation: Synthesize abundant and high-quality data for specific domains or use-cases**
  - Distillation: Design paradigms to prompt LLM to generate high-quality data to train a student LLM with less parameters to mimic the target model's generation ability.

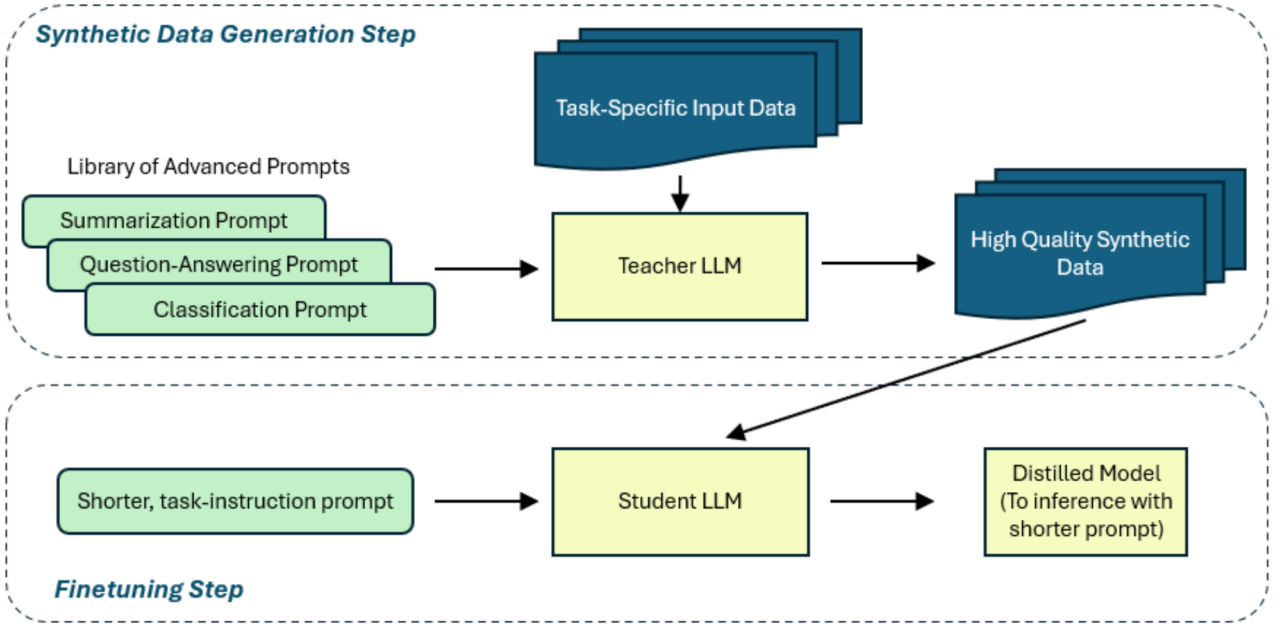


Figure 1: Overview of the proposed methodology for model distillation. Synthetic data is generated using advanced, task-engineered prompt while fine-tuning (hence inferencing) of student model uses shorter, less expensive vanilla prompt.

Data Synthesis and Distillation For LLM training

- Pretraining Data Augmentation: Design paradigms to prompt LLM to generate synthetic pre-training data for training performant LLM.
- Rephrasing: Rephrase raw corpora into styled corpora

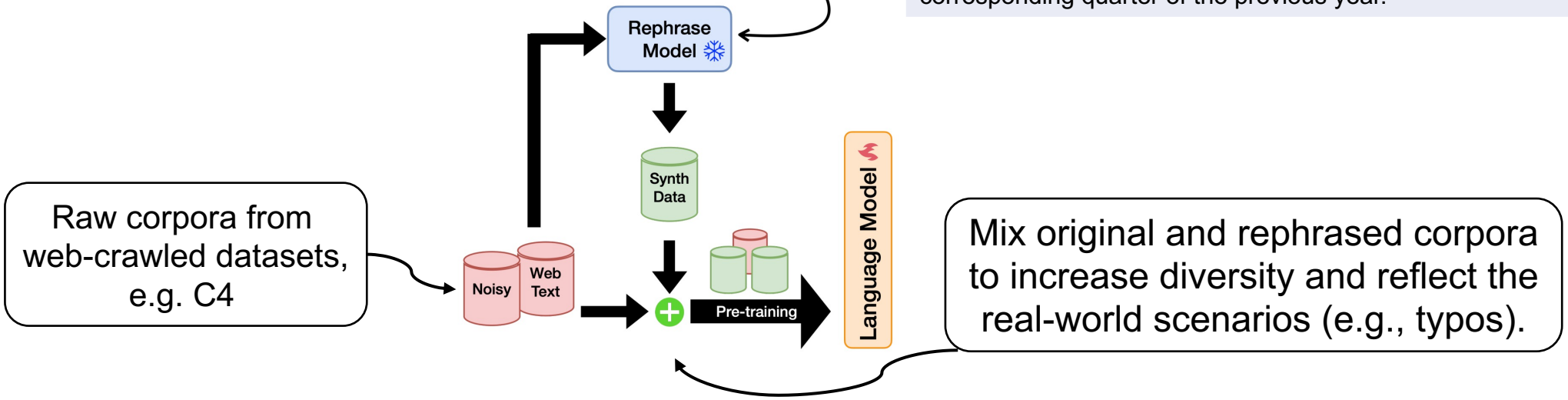
Rephrase raw corpora into various styles, from easier to harder, e.g.:

1. Naïve vocabulary and sentence structures.
2. Standardized encyclopedia-style expression.
3. Complex academic-style expression.
4. Multi-turn dialogue.

Example

Original: The stock rose $2.11, or about 11 percent, to close Friday at $21.51 on the New York Stock Exchange. Revenue in the first quarter of the year dropped 15 percent from the same period a year earlier.

Medium: The stock experienced an increase of approximately 11 percent, closing at $21.51 on the New York Stock Exchange on Friday, with a rise of $2.11. During the initial three months of the current year, there was a 15 percent decrease in revenue compared to the corresponding quarter of the previous year.

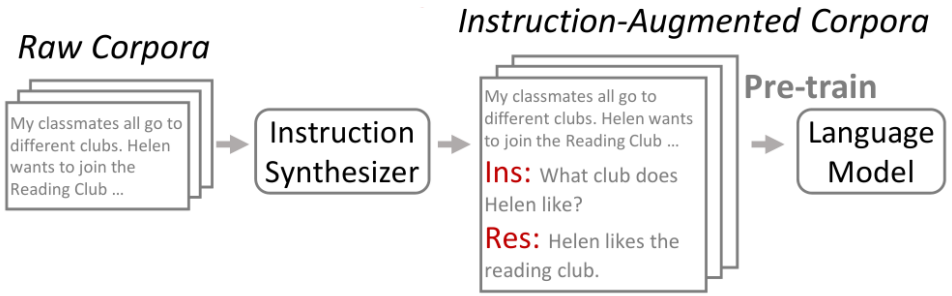


Maini, P., Seto, S., Bai, H. et al.: Rephrasing the web: A recipe for compute and data-efficient language modeling.

Data Synthesis and Distillation For LLM training

- Pretraining Data Augmentation: Design paradigms to prompt LLM to generate synthetic pre-training data for training performant LLM.
- Augmentation: Turn raw corpora into instruction-augmented corpora.

  Fine-tune an Instruction Synthesizer LLM that outputs various instruction-response pairs given the raw corpora for LLM pre-training



1. Curate a diverse set of existing NLP datasets where each example includes a context (raw text) and associated tasks (e.g., QA, sentiment, reasoning).
2. Fine-tune an instruction synthesizer on the curated data.
3. Convert raw corpora into instruction-augmented corpora by interleaving raw text with its synthesized instruction-response pairs

Cheng, D., Gu, Y., Huang, S., Bi, J., Huang, M., & Wei, F. (2024). Instruction pre-training: Language models are supervised multitask learners.

