



# STAT 453: Introduction to Deep Learning and Generative Models

---

Ben Lengerich

Lecture 22: Unsupervised training of LLMs

November 19, 2025

Reading: See course homepage



# Today

---

- HW4 due this Friday
- [Project Presentation Sign-up](#)
  - **4 minute presentations!**
- Final Exam
  - December 17<sup>th</sup>, 5:05-7:05PM
  - Science 180



# Today

---

- Unsupervised training of LLMs
  - Emergent Capabilities
  - Challenges of MLE-based unsupervised training



# Last Time: From GPT-1 to GPT-4

---

- **Architecture:**

- **Scale:** Variety of options, with biggest (1.5B params → >1T params):
  - Block size (max context): 512 → 128k → 上文窗口大小 (决定模型能在同一时间处理多少内容)
  - Layers: 12 → >96
  - Attention Heads: 12 → >96
  - Embedding Dim: 768 → >12,288
  - Vocab: 40k → >50k tokens
- Tokenizer: Includes image patches for multimodal
- **Mixture-of-Experts**

- **Training:**

- Dataset: BookCorpus (5GB) → Private 13T tokens (~50TB)
- Reinforcement learning for alignment



# Today: Unsupervised Training of LLMs

---

- **Architecture:**

- **Scale:** Variety of options, with biggest (1.5B params → >1T params):
  - Block size (max context): 512 → 128k
  - Layers: 12 → >96
  - Attention Heads: 12 → >96
  - Embedding Dim: 768 → >12,288
  - Vocab: 40k → >50k tokens
- Tokenizer: Includes image patches for multimodal
- **Mixture-of-Experts**

- **Training:**

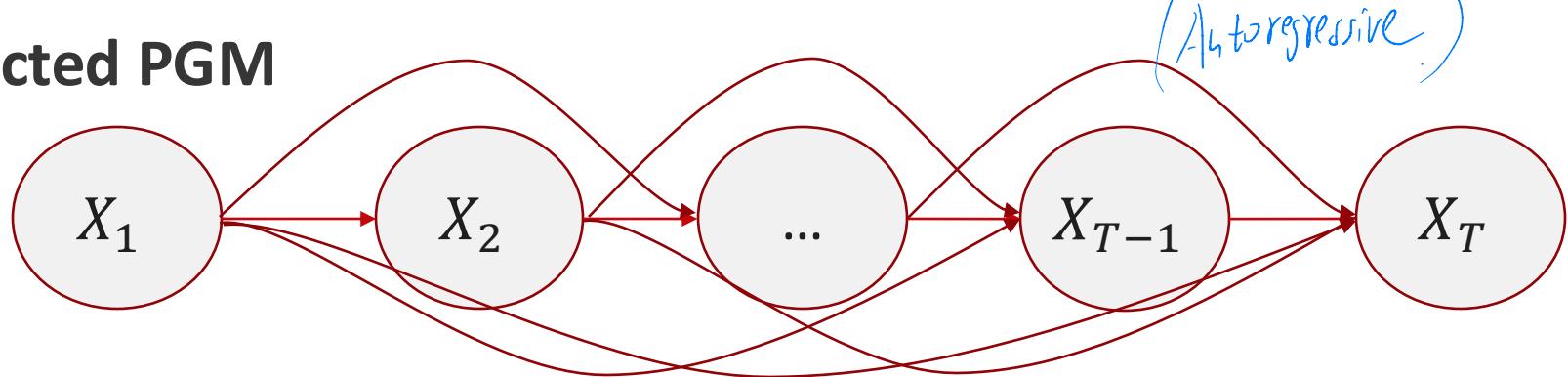
- Dataset: BookCorpus (5GB) → Private 13T tokens (~50TB)
- Reinforcement learning for alignment



# Unsupervised Training of LLMs

# Recall GPT training objective: MLE

- Directed PGM



$$P_\theta(X) = \prod_i \prod_t P_\theta(X_{i,t} \mid X_{i,<t})$$

★ Probabilistic objective: Max log-likelihood of observed seqs

$$\max_\theta \sum_i \sum_t \log P_\theta(X_{i,t} \mid X_{i,<t})$$

[Radford et al., [Improving Language Understanding by Generative Pre-Training](#)]



# We've had MLE-based Language Models for a while...

---

**Large Language Models in Machine Translation** 2007

**Thorsten Brants    Ashok C. Popat    Peng Xu    Franz J. Och    Jeffrey Dean**

Google, Inc.  
1600 Amphitheatre Parkway  
Mountain View, CA 94303, USA  
`{brants, popat, xp, och, jeff}@google.com`



Some fun:  
<https://github.com/LRitzdorf/TheJeffDeanFacts>



# We've had MLE-based Language Models for a while...

## Large Language Models in Machine Translation 2007

This paper reports on the benefits of large-scale statistical language modeling in machine translation. A distributed infrastructure is proposed which we use to train on up to 2 trillion tokens, resulting in language models having up to 300 billion  $n$ -grams. It is capable of providing smoothed probabilities for fast, single-pass decoding. We introduce a new smoothing method, dubbed *Stupid Backoff*, that is inexpensive to train on large data sets and approaches the quality of Kneser-Ney Smoothing as the amount of training data increases.

