Pushing the Frontiers of Foundation Models
(in NLP)

Sewon Min
sewonmin.com
Berkeley UNIVERSITY OF CALIFORNIA BAIR Ai2

Context of this talk

Goal: “design an AI model that is **performant** and **general**”

- Performant: **“depth”** of intelligence
- General:**“breath”** of intelligence
- We’ll not use the term “AGI” as it’s often a marketing term
- Care about “capabilities” as a function of an AI model — we don’t care about “humanlike” or “consciousness”

Where are we? (from GDM)

	Narrow	General
Emerging Equal to or somewhat better than an unskilled human	Emerging Narrow AI	Emerging General AI
Competent At least 50th percentile of skilled adults	Competent Narrow AI	Competent General AI
Expert At least 90th percentile of skilled adults	Expert Narrow AI	Expert General AI
Superhuman Outperforms 99.9th percentile of humans	Superhuman Narrow AI	Superhuman General AI

Morris et al. 2025. "Position: Levels of AGI for Operationalizing Progress on the Path to AGI"

Where are we? (from GDM)

	Narrow	General
Emerging Equal to or somewhat better than an unskilled human	Emerging Narrow AI GOFAI (Boden, 2014); SHRDLI (Winograd, 1971)	Emerging General AI ChatGPT, Bard, Llama 2, Gemini
Competent At least 50th percentile of skilled adults	Competent Narrow AI Siri, Alexa, or Google Assistant, IBM Watson	Competent General AI **Not yet achieved**
Expert At least 90th percentile of skilled adults	Expert Narrow AI Spelling & grammer checkers, Dall-E 2	Expert General AI **Not yet achieved**
Superhuman Outperforms 99.9th percentile of humans	Superhuman Narrow AI AlphaGo, AlphaFold, AlphaZero	Superhuman General AI **Not yet achieved**

Morris et al. 2025. "Position: Levels of AGI for Operationalizing Progress on the Path to AGI"

Where are we? (from GDM)

	Narrow	General
Emerging Equal to or somewhat better than an unskilled human	Emerging Narrow AI GOFAI (Boden, 2014); SHRDLI (Winograd, 1971)	Emerging General AI ChatGPT, Bard, Llama 2, Gemini
Competent At least 50th percentile of skilled adults	Competent Narrow AI Siri, Alexa, or Google Assistant, IBM Watson	Competent General AI **Not yet achieved**
Expert At least 90th percentile of skilled adults	Expert Narrow AI Spelling & grammer checkers, Dall-E 2	Expert General AI **Not yet achieved**
Superhuman Outperforms 99.9th percentile of humans	Superhuman Narrow AI AlphaGo, AlphaFold, AlphaZero	Superhuman General AI **Not yet achieved**

Morris et al. 2025. "Position: Levels of AGI for Operationalizing Progress on the Path to AGI"

Where are we? (from GDM)

	Narrow	General
Emerging Equal to or somewhat better than an unskilled human	Emerging Narrow AI GOFAI (Boden, 2014); SHRDLI (Winograd, 1971)	Emerging General AI ChatGPT, Bard, Llama 2, Gemini
Competent At least 50th percentile of skilled adults	Competent Narrow AI Siri, Alexa, or Google Assistant, IBM Watson	Competent General AI **Not yet achieved** **(Prediction: Will achieve in ~5 years)**
Expert At least 90th percentile of skilled adults	Expert Narrow AI Spelling & grammer checkers, Dall-E 2	Expert General AI **Not yet achieved (???)**
Superhuman Outperforms 99.9th percentile of humans	Superhuman Narrow AI AlphaGo, AlphaFold, AlphaZero	Superhuman General AI **Not yet achieved (???)**

If we are not there yet after 5 years, why would that be?

What would hold us back?

1. **Scaling laws don't extend** as continuously as once expected
2. Scaling laws do extend continuously, but **are prohibitively expensive** (cost, power, environmental harms)
3. **Data scaling laws slow down** because high-value datasets are proprietary and fragmented (no single entity owns all the data)
4. **Non-technical limits**, e.g., policy, regulation, and geopolitical tensions

Claim: NLP/ML researchers should work on solutions that matter 5+ years from now, not the next 5 months.

What would make us overcome these barriers?

Three Key Trends Ahead

1. Scaling laws don’t extend as continuously as once expected
2. Scaling laws do extend continuously, but are prohibitively expensive (cost, power, environmental harms)
3. Data scaling laws slow down because high-value datasets are proprietary and fragmented (no single entity owns all the data)
4. Non-technical limits, e.g., policy, regulation, and geopolitical tensions

- Learning from limited data
  (for problems 1, 2, and 3)
- More end-to-end
  (for problems 1 and 2)
- Breaking end-to-end
  (for problems 2 and 3)

Three Key Trends Ahead

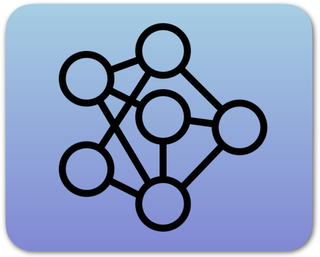
1. Scaling laws don't extend as continuously as once expected
2. Scaling laws do extend continuously, but are prohibitively expensive (cost, power, environmental harms)
3. Data scaling laws slow down because high-value datasets are proprietary and fragmented (no single entity owns all the data)
4. Non-technical limits, e.g., policy, regulation, and geopolitical tensions

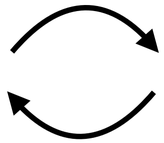
- Learning from limited data
  (for problems 1, 2, and 3)
- More end-to-end
  (for problems 1 and 2)
- Breaking end-to-end
  (for problems 2 and 3)

