





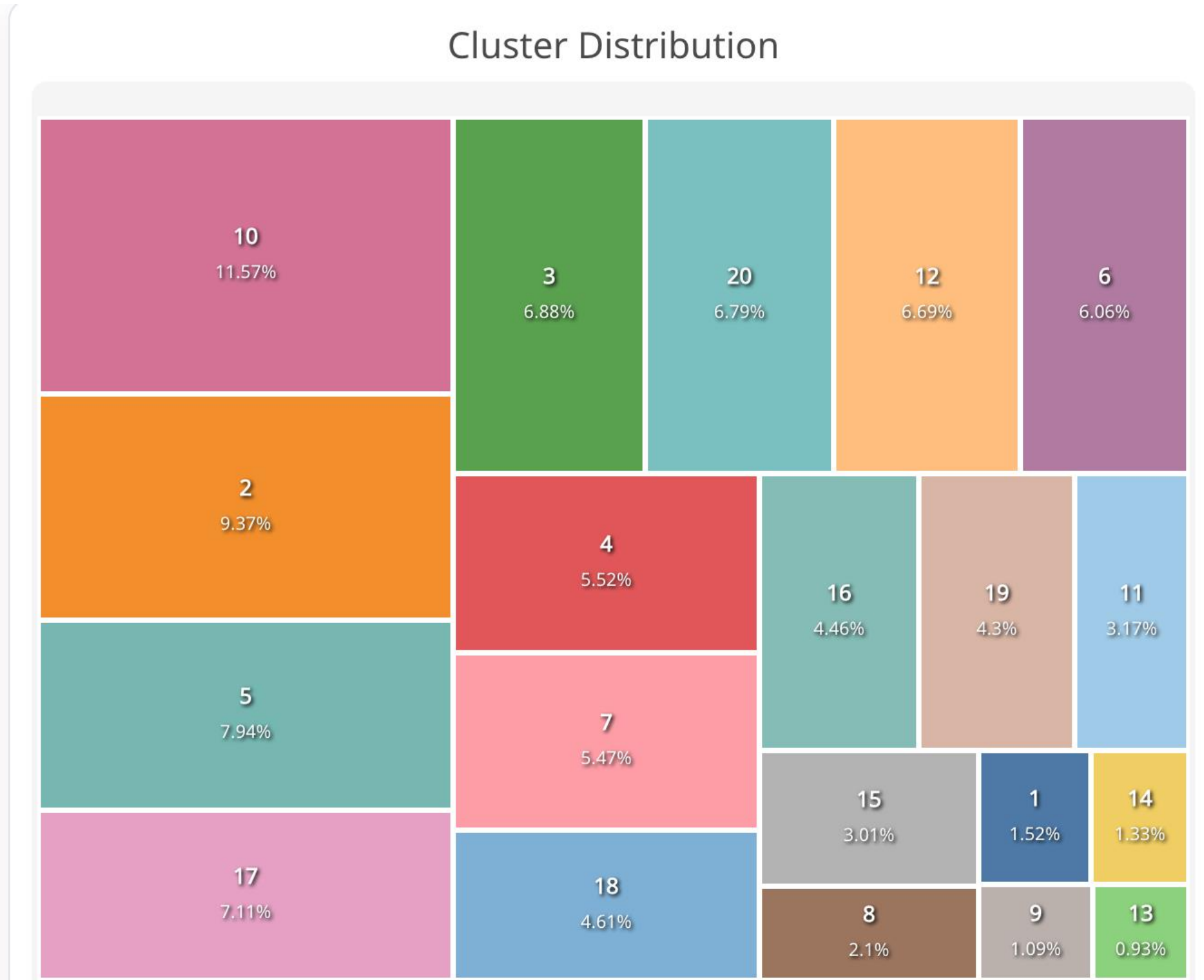
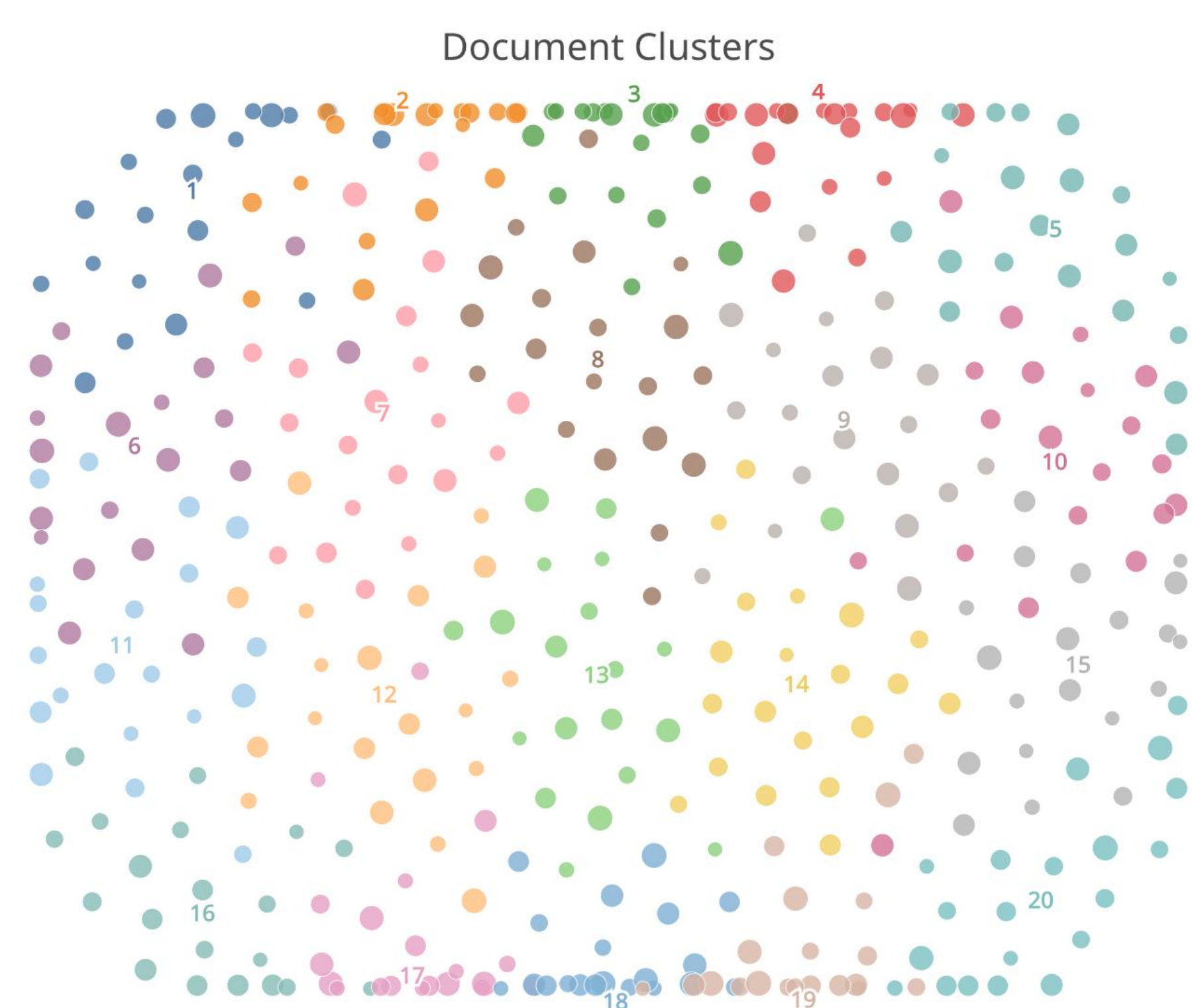
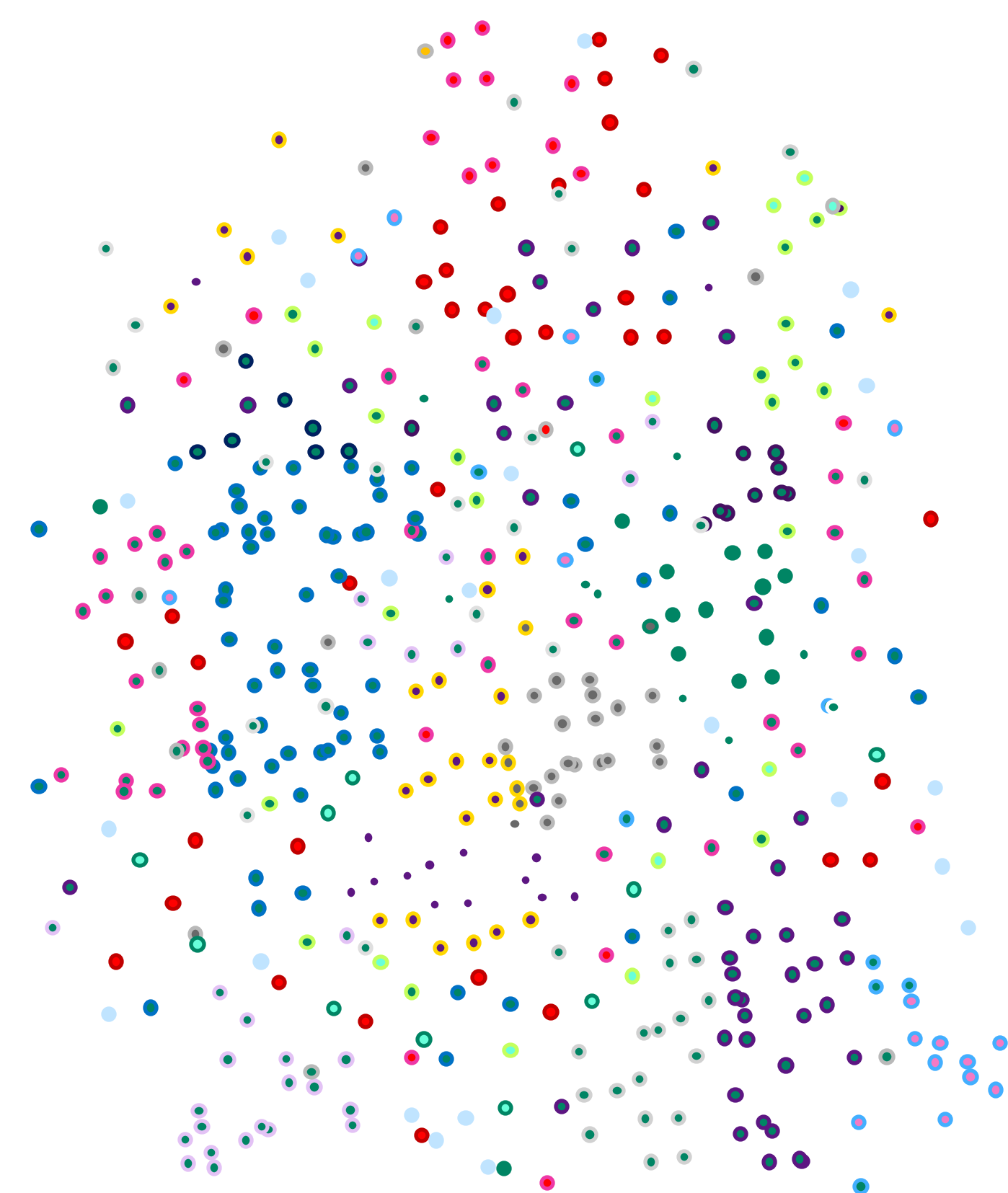


