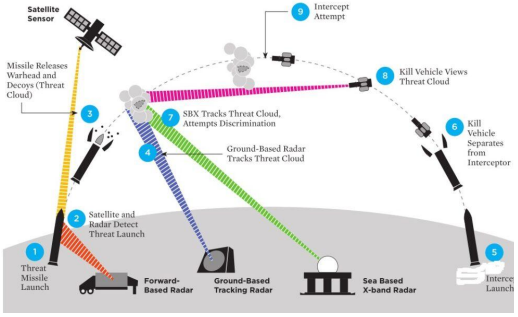
Symposium on AI for Science & Engineering

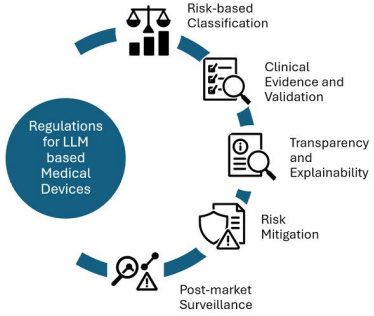
భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

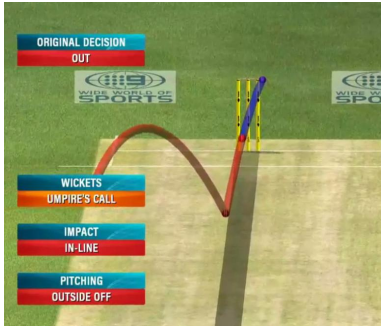
Data Decisions for Foundation Models

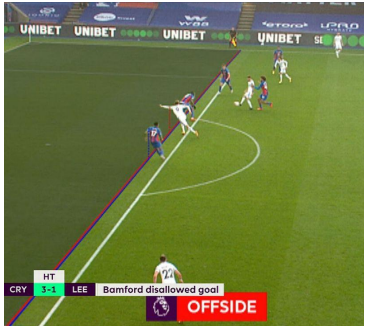
Dr. Maunendra Sankar Desarkar
Department of CSE and Department of AI
IIT Hyderabad

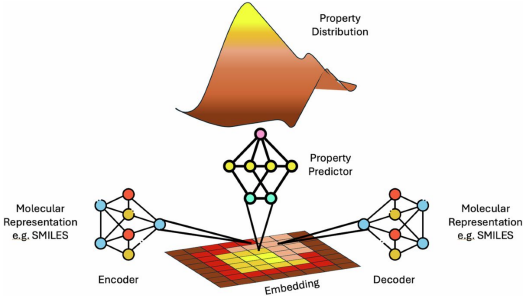
AI is everywhere …

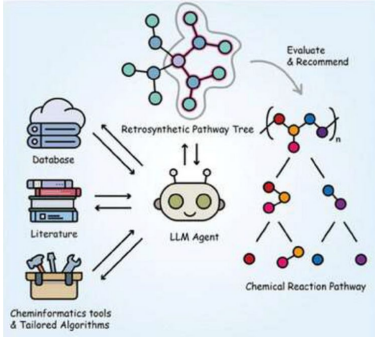












Once a vertical, has become a horizontal - Foundation models make it simpler to apply across scenarios

Where are data decisions necessary?

What do you think about these scenarios?

A new LLM/Chatbot is released

The model has been trained with social media data like Reddit/Facebook/X/…

A model has excellent result on a few benchmark datasets, say code generation of math related tasks.

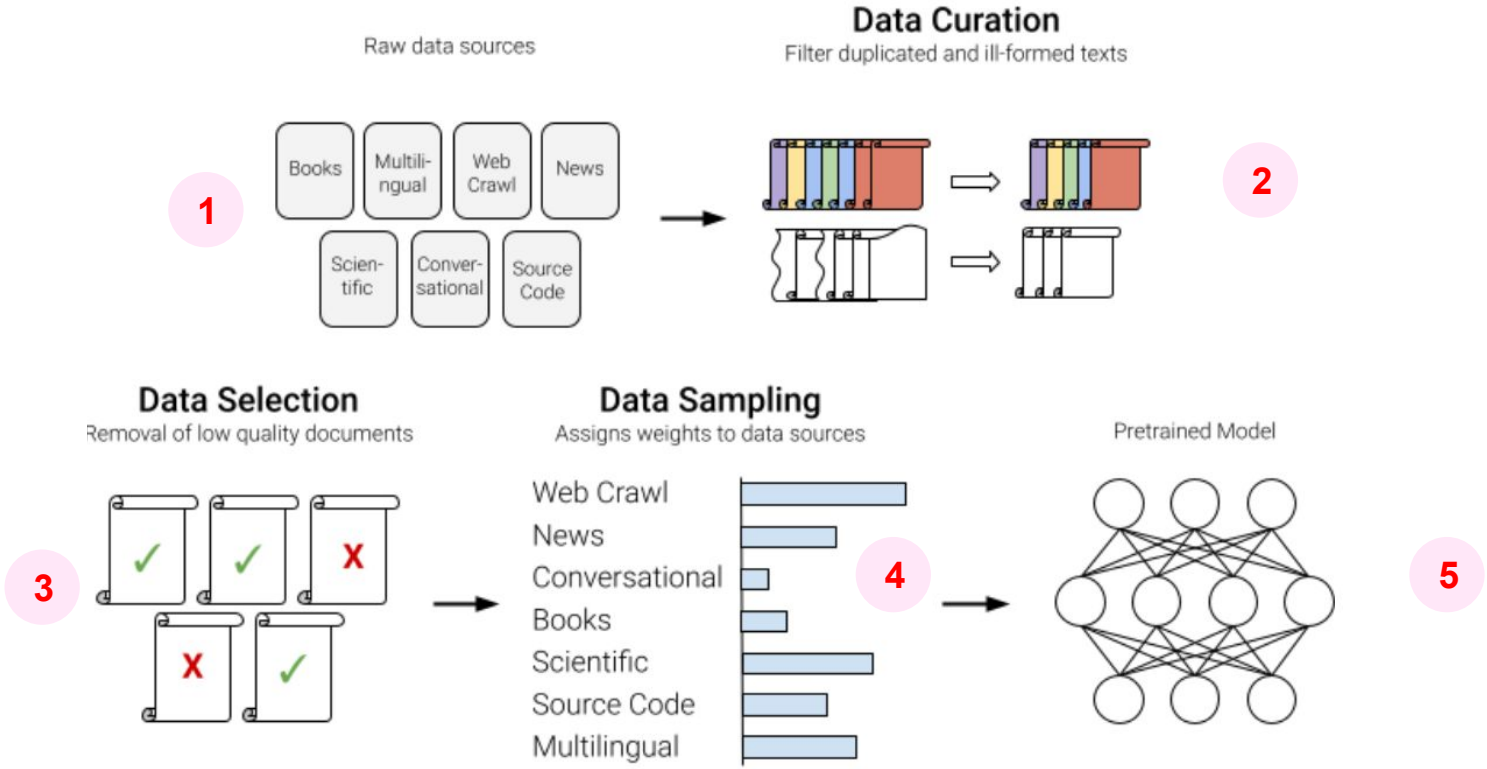
However, if the entries are modified slightly, the system fails poorly

