Fall 2024 COS597R:

Deep Dive into Large Language Models

Danqi Chen, Sanjeev Arora

PRINCETON UNIVERSITY

Lecture 5: Emergent behaviors in LLMs and our current understanding

https://princeton-cos597r.github.io/

"Emergence"

Wikipedia

In philosophy, systems theory, science, and art, **emergence** occurs when a complex entity has properties or behaviors that its parts do not have on their own, and emerge only when they interact in a wider whole.

Emergence plays a central role in theories of integrative levels and of complex systems. For instance, the phenomenon of life as studied in biology is an emergent property of chemistry and physics.

"More is different"
[Philip Anderson, 1971]

The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles."

"Emergence"

Wikipedia

In philosophy, systems theory, science, and art, **emergence** occurs when a complex entity has properties or behaviors that its parts do not have on their own, and emerge only when they interact in a wider whole.

Emergence plays a central role in theories of integrative levels and of complex systems. For instance, the phenomenon of life as studied in biology is an emergent property of chemistry and physics.

"Weak"

weak emergence is a type of emergence in which the emergent property is amenable to computer simulation or similar forms of after-the-fact analysis (for example, the formation of a traffic jam, the structure of a flock of starlings in flight or a school of fish, or the formation of galaxies).

"Strong"

(possibly unscientific?)

The whole is other than the sum of its parts. It is argued then that no simulation of the system can exist, for such a simulation would itself constitute a reduction of the system to its constituent parts