CLIMB: CLustering-based Iterative Data Mixture Bootstrapping for Language Model Pre-training

Shizhe Diao
May 21 2025

Trends and Challenges in Pre-training Data Curation

Trend	Research Question
Collect from multiple known sources    → The Pile	How to mix and balance domains when domain labels are known?
Crawl the open web  →  Common Crawl  → Nemotron-CC	How to organize and mix data to construct a good data blend when domain labels are missing?



Expectation

- **Quality** and **Efficiency**: High quality data leading to better data efficiency
- **Diversity**: Data diversity matters, we need domain-aware data
- **Search**: Neural Data mixtures as an optimization problem



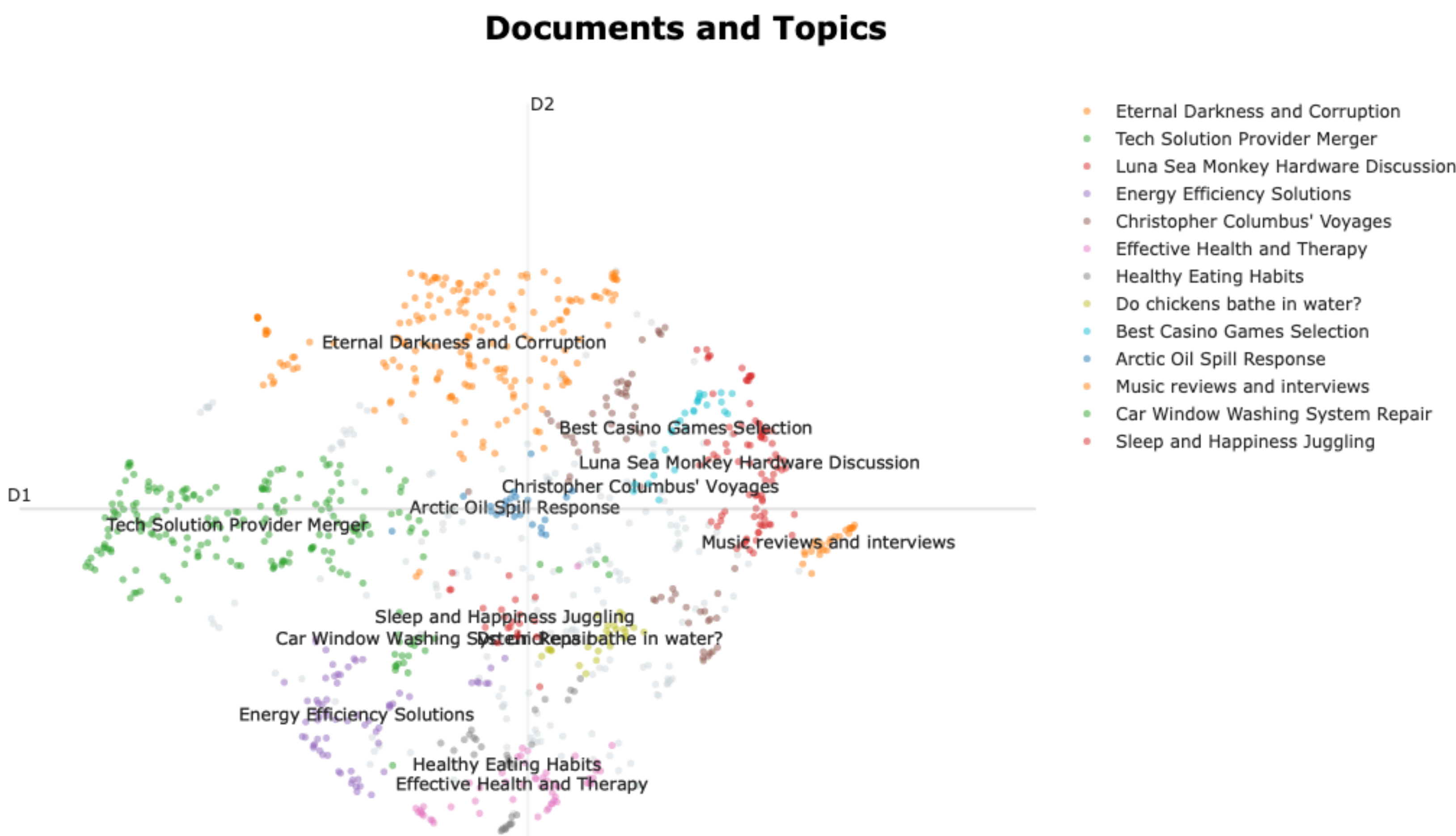
Pilot Study

Exploration

- Data Organization: **CLIMB-Clustering**
 - Clustering
 - Cluster huge data into different clusters based on semantic meaning
 - Data quality and data diversity
 - Neural Classifier
 - Prompt Teacher model to generate data to train classifiers
 - More fine-grained scores with clear criteria and efficient classifiers
- Data Mixing: **CLIMB-Search**
 - Lightweight Proxy Model
 - Predictor Training

Data Clustering

- Step 1: Embedding. Convert texts to embedding vectors
 - Embedding Model: stella_en_400M_v5 (400M, dim=1024)
- Step 2: Clustering.
 - Clustering algorithm: K-Means
- Step 3: Cluster Pruning.
 - Verify each cluster based on some proxy
 - High- (keep), Medium- (drop), Low- (drop) quality clusters



Massive Text Embedding Benchmark (MTEB) Leaderboard. To submit, refer to the [MTEB GitHub repository](#) 🤖 Refer to the [MTEB paper](#) for details on metrics, tasks and models. Also check out [MTEB Arena](#) ✂

Search Bar (separate multiple queries with `;`)

Search for a model and press enter...

Model types

☒ Open ☒ Proprietary ☒ Sentence Transformers ☒ Cross-Encoders ☒ Bi-Encoders

☒ Uses Instructions ☒ No Instructions

Model sizes (in number of parameters)

☒ <100M ☒ 100M to 250M ☒ 250M to 500M

☒ 500M to 1B ☒ >1B

Overall Bitext Mining Classification Clustering Pair Classification Reranking Retrieval STS Summarization Retrieval w/Instructions

English Chinese French Polish Russian

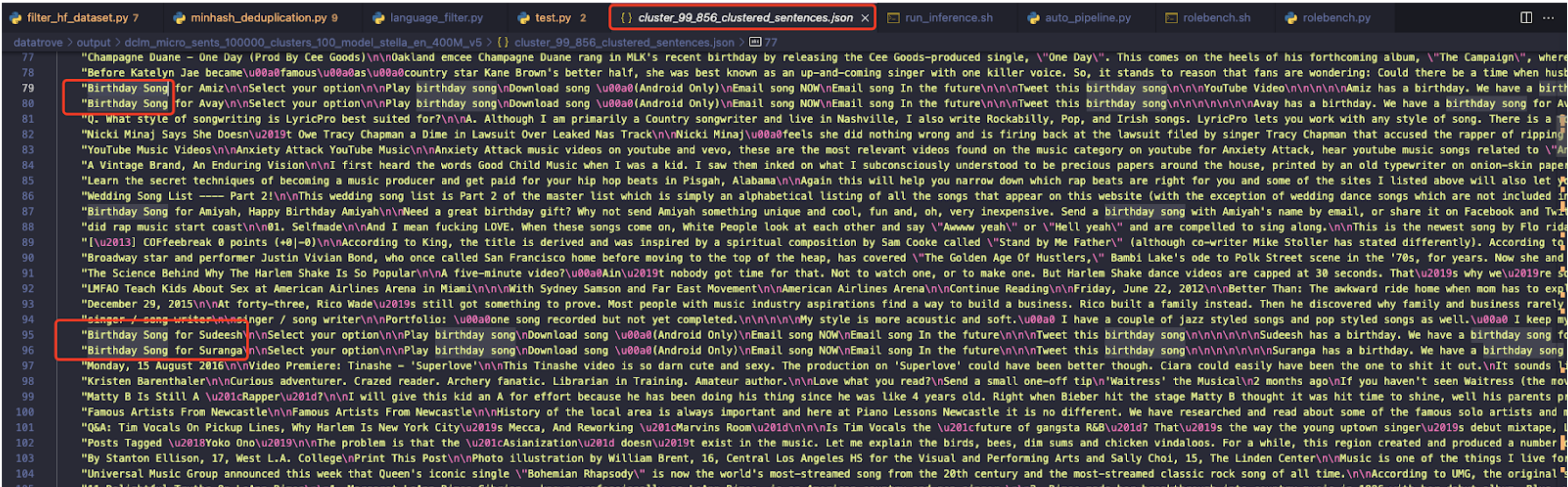
Overall MTEB English leaderboard 🏆

- Metric: Various, refer to task tabs
- Languages: English

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	PairClassification Average (3 datasets)
1	NV-Embed-v2	7851	29.25	4096	32768	72.31	90.37	58.46	88.67
2	bge-en-ic1	7111	26.49	4096	32768	71.67	88.95	57.89	88.14
3	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19	87.63	57.69	88.07