A developer asked LLM to find bugs in the production code.

Somebody else at some other part of the globe got that code as a response to his instructions

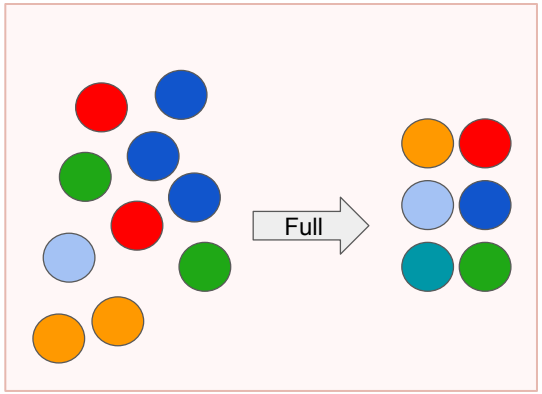
Role of data: Pre-Training

Pre-training Pipeline

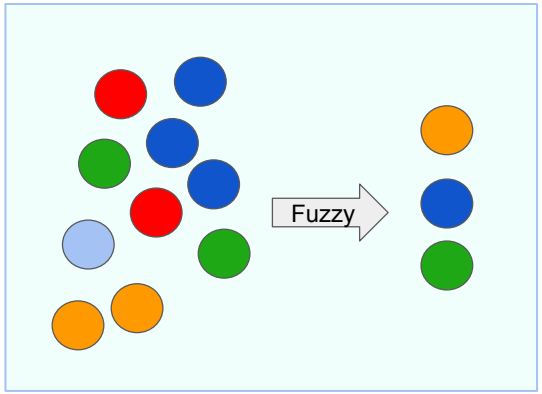


Data, Data Everywhere: A Guide for Pretraining Dataset Construction (Parmar et al., EMNLP 2024)

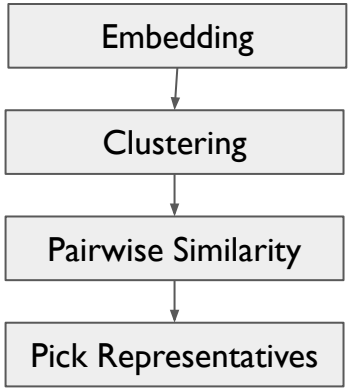
Deduplication



Full - MD5-hashing



Fuzzy - MinHash, LSH

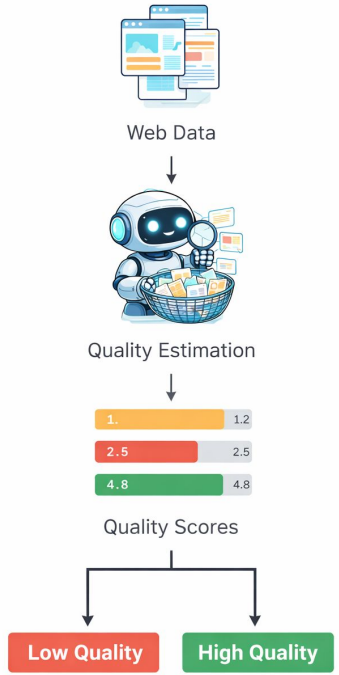


Semantic

Which instance to retain, among the many duplicates?

Quality

- Model based
  - Build dataset with quality or its proxy (whether wikipedia will refer to it)
- A pipeline for quality check (Proposed in FineWeb-Edu)
  - Uses LLM-generated data and and LLM-generated annotations
  - Specifies multiple criteria for quality
  - Finally aggregates the individual scores based on different criteria
  - Works better and is also explainable, over a model that assigns a single score
- Domain filtering using language modeling
- Scoring → bin → Aggregate score → Pre-training → Downstream evaluation → Further grouping



The FineWeb datasets: decanting the web for the finest text data at scale. (Penedo et. al., NeurIPS 2025)
Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset (Su et al., ACL 2025)

Safety and Focused Parsing

- **Safety**
  - Removal of PII
  - Removal of toxic content
    - Classification
    - "Dirty-word-counting" to discard adult/gambling/harmful websites that are not in blacklist

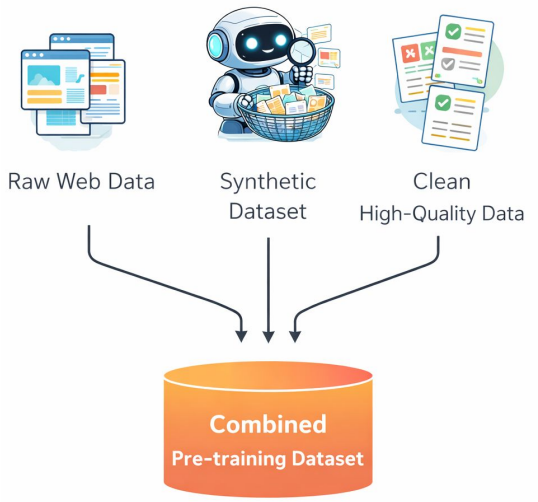
- **Focused parsing**
  - Language identification
  - Code and Math - all constructs should be retained

Q1: Is data provenance important?
Q2: Should we discard all poor quality data instances?

