MixtureVitae

Provenance-structured, performant pretraining from permissive-first sources

422B tokens · three provenance tiers · best permissive results at 1.7B / 300B · strongly outperforms non-permissive on math/code

Huu Nguyen*, Victor May*, Harsh Raj*, Marianna Nezhurina, Yishan Wang, Yanqi Luo,
Minh Chien Vu, Taishi Nakamura, Ken Tsui, Van Khue Nguyen, David Salinas,
Aleksandra Krasnodębska, Christoph Schuhmann, Mats Leon Richter, Xuan-Son Vu, Jenia Jitsev
(* equal contribution)

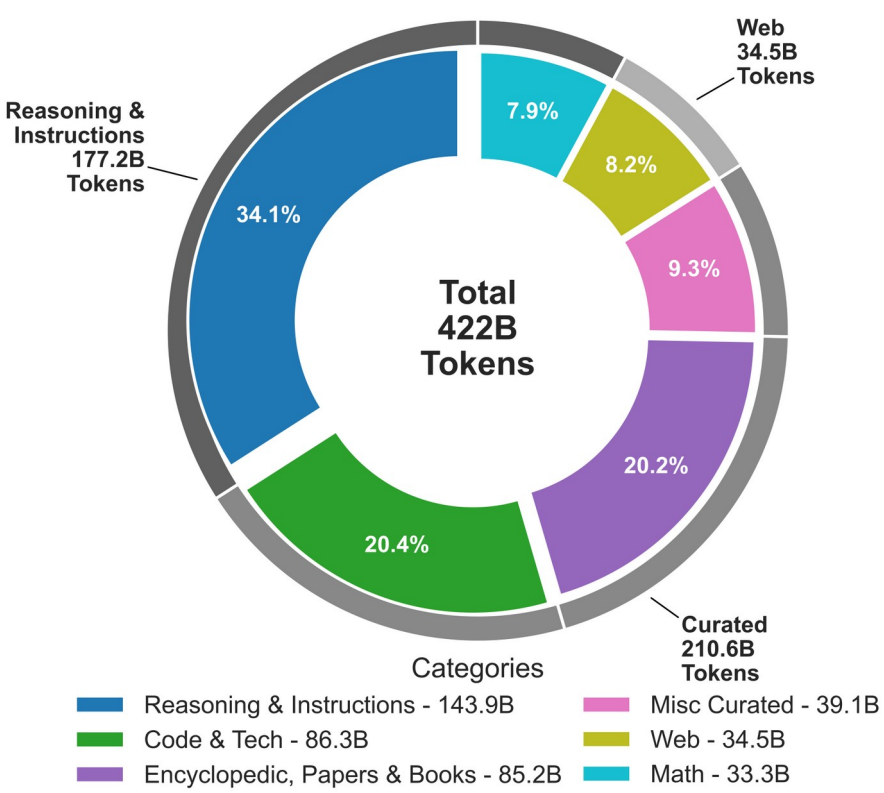
MOTIVATION

Why MixtureVitae is different

- Not just a permissive corpus - a **provenance-structured** one
- **422B** tokens incl. synth data (eg compared to Comma-0.1 **~316B,** no synth)
- Front loads reasoning/instruction data into **pretraining**
- Best permissive results, competitive with strong mixed-license baselines
- Strongly outperforms all baselines on math/code, without degradation on language

