

OpenSeek:Open Source Drives the Next Generation of AI Models

- Liu Guang
- BAAI Data Research Team

<https://github.com/FlagAI-Open/OpenSeek>

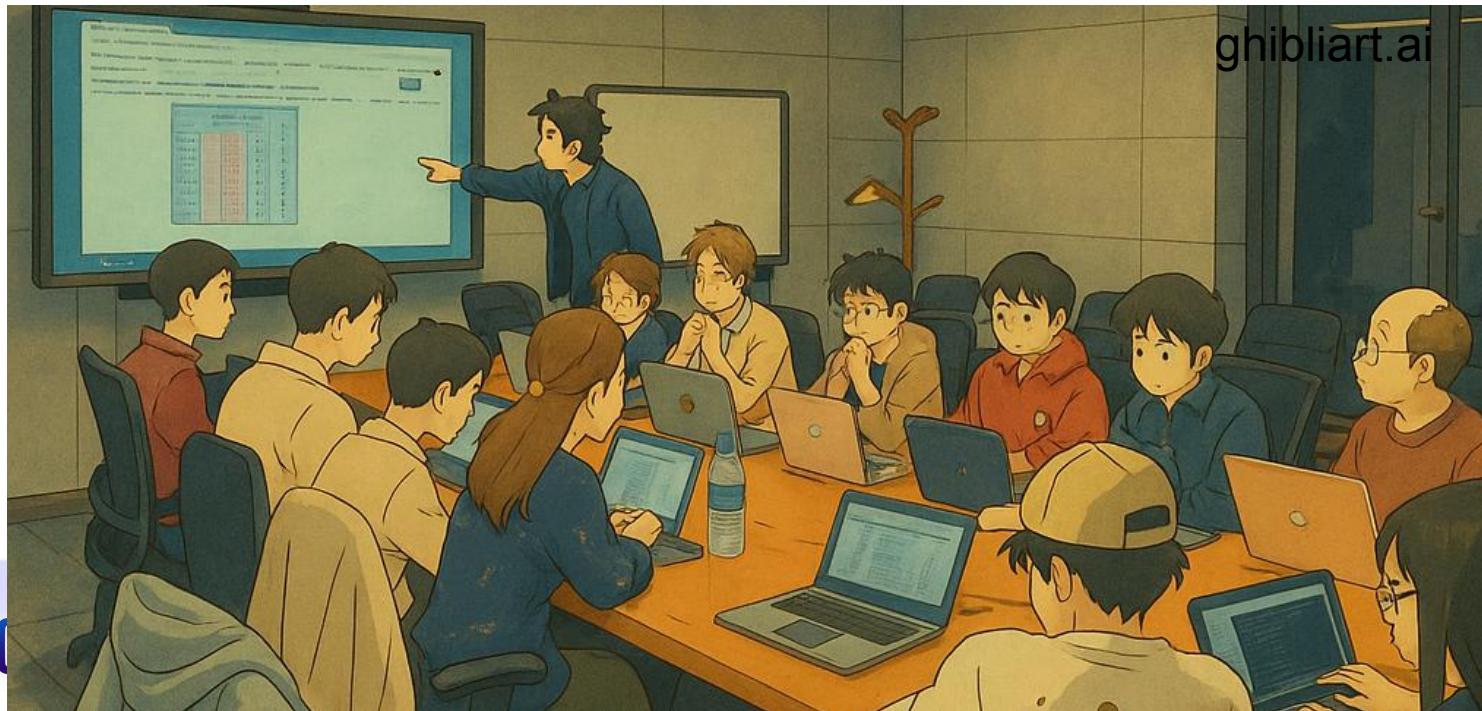
Working Groups

● Todo-Data 16 ...
This item hasn't been started

● OpenSeek #19
PreTraining Data Preprocess: Information Density - Entity & Keyword Coverage

● Todo-Infra 18 ...
● OpenSeek #23
Distributed Training System: Enhance the distributed training documentation

● Todo-Algorithm 1 ...
● OpenSeek #33 ...
Distributed Training System: Support for DeepSeek NAS or Kimi MoBA etc.



OpenSeek

System

Multi-Chip
Support for
Efficient
Training of
DeepSeek-
like model

Data

10TB-Level
Multilingual
High-Quality
Data (CCI
4.0)

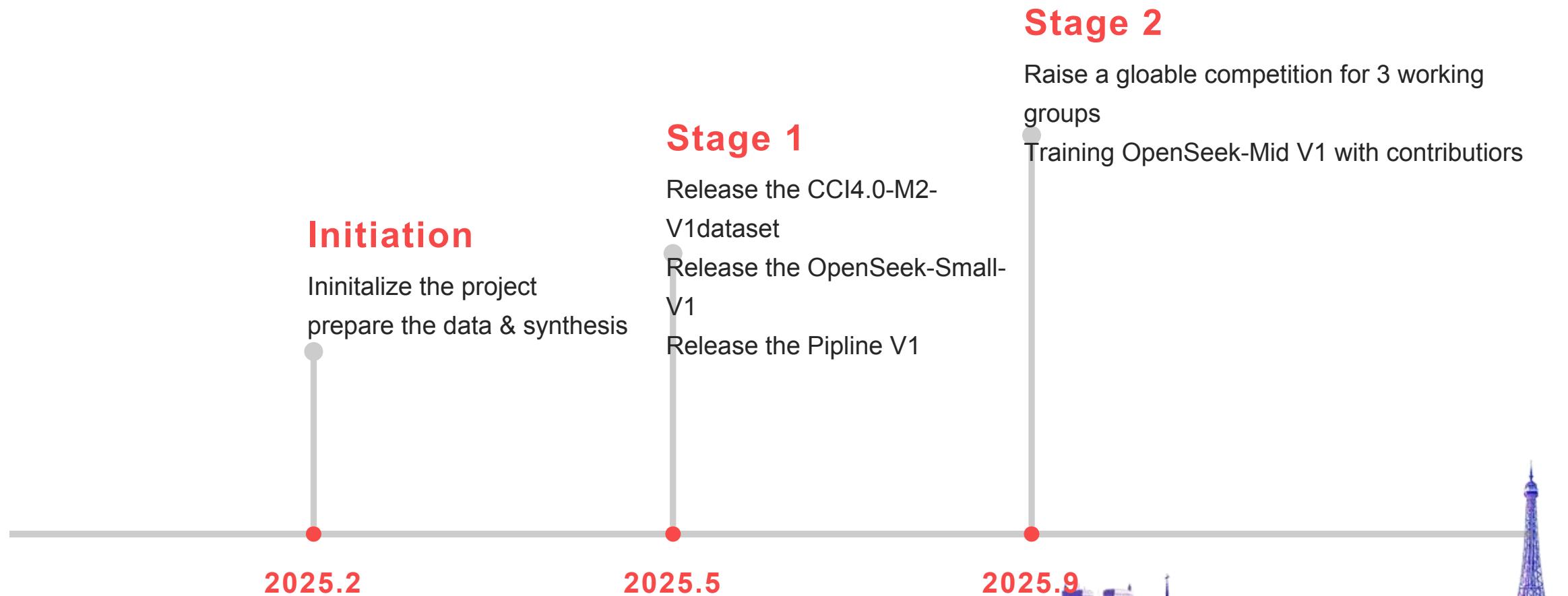
Algo

Improvements in
Data Mixture,
Model
Architecture,
Training
Algorithms



Time Line

From one organization to Community-Driven Open Source



BigScience

Research Workshop | May 2021 - May 2022

Timeline and Objectives

- Duration of 1 year
- Construction of the open-source multilingual model BLOOM and dataset ROOTS
- Inspired by the collaboration model of CERN and LHC

Organization and Collaboration

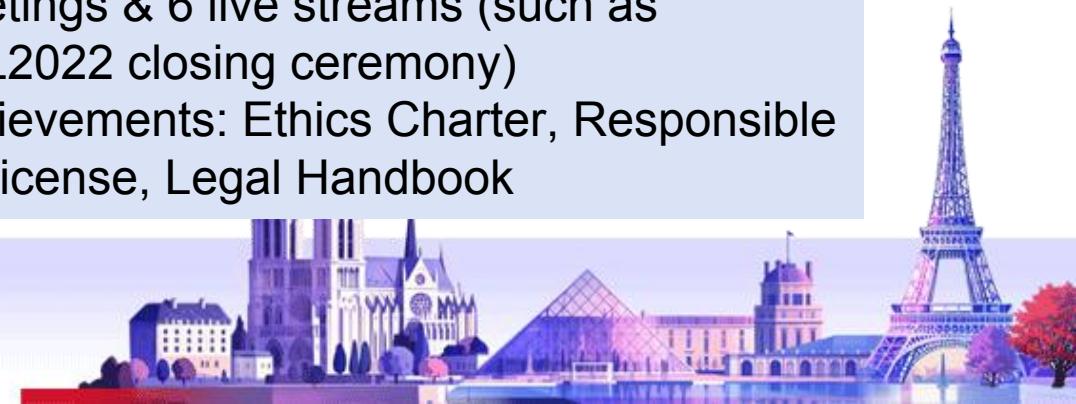
- Informal open collaboration, no legal entity, consensus-based decision-making
- All processes are public on Notion
- Divided into 30 working groups (assessment, data preparation, governance, ethics, etc.)

Initiation and Participation

- Initiating organizations: Hugging Face, GENCI, IDRIS
- Global participation: 60 countries, over 1000 researchers, more than 250 institutions

Achievements

- Communication: Regular biweekly meetings & 6 live streams (such as ACL2022 closing ceremony)
- Achievements: Ethics Charter, Responsible AI License, Legal Handbook



OpenSeek

System

Multi-Chip
Support for
Efficient
Training of
DeepSeek-
like model

Data

10TB-Level
Multilingual
High-Quality
Data (CCI
4.0)

Algo

Improvements in
Data Mixture,
Model
Architecture,
Training
Algorithms



Key Features of DeepSeek V3

- **MOE:** 1 shared expert, 256 routing experts (activating 8 each time), with the first three layers being dense structures
- **MLA:** Efficiently supports long sequences
- **MTP:** Multi-token prediction training objective, enhances downstream metrics
- **FP8 training:** ultimate targeted optimization
- **DualPipe** distributed training strategy

