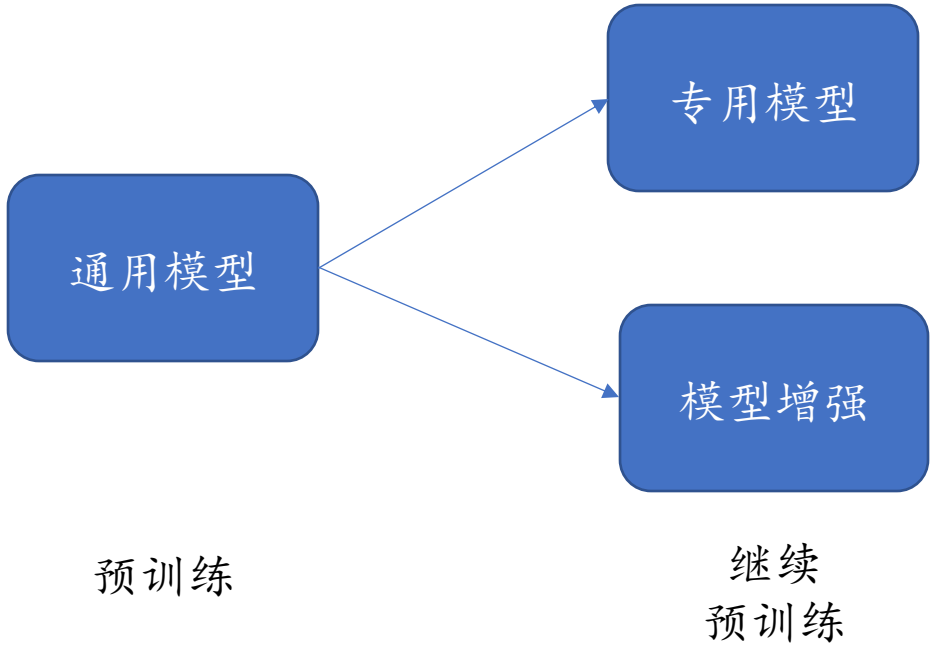
预训练之具体流程

《大语言模型》编写团队：赵鑫

继续预训练

- 继续预训练已经成为大模型增量研发的重要技术
  - 英文转多语
    - Llama 3 中文化
  - 通用转专用
    - CodeLlama
  - 拓长长文本
    - 16K转64K



继续预训练

➢ Llama 3 继续预训练初期实验

Detail		Benchmark				
		ceval	cmmlu	mmlu	humaneval	mbpp
Meta-Llama-3-8B		49.43	51.03	60.91	36.59	47.00
llama3 0:1	5b	47.84	47.93	56.94	29.88	42.60
	10b	47.12	47.62	58.26	33.54	44.40
llama3 1:4	3b	52.17	52.44	59.47	34.15	42.20
	4b	51.48	51.51	59.31	30.49	42.20
	5b	54.56	52.07	54.32	32.32	43.60
	6b	53.77	52.96	56.95	35.37	42.80
	7b	51.42	51.97	57.77	26.83	43.40
	8b	51.30	52.73	58.25	27.44	44.60
	9b	51.40	53.00	58.99	28.05	43.20
	10b	51.97	54.08	58.14	34.15	43.20

实验观察

- 出现了“灾难性遗忘”问题
- 中文榜单会出现震荡但整体呈现升高趋势
- 英文榜单均受到损害，部分英文榜单波动性比较大

推测原因

- Llama 3参数已经相对稳定，继续预训练破坏了原始的收敛
- 新训练数据分布与旧数据分布存在一定的差异

继续预训练

➢ Llama 3 继续预训练初期实验

Detail		Benchmark				
		ceval	cmmlu	mmlu	humaneval	mbpp
Meta-Llama-3-8B		49.43	51.03	60.91	36.59	47.00
llama3 0:1	5b	47.84	47.93	56.94	29.88	42.60
	10b	47.12	47.62	58.26	33.54	44.40
llama3 1:4	3b	52.17	52.44	59.47	34.15	42.20
	4b	51.48	51.51	59.31	30.49	42.20
	5b	54.56	52.07	54.32	32.32	43.60
	6b	53.77	52.96	56.95	35.37	42.80
	7b	51.42	51.97	57.77	26.83	43.40
	8b	51.30	52.73	58.25	27.44	44.60
	9b	51.40	53.00	58.99	28.05	43.20
	10b	51.97	54.08	58.14	34.15	43.20