Data Synthesis and Distillation For LLM training

- Pretraining Data Augmentation: Design paradigms to prompt LLM to generate synthetic pre-training data for training performant LLM.
- Domain Data Synthesis: Turn raw corpora into instruction-augmented corpora.

  Synthesize Q&A pairs based on domain and language:

  1. For scientific QA data, prompt LLM with scientific data (e.g., Mathematics Stack Exchange) to generate QA pairs (a self-contained problem and comprehensive, step-by-step solution).
  2. For code QA data, prompt with examples from LeetCode to generate new, high-quality coding problems and their corresponding solutions.

  Mix and sample Chinese, English and synthetic data to enhance model's scientific ability.

Prompt for Scientific QA Synthesis

Instruction
Please gain inspiration from the following {Discipline Placeholder} content to create a high-quality {Discipline Placeholder} problem and solution. Present your output in two distinct sections: [Problem] and [Solution].

{Discipline Placeholder} Content
{Seed Snippet Placeholder}

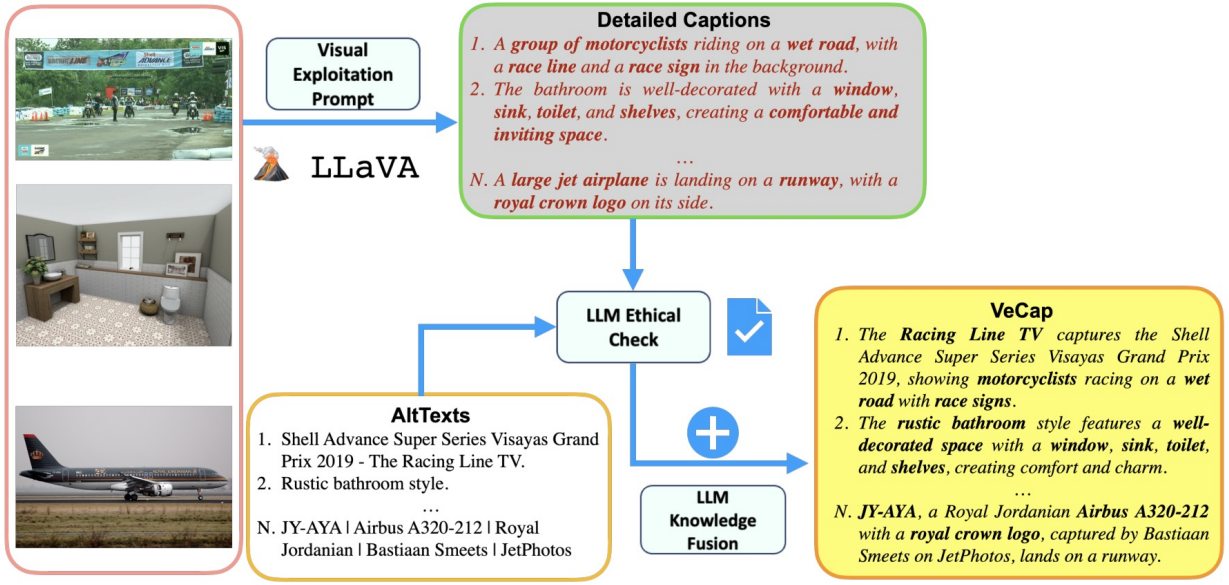
Guidelines
[Problem]: This should be **completely self-contained**, providing all the contextual information one needs to understand and solve the problem.

[Solution]: Present a comprehensive, step-by-step solution that solves the problem **correctly** and educates the student, around 250-350 words long. Clearly articulate the reasoning and methods used at each step, providing insight into the problem-solving process. Take care to format any equations properly using LaTeX or appropriate notation.

Chen, J., Chen, Z., Wang, J., Zhou, K., Zhu, Y., Jiang, J., ... & Wen, J. R. (2024). Towards effective and efficient continual pre-training of large language models. arXiv preprint arXiv:2407.18743.

Data Synthesis and Distillation For LLM training

- Pretraining Data Augmentation: Design paradigms to prompt LLM to generate synthetic pre-training data for training performant LLM.
- Image Caption Synthesis: Augment image captions for uncaptioned or simple-captioned source images via multimodal models
  - Rewrite existing captions using textual LLM for diversity
  - Write a detailed visual description of the image content (e.g., color, shape, surroundings, etc.) using multimodal LLM and fuse them with the original caption for more comprehensive final caption



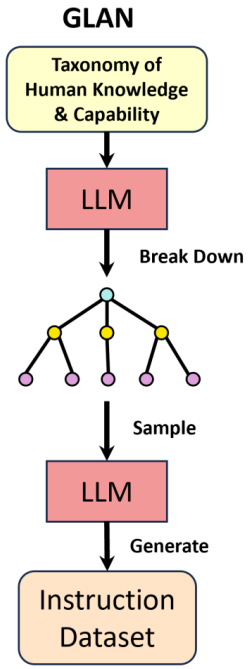
Fan, L., Krishnan, D., Isola, P., Katabi, D., & Tian, Y. (2023). Improving clip training with language rewrites. Advances in Neural Information Processing Systems, 36, 35544-35575.
Lai, Z., Zhang, H., Zhang, B., Wu, W., Bai, H., Timofeev, A., ... & Cao, M. (2024, September). Veclip: Improving clip training via visual-enriched captions. In European Conference on Computer Vision (pp. 111-127). Cham: Springer Nature Switzerland.

Data Synthesis and Distillation For LLM training

- SFT Task Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to improve specific domains (math, medicine, etc.), align LLM's knowledge to instructions, etc.
- Domain Data Synthesis: Improve specific domains (math, medicine) and enhance data diversity

  Synthesize domain data gradually and by domains in a course outline way for systematic coverage of knowledge

  1. Classify data into disciplines (e.g., math, chemistry, computer science)
  2. Generate syllabus (e.g., "Introduction to Calculus") and key concepts (e.g., "Limits") for a specific subject
  3. Create diverse questions and answers based on the concepts using LLM.



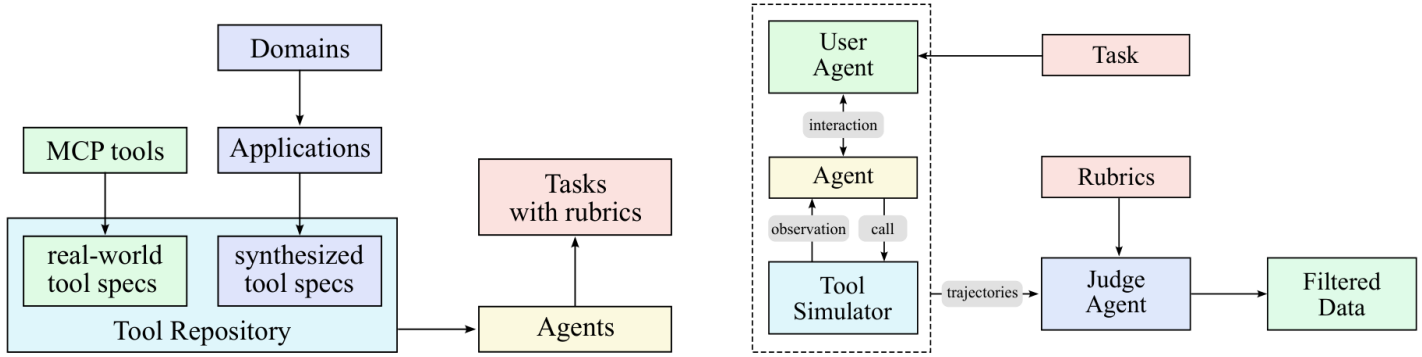
Li, H., Dong, Q., Tang, Z. et al.: Synthetic data (almost) from scratch: Generalized instruction tuning for language models. arXiv preprint arXiv:2402.13064 (2024)

Data Synthesis and Distillation For LLM training

- SFT Task Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to improve specific domains (math, medicine, etc.), align LLM's knowledge to instructions, etc.
- Agentic Data Synthesis: Make LLM solve complex tasks with unfamiliar tools

  Generate agentic training data that mimics the real-world tool-use scenarios:

  1. Build a diverse tool repository with MCP tools fetched from GitHub and synthetically generated tools created by domains (e.g., financial trading, robotics).
  2. Synthesize agents via system prompts for sampled toolsets and generates tasks paired with explicit success rubrics, covering a range from simple to complex operations.
  3. Produce realistic multi-turn interaction trajectories where agents use tools to complete tasks using multi-agent simulation.

