北京大学
PEKING UNIVERSITY

Data Preparation for LLM Pre-training

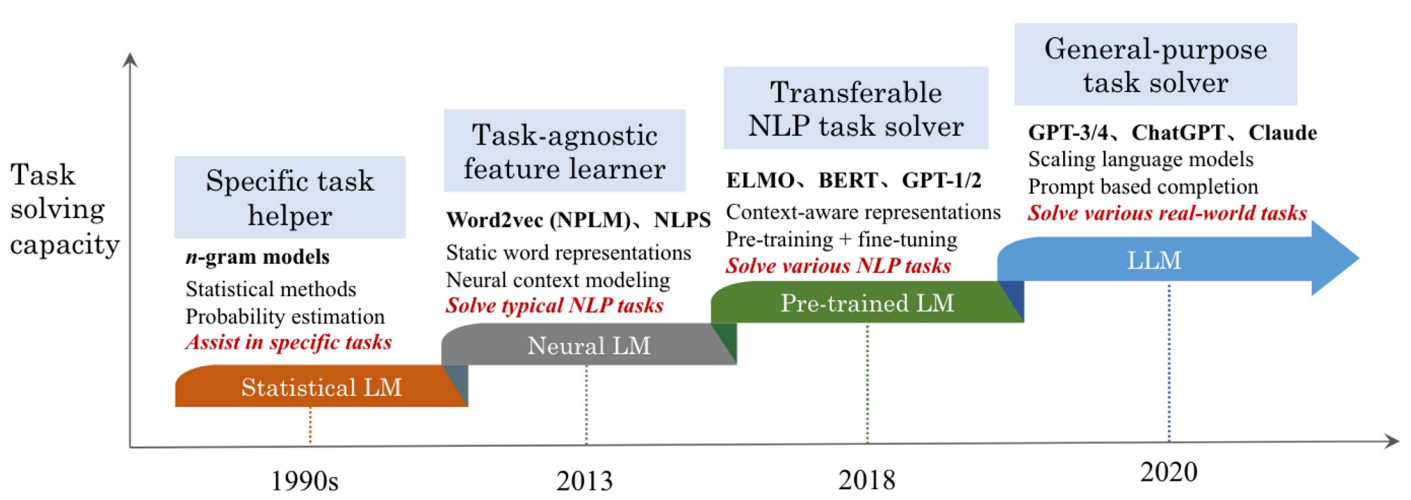
Keer Lu and Kun Yuan

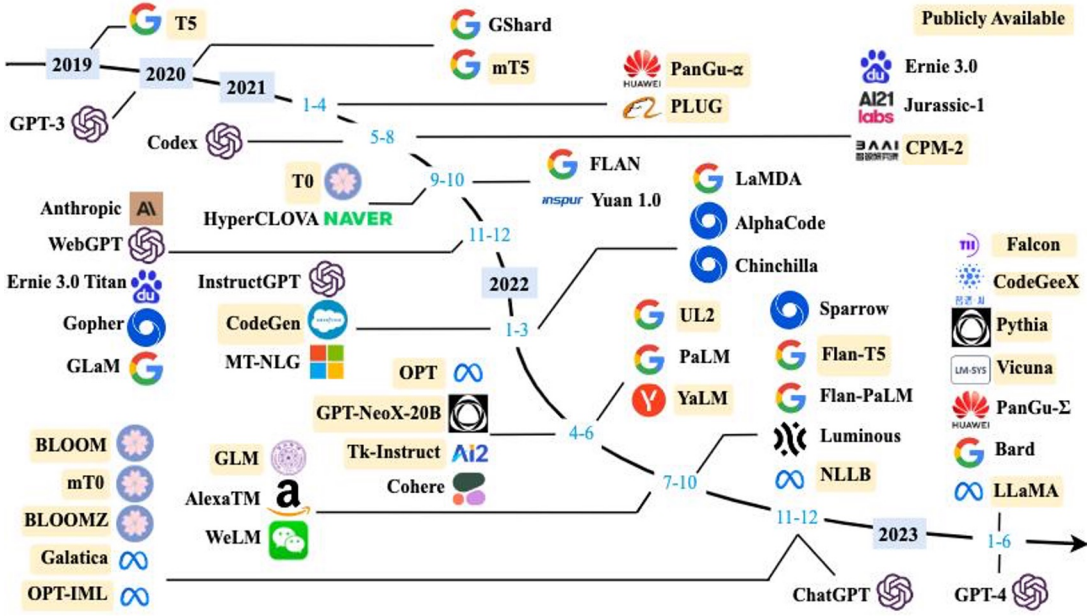
Center for Machine Learning Research @ Peking University

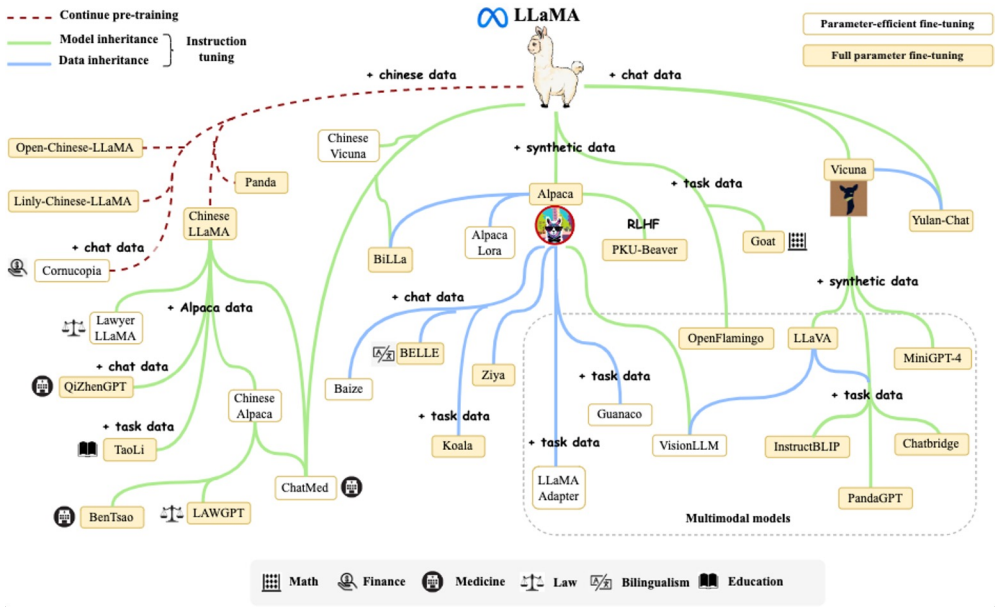
Contents

Background

- Evolution of LM
  - Statistical LM (SLM)
  - Neural LM (NLM)
  - Pre-trained LM (PLM)
  - LLM