实验观察

- 出现了“灾难性遗忘”问题
- 中文榜单会出现震荡但整体呈现升高趋势
- 英文榜单均受到损害，部分英文榜单波动性比较大

推测原因

- Llama 3参数已经相对稳定，继续预训练破坏了原始的收敛
- 新训练数据可能不如原始数据质量以及适配性更好

解决方法

- 合成增强特定能力的高质量数据（强化中英文数理能力）
- 设计更为合适的预训练数据课程（保持原始英文能力）

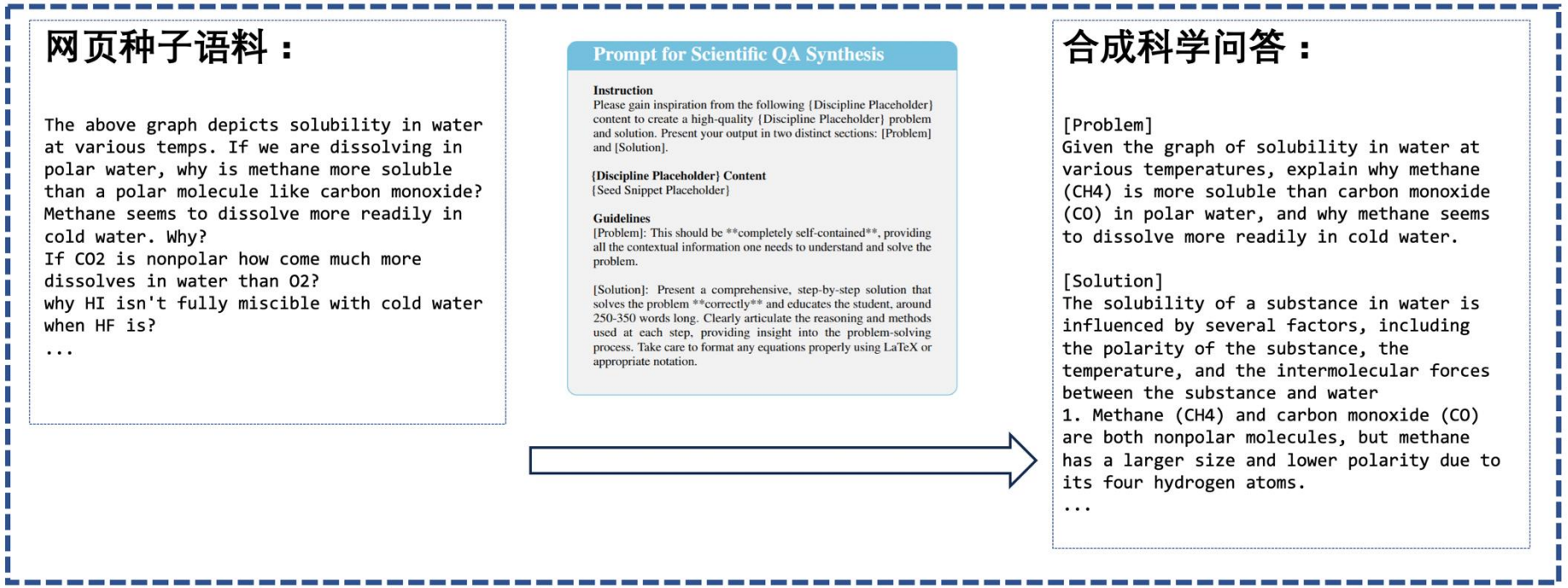
面向特定学科的数据合成方法

- 提升数理科学能力需要对应提供相关训练数据
  - 根据域名从开源网页数据集（如Dolma）中筛选科学相关种子语料

Subject	Domains
Math	math.stackexchange.com ...
Physics	physicsforums.com ...
Chemistry	chemcollective.org ...
Biology	biology.stackexchange.com …
Astronomy	skyandtelescope.org …
Earth science	earthscience.stackexchange.com …
Computer science	stackoverflow.com …
Medical science	health.stackexchange.com …
General education	indiabix.com …