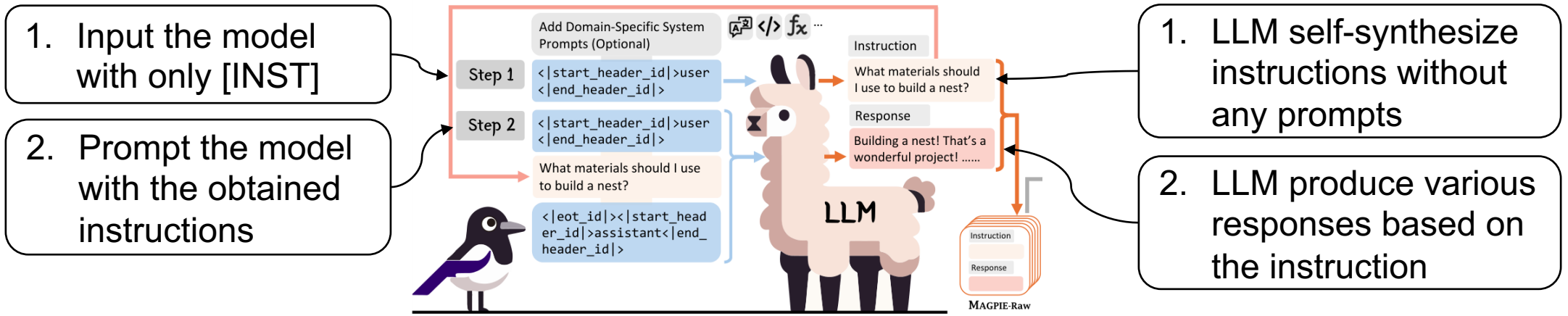


(a) Synthesizing tool specs, agents and tasks (b) Generating agent trajectories

Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., ... & Zhang, H. (2025). Kimi k2: Open agentic intelligence. arXiv preprint arXiv:2507.20534.

Data Synthesis and Distillation For LLM training

- SFT Task Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to improve specific domains (math, medicine, etc.), align LLM's knowledge to instructions, etc.
- Self-Synthesis of Instructed LLM: Make LLM self-synthesize user input by prompting the model with partial instruction template right before the user input.
  - Instructed LLMs were fine-tuned on templates like `...[INST]Instruction[/INST]...`, enabling LLM autoregressive output.



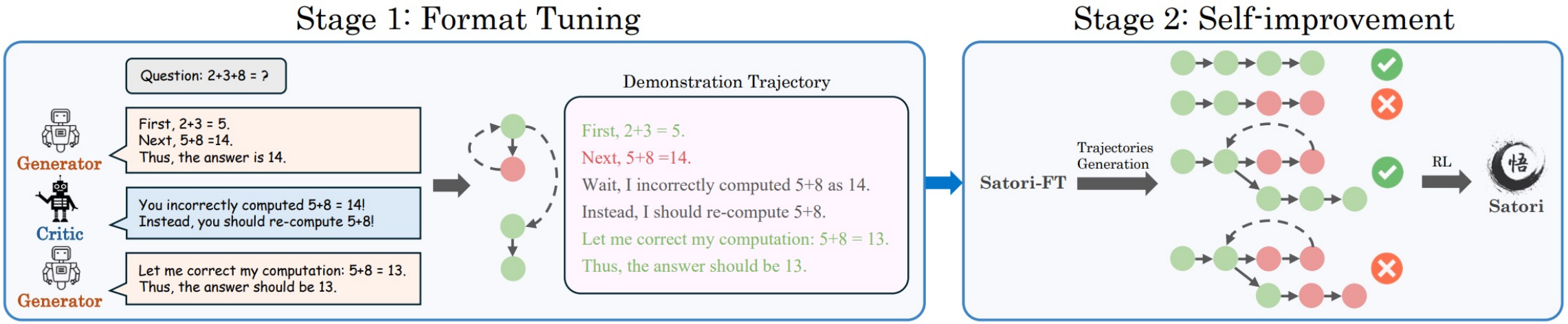
3. The instructions and responses form the dataset

Xu, Z., Jiang, F., Niu, L., Deng, Y., Poovendran, R., Choi, Y., & Lin, B. Y. (2024). Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. arXiv preprint arXiv:2406.08464.

Data Synthesis and Distillation For LLM training

- SFT Reasoning Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to enhance LLM's reasoning ability.
- LLM Exploring: Make models perform self-reflection, error correction, and alternative solution exploration during reasoning.

  Multiple LLM agents generate "Continue-Reflect-Explore" COAT-formatted reasoning chains to fine-tune a base model for COAT-formatted syntax mastery.



Shen, M., Zeng, G., Qi, Z. et al.: Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search. arXiv preprint arXiv:2502.02508 (2025)

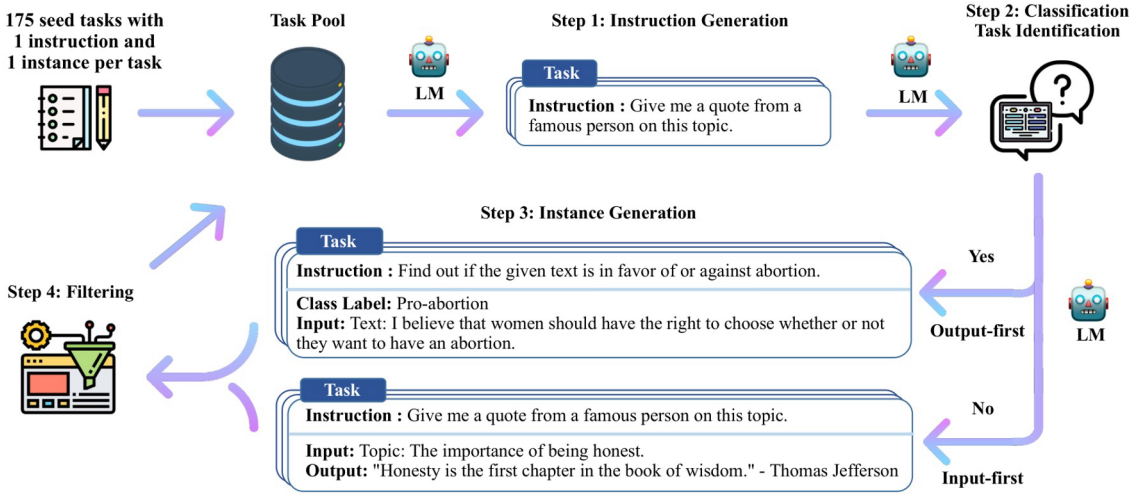
Data Synthesis and Distillation For LLM training

- SFT Reasoning Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to enhance LLM's reasoning ability.
- Student Model Collaboration Distillation: Improve data labeling by using multiple student models with generalized knowledge learned from teacher model.
  1. Generate initial pseudo-labels for unlabeled data using GPT-3.5.
  2. Split data with pseudo-labels into two subsets. Iteratively train two student models on the corresponding subset and use these models to generate new pseudo-labels for the other subset, achieving less noise and higher generalizability.
  3. Train a single model on all the data with refined labels, achieving near-supervised performance with 50 labeled samples (vs. 500 required traditionally).

Zhao, J., Zhao, W., Drozdov, A., Rozonoyer, B., Sultan, M. A., Lee, J. Y., ... & McCallum, A. (2023). Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. arXiv preprint arXiv:2311.08640.