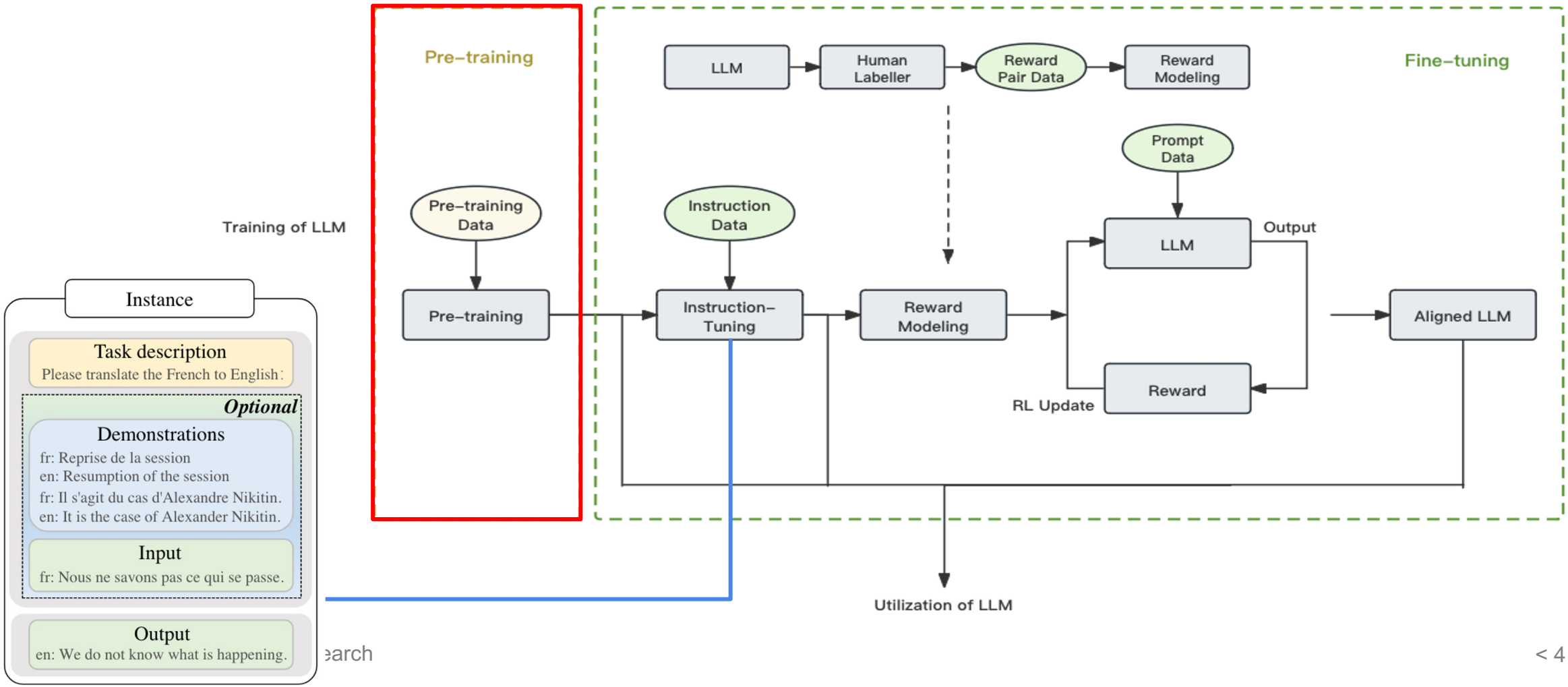






Background

- Training Process of LLMs



Background

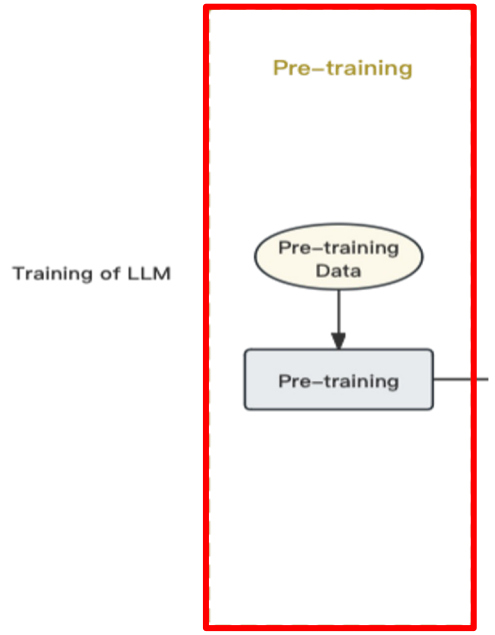
- Training Process of LLMs —— Pre-training Stage
  - Description
    - Initial stage of learning for LLMs
    - Tasks
      - Language Modeling

$$\mathcal{L}_{LM}(\mathbf{x}) = \sum_{i=1}^{n} \log P(x_i|\mathbf{x}_{<i}).$$

      - Denoising Autoencoding (DAE)

$$\mathcal{L}_{DAE}(\mathbf{x}) = \log P(\tilde{\mathbf{x}}|\mathbf{x}_{\setminus\tilde{\mathbf{x}}}).$$

  - Data for Pre-training
    - a large amount of unlabeled text data

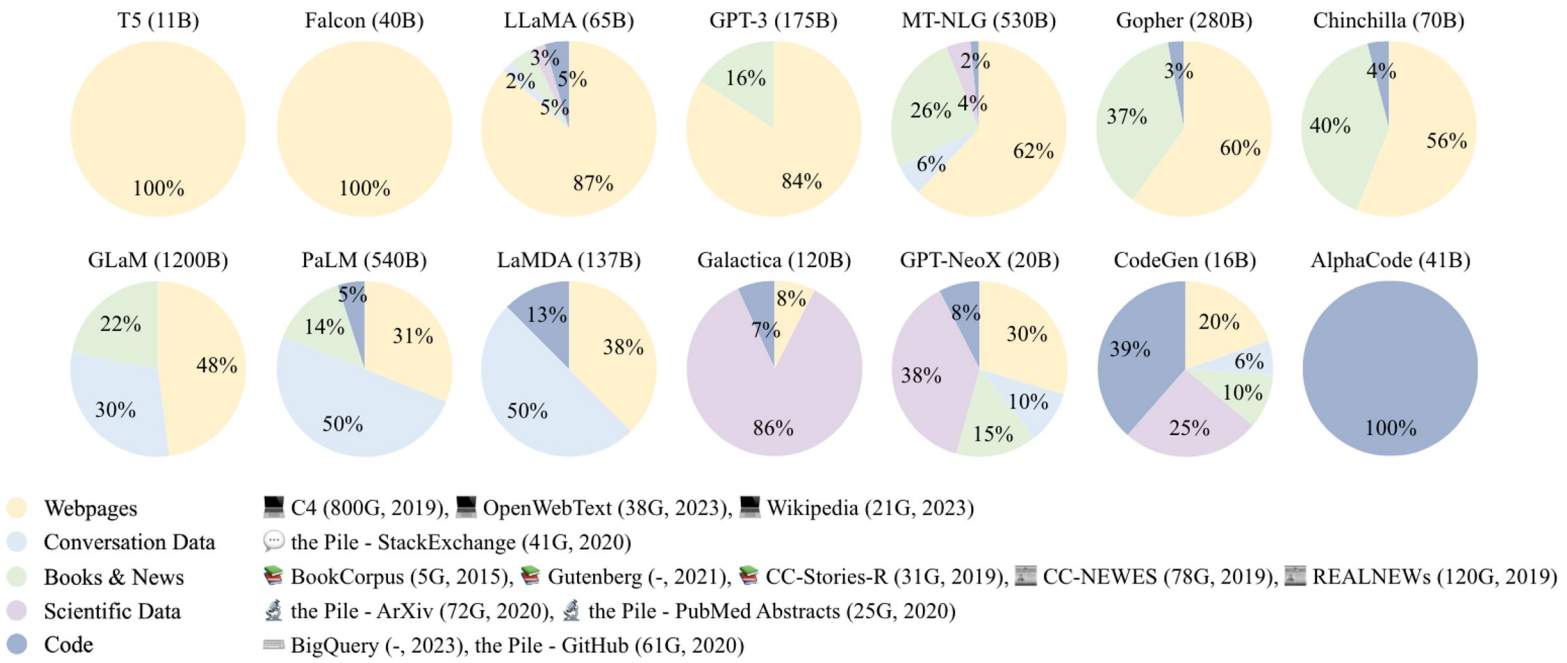


Model	Parameters	Tokens
PaLM2	340B	3.6T
LAMMA2-70B	70B	2T
GPT-4	175B	13T
Baichuan2-13B	13B	2.6T

[1] https://www.datalearner.com/ai-models/llm-evaluation

Data Source for Pre-training

- LLMs require a higher volume of training data that covers a broad range of content
- Data Source for Pre-training
  - **General Text Data**
    - Webpages
    - Conversation Text
    - Books
  - **Specialized Text Data**
    - Multilingual Text
    - Scientific Text
    - Code



[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

Data Source for Pre-training

- **General Text Data**
  - **Webpages**
    - eg. CommonCrawl [2]
      - The corpus contains raw web page data, metadata extracts, and text extracts.
      - Common Crawl data is stored on Amazon Web Services' Public Data Sets and on multiple academic cloud platforms across the world.
  - **Conversation text**
    - eg. PushShift.io Reddit corpus [3]
      - submissions and comments posted on Reddit between June 2005 and April 2019
  - **Books**
    - eg. Books3, Bookcorpus2 [4]

[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.
[2] "Common crawl." [Online]. Available: https:// commoncrawl.org/
[3] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020. AAAI Press, 2020, pp. 830–839.
[4] Gao L, Biderman S, Black S, et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling[J].