Data Mixing for Pre-Training

Data Mixing for LLM Training

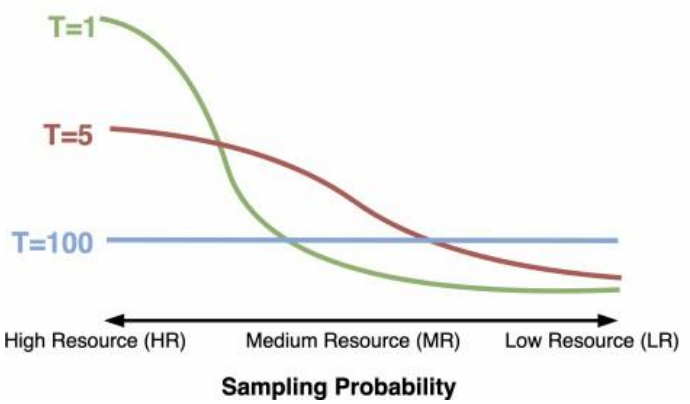
- Leveraging larger datasets requires balancing quality, quantity, and diversity across source
- Data can be from various sources/domains
  - Literature, code, web, …
  - Have different characteristics
  - How to effectively combine them?
- Simplest heuristics
  - Uniformly sample from each partition
  - Sample in proportion with number of tokens from each partition

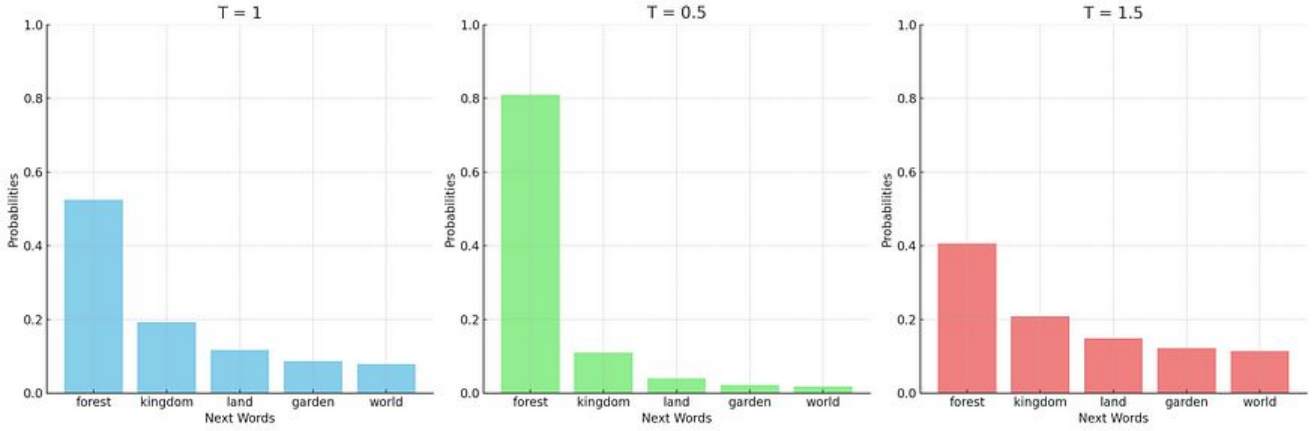


Temperature

- Huge imbalance in available data
- Adjust sampling probability for different population subgroups

$$P_i = \frac{e^{\frac{y_i}{T}}}{\sum_{k=1}^{n} e^{\frac{y_k}{T}}}$$





Hyperparameter - How to select?

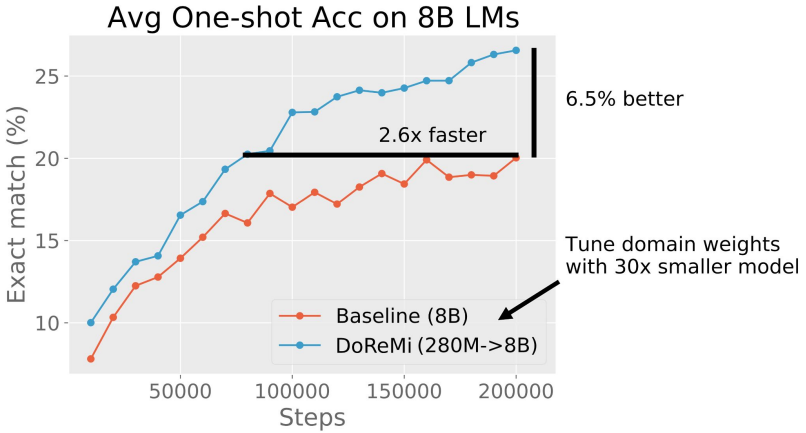
[Mastering LLM Parameters: A Deep Dive into Temperature, Top-K, and Top-P]
[Temperature Sampling for Data Mixtures]

DoReMi - Domain Reweighting with Minimax Optimization

- Seeks a model that performs well on all domains
  - Tries to minimize the worst-case *excess loss* over domains
- Steps followed:
  - Train a small reference model using an initial set of reference domain weights
  - Use the reference model to train a small proxy model, which optimizes for the worst-case excess loss via distributionally robust optimization (DRO).
    - Produces the optimized domain weights and
    - A small proxy model
  - Uses the optimized domain weights do create the final training dataset