DATASET

This is intentionally not a web text dominant composition



422B

total tokens

177.2B

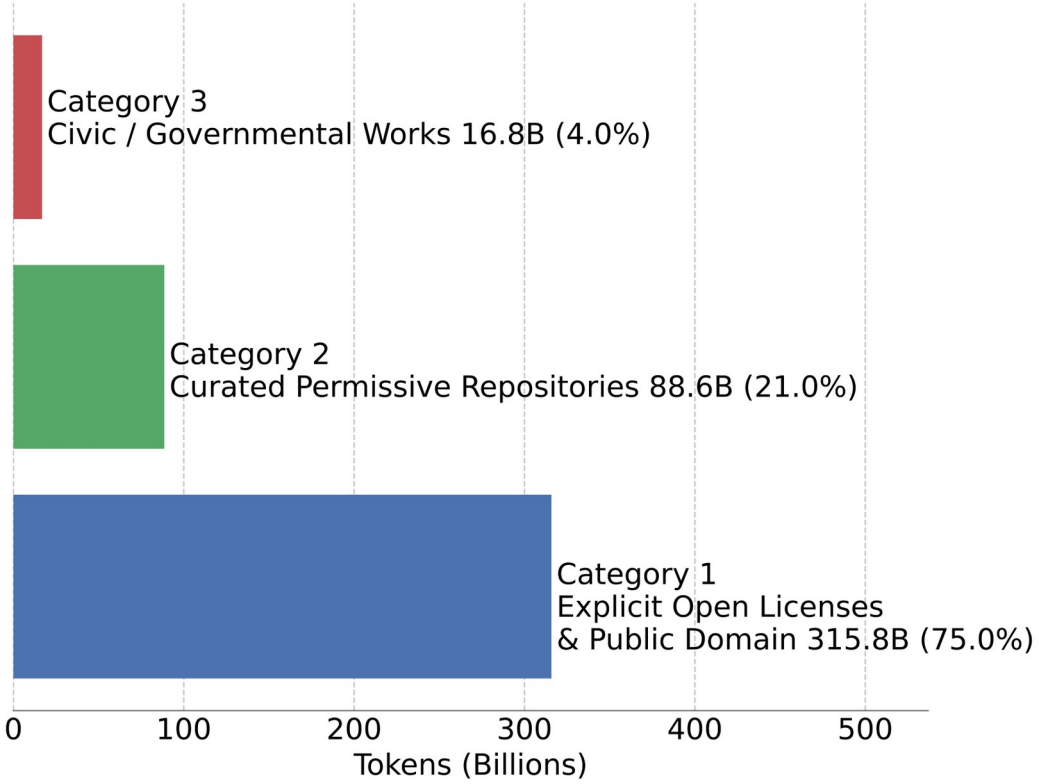
reasoning & instruction

34.5B

raw web — much smaller fraction than usual

DATASET

Three provenance tiers, with user-selectable risk



Tier 1 — 315.8B tokens (75%)

Explicit open licenses + public domain

Tier 2 — 88.6B tokens (21%)

Curated permissive sources with partial opacity

Tier 3 — 16.8B tokens (4%)

