L11: Synthetic Data Powering Pretraining

UC Berkeley EE 194/290-16: Scalable AI

Eric W. Tramel, Ph.D.

February 24, 2026

Principal Research Scientist, NVIDIA
NeMo Data Designer

Who Am I?

Eric W. Tramel, Ph.D. — 18 Years of ML R&D

- Compressed Sensing → Statistical Physics → RBMs → Federated Learning → LLMs
- Postdoc at École Normale Sup.
- 3× Startup Veteran
- 2× Medical AI (Owkin, Unlearn)
- Edge ML @ Amazon Alexa
- Acquired by NVIDIA
- Principal RS, Synthetic Data Research (& all fun things ML)

Outline

Introduction

The Data Wall and the Scaling Imperative

“But Won’t the Models Collapse?”

Model-Amplified Data — What Works and Why

Seeding Capabilities — Synthetic Code and Reasoning

The Future — From Passive Collection to Active Curriculum Design

Introduction

“Is Pretraining Over?”

“Pre-training as we know it will unquestionably end. . . because we have but one internet.”

— Ilya Sutskever @ NeurIPS, December 2024

“Contra the popular belief that scaling is over. . . the team delivered a drastic jump. The delta between 2.5 and 3.0 is as big as we’ve ever seen. No walls in sight!”

— Oriol Vinyals (Gemini co-lead), November 2025

The Data Wall and the Scaling Imperative

Scaling Laws — The Starting Point

- Kaplan et al. (2020): power-law relationships between data, parameters, and loss — held across 7 orders of magnitude of compute
- Hoffmann et al. (2022) "Chinchilla": $D \approx 20N$ for compute-optimal dense training. A 70B model needs ∼1.4T tokens.
- Key point: these laws were derived for **dense transformers** where $P_{\text{total}} = P_{\text{active}}$

Every generation of larger models demands proportionally more training tokens.

Architecture Matters — Dense vs. MoE

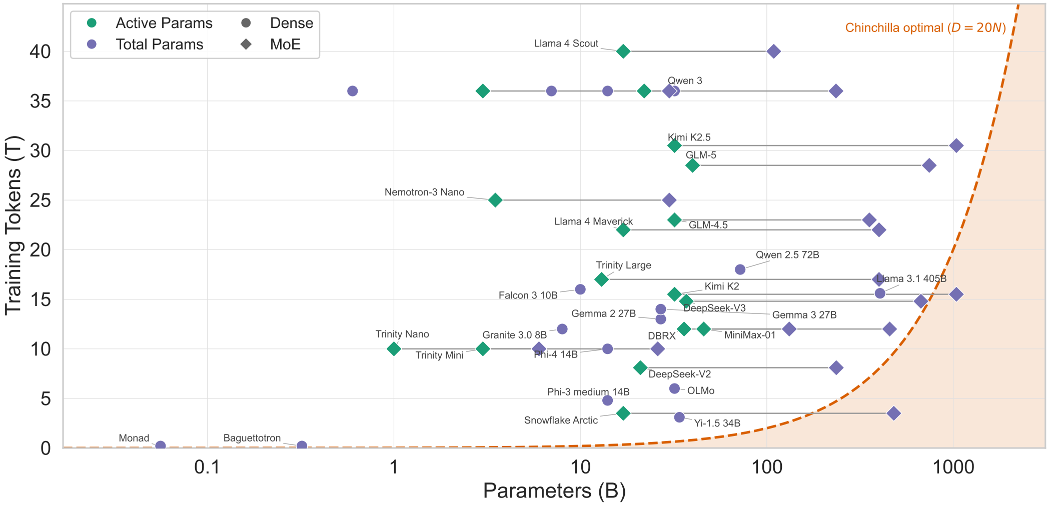
- Modern frontier models are now exclusively MoEs: inference economics.
- FLOPs per token savings: $\sim 2 \cdot P_{\text{active}}$ (inference), $\sim 6 \cdot P_{\text{active}}$ (training)
- MoE decouples P_{total} from P_{active}: complicates scaling law interpretations.
- MoE scaling laws complex (optimal shared expert count? Top-K? Passive Experts?)
- Abnar et al. (2025): optimal sparsity increases with compute budget