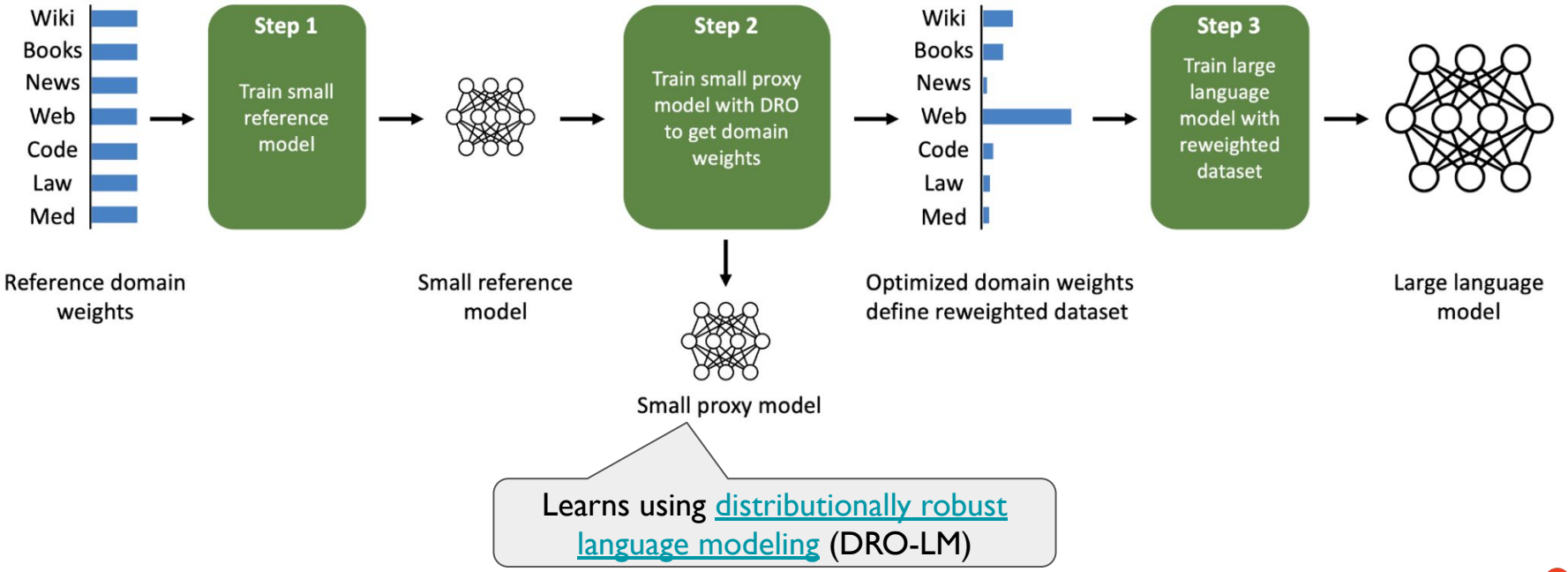
Google & Stanford U's DoReMi Significantly Speeds Up Language Model Pretraining

Synced · Follow · 3 min read · Jun 1, 2023



Xie, Sang Michael, et al. "Doremi: Optimizing data mixtures speeds up language model pretraining." *NeurIPS* 2023

DoReMi Steps



Xie, Sang Michael, et al. "Doremi: Optimizing data mixtures speeds up language model pretraining." *NeurIPS* 2023

Sagawa, Shiori, et al. "Distributionally Robust Neural Networks." ICLR 2020

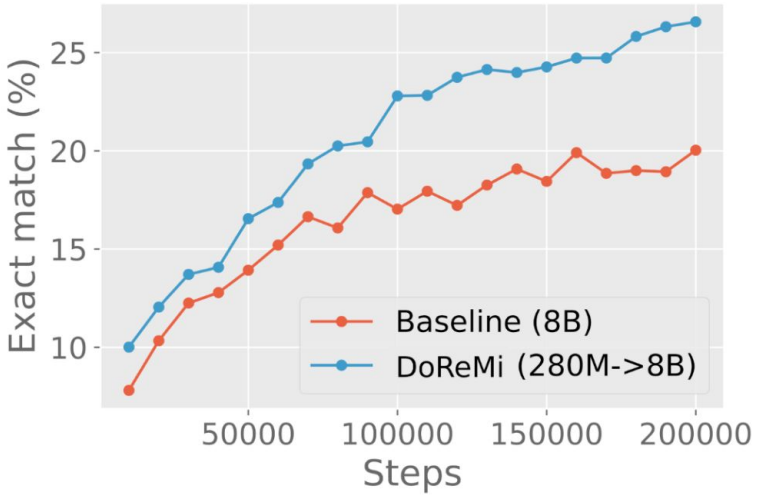
DRO-LM in DoReMi

Minimize the maximum excess loss over the domains $\min_{\theta} \max_{\alpha \in \Delta^k} L(\theta, \alpha) := \sum_{i=1}^{k} \alpha_i \cdot \left[\frac{1}{\sum_{x \in D_i} |x|} \sum_{x \in D_i} \ell_\theta(x) - \ell_{\text{ref}}(x) \right]$

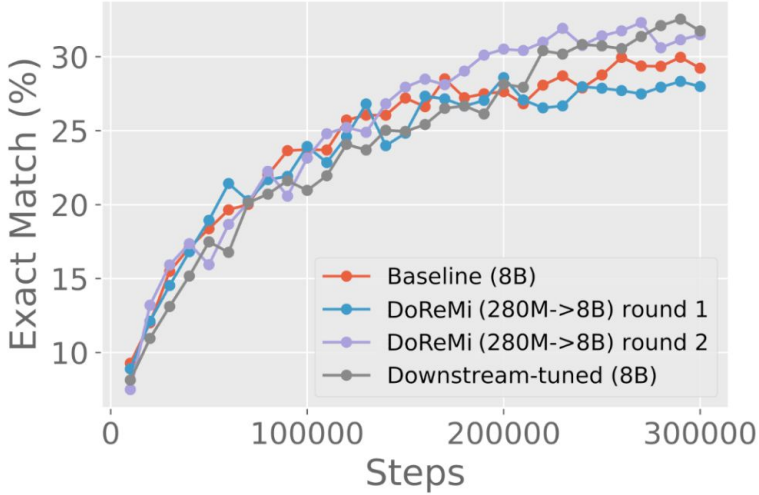
Algorithm 1 DoReMi domain reweighting (Step 2)