Data Source for Pre-training

北京大学 PEKING UNIVERSITY

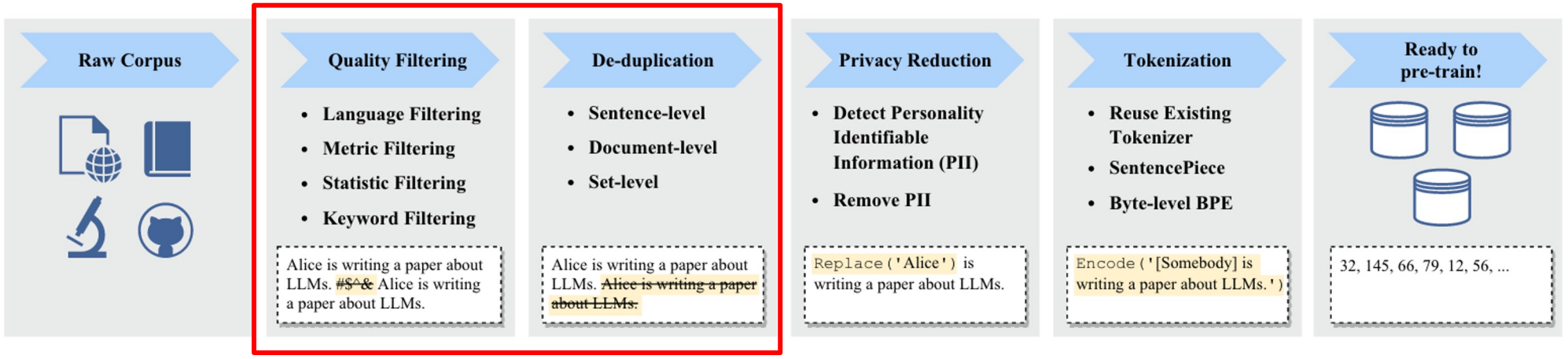
- **Specialized Text Data**
  - **Multilingual text**
    - eg. ROOTS [2]
      - **R**esponsible **O**pen-science **O**pen-collaboration **T**ext **S**ources (ROOTS) corpus
      - a 1.6TB dataset spanning 59 languages
  - **Scientific text**
    - arXiv papers, scientific textbooks, math webpages etc.
    - require specific tokenization and preprocessing techniques to transform these different formats of data into a unified form that can be processed by language models
  - **Code**
    - eg. Stack Exchange [3], GitHub

[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.
[2] Laurençon H, Saulnier L, Wang T, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset[J]. Advances in Neural Information Processing Systems, 2022, 35: 31809-31826.
[3] F. F. Xu, U. Alon, G. Neubig, and V. J. Hellendoorn, "A systematic evaluation of large language models of code," in MAPS@PLDI, 2022.

Data Processing - Overview

北京大学 PEKING UNIVERSITY

- **Pipeline of Data Processing**
  - Quality Filtering
  - Deduplication
  - Sensitive Information Detection
    - privacy reduction
    - toxicity filtering
    - bias filtering
  - Tokenization



[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

Data Processing - Deduplication

北京大学 PEKING UNIVERSITY

- **Motivation**
  - large amount of redundant data
  - unstable training process
- **Classification**
  - exact-based
  - fuzzy-based
  - embedding-based (model-based)

[1] Wang Z, Zhong W, Wang Y, et al. Data management for large language models: A survey[J]. arXiv preprint arXiv:2312.01700, 2023.
[2] Albalak A, Elazar Y, Xie S M, et al. A Survey on Data Selection for Language Models[J]. arXiv preprint arXiv:2402.16827, 2024.

Data Processing - Deduplication

- **Deduplication** —— exact&fuzzy-based
  - *The RefinedWeb Dataset for Falcon LLM* [2]
    - fuzzy deduplication: MinHash [3]
      - params: 9000 hashers / doc (20 × 450) + 5-gram
    - exact substring deduplication: suffix arrays

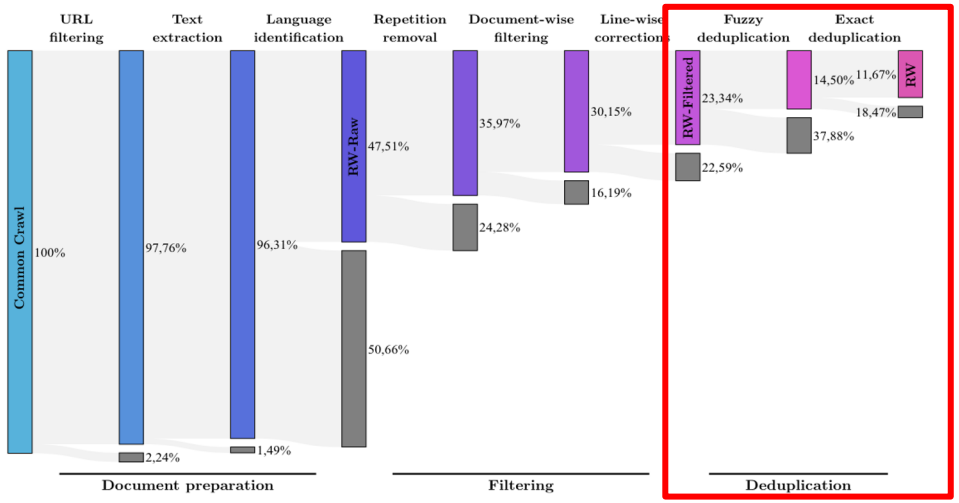
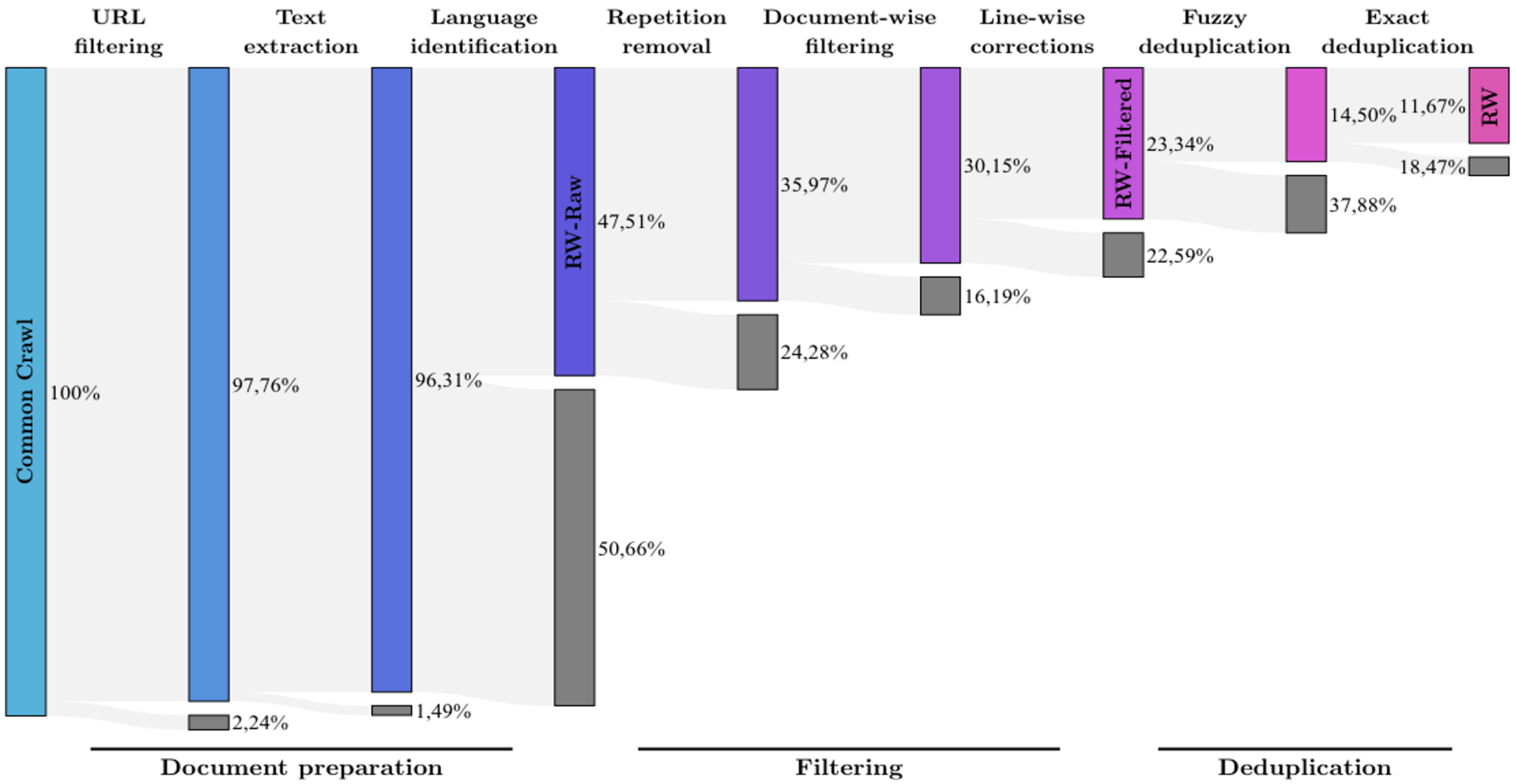


Table 17. **Matches found by exact substring deduplication** (in *italics*).