Data Synthesis and Distillation For LLM training

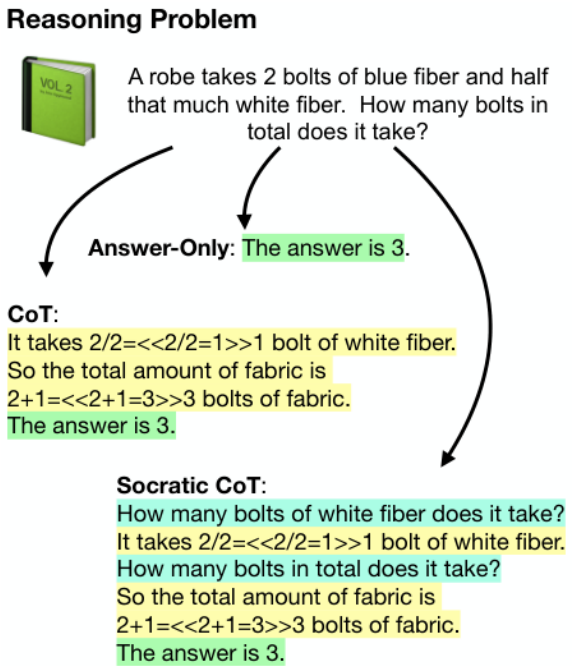
- SFT Reasoning Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to enhance LLM's reasoning ability.
- Self-Instruct: Improve instruction following using the model's generated data.
  - Prompt model to generate novel instructions from in-context examples sampled from seed task pool and determine its task type with few-shot labeled examples.
    - Input-first approach for non-classification tasks: generates an input based on the instruction, then produces the corresponding output.
    - Output-first approach for classification tasks: picks a class label, then generates an input that matches that label, improving label balance.



Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560.

Data Synthesis and Distillation For LLM training

- SFT Reasoning Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to enhance LLM's reasoning ability.
- Chain-of-Thought Distillation: Enable CoT-like reasoning abilities for smaller models by decomposing problems into smaller ones and guiding reasoning steps through subproblems.
  1. Generate step-by-step reasoning traces (full CoT explanations or subquestion-solution pairs) for problems using a larger model.
  2. Decompose problems into subquestion and frame each reasoning step as a subquestion followed by its solution.
  3. Train two separate models: a Question Generator (QG) to produce subquestions and a Question Answerer (QA) to solve them step by step.



Shridhar, K., Stolfo.. Distilling reasoning capabilities into smaller language models. Findings of the Association for Computational Linguistics: ACL 2023, 7059-7073.

Data Synthesis and Distillation For LLM training

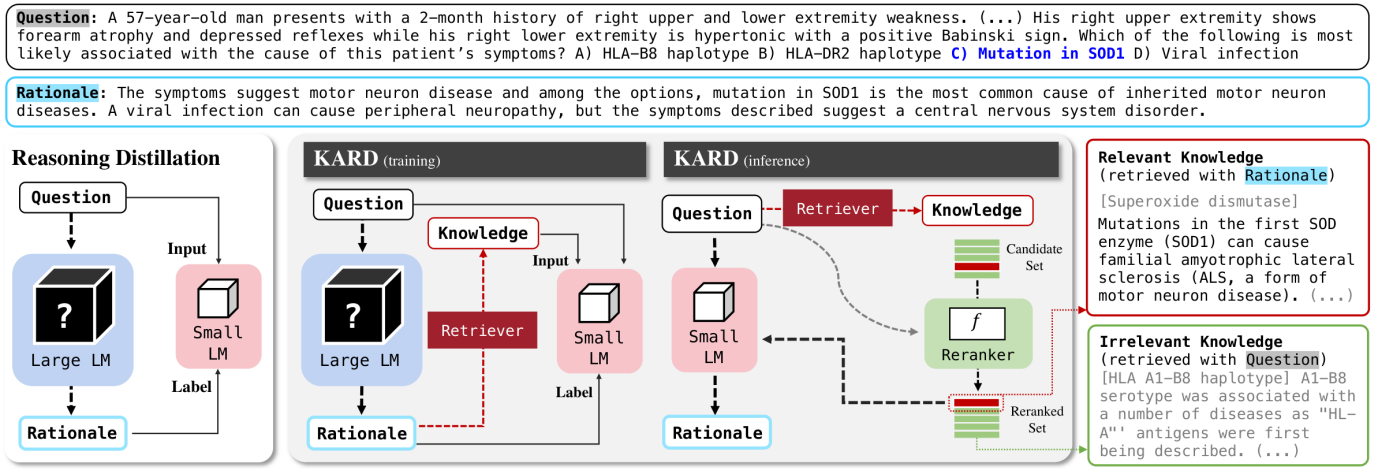
- SFT Reasoning Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to enhance LLM's reasoning ability.
- Self-Consistent Chain-of-Thought Distillation: Make teacher model generate rationales consistent with correct answer and student model focus on rationales.
  - For teacher model, compare token likelihoods when the correct answer is provided versus when a perturbed (e.g., empty or incorrect) answer is given and prefer tokens that are more plausible only under the correct answer.
  - The student model is trained on two objectives:
    - factual reasoning, where it learns to generate a rationale and predict the correct answer; and
    - counterfactual reasoning, where it learns to predict the (incorrect) answer associated with a counterfactual rationale.

    This teaches the student to base its predictions on the content of the rationale rather than shortcuts.

Wang, P., Wang, Z., Li, Z., Gao, Y., Yin, B., & Ren, X. (2023). Scott: Self-consistent chain-of-thought distillation. arXiv preprint arXiv:2305.01879.

Data Synthesis and Distillation For LLM training

- SFT Reasoning Data Augmentation: Design paradigms to prompt LLM to generate fine-tuning data to enhance LLM's reasoning ability.
- Knowledge-Augmented Reasoning Distillation: Adapt small models to knowledge-intensive reasoning tasks with complex reasoning and extensive knowledge.
  - Reasoning Distillation: Prompt teacher model to generate multiple reasoning steps (rationales) for each training question, along with correct answers.
  - RAG: During training, retrieve passages relevant to each rationale from a knowledge base using the rationale as a query.



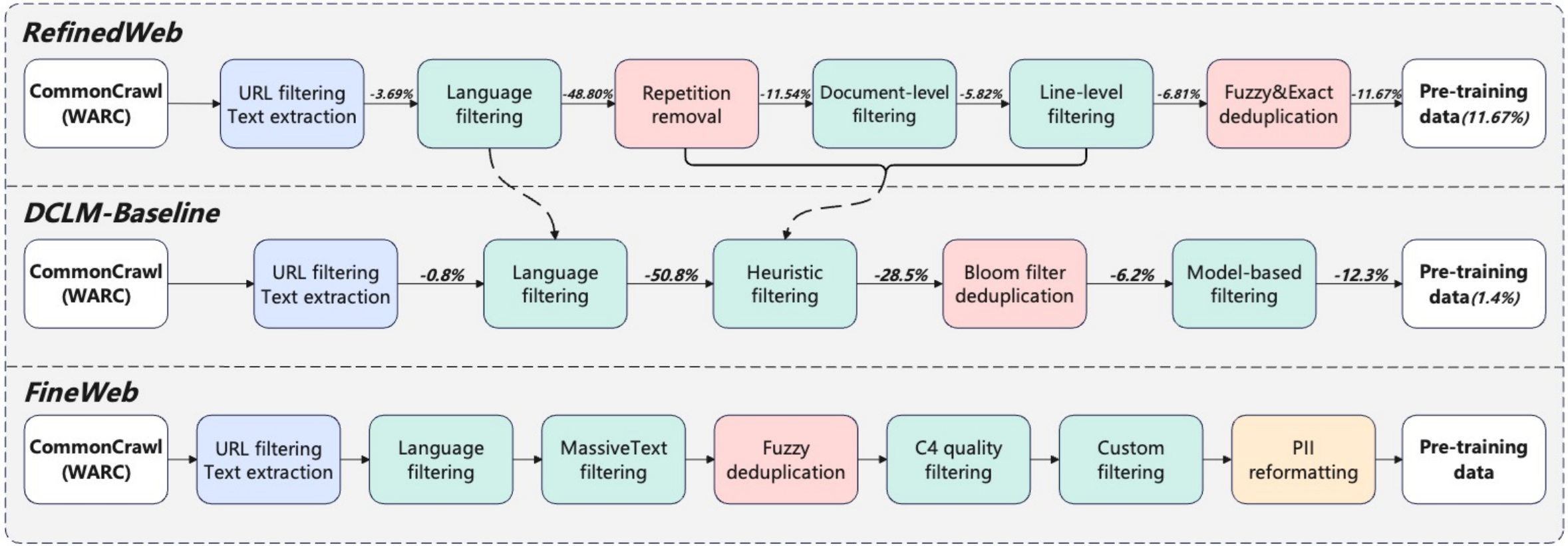
Kang, M., Lee, S., Baek, J., Kawaguchi, K., & Hwang, S. J. (2023). Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. Advances in Neural Information Processing Systems, 36, 48573-48602.