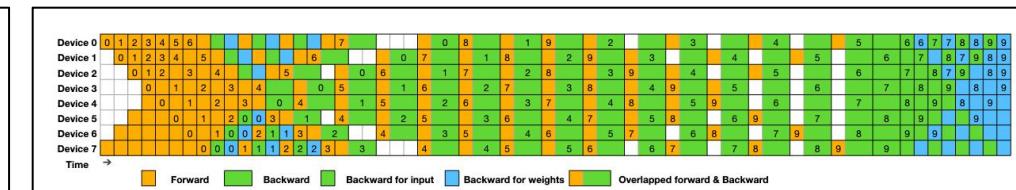
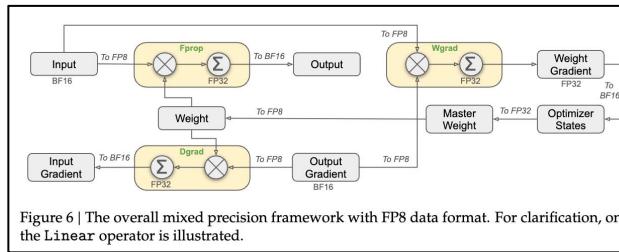
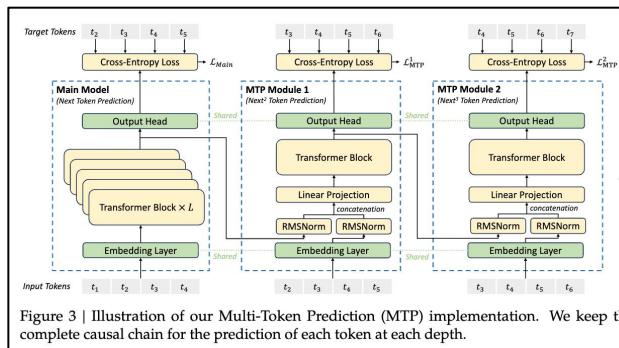
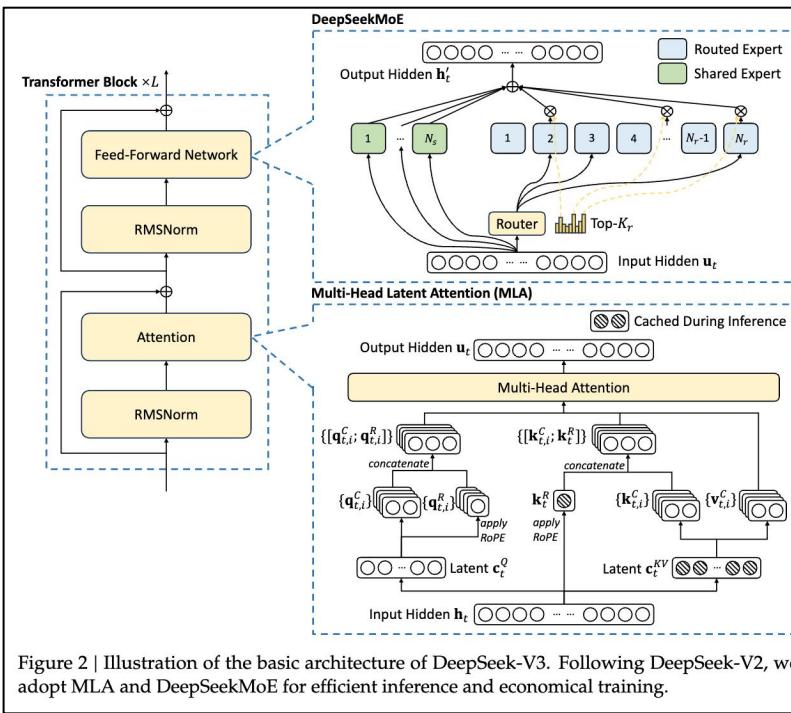


Figure 5 | Example DualPipe scheduling for 8 PP ranks and 20 micro-batches in two directions. The micro-batches in the reverse direction are symmetric to those in the forward direction, so we omit their batch ID for illustration simplicity. Two cells enclosed by a shared black border have mutually overlapped computation and communication.

Method	Bubble	Parameter	Activation
1F1B	$(PP - 1)(F + B)$	1x	PP
ZB1P	$(PP - 1)(F + B - 2W)$	1x	PP
DualPipe (Ours)	$(\frac{PP}{2} - 1)(F \& B + B - 3W)$	2x	PP + 1

Objectives & Tasks

With support from the open-source community, flagscale aims to reproduce DeepSeek V3 & R1's distributed training system, ensuring stable and performant end-to-end training.

DONE	IN PROCESS	TODO
<ul style="list-style-type: none">• Model Architecture Support<ul style="list-style-type: none">◦ MLA, DeepSeek MoE, MTP etc.• Huggingface Compatibility<ul style="list-style-type: none">◦ Conversion ckpt between FlagScale and Huggingface parameters• MoE Parallelism Optimization<ul style="list-style-type: none">◦ Performance analysis of the current DeepSeekMoE distributed training implementation• Multi-Chip Support<ul style="list-style-type: none">◦ Integration of the FlagGems Triton operator library and corresponding training accuracy validation• Distributed Training Process Display<ul style="list-style-type: none">◦ Process and display of records related to distributed training• Large-Scale Stability<ul style="list-style-type: none">◦ Develop tools for detecting slow nodes, faulty nodes, and NCCL errors in large-scale clusters◦ Implement a distributed log consolidation mechanism◦ Improve the monitoring system of distributed training	<ul style="list-style-type: none">• Usability Improvement<ul style="list-style-type: none">◦ Enhance the distributed training documentation◦ Improve the installation and usage• Pipeline Parallelism Optimization<ul style="list-style-type: none">◦ Support for DualPipe pipeline parallelism• Long Sequence Optimization<ul style="list-style-type: none">◦ Performance analysis of current long sequence handling◦ Support for DeepSeek NAS or Kimi MoBA etc.• Distributed Reinforcement Learning<ul style="list-style-type: none">◦ Research and design a solution can be easily implemented in FlagScale◦ Implementation of a distributed reinforcement learning system to support efficient DeepSeek R1	<ul style="list-style-type: none">• Pipeline Parallelism Optimization<ul style="list-style-type: none">◦ Achieve more efficient pipeline parallelism• MoE Parallelism Optimization<ul style="list-style-type: none">◦ Improve communication algorithms to achieve more efficient MoE parallelism optimization• Long Sequence Optimization<ul style="list-style-type: none">◦ Collaborate with algorithm teams to achieve more efficient long sequence optimization• Multi-Chip Support<ul style="list-style-type: none">◦ Implement customized parallel and optimization strategies for specific chips• Large-Scale Stability<ul style="list-style-type: none">◦ Visualization of the communication flows and scheduling relationships in complex large-scale distributed clusters• FP8 Training<ul style="list-style-type: none">◦ Implementation of the FP8 operators required in DeepSeek V3, with support for validation during the training process◦ Implement more innovative FP8 training solutions

OpenSeek

System

Multi-Chip
Support for
Efficient
Training of
DeepSeek-
like model

Data

10TB-Level
Multilingual
High-Quality
Data (CCI
4.0)

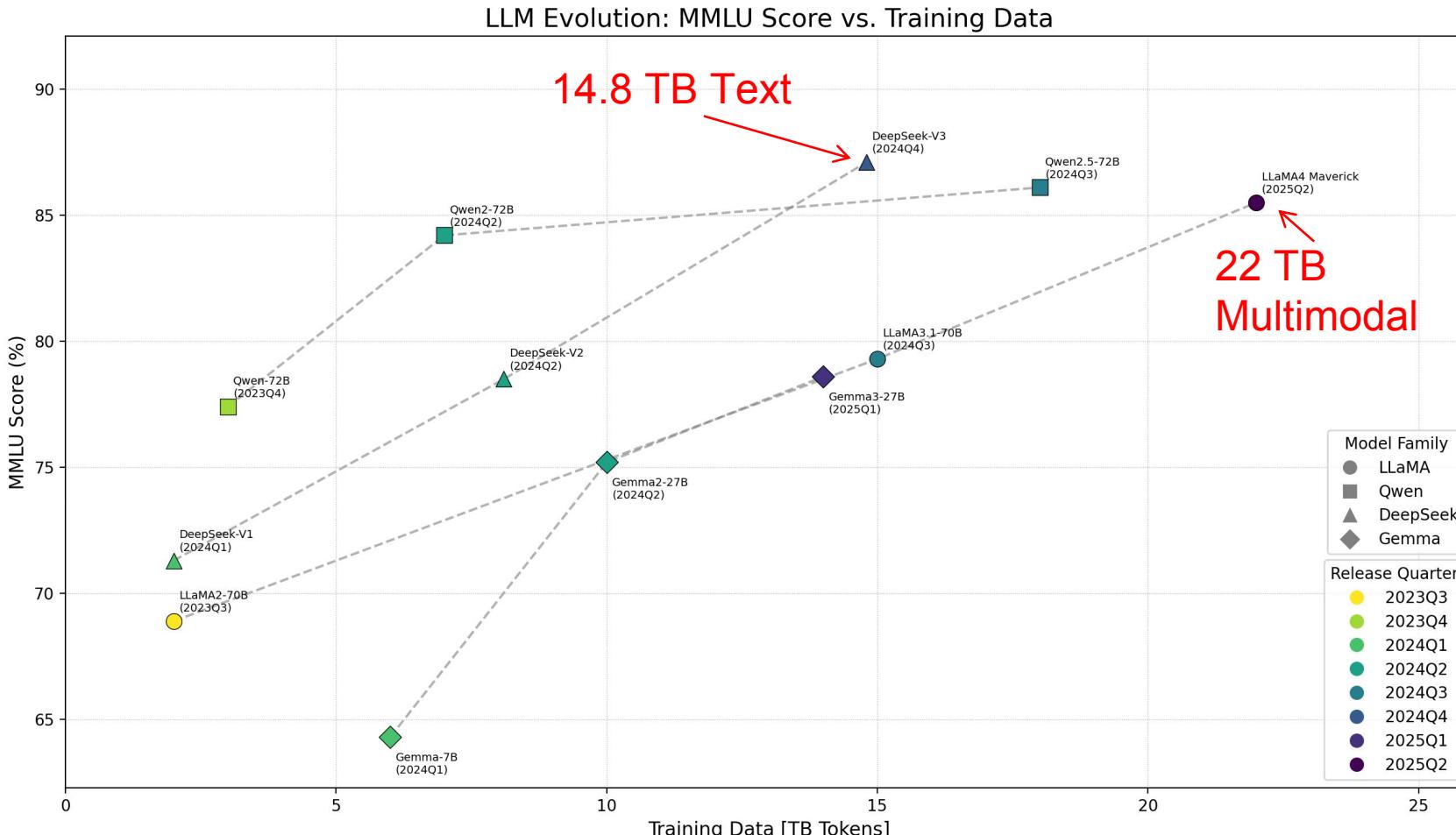
Algo

Improvements in
Data Mixture,
Model
Architecture,
Training
Algorithms

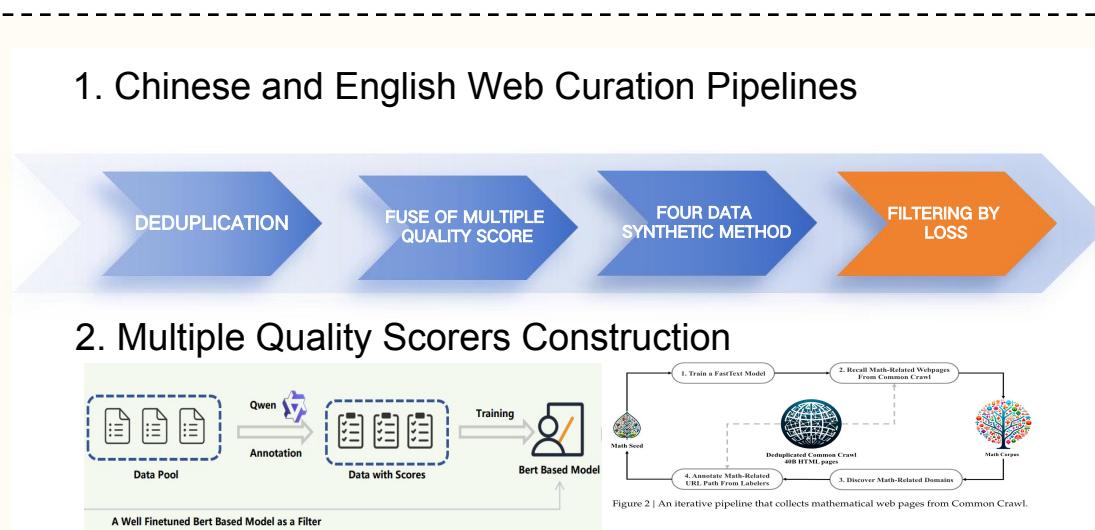


Pretraining Data Scale

- Model performance and data volume show non-linear growth, but there are "highly efficient" models.