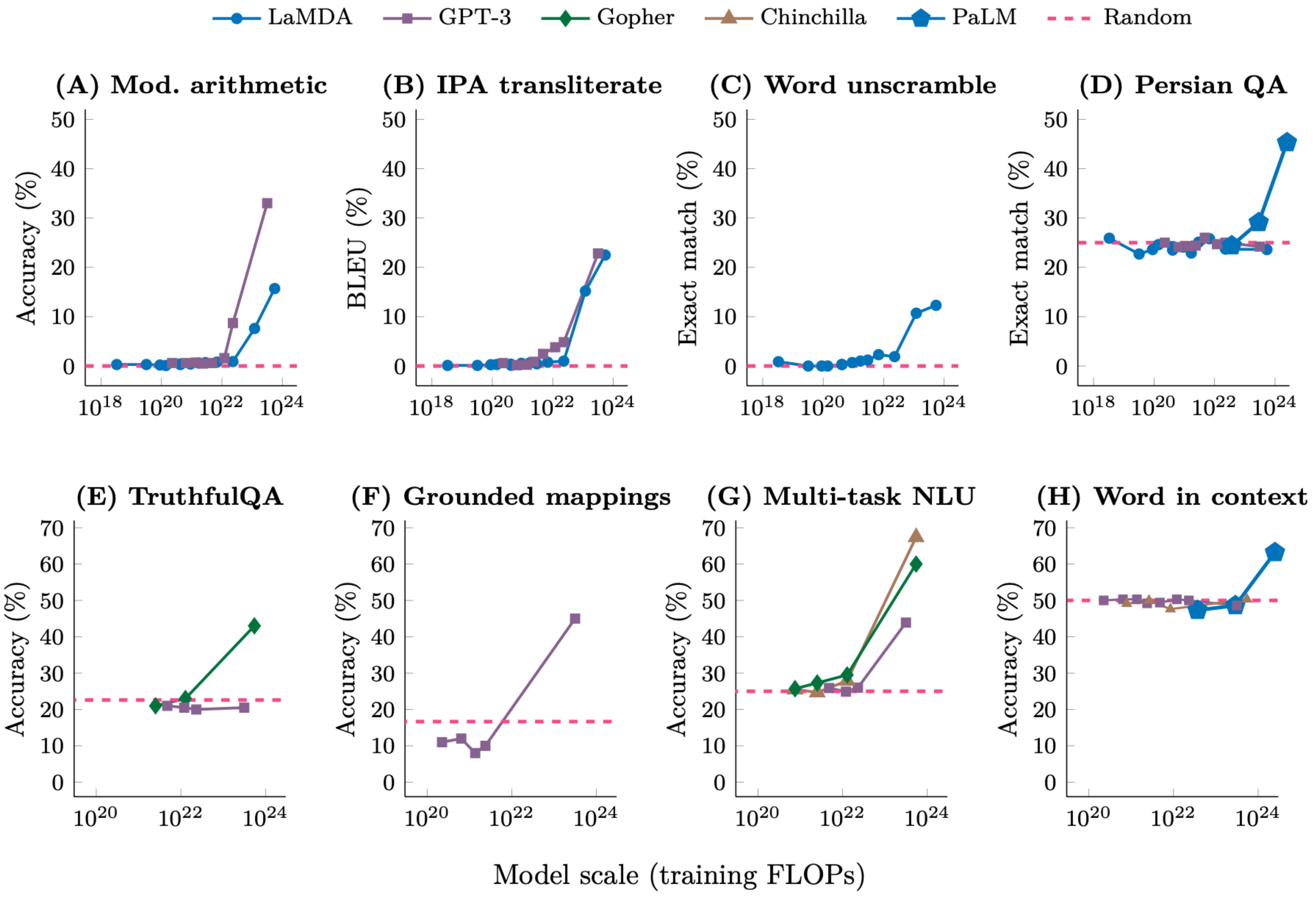
The “emergence” phenomenon in LLMs

Emergent Abilities of Large Language Models , Wei et al’21

From the abstract..

Scaling up language models has been shown to predictably improve performance and sample efficiency on a wide range of downstream tasks. This paper instead discusses an unpredictable phenomenon that we refer to as *emergent abilities* of large language models. We consider an ability to be emergent if it is not present in smaller models but is present in larger models. 1.
Thus, emergent abilities cannot be predicted simply by extrapolating the performance of smaller models. The existence of such emergence raises the question of whether additional 1.
scaling could potentially further expand the range of capabilities of language models.

Scaling up makes LLMs qualitatively different



Strange case of Chain-of-Thought (also emergent)

[Wei et al '22]

Question: 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?

Answer: 7

Question: 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?

Answer: Let's think step by step..

Each can has 3 tennis balls and so 2 cans have 3 x 2 =6 tennis balls. Since Roger started with 5 tennis balls he now has 5 + 6 =11 tennis balls.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei **Xuezhi Wang** **Dale Schuurmans** **Maarten Bosma**
Brian Ichter **Fei Xia** **Ed H. Chi** **Quoc V. Le** **Denny Zhou**

Emergent tasks related to picking up capabilities (either in-context learning or SFT)