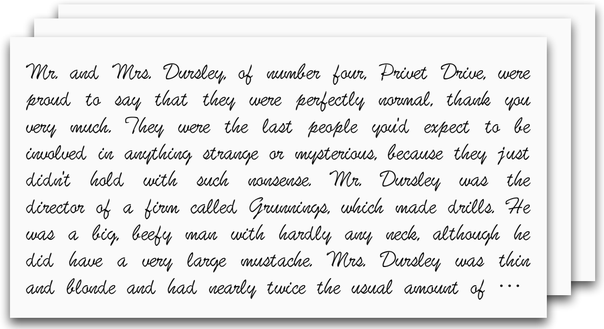
Current approaches

Billions—trillions of words

Current approaches





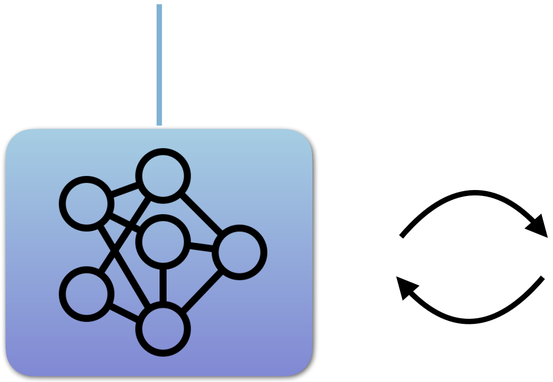


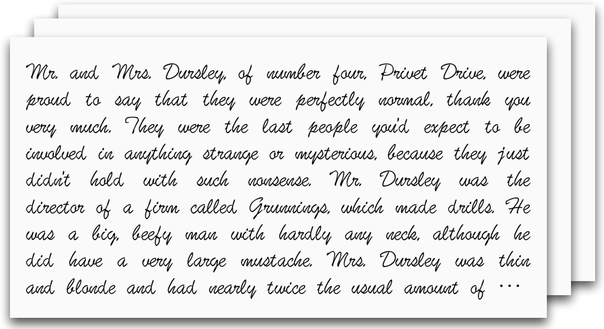
10+ billion parameters

Billions—trillions of words

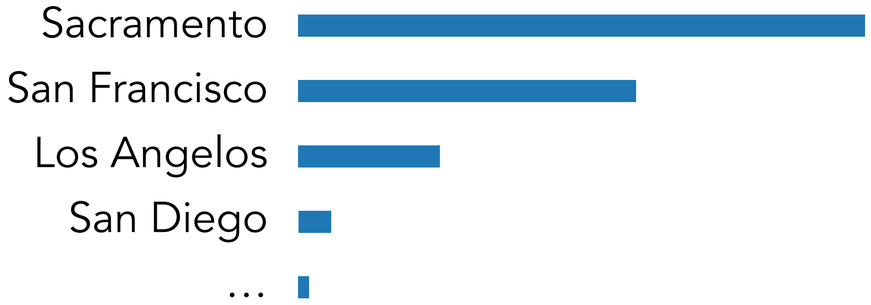
Current approaches

The capital city of California is



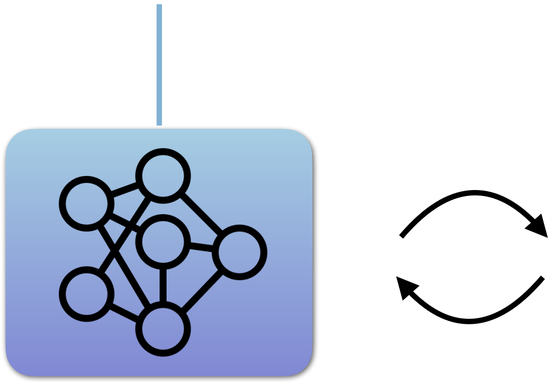


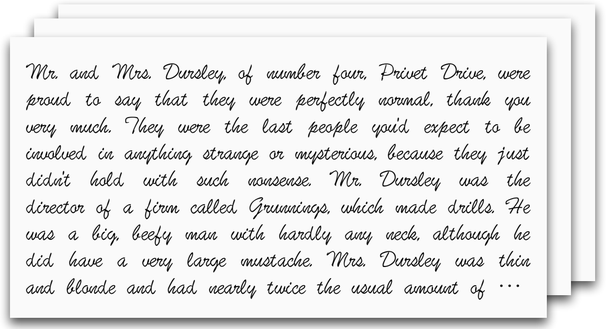
Billions—trillions of words



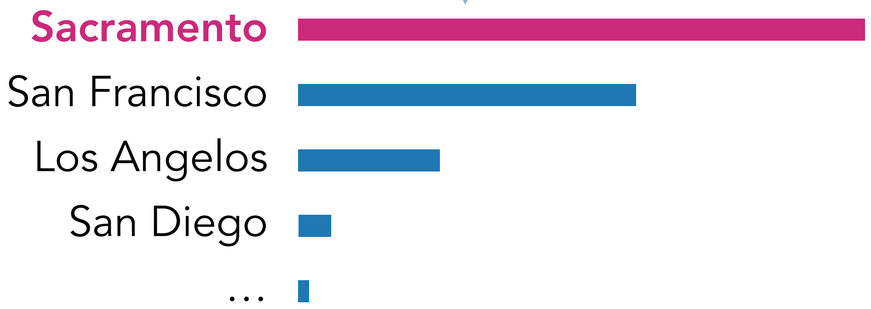
Current approaches

The capital city of California is

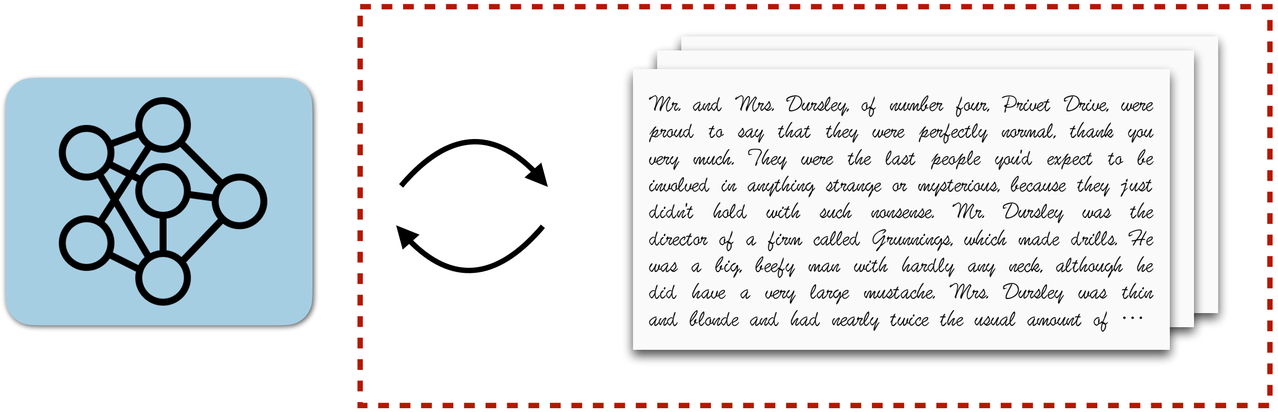




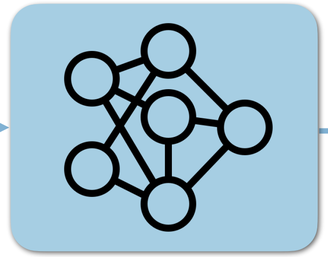
Billions—trillions of words