Mark C. Popat Peng Xu Franz J. Och Jeffrey Dean

Google, Inc.

1600 Amphitheatre Parkway

Mountain View, CA 94303, USA

popat, xp, och, jeff}@google.com

$$P(w_1^L) = \prod_{i=1}^L P(w_i | w_1^{i-1}) \approx \prod_{i=1}^L \hat{P}(w_i | w_{i-n+1}^{i-1})$$

→  $N$ -gram 的核心原理，这里的  $N$  就是指的是看前面几个词

理想情况，代表一句出现的概率是每个词出现概率的乘积（前提是基于它前面的所有词）

$N$ -gram 模型的做法：  
只看前面  $n$  个词，完全不管更前面的内容（瓶颈在于坏文）



# We've had MLE-based Language Models for a while...

## Large Language Models in Machine Translation 2007

Thorsten Brants Ashok C. Popat Peng Xu Franz J. Och Jeffrey Dean

Google, Inc.

1600 Amphitheatre Parkway

Mountain View, CA 94303, USA

{brants, popat, xp, och, jeff}@google.com

- Why wasn't this the LLM moment?

- Modeled n-grams, not *relationships* of token embeddings
- Transformer architecture
- Scale, GPU acceleration

做字符匹配和统计  
(把单词看作独立的字符, 只是机械地记录了“我的”后面接“东西”的概率, 没有泛化能力(语义))



# MLE → Emergent Understanding?

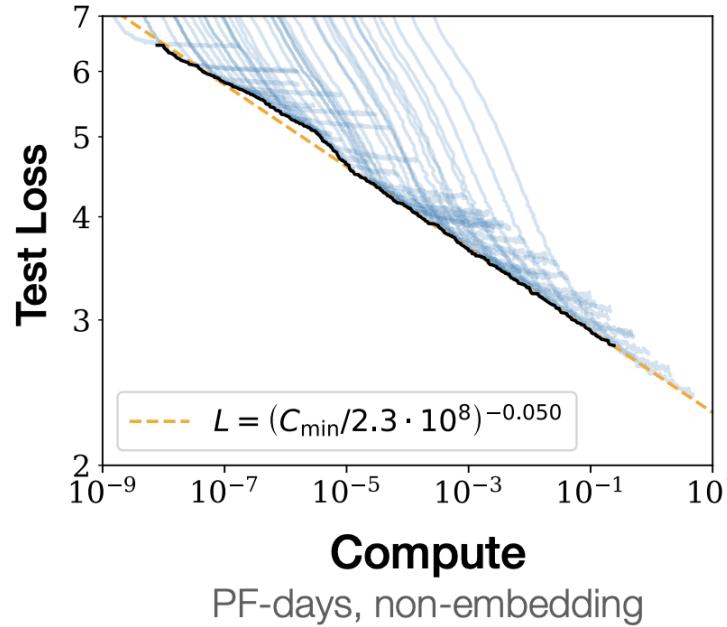
---

- "The simplest way to predict the next token is to understand what happened throughout the context."
- To predict the word "*is*" in "*The capital of France* \_\_\_\_ *Paris.*", the model must:
  - Resolve subject-verb agreement
  - Recognize a factual structure
  - Know the topic is geography



# Scale & Emergent Capabilities

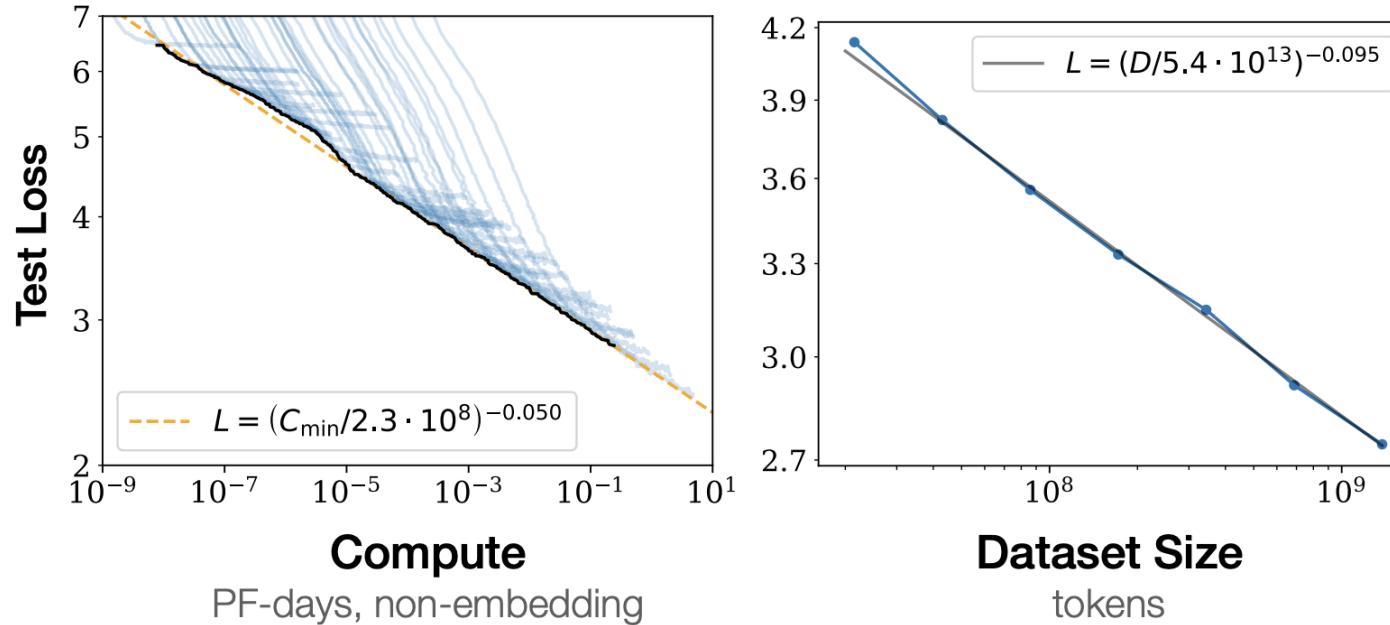
# What happens as we scale training?