Data Synthesis and Distillation For LLM training

- **Takeaways**
  - Data synthesis using large language models (LLMs) is a powerful approach to generate high-quality, diverse, and privacy-preserving training data, addressing challenges like data scarcity, imbalance, and sensitive information exposure.
  - Data Distillation enables LLMs with less parameters has similar performance of larger LLM.
  - Synthetic data for pretraining has to be controlled under certain ratio, should be mixed with authentic data, otherwise it will even does harm to performance.
  - High quality and well-formatted reasoning data are keys to high reasoning performance.

End-to-End Pipelines

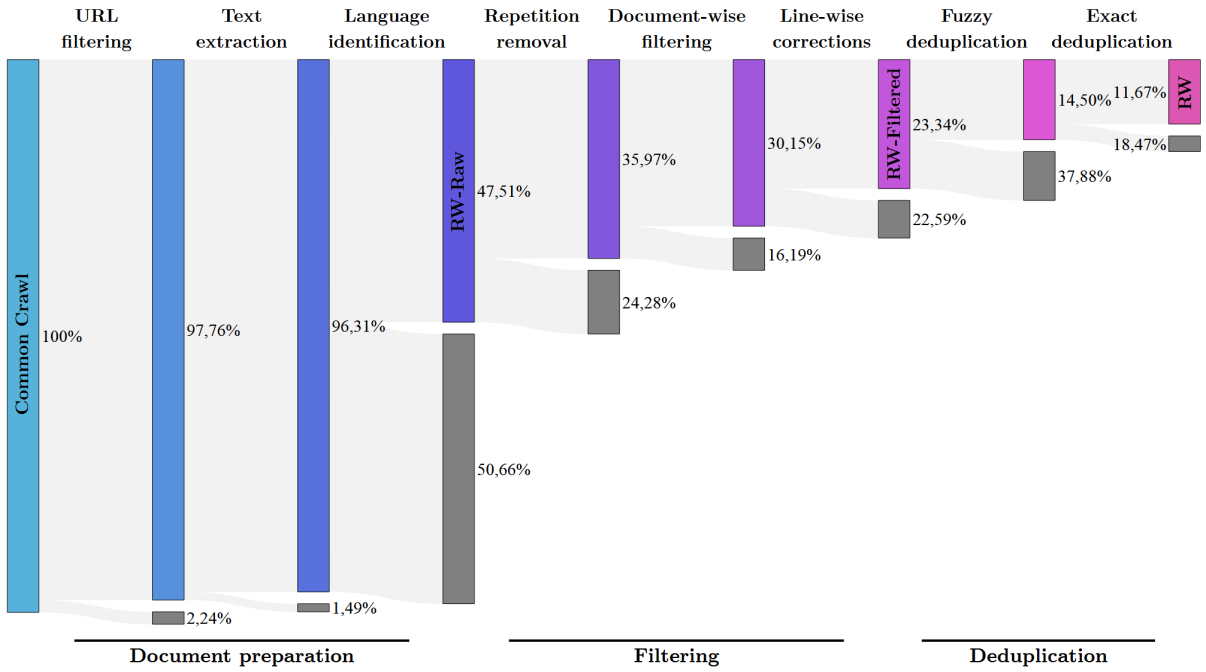
- End-to-End Pipelines: Orchestrate data processing operations that transform **raw data** into **high-quality LLM pre-training data**.



End-to-End Pipelines

- **RefinedWeb Pipeline**

1. **Data acquisition**
   - URL filtering: blocklist & URL score
   - Lang identification: FastText (like Word2Vec)
   - Text extraction: Regx Lib (Trafilatura)
2. **Data filtering**
   - Document-level filtering: Rule-based.
   - Line-level filtering: Rule-based.
3. **Data deduplication**
   - Fuzzy deduplication: Minhash.
   - Exact deduplication: Suffix array.



The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data only[J]. arXiv, 2023.

End-to-End Pipelines

- **DCLM-Baseline Pipeline**
  - Adopt RefinedWeb's heuristic filtering.
  - Use Bloom filter deduplication, offering comparable performance to MinHash with higher efficiency on large-scale datasets.
  - Compared to RefinedWeb, additionally apply model-based filtering, retaining only **1.4%** of raw data (vs. **11.67%** in RefinedWeb).

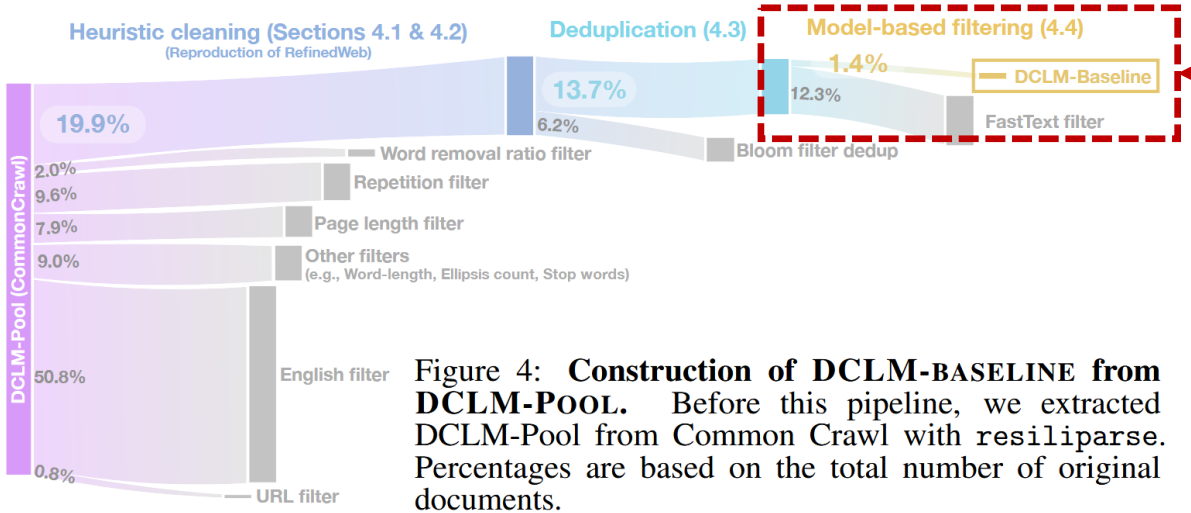


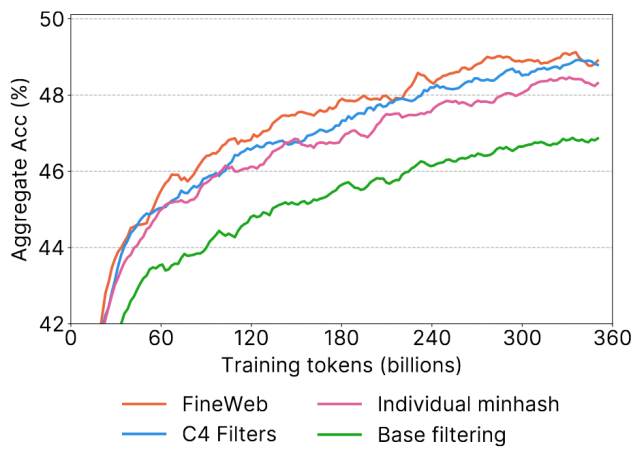
Figure 4: **Construction of DCLM-BASELINE from DCLM-POOL.** Before this pipeline, we extracted DCLM-Pool from Common Crawl with `resiliparse`. Percentages are based on the total number of original documents.