MoE Makes the Data Wall Worse

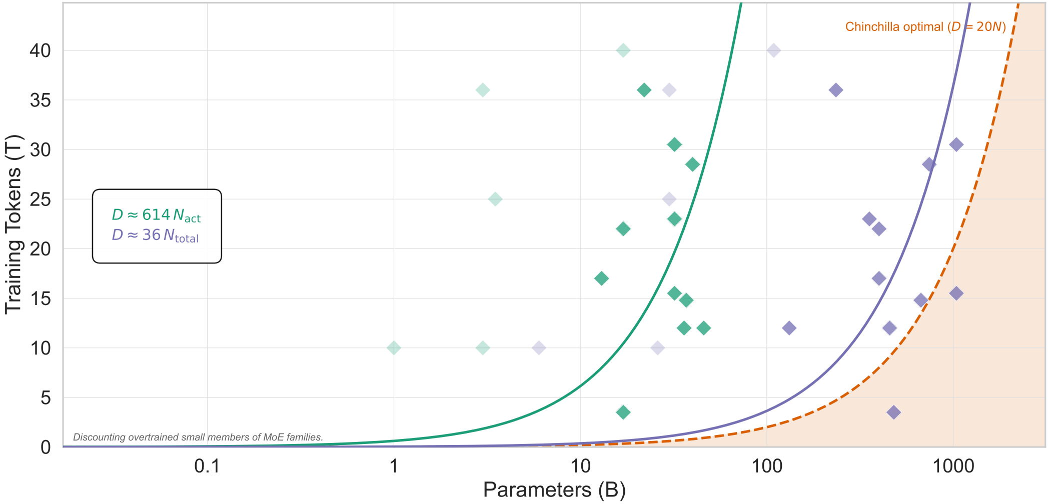
- Sparse MoE models are more prone to overfitting than dense (Fedus et al., 2022; Zoph et al., 2022)
- Each expert sees $\sim 1/E$ of training tokens under balanced routing; repeats hit harder
- MoE data needs scale with *total* parameters, not active — token requirements grow more aggressively than dense (Ludziejewski et al., 2025; G. Zhao et al., 2025)
- Practitioner experience: past ~ 2 epochs even more detrimental for MoE than dense

MoE doesn't relieve the data wall. It *accelerates* it.

The Data Wall — What Industry Actually Uses



The Data Wall — What Industry Actually Uses

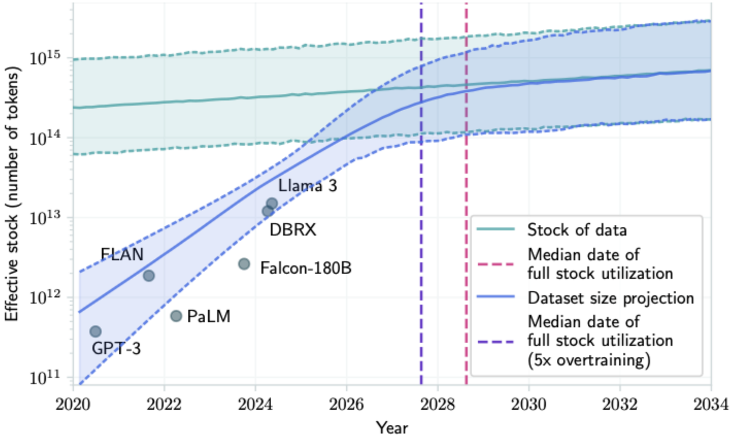


Why So Much Data?

1. **Inference cost dominates training.** Trained once, served billions of times (Sardana et al., 2024)
2. **Overtraining laws hold.** (Gadre et al., 2024): up to 640× Chinchilla, 0.7% relative error.
3. **Repetition hurts.** $\sim$4 epochs degrades dense (Muennighoff et al., 2023), $\sim$2 for MoE — you need *novel* tokens, not more passes.

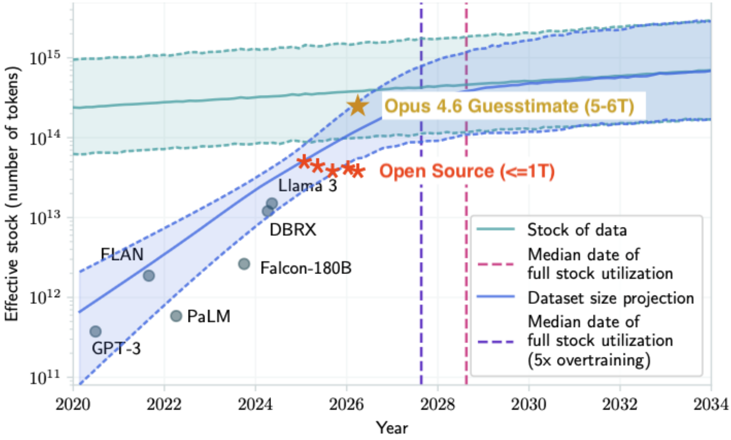
The pressure on D is structural, not optional, and you can't fake it with repetition.

The Data Stock: What Ilya Saw?