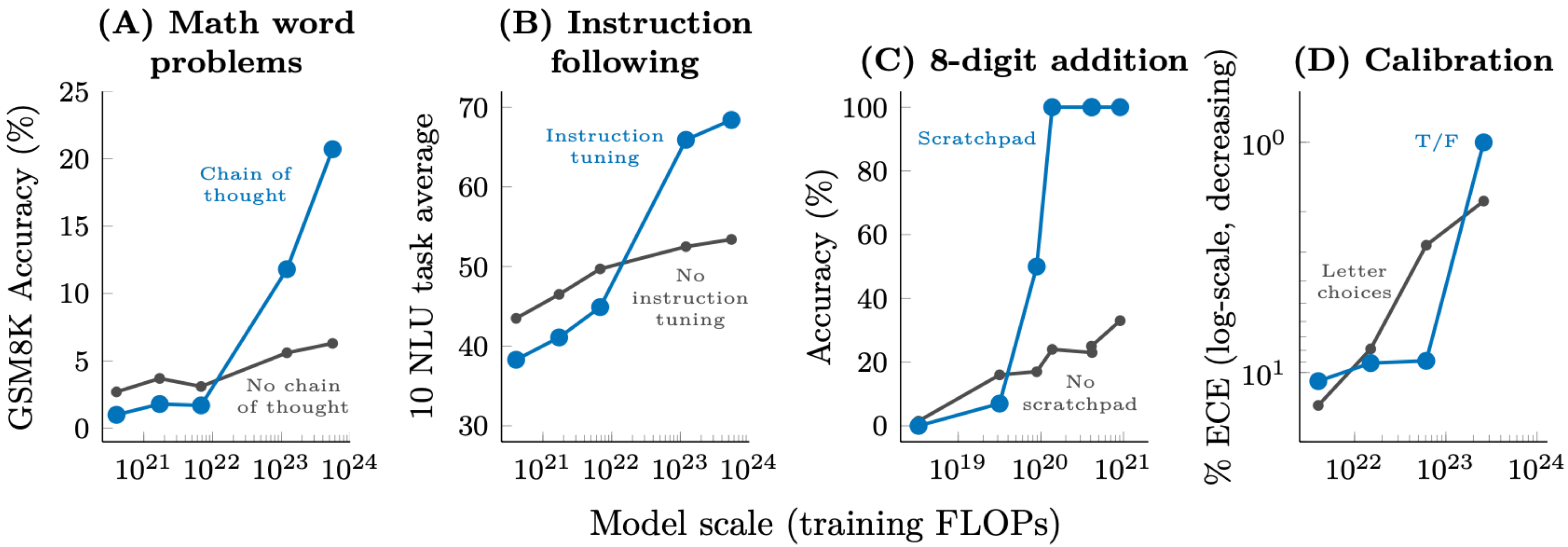


Figure 3: Specialized prompting or finetuning methods can be emergent in that they do not have a positive effect until a certain model scale. A: Wei et al. (2022b). B: Wei et al. (2022a). C: Nye et al. (2021). D: Kadavath et al. (2022). An analogous figure with number of parameters on the x-axis instead of training FLOPs is given in Figure 12. The model shown in A-C is LaMDA (Thoppilan et al., 2022), and the model shown in D is from Anthropic.

Go to page 14

Doing really well on next-word prediction requires general purpose skills (grammar, world knowledge, etc.)

The glass fell off the table onto the marble floor and ….

Human: “shattered”

Model: “bounced”

Winograd Schemas [1971]

The city councilmen denied the demonstrators a permit because they feared violence. Q: Who feared violence? A: Demonstrators B: Councilmen

Such tests were considered hard for many years. They became trivial from 10x LLM scaling, over just a year or two.

Possible explanations why scale may help

(Sec 5.1)

The glass fell of the table on the marble floor and (a) shattered (b) bounced.

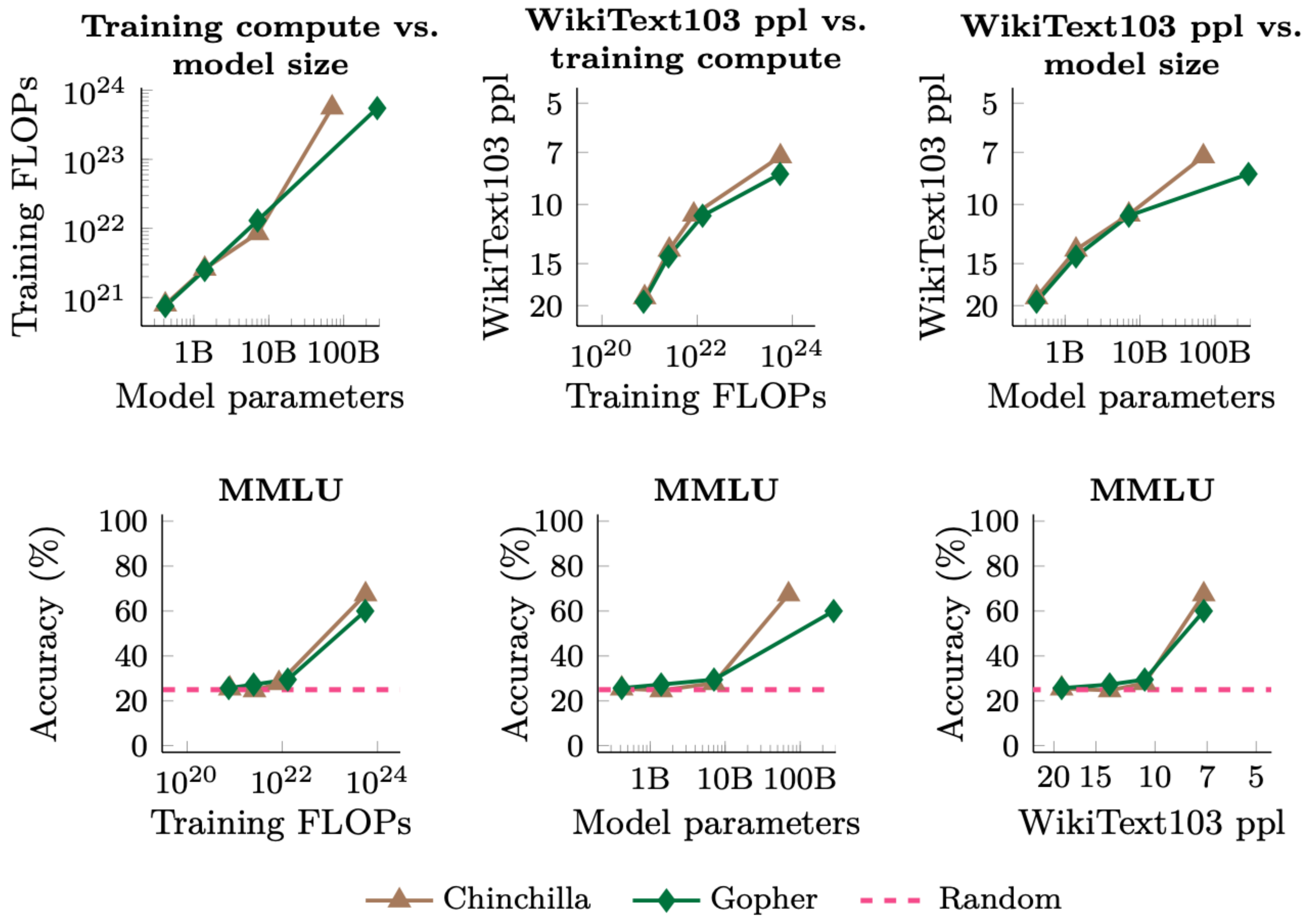
1. Bigger models => higher depth. Maybe this enables multi-step reasoning?