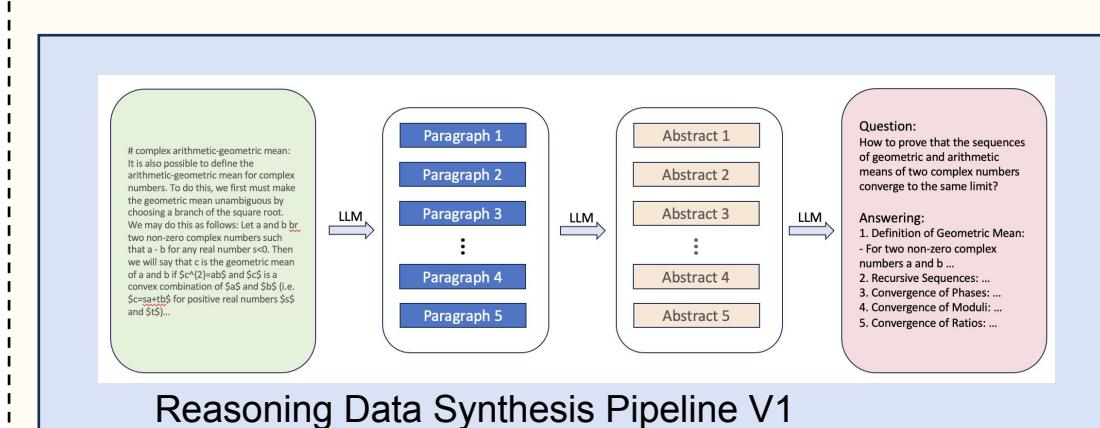


CCI4.0-M2-Base V1

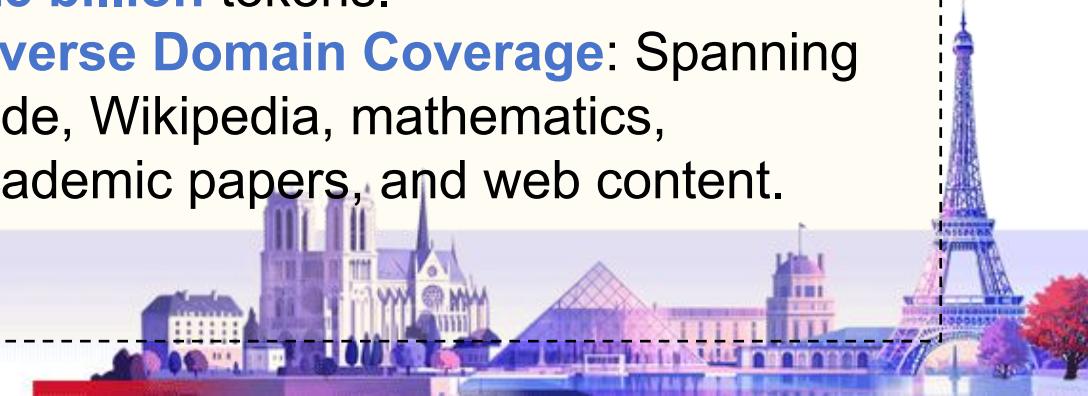


- **Large-Scale Pretrain Dataset: 10 billion** documents encompassing 6400 billion tokens.
- **Multiple Quality Scorers:** Three customized classifiers for Chinese and English
- **Multiple Tagging and Bucketing:** Domain Tagging and PPL Bucketing

CCI4.0-M2-CoT V1

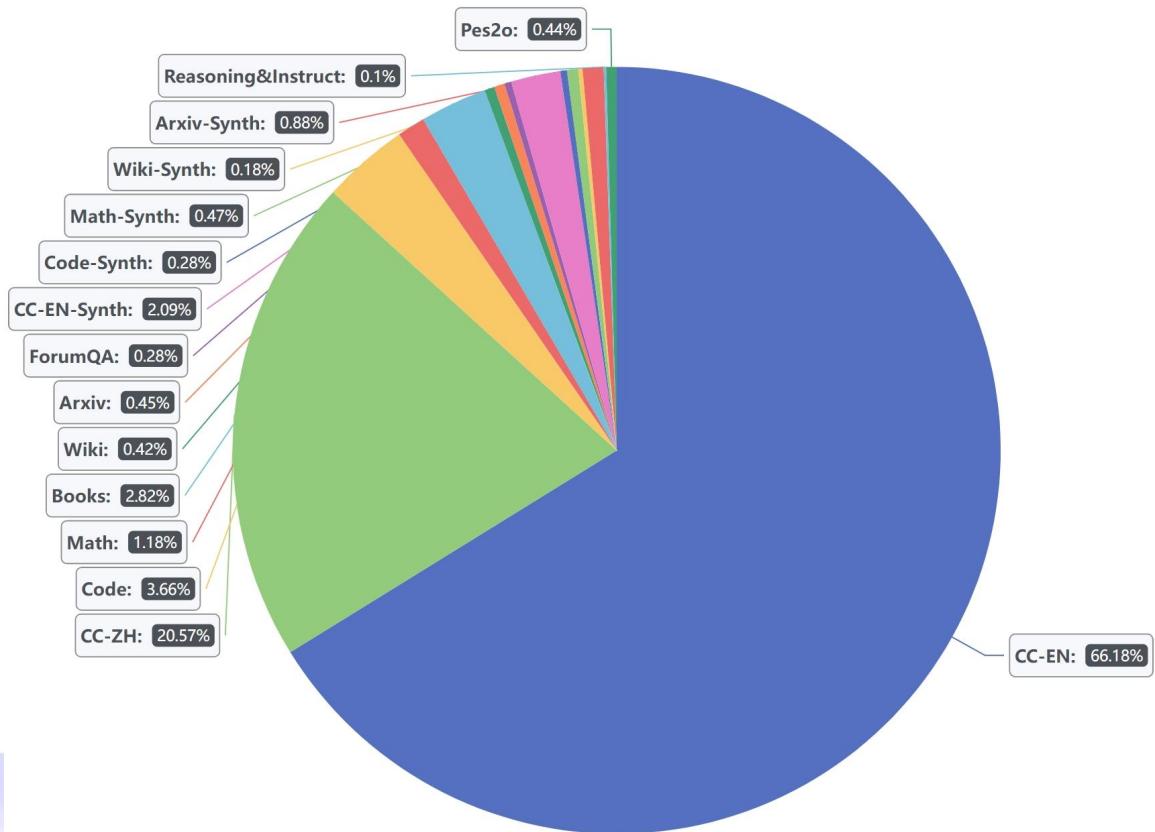


- **Large-Scale Synthesis: 430 million** samples encompassing approximately **425 billion** tokens.
- **Diverse Domain Coverage:** Spanning code, Wikipedia, mathematics, academic papers, and web content.



Overall of CCI4.0-M2 V1

- Total 6.4TB tokens
 - EN Webpage~66.17%, ZH Webpage~20.57%, CoT ~ 4%, other (e.g., code, book and wiki) ~ 9%



Domain Name	Tokens(B)	Ratio(%)
CC-EN	4246.08	66.17
CC-ZH	1320.19	20.57
Code	234.94	3.66
Math	75.72	1.18
Books	180.98	2.82
Wiki	26.74	0.42
Arxiv	28.81	0.45
ForumQA	17.78	0.28
CC-EN-Synth	134.31	2.09
Code-Synth	18.11	0.28
Math-Synth	29.99	0.47
Wiki-Synth	11.32	0.18
Arxiv-Synth	56.78	0.88
Reasoning&Instruct	6.35	0.10
Pes20	28.40	0.44
Total	6416.49	100.00

EN Webpage



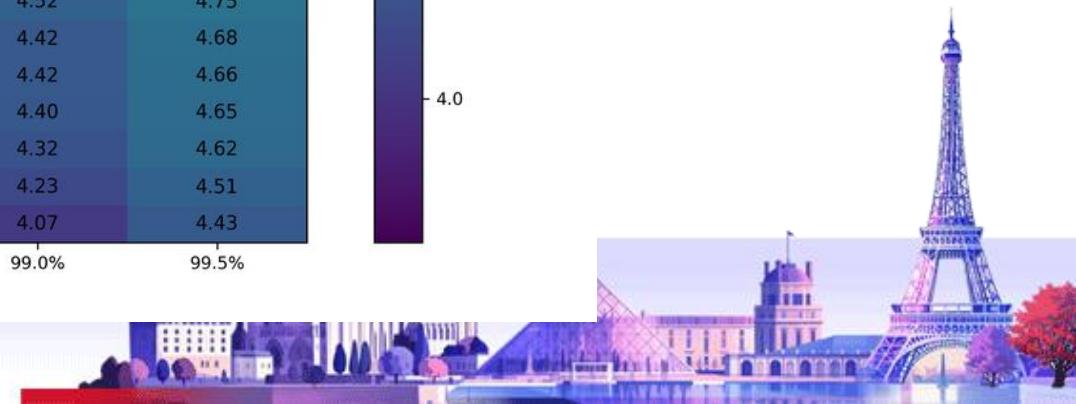
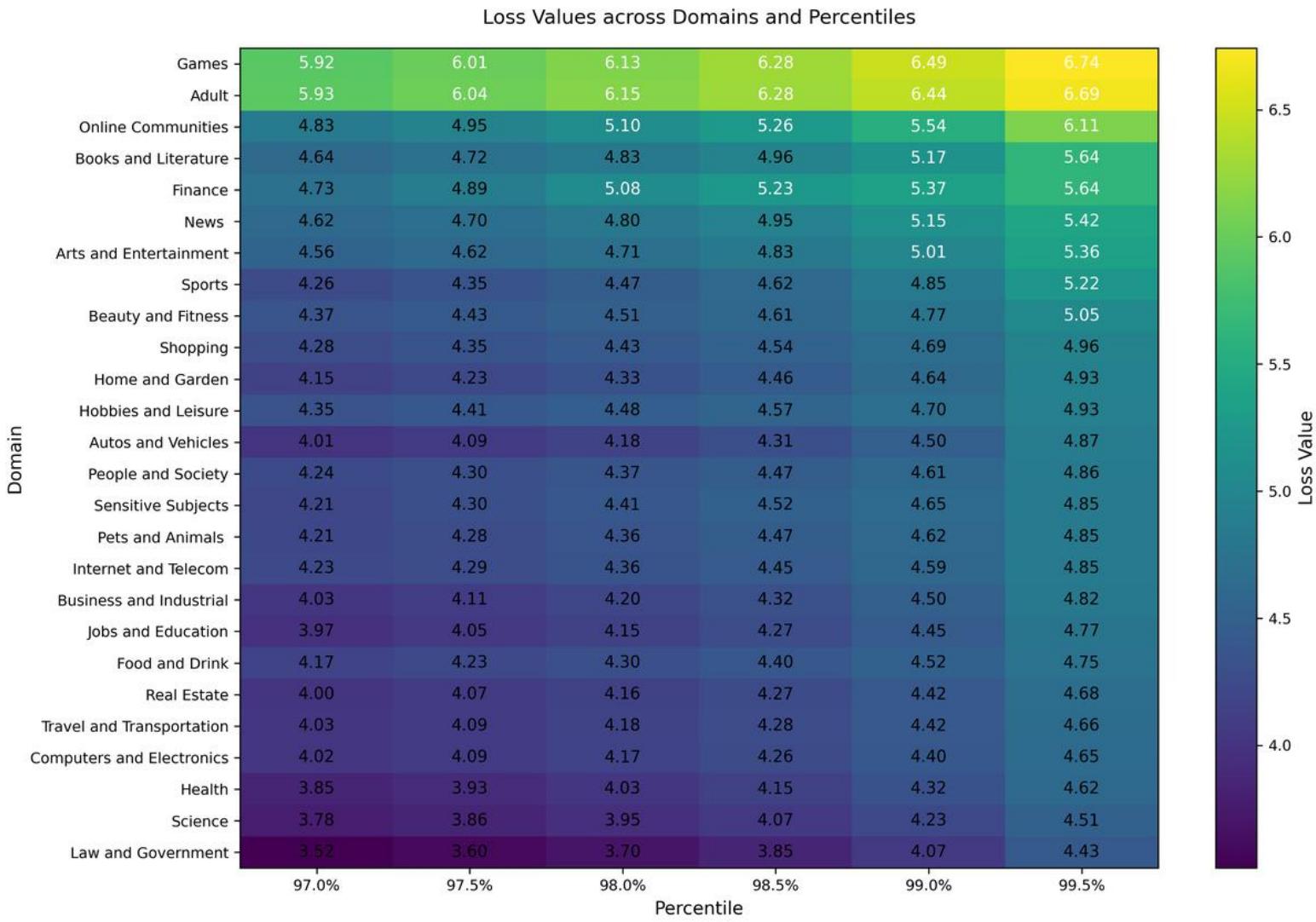
1. Calculate sample loss using the Qwen2.5-3B model by domain.
2. Remove samples above the 99.5th percentile by domain.