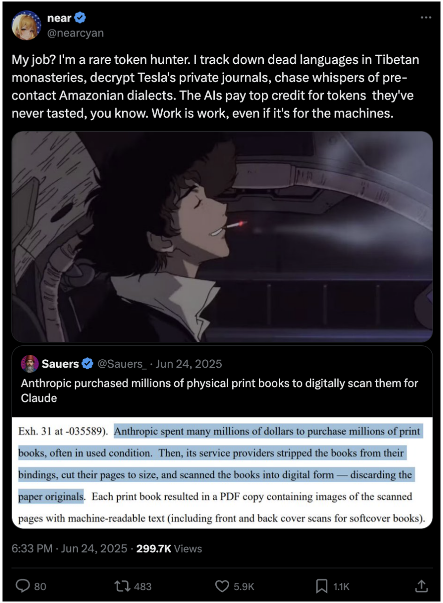


from Villalobos et al., 2024

The Data Stock: A 2026 Open-Source Update?



from Villalobos et al., 2024



near
@nearcyan

My job? I'm a rare token hunter. I track down dead languages in Tibetan monasteries, decrypt Tesla's private journals, chase whispers of pre-contact Amazonian dialects. The AIs pay top credit for tokens they've never tasted, you know. Work is work, even if it's for the machines.

Sauers @Sauers_ · Jun 24, 2025

Anthropic purchased millions of physical print books to digitally scan them for Claude

Exh. 31 at -035589). Anthropic spent many millions of dollars to purchase millions of print books, often in used condition. Then, its service providers stripped the books from their bindings, cut their pages to size, and scanned the books into digital form — discarding the paper originals. Each print book resulted in a PDF copy containing images of the scanned pages with machine-readable text (including front and back cover scans for softcover books).

6:33 PM · Jun 24, 2025 · **299.7K** Views

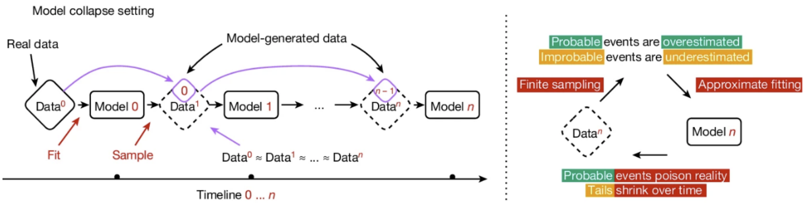
The Synthetic Lever

- Whether dense or MoE, economic pressure pushes toward **more data tokens**
- MoE makes it worse — designed to trade parameters for data
- Inference economics make overtraining *structural*, not a choice
- Synthetic Data Economic Thesis: ↑ unique human token scarcity, ↓ inference costs.

If human token costs scale harder than inference costs, you have two options: stop scaling, or **generate more data.** ***But how***?

“But Won’t the Models Collapse?”

The Model Collapse Concern



[1]

- Shumailov et al. (2024): recursive self-training loses tail diversity. Published in *Nature*.
- The failure mode: **recursive multi-generation training** where each generation *replaces* previous data
- Two symptoms: (1) loss of tail diversity, (2) convergence toward homogeneous outputs

Does this apply to what practitioners actually do?

[1] Figure 1 from Shumailov et al. (2024).

Mitigating Collapse

- **Mixing with real data prevents collapse.** (Gerstgrasser et al., 2024)
- **Rephrasing $\neq$ textbook-style.** (Kang et al., 2025): rephrased data shows no degradation; textbook-style degrades at high fractions.

Model-Amplified Data — What Works and Why

Early Forays — The Phi Series

The first serious attempts at synthetic pretraining: **construct knowledge from scratch.**

- Gunasekar et al. (2023) **Phi-1** (2023): 1.3B model trained on ∼7B synthetic textbook tokens → 50.6% HumanEval, rivaling 10× larger models
- **Phi-2** (2023): 2.7B model, ∼250B tokens (mix of synthetic + filtered web)
- Abdin et al. (2024) **Phi-4** (2024): 14B model, ∼40% synthetic (∼400B tokens). Beat GPT-4o on GPQA and MATH.
- **Cosmopedia** (Ben Allal et al., 2024): community attempt to replicate Phi-style data at scale — reached 25B tokens

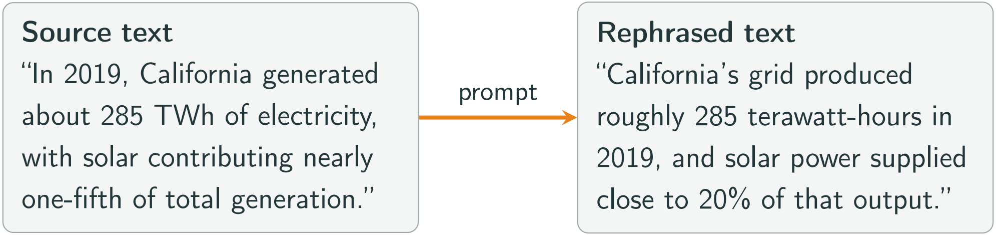
The trajectory: 7B → 250B → 400B tokens. Impressive, but still **orders of magnitude short** of the $\mathcal{O}(10T)$ frontier models need today.