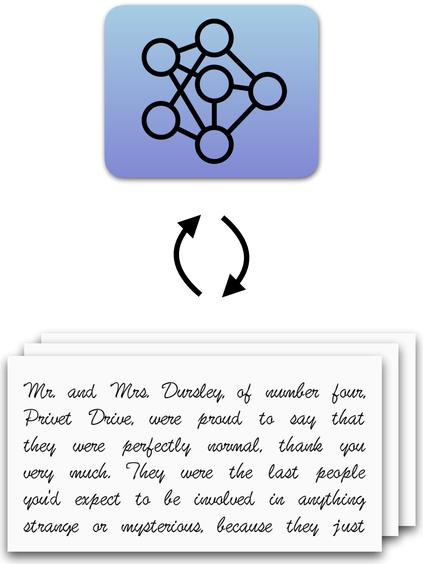
Current approaches



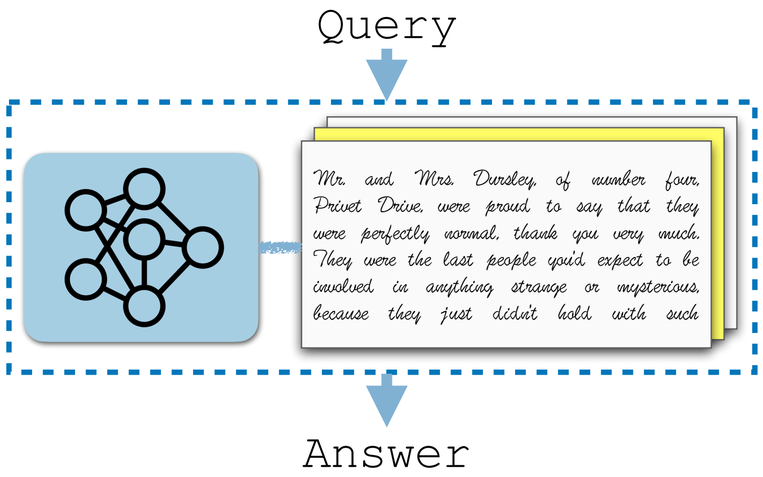
Current approaches

Q: What's the capital city of California?  A: Sacramento

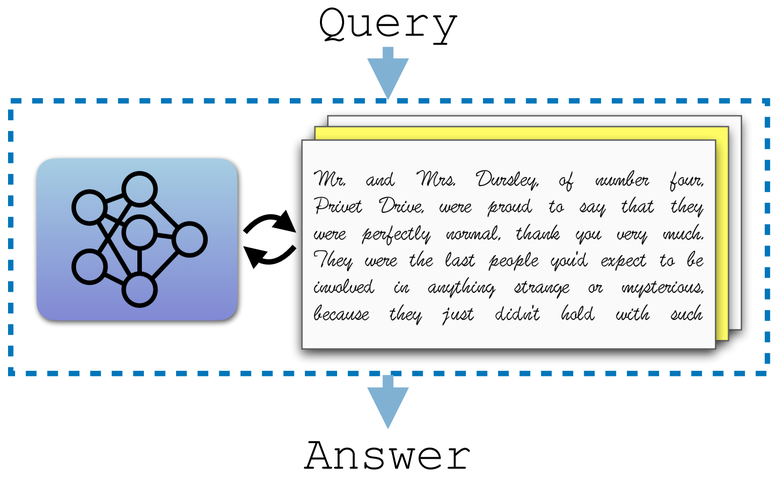
Claim: We Should Re-Use the Data as Much as Possible



Training

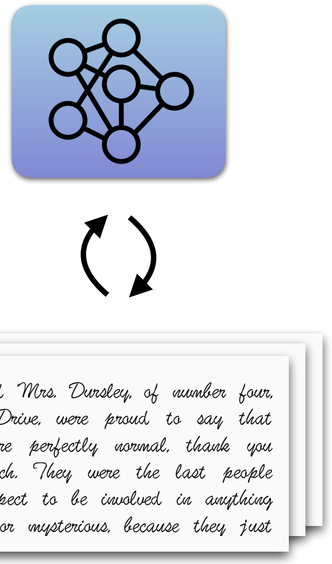


Test-time Inference
(Retrieval)



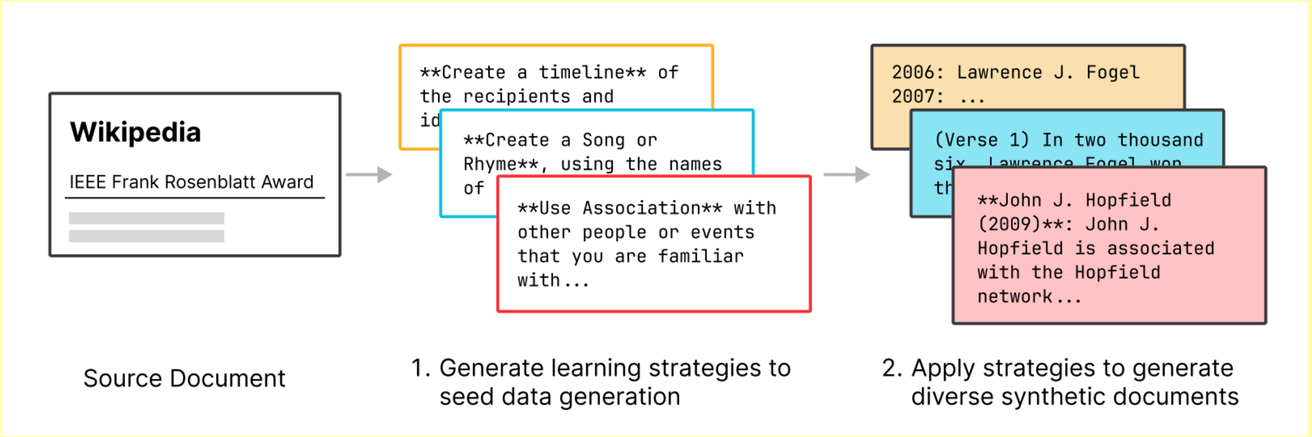
Test-time Training

Claim: We Should Re-Use the Data as Much as Possible

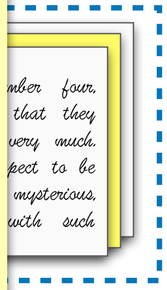


Training

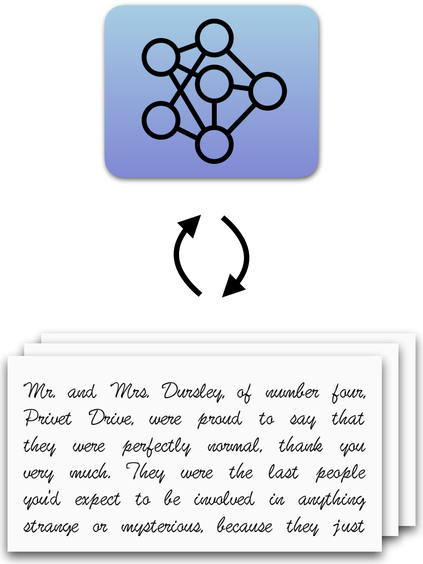
Goal: Establish new scaling laws with repeated data & synthetically generated data