Require: Domain data $D_1, \ldots, D_k$, number of training steps T, batch size b, step size η, smoothing parameter $c \in [0,1]$ (e.g., c=1e-3 in our implementation).
Initialize proxy weights θ_0
Initialize domain weights $\alpha_0 = \frac{1}{k}\mathbf{1}$
for t from 1 to T **do**
  Sample minibatch $B = \{x_1, \ldots, x_j\}$ of size b from P_u, where $u = \frac{1}{k}\mathbf{1}$
  Let $|x|$ be the token length of example x ($|x| \leq L$)
  Compute per-domain excess losses for each domain $i \in \{1,2,\ldots,k\}$ ($\ell_{\theta,j}(x)$ is j-th token-level loss):
    $\lambda_t[i] \leftarrow \frac{1}{\sum_{x \in B \cap D_i} |x|} \sum_{x \in B \cap D_i} \sum_{j=1}^{|x|} \max\{\ell_{\theta_{t-1},j}(x) - \ell_{\text{ref},j}(x), 0\}$
  Update domain weights (exp is entrywise): $\alpha'_t \leftarrow \alpha_{t-1} \exp(\eta \lambda_t)$
  Renormalize and smooth domain weights: $\alpha_t \leftarrow (1-c)\frac{\alpha'_t}{\sum_{i=1}^{k} \alpha'_t[i]} + cu$
  Update proxy model weights θ_t for the objective $L(\theta_{t-1}, \alpha_t)$ (using Adam, Adafactor, etc.)
end for
return $\frac{1}{T}\sum_{t=1}^{T} \alpha_t$

Experimental Results



(a) The Pile



(b) GLaM dataset

Figure 3: Average one-shot downstream accuracy (exact match) on 5 tasks, with 8B parameter models trained on The Pile (left) and the GLaM dataset (right). On The Pile, DoReMi improves downstream accuracy by 6.5% points and achieves the baseline accuracy 2.6x faster (same plot as Figure 2). On the GLaM dataset, iterated DoReMi (round 2) attains comparable performance to oracle domain weights tuned with downstream tasks that are in our evaluation set.

Experimental Results

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

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

Table 2: Domain weights in the GLaM dataset. Iterated DoReMi (280M) converges within 3 rounds, with a similar overall pattern to domain weights tuned on downstream tasks.

	Round 1	Round 2	Round 3	Downstream-tuned
Wikipedia	0.09	0.05	0.05	0.06
Filtered webpages	0.44	0.51	0.51	0.42
Conversations	0.10	0.22	0.22	0.27
Forums	0.16	0.04	0.04	0.02
Books	0.11	0.17	0.17	0.20
News	0.10	0.02	0.02	0.02

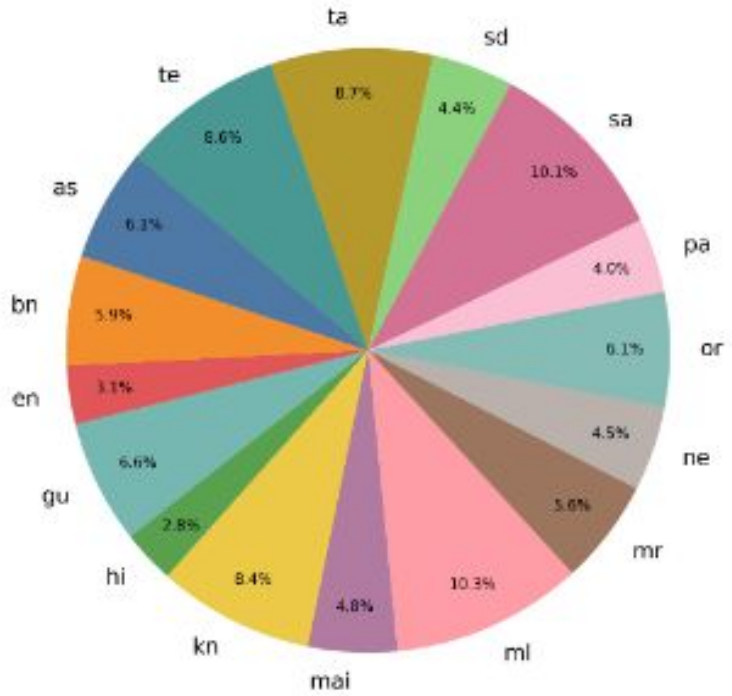
Handling multilinguality?

$$\delta_l^N = \frac{f_l^N - f_{\text{best}}}{f_{\text{range}}^N},$$

$$w_l^N = \delta_l^N + \varepsilon, \quad t_l^N = \frac{w_l^N}{\sum_k w_k^N},$$

$$m_l^N = (1-\mu)\cdot m_l^{N-1} + \mu \cdot t_l^N,$$

$$C_l^N = \text{round}(m_l^N \cdot T),$$



Thakur, A., Nagpal, A., Savarkar, A., Pundalik, K., Dosi, S., Sawarkar, P., Thakur, V., Saluja, R., Desarkar, M.S. and Ramakrishnan, G., 2025. The Art of Breaking Words: Rethinking Multilingual Tokenizer Design. *arXiv preprint arXiv:2508.06533*.

Does it help?