面向特定学科的数据合成方法

- 设计 prompt 基于种子语料合成科学相关问答数据
  - 可使用开源小模型进行数据合成，降低推理开销



Llama3-SynE: 完整的继续预训练过程

➢ 数据配比

Dataset	English	Chinese	Volume
Web Pages	✓	✓	45.18B
Encyclopedia	✓	✓	4.92B
Books	✓	✓	15.74B
QA Forums	✓	✓	4.92B
Academic Papers	✓	×	7.93B
Mathematical Corpora	✓	×	7.93B
Code	✓	×	11.88B
Synthetic Data	✓	×	1.50B
Total	-	-	100.00B

训练数据主要由网页、图书和代码占据最高比例，进行了更为严格的数据清洗，最后混入了大概1.5B词元的合成数据（包括多学科+代码）

Llama3-SynE: 完整的继续预训练过程

- 数据课程（粗粒度）
  - 第一阶段：双语适配
    - 逐步让Llama3适应中文语料
    - 中英文比例为 2:8
    - 课程策略：PPL排序
  - 第二阶段：合成数据加强阶段
    - 融入合成数据，加强任务解决能力
    - 中文：英文：合成=1：7：2

Strategy	Bilingual Adpatation	Synthetic Enhancement
Topic-based Data Mixture	✓	×
PPL-based Data Curriculum	✓	×
Scientific Data Synthesizing	×	✓
Training Data Volume	92.5B	7.5B

整体课程策略和数据量

Llama3-SynE: 完整的继续预训练过程

- 数据课程（细粒度）
  - 第一阶段：双语适配（整体采用PPL由低到高进行排序）

Language	Topic
English	Mathematics and Physics Computer Science and Engineering Biology and Chemistry History and Geography Law and Policy Philosophy and Logic Economics and Business Psychology and Sociology Security and International Relations Medicine and Health Others
Chinese	Biology and Chemistry Computer Science and Engineering Economics and Business History and Geography Law and Policy Mathematics and Physics Medicine and Health Philosophy Arts and Culture Project and Practical Management Psychology Sociology and Education Others

将网页分类打标签

计算不同类别的PPL变化

$$\Delta p_i = p_i^{(t)} - p_i^{(t-1)}, \quad i = 1, \ldots, n,$$

$$\delta_{p_i} = \frac{\Delta p_i}{\max(|\Delta p_i|)}, \quad i = 1, \ldots, n.$$

根据PPL变化对应调整比例

$$f_i = 1 + \alpha \cdot \delta_{p_i} \cdot w_i,$$

$$r_i^{(t)} = \frac{r_i^{(t-1)} \cdot f_i}{\sum_{j=1}^{n} r_j^{(t-1)} \cdot f_j}.$$

动态调整类别采样比例

Llama3-SynE: 完整的继续预训练过程

- 数据课程（细粒度）
  - 第二阶段：合成数据（中文：英文：合成=1：7：2）

Category	Discipline	Num. Synthetic Data
Scientific	Mathematics	207,448
	Physics	241,516
	Chemistry	30,838
	Biology	25,103
	Astronomy	24,060
	Earth Science	7,936
	Medical Science	8,199
	Computer Science	475,566
	General Education	572,478
Code	-	1,385,696

- 本部分实验不再采用PPL的课程排序，直接随机采样
- 合成数据以QA对形式出现在训练数据中，和普通文档一样对待
- 不同学科由于原始数据的采集，可能会出现分布不均衡问题

Llama3-SynE: 完整的继续预训练过程

- 最终评测结果
  - 英文能力与原始模型相当
  - 中文、数学能力显著提升

Models	CEval	CMMLU	MMLU	MATH	GSM8K	HumanEval	MBPP
Llama-3-8B	49.43	51.03	**66.60**	16.20	54.40	36.59	**47.00**
Llama-3-SynE	**58.24**	**57.34**	65.19	**28.20**	**60.80**	**42.07**	45.6
	中文能力		英文能力	数学能力（英文）		代码能力（英文）	