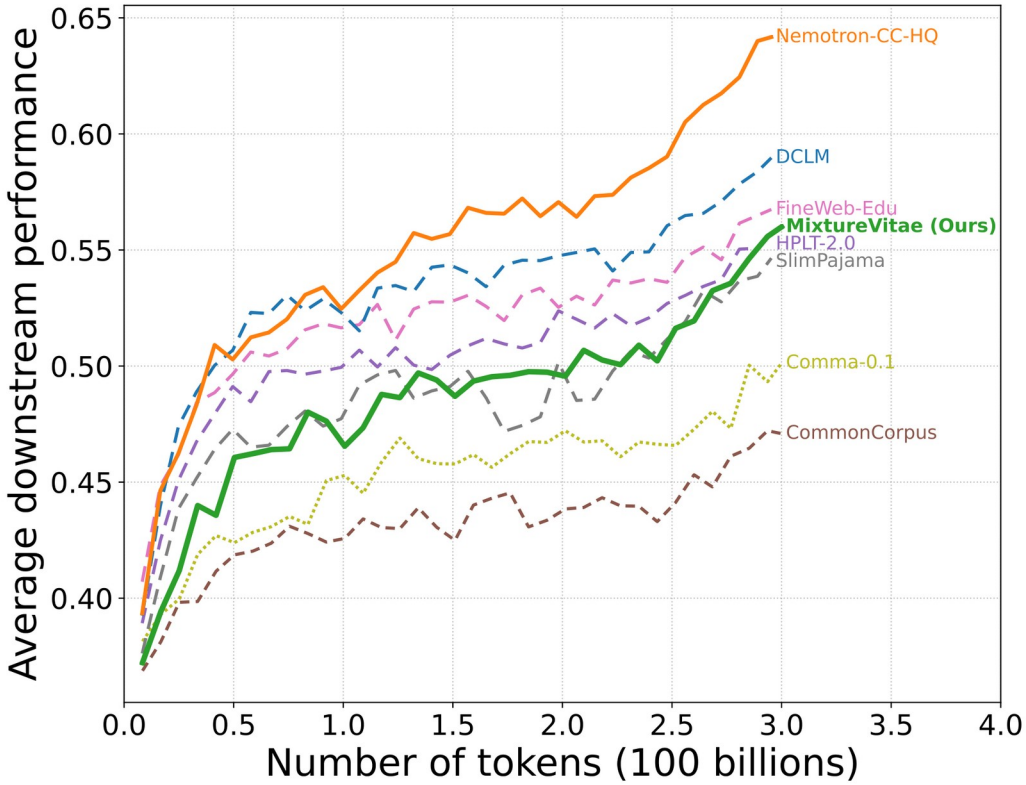
Civic / governmental works

Tier 2(b) is a small slice and can be excluded without loss.

Risk-mitigated, not risk-free.

RESULTS

Best permissive corpus across 11 tasks



Average score at 300B tokens

MixtureVitae	**0.56**
DCLM	0.59
FineWeb-Edu	0.57
Comma-0.1	0.50
CommonCorpus	0.47

RESULTS

A base model with unusually strong math and code

GSM8K

0.53

SmolLM2-Instruct: 0.37

MixtureVitae base ↑

HumanEval

0.32

SmolLM2-Instruct: 0.28

MixtureVitae base ↑

MBPP

0.38

SmolLM2-Instruct: 0.37

MixtureVitae base ↑

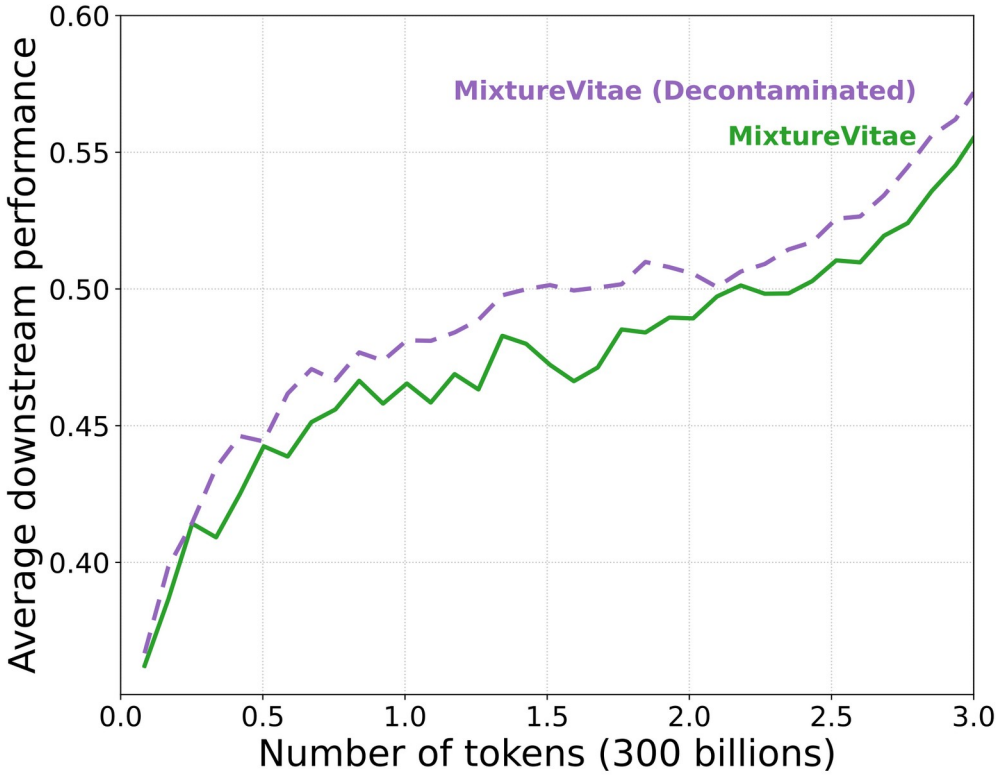
Matches or beats SmolLM2-1.7B-Instruct on all three — using 300B tokens vs. ~11T