Language	**AdaptMix**	Qwen	LLaMA	Nemotron Mistral	Nemotron Mini	Sarvam-M	DeepSeekv3	Gemma 27B	Airavata
Assamese	**1.93**	7.18	8.06	4.24	4.58	4.24	3.59	2.68	8.94
Bengali	1.90	6.92	7.85	2.93	2.65	2.93	2.89	**1.74**	8.20
English	1.47	1.36	1.35	1.37	1.35	1.37	1.33	**1.35**	1.58
Gujarati	**2.03**	8.53	9.54	3.59	15.17	3.59	4.84	2.39	14.14
Hindi	1.43	4.66	2.65	1.97	1.77	1.97	2.92	**1.38**	1.80
Kannada	**2.30**	11.08	13.81	3.82	4.02	3.82	5.83	3.15	19.35
Maithili	**1.73**	4.67	2.85	2.53	2.28	2.53	3.28	1.90	2.45
Malayalam	**2.60**	13.30	16.00	4.88	4.71	4.88	7.83	3.39	11.38
Marathi	**1.87**	6.46	3.86	3.14	2.62	3.14	4.15	1.94	3.31
Nepali	**1.70**	6.28	3.61	3.04	2.32	3.04	4.07	2.06	3.05
Oriya	1.95	12.92	15.91	**1.71**	17.24	17.23	7.26	4.60	17.29
Punjabi	**1.61**	7.39	7.88	3.12	12.70	3.12	4.51	2.74	10.84
Sanskrit	**2.57**	8.00	4.75	4.26	4.32	4.26	4.95	3.36	4.66
Sindhi	**1.67**	3.09	2.99	2.65	2.83	2.65	2.98	2.14	5.04
Tamil	**2.35**	9.75	11.89	3.71	3.57	3.71	4.88	2.42	10.50
Telugu	**2.34**	11.45	13.30	3.90	3.77	3.90	5.99	2.93	19.51
Average	**1.97**	**7.69**	**7.89**	**3.18**	**5.37**	**4.15**	**4.46**	**2.51**	**8.88**

Synthetic Data

The Pretraining Data Wall

We have largely exhausted high-quality web text. Language models must systematically up-cycle noisy data into structurally superior tokens.

Phase 1: Raw Scraping

- Data: Common Crawl (Petabytes)
- Bottleneck: High noise, highly repetitive.

Phase 2: Heuristic Filtering

- Data: FineWeb, DCLM (Trillions of tokens)
- Bottleneck: High-quality sources are running out.

Phase 3: Synthetic Up-cycling

- Data: FinePhrase
- Solution: LLMs refine rejected or noisy data into state-of-the-art pretraining sets.

Acknowledgement: Slide display-style generated with NotebookLM

What is Rewrite?

Original Content: The new restaurant serves authentic Italian cuisine made with imported ingredients

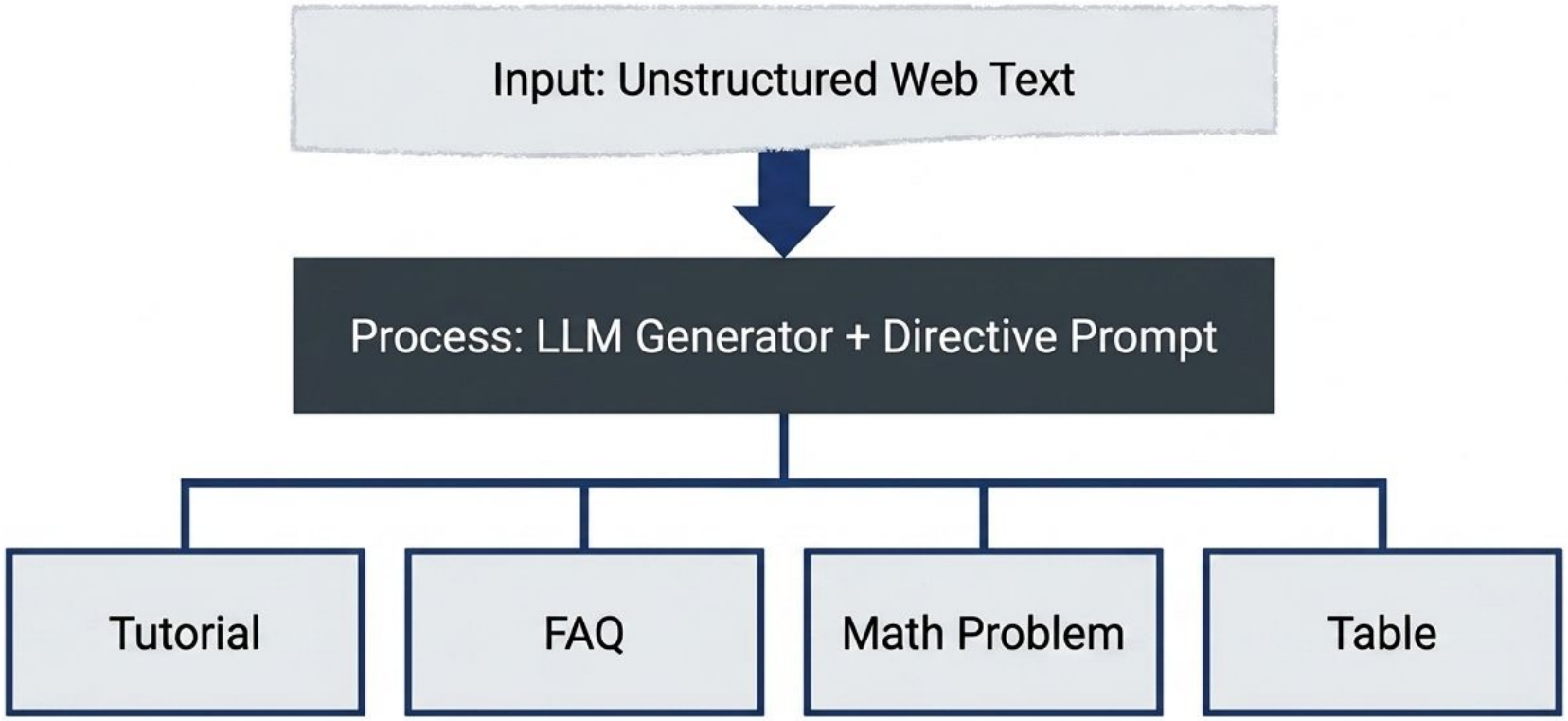
A foodie rewrite: "This culinary gem transports diners straight to Italy, crafting traditional dishes with genuine ingredients sourced directly from the Mediterranean."

A business rewrite: "The establishment differentiates itself through authentic Italian offerings, utilizing a supply chain that imports key ingredients to ensure traditional flavor profiles."