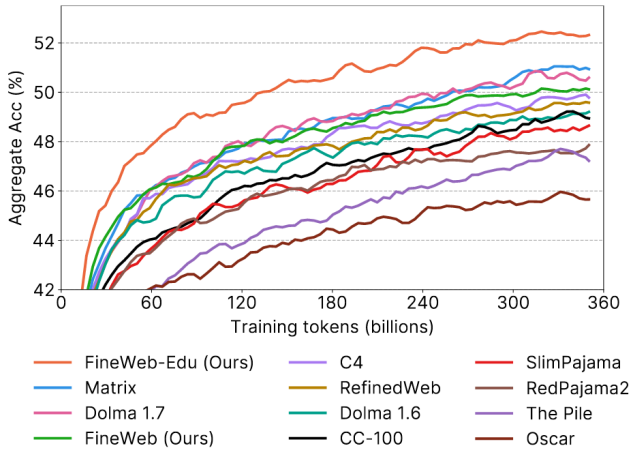
A fastText classifier trained on instruction-formatted data, including diverse data formats (OpenHermes 2.5) and QA samples (ExplainLikeImFive).

Datacomp-lm: In search of the next generation of training sets for language models[J]. NIPS, 2024.

End-to-End Pipelines

- **FineWeb Pipeline**
  - Adopts heuristic filtering from MassiveText and C4 Dataset.
  - Conducts individual Minhash deduplication for each CommonCrawl snapshot.
  - Develops additional custom heuristic filters (e.g. fraction of lines ending with punctuation) through a systematic process for better performance.
  - PII (email addresses and public IP addresses) is anonymized using regex patterns.

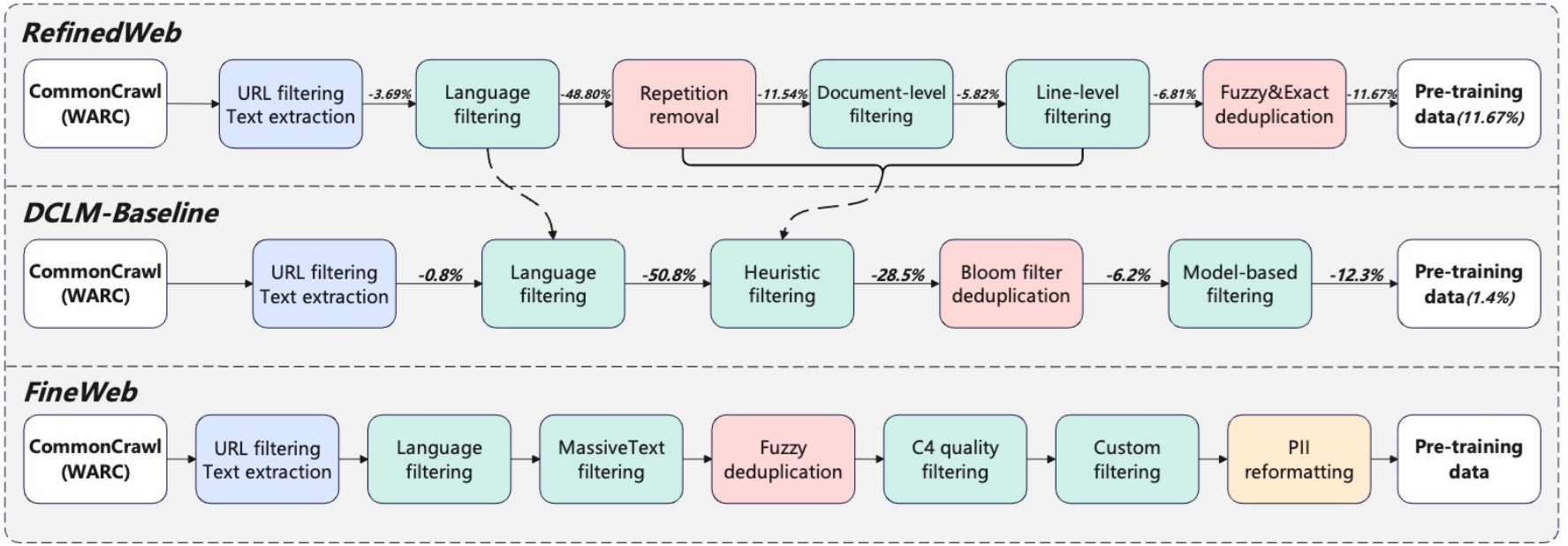




The fineweb datasets: Decanting the web for the finest text data at scale[J]. NIPS, 2024.

End-to-End Pipelines

- **Designing Principles**
  - The trade-off between data quality and quantity.
  - Dependencies across the processing operations (e.g., text extraction necessarily preceding operations like deduplication and filtering).
  - Efficiency optimization (e.g., conducting computationally intensive steps like model-based filtering after lightweight processing steps like URL filtering).



Data Packing For Data-Centric Training

- **Challenge: How to conduct data-centric training on the basis of a high-quality dataset.**

Data Packing Is Required During Pre-Training: Data Packing combines texts to ensure uniform input lengths in pre-training, improving coherence and reducing padding and truncation.