4	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31	89.05	56.17	88.07
5	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24	86.58	56.92	85.79
6	stella_en_400M_v5	435	1.62	8192	8192	70.11	86.67	56.7	87.74
7	bge-multilingual-gemma2	9242	34.43	3584	8192	69.88	88.08	54.65	85.84
8	NV-Embed-v1	7851	29.25	4096	32768	69.32	87.35	52.8	86.91
9	ling-Embed-Mistral	7111	26.49	4096	32768	68.17	80.2	51.42	88.35
10	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67	88.54

Qualitative Analysis of Data Clustering

- Finding #1: clustering could deduplicate docs (semdedup).



- Finding #2: identify bad data semantically

Cluster 163: Advertisement

```
[62]: import pandas as pd
df = pd.DataFrame({"Document": corpus, "Topic": topics})

df[df.Topic == 163]
```

	Document	Topic
2929	JulesB Discount Code\nVisit JulesB >>\nCheck o...	163
3569	Discount Golf Store Promo Code & Voucher Code\...	163
4379	Funyroot Coupon Code and Promo Code May 2021\n...	163
8275	Kind TechGroup - Coupons & Deals, October 2020...	163
11015	FORUMS DEALS COUPONS S...	163
13931	Posted in: Mobile phones\nSony offers 10% disc...	163
17333	Welcome to AmeriMark! (log in or create an acc...	163

```
grammar_error_doc = "Thiss is a txt taht shows how speling and gramer errors can make it hard for rader to understnd the mesage. Sometimes the

ads_doc = "Looking for the best deals on laptops? Visit www.example.com today and get up to 50% off on all electronics! Our store offers a w

repeat_sent_doc = "The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over

excessive_special_chars_doc = "###@%***This!!!!?? is a text full of ~~~special$$$!!! characters &&& that don't really contribute**** an

non_human_doc = "00110011Error^^_non_text_alpha@@##_404signal***noise!!!text replaced<>123__&invalid^^code detected!!!retry system i

new_docs = [grammar_error_doc, ads_doc, repeat_sent_doc, excessive_special_chars_doc, non_human_doc]
# Create embeddings
sentence_model = SentenceTransformer(MODEL_NAME, trust_remote_code=True)
new_embeddings = sentence_model.encode(new_docs, show_progress_bar=True)

topic_model.transform(new_docs, new_embeddings)

2024-09-21 17:29:26,276 - BERTopic - Dimensionality - Reducing dimensionality of input embeddings.
2024-09-21 17:29:26,302 - BERTopic - Dimensionality - Completed ✓
2024-09-21 17:29:26,303 - BERTopic - Clustering - Approximating new points with `hdbscan_model`
2024-09-21 17:29:26,305 - BERTopic - Cluster - Completed ✓

[[-1, 163, -1, -1, 54], it is `page not found` cluster
array([0.          , 1.          , 0.          , 0.          , 0.69716257],
      dtype=float32))
```