IFEval: 0.19 · no instruction tuning applied — base model only

A single-stage pretrained base model already behaves unusually well on math and code.

VALIDITY

The gains survive decontamination



GSM8K

0.53 → 0.54

after decontamination

MBPP

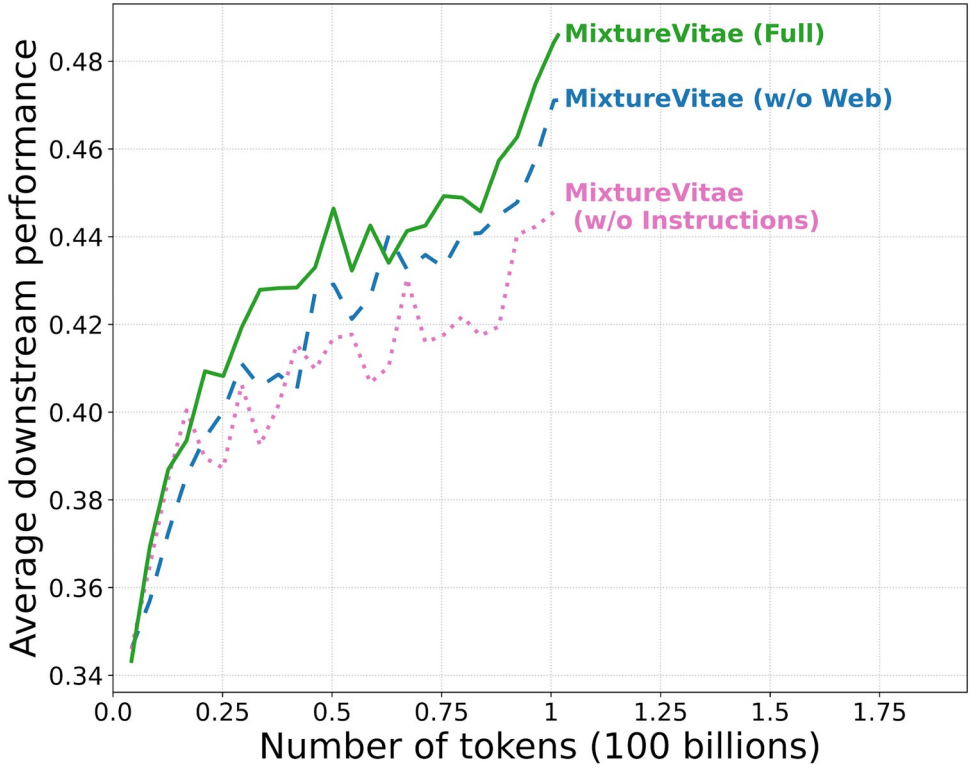
0.38 → 0.38

unchanged

Retraining after removing overlap-heavy shards **does not** change performance

ANALYSIS

Instruction and reasoning data drive most of the gain



Especially strong drop on math/code, eg GSM8K :

Full mix

0.47

Without web

0.41

Without instructions and reasoning

0.03

SUMMARY

What to remember

1. MixtureVitae is provenance-structured, not just permissive.

2. It is the strongest permissive corpus tested, overall competitive with mixed-license baselines, outperforming all baselines by large margins specifically on math/code

3. Its math/code gains survive decontamination and come mainly from reasoning-heavy pretraining.

MixtureVitae shows that permissive-first pretraining without broad non-permissive web scrapes can recover strong capability by front-loading reasoning/instruct data