1. Diverse Question-Answer (QA) pairs
2. Distill: rewrite the text into a concise and clear passag
3. Extract knowledge: rewrite knowledge from the text and disregard uninformative content;
4. Knowledge list: extract key information from the text as an organized list.

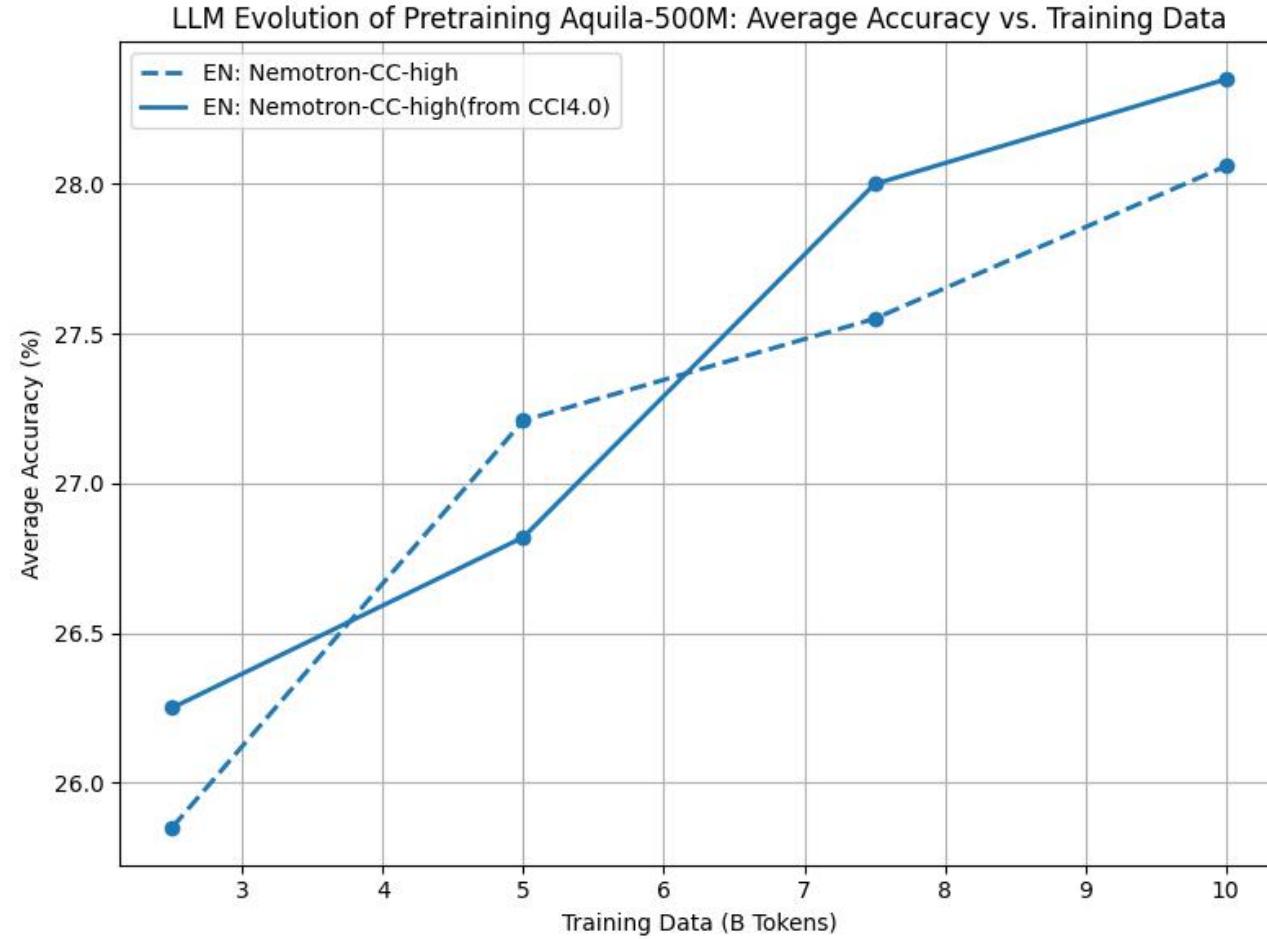


BAI
智源研究院

LLM loss on different domains GOSIM



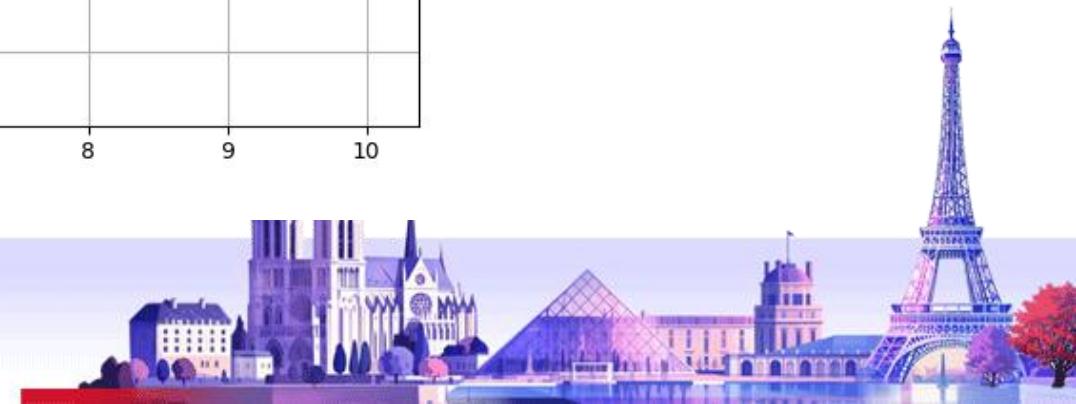
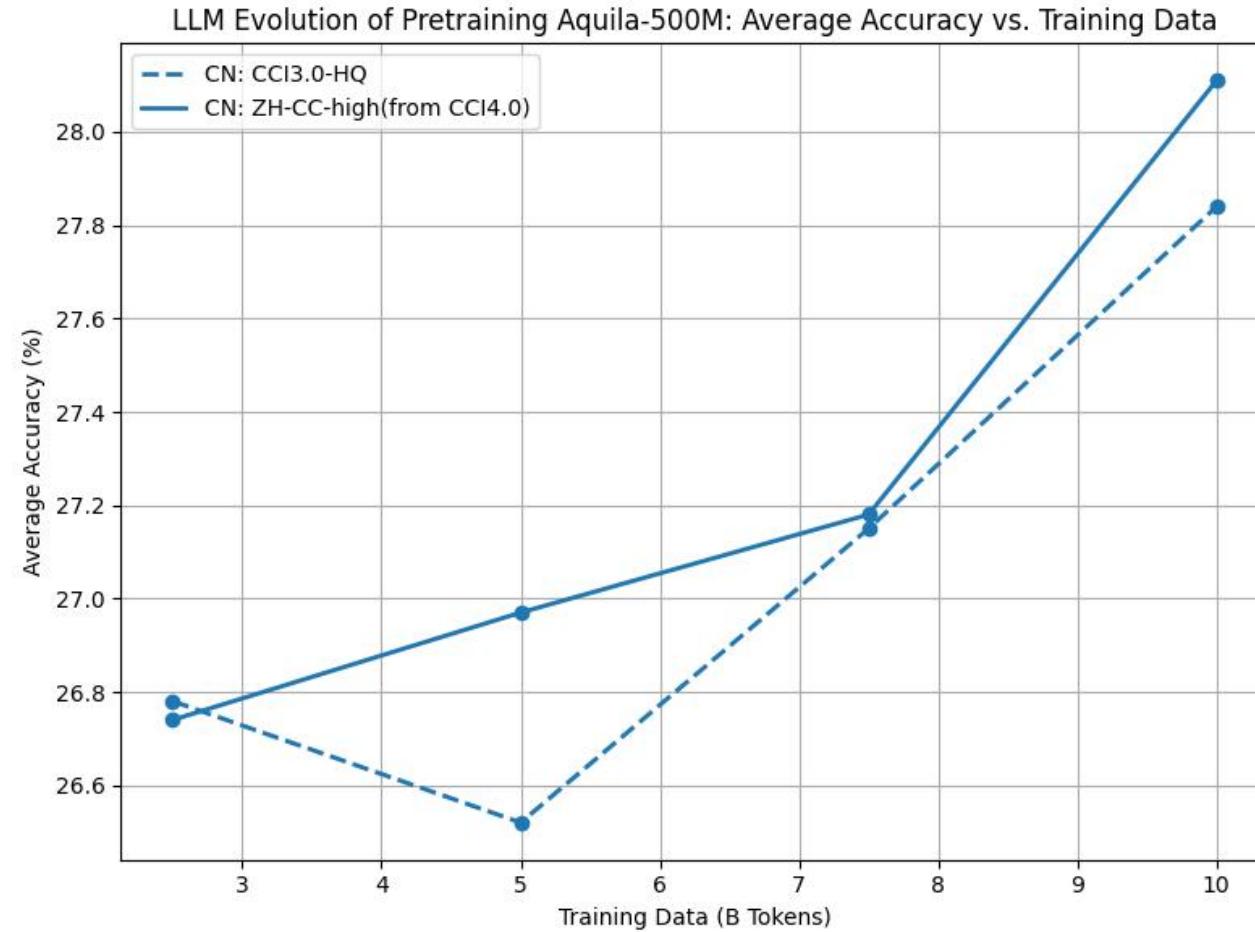
Experimental Results