e.g., Best-fit Packing

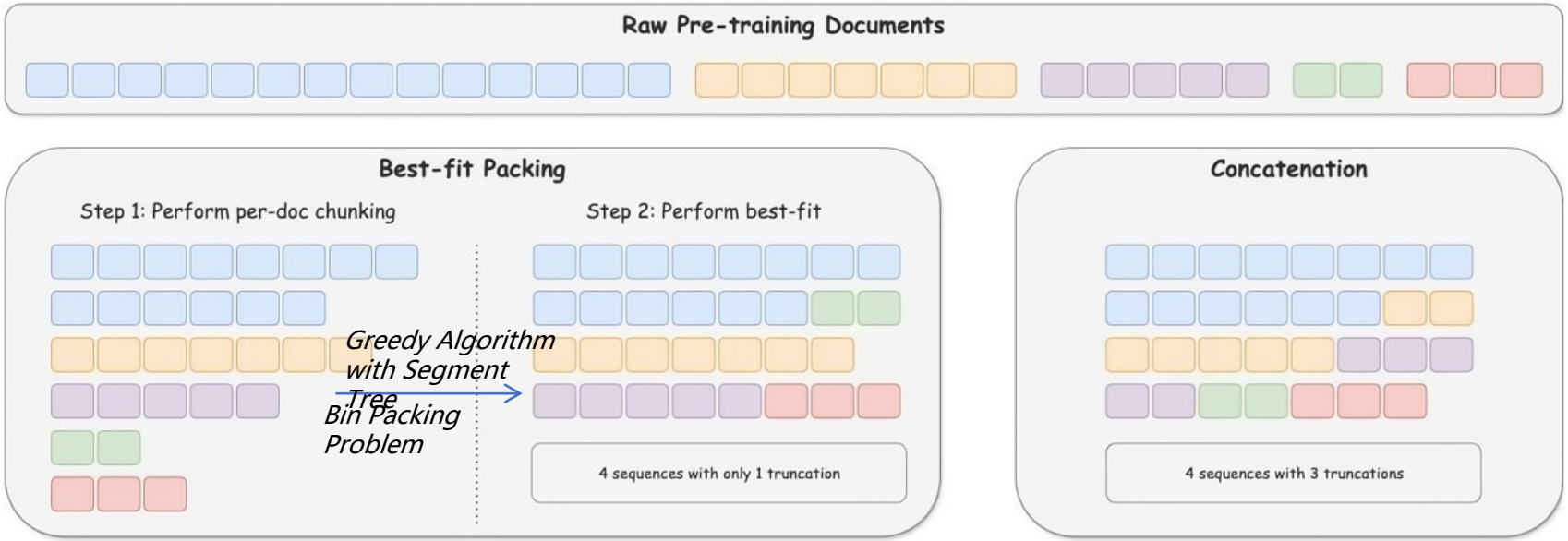


Figure 1. An illustration of the proposed Best-fit Packing compared with concatenation (baseline). We set max sequence length to 8 tokens in this example. **Top:** Original training documents. Each box stands for a token. Contiguous boxes in the same color represent a document. There are five documents of lengths 14, 7, 5, 2, 3, respectively. **Bottom-left:** Best-fit Packing. In step 1, we segment the long document (e.g., blue) into chunks with ≤ 8 tokens. In step 2, we group chunks into training sequences in a smart way that results in the smallest number of sequences. We do not break any chunk in the second step. In total, only one document was truncated and this is necessary to meet the max sequence length requirement. **Bottom-right:** The concatenation approach. 3 out of the 5 documents are truncated.

Hantian Ding, Zijian Wang, et al. Fewer Truncations Improve Language Modeling. ICML, 2024

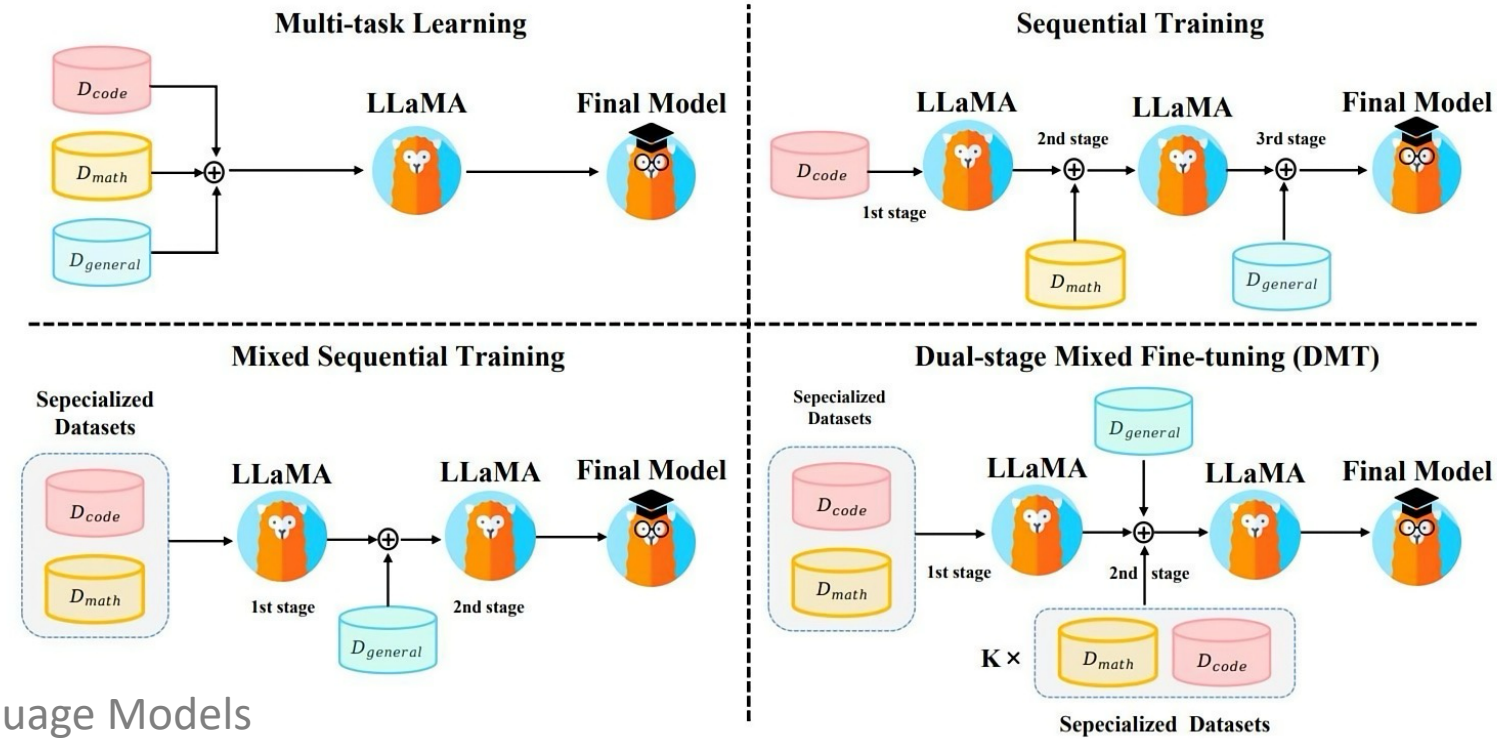
Training Strategy For Data-Centric Training

- **Challenge: How to conduct data-centric training on the basis of a high-quality dataset.**

During Training Need Appropriate Strategy: Different training strategies lead to different training results.

To balance general and specialized abilities:

- 1) Fine-tunes on specific datasets.
- 2) Fine-tunes on mixed data.



Guanting Dong, et al. How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition. ACL, 2024

Data Shuffling For Data-Centric Training

- **Challenge: How to conduct data-centric training on the basis of a high-quality dataset.**

Data Shuffling To Help Training: Data shuffling means that different data needs to be selected and provided to LLMs at various stages. (e.g., in different epochs for SFT).

e.g., Velocitune

$$V_t[i] = \frac{\ell_t[i] - \ell_{\text{target}[i]}}{\ell_{\text{init}[i]} - \ell_{\text{target}[i]}}$$

$V_t[i]$ is the learning velocity for domain i at step t

$\ell_t[i]$ is the current loss for domain i

$\ell_{\text{target}[i]}$ is the target loss for domain i, predicted by the scaling law

$\ell_{\text{init}[i]}$ is the initial loss for domain i, calculated before training starts.

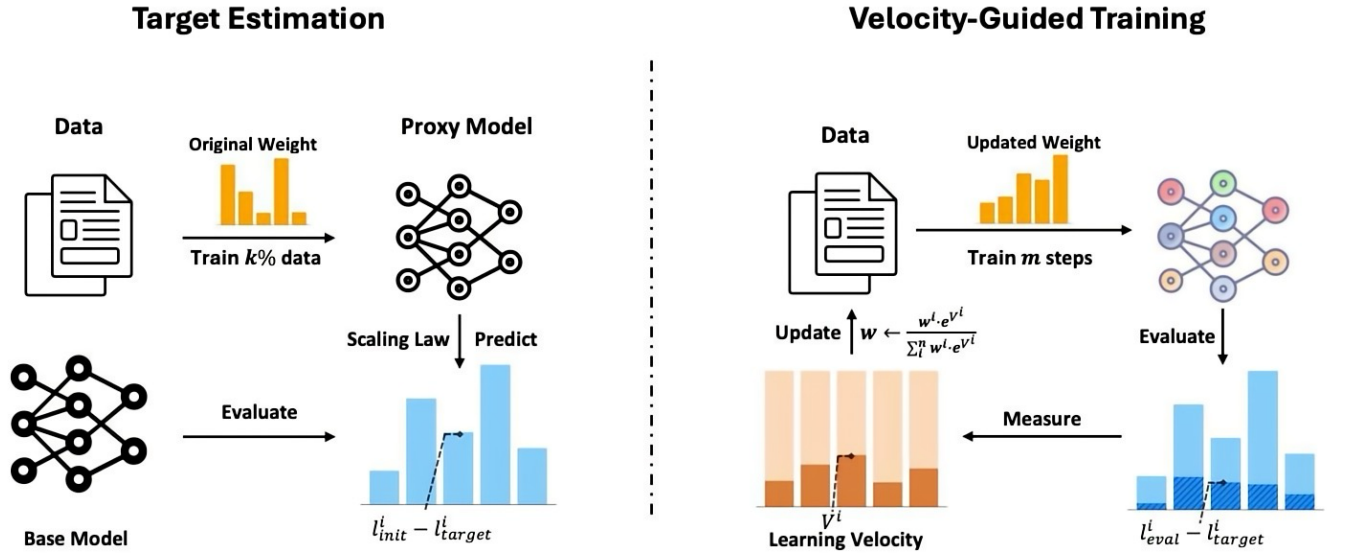


Figure 1: The overall pipeline of Velocitune. Initially, a proxy model is trained using the original domain weights on a subset of the data. Following this, the initial loss is collected by evaluating the base model, while the target loss is determined by extrapolating the evaluation loss of the proxy model. In the second phase, we calculate the learning velocity by rescaling the learning progress between the initial and target losses. This learning velocity is then used to update the domain weights effectively.