Qualitative Analysis of Data Clustering

- Finding #1: clustering could deduplicate docs ([sem dedup](#)).

```
filter_hf_dataset.py 7  minhash_deduplication.py 9  language_filter.py  test.py 2  cluster_99_856_clustered_sentences.json  x  run_inference.sh  auto_pipeline.py  rolebench.sh  rolebench.py

datatrove > output > dclm_micro_sents_100000_clusters_100_model_stella_en_400M_v5 > {} cluster_99_856_clustered_sentences.json > 77

77  "Champagne Duane - One Day (Prod By Cee Goods)\n\nOakland emcee Champagne Duane rang in MLK's recent birthday by releasing the Cee Goods-produced single, \"One Day\". This comes on the heels of his forthcoming album, \"The Campaign\", where
78  "Before Katelyn Jae became\u00a0famous\u00a0as\u00a0country star Kane Brown's better half, she was best known as an up-and-coming singer with one killer voice. So, it stands to reason that fans are wondering: Could there be a time when hust
79  "Birthday Song for Amiz\n\nSelect your option\n\nPlay birthday song\nDownload song \u00a0(Android Only)\nEmail song NOW\nEmail song In the future\n\n\nTweet this birthday song\n\n\nYouTube Video\n\n\n\nAmiz has a birthday. We have a birth
80  "Birthday Song for Avay\n\nSelect your option\n\nPlay birthday song\nDownload song \u00a0(Android Only)\nEmail song NOW\nEmail song In the future\n\n\nTweet this birthday song\n\n\n\n\nAvay has a birthday. We have a birthday song for Av
81  "Q. What style of songwriting is LyricPro best suited for?\n\nA. Although I am primarily a Country songwriter and live in Nashville, I also write Rockabilly, Pop, and Irish songs. LyricPro lets you work with any style of song. There is a
82  "Nicki Minaj Says She Doesn't Owe Tracy Chapman a Dime in Lawsuit Over Leaked Nas Track\n\nNicki Minaj\u00a0feels she did nothing wrong and is firing back at the lawsuit filed by singer Tracy Chapman that accused the rapper of ripping
83  "YouTube Music Videos\n\nAnxiety Attack YouTube Music\n\nAnxiety Attack music videos on youtube and vevo, these are the most relevant videos found on the music category on youtube for Anxiety Attack, hear youtube music songs related to \"Ar
84  "A Vintage Brand, An Enduring Vision\n\nI first heard the words Good Child Music when I was a kid. I saw them inked on what I subconsciously understood to be precious papers around the house, printed by an old typewriter on onion-skin paper
85  "Learn the secret techniques of becoming a music producer and get paid for your hip hop beats in Pisgah, Alabama\n\nAgain this will help you narrow down which rap beats are right for you and some of the sites I listed above will also let y
86  "Wedding Song List ---- Part 2!\n\nThis wedding song list is Part 2 of the master list which is simply an alphabetical listing of all the songs that appear on this website (with the exception of wedding dance songs which are not included in
87  "Birthday Song for Amiyah, Happy Birthday Amiyah\n\nNeed a great birthday gift? Why not send Amiyah something unique and cool, fun and, oh, very inexpensive. Send a birthday song with Amiyah's name by email, or share it on Facebook and Twi
88  "did rap music start coast\n\n01. Selfmade\n\nAnd I mean fucking LOVE. When these songs come on, White People look at each other and say \"Awww yeah\" or \"Hell yeah\" and are compelled to sing along.\n\nThis is the newest song by Flo rida
89  "[\u00a02013] COFfeebreak 0 points (+0|-0)\n\nAccording to King, the title is derived and was inspired by a spiritual composition by Sam Cooke called \"Stand by Me Father\" (although co-writer Mike Stoller has stated differently). According to
90  "Broadway star and performer Justin Vivian Bond, who once called San Francisco home before moving to the top of the heap, has covered \"The Golden Age Of Hustlers,\" Bambi Lake's ode to Polk Street scene in the '70s, for years. Now she and
91  "The Science Behind Why The Harlem Shake Is So Popular\n\nA five-minute video?\u00a0Ain't nobody got time for that. Not to watch one, or to make one. But Harlem Shake dance videos are capped at 30 seconds. That\u00a02019s why we\u00a02019re st
92  "LMFAO Teach Kids About Sex at American Airlines Arena in Miami\n\nWith Sydney Samson and Far East Movement\n\nAmerican Airlines Arena\n\nContinue Reading\n\nFriday, June 22, 2012\n\nBetter Than: The awkward ride home when mom has to exp
93  "December 29, 2015\n\nAt forty-three, Rico Wade\u00a02019s still got something to prove. Most people with music industry aspirations find a way to build a business. Rico built a family instead. Then he discovered why family and business rarely
94  "singer / song-writer\n\nsinger / song writer\n\nPortfolio: \u00a0one song recorded but not yet completed.\n\n\n\n\nMy style is more acoustic and soft.\u00a0 I have a couple of jazz styled songs and pop styled songs as well.\u00a0 I keep my
95  "Birthday Song for Sudeesh\n\nSelect your option\n\nPlay birthday song\nDownload song \u00a0(Android Only)\nEmail song NOW\nEmail song In the future\n\n\nTweet this birthday song\n\n\n\n\nSudeesh has a birthday. We have a birthday song fo
96  "Birthday Song for Suranga\n\nSelect your option\n\nPlay birthday song\nDownload song \u00a0(Android Only)\nEmail song NOW\nEmail song In the future\n\n\nTweet this birthday song\n\n\n\n\nSuranga has a birthday. We have a birthday song
97  "Monday, 15 August 2016\n\nVideo Premiere: Tinashe - 'Superlove'\n\nThis Tinashe video is so darn cute and sexy. The production on 'Superlove' could have been better though. Ciara could easily have been the one to shit it out.\n\nIt sounds l
98  "Kristen Barenthaler\n\nCurious adventurer. Crazy reader. Archery fanatic. Librarian in Training. Amateur author.\n\nLove what you read?\nSend a small one-off tip\n'Waitress' the Musical\n2 months ago\nIf you haven't seen Waitress (the nov
99  "Matty B Is Still A \u00a0201cRapper\u00a0201d?\n\nI will give this kid an A for effort because he has been doing his thing since he was like 4 years old. Right when Bieber hit the stage Matty B thought it was hit time to shine, well his parents pr
100 "Famous Artists From Newcastle\n\nFamous Artists From Newcastle\n\nHistory of the local area is always important and here at Piano Lessons Newcastle it is no different. We have researched and read about some of the famous solo artists and
101 "Q&A: Tim Vocals On Pickup Lines, Why Harlem Is New York City\u00a02019s Mecca, And Reworking \u00a0201cMarvins Room\u00a0201d\n\n\n\nIs Tim Vocals the \u00a0201cfuture of gangsta R&B\u00a0201d? That\u00a02019s the way the young uptown singer\u00a02019s debut mixtape, l
102 "Posts Tagged \u00a02018Yoko Ono\u00a02019\n\nThe problem is that the \u00a0201cAsianization\u00a0201d doesn't exist in the music. Let me explain the birds, bees, dim sums and chicken vindaloos. For a while, this region created and produced a number
103 "By Stanton Ellison, 17, West L.A. College\n\nPrint This Post\n\nPhoto illustration by William Brent, 16, Central Los Angeles HS for the Visual and Performing Arts and Sally Choi, 15, The Linden Center\n\nMusic is one of the things I live for
104 "Universal Music Group announced this week that Queen's iconic single \"Bohemian Rhapsody\" is now the world's most-streamed song from the 20th century and the most-streamed classic rock song of all time.\n\nAccording to UMG, the original's
```