ZH Webpage



Experimental Results

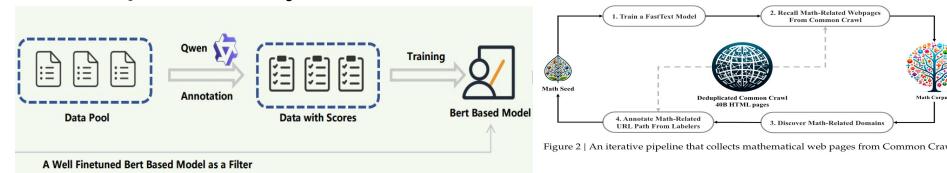


CCI4.0-M2-Base V1

1. Chinese and English Web Curation Pipelines

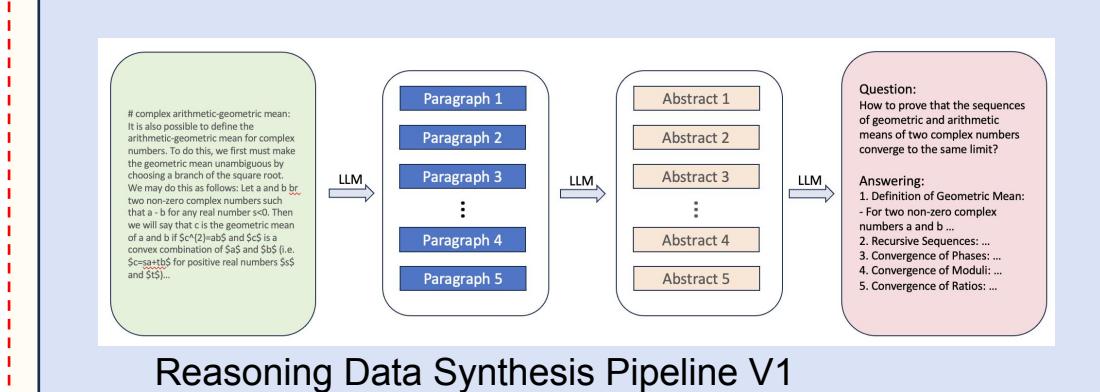


2. Multiple Quality Scorers Construction

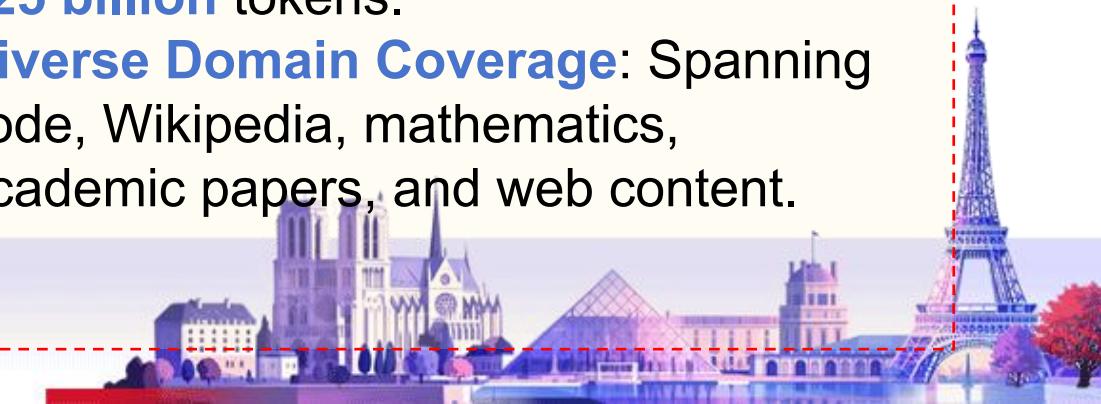


- **Large-Scale Pretrain Dataset: 10 billion** documents encompassing 6400 billion tokens.
- **Multiple Quality Scorers:** Three customized classifiers for Chinese and English
- **Multiple Tagging and Bucketing:** Domain Tagging and PPL Bucketing

CCI4.0-M2-CoT V1



- **Large-Scale Synthesis: 430 million** samples encompassing approximately **425 billion** tokens.
- **Diverse Domain Coverage:** Spanning code, Wikipedia, mathematics, academic papers, and web content.



Motivation of CCI4.0-M2-CoT

- The reasoning ability of LLMs originates from the **pre-training stage**, activated by RL training^{1, 2, 3}.
- Massive raw corpus containing complex human reasoning process, but lack **of generalized and effective synthesis method** to extract these reasoning process.

¹Cognitive Behaviors that Enable Self-Improving Reasoners

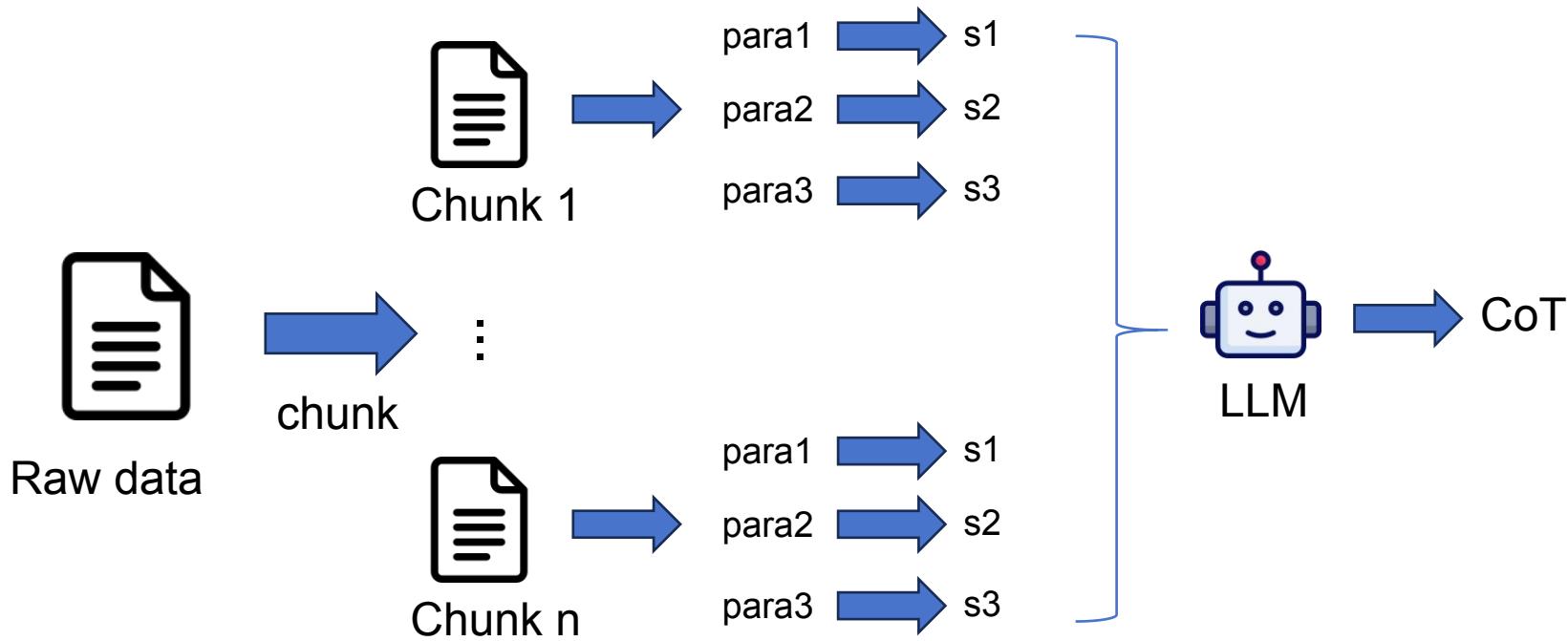
²Does Reinforcement Learning Really Incentivize Reasoning Beyond the Base Model

³Rethinking Reflection in Pre-training



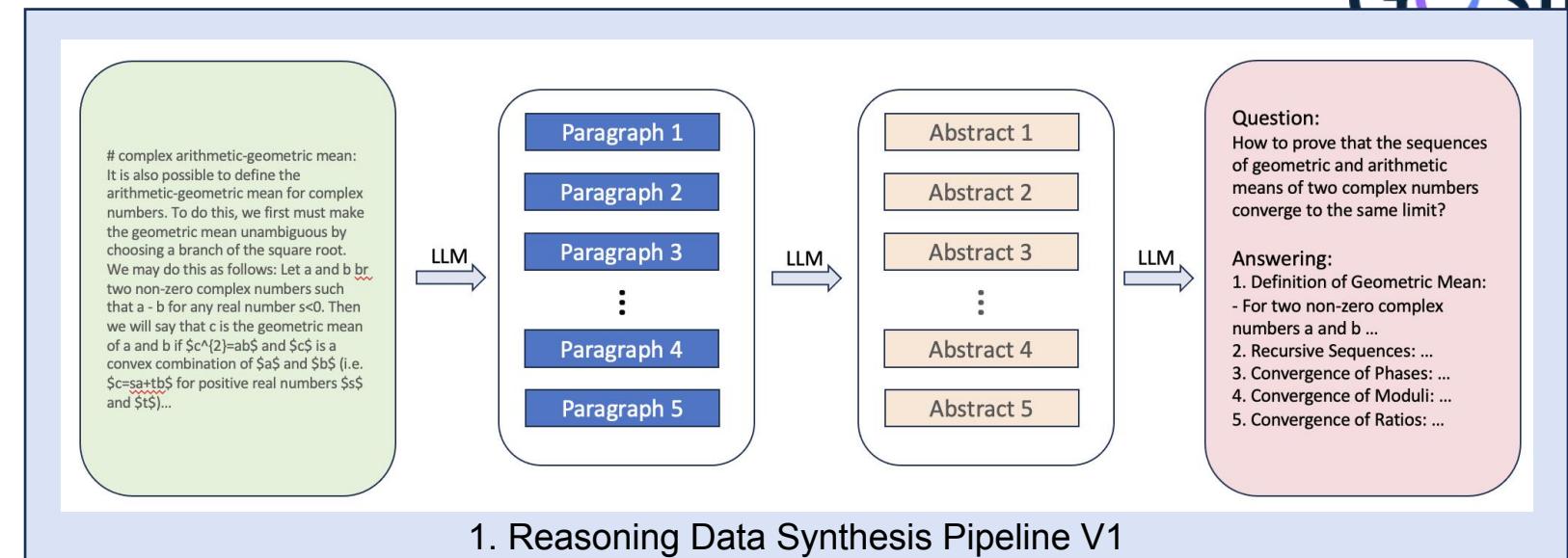
CoT Template extraction

- Object: extract reasoning template from raw data



```
{"id": 1, "conversations": [{"from": "human", "value": "Generate a passage that solves the system of equations y = -1/3x + 2 and x + 3y = 3, and determine the correct solution among the given options or identify if there is no solution. The passage should follow a logical sequence of steps to arrive at the final result."}, {"from": "gpt", "value": "<think>Step 1: Introduce the System of Equations\n- Start by introducing the system of equations: y = -1/3x + 2 and x + 3y = 3.\n\nStep 2: Correct the First Equation\n- Mention that the first equation is correctly stated as y = -1/3x + 2.\n\nStep 3: Transform the Second Equation\n- Suggest transforming the second equation by subtracting x from both sides, resulting in 3y = -x + 3.\n\nStep 4: Solve the System Using Substitution\n- Explain that one can solve the system by substituting y = -1/3x + 2 into the second equation, leading to the equation x + 3(-1/3x + 2) = 3.\n- Solve for x, resulting in x = 0.\n- Substitute x = 0 back into the first equation to find y = 2.\n\nStep 5: Verify the Solution\n- Note that the solution (x = 0, y = 2) does not match the given answer options.\n- Confirm that the correct answer based on the given options is (0, 1), despite the confusion over the calculations.\n\nStep 6: Conclude on the System's Solution\n- Conclude that the system of equations has no solution, as the lines are parallel, and the given answer choices are incorrect.\n\nStep 7: Suggest Further Resources\n- Suggest finding more explanations on OpenStudy for further clarification./<think>\n## anonymous 3 years ago solve the system: y= -1/3 + 2 and x + 3y = 3\n1. anonymous\n-1/3x+2=n1\n2. anonymous\nnx+3y=3 substrct x from both sides what would you get?\nn3. anonymous\nn3y=-x+3?\nn4. anonymous\nnid set them both equal to y then solve like that\nn5. anonymous\nnar you can sub y=-1/3x+2 in to other equation x+3(-1/3 x+x)=3 solve for x\nn6. anonymous"}, {"from": "human", "value": "What is the final answer?"}, {"from": "gpt", "value": "The final answer is (0, 1)."}]
```