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

“Scaling Laws for Neural Language Models”. Kaplan et al 2021

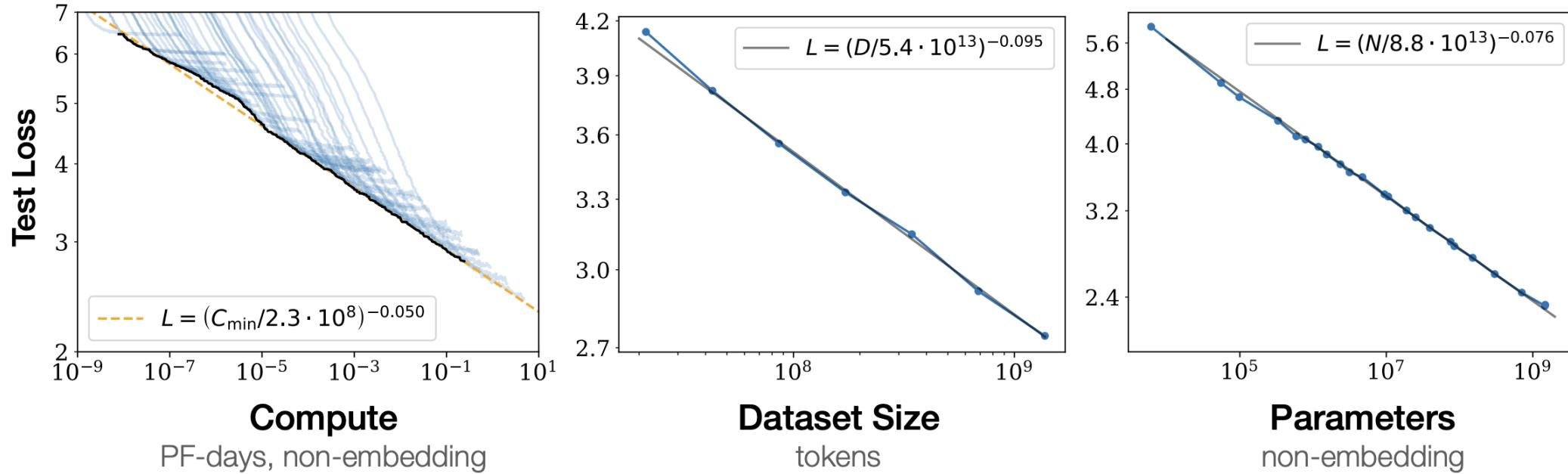
# What happens as we scale training?



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

“Scaling Laws for Neural Language Models”. Kaplan et al 2021

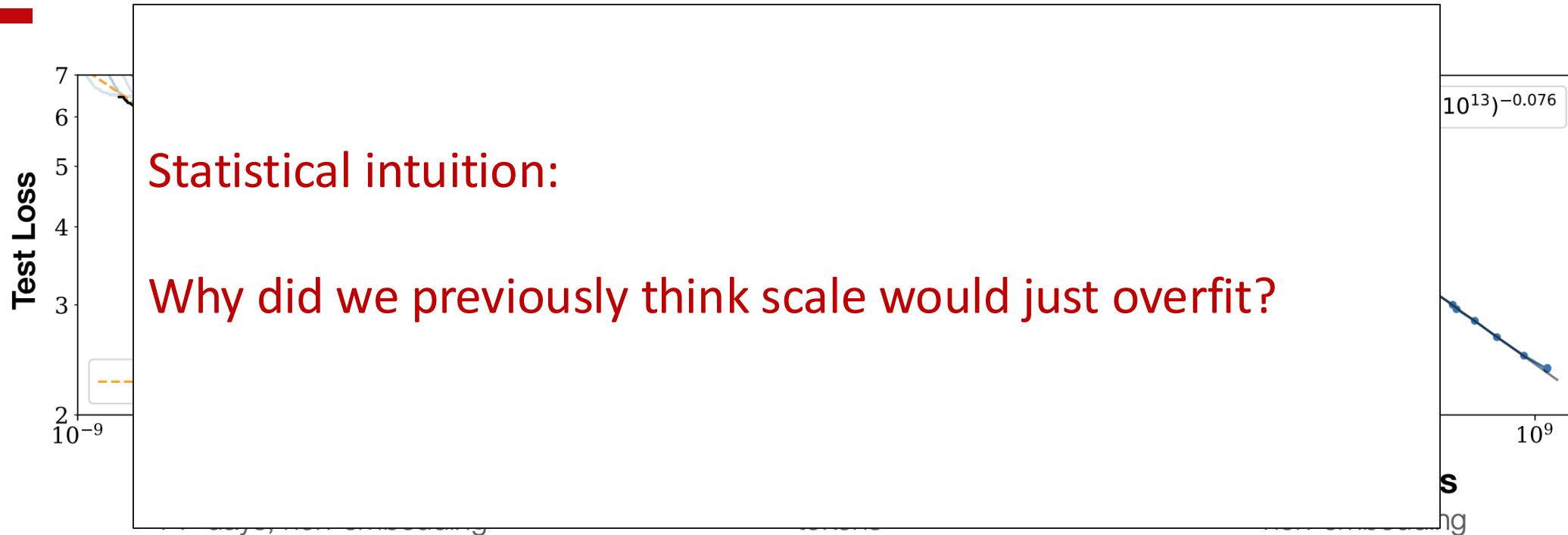
# What happens as we scale training?