Qualitative Analysis of Data Clustering

- Finding #2: identify bad data semantically

Cluster 163: Advertisement

```
[62]: import pandas as pd
df = pd.DataFrame({"Document": corpus, "Topic": topics})

df[df.Topic == 163]
```

	Document	Topic
2929	JulesB Discount Code\nVisit JulesB >>\nCheck o...	163
3569	Discount Golf Store Promo Code & Voucher Code\...	163
4379	Funyroot Coupon Code and Promo Code May 2021\n...	163
8275	Kind TechGroup - Coupons & Deals, October 2020...	163
11015	FORUMS DEALS COUPONS S...	163
13931	Posted in: Mobile phones\nSony offers 10% disc...	163
17333	Welcome to AmeriMark! (log in or create an acc...	163

```
grammar_error_doc = "Thiss is a txt taht shows how speling and gramer errors can make it hard for rader to understnd the mesage. Somtimes the
ads_doc = "Looking for the best deals on laptops? Visit www.example.com today and get up to 50% off on all electronics! Our store offers a w
repeat_sent_doc = "The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over
excessive_special_chars_doc = "###@@@%%This!!!??? is a text full of ~special$$$!!! characters &&& that don't really contribute**** an
non_human_doc = "00110011Error_^^__non_text__alpha@@##__404signal***noise!!!text replaced<>123__&&invalid^^code detected!!!retry system i

new_docs = [grammar_error_doc, ads_doc, repeat_sent_doc, excessive_special_chars_doc, non_human_doc]
# Create embeddings
sentence_model = SentenceTransformer(MODEL_NAME, trust_remote_code=True)
new_embeddings = sentence_model.encode(new_docs, show_progress_bar=True)
```

```
topic_model.transform(new_docs, new_embeddings)

2024-09-21 17:29:26,276 - BERTopic - Dimensionality - Reducing dimensionality of input embeddings.
2024-09-21 17:29:26,302 - BERTopic - Dimensionality - Completed ✓
2024-09-21 17:29:26,303 - BERTopic - Clustering - Approximating new points with `hdbscan_model`
2024-09-21 17:29:26,305 - BERTopic - Cluster - Completed ✓

([-1, 163, -1, -1, 54], it is `page not found` cluster
array([0.          , 1.          , 0.          , 0.          , 0.69716257],
      dtype=float32))
```


Neural Classifier

- Step 1: Leverage larger teacher model (Nemotron-340B) evaluating a subset of pre-training data.
- Step 2: With teacher’s generation, train fasttext-based quality classifiers
- Step 3: Filter cluster with model-based quality classifiers

Evaluation Criteria for Pre-Training Data

You are an expert evaluator assessing a text for suitability as pre-training data for a large language model. For each criterion, start from 0 points. Then add points based on the conditions described. If no conditions are met, the score remains 0. Please evaluate the given text using the rating scale below. Assign a score from 0 to 5 for each criterion, and reference the expanded guidelines under each category to determine the appropriate rating:

Rating Scale:

- 0: Does not meet the criterion at all
- 1: Partially meets the criterion
- 2: Fairly meets the criterion
- 3: Mostly meets the criterion
- 4: Fully meets the criterion
- 5: Exceeds the criterion

Criteria and Expanded Guidelines:

- Quality:** The text is natural, clean, and free from severe grammatical errors, spelling mistakes, syntactical issues, repetitive phrasing, or random symbols.
 - +1: Correct basic spelling and mostly proper grammar, despite minor slips.
 - +1: Coherent sentence structures, no glaring syntactical breakdowns.
 - +1: Natural language, free from repetitive phrasing, easy to read.
 - +1: Polished, no major grammatical errors or spelling mistakes.
 - +1: Professional-level writing quality, free from unnatural phrasing.
- Advertisement:** The text should avoid excessive promotional language or overt advertising.
 - +1: Minimal promotional elements, not distracting.
 - +1: Subtle promotional aspects, not overshadowing content.
 - +1: Mostly neutral with slight marketing-like language.
 - +1: Almost free from advertisements, at most one mild reference.
 - +1: No detectable promotional content.
- Informational Value:** The text provides accurate insights, useful facts, or relevant knowledge.
 - +1: At least one accurate fact or relevant information.
 - +1: Multiple useful pieces of information.
 - +1: Enhances understanding, presents explanations.
 - +1: Substantial, well-structured, reliable information.
 - +1: Exceptional depth, authoritative content.
- Educational Value:** Assess if the text is beneficial for structured learning.
 - +1: Basic educational relevance, even if mixed with non-academic content.
 - +1: Addresses education but lacks strong alignment with standards.
 - +1: Suitable for educational use, introduces key concepts.
 - +1: Highly relevant for structured learning, minimal extraneous content.
 - +1: Outstanding educational value, clear, easy-to-follow insights.

Final Output Format:

```
{
  "quality": < integer 0-5 >,
  "advertisement": < integer 0-5 >,
  "informational_value": < integer 0-5 >,
  "educational_value": < integer 0-5 >,
}
```

Content to evaluate:

INPUT_DOC

Qualitative Analysis of Clustering-based Data Filtering

- Step 1: Clustering Nemotron-CC-HQ (500B tokens) into 1000 clusters
- Step 2: Train several classifiers: advertisement > 2, educational > 1 based on the annotations from Nemotron-340B
- Step 3: Apply classifiers to filter bad clusters out

	HQ	Filtered out	Total
# of clusters	875	125	1000
# of samples	661M	84M (10%)	746M

Advertisement

(\n<<<EVERY ORDER FROM THIS SALE RECEIVES A
FREE GIFT VALUED AT \$25!>>>\n\nI have this, bought
from COTD, and the biggest problem is, after changing
bits a few times, the bits keep falling out as there's
not enough inside the holder to grab them.