Llama3-SynE: 完整的继续预训练过程

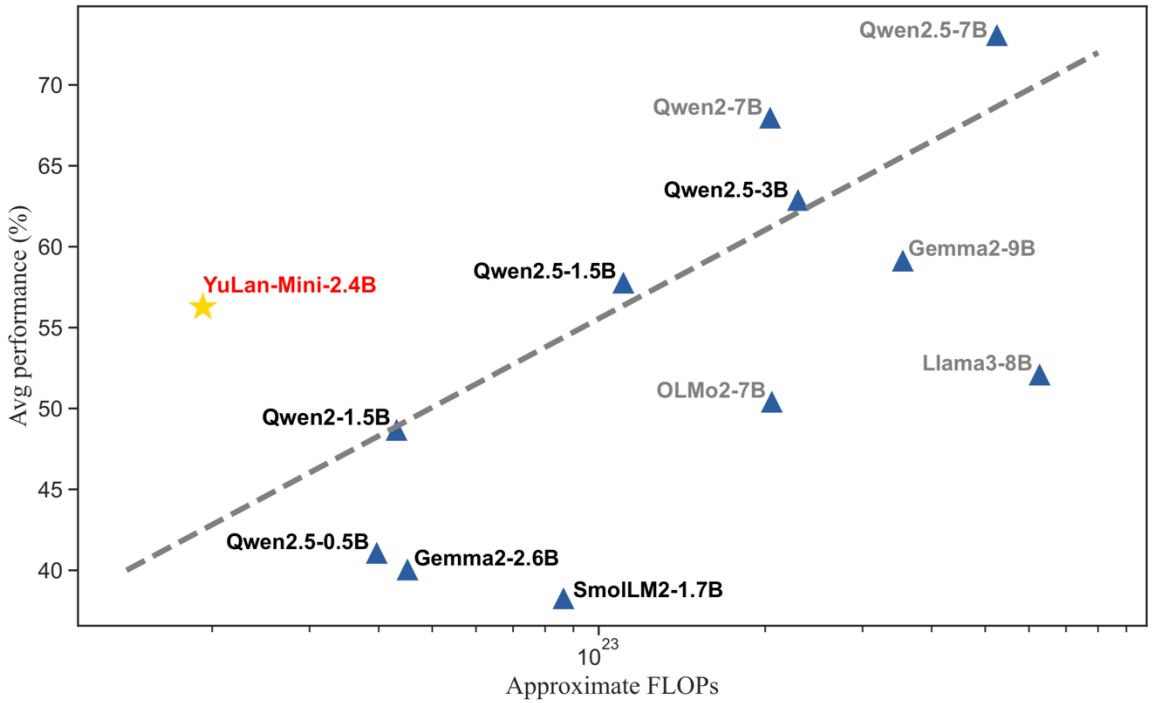
- 最终评测结果
  - 中文科学能力显著提升
  - 英文科学能力和推理能力有整体提升

Models	GaoKao-MathQA	GaoKao-Chemistry	GaoKao-Biology	SciEval	SciQ	ARC	SAT-Math
Llama-3-8B	27.92	32.85	43.81	65.47	90.90	84.51	38.64
Llama-3-SynE	**31.05**	**51.21**	**69.52**	**69.60**	**91.20**	**86.28**	**43.64**