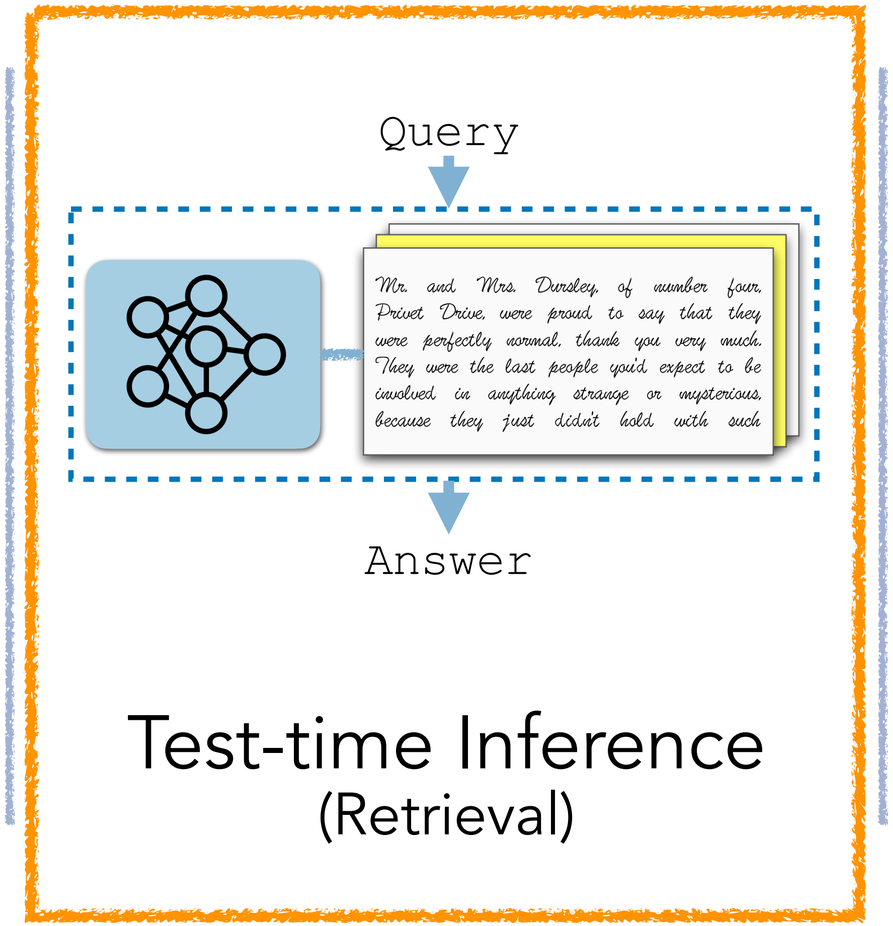
- Muennighoff et al. 2023. "Scaling Data-Constrained Language Models"
- Kim et al. 2025. "Pre-training under infinite compute"
- Lin et al. 2025. "Learning Facts at Scale with Active Reading"



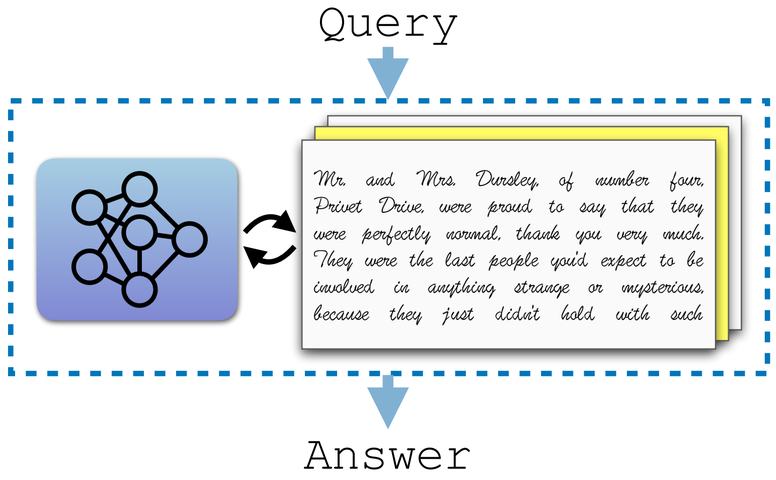
Claim: We Should Re-Use the Data as Much as Possible



Training



Test-time Inference
(Retrieval)

