1 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 (4.91), starcoder (2.92), smollm-python (0.20), proof-pile-2 (1.52), automathtext (1.12), open-web-math-pro (0.20), cosmopedia (1.01), mathtext (0.12)

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)

3 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 (4.86), 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)

4 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 (4.14), starcoder (2.92), smollm-python (0.20), proof-pile-2 (1.52), automathtext (1.12), open-web-math-pro (0.24), cosmopedia (0.98), mathtext (0.12)

5 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 (1.14), open-web-math-pro (0.24), cosmopedia (0.94), 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)

7 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 (1.60), automathtext (1.17), open-web-math-pro (0.28), cosmopedia (0.77), mathtext (0.12), metamathqa (0.01), orca-math (0.01), yulan-mini-syn-math-inst (0.02)

8 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 (1.64), automathtext (1.17), open-web-math-pro (0.32), cosmopedia (0.53), fineweb-math (0.16), mathtext (0.12), metamathqa (0.01), orca-math (0.01), yulan-mini-syn-math-inst (0.02)

9 dclm (1.80), fineweb-edu (16.20), english-books (1.20), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.40), 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 (4.86), starcoder (2.92), smollm-python (0.20), proof-pile-2 (1.64), automathtext (1.17), open-web-math-pro (0.32), cosmopedia (0.33), fineweb-math (0.16), mathtext (0.12), metamathqa (0.01), orca-math (0.01), yulan-mini-syn-math-inst (0.02), yulan-mini-syn-math-doc (0.22)

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)

11 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), 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 (4.56), starcoder (2.92), smollm-python (0.20), mnbvc-code (0.04), yulan-mini-syn-code-inst (0.28), proof-pile-2 (1.64), automathtext (0.93), open-web-math-pro (0.41), cosmopedia (0.16), fineweb-math (0.33), dclm-math (0.12), mathtext (0.12), metamathqa (0.01), orca-math (0.01), yulan-mini-syn-math-inst (0.02), yulan-mini-syn-math-doc (0.25)

12 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), 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 (4.44), starcoder (2.92), smollm-python (0.20), mnbvc-code (0.16), yulan-mini-syn-code-inst (0.28), proof-pile-2 (1.64), automathtext (0.63), open-web-math-pro (0.41), cosmopedia (0.03), fineweb-math (0.50), dclm-math (0.11), mathtext (0.12), basic-math-10m (0.04), metamathqa (0.01), orca-math (0.01), yulan-mini-syn-math-inst (0.06), yulan-mini-syn-math-doc (0.44)

13 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), 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 (4.44), starcoder (2.92), smollm-python (0.20), mnbvc-code (0.16), yulan-mini-syn-code-inst (0.28), proof-pile-2 (1.64), automathtext (0.58), open-web-math-pro (0.41), fineweb-math (0.51), dclm-math (0.15), mathtext (0.12), basic-math-10m (0.04), metamathqa (0.01), yulan-mini-syn-math-inst (0.08), yulan-mini-syn-math-doc (0.44)

14 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), cicg-news (0.76), cn-baike (0.33), mnbvc-news (0.08), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.44), starcoder (2.92), smollm-python (0.20), mnbvc-code (0.16), yulan-mini-syn-code-inst (0.28), proof-pile-2 (1.64), automathtext (0.57), open-web-math-pro (0.41), cosmopedia (0.02), fineweb-math (0.51), dclm-math (0.15), mathtext (0.12), basic-math-10m (0.04), metamathqa (0.01), yulan-mini-syn-math-inst (0.09), yulan-mini-syn-math-doc (0.44)

15 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), cicg-news (0.76), cn-baike (0.27), mnbvc-news (0.08), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.37), starcoder (2.32), smollm-python (0.20), mnbvc-code (0.83), yulan-mini-syn-code-inst (0.28), proof-pile-2 (1.34), automathtext (0.65), open-web-math-pro (0.41), cosmopedia (0.02), fineweb-math (0.42), dclm-math (0.40), mathtext (0.12), basic-math-10m (0.04), metamathqa (0.01), yulan-mini-syn-math-inst (0.13), yulan-mini-syn-math-doc (0.46)

16 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), cicg-news (0.76), cn-baike (0.27), mnbvc-news (0.10), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.27), starcoder (2.12), smollm-python (0.20), mnbvc-code (1.13), yulan-mini-syn-code-inst (0.28), proof-pile-2 (1.03), automathtext (0.68), open-web-math-pro (0.36), cosmopedia (0.02), fineweb-math (0.46), dclm-math (0.51), mathtext (0.12), basic-math-10m (0.04), yulan-mini-syn-math-inst (0.20), yulan-mini-syn-math-doc (0.57)