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

“Scaling Laws for Neural Language Models”. Kaplan et al 2021

# What happens as we scale training?

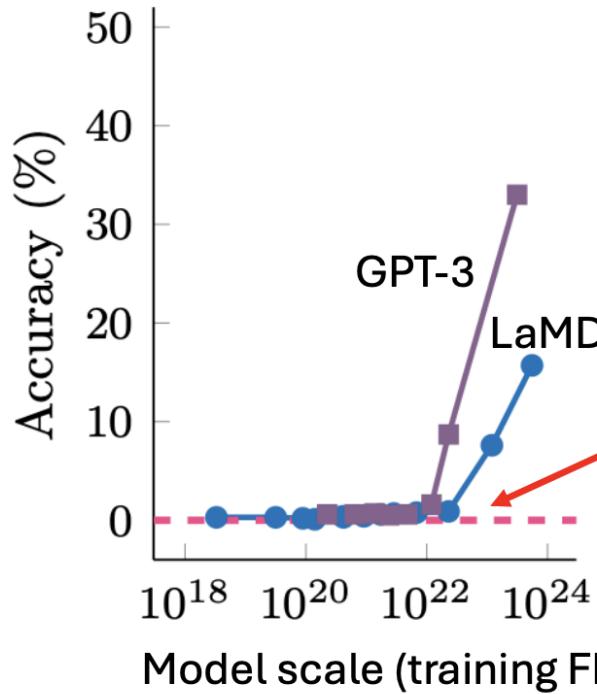


**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

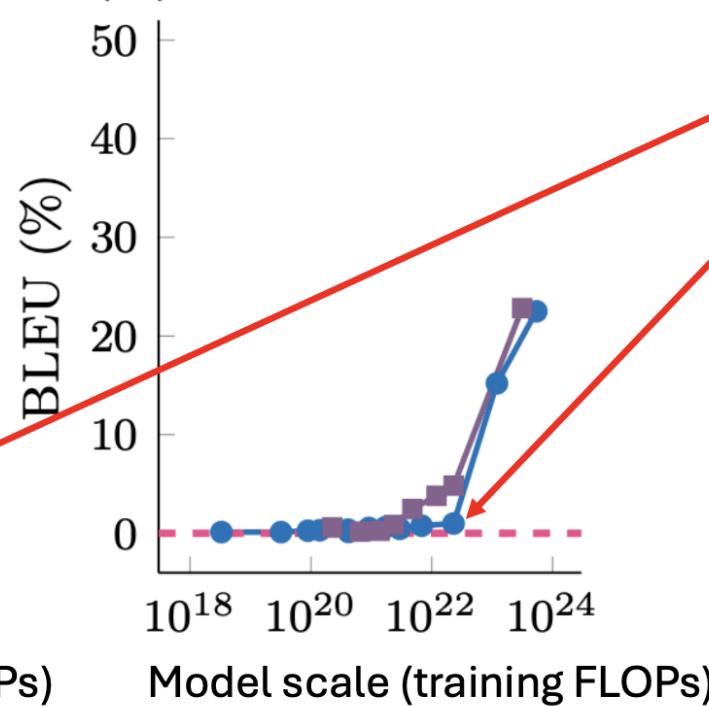
“Scaling Laws for Neural Language Models”. Kaplan et al 2021