The Scaling Problem with Knowledge Construction

- **Doesn't scale to trillions.** Enumerating "things a model should know" is unbounded.
- **Distribution narrowing.** Misses capabilities "tacitly expected" by users (Langlais, 2026)
- **Microsoft's own trajectory**: subsequent Phi versions returned to mixing with web data
- **Key insight**: the web already has the knowledge & long tail. The problem isn't just content, it's *presentation* and *reinforcement*.

What if instead of constructing knowledge, you just *repackaged* it?

Rephrasing in One Example



Source text

"In 2019, California generated about 285 TWh of electricity, with solar contributing nearly one-fifth of total generation."

prompt

Rephrased text

"California's grid produced roughly 285 terawatt-hours in 2019, and solar power supplied close to 20% of that output."

Facts preserved: state = California, year = 2019, amount = 285 TWh, solar share ≈ 20%

Data Augmentation: Rephrasing changes wording and structure while preserving the underlying information.

Rephrasing — Transform, Don't Replace

- Maini, Seto, et al. (2024): rephrased C4 into 4 styles via Mistral-7B → $\sim 3\times$ training speedup (85B rephrased $\approx$ 300B real-only)
- Why it scales: no need to enumerate topics; small models work; parallelizable
- Su et al. (2024) Nemotron-CC: 6.3T tokens, **1.9T synthetic**, quality-conditioned rephrasing
- Maini, Dorna, et al. (2025) BeyondWeb: per-token information density as the primary mechanism
- Scale comparison: Phi-1 = ∼7B → Nemotron-CC = 1.9T → BeyondWeb & Qwen3 = "trillions".

The Empirical Sweet Spot

- Kang et al. (2025): 1,000+ LLM variants. Pure synthetic does NOT outperform Common Crawl alone. However,

  $^1/_3$ rephrased + $^2/_3$ web data → **5–10× convergence acceleration**

Synthetic data is a **multiplier** on real data,
not a replacement for it (yet).

Small Generators, Big Returns

- Maini, Dorna, et al. (2025): 3B generators produce data statistically indistinguishable from 8B, at **2.7× lower cost**
- OLMo-2-7B (lowest benchmarks) produced the **highest-quality synthetic data**
- Four model families within 1pp of each other — generator family doesn't matter much
- **Why?** the generator's job is **reformatting, not reasoning** — a fundamentally easier task.

Today, invest in **source quality filtering upstream**, not generator size.

The Numbers — Training Efficiency Gains

Matching RedPajama @180B accuracy:

	RedPajama	BeyondWeb @8B	BeyondWeb @3B
Generator	N/A	$\sim$8B	3B
Tokens to target	180B	23.2B	23.2B
Gen FLOPs	0	3.71×10^{20}	1.39×10^{20}
Train FLOPs	8.64×10^{21}	1.11×10^{21}	1.11×10^{21}
Gen/Train	0%	33%	13%
Total FLOPs	8.64×10^{21}	1.48×10^{21}	1.25×10^{21}
vs. RedPajama	100%	17%	**14%**

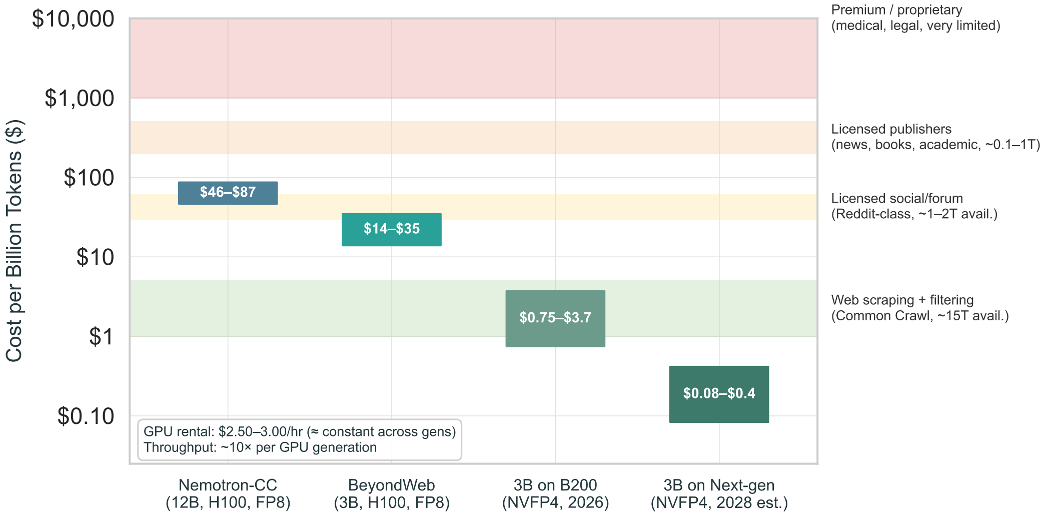
What Does Synthetic Data Cost?