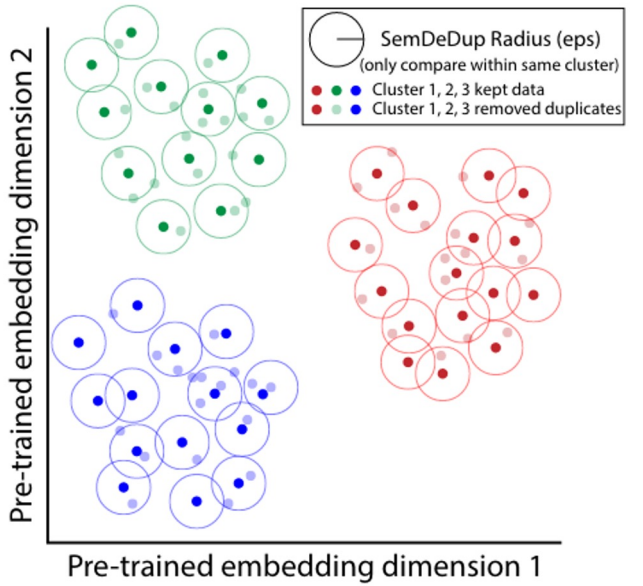
it appears there is a transfer of ranking signals in this relationship. Supporting this finding is a quote from Google's guidelines: *Using JavaScript to redirect users can be a legitimate practice. For example, if you redirect users to an internal page once they're logged in, you can use JavaScript to do so. When examining JavaScript or other redirect methods to ensure your site adheres to our guidelines, consider the intent. Keep in mind that 301 redirects are best when moving your site, but you could use a JavaScript redirect for this purpose if you don't have access to your website's server.* NOTE: Their experiment is based on a live page with status code 200 and NOT an inactive page. So if you want to implement this for legacy	Some examples of sneaky redirects include: - Search engines shown one type of content while users are redirected to something significantly different. - Desktop users receive a normal page, while mobile users are redirected to a completely different spam domain. *Using JavaScript to redirect users can be a legitimate practice. For example, if you redirect users to an internal page once they're logged in, you can use JavaScript to do so. When examining JavaScript or other redirect methods to ensure your site adheres to our guidelines, consider the intent. Keep in mind that 301 redirects are best when moving your site, but you could use a JavaScript redirect for this purpose if you don't have access to your website's server.*
Find Palm Beache FL homes for sale and other Palm Beach real estate on homesofthepalmbeaches.com. Browse and search Palm Beach houses, condos, townhomes and single-family homes by community , building, or location. *Our extensive database of real estate listings provide the most comprehensive property details including home values, features and local school and neighborhood info so you can be sure that you have nearly all the facts you need upfront. Search* homesofthepalmbeaches.com today! Want a closer look at what other Palm Beach properties are available?	Search Stuart houses, condos, townhomes and single-family homes by price and location. *Our extensive database of real estate listings provide the most comprehensive property details including home values, features and local school and neighborhood info so you can be sure that you have nearly all the facts you need upfront. Search* Stuart Listings today! Want a closer look at what other Stuart properties are available? Also search our listings for the Newest Stuart Listings and Stuart Homes with Price Reductions now. Stuart FL Homes for Sale - Stuart Real Estate Listings FREE to search Stuart Property
To find the correct size you should measure your foot from the heel to the toe point. Add approximately 1 - 1,5cm to get the actual inner sole length. Measure both feet and fit shoes to the larger foot. Measure feet at the end of the day, when your feet are at their largest. Lente shoes are women's easy slip-on leisure shoes for everyday use. These lightweight shoes have a breathable textile mesh upper made of recycled PET bottles and cool Lycra lining.	*To find the correct size you should measure your foot from the heel to the toe point. Add approximately 1 - 1,5cm to get the actual inner sole length. Measure both feet and fit shoes to the larger foot. Measure feet at the end of the day, when your feet are at their largest.* Enjoy your summer days with Masera leisure sneakers. These low-cut women's sneakers are extremely lightweight thanks to phylon midsole and breathable textile mesh upper
This bandana makes the perfect addition to every fur babies birthday collection! With its sparkly crown pattern, your pup will be ready for every birthday celebration! *With snaps for security, this bandana is made with love, down to the very last stitch ! Fabric: cotton Care Instructions: Hand wash only, iron as needed, on low heat Always supervise your pup while wearing Faithful Paws Co. accessories, as it could become a choking hazard if consumed.*	This bandana makes the perfect addition to every fur babies summer collection! With its vibrant watercolor popsicle pattern, your pup will be ready for every summer cookout! *With snaps for security, this bandana is made with love, down to the very last stitch ! Fabric: cotton Care Instructions: Hand wash only, iron as needed, on low heat Always supervise your pup while wearing Faithful Paws Co. accessories, as it could become a choking hazard if consumed.*

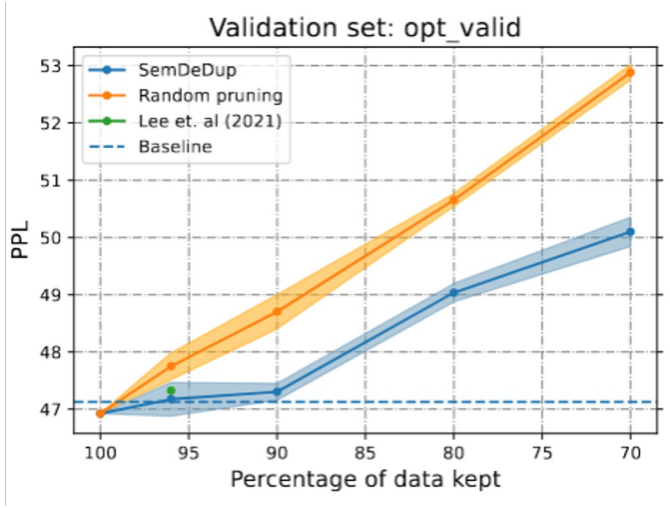
[1] Wang Z, Zhong W, Wang Y, et al. Data management for large language models: A survey[J]. arXiv preprint arXiv:2312.01700, 2023.
[2] Penedo G, Malartic Q, Hesslow D, et al. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only[J]. arXiv preprint arXiv:2306.01116, 2023.
[3]Albalak A, Elazar Y, Xie S M, et al. A Survey on Data Selection for Language Models[J]. arXiv preprint arXiv:2402.16827, 2024.


URL filtering
Text extraction
Language identification
Repetition removal
Document-wise filtering
Line-wise corrections
Fuzzy deduplication
Exact deduplication
Common Crawl
100%
97,76%
2,24%
96,31%
1,49%
RW-Raw
47,51%
50,66%
35,97%
24,28%
30,15%
16,19%
RW-Filtered
23,34%
22,59%
14,50%
37,88%
11,67%
18,47%
RW
Document preparation
Filtering
Deduplication

Data Processing - Deduplication

- **Deduplication** —— embedding-based
  - *SemDeDup* [1]
    - Idea
      - identify semantic duplicates by embeddings from pre-trained models
    - Dataset
      - C4
    - Pipeline (for natural language)
      - **embedding data**, using OPT model
      - get k clusters via **k-means** (k = 11,000)
      - compute cosine similarities
        - retain the lowest one to the centroid
        - remove the rest





[1] Abbas A, Tirumala K, Simig D, et al. SemDeDup: Data-efficient learning at web-scale through semantic deduplication[J]. arXiv preprint arXiv:2303.09540, 2023.

Data Processing - Quality Filtering

- **Motivation**
  - remove low-quality data from collected corpus
- **Classification**
  - classifier-based
  - heuristic-based
  - metric-based

[1] Wang Z, Zhong W, Wang Y, et al. Data management for large language models: A survey[J]. arXiv preprint arXiv:2312.01700, 2023.
[2] Albalak A, Elazar Y, Xie S M, et al. A Survey on Data Selection for Language Models[J]. arXiv preprint arXiv:2402.16827, 2024.