# Smooth improvements → sharp emergent ability?

(A) Mod. arithmetic



(B) IPA transliterate

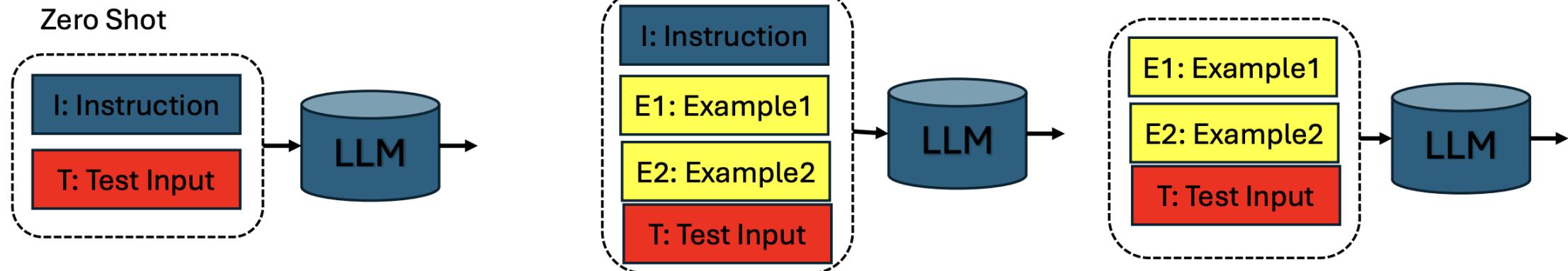


An ability is emergent if it is not present in smaller models but is present in larger models [Wei, et al (2022). Emergent Abilities of Large Language Models]



# Example of emergence: In-Context Learning

I: Instruction	Translate English to French	
E1: Example1	[en]: A discomfort which lasts.	[fr]: Un malaise qui dure
E2: Example2	[en]: HTML is a language for formatting formatage	[fr]: HTML est un langage de
T: Test Input	[en]: After you become comfortable with formatting [fr]:	