The total compute budget:

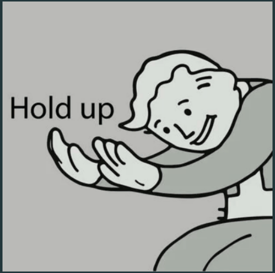
$$C_{\text{total}} = \underbrace{6 \cdot P_{\text{train}} \cdot D_{\text{train}}}_{\text{training}} + \underbrace{2 \cdot P_{\text{gen}} \cdot D_{\text{gen}}}_{\text{generation}}$$

- Concrete example: generating 1.9T tokens with 12B generator $\approx 4.6 \times 10^{22}$ FLOPs. Training 8B on 6.3T tokens $\approx 3.0 \times 10^{23}$ FLOPs. Generation is $\sim$15% of training cost.
- **Generation cost is small** — especially with small generators

The Falling Cost of Synthetic Tokens



“Great! We can just push 5T tokens through a prompt 100 times for less than $100k and scale to 15T params!”



Diversity Beats Volume

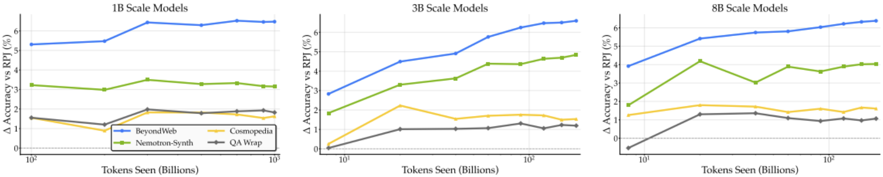
- Lin et al. (2025): diverse study strategies (timelines, concept maps, mnemonics) — 8B matched 405B Llama 3.1 on tail-fact recall
- Kimi Team (2025) Kimi K2:

$$10 \text{ rephrasings} \times 1 \text{ epoch} \geq 1 \text{ rephrasing} \times 10 \text{ epochs}$$

- Maini, Dorna, et al. (2025): single-strategy learning curves **flatten during training**. Multi-strategy does not.

More tokens in the same format saturate.
Diverse generation strategies maintain learning signal.

Evidence from BeyondWeb



- RQ1: Summarization ≈ Cosmopedia (46.7% vs. 47.1%)
- RQ2: Naive continuation = 46.2% (ceiling). BeyondWeb = 50.4% (**+4.2pp above ceiling**)
- RQ5: Multi-strategy maintains positive slopes at 50× beyond Chinchilla-optimal.

Vary the **transformation**, not just the input.

The Bitter Lesson of Synthetic Pretraining

- **Elaborate approach** (Phi-style): frontier generators, curriculum design, quality classifiers. High cost/token. Doesn't scale past hundreds of billions.
- **Simple approach** (rephrasing): small cheap models, straightforward prompts, process the entire web. Scales to trillions.
- **The data says the simple approach wins.**

The Open SoTA recipe: good source data + cheap diverse rephrasing + principled mixing.

The Compute is Coming

If inference costs are falling ∼10× per GPU generation:

GPU Generation	Cost / Billion Tokens	vs. H100
H100 (2024, FP8)	\$14–\$35	1×
B200 (2026, NVFP4)	\$0.75–\$3.70	∼10× cheaper
Next-gen (2028, NVFP4)	\$0.08–\$0.40	∼100× cheaper

Rephrasing exploits cheap inference. As inference gets **OOMs cheaper**, what *new* data methodologies become viable?

Seeding Capabilities — Synthetic Code and Reasoning

Beyond Rephrasing — Seeding Capabilities

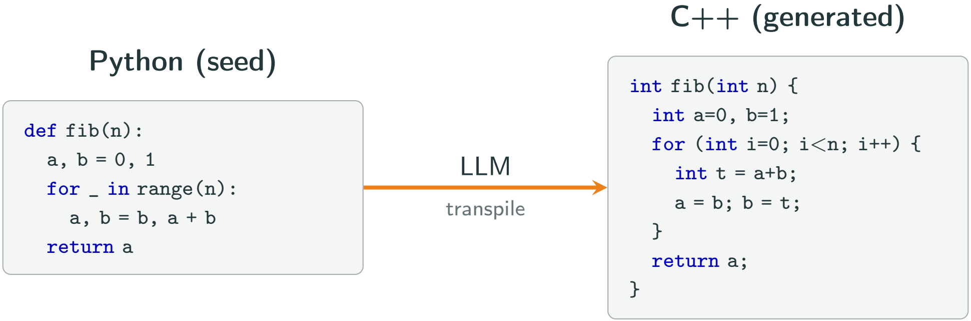
- Rephrasing works for general NL — but what about capabilities that barely exist in web text?
- Purpose-built synthetic data mixed into **bulk pretraining**, not SFT — capabilities seeded here are more durable than those added in post-training
- Two domains with strong evidence: **code** (verifiable) and **reasoning traces**