Paper
arxiv.org/abs/2509.25531 (ArXiv)
https://openreview.net/forum?id=SyCcUNUUMf
(TMLR)

Website
mixturevitae.github.io

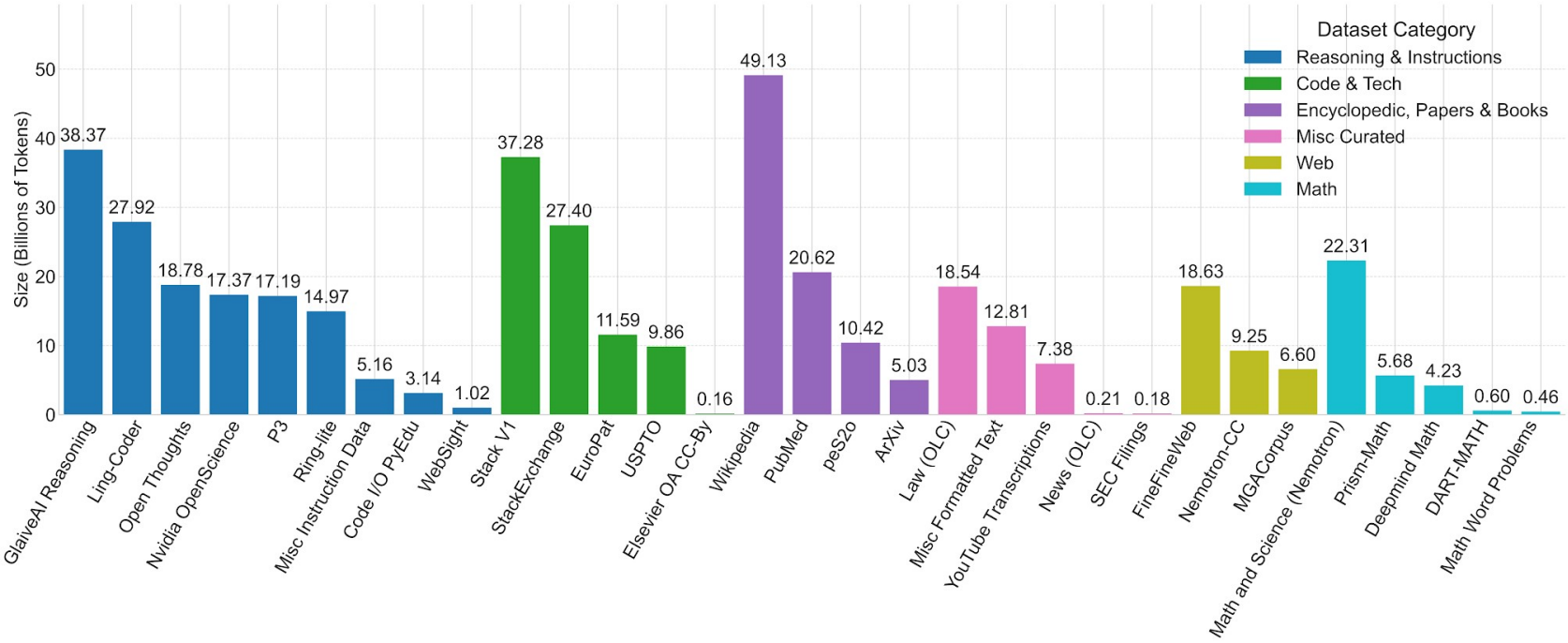
Dataset
hf.co/datasets/ontocord/MixtureVitae-v1