# Example of emergence: Chain-of-Thought

## Standard Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. X

## Chain-of-Thought Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

Wei, et al. (2023) Chain-of-Thought Prompting Elicits Reasoning in LLMs

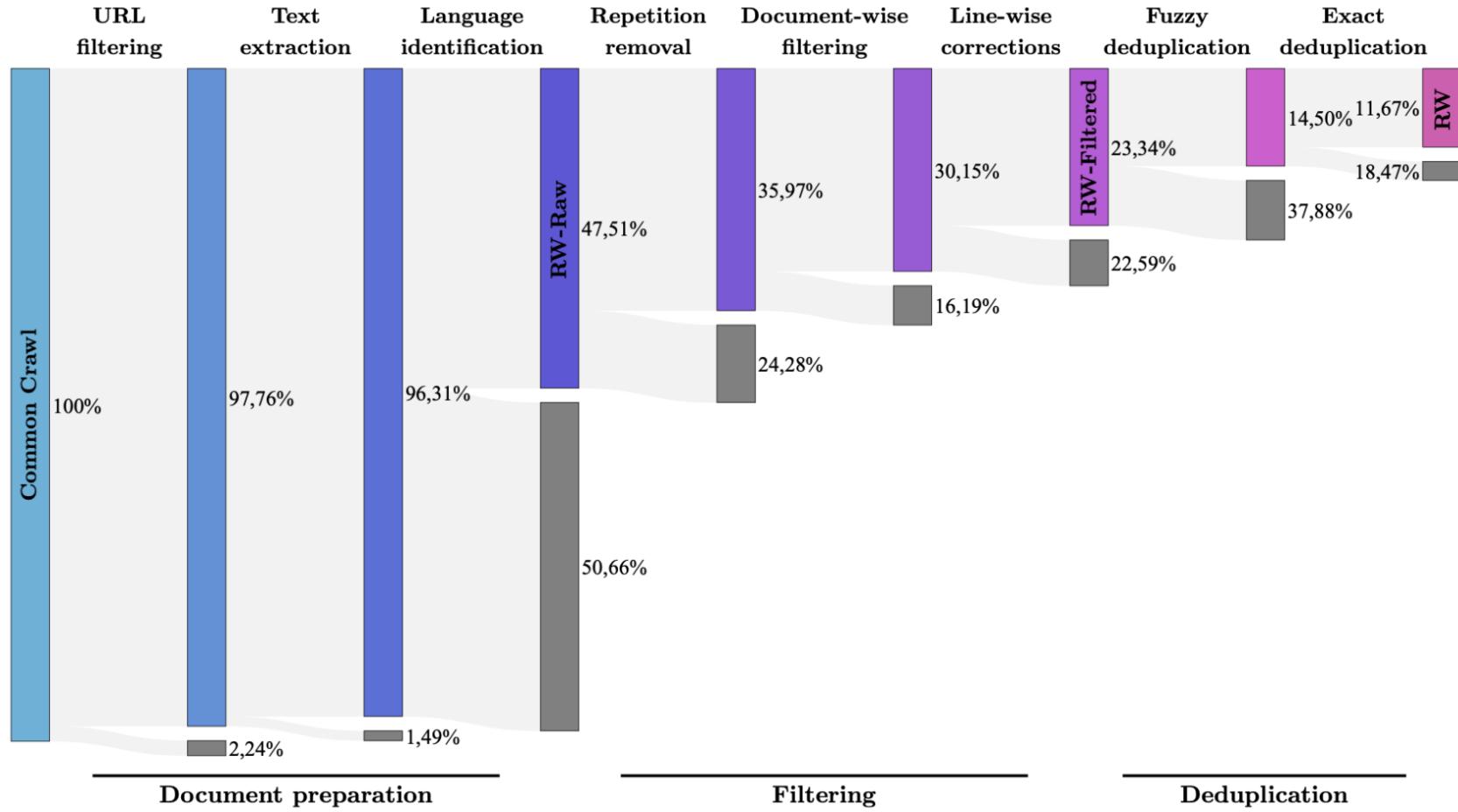


# Why does this work? Some hypotheses

---

- Task identification?
  - Xie et al. (2021). An explanation of in-context learning as implicit Bayesian inference
  - Raventos, et al. (2023). Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression
- Some kind of "learning" without model updates?
  - Akyurek, et al. (2024). In-context language learning: architectures and algorithms
  - von Oswald, et al. (2023). Transformers learn in-context by gradient descent
- Both?
  - Pan, et al. (2023). What in-context learning "learns" in-context: disentangling task recognition and task learning

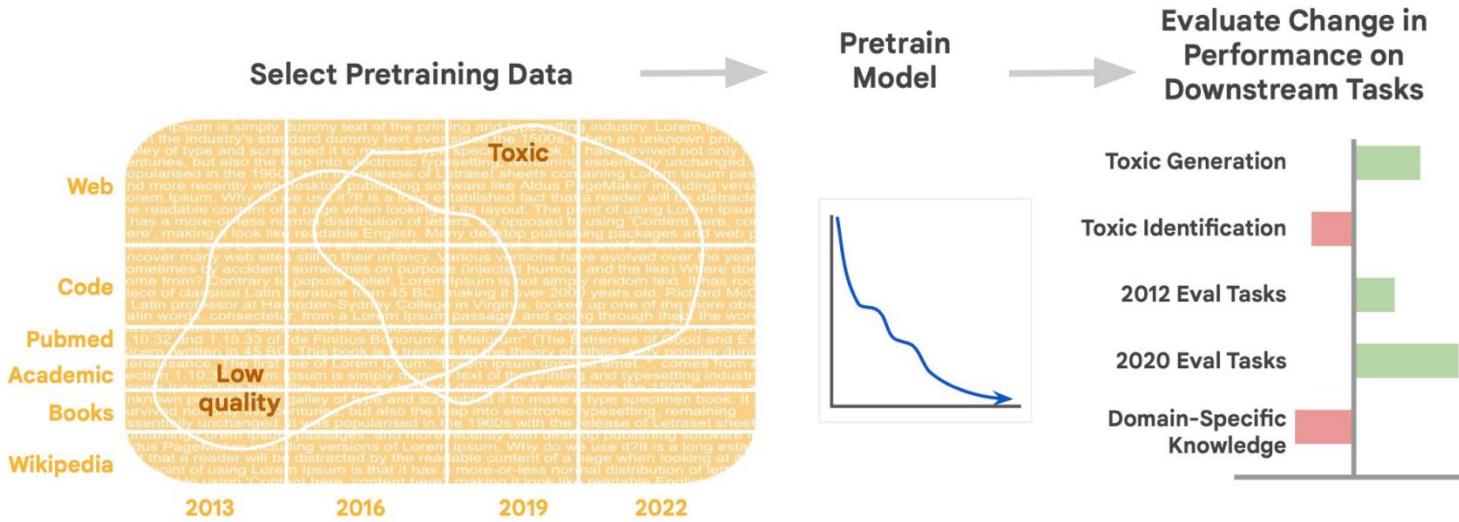
# Scale is difficult...for example, data filtering



Penedo, et al. (2023) The Refined Web dataset for Falcon LLM

# Scale is difficult...for example, data filtering

A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity [Longpre et al, NAACL 2024]

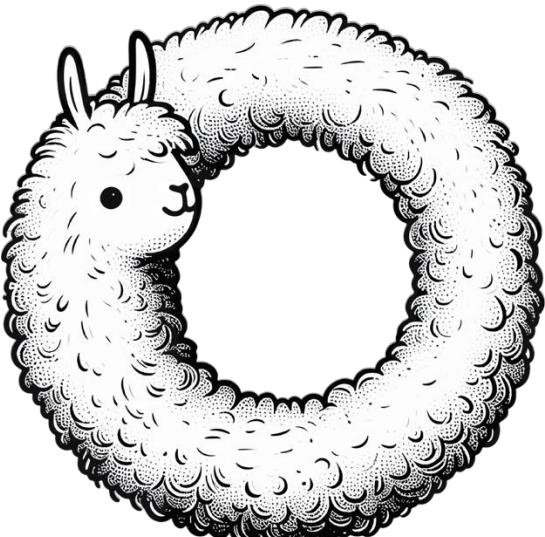


Some findings: strongly encourage to read the paper!

- “temporal shift between evaluation data and pretraining data leads to performance degradation, which is not overcome by finetuning”
- “a trade-off between performance on standard benchmarks and risk of toxic generations... there does not exist a one-size-fits-all solution to filtering.”