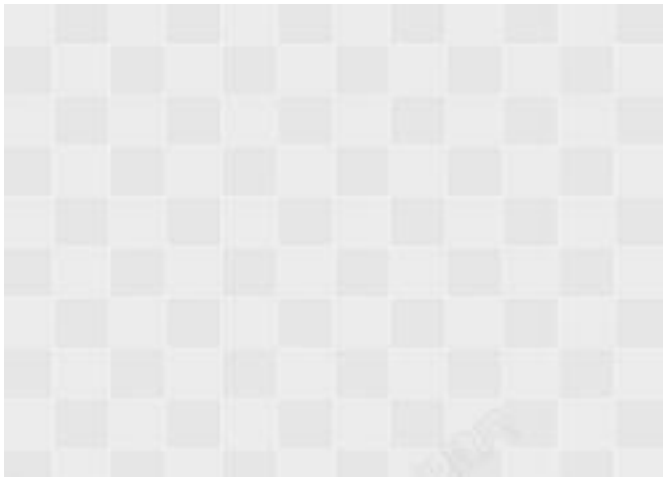
{"text": "paid \$21.87 for an iPaad 2657 which only
cost me \$62.81 to buy. Here is the website we use to
get it all from,"}

Garbage Content

```
(base) sdiao@cw-dfw-cs-001-dc-02:/lustre/fsw/portfolios/llmservi  
ce/users/sdiao/data/nemotron-cc-hq-clusters$ head cluster_171.js  
onl  
{ "text": "You" }  
{ "text": "You" }  
{ "text": "that requestMEMO THAT REQUESTMEMO THAT REQUEST SAMPLE"  
}  
{ "text": "it" }  
{ "text": "T" }  
{ "text": "less" }  
{ "text": "One" }  
{ "text": "PLA" }  
{ "text": "can" }  
{ "text": "represent" }
```

```
(base) sdiao@cw-dfw-cs-001-dc-02:/lustre/fsw/portfolios/llmservi  
ce/users/sdiao/data/nemotron-cc-hq-clusters$ h  
ead cluster_99.jsonl  
{ "text": "3307 group" }  
{ "text": "51" }  
{ "text": "7 hours 7 hours" }  
{ "text": "0495504297\n\n9780495504290 «Show less" }  
{ "text": "Rights Skripsi No. 03010021/ELK/2004; Adry (23400" }  
{ "text": "44.4 #168 1:19AM193 6:38PM" }  
{ "text": "ID#: 82344; Salary:1999.00; Salary:--; Salary:--" }  
{ "text": "8 gets stuck131416including" }  
{ "text": "7.07.2011\n\nX Minus Oneaired" }  
{ "text": "5,658" }
```

Adult Content



CLIMB-Clustering

Clustering-based Data Filtering

- Results on Nemotron-CC-HQ (500B tokens) data looks promising
 - CLIMB-Clustering does not hurt performance, offers gains to MMLU and Math.
 - It also has great potential to reduce the model's toxicity.
 - Apply the same process to larger, lower-quality datasets, expect even greater improvements
- Note: the original data is of high-quality.

Data	Avg_MMLU	Avg_code	Avg_Math	Avg_other	Avg.
Nemotron-CC-HQ	35.80	26.68	21.91	49.71	29.14
-good	36.53	26.54	22.82	49.89	29.57

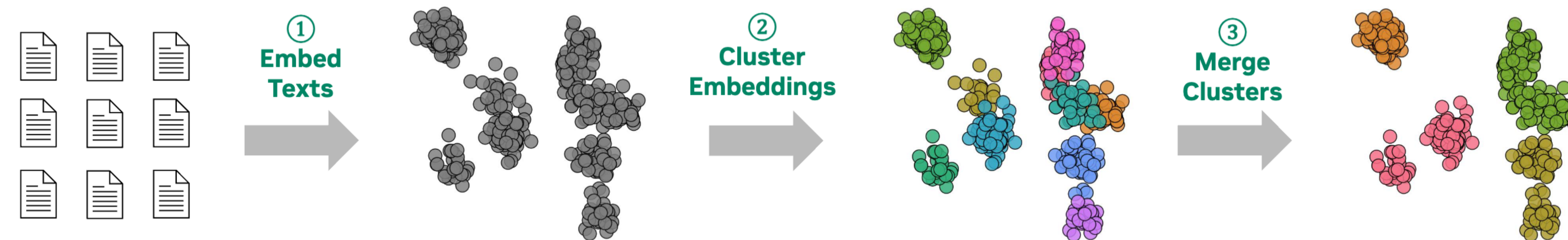


CLIMB: the approach

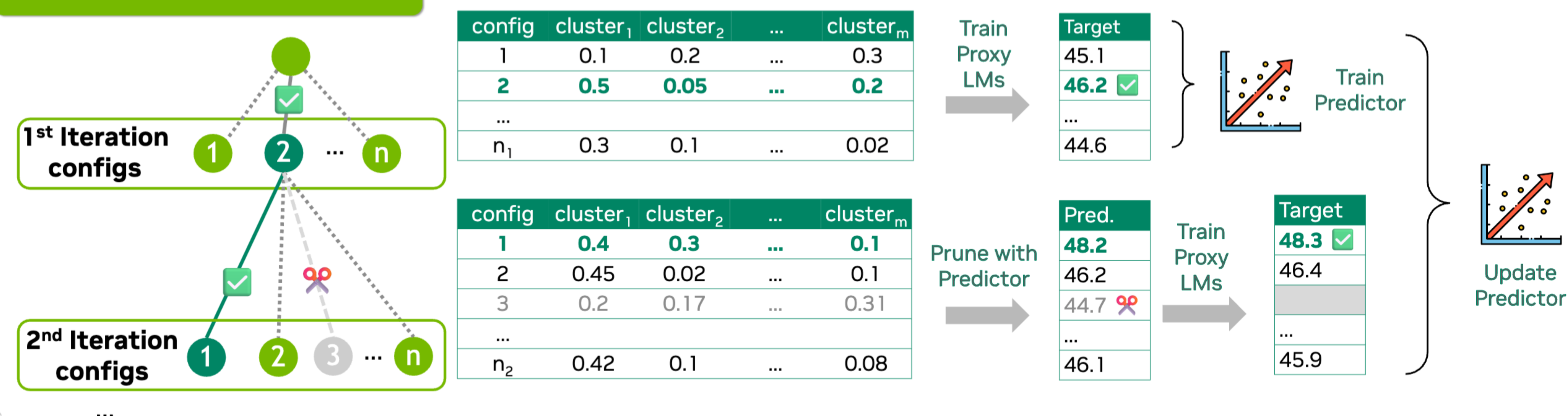
CLIMB algorithm

1. Embedding all documents into semantic space with embedding models.
2. Grouping all embeddings into N clusters via clustering algorithm (e.g., K-Means).
3. Identifying and filtering bad clusters with quality filters and merging good clusters into M super-clusters.
4. With the remaining high-quality clusters, searching the optimal mixture weights to combine them. This step, we will train proxy models and a regression model to predict the optimal mixture weights.

(a) Data Preprocessing



(b) Mixture Bootstrapping



(c) Optimal Mixture Weights

Use the predictor after K iterations to get the optimal data mixture weights.

Experiments

- Pre-training
 - Phase1
 - Phase2
 - Phase3
 - Phase4
- Implementation
 - Embedding Model: NovaSearch/stella_en_400M-v5
 - Clustering: K-means into 1000 clusters
 - Cluster Merging: reduce the clusters from 1000 to 240 with neural classifiers, then group into 21 clusters based on distance.
 - Iterative Bootstrapping: 64, 32, 16 searches in the first, second, and third iterations.
 - Predictor Training: LightGBM regression model, L1 and L2 regularization, early stopping, max_depth=4 to prevent overfitting.

Phase2 Pre-training

- Starting point: Phase1 Pre-trained on 10T tokens (dclm and txt360) using warmup-stable-decay LR schedule
- Proxy Model: 350M Transformer, Target Model: 1B Transformer
- Starter Data: Nemotron-CC + smollm-corpus
- CLIMB: 21 clusters containing 800B tokens

Table 1 | Comparison with data mixture methods. All models are continuously trained on the same number of tokens (40B). The best results are highlighted in **bold**. Base refers to the model before training and serves as the starting point for all other models. We report perplexity for wiki and lambda, accuracy for arc_e, winogrande, siqa, accuracy_norm for piqa, arc_c, hellaswag.