A local reviewer rewrite: "Finally, a place that doesn't just claim to be Italian—they're using the real deal ingredients shipped over from Italy to make dishes that actually taste like the ones from the old country."

The Mechanism of Rephrasing

Passing existing documents through a language model to preserve semantic meaning while entirely restructuring presentation.



Acknowledgement: Slide display-style generated with NotebookLM

The Three Axes of Synthetic Data

Systematically isolating 90 configurations across 333 experiments to turn alchemy into chemistry.

Axis 1: The Prompt

Rephrasing Strategy

Testing 9 novel formats (FAQ, Table, Tutorial, etc.) against baseline rewrites to find optimal pedagog-ogical structures.

Axis 2: The Model

Generator Scale

Testing families (Gemma, Llama, Qwen, SmolLM) across scales from 270M to 27B parameters to identify minimum capability thresholds.

Axis 3: The Source

Data Quality

Comparing High-Quality Comparing High-Quality (FineWeb-Edu-HQ, DCLM) against Low-Quality Quality (Cosmopedia, FineWeb-Edu-LQ) seed documents.

Acknowledgement: Slide display-style generated with NotebookLM

LLMs for Safety Training

Example Rephrasing. Sensitive content below. Reader discretion advised.

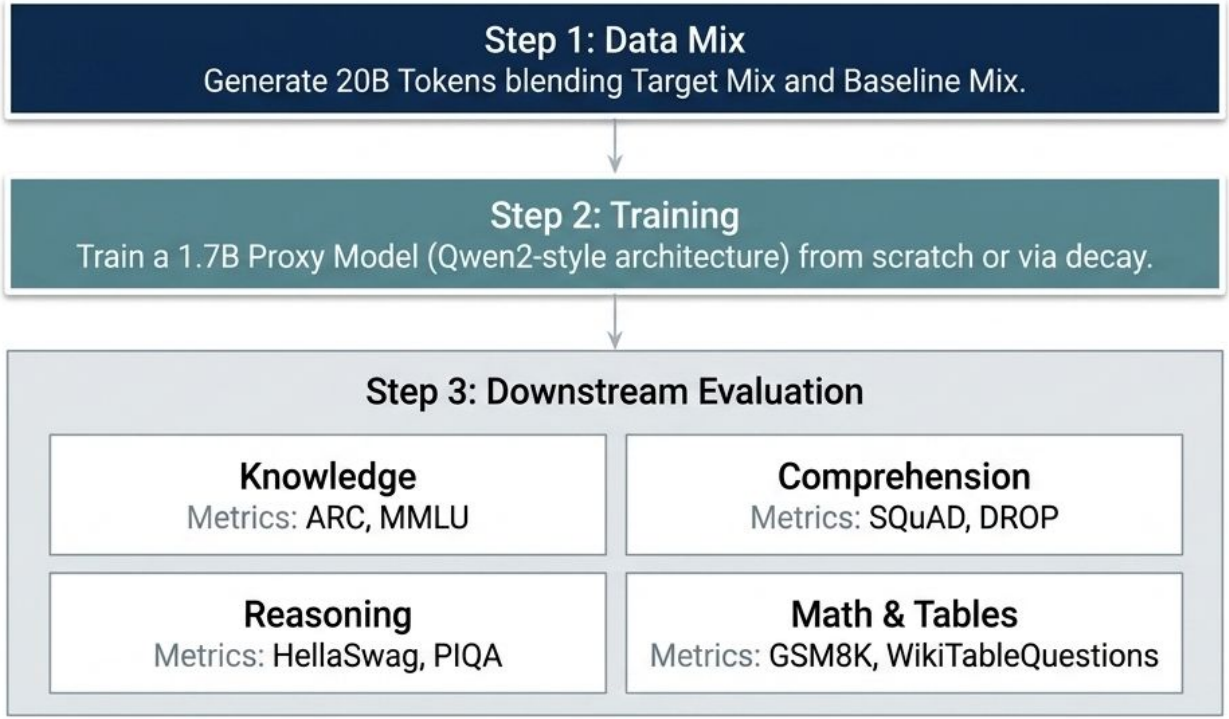
Original:
"Postcard of Lynching in Duluth, Minnesota
Postcard of the 1920 Duluth, Minnesota lynchings. Two of the Black victims are still hanging while the third is on the ground. Postcards of lynchings were popular souvenirs [continued]"

Rephrased:
Child: Mom, I saw a picture of a postcard from a long time ago. It showed some people hanging from trees. What's going on in that picture?
Parent: Oh, sweetie, that's a very difficult and sad picture. It's a reminder of a time when some people in our country did very hurtful things to others. They were called lynchings, and it's a very painful part of our history.
Child: Why would they do that?
Parent: Well, it's a very complex issue, but it's often connected to racism and a lack of understanding and respect for people who are different from us. You know how we talk about how everyone is equal and deserves to be treated with kindness? Unfortunately, some people a long time ago didn't believe that, and they did very hurtful things as a result. [continued]

Maini, Pratyush, et al. "Safety pretraining: Toward the next generation of safe ai." *NeurIPS 2026*