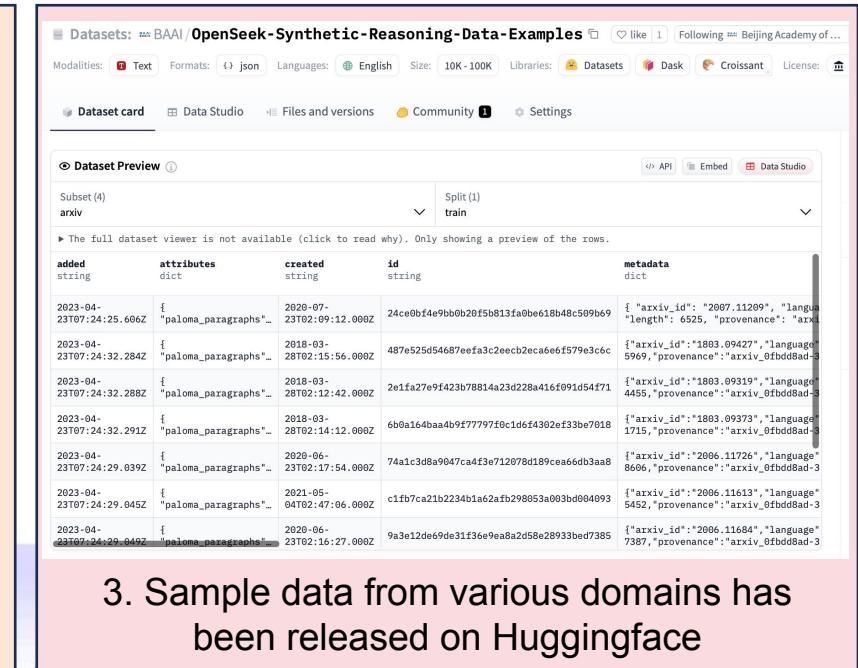
- We constructed the first version of a pipeline for synthesizing reasoning data from raw documents.
- We synthesized **430 million rows**, about **425B tokens** of synthetic data. Contains domains such as code, wiki, math, papers, web, etc.
- We released the first version of the sample synthetic reasoning data on the Huggingface.



1. Reasoning Data Synthesis Pipeline V1

类型	状态	行数(*预估)	硬盘大小(G)	tokens(B)(*预估)
Math	已完成	7,800,000	129	13.47
Code	已完成	2,992,343	11	2.64
Arxiv	已完成	1,529,340	152	29.40
Math	已完成	5,749,146	141	15.73
Math	已完成	18,025,000	135	12.76
Wiki	已完成	4,326,000	57	4.67
Code	已完成	1,839,000	32	1.76
Code	已完成	4,007,500	9.7	2.38
Code	已完成	9,045,500	35	8.41
Wiki	已完成	9,674,640	135	11.73
Code	已完成	10,195,600	74	17.20
CC-EN	已完成	353,329,000	3700	276.00
Arxiv	已完成	1,525,500	128	29.59
		430,038,569	4738.7	425.74

2. Synthesized 430 million rows of data



Datasets: BAAI/OpenSeek-Synthetic-Reasoning-Data-Examples

Modalities: Text Formats: json Languages: English Size: 10K-100K Libraries: Datasets Dask Croissant License: CC-BY-NC-ND

Dataset card Data Studio Files and versions Community Settings

Subset (4) arxiv Split (1) train

The full dataset viewer is not available (click to read why). Only showing a preview of the rows.

added	attributes	created	id	metadata
2023-04-23T07:24:25.606Z	{"paloma_paragraphs": ...}	2020-07-23T02:09:12.000Z	24ce0bf1e9bb0b20f5b813fa0be618b48c509b69	{"arxiv_id": "2007.11209", "language": "en", "length": 6525, "provenance": "arxiv_0fbdd8ad-3"}
2023-04-23T07:24:32.284Z	{"paloma_paragraphs": ...}	2018-03-28T02:15:56.000Z	487e525d54687eefa3czech2ceca6e6f579e3c6c	{"arxiv_id": "1803.09427", "language": "en", "length": 5969, "provenance": "arxiv_0fbdd8ad-3"}
2023-04-23T07:24:32.288Z	{"paloma_paragraphs": ...}	2018-03-28T02:12:42.000Z	2eifa27e9f1423b78814a23d228a416f091d547f1	{"arxiv_id": "1803.09319", "language": "en", "length": 4455, "provenance": "arxiv_0fbdd8ad-3"}
2023-04-23T07:24:32.291Z	{"paloma_paragraphs": ...}	2018-03-28T02:14:12.000Z	6b0a164baa4bf7797f0c1d6f4302ef33be7018	{"arxiv_id": "1803.09373", "language": "en", "length": 1715, "provenance": "arxiv_0fbdd8ad-3"}
2023-04-23T07:24:29.039Z	{"paloma_paragraphs": ...}	2020-06-23T02:17:54.000Z	74a1c3d8a9047ca4f3e712078d189cea6db3aa8	{"arxiv_id": "2006.11726", "language": "en", "length": 8606, "provenance": "arxiv_0fbdd8ad-3"}
2023-04-23T07:24:29.045Z	{"paloma_paragraphs": ...}	2021-05-04T02:47:06.000Z	c1fb7ca21b2234ba162af2fb298053a003bd004093	{"arxiv_id": "2006.11613", "language": "en", "length": 5452, "provenance": "arxiv_0fbdd8ad-3"}
2023-04-23T07:24:29.049Z	{"paloma_paragraphs": ...}	2020-06-23T02:16:27.000Z	9a3e12de69de31f36e9ea8a2d58e28933bed7385	{"arxiv_id": "2006.11684", "language": "en", "length": 7387, "provenance": "arxiv_0fbdd8ad-3"}

3. Sample data from various domains has been released on Huggingface



One-Stage Evaluation: SFT GOSIM

	Math-500	AIME-2024	Omni-Math
Llama3.1-8B	4.8	0.0	2.8
FineMath	5.1	0.0	4.6
FineMath + Q	12.6	0.0	11.7
FineMath + Q + CoT	19.2	6.7	16.4

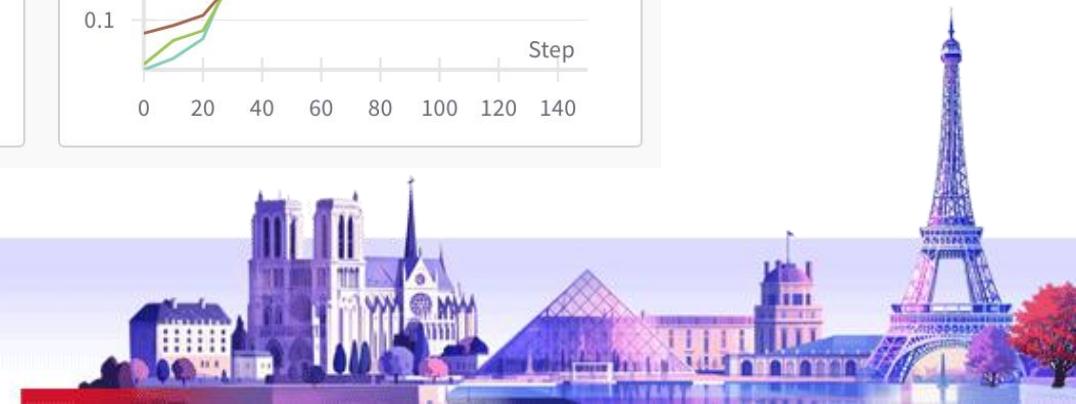
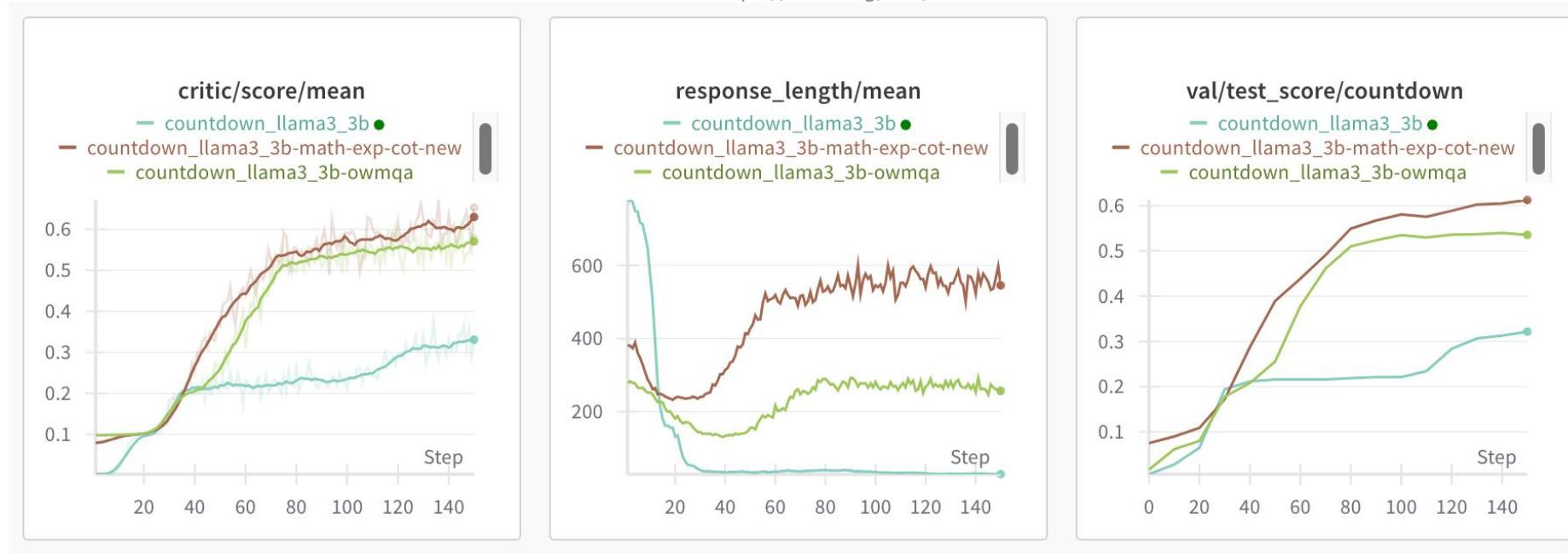


Two-Stage Evaluation: SFT+GRPO

- Llama often failed in long CoT generation
- 100k random sampled data from CCI4.0 outperform generated data

Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs

<https://arxiv.org/abs/2503.01307>



OpenSeek

System

Multi-Chip
Support for
Efficient
Training of
DeepSeek-
like model

Data

10TB-Level
Multilingual
High-Quality
Data (CCI
4.0)