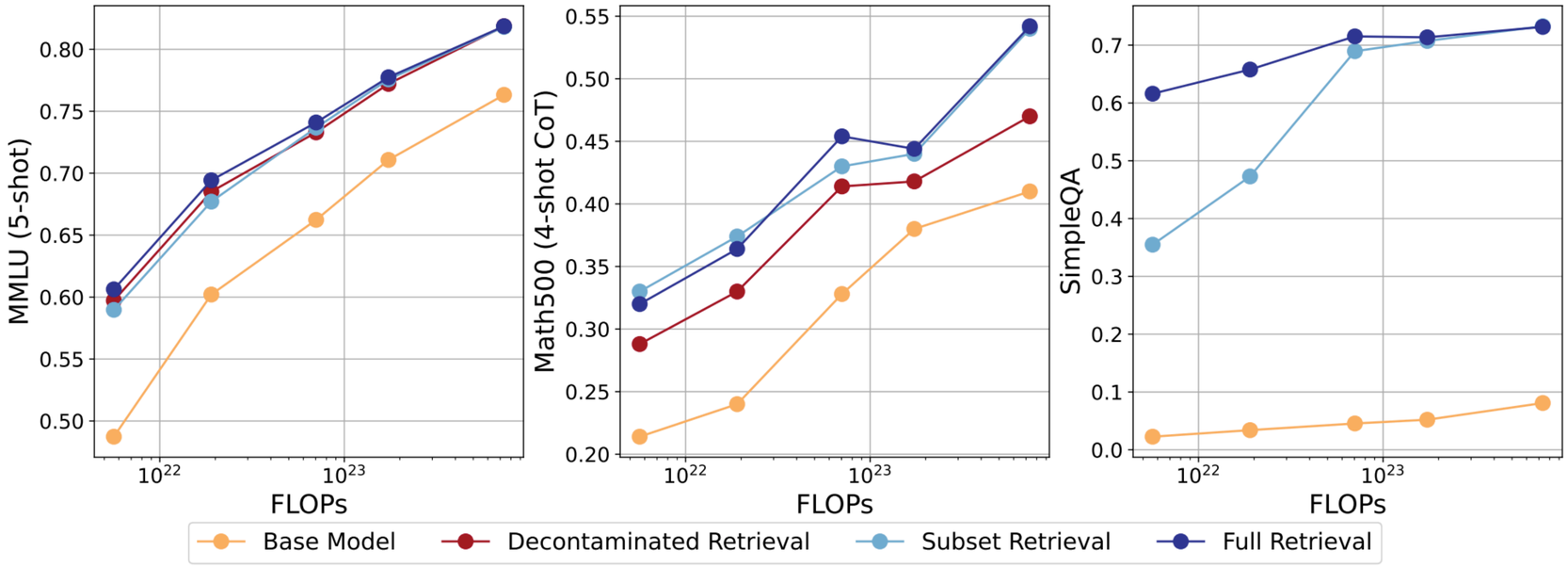


Test-time Training

Claim: We Should Re-Use the Data as Much as Possible

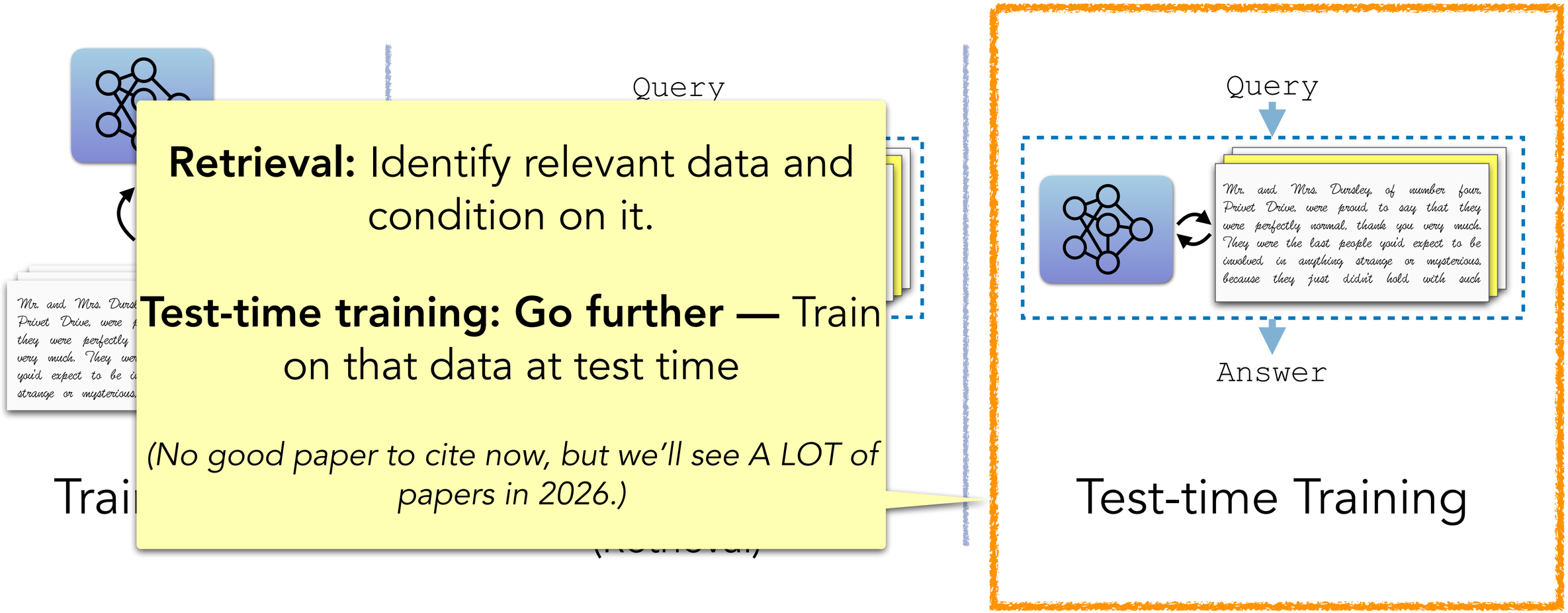
"[R]etrieval acts as a ~5x compute multiplier versus pre-training alone."

— Fang et al. 2025



- Shao et al. 2024. "Scaling Retrieval-Based Language Models with a Trillion-Token Datastore"
- Lyu et al. 2025. "Frustratingly Simple Retrieval Improves Challenging, Reasoning-Intensive Benchmarks"
- Fang et al. 2025. "Reusing Pre-Training Data at Test Time is a Compute Multiplier"

Claim: We Should Re-Use the Data as Much as Possible



Retrieval: Identify relevant data and condition on it.

Test-time training: Go further — Train on that data at test time

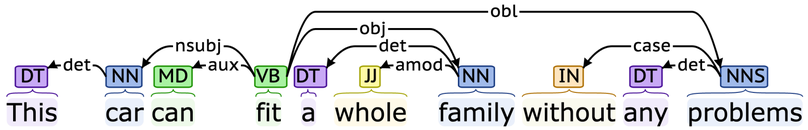
(No good paper to cite now, but we'll see A LOT of papers in 2026.)

Three Key Trends Ahead

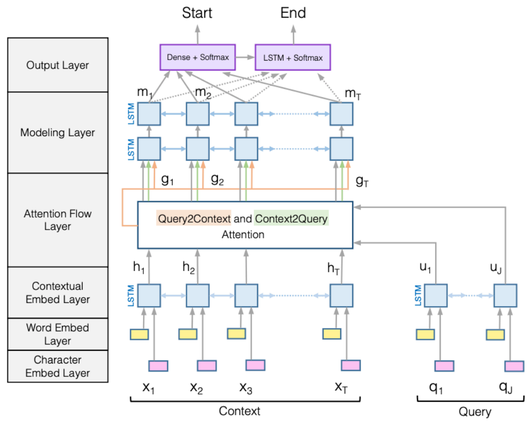
1. Scaling laws don't extend as continuously as once expected
2. Scaling laws do extend continuously, but are prohibitively expensive (cost, power, environmental harms)
3. Data scaling laws slow down because high-value datasets are proprietary and fragmented (no single entity owns all the data)
4. Non-technical limits, e.g., policy, regulation, and geopolitical tensions

- Learning from limited data
  (for problems 1, 2, and 3)
- More end-to-end
  (for problems 1 and 2)
- Breaking end-to-end
  (for problems 2 and 3)

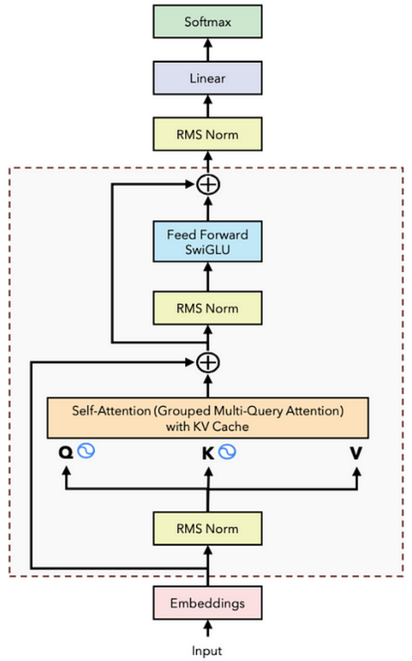
Trends in AI Always Have Shifted Toward More End-to-End