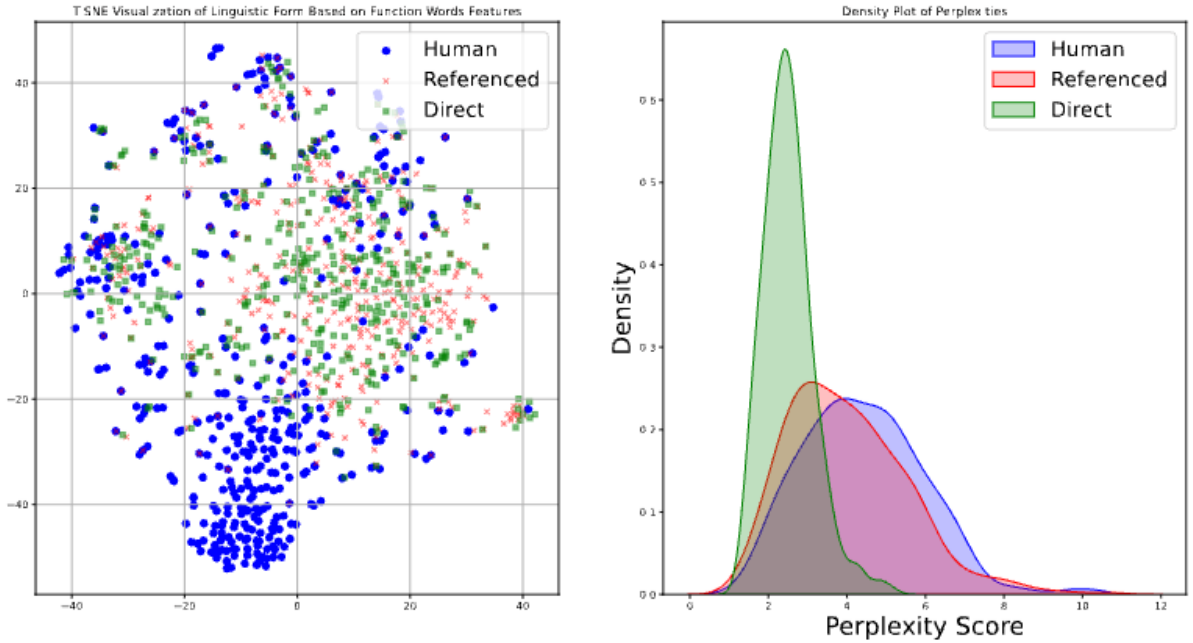
Zheheng Luo, Xin Zhang,et al. Velocitune: A Velocity-based Dynamic Domain Reweighting Method for Continual Pre-training. ACL, 2025

Takeaways

- Data Packing is essentially a bin packing problem in the field of optimization, necessitating the use of efficient algorithms with low time complexity.
- Training Strategy requires an appropriate workflow that explores suitable domain compositions and mixing ratios at each stage.
- Data Shuffling involves monitoring training signals such as loss, gradients... allowing for dynamic adjustment of data sampling ratios throughout the training process.
- Open problems:
  - More explainable Training Strategy
  - More unified metrics for monitoring training status in Data Shuffling

Future Opportunities

- **Task-Specific Data Selection for Efficient Pretraining**
  - Inclusion of irrelevant data not only increases training time but also impedes the model's adaptability to specific tasks → Adaptive data selection strategies

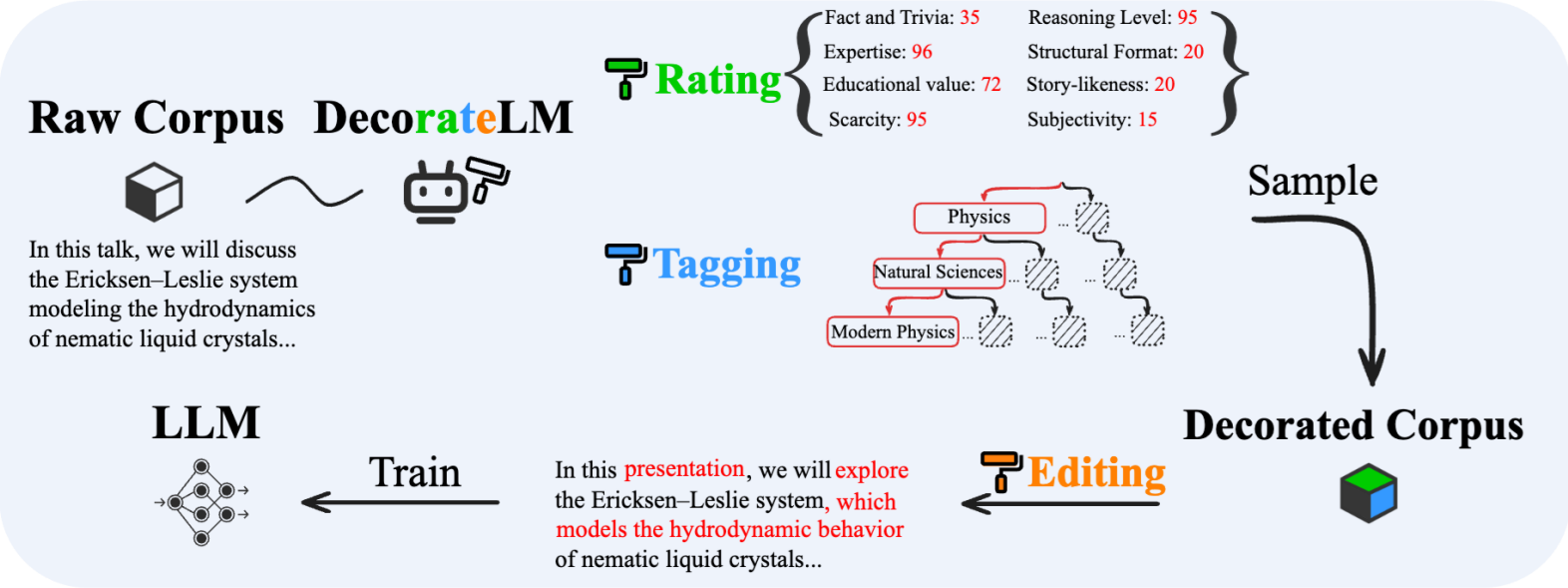


Align with Human Preference → Specific Tasks

SCAR: Data Selection via Style Consistency-Aware Response Ranking for Efficient Instruction-Tuning of Large Language Models. arXiv, 2024.

Future Opportunities

- **Predictive Pipeline Selection / Processing Agent Design**
  - Experimentally decide the pipeline is resource-intensive → Predict optimal preprocessing configurations in advance or design agentic processing method



Empirical Pipelines → Data(-agent) driven Pipelines

DecorateLM: Data Engineering through Corpus Rating, Tagging, and Editing with Language Models. EMNLP, 2024.

Thanks