Data Processing - Quality Filtering

- **Quality Filtering** —— classifier-based
  - Description: train a selection classifier based on high quality texts
  - *GPT-3* [1]
    - Idea
      - classifier trained on **logistic regression** → filter Common Crawl
    - Corpus for Classifier Training
      - positive examples: WebText, Wikipedia and web books corpus
      - negative examples: Unfiltered Common Crawl
    - Pipeline
      - use the classifier to score documents, keep the document when

$$\texttt{np.random.pareto}(\alpha) > 1 - \texttt{document\_score}$$

      - chose α = 9

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P et al. (2020). Language Models are Few-Shot Learners. arXiv e-prints, arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165

Data Processing - Quality Filtering

北京大学 PEKING UNIVERSITY

Quality Filtering: leverage well crafted rules && statistical regularities

- **Dolma** [1]
  - 200T raw text → 3T tokens English corpus

Source	Doc Type	UTF-8 bytes (GB)	Documents (millions)	Unicode words (billions)	Llama tokens (billions)
Common Crawl	web pages	9,022	3,370	1,775	2,281
The Stack	code	1,043	210	260	411
C4	web pages	790	364	153	198
Reddit	social media	339	377	72	89
PeS2o	STEM papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7	4.3
Total		**11,519**	**4,367**	**2,318**	**3,059**

- Common Crawl
  - Fraction of characters in most common ngram greater than a threshold[48]
  - Fraction of characters in duplicate ngrams greater than a threshold[49]
  - Contains fewer than 50 or more than 100K words
  - Median word length is less than 3 or greater than 10
  - Symbol to word ratio greater than 0.10
  - Fraction of words with alpha character less than 0.80
  - Contains fewer than 2 of a set of required words[50]
  - Fraction of lines in document starting with bullet point greater than 0.90
  - Fraction of lines in document ending with ellipsis greater than 0.30
  - Fraction of lines in document that are duplicated greater than 0.30
  - Fraction of characters in duplicated lines greater than 0.30

- The Stack
  - Contains XML template code.
  - HTML code-to-text ratio <= 0.2.
  - Java, Javascript, Python code-to-comment ratio <= 0.01 or > 0.8.

[1] Soldaini L, Kinney R, Bhagia A, et al. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research[J]. arXiv preprint arXiv:2402.00159, 2024.

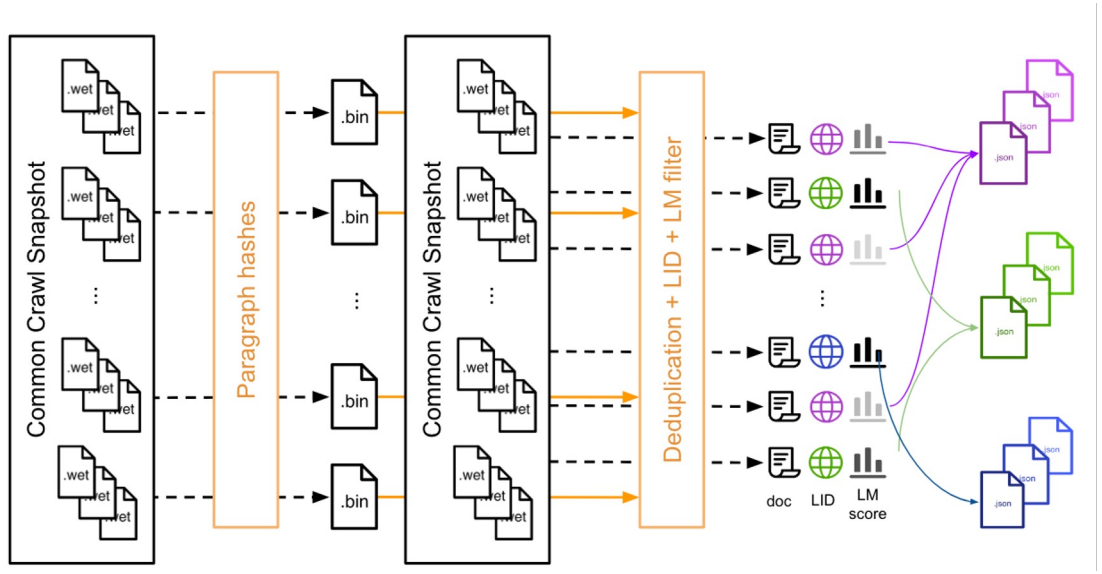
Data Processing - Quality Filtering

- **Quality Filtering** —— metric-based
  - Description: utilize the capabilities of a well-trained model to assess data quality
  - *CCNet(2019), Scaling Data-Constrained Language Models(2023)*
    - Idea
      - use perplexity(PPL) as a filtering criterion
    - Pipeline
      - Language Model Training
        - SentencePiece tokenizer
        - 5-gram Kneser-Ney model
      - Perplexity Calculation

$$p(s) = p(w_1, w_2, \dots w_n) = \prod_{i=1}^{n} p(w_i | w_1, w_2, \dots w_{i-1})$$

$$perplexity = p(s)^{-\frac{1}{n}} = p(w_1, w_2, \dots, w_n)^{-\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{p(w_i | w_1, w_2, \dots w_{i-1})}}$$

      - Filtering
        - divide documents into parts: head, middle, and tail



[1] Wenzek G, Lachaux M A, Conneau A, et al. CCNet: Extracting high quality monolingual datasets from web crawl data[J]. arXiv preprint arXiv:1911.00359, 2019.

[2] Muennighoff N, Rush A, Barak B, et al. Scaling data-constrained language models[J]. Advances in Neural Information Processing Systems, 2024, 36.

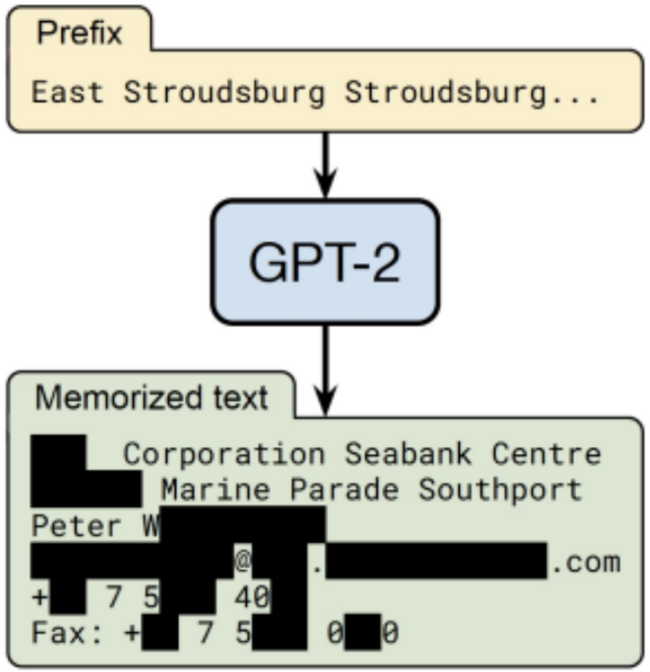
Data Processing - Sensitive Information Detection

- **Motivation**
  - detect and filter specific information in text corpus for better pre-training
- **Classification**
  - privacy reduction
  - toxicity filtering
  - bias filtering

[1] Wang Z, Zhong W, Wang Y, et al. Data management for large language models: A survey[J]. arXiv preprint arXiv:2312.01700, 2023.

Data Processing - Sensitive Information Detection

- **Privacy Reduction**
  - remove *personal identifiable information (PII)*
    - names, phone numbers, addresses etc.
  - common detection method: heuristic rule-based filtering
    - eg. Google Cloud NLP [1]
      - use regular expressions or dictionary matching