中文高考学科测试　　英文科学&推理

YuLan-Mini: 完整的预训练过程

- 一个2.4B参数的小模型
  - 特点：数据细节全部开放，设计了一个高效的数据训练流程

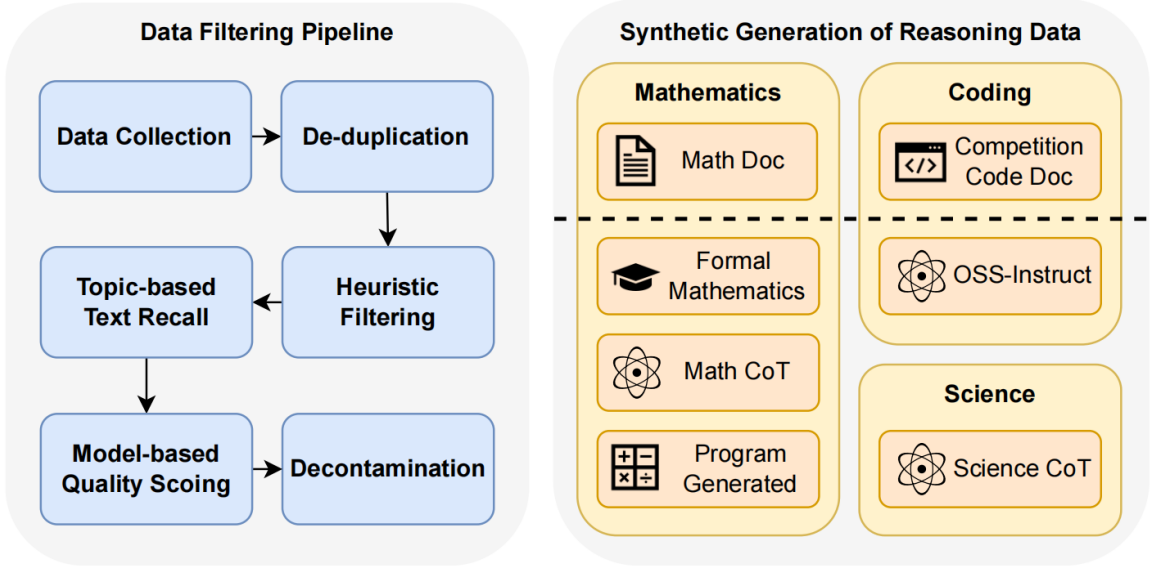


测试基准包括：GSM8K, MATH-500, HumanEval, MBPP, MMLU, ARC-Challenge, HellaSwag, and CEval

YuLan-Mini: 完整的预训练过程

- 整体数据策略
  - 数据筛选: 去重、启发式过滤、质量打分器、主题分类器、去污染
  - 数据合成: 综合尝试了各种合成数据

严格的文本质量控制



深入探索了合成数据

YuLan-Mini: 完整的预训练过程

- 数据配比
  - 训练数据: 1.08T 高质量 Tokens （多种数据源）
  - 数据配比: 数据课程控制由易到难

Type	Source	Volume
Web Pages	FineWeb-Edu, DCLM, Chinese-FineWeb-Edu	559.76B
Math (Pretrain)	AutoMathText, Proof-Pile-2, OpenWebMath Pro	85.00B
Code (Pretrain)	the-stack-v2, StarCoder	202.44B
General Knowledge	arXiv, StackExchange, English News	121.87B
Books	CBook, Gutenberg, LoC-PD-Books	52.13B
Encyclopedia	Wikipedia, Baidu-Baike	14.80B
Open-Source Instruction	SlimOrca, OpenMathInstruct-1, JiuZhang3.0	11.64B
Synthetic Pretrain Data (Ours)	Synthetic document (seed: AutoMathText, LeetCode)	8.76B
Synthetic Instruction (Ours)	Reasoning (seed: MetaMathQA, DeepMind Math, ...)	23.52B
Total	-	1,080B