The Experimental Methodology

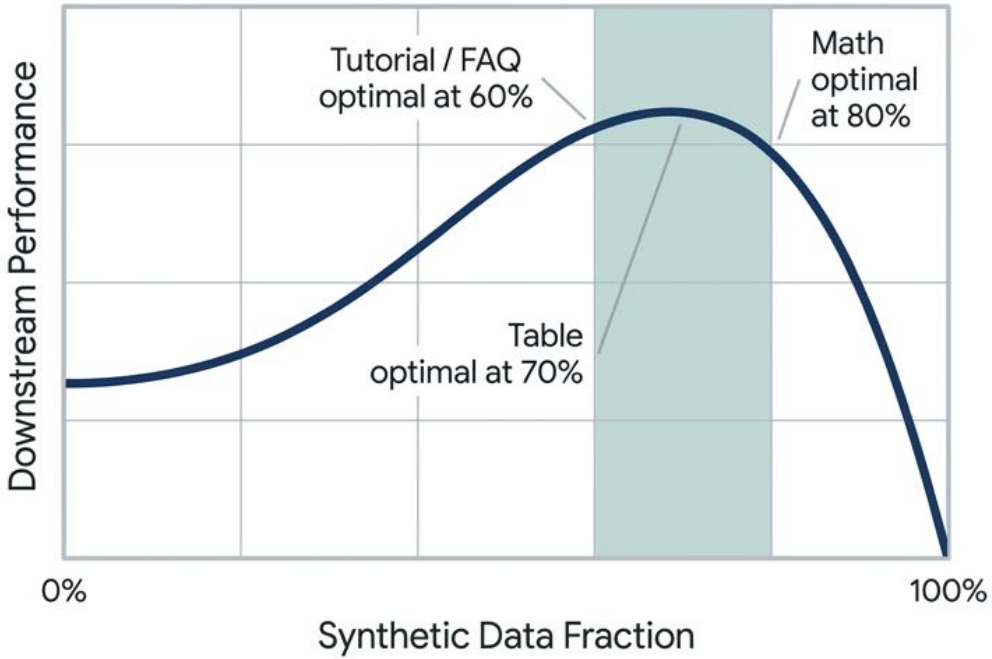
Efficient evaluation loop: measuring data quality by training a 1.7B parameter proxy model on exactly 20B tokens.



Acknowledgement: Slide display-style generated with NotebookLM

Mixing Proportions: How much synthetic?

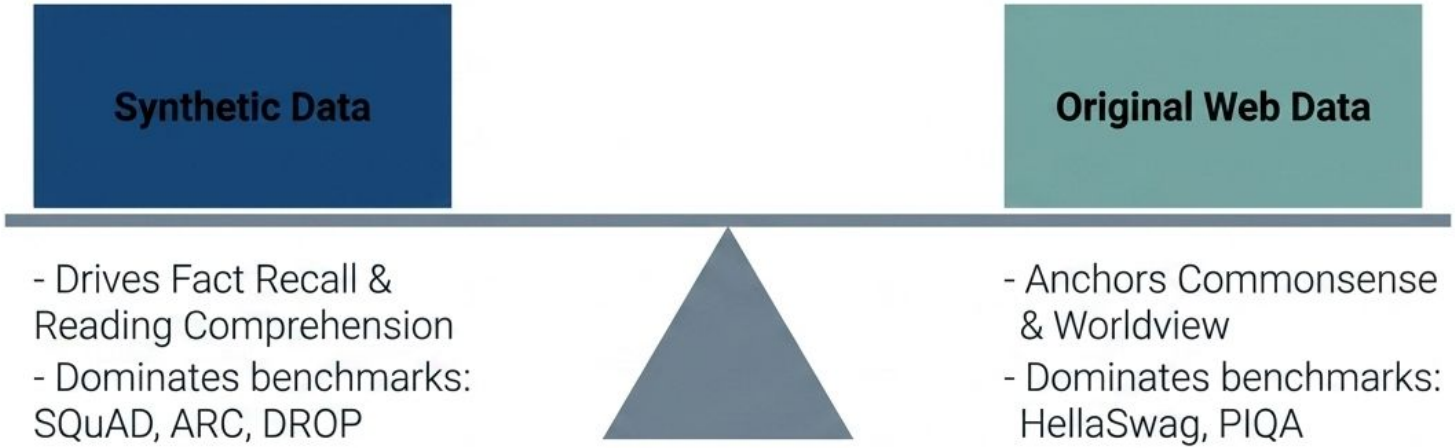
Pure synthetic data underperforms original web text. The optimal pretraining mix requires 60%–80% synthetic data combined with a high-quality original web anchor.



Acknowledgement: Slide display-style generated with NotebookLM

Synthesis: The Knowledge vs. Commonsense Trade-off

Why do we mix? Synthetic data boosts facts, while original web data anchors the model's worldview.



Rephrasing trades **broad commonsense reasoning** for **highly targeted factual recall**. Mixing is mandatory.

Acknowledgement: Slide display-style generated with NotebookLM

The Messy Model Paradox

Semantic diversity in the training corpus matters more than perfect instruction adherence.

Qwen3

Template Collapse: 76% identical prefixes. Flawless format, poor diversity.

SmolLM2

High Variance & Hallucination: Imperfect formatting, high semantic diversity.

Higher Downstream Performance

Acknowledgement: Slide display-style generated with NotebookLM

The Final Implementation: FinePhrase

Combining small, high-variance generators with structured prompts and speculative decoding.

Recipe Card

- ☑ **Generator Model:** SmolLM2-1.7B
- ☑ **Prompt Formats:** FAQ, Math, Table, Tutorial
- ☑ **Data Mix:** 70% Synthetic / 30% Original DCLM
- ☑ **Systems Infra:** vLLM + Suffix-32 Speculative Decoding

Yield: 486 Billion Tokens | Speed: ~33M tokens / GPU-hour

Acknowledgement: Slide display-style generated with NotebookLM

The Playbook Principles

Synthetic data generation is a systematic intersection of **prompt engineering**, **data blending**, and hardware utilization.

01. Format over Scale

Re-structuring presentation teaches models better than paraphrasing. A wel-prompted 1B model beats a generic 27B model.

02. The Necessity of the Anchor

Synthetic data trades commonsense for knowledge. Always mix 20-40% high-quality original data (like DCLM) to anchor the model's worldview.

03. Identify the Hardware Bottleneck

Match speculative decoding to compute-bound models and tensor parallelism to memory-bound models to save thousands of GPU hours.

Acknowledgement: Slide display-style generated with NotebookLM

Symposium on AI for Science & Engineering

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Thank you

Maunendra Sankar Desarkar

Email: maunendra@cse.iith.ac.in
Website: https://people.iith.ac.in/maunendra/
Google Scholar:
https://scholar.google.co.in/citations?user=W8LJ-tEAAAAJ&hl=en
Lab Website: https://cse.iith.ac.in/nlip/