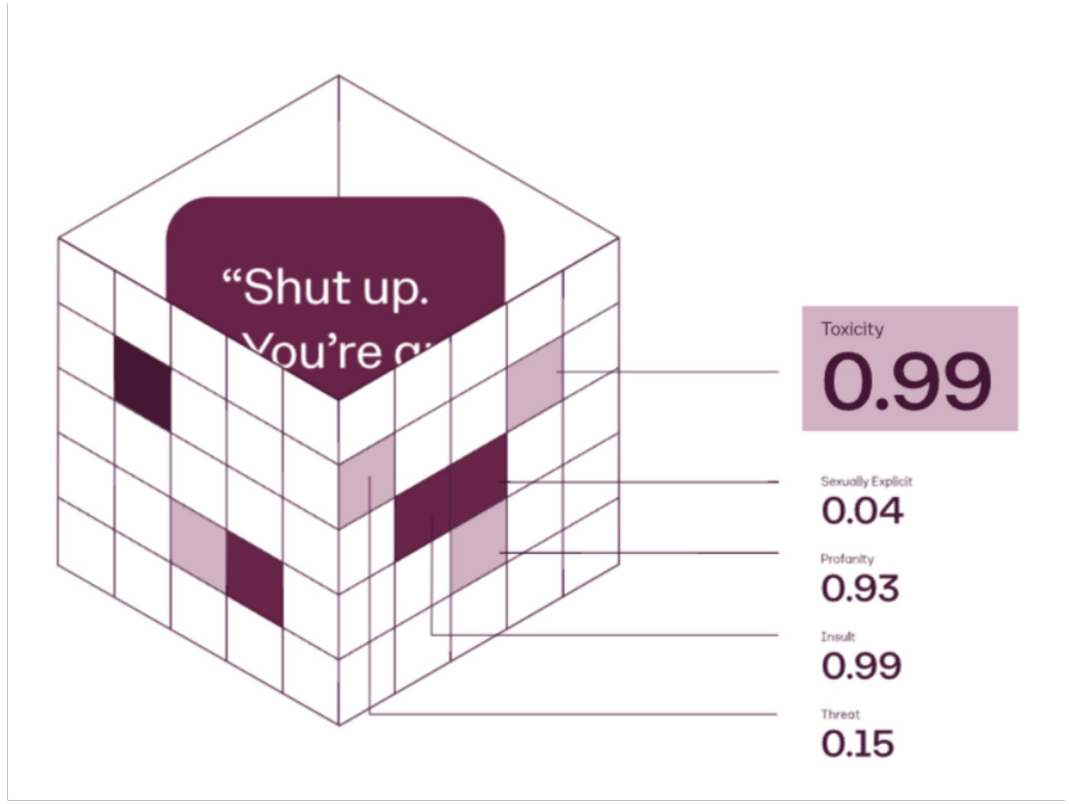


  - [2] finds that after deduplication of the training data, language models exhibit a reduction in the generation of sensitive information, leading to significantly enhanced security against privacy attacks.

[1] Google Cloud NLP. Google Cloud infotype detector, 2023a. URL https://cloud.google.com/dlp/docs/ infotypes-reference.
[2] Kandpal N, Wallace E, Raffel C. Deduplicating training data mitigates privacy risks in language models[C]//International Conference on Machine Learning. PMLR, 2022: 10697-10707.

Data Processing - Sensitive Information Detection

- **Toxicity Filtering**
  - description of *toxicity*
    - rude, dis-respectful, or unreasonable
  - detection method
    - traditional method
      - heuristic rule-based filtering
      - N-gram classifier
      - etc.
    - advanced method
      - Jigsaw's Perspective API [1]
        - a deep learning way to detect toxic information online
          - train a BERT-based model and distill into CNNs to reduce the computation time
        - more accurate than traditional ways



[1] https://www.perspectiveapi.com/

Data Processing - Sensitive Information Detection

北京大学 PEKING UNIVERSITY

- **Bias Filtering**
  - description of *bias*
    - gender bias, racial bias, religious bias etc.
  - debias techniques
    - Counterfactual Data Augmentation(CDA)Dropout, Iterative Nullspace Projection(INLP), Sentence Debias and Self-Debias
  - Improvements on bias benchmarks by using debiasing strategies are often accompanied by a decrease in language modeling ability[1]
  - Self-Debias is the strongest debiasing technique, obtaining improved scores on all bias benchmarks [1]

[1] Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models[J]. arXiv preprint arXiv:2110.08527, 2021.

Data Processing - Sensitive Information Detection

北京大学 PEKING UNIVERSITY

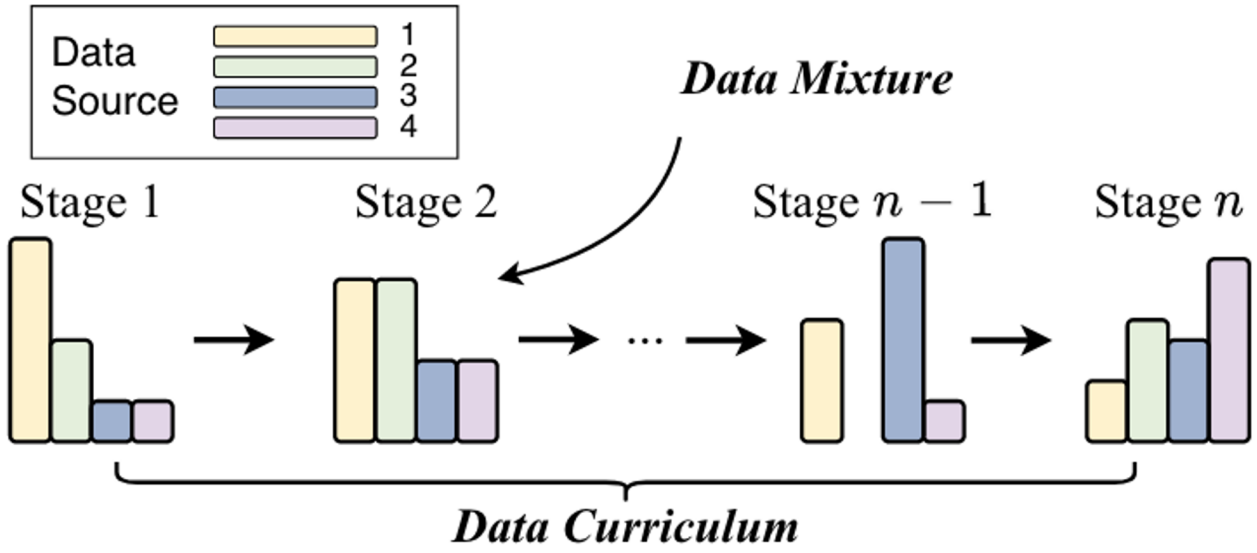
- **Bias Filtering**
  - CDA [1]: swap bias attribute words (he/she) for data augmentation
  - Dropout [2]: increased dropout regularization reduces gender bias within BERT and ALBERT
  - INLP [3]: train a linear classifier to predict the bias property for representations
  - Sentence Debias [4]: project representations onto linear subspace of bias and remove them
  - Self-Debias [5]: prompt LLMs to generate biased text first and then prompt LLMs for non-discriminative words

[1] Zmigrod R, Mielke S J, Wallach H, et al. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology[J]. arXiv preprint arXiv:1906.04571, 2019.
[2] Webster K, Wang X, Tenney I, et al. Measuring and reducing gendered correlations in pre-trained models[J]. arXiv preprint arXiv:2010.06032, 2020.]
[3] Ravfogel S, Elazar Y, Gonen H, et al. Null it out: Guarding protected attributes by iterative nullspace projection[J]. arXiv preprint arXiv:2004.07667, 2020.
[4] Liang P P, Li I M, Zheng E, et al. Towards debiasing sentence representations[J]. arXiv preprint arXiv:2007.08100, 2020.
[5] Schick T, Udupa S, Schütze H. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 1408-1424.

Data Scheduling

- **Data Composition**
  - **Motivation**
    - set a suitable distribution to mix data source
  - **Classification**
    - Based on Statistics
    - Based on Influence Score
    - Based on Proxy Model

- **Data Curriculum**
  - schedule the order that specific data is presented to LLMs for pre-training



[1] Wang Z, Zhong W, Wang Y, et al. Data management for large language models: A survey[J]. arXiv preprint arXiv:2312.01700, 2023.
[2] Albalak A, Elazar Y, Xie S M, et al. A Survey on Data Selection for Language Models[J]. arXiv preprint arXiv:2402.16827, 2024.