Algo

Improvements in
Data Mixture,
Model
Architecture,
Training
Algorithms



Experimental settings

- Phi4-like
- webpage 1:1 | knowledge 1:1.5 | rewrite 1:4 | synthesis 1:10

DataSet Type	Tokens(B)	Weight(%)	Epochs	Tokens(B)	Weight(%)
CC-EN	172.4625562	2.69	1	172.4625562	1.26
CC-EN	301.6132778	4.70	1	301.6132778	2.20
CC-EN	22.97631421	0.36	1	22.97631421	0.17
CC-EN-Rewrite	22.36849674	0.35	4	89.47398696	0.65
CC-EN-Rewrite	64.34685559	1.00	4	257.3874224	1.88
CC-ZH	1320.19	20.57	1	1320.19	9.65
Code	234.941	3.66	2	469.882	3.43
Math	75.7196	1.18	1.5	113.5794	0.83
Books	180.978	2.82	1.5	271.467	1.98
Wiki	26.739	0.42	1.5	40.1085	0.29
Arxiv	28.8105	0.45	1.5	43.21575	0.32
ForumQA	17.7777	0.28	1.5	26.66655	0.19
CC-EN-Synth	134.311	2.09	10	1343.11	9.82
Code-Synth	18.1101	0.28	10	181.101	1.32
Math-Synth	29.9877	0.47	10	299.877	2.19
Wiki-Synth	11.3165	0.18	10	113.165	0.83
Arxiv-Synth	56.7782	0.88	10	567.782	4.15
Reasoning&Instruct	6.34884	0.10	10	63.4884	0.46
Pes2o	28.40089493	0.44	1.5	42.60134239	0.31

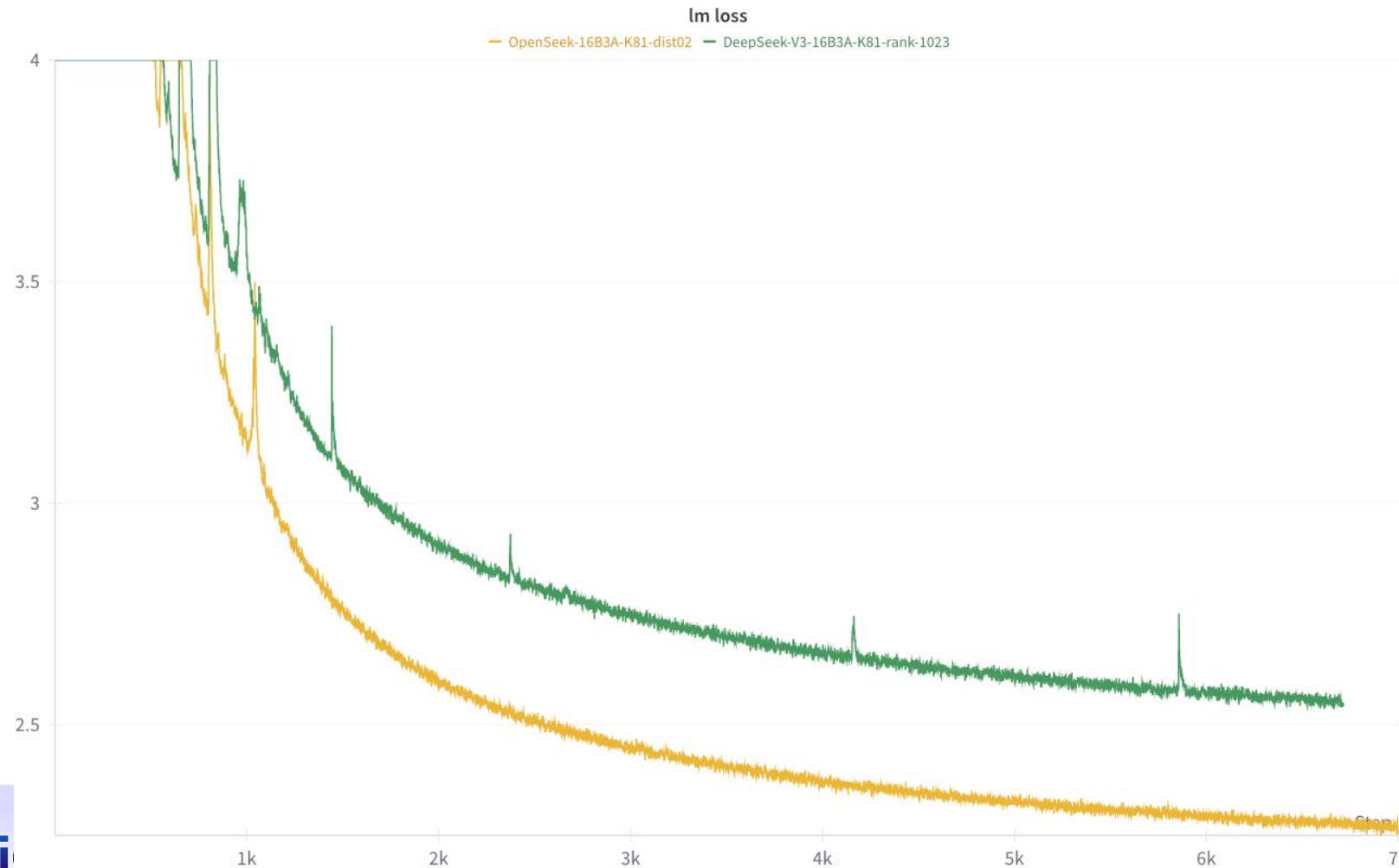
Data Source	Fraction of Training	Unique Token Count	Number of Epochs
Web	15%	1.3T	1.2
Web rewrites	15%	290B	5.2
Synthetic	40%	290B	13.8
Code data	20%	820B	2.4
Acquired sources	10%	580B	1.7

Table 5: Data mixture for pretraining.



Loss Curve

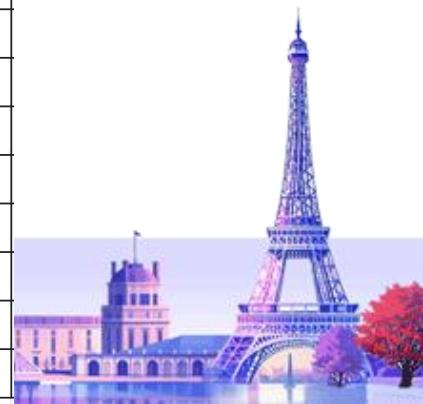
- Green: w/o upsampling
- Yellow: Phi4-like



Experimental results

- webpage 1:1 | knowledge 1:1.5 | rewrite 1:4 | synthesis 1:10
- AVG 39.05->39.84(+2.02%)

备注	王良栋 5226	w/o upsampling	Phi4-like		
Category	Metrics (shots)	Aquila-16B-A3B-K81-NoMTP-100B	Aquila-16B-A3B-K81-NoMTP-200B	Aquila-16B-A3B-K81-NoMTP-dist02-100B	Aquila-16B-A3B-K81-NoMTP-dist02-200B
English-Commonsense Reasoning	TruthfulQA (0-shot)	0.4114	0.4214	0.4167	0.4235
	Winogrande (5-shot)	0.5975	0.6322	0.6235	0.6093
	CommonsenseQA (5-shot)	0.2023	0.1925	0.2023	0.2121
	PIQA (5-shot)	0.7612	0.7797	0.7524	0.7699
	OpenBookQA (5-shot)	0.2640	0.2920	0.3020	0.3040
	BoolQ (5-shot)	0.5615	0.6740	0.6743	0.6743
English-Problem-Solving	ARC Easy (5-shot)	0.7222	0.7437	0.7218	0.7513
	ARC Challenge (5-shot)	0.3703	0.3891	0.3916	0.4224
	MMLU (5-shot)	0.2707	0.2709	0.2599	0.2681
English-Mathematics	GSM8K (5-shot)	0.0349	0.0705	0.0644	0.1281
	Minerva Math (4-shot)	0.0118	0.0168	0.0222	0.0262
Chinese	CEval (5-shot)	0.2481	0.2355	0.2474	0.2489
	CMMLU (5-shot)	0.2469	0.2481	0.2520	0.2491
Average Metrics	Average-English(w/o Math)	0.4633	0.4896	0.4812	0.4925
	Average-English	0.3900	0.4153	0.3959	0.4233
	Average-Chinese	0.2475	0.2418	0.2520	0.2490
	Average	0.3696	0.3905	0.3856	0.3984
	Average(w/o Math)	0.4274	0.4483	0.4426	0.4519



Experimental results

- Settings: std=6e-3|router dtype=fp32|aux loss=0.001|mtp_coeff=0.3
- AVG 37.83->43.94(+16.1%)

Category	Metrics (shots)	before			after		
		Aquila-16B-A3B-K77-NoMTP-100B	Aquila-16B-A3B-K77-NoMTP-145B	Aquila-16B-A3B-K77-NoMTP-200B	Aquila-16B-A3B-K77-100B	Aquila-16B-A3B-K77-145B	Aquila-16B-A3B-K77-200B
English-Commonsense Reasoning	HellaSwag (5-shot)	0.4414	0.4588	0.4643	0.4568	0.4751	0.4864
	TruthfulQA (0-shot)	0.3735	0.3849	0.3941	0.3438	0.3660	0.3821
	Winogrande (5-shot)	0.5927	0.5959	0.6038	0.6062	0.6117	0.6219
	CommonsenseQA (5-shot)	0.2056	0.2039	0.1998	0.2531	0.4283	0.4988
	PIQA (5-shot)	0.7274	0.7432	0.7535	0.7454	0.7486	0.7503
	OpenBookQA (5-shot)	0.2760	0.2720	0.2980	0.3180	0.3160	0.3200
	BoolQ (5-shot)	0.6294	0.6768	0.6670	0.6471	0.5596	0.6826
English-Problem-Solving	ARC Easy (5-shot)	0.7029	0.7096	0.7294	0.7264	0.7483	0.7496
	ARC Challenge (5-shot)	0.3703	0.3737	0.3959	0.4053	0.4044	0.4198
	MMLU (5-shot)	0.2602	0.2817	0.2694	0.3397	0.3839	0.4053
English-Mathematics	GSM8K (5-shot)	0.0182	0.0220	0.0190	0.0136	0.0303	0.0318
	Minerva Math (4-shot)	0.0094	0.0098	0.0148	0.0080	0.017	0.0164
Chinese	CEval (5-shot)	0.2645	0.2712	0.2348	0.3076	0.3566	0.3990
	CMMLU (5-shot)	0.2455	0.2466	0.2529	0.2856	0.3453	0.3871
Average Metrics	Average-English(w/o Math)	0.4579	0.4700	0.4775	0.4842	0.5042	0.5317
	Average-English	0.3839	0.3944	0.4007	0.4053	0.4241	0.4471
	Average-Chinese	0.2550	0.2589	0.2438	0.2966	0.3509	0.3930
	Average	0.3655	0.3750	0.3783	0.3897	0.4137	0.4394
	Average(w/o Math)	0.4241	0.4349	0.4386	0.4529	0.4787	0.5086