Synthetic Code in Pretraining

- Abdin et al. (2024) Phi-4: instruction reversal — generate problem from code, keep matching pairs
- NVIDIA (2025b) Nemotron-Code-v2: 340B tokens, 5 synthetic techniques (QA, code review, pedagogy, style rewriting, transpilation)
- Hui et al. (2024) Qwen2.5-Coder: 5.5T tokens, self-bootstrapping + Text-Code Grounding (+5.2pp on HumanEval+MBPP)
- Fujii et al. (2025) SwallowCode: **+17pp on HumanEval** vs. quality-filtered code. **Rewriting beats filtering.**

Code's structural advantage: you can **verify** synthetic code via execution.

Example: Code Transpilation



Python (seed)

```python
def fib(n):
  a, b = 0, 1
  for _ in range(n):
    a, b = b, a + b
  return a
```

LLM

transpile

C++ (generated)

```cpp
int fib(int n) {
  int a=0, b=1;
  for (int i=0; i<n; i++) {
    int t = a+b;
    a = b; b = t;
  }
  return a;
}
```

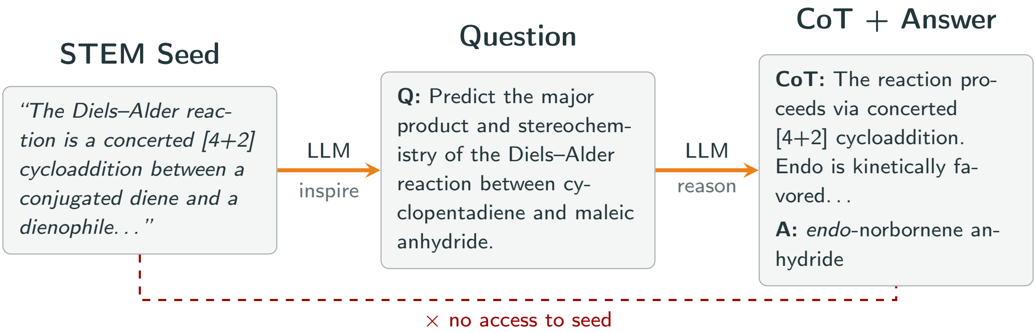
One of 5 code synthesis strategies in NVIDIA (2025b) (340B tokens). Both versions become pretraining data — model learns cross-language structure. Key advantage: synthetic code pairs can be **verified via lint & execution**.

Front-Loading Reasoning

- Akter et al. (2025): 8B model, 1T tokens. 80% general + **20% reasoning data** (17% code, 56% math, 27% science)
- Code: **+9pp over baseline** in pretraining (40.89% → 49.94%)
- Gains **compound through post-training**: SFT code 7.09% → 16.75%, RL LiveCodeBench 13.16% → 32.43%
- **The compounding effect is the key finding.** Reasoning patterns in pretraining serve as scaffolding for SFT and RL.

Curriculum design principle: **what you put in pretraining shapes what post-training can build on.**

Example: Synthetic Reasoning Data (RQA)



STEM Seed

“The Diels–Alder reaction is a concerted [4+2] cycloaddition between a conjugated diene and a dienophile. . . ”

Question

Q: Predict the major product and stereochemistry of the Diels–Alder reaction between cyclopentadiene and maleic anhydride.

CoT + Answer

CoT: The reaction proceeds via concerted [4+2] cycloaddition. Endo is kinetically favored. . .

A: *endo*-norbornene anhydride

NVIDIA (2025a): decoupled generation forces reasoning from knowledge, not extraction. 9M samples across 8 STEM domains; at 4–8% of mix, lifts MMLU-Pro, Math-500, **and** HumanEval.

The Future — From Passive Collection to Active Curriculum Design

Memory → Logic → Simulation

Langlais (2026) framework:

1. **Memory** — rephrasing for knowledge retention. Mature: BeyondWeb, Nemotron-CC.
2. **Logic** — reasoning primitives. Emerging: Front-Loading (+9% code), RLP (Hatamizadeh et al., 2025) (+19% math/science).
3. **Simulation** — model entire environments. Frontier: only Agentic CPT so far.

Unifying insight: **passive data collection → active curriculum design.**

Echo Chamber (R. Zhao et al., 2025): RL amplifies patterns already in pretraining ⇒ **pretraining determines the ceiling.**

The Direction Is Clear

- **Web-only pretraining is hitting limits, but pretraining itself is being transformed.**
- The field has moved from passive data collection to active curriculum design.
- **Today**: filtered web + diverse rephrasing + small generators + principled mixing.
- **Tomorrow**: reasoning-enriched pretraining, multimodal synthetic data, agent trajectories.
- Goal: Maximize compute to optimize the pretraining environment, shifting costs from training dense fact-machines to lighter, adaptive "cognitive core" (Karpathy, 2025).