2. Larger models => More capacity to remember world-knowledge (e.g., properties of glass, marble etc.), grammar rules, etc.

3. Many NLP tasks are graded using exact or approximate string matching (eg BLEU scores). Good score requires getting many “matches”, which is a discrete metric, not continuous. It could appear discontinuously as model is scaled up.

4. The paper reports that cross-entropy loss on correct answers grows continuously during scaling, even though the discrete score is continuous.

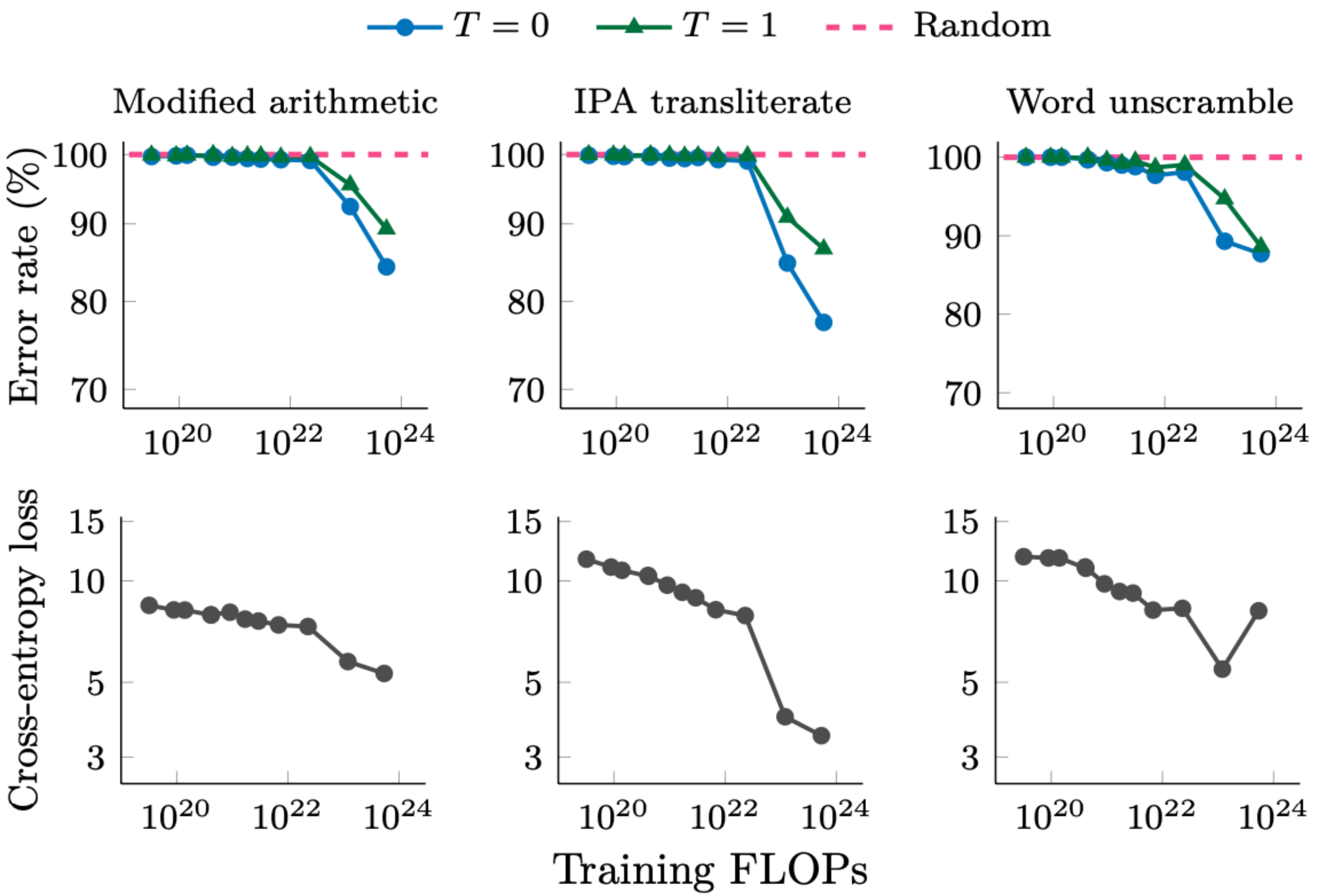
Role of Perplexity/cross entropy?