17 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), cicg-news (0.76), cn-baike (0.27), mnbvc-news (0.10), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.45), starcoder (2.12), smollm-python (0.20), mnbvc-code (1.13), yulan-mini-syn-code-inst (0.28), proof-pile-2 (0.55), automathtext (0.68), cosmopedia (0.02), fineweb-math (0.46), dclm-math (1.02), mathtext (0.12), basic-math-10m (0.04), yulan-mini-syn-math-inst (0.36), yulan-mini-syn-math-doc (0.57)

18 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), cicg-news (0.76), cn-baike (0.27), mnbvc-news (0.10), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.45), starcoder (2.12), smollm-python (0.20), mnbvc-code (1.13), yulan-mini-syn-code-inst (0.31), opencoder-llm-math-web (1.54), automathtext (0.00), cosmopedia (0.02), fineweb-math (0.17), dclm-math (0.82), mathtext (0.12), basic-math-10m (0.04), yulan-mini-syn-math-inst (0.52), yulan-mini-syn-math-doc (0.56)

19 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (1.24), cosmopedia-v2 (0.78), cicg-news (0.76), cn-baike (0.02), mnbvc-news (0.02), cn-book (0.24), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.41), starcoder (2.12), smollm-python (0.20), mnbvc-code (1.06), opencoder-llm-annealing (0.08), yulan-mini-syn-code-inst (0.31), opencoder-llm-math-web (0.38), cosmopedia (0.02), fineweb-math (0.10), dclm-math (0.82), infimm-webmath (1.23), mathtext (0.12), basic-math-10m (0.04), yulan-mini-syn-math-inst (0.55), yulan-mini-syn-math-doc (0.56)

20 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.40), dolma (0.84), opencoder-llm-fineweb-corpus (0.40), cosmopedia-v2 (0.78), cicg-news (0.76), mnbvc-news (0.02), cn-book (0.26), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.33), starcoder (2.12), smollm-python (0.20), mnbvc-code (0.92), opencoder-llm-annealing (0.16), yulan-mini-syn-code-inst (0.45), opencoder-llm-math-web (0.24), cosmopedia (0.02), fineweb-math (0.04), dclm-math (0.56), infimm-webmath (1.62), mathtext (0.12), basic-math-10m (0.01), yulan-mini-syn-math-inst (0.67), yulan-mini-syn-math-doc (0.53)

21 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), wikipedia (0.30), dolma (0.84), opencoder-llm-fineweb-corpus (0.50), cosmopedia-v2 (0.78), cicg-news (0.76), mnbvc-news (0.02), cn-book (0.26), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.19), starcoder (2.12), smollm-python (0.20), mnbvc-code (0.80), opencoder-llm-annealing (0.41), yulan-mini-syn-code-inst (0.45), opencoder-llm-math-web (0.24), cosmopedia (0.02), dclm-math (0.23), infimm-webmath (2.00), mathtext (0.12), yulan-mini-syn-math-inst (0.70), yulan-mini-syn-math-doc (0.53)

22 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), dolma (0.84), opencoder-llm-fineweb-corpus (0.80), cosmopedia-v2 (0.78), cicg-news (0.76), mnbvc-news (0.02), cn-book (0.26), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.60), starcoder (1.10), smollm-python (0.20), mnbvc-code (1.13), opencoder-llm-sft-s1 (0.20), opencoder-llm-annealing (0.38), yulan-mini-syn-code-inst (0.56), opencoder-llm-math-web (0.24), cosmopedia (0.02), dclm-math (0.22), infimm-webmath (1.98), mathtext (0.06), yulan-mini-syn-math-inst (0.78), yulan-mini-syn-math-doc (0.53)

23 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), dolma (0.84), opencoder-llm-fineweb-corpus (0.80), cosmopedia-v2 (0.78), cicg-news (0.76), mnbvc-news (0.02), cn-book (0.26), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.16), starcoder (0.80), mnbvc-code (0.85), opencoder-llm-sft-s1 (0.20), opencoder-llm-sft-s2 (0.15), opencoder-llm-annealing (1.20), yulan-mini-syn-code-inst (0.56), opencoder-llm-math-web (0.24), cosmopedia (0.02), dclm-math (0.22), infimm-webmath (2.06), mathtext (0.04), lean (0.02), yulan-mini-syn-math-inst (0.92), yulan-mini-syn-math-doc (0.56)

24 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), dolma (0.84), opencoder-llm-fineweb-corpus (0.80), cosmopedia-v2 (0.78), cicg-news (0.76), mnbvc-news (0.02), cn-book (0.26), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.07), starcoder (0.80), mnbvc-code (0.85), opencoder-llm-sft-s1 (0.25), opencoder-llm-sft-s2 (0.08), opencoder-llm-annealing (1.27), yulan-mini-syn-code-inst (0.56), opencoder-llm-math-web (0.24), cosmopedia (0.02), dclm-math (0.19), infimm-webmath (2.11), lean (0.04), yulan-mini-syn-math-inst (0.96), yulan-mini-syn-math-doc (0.56)