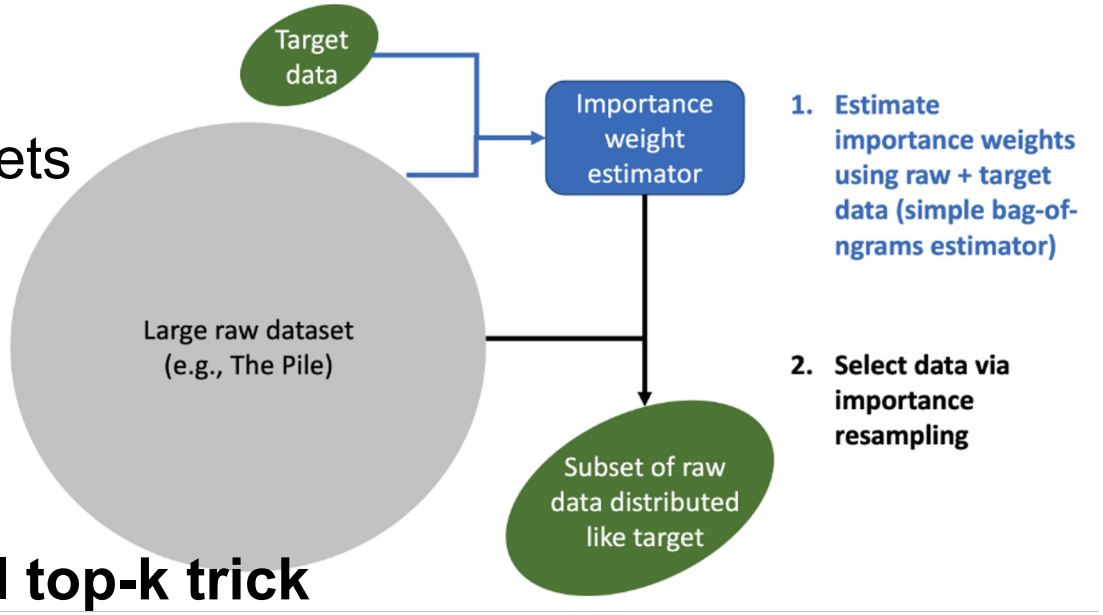
Data Scheduling - Data Composition

- **Based on Statistics**
  - Description: use heuristic or manually determined domain weights
  - *Data Selection with Importance Resampling (DSIR)* [1]
    - Idea
      - large raw dataset + smaller target dataset
      - select a subset that is distributed like the target in some feature space
    - Pipeline
      - hashed **n-gram** features
        - unigram + bigram, 1000 buckets
      - compute importance weights
        - data metric: **KL reduction**

$$D_{KL}(p||q) = \sum_{j=1}^{n} p(x_j) ln \frac{p(x_j)}{q(x_j)}$$

$$D_{KL}(target||random) - D_{KL}(target||selected)$$

      - resample
        - domain distribution + **Gumbel top-k trick**

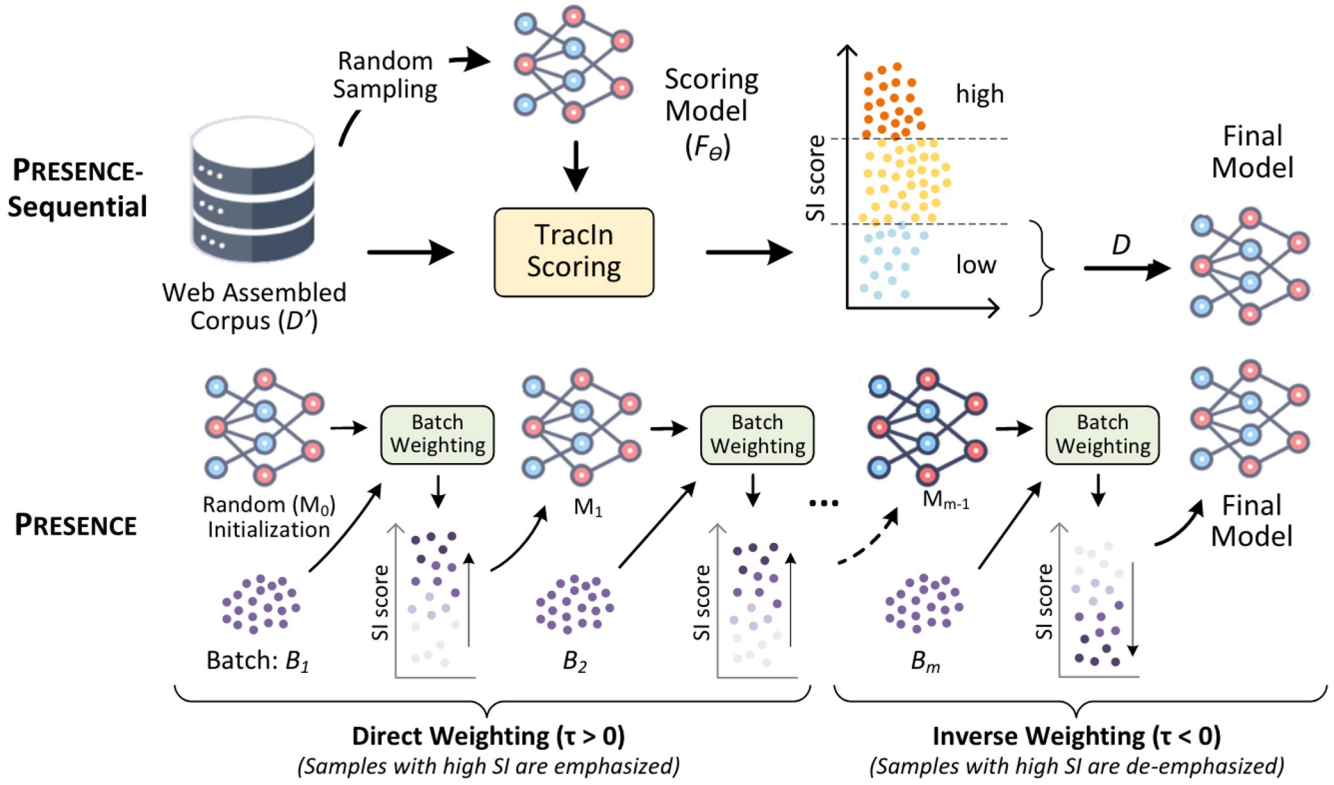


[1] Xie S M, Santurkar S, Ma T, et al. Data selection for language models via importance resampling[J]. Advances in Neural Information Processing Systems, 2024, 36.

Data Scheduling - Data Composition

Based on Influence Score

- Description: utilize influence scores approximation to calculate the impact



Score	Description
VoG (Agarwal et al., 2022)	Variance of gradients of model outputs with respect to the inputs.
EL2N (Paul et al., 2021)	Norm of the margin of confidence between the model prediction and the one-hot label.
Forgetting Score (Toneva et al., 2019)	How often an example is forgotten i.e. goes from being classified correctly at checkpoint i to incorrectly at $i+1$.
PVI (Ethayarajh et al., 2022)	Fine-grained information-theoretic quantity whose expectation value is the amount of usable information (in bits) by the model.
TracIn (Garima et al., 2020)	Influence of any example z towards another example z' by tracking their gradient dot products. We generate the self-influence scores where $z = z'$.

[1] Thakkar M, Bolukbasi T, Ganapathy S, et al. Self-Influence Guided Data Reweighting for Language Model Pre-training[J]. arXiv preprint arXiv:2311.00913, 2023.

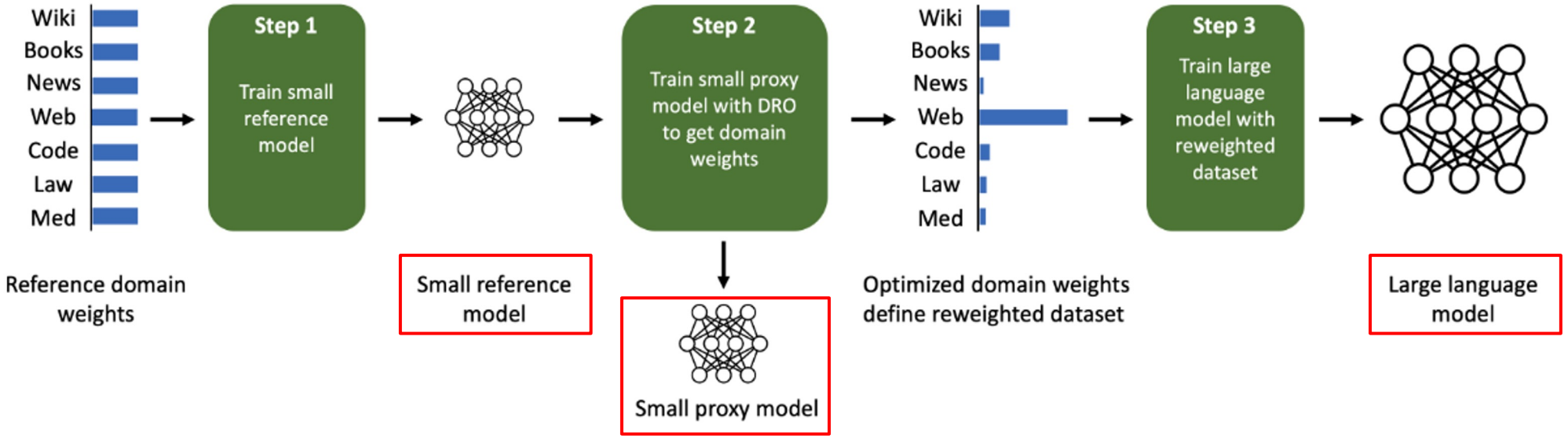
[2] Anand N, Tan J, Minakova M. Influence Scores at Scale for Efficient Language Data Sampling[J]. arXiv preprint arXiv:2311.16298, 2023.