Size	Model	Proxy	wiki	lambda	piqa	arc_c	arc_e	hellaswag	winogrande	siqa	avg.
350M	Base	-	22.70	8.87	70.03	28.11	56.12	51.16	54.48	40.75	50.11
	Random	-	20.92	9.85	71.16	30.54	62.50	52.14	55.40	41.29	52.17
	Doremi	350M	19.41	10.39	70.29	33.53	66.41	52.25	55.95	41.86	53.38
	RegMix	350M	20.93	10.32	71.92	33.42	66.12	53.69	55.27	42.23	53.78
	CLIMB	350M	19.67	9.29	72.21	34.87	67.25	55.32	56.79	42.54	54.83
1B	Base	-	17.79	6.65	73.89	34.92	66.77	62.12	59.82	41.26	56.46
	Random	-	17.82	6.53	74.05	37.12	70.24	62.90	60.77	42.48	57.93
	Doremi	350M	15.78	6.33	74.91	40.01	72.34	63.53	61.08	43.09	59.16
	RegMix	350M	16.19	6.62	75.22	40.42	71.32	64.73	62.33	42.22	59.37
	CLIMB	350M	15.96	6.44	75.78	40.98	72.97	66.01	63.32	43.37	60.41

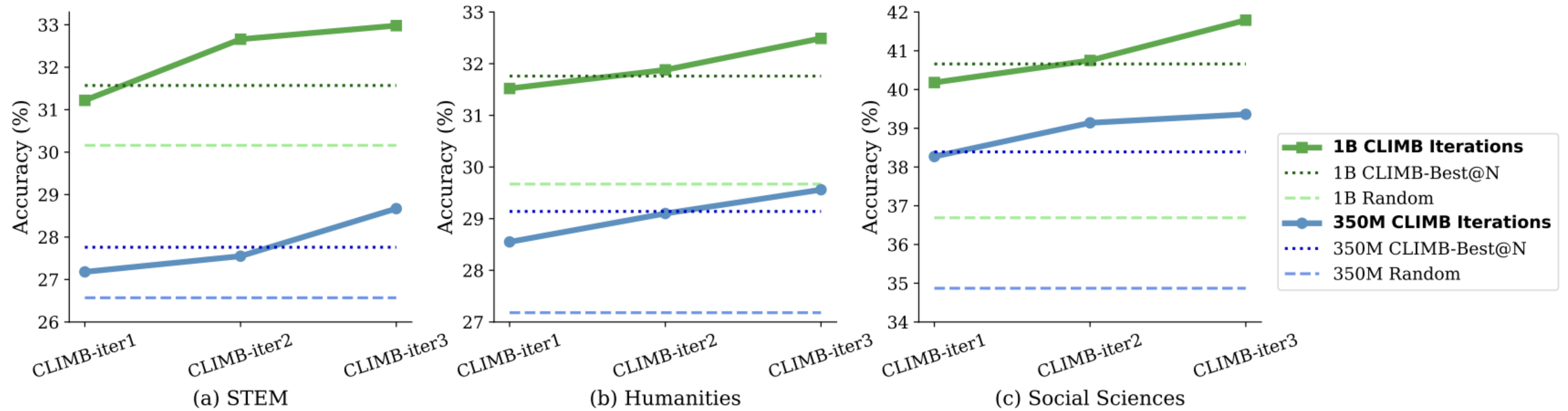
Phase2 Pre-training

- Starting point: Phase1 Pre-trained on 10T tokens (dclm and txt360) using warmup-stable-decay LR schedule
- Proxy Model: 350M Transformer, Target Model: 1B Transformer
- Starter Data: Nemotron-CC + smollm-corpus
- CLIMB: 21 clusters containing 800B tokens

Table 2 | Comparison with state-of-the-art language models on general reasoning benchmarks. CLIMB is continuously trained on 400B tokens with the optimal data mixture. The best results are highlighted in **bold**.

Model	Size	piqa	arc_c	arc_e	hellaswag	winogrande	siqa	mmlu	obqa	boolq	race	lambda	truthfulqa	Avg.
Qwen2.5	490M	69.96	32.42	64.60	52.14	56.59	44.22	33.03	35.20	62.29	34.93	52.51	39.74	48.14
SmolLM	360M	71.49	36.00	70.08	53.52	56.75	41.20	32.98	37.60	55.29	34.74	45.76	37.93	47.78
CLIMB (Ours)	350M	72.52	35.07	67.38	56.27	57.93	42.88	33.28	36.60	62.29	33.39	52.62	36.86	48.93
TinyLlama	1.1B	73.29	30.12	60.31	59.19	59.12	40.63	31.60	36.00	57.83	36.46	58.84	37.60	48.42
AMD-OLMo	1.2B	75.63	33.70	65.95	63.61	61.64	44.17	31.92	35.80	60.58	34.64	59.31	32.22	49.93
Llama-3.2	1.2B	74.59	36.26	65.49	63.67	60.69	42.99	35.40	37.20	63.98	37.80	62.99	37.67	51.56
CLIMB (Ours)	950M	75.46	40.96	73.57	66.90	63.54	43.55	36.47	41.20	66.02	36.65	59.05	39.06	53.54

Analysis – Domain Adaptation



Ablation Study

- 1. More compute in search, better performance
- 2. Balance search depth and breadth.
- 3. Proxy model size could be reduced.
- 4. Robust to number of clusters
- 5. Robust to initialization method