Synthetic pretraining is the way frontier models are built.

Thoughts? Feedback? Corrections?

Eric W. Tramel: etramel@nvidia.com

http://eric-tramel.github.io

References i

- Abdin, Marah et al. (2024). **“Phi-4 Technical Report”.** In: *arXiv preprint arXiv:2412.08905*. DOI: 10.48550/arXiv.2412.08905. eprint: 2412.08905. URL: https://arxiv.org/abs/2412.08905.
- Abnar, Samira et al. (2025). **“Parameters vs FLOPs: Scaling Laws for Optimal Sparsity for Mixture-of-Experts Language Models”.** In: *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. DOI: 10.48550/arXiv.2501.12370. eprint: 2501.12370. URL: https://proceedings.mlr.press/v267/abnar25a.html.

References ii

- Akter, Syeda Nahida et al. (2025). **“Front-Loading Reasoning: The Synergy between Pretraining and Post-Training Data”.** In: *arXiv preprint arXiv:2510.03264*. DOI: 10.48550/arXiv.2510.03264. arXiv: 2510.03264 [cs.LG]. URL: https://arxiv.org/abs/2510.03264.
- Ben Allal, Loubna et al. (Feb. 2024). ***Cosmopedia*.** URL: https://huggingface.co/datasets/HuggingFaceTB/cosmopedia.
- Fedus, William et al. (2022). **“Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity”.** In: *Journal of Machine Learning Research* 23.120, pp. 1–39. DOI: 10.48550/arXiv.2101.03961. eprint: 2101.03961. URL: https://jmlr.org/papers/v23/21-0998.html.

References iii

- Fujii, Kazuki et al. (2025). **"Rewriting Pre-Training Data Boosts LLM Performance in Math and Code".** In: *arXiv preprint arXiv:2505.02881*. DOI: 10.48550/arXiv.2505.02881. arXiv: 2505.02881 [cs.CL]. URL: https://arxiv.org/abs/2505.02881.
- Gadre, Samir Yitzhak et al. (2024). **"Language models scale reliably with over-training and on downstream tasks".** In: *arXiv preprint arXiv:2403.08540*. DOI: 10.48550/arXiv.2403.08540. eprint: 2403.08540. URL: https://arxiv.org/abs/2403.08540.

References iv

- Gerstgrasser, Matthias et al. (2024). **"Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data".** In: *arXiv preprint arXiv:2404.01413*. DOI: 10.48550/arXiv.2404.01413. eprint: 2404.01413. URL: https://arxiv.org/abs/2404.01413.
- Gunasekar, Suriya et al. (2023). **"Textbooks Are All You Need".** In: *arXiv preprint arXiv:2306.11644*. DOI: 10.48550/arXiv.2306.11644. eprint: 2306.11644. URL: https://arxiv.org/abs/2306.11644.

References v

- Hatamizadeh, Ali et al. (2025). “**RLP: Reinforcement as a Pretraining Objective**”. In: *arXiv preprint arXiv:2510.01265*. DOI: 10.48550/arXiv.2510.01265. eprint: 2510.01265. URL: https://arxiv.org/abs/2510.01265.
- Hoffmann, Jordan et al. (2022). “**Training Compute-Optimal Large Language Models**”. In: *arXiv preprint arXiv:2203.15556*. DOI: 10.48550/arXiv.2203.15556. eprint: 2203.15556. URL: https://arxiv.org/abs/2203.15556.
- Hui, Binyuan et al. (2024). “**Qwen2.5-Coder Technical Report**”. In: *arXiv preprint arXiv:2409.12186*. DOI: 10.48550/arXiv.2409.12186. eprint: 2409.12186. URL: https://arxiv.org/abs/2409.12186.

References vi

- Kang, Feiyang et al. (2025). “**Demystifying Synthetic Data in LLM Pre-training: A Systematic Study of Scaling Laws, Benefits, and Pitfalls**”. In: *arXiv preprint arXiv:2510.01631*. DOI: 10.48550/arXiv.2510.01631. eprint: 2510.01631. URL: https://arxiv.org/abs/2510.01631.
- Kaplan, Jared et al. (2020). “**Scaling Laws for Neural Language Models**”. In: *arXiv preprint arXiv:2001.08361*. DOI: 10.48550/arXiv.2001.08361. eprint: 2001.08361. URL: https://arxiv.org/abs/2001.08361.