Code
github.com/ontocord/mixturevitae

Appendix

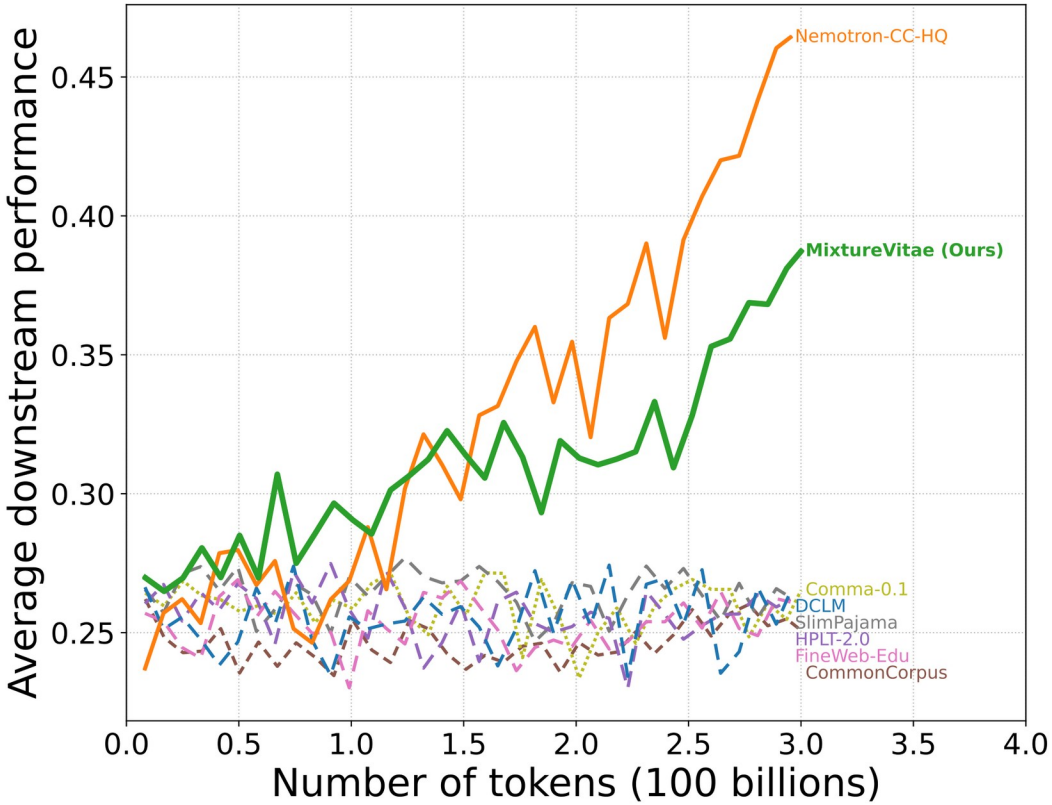
Where the tokens actually come from

The mixture is broad, but a few large reasoning, code, and encyclopedic sources do most of the work.



Reasoning gains appear early, not just at the end

MixtureVitae is the only permissive corpus with a strong MMLU signal at this scale.



Math/code/instruct: outperforming strong baselines

1.7B reference model trained on MixtureVitae leaves all strong baselines far behind

Training Dataset	Tokens	IF-Eval	GSM8K	HumanEval	MBPP	Average
Models Trained with `open-sci-ref` for 300B Tokens						
MixtureVitae	300B	0.19	**0.53**	**0.32**	**0.38**	**0.36**
Comma-0.1	300B	0.19	0.06	0.13	0.22	0.15
CommonCorpus	300B	0.13	0.02	0.05	0.05	0.06
C4	300B	0.20	0.02	0.00	0.00	0.06
SlimPajama	300B	0.14	0.02	0.05	0.00	0.05
HPLT-2.0	300B	0.17	0.02	0.00	0.00	0.05
DCLM	300B	0.13	0.02	0.01	0.01	0.04
Nemotron-CC-HQ	300B	0.09	0.03	0.02	0.00	0.03
Models Trained with `open-sci-ref` for 1T Tokens						
FineWeb-Edu	1T	0.20	0.03	0.00	0.00	0.06
Nemotron-CC-HQ	1T	0.13	0.03	0.01	0.04	0.05
DCLM	1T	0.15	0.03	0.00	0.01	0.05
Other Models						
SmolLM2-1.7B	11T	0.18	0.31	0.01	0.35	0.21
SmolLM2-1.7B-Instruct	11T	**0.28**	0.37	0.28	0.37	0.33