Pre-2014..?
I don't know



2014 - 2020

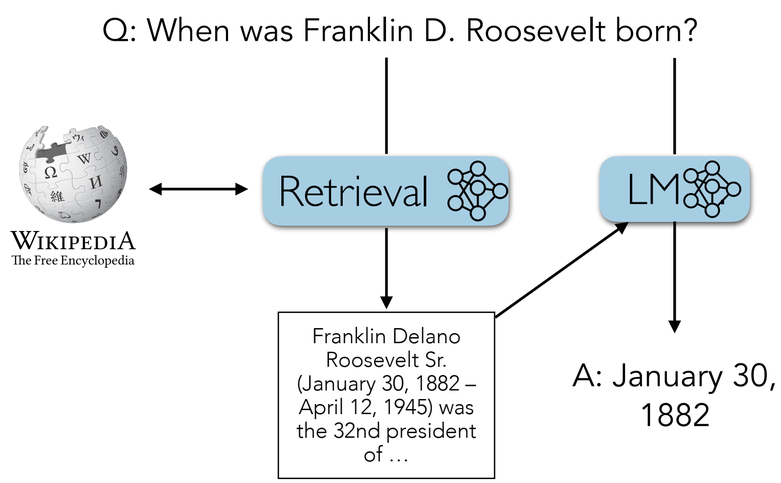


Now

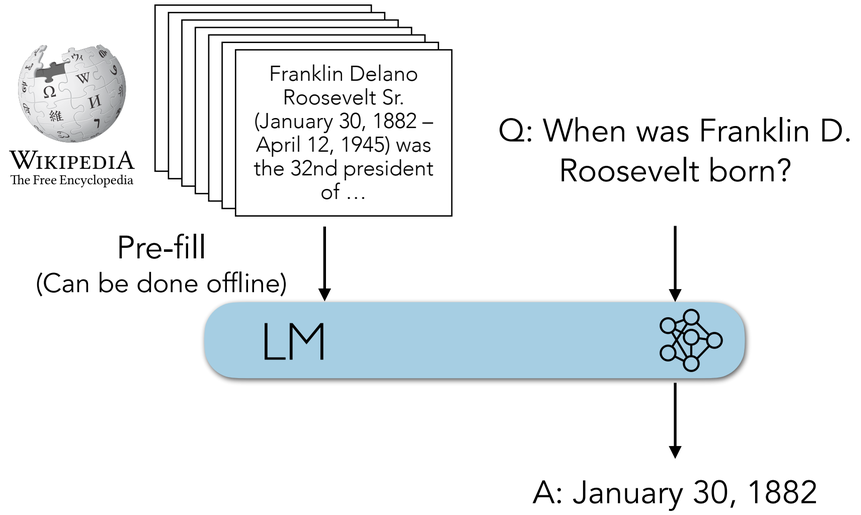
Claim: We're actually not sufficiently end-to-end.
Being more end-to-end will improve scaling laws.

How to Make it Even More End-to-End?

(1) Removing Stages in RAG



Current RAG



End-to-end RAG

(More end-to-end & more suitable when a task requires reasoning across a large chunk from the corpus)

- Lee et al. 2024. "Can Long-Context Language Models Subsume Retrieval, RAG, SQL, and More?"
- Gupta et al. 2025. "Scalable In-context Ranking with Generative Models"

How to Make it Even More End-to-End?
(2) Removing Tokenizer

Word-level: `Berkeley` `is` `a` `city` `on` `the` `eastern` `shore` `of` `San` `Francisco` `Bay`

Subword-level *(Current dominant)*: `Ber` `keley` `is` `a` `city` `on` `the` `east` `ern` `shore` `of` `San` `Fran` `cisco` `Bay`

Char-level: `B` `e` `r` `k` `e` `l` `e` `y` `i` `s` `a` `c` `i` `t` `y` `o` `n` `t` `h` `e` …

Byte-level: `0` `1` `1` `0` `1` `1` `0` `0` `0` `1` `0` `0` `1` `1` `1` `0` `1` `0` `0` `1` `1` `0` `0` `0` `1` `0` `0` `1` `1` `1` `0` `1` `0` `0` `1` …

- Yu et al. 2023. "MEGABYTE: Predicting Million-byte Sequences with Multiscale Transformers”
- Pagnoni et al. 2025. "Byte Latent Transformer: Patches Scale Better Than Tokens”

How to Make it Even More End-to-End?

(3) Removing Text Extraction (OCR)

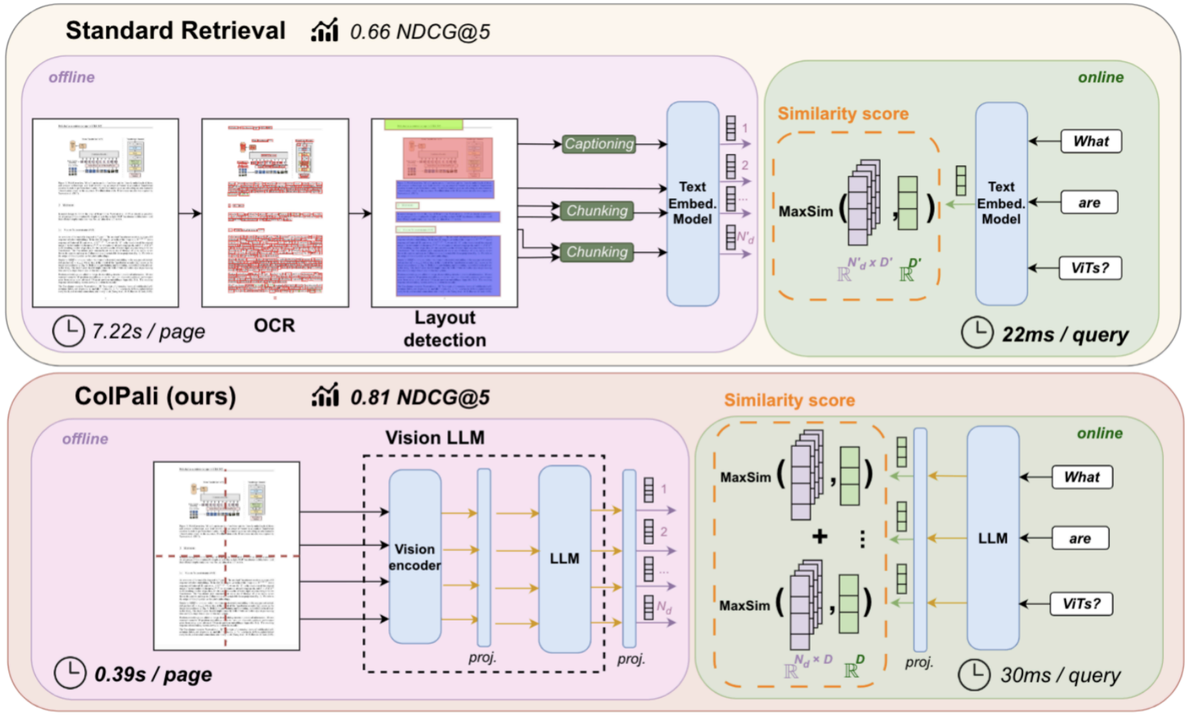
Text-only LMs (GPT, LLaMA)