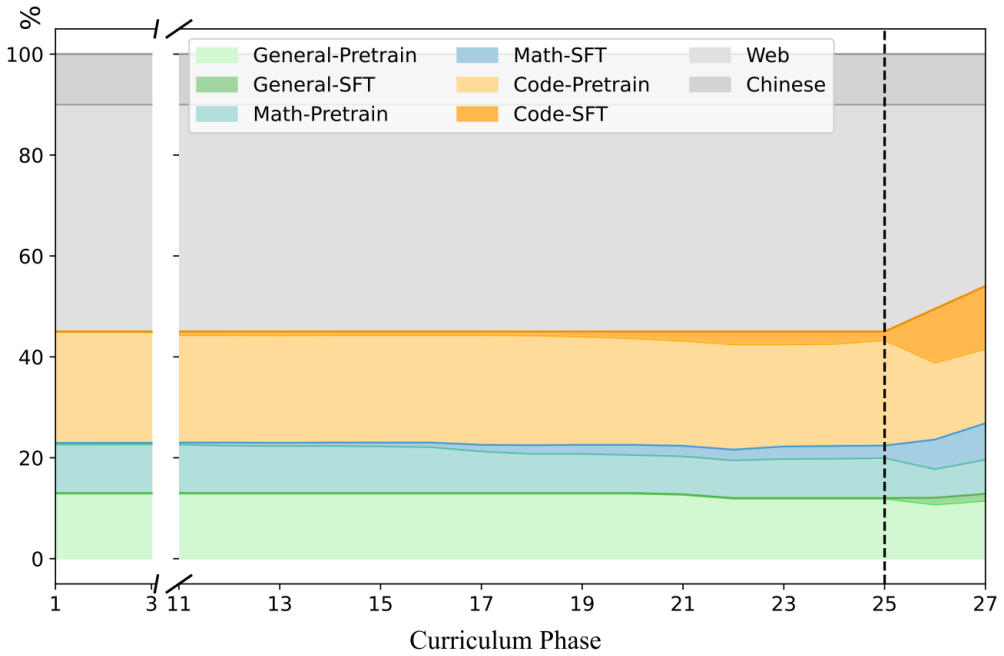
沿用了部分Yulan-3的中文数据，使用高质量语料，进行了丰富多样的数据合成，采取从易到难的数据课程，仅仅使用1.08T数据就在2.4B小模型上取得了先进的效果

YuLan-Mini: 完整的预训练过程

- 数据配比
  - 系统探索了推理数据的合成方法
    - 数学
      - 数学文档、CoT数据、Long CoT数据、形式推理数据（Lean）、数值推理数据
    - 代码
      - 竞赛代码合成（LeetCode）、OSS-Instruct（MagicCoder的代码指令合成方法）
    - 科学
      - 基于科学文档的科学问答（CoT形式）、基于学科试题的复杂推理（Long CoT形式）
    - 反思
      - 利用大模型创建错误反思过程

YuLan-Mini: 完整的预训练过程

- 数据课程
  - 三段式训练：热身训练（10B）、稳定训练（990B）和退火训练（80B）
    - 共划分为27个阶段，每个阶段40B数据



每隔40B数据根据测试结果微调一下数据训练分布，尽量连续阶段数据分布的平滑过渡。设置单独的退火阶段，加入高质量数据和指令数据，用于显著提升模型评测性能。

YuLan-Mini: 完整的预训练过程

- 数据课程
  - 全部开放：不同集合的采样配比

2 dclm (1.80), fineweb-edu (16.20), english-books (1.60), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cicg-news (0.76), cn-baike (0.39), mnbvc-news (0.08), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.90), starcoder (2.92), smollm-python (0.20), proof-pile-2 (1.52), automathtext (1.12), open-web-math-pro (0.20), cosmopedia (1.02), mathtext (0.12)

6 dclm (1.80), fineweb-edu (16.20), english-books (1.60), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cicg-news (0.76), cn-baike (0.39), mnbvc-news (0.08), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.90), starcoder (2.92), smollm-python (0.20), proof-pile-2 (1.54), automathtext (1.16), open-web-math-pro (0.26), cosmopedia (0.82), mathtext (0.12), metamathqa (0.02), orca-math (0.02), yulan-mini-syn-math-inst (0.04)

10 dclm (1.80), fineweb-edu (16.20), english-books (1.00), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.60), cicg-news (0.76), cn-baike (0.39), mnbvc-news (0.08), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.85), starcoder (2.92), smollm-python (0.20), yulan-mini-syn-code-inst (0.03), proof-pile-2 (1.64), automathtext (1.17), open-web-math-pro (0.32), cosmopedia (0.29), fineweb-math (0.20), mathtext (0.12), metamathqa (0.01), orca-math (0.01), yulan-mini-syn-math-inst (0.02), yulan-mini-syn-math-doc (0.22)