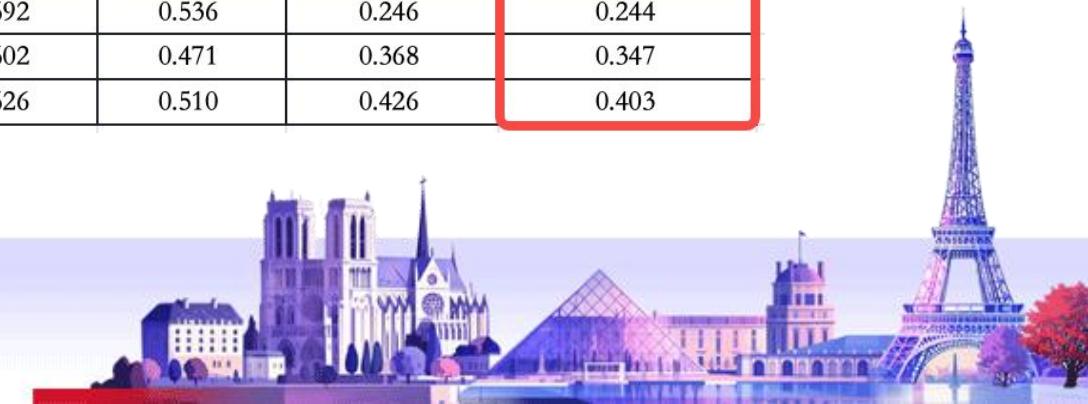




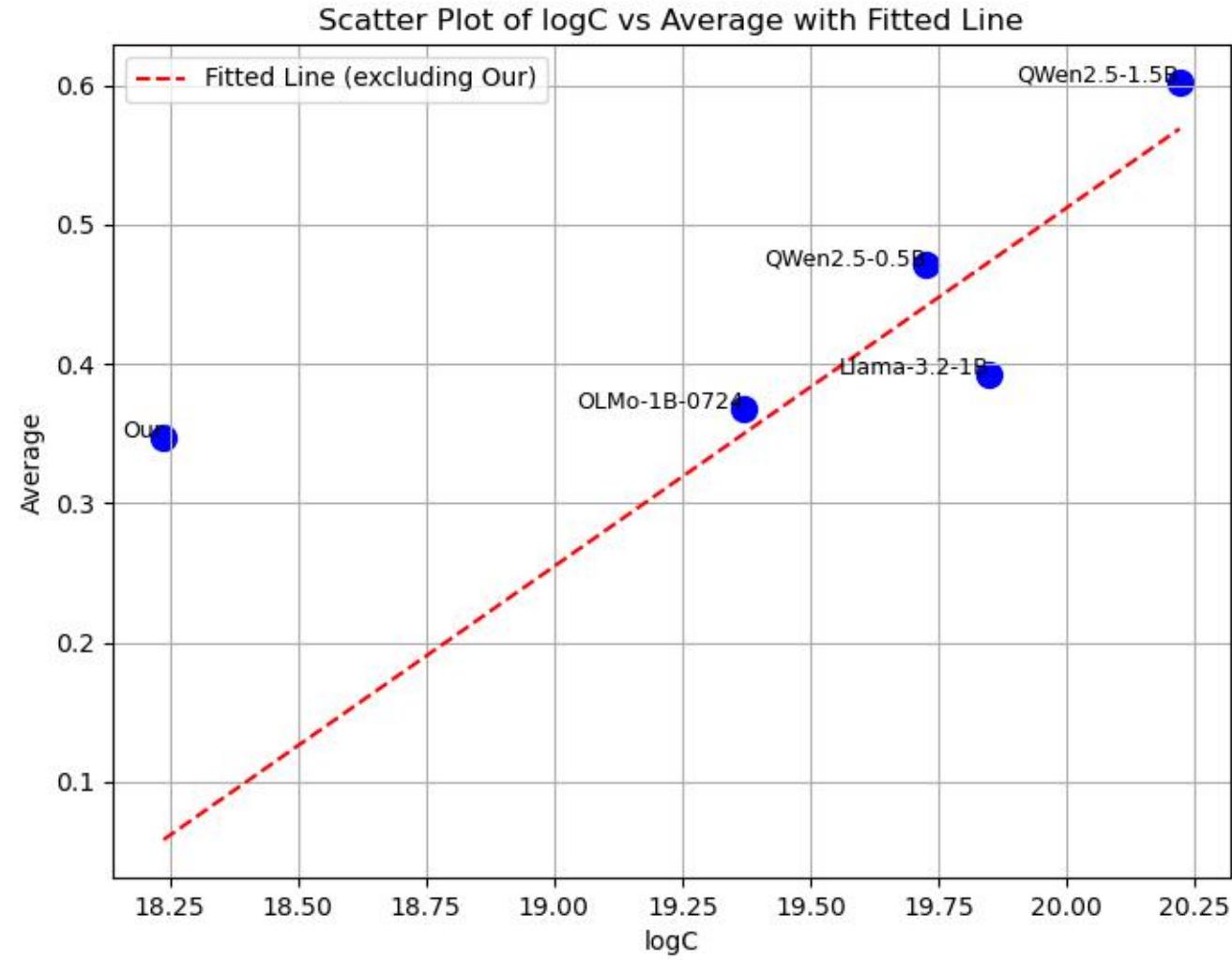
智源研究院

Release of OpenSeek-Small V1 GOSIM

		Llama-3.2-1B	Qwen2.5-1.5B	Qwen2.5-0.5B	OLMo-1B-0724	OpenSeek-Small-V1
English-Commonsense Reasoning	Training Tokens(TB)	9.2	18	18	3.05	0.72
	HellaSwag (5-shot)	0.483	0.501	0.401	0.491	0.389
	TruthfulQA (0-shot)	0.377	0.466	0.399	0.403	0.399
	Winogrande (5-shot)	0.621	0.645	0.568	0.629	0.554
	CommonsenseQA (5-shot)	0.312	0.745	0.549	0.195	0.205
	PIQA (5-shot)	0.751	0.761	0.711	0.746	0.720
	OpenBookQA (5-shot)	0.296	0.334	0.272	0.308	0.256
	BoolQ (5-shot)	0.659	0.777	0.657	0.651	0.617
English-Problem-Solving	ARC Easy (5-shot)	0.694	0.804	0.678	0.611	0.624
	ARC Challenge (5-shot)	0.353	0.485	0.337	0.306	0.316
	MMLU (5-shot)	0.312	0.617	0.482	0.287	0.265
English-Mathematics	GSM8K (5-shot)	0.064	0.619	0.350	0.016	0.018
	Minerva Math (4-shot)	0.018	0.288	0.116	0.018	0.001
Chinese	CEval (5-shot)	0.278	0.695	0.542	0.234	0.242
	CMMLU (5-shot)	0.269	0.688	0.530	0.257	0.247
Average	Average-English(w/o Math)	0.486	0.613	0.505	0.463	0.434
	Average-English	0.412	0.587	0.460	0.388	0.364
	Average-Chinese	0.273	0.692	0.536	0.246	0.244
	Average	0.392	0.602	0.471	0.368	0.347
	Average(w/o Math)	0.451	0.626	0.510	0.426	0.403

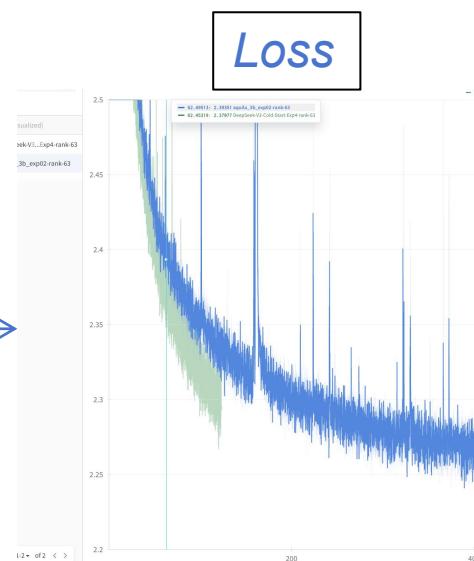


Release of OpenSeek-Small V1 GOSIM



Release of Pipeline

- We will release a set of code for verifying the data effect on small - sized models. Also, this code can be used for evaluating the effects of training strategies and model structures. The evaluation is based on model loss and other quantitative indicators.



Metrics	K76数据集及配比	K77数据集及配比
HellaSwag	52.40	53.86
ARC(Average)	49.29	50.44
PIQA	72.91	73.12
MMLU(cloze)	33.43	33.26
CommonsenseQA	42.42	40.38
TriviaQA	7.79	14.99
Winograde	54.30	54.22
OpenBookQA	38.00	37.4
GSM8K(5-shot)	0.99	1.21
SIQA	42.73	42.73
CEval	32.52	33.98
CMMU	32.85	34.38
Average-English	39.43	40.16
Average-Chinese	32.69	34.18
Average	36.06	37.17

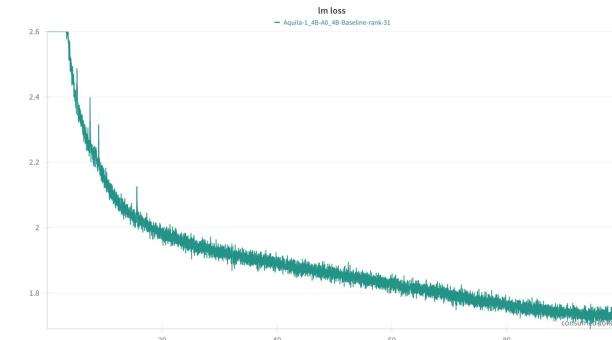
We will provide the **baseline curves** and corresponding **evaluation results**.



Baseline in PipeLine

- Model: 1.6B0.4A deepseek-v3-like
- Hyper: LR (3e-3 -> 3e-5), bsz=4M;
topk=6/64, share expert=2
- Data Mixtrue: En web (66.2%) 、 Cn Web(20.6%) 、 code (3.7%) 、 math (1.2%) 、 arxive (2.8%) 、 wiki (1.5%) 、 Synthesis (4%)
- Resource: GPU Ram >40G

- Loss



- Evaluation

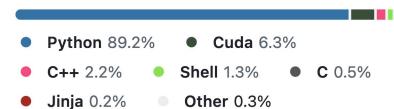
Category	Metrics (shots)	Llama-3.2-1B	Qwen2.5-1.5B	Qwen2.5-0.5B	OLMo-1B-0724	1_4B_A0_4B-100B
English-Commonsense Reasoning	HellaSwag (5-shot)	0.483	0.501	0.401	0.491	0.345
	TruthfulQA (0-shot)	0.377	0.466	0.399	0.403	0.385
	Winogrande (5-shot)	0.621	0.645	0.568	0.629	0.528
	CommonsenseQA (5-shot)	0.312	0.745	0.549	0.195	0.188
	PIQA (5-shot)	0.751	0.761	0.711	0.746	0.687
	OpenBookQA (5-shot)	0.296	0.334	0.272	0.308	0.224
	BoolQ (5-shot)	0.659	0.777	0.657	0.651	0.525
English-Problem-Solving	ARC Easy (5-shot)	0.694	0.804	0.678	0.611	0.602
	ARC Challenge (5-shot)	0.353	0.485	0.337	0.306	0.265
	MMLU (5-shot)	0.312	0.617	0.482	0.287	0.266
English-Mathematics	GSM8K (5-shot)	0.064	0.619	0.350	0.016	0.073
	Minerva Math (4-shot)	0.018	0.288	0.116	0.018	0.035
Chinese	CEval (5-shot)	0.278	0.695	0.542	0.234	0.249
	CMMLU (5-shot)	0.269	0.688	0.530	0.257	0.256
Average Metrics	Average-English(w/o Math)	0.486	0.613	0.505	0.463	0.402
	Average-English	0.412	0.587	0.460	0.388	0.344
	Average-Chinese	0.273	0.692	0.536	0.246	0.256
	Average	0.392	0.602	0.471	0.368	0.331
	Average(w/o Math)	0.451	0.626	0.510	0.426	0.377

Working Groups

Contributors 8



Languages



- **System:** FlagScale Team
- **Data:**
 - Wang Liangdong
 - Li Jijie
 - Yu Yang
- **Algorithm:**
 - Li Jijie/Xu Yao/Chen Jiabei
 - Wang Liangdong



Community

200+contributor
40+organization



<https://github.com/FlagAI-Open/OpenSeek>

Biweekly Meeting

- **3rd:** 2025-03-20 <https://meeting.tencent.com/crm/NXwDAyLG59>
- **2nd** 2025-02-27 : <https://meeting.tencent.com/crm/2pxo8BBDb7>
- **1st:** 2025-02-20 : https://meeting.tencent.com/v2/cloud-record/share?id=e188482b-0105-43f9-b8e7-cf5f1e4d136b&from=3is-single=false&record_type=2

Operation
Wang Songjing
Liao Feng/Bai Yu



THANK YOU