The quick brown fox jumps over the lazy dog. → LM

Screenshot LMs (Donut, Pix2Struct)

→ Screenshot LM

Gao et al. 2024. "Improving Language Understanding from Screenshots"



Faysse et al. 2025. "ColPali: Efficient Document Retrieval with Vision Language Models"

Three Key Trends Ahead

1. Scaling laws don't extend as continuously as once expected
2. Scaling laws do extend continuously, but are prohibitively expensive (cost, power, environmental harms)
3. Data scaling laws slow down because high-value datasets are proprietary and fragmented (no single entity owns all the data)
4. Non-technical limits, e.g., policy, regulation, and geopolitical tensions

- Learning from limited data
  (for problems 1, 2, and 3)
- More end-to-end
  (for problems 1 and 2)
- Breaking end-to-end
  (for problems 2 and 3)

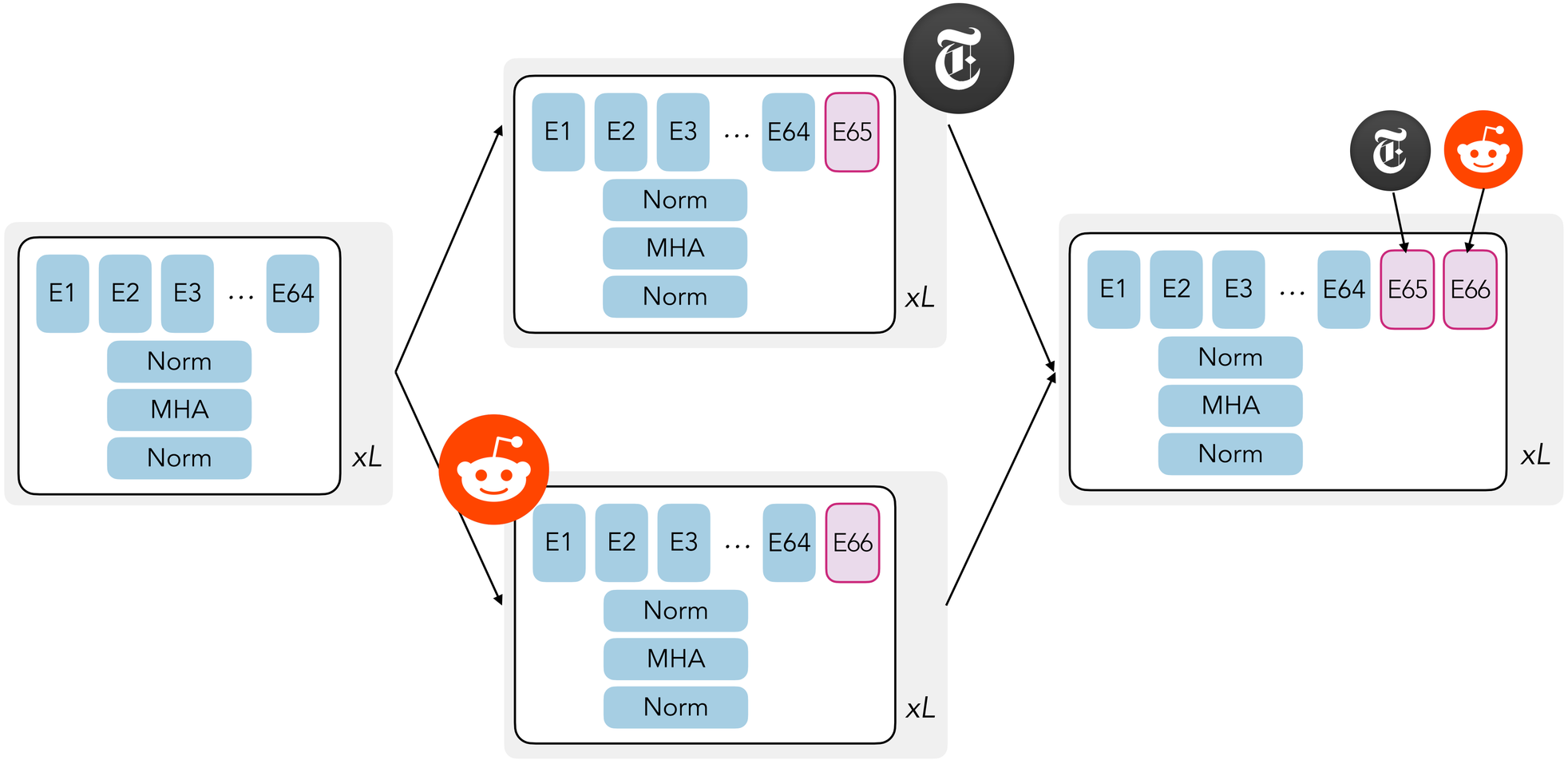
Data is the #1 Key Driver

- Currently, the “winner” is whoever has the largest, highest-quality data.
- With rising competition and data running out, we can’t simply rely on this paradigm.
- Everyone will have their own proprietary data, and we can’t assume a single entity will own all the best data (nor is that desirable).

Claim: We need a new architecture that enables “collaborative” development of an AI model.

Current training methods don’t support this (they require direct, end-to-end access to all data throughout training).