Data Scheduling - Data Composition

- **Based on Proxy Model**

How to design a proxy model?😢

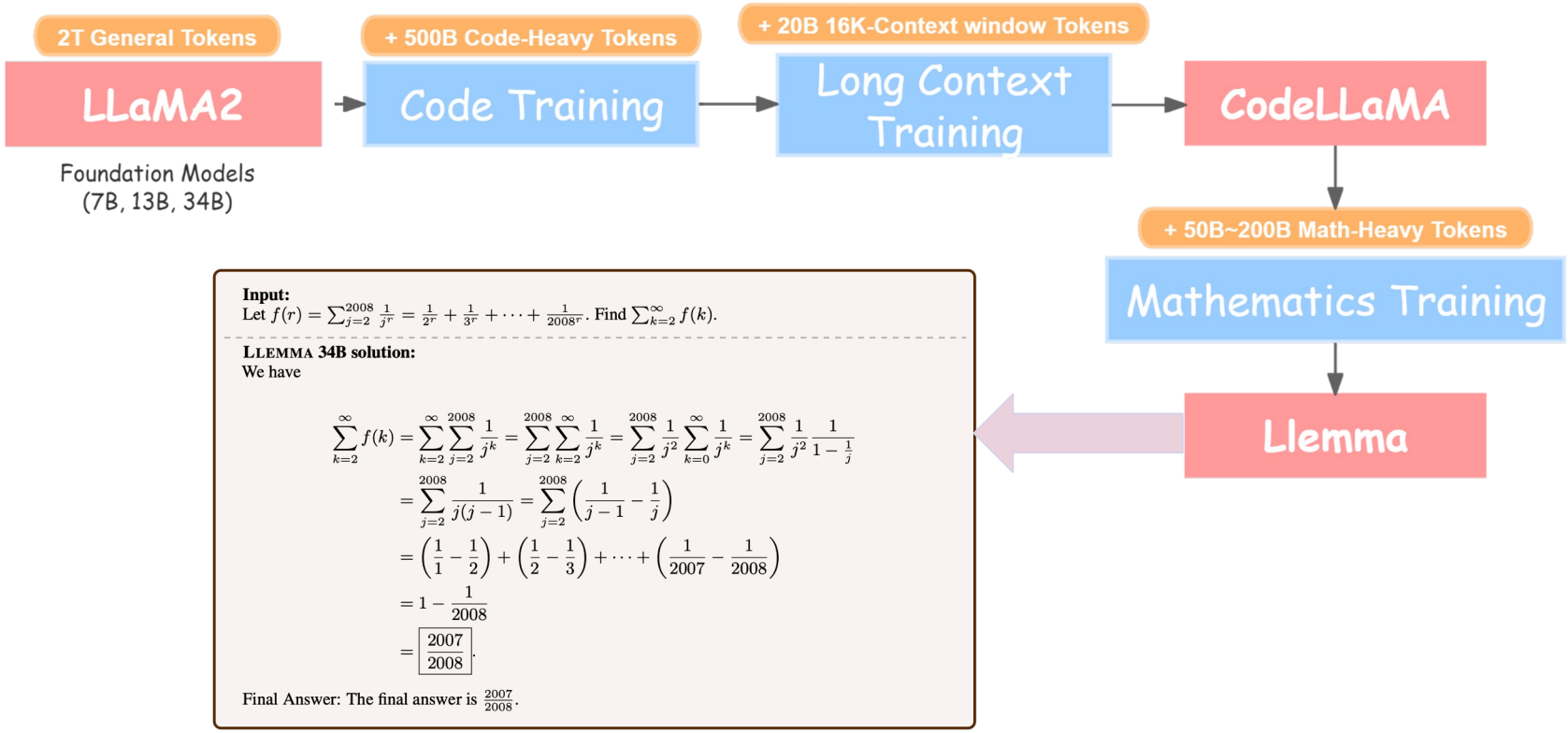
  - Description: train a small proxy model → transfer domain weights
  - ***DoReMi*** [1]
    - **Do**main **Re**weighting with **Mi**nimax Optimization
    - insight
      - group distributionally robust optimization (Group DRO) → train proxy model
      - resample dataset with gained domain weights → train full sized model



[1] Xie S M, Pham H, Dong X, et al. Doremi: Optimizing data mixtures speeds up language model pretraining[J]. Advances in Neural Information Processing Systems, 2024, 36.

Data Scheduling - Data Curriculum

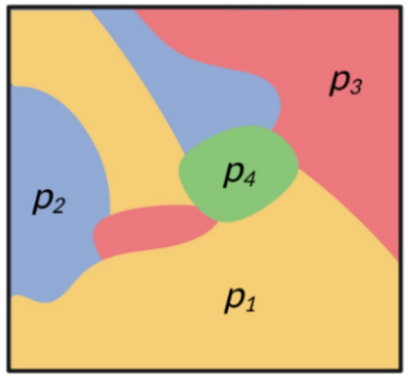
- Description
  - schedule the presenting order of specific data (basic → target)
  - focus on continual pre-training
- Application Scenarios
  - Coding
    - *CodeLLaMA* [2]
  - Mathematics
    - *Llemma* [3]
  - Long Context
    - *CodeLLaMA* [2]



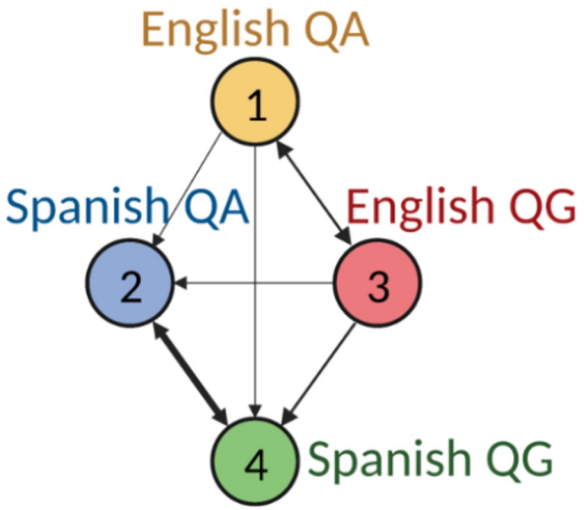
[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.
[2] Roziere B, Gehring J, Gloeckle F, et al. Code llama: Open foundation models for code[J]. arXiv preprint arXiv:2308.12950, 2023.
[3] Azerbayev Z, Schoelkopf H, Paster K, et al. Llemma: An open language model for mathematics[J]. arXiv preprint arXiv:2310.10631, 2023.

Data Scheduling - Data Composition + Data Curriculum

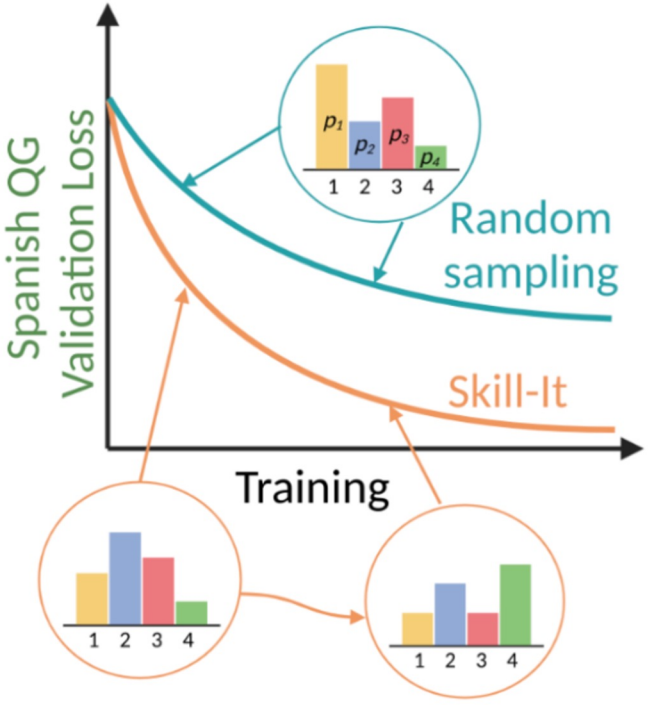
- ***Skill-it!*** [1]
  - *SKILL-IT*: online data sampling algorithm
  - data curriculum (ordered skill)
    - motivation: LMs follow a natural order when learning a set of skills



Data



Ordered skill set



[1] Chen M, Roberts N, Bhatia K, et al. Skill-it! a data-driven skills framework for understanding and training language models[J]. Advances in Neural Information Processing Systems, 2024, 36.

Thanks