RESULTS

APPENDIX

Language evals intact, even catching up with non-permissive

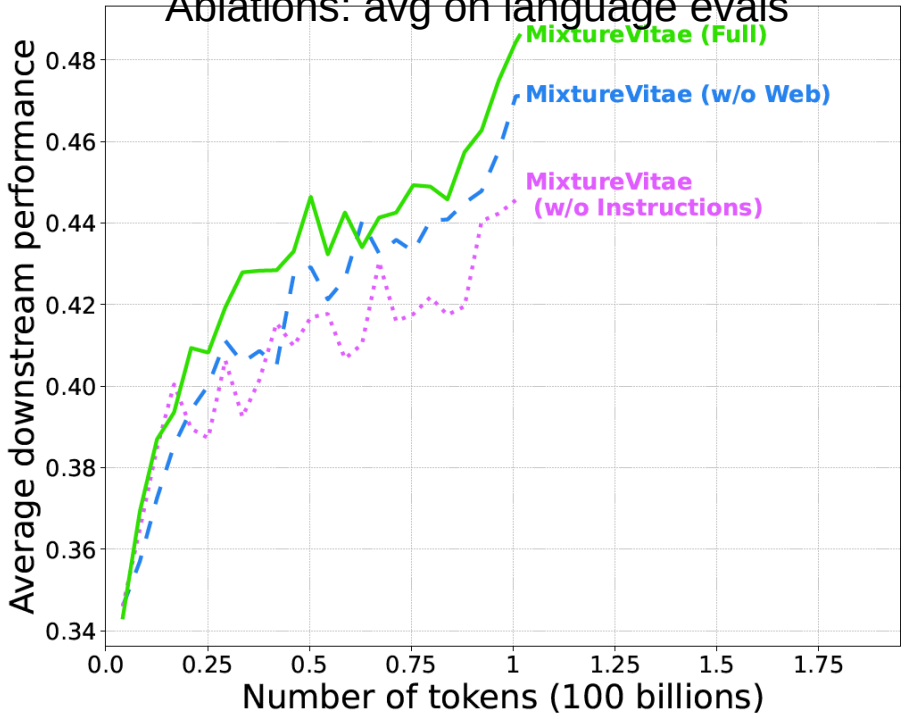
1.7B reference model trained on MixtureVitae outperforms permissive and catches up non-permissive data

Benchmark	MixtureVitae (permissive)	Comma-0.1 (permissive)	CommonCorpus (permissive)	FineWeb-Edu (mixed-license)	DCLM (mixed-license)
COPA	*0.72*	0.71	0.71	0.82	**0.89**
Lambada	0.49	*0.54*	0.49	0.56	**0.69**
OpenBookQA	*0.36*	0.33	0.31	**0.42**	0.39
Winogrande	0.59	*0.60*	0.56	0.62	**0.65**
MMLU	***0.39***	0.27	0.25	0.26	0.26
ARC-Challenge	*0.41*	0.36	0.32	**0.48**	0.44
ARC-Easy	*0.72*	0.63	0.61	**0.77**	0.75
BoolQ	***0.73***	0.62	0.62	0.65	0.72
CommonSense-QA	***0.49***	0.21	0.19	0.21	0.20
HellaSwag	*0.55*	0.53	0.45	0.67	**0.71**
PIQA	*0.71*	*0.71*	0.66	0.77	**0.78**
Average	*0.56*	0.50	0.47	0.57	**0.59**

Strong performance comes from Reasoning/Instruct subset

100B tokens ablations on 1.7B show that removing Reasoning/Instruct subset causes strong perf drop

Ablations: avg on language evals



Ablations: math/code/instruct evals

Training Dataset	IF-Eval	GSM8K	MBPP	Average
MixtureVitae	0.14	**0.47**	**0.34**	**0.25**
MixtureVitae (w/o Web)	0.18	0.41	0.33	**0.25**
MixtureVitae (w/o Instructions)	**0.19**	0.03	0.14	0.14