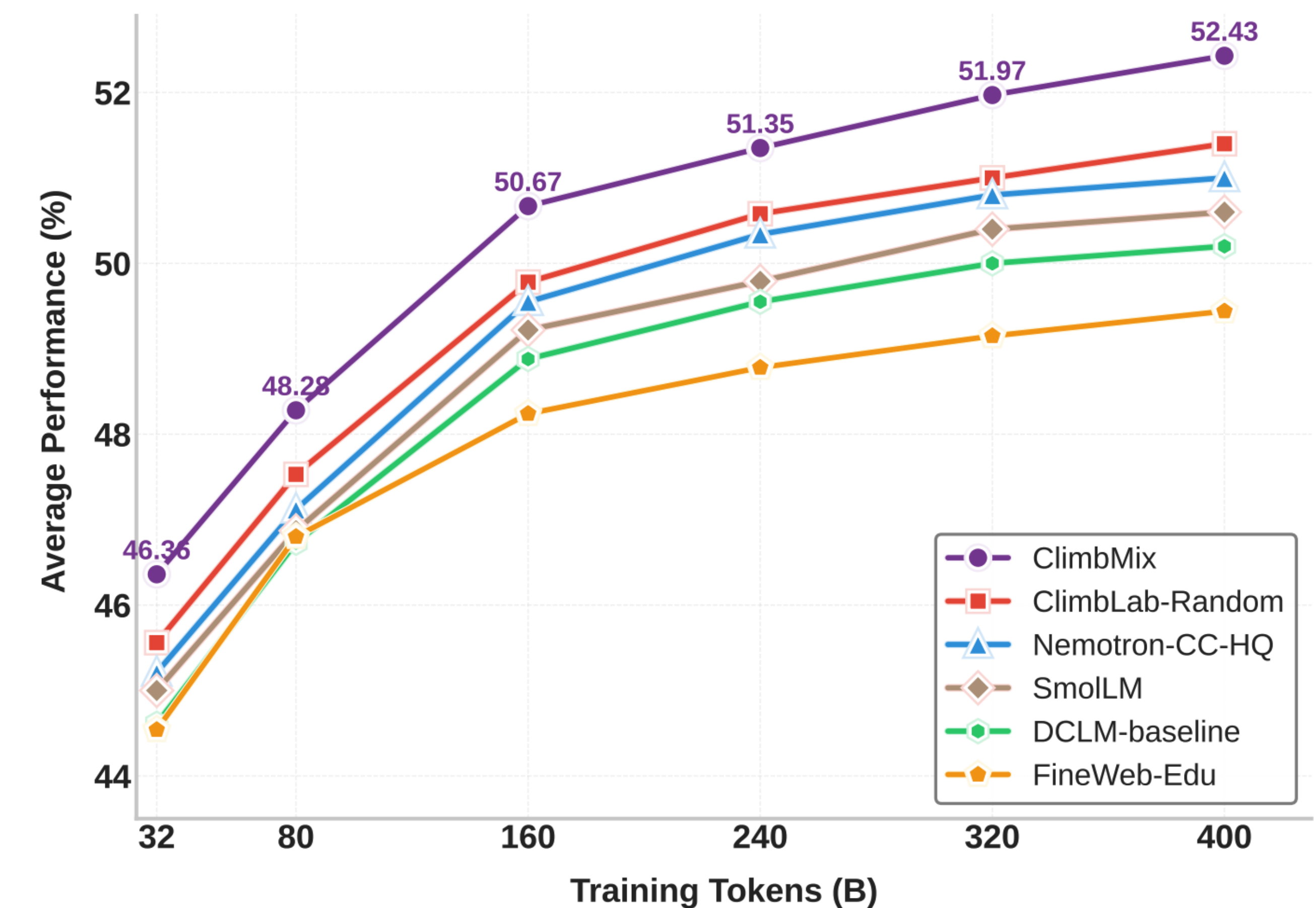
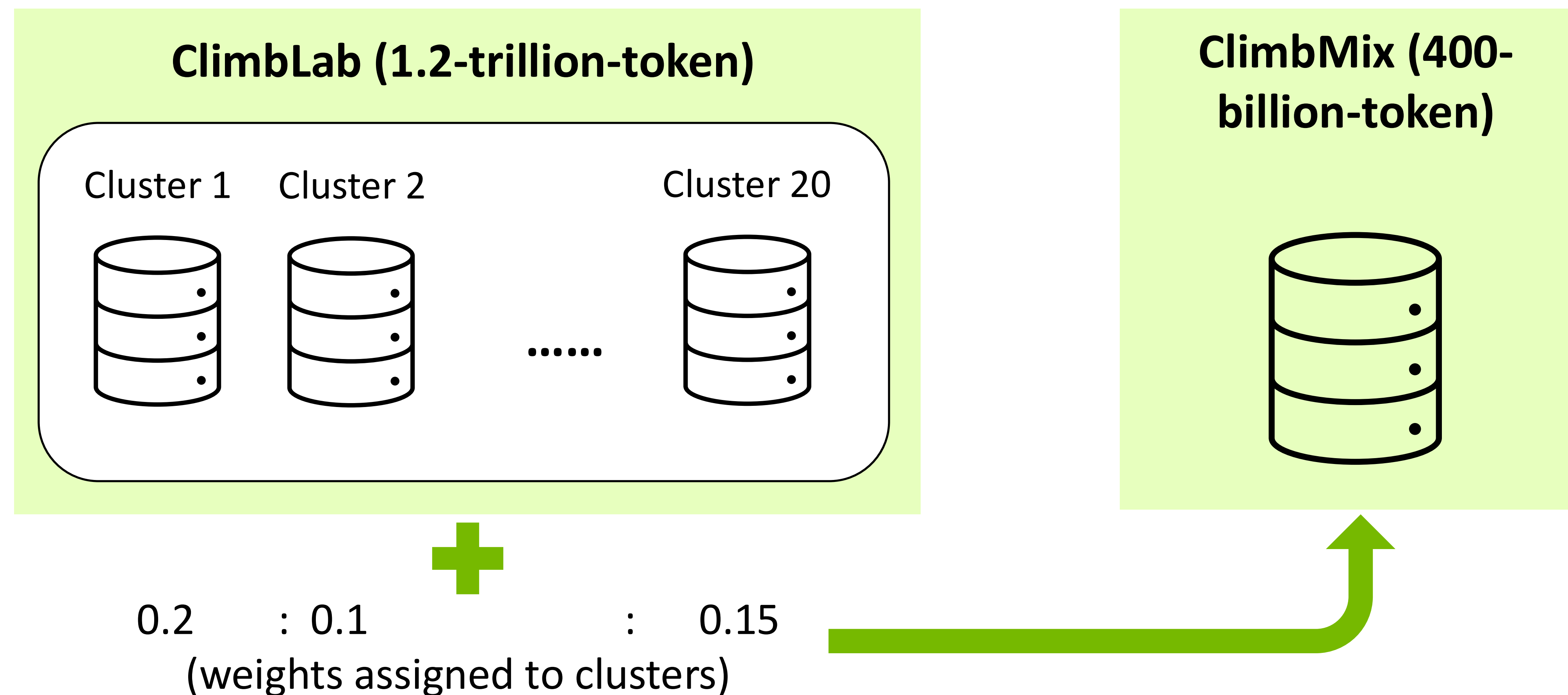
Setting	Model	Proxy	Comp.	piqa	arc_c	arc_e	hellaswag	winogrande	siqa	Avg.
Abl.comp	CLIMB	350M	100%	75.78	40.98	72.97	66.01	63.32	43.37	60.41
	CLIMB	350M	150%	76.23	41.28	73.16	66.41	63.53	43.71	60.72
	CLIMB	350M	200%	76.51	42.31	73.41	66.81	63.70	43.99	61.12
Abl.allo	CLIMB	350M	6:1	75.32	40.80	72.91	65.51	62.84	42.93	60.05
	CLIMB	350M	4:2:1	75.78	40.98	72.97	66.01	63.32	43.37	60.41
	CLIMB	350M	2:2:1:1	75.36	40.77	72.88	65.86	62.97	43.02	60.14
Abl.proxy	CLIMB	62M	100%	75.41	40.56	72.82	65.76	63.23	42.89	60.11
	CLIMB	132M	100%	75.56	40.93	72.94	65.57	63.09	43.07	60.19
	CLIMB	350M	100%	75.78	40.98	72.97	66.01	63.32	43.37	60.41
Abl.clus	48-21cluster	350M	100%	75.89	39.91	71.92	65.87	63.21	42.62	59.90
	64-21cluster	350M	100%	75.87	40.34	72.44	65.39	63.14	43.55	60.12
	100-21cluster	350M	100%	76.13	40.73	72.57	66.13	63.39	43.70	60.44
	1000-21cluster	350M	100%	75.78	40.98	72.97	66.01	63.32	43.37	60.41
	2000-21cluster	350M	100%	75.37	41.33	72.47	65.79	63.46	42.99	60.24
	1000-15cluster	350M	100%	75.94	41.33	73.34	66.28	63.62	43.05	60.59
	1000-30cluster	350M	100%	76.03	40.49	72.66	65.78	63.45	43.12	60.25
Abl.init	Random	350M	100%	75.42	40.12	72.47	65.73	64.27	43.22	60.21
	Dirichlet	350M	100%	75.78	40.98	72.97	66.01	63.32	43.37	60.41

New Datasets

Phase1 pre-training

Starting point: smollm (300B) + Nemotron-CC-HQ (500B) + Nemotron-CC-HQ-Synthetic (500B)

1. **ClimbLab**: a 1.3-Trillion-token corpus with 20 clusters. Based on Nemotron-CC and SmolLM-corpus, we apply CLIMB-clustering algorithm to cluster and filter data. This is a good playground for data mixing research.
2. **ClimbMix**: a 400-Billion-token corpus for efficient LM pre-training. Based on ClimbLab, we apply CLIMB-search algorithm to identify the optimal data mixture and create a compact yet powerful corpus.





Thank you!