25 dclm (1.80), fineweb-edu (16.20), english-books (0.82), pes2o (0.80), arxiv (1.20), dolma (0.84), opencoder-llm-fineweb-corpus (0.80), cosmopedia-v2 (0.78), cicg-news (0.76), mnbvc-news (0.02), cn-book (0.26), cn-legal-case-law (0.36), zhihu-qa (0.12), the-stack-v2 (4.11), starcoder (0.80), mnbvc-code (0.78), opencoder-llm-sft-s1 (0.25), opencoder-llm-sft-s2 (0.08), opencoder-llm-annealing (1.30), ioccc (0.00), yulan-mini-syn-code-inst (0.54), opencoder-llm-math-web (0.24), cosmopedia (0.02), fineweb-math (0.20), dclm-math (0.17), infimm-webmath (2.11), lean (0.04), yulan-mini-syn-math-inst (0.93), yulan-mini-syn-math-doc (0.45)

YuLan-Mini: An Open Data-efficient Language Model

26 dclm (1.62), fineweb-edu (14.58), english-books (0.74), pes2o (0.72), arxiv (1.08), dolma (0.48), opencoder-llm-fineweb-corpus (1.00), cosmopedia-v2 (0.70), cicg-news (0.68), wizardlm-evol-instruct-v2-196k (0.04), less-data (0.04), claude-3-opus-claude-3.5-sonnnet-9k (0.01), slimorca (0.16), tulu-v3.1-mix-preview-4096-olmoe (0.20), supernova (0.03), magpie-reasoning-150k (0.09), spurline (0.01), celestia (0.04), mnbvc-news (0.01), cn-book (1.00), zhihu-qa (0.05), chinese-porety (0.03), the-stack-v2 (3.16), starcoder (0.18), mnbvc-code (0.70), opencoder-llm-sft-s1 (0.62), opencoder-llm-annealing (1.09), magicoder-oss (0.06), textbook-quality-programming (0.06), yulan-mini-syn-code-inst (3.00), code-290k-sharegpt (0.08), evol-codealpaca-v1 (0.04), magicoder-evol-instruct-110k (0.04), mathcodeinstruct (0.03), codefeedback-filtered-instruction (0.08), python-code-23k-sharegpt (0.01), evol-instruct-code-80k-v1 (0.03), codeexercise-python-27k (0.02), xcoder-80k (0.04), leetcode-solution-python (0.00), tachibana (0.03), opencoder-llm-math-web (0.24), cosmopedia (0.12), infimm-webmath (1.33), ape210k (0.01), polytope (0.03), yulan-mini-syn-math-inst (1.44), yulan-mini-syn-math-doc (0.58), mammothmathinstruct (0.04), openmathinstruct-1 (0.35), fol-nli (0.12), mathscaleqa-2m (0.28)

27 dclm (1.44), fineweb-edu (12.96), english-books (1.48), pes2o (0.64), arxiv (0.96), dolma (0.20), opencoder-llm-fineweb-corpus (1.08), cosmopedia-v2 (0.70), cicg-news (0.61), wizardlm-evol-instruct-v2-196k (0.04), long-cot (0.65), slimorca (0.04), tulu-v3.1-mix-preview-4096-olmoe (0.25), evolkit-20k (0.02), orca-agentinstruct (0.49), transcript (0.01), spurline (0.01), titanium (0.02), celestia (0.06), cn-book (1.40), zhihu-qa (0.05), ruozhiba (0.00), chinese-porety (0.04), the-stack-v2 (1.50), mnbvc-code (0.38), opencoder-llm-sft-s1 (0.16), opencoder-llm-annealing (1.13), magicoder-oss (0.11), textbook-quality-programming (0.05), longwanjuan-github (1.82), yulan-mini-syn-code-inst (3.75), code-290k-sharegpt (0.04), evol-codealpaca-v1 (0.03), magicoder-evol-instruct-110k (0.03), mathcodeinstruct (0.02), codefeedback-filtered-instruction (0.01), self-oss-instruct-sc2-exec-filter-50k (0.02), xcoder-80k (0.04), tulu-code (0.02), codefuse-evol-instruct-clean (0.03), proof-pile-2 (0.60), opencoder-llm-math-web (0.45), cosmopedia (0.02), dclm-math (0.06), infimm-webmath (1.34), yulan-mini-syn-math-inst (1.72), yulan-mini-syn-math-doc (0.25), mammothmathinstruct (0.02), openmathinstruct-1 (0.02), tulu-math (0.14), tulu-math-grade (0.03), tulu-algebra (0.02), fol-nli (0.12), reasoning-0.01 (0.02), gretel-math-gsm8k-v1 (0.01), mathscaleqa-2m (0.40)