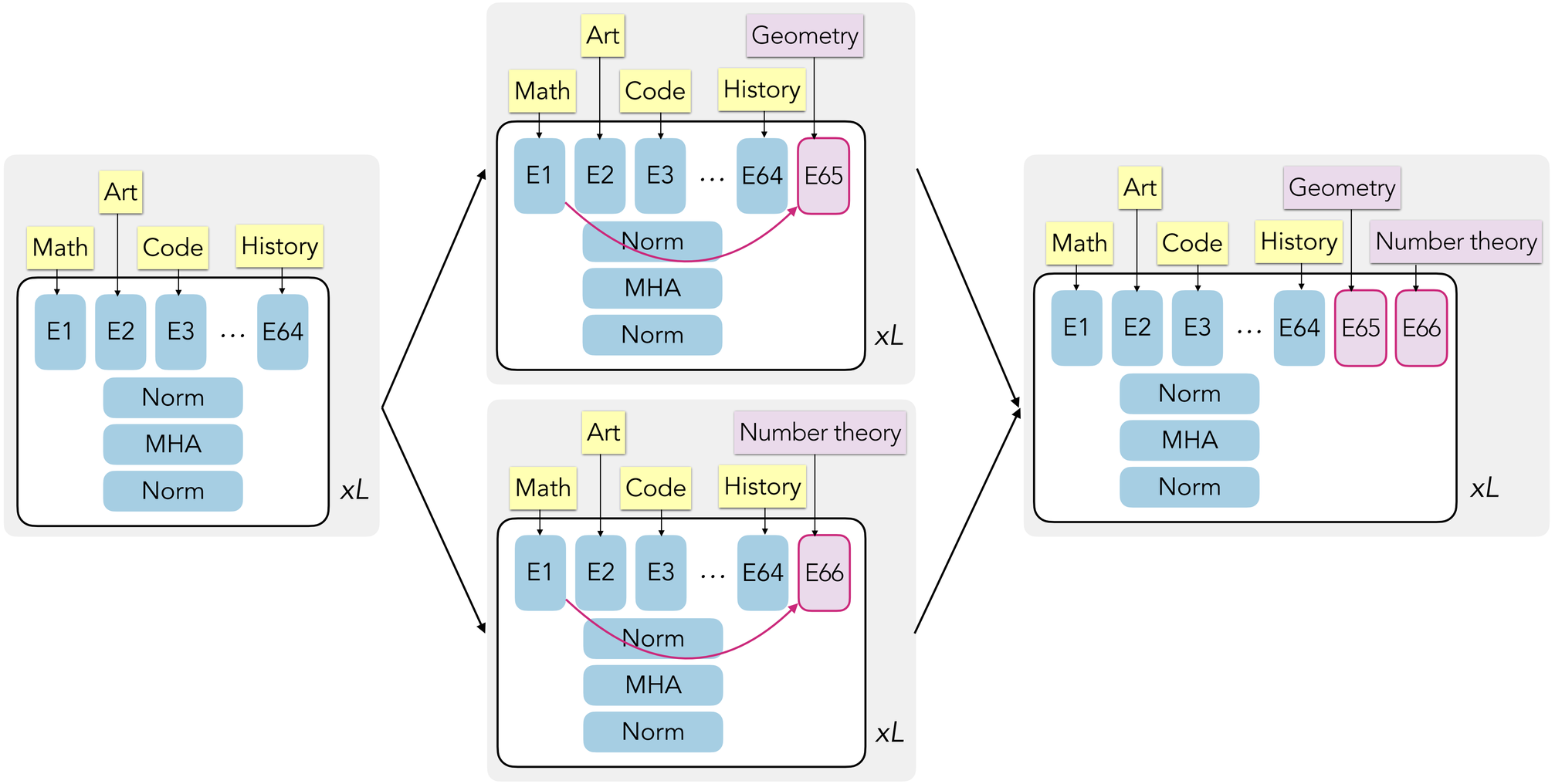
A “Hub” Model for Collaborative AI Training



Colin Raffel. 2021. “A Call to Build Models Like We Build Open-Source Software”

Shi et al. 2025. "FlexOlmo: Open Language Models for Flexible Data Use"

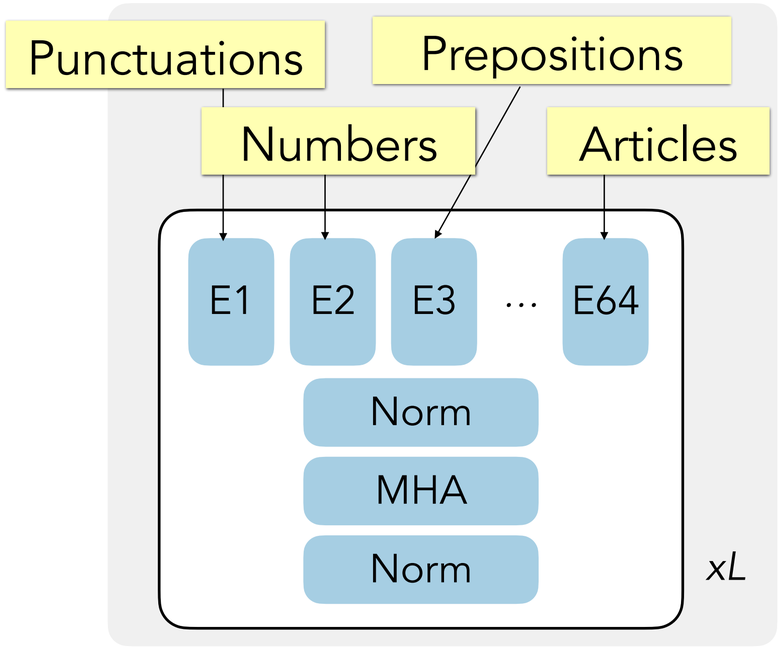
A “Hub” Model for Collaborative AI Training



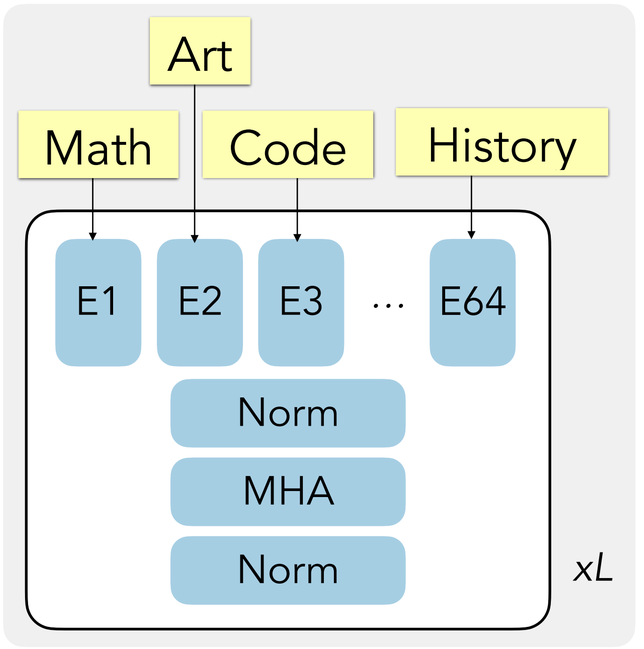
Colin Raffel. 2021. “A Call to Build Models Like We Build Open-Source Software”

Shi et al. 2025. "FlexOlmo: Open Language Models for Flexible Data Use"

Prerequisite: How to Build a Base Model



Current Mixture-of-Experts



Desired Mixture-of-Experts

Colin Raffel. 2021. "A Call to Build Models Like We Build Open-Source Software"
Shi et al. 2025. "FlexOlmo: Open Language Models for Flexible Data Use"

Summary

Focus on Problems in 2030+ (assuming we didn't achieve AGI) —
limited scaling laws, data running out, data being proprietary and fragmented

Learning from limited data

- For breaking scaling laws limits, and making better use of available data
- Example ideas: Repeating data, synthetic data, retrieval, & test-time training

More end-to-end

- For improving scaling laws, following the "end-to-end" trends
- Example ideas: Removing stages in RAG, removing tokenizer, removing text extraction (OCR)

Breaking end-to-end

- For breaking data scaling laws limits, and broadening the usable data
- Example ideas: A "Hub" model through Mixture-of-Experts

Thank you for listening!

sewonmin.com

sewonm@berkeley.edu