YuLan-Mini: 完整的预训练过程

- 数据课程
  - 退火训练：学习率从10^-2 衰减到 5.22 × 10 ^-5，共80B数据
    - 引入了高质量数据（包括合成数据）、长文本数据和指令数据

Domain	Type	Dataset	Volume
Mix	Pretrain	FineWeb-Edu, CBook, arXiv	64.65B
Math	(1) CoT	Deepmind-Math, MathInstruct	3.07B
	(2) Long CoT	Numina, AMPS, Platypus	0.61B
	(3) Formal math	Lean-GitHub, Lean-WorkBook, DeepSeek-Prover-V1	0.10B
	(4) Curated	Tulu v3, MathInstruct	1.42B
Code	(1) CoT	OSS-Instruct (seed: the-Stack-v2), OpenCoder-LLM	6.66B
	(2) Curated	LeetCode, XCoder-80K	2.39B
Science	(1) Long CoT	Camel-ai	0.04B
	(2) Curated	EvolKit-20k, Celestia, Supernova	1.06B
Total	-	-	80B

YuLan-Mini: 完整的预训练过程

- 最终评测结果
  - 训练高效性、优秀数学和代码能力

Models	Model Size	# Train Tokens	Context Length	MATH 500	GSM 8K	Human Eval	MBPP	RACE Middle	RACE High	RULER
MiniCPM	2.6B	1.06T	4K	15.00	53.83	50.00*	47.31	56.61	44.27	N/A
Qwen-2	1.5B	7T	128K	22.60	46.90*	34.80*	46.90*	55.77	43.69	60.16
Qwen2.5	0.5B	18T	128K	23.60	41.60*	30.50*	39.30*	52.36	40.31	49.23
Qwen2.5	1.5B	18T	128K	**45.40**	**68.50***	37.20*	60.20*	**58.77**	44.33	68.26
Gemma2	2.6B	2T	8K	18.30*	30.30*	19.50*	42.10*	-	-	N/A
StableLM2	1.7B	2T	4K	-	20.62	8.50	17.50	56.33	**45.06**	N/A
SmolLM2	1.7B	11T	8K	11.80	-	23.35	45.00	55.77	43.06	N/A
Llama3.2	3.2B	9T	128K	7.40	-	29.30	49.70	55.29	43.34	**77.06**
YuLan-Mini	2.4B	1.04T	4K	32.60	66.65	61.60	**66.70**	55.71	43.58	N/A
	2.4B	1.08T	28K	37.80	68.46	**64.00**	65.90	57.18	44.57	51.48

YuLan-Mini: 完整的预训练过程

- 最终评测结果
  - 强大的通用能力

Models	LAMBADA	MMLU	CMMLU	CEval	Hella Swag	Wino Grande	Story Cloze	ARC-e	ARC-c
MiniCPM-2.6B	61.91	53.37	48.97	48.24	67.92	65.74	78.51	55.51	43.86
Qwen2-1.5B	64.68	55.90	**70.76**	**71.94**	66.11	66.14	77.60	62.21	42.92
Qwen2.5-0.5B	52.00	47.50	52.17	54.27	50.54	55.88	71.67	56.10	39.51
Qwen2.5-1.5B	62.12	60.71	67.82	69.05	67.18	64.48	76.80	**71.51**	53.41
Gemma2-2.6B	-	52.20*	-	28.00*	74.60*	**71.50***	-	-	**55.70***
StableLM2-1.7B	66.15	40.37	29.29	26.99	69.79	64.64	78.56	54.00	40.78
SmolLM2-1.7B	67.42	51.91	33.46	35.10	72.96	67.40	**79.32**	44.82	35.49
Llama3.2-3B	**69.08**	**63.40**	44.44	44.49	**75.62**	67.48	76.80	70.12	48.81
YuLan-Mini	64.72	51.79	48.35	51.47	68.65	67.09	76.37	69.87	50.51
	65.67	49.10	45.45	48.23	67.22	67.24	75.89	67.47	49.32

谢谢