References vii

- Karpathy, Andrej (Oct. 2025). ***Andrej Karpathy: AGI Is Still a Decade Away***. Dwarkesh Podcast interview (hosted by Dwarkesh Patel). URL: https://www.dwarkesh.com/p/andrej-karpathy.
- Kimi Team (2025). "**Kimi K2: Open Agentic Intelligence**". In: *arXiv preprint arXiv:2507.20534*. DOI: 10.48550/arXiv.2507.20534. arXiv: 2507.20534 [cs.LG]. URL: https://arxiv.org/abs/2507.20534.
- Langlais, Pierre-Carl (Feb. 2026). ***Synthetic Pretraining***. Vintage Data blog. Retrieved February 2026. URL: https://vintagedata.org/blog/posts/synthetic-pretraining.

References viii

- Lin, Jessy et al. (2025). **“Learning Facts at Scale with Active Reading”.** In: *arXiv preprint arXiv:2508.09494*. Meta WikiExpert model and 1T token dataset released. DOI: 10.48550/arXiv.2508.09494. eprint: 2508.09494. URL: https://arxiv.org/abs/2508.09494.
- Ludziejewski, Jan et al. (2025). **“Joint MoE Scaling Laws: Mixture of Experts Can Be Memory Efficient”.** In: *Proceedings of the 42nd International Conference on Machine Learning*. DOI: 10.48550/arXiv.2502.05172. eprint: 2502.05172. URL: https://proceedings.mlr.press/v267/ludziejewski25a.html.

References ix

- Maini, Pratyush, Vineeth Dorna, et al. (2025). “**BeyondWeb: Lessons from Scaling Synthetic Data for Trillion-scale Pretraining**”. In: *arXiv preprint arXiv:2508.10975*. DOI: 10.48550/arXiv.2508.10975. arXiv: 2508.10975 [cs.CL]. URL: https://arxiv.org/abs/2508.10975.
- Maini, Pratyush, Skyler Seto, et al. (2024). “**Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling**”. In: *arXiv preprint arXiv:2401.16380*. DOI: 10.48550/arXiv.2401.16380. eprint: 2401.16380. URL: https://arxiv.org/abs/2401.16380.

References x

- Muennighoff, Niklas et al. (2023). **“Scaling Data-Constrained Language Models”.** In: *Advances in Neural Information Processing Systems*. Vol. 36. DOI: 10.48550/arXiv.2305.16264. arXiv: 2305.16264. URL: https://papers.nips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf.
- NVIDIA (Dec. 2025a). ***Nemotron 3 Nano: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning*.** Tech. rep. arXiv preprint arXiv:2512.20848. NVIDIA Corporation. DOI: 10.48550/arXiv.2512.20848. eprint: 2512.20848. URL: https://arxiv.org/abs/2512.20848.

References xi

- NVIDIA (2025b). ***Nemotron-Pretraining-Code-v2.*** Hugging Face. URL: https://huggingface.co/datasets/nvidia/Nemotron-Pretraining-Code-v2.

- Sardana, Nikhil et al. (2024). **"Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws".** In: *arXiv preprint arXiv:2401.00448*. DOI: 10.48550/arXiv.2401.00448. eprint: 2401.00448. URL: https://arxiv.org/abs/2401.00448.

- Shumailov, Ilia et al. (2024). **"AI models collapse when trained on recursively generated data".** In: *Nature* 631, pp. 755–759. DOI: 10.1038/s41586-024-07566-y. URL: https://doi.org/10.1038/s41586-024-07566-y.

References xii

- Su, Dan et al. (2024). **"Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset".** In: *arXiv preprint arXiv:2412.02595*. DOI: `10.48550/arXiv.2412.02595`. eprint: `2412.02595`. URL: `https://arxiv.org/abs/2412.02595`.
- Villalobos, Pablo et al. (21–27 Jul 2024). **"Position: Will we run out of data? Limits of LLM scaling based on human-generated data".** In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 52213–52249. URL: `https://proceedings.mlr.press/v235/villalobos24a.html`.

References xiii

- Zhao, Guoliang et al. (2025). **“Towards a Comprehensive Scaling Law of Mixture-of-Experts”.** In: *arXiv preprint arXiv:2509.23678*. DOI: 10.48550/arXiv.2509.23678. eprint: 2509.23678. URL: https://arxiv.org/abs/2509.23678.
- Zhao, Rosie et al. (2025). **“Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining”.** In: *arXiv preprint arXiv:2504.07912*. DOI: 10.48550/arXiv.2504.07912. arXiv: 2504.07912 [cs.LG]. URL: https://arxiv.org/abs/2504.07912.

References xiv

Zoph, Barret et al. (2022). **“ST-MoE: Designing Stable and Transferable Sparse Expert Models”.** In: *arXiv preprint arXiv:2202.08906*. DOI: 10.48550/arXiv.2202.08906. eprint: 2202.08906. URL: https://arxiv.org/abs/2202.08906.