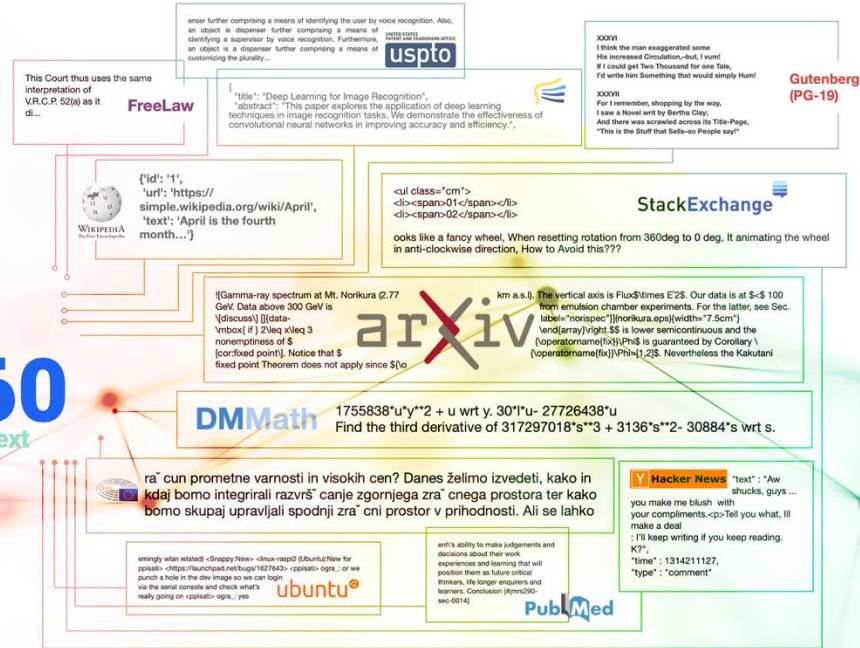
# A great learning opportunity



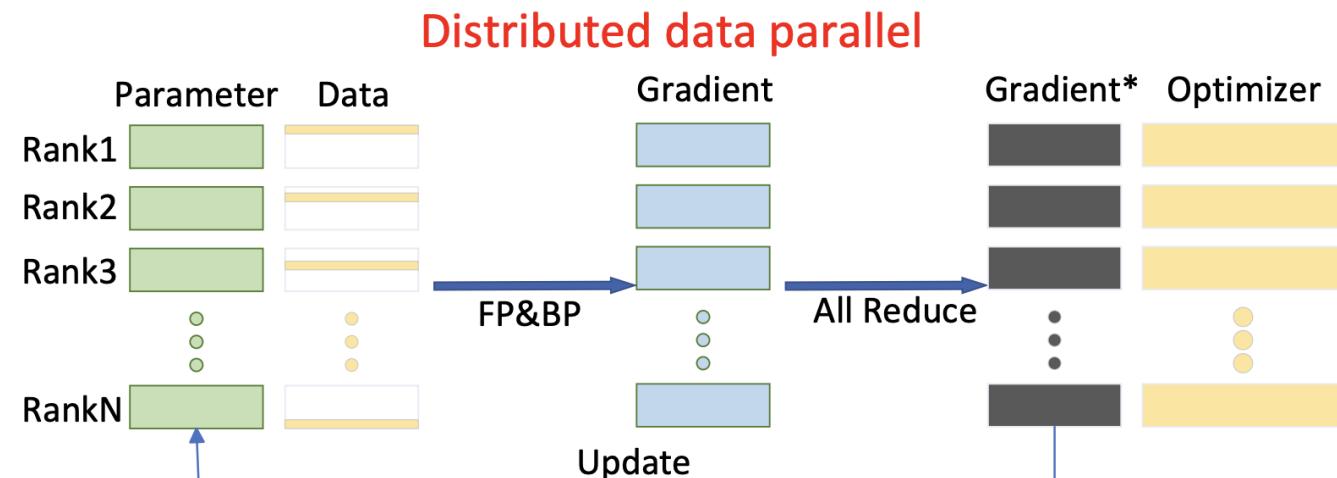
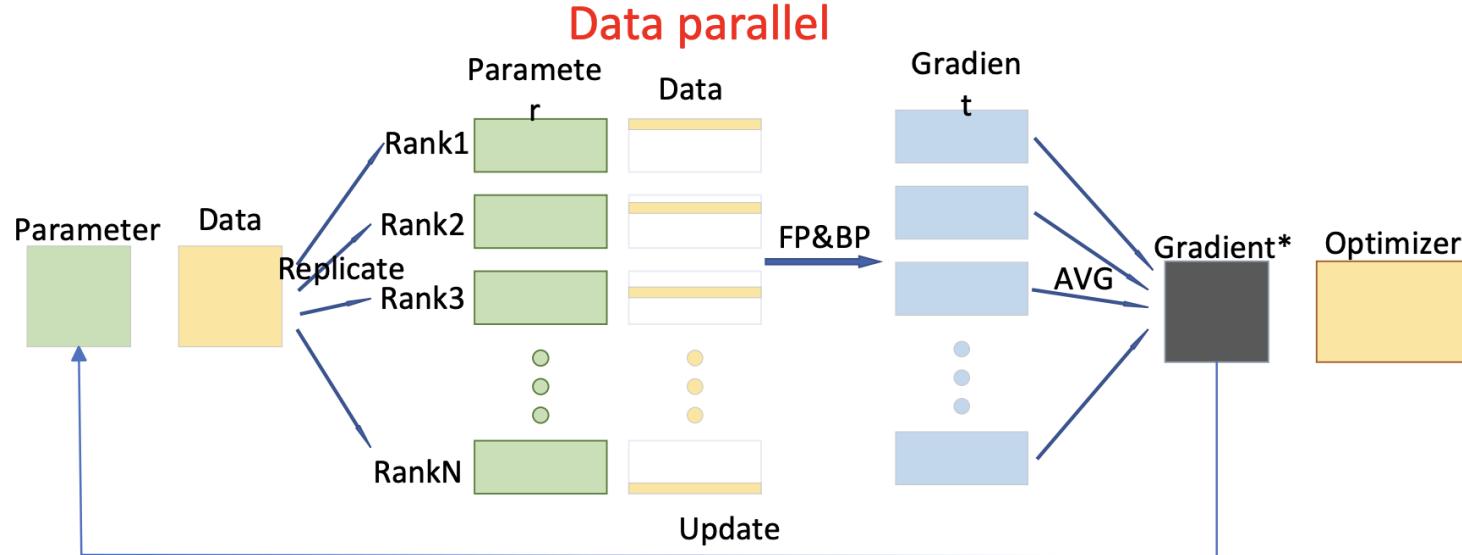
**LLM360**



TxT360



# Scale is difficult...for example, parallel training



[Understanding LLMs: A Comprehensive Overview from Training to Inference](#)



# Challenges of MLE-based unsupervised training



# MLE: A KL-Divergence view

---

$$KL(P_{data} \parallel P_{\theta}) = E_{x \sim P_{data}}[-\log P_{\theta}(x)] + const.$$

$$\operatorname{argmax}_{\theta} E_{x \sim P_{data}}[\log P_{\theta}(x)] = \operatorname{argmin}_{\theta} KL(P_{data} \parallel P_{\theta})$$

- Must spread probability mass to cover observed sequences, even if incoherent
  - Repetition and genericity ("The man said the man said...")
  - Poor calibration on out-of-distribution prompts
  - Memorization of rare patterns
- Sampling strategies matter
- Entropy / confidence



# What does MLE not do?

---

- No **semantics**
- No **task goals**
- No **explicit reward**



# Entropy and Confidence

---

- Token-level entropy:

$$\begin{aligned} H_t &= - \sum_v P(x_t = v \mid x_{<t}) \log P(x_t = v \mid x_{<t}) \\ &= E_v[-\log P(x_t = v \mid x_{<t})] \end{aligned}$$

- Low entropy = high confidence.

Do we want our model to be low or high entropy?



# Recall from Variational Inference

---

$$\log p(x | \theta) = E_{z \sim q}[\log p(x, z | \theta)] + H(q) + KL(q(z | x) || p(z | x, \theta))$$

$$\log p(x | \theta) \geq E_{z \sim q}[\log p(x, z | \theta)] + H(q)$$



A horizontal red bracket is positioned below the first two terms of the equation, spanning from the start of the inner expectation to the end of the term plus  $H(q)$ .

“ELBO”: Evidence Lower Bound

Maximizing ELBO → Increased entropy of  $q(z)$

Does Maximizing Likelihood → Increased entropy of  $p(x | \theta)$ ?



# MLE and Entropy

---

$$\widehat{\theta_{MLE}} = \operatorname{argmax}_{\theta} E_{x \sim P_{data}} [\log P_{\theta}(x)]$$

- Directly optimize the data distribution  $P_{\theta}(x)$  without explicitly introducing auxiliary variables or explicitly controlling the entropy of  $x$
- Often implicitly pushes  $P_{\theta}(x)$  toward low-entropy distributions
- Mode collapse / degeneration / “memorization”



# MLE and Entropy

**Context:** In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

## Pure Sampling:

They were cattle called **Bolivian Cavalleros**; they live in a remote desert **uninterrupted by town**, and they speak **huge, beautiful, paradisiacal Bolivian linguistic thing**. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

## Beam Search, $b=32$ :

"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

THE CURIOUS CASE OF NEURAL TEXT  
DeGENERATION [Holtzman 2020]



# MLE and Entropy

---

$$\widehat{\theta}_{MLE} = \operatorname{argmax}_{\theta} E_{x \sim P_{data}} [\log P_{\theta}(x)]$$

- Solutions:
  - Explicitly add a penalty for low entropy:

$$\widehat{\theta} = \operatorname{argmax}_{\theta} E_{x \sim P_{data}} [\log P_{\theta}(x)] + \lambda \sum_t H[P_{\theta}(x_t | x_{<t})]$$

- Smooth labels:

$$\tilde{y}_\epsilon = \begin{cases} 1 - \epsilon, & y = 1 \\ \frac{\epsilon}{|V| - 1}, & y = 0 \end{cases}$$

- Contrastive losses:
- Scheduled sampling / noise contrastive objectives
- Risk minimization, utility, preference-based losses

Questions?