5.3 Another view of emergence

While scale (e.g., training FLOPs or model parameters) has been highly correlated with language model performance on many downstream metrics so far, scale need not be the only lens to view emergent abilities. For example, the emergence of task-specific abilities can be analyzed as a function of the language model's perplexity on a general text corpus such as WikiText103 (Merity et al., 2016). Figure 4 shows such a plot with WikiText103 perplexity of the language model on the x-axis and performance on the MMLU benchmark on the y-axis, side-by-side with plots of training FLOPs and model parameters on the x-axis.

Role of c-e loss (contd.)



string-manipulation type tasks

Do LLMs “understand”? Are they producing novel text?

Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

(Will be one of the debate papers next week)

"Stochastic Parrot" Debate

was published, BERT was picked up by the NLP community and applied with great success to a wide variety of tasks [e.g. 2, 149].

However, no actual language understanding is taking place in LM-driven approaches to these tasks, as can be shown by careful manipulation of the test data to remove spurious cues the systems are leveraging [21, 93]. Furthermore, as Bender and Koller [14] argue from a theoretical perspective, languages are systems of signs [37], i.e. pairings of form and meaning. But the training data for LMs is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized.

"Stochastic parrots" ??
[Bender et al'21]

credit: DALLE-3

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Strange case of Chain-of-Thought (also emergent)

[Wei et al ’22]

Question: 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?

Answer: 7

Question: 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?

Answer: Let’s think step by step..

Each can has 3 tennis balls and so 2 cans have 3 x 2 =6 tennis balls. Since Roger started with 5 tennis balls he now has 5 + 6 =11 tennis balls.

Did the LLM actually “think” or did it parrot back patterns?

Creativity out of AI?

choices. When you give a generative-A.I. program a prompt, you are making very few choices; if you supply a hundred-word prompt, you have made on the order of a hundred choices.

If an A.I. generates a ten-thousand-word story based on your prompt, it has to fill in for all of the choices that you are not making. There are various ways it can do this. One is to take an average of the choices that other writers have made, as represented by text found on the Internet; that average is equivalent to the least interesting choices possible, which is why A.I.-generated text is often really bland. Another is to instruct the program to engage in style mimicry, emulating the choices made by a specific writer, which produces a highly derivative story. In neither case is it creating interesting art.

Why AI Isn't Going to Make Art

Ted Chiang, New Yorker Aug'24

Qs: Point out misconceptions about LLMs in this para.

Some evidence of “creativity”?

(Note: we don’t know the training corpus of the frontier models..)

Compositional capability: SkillMix Evaluation

***N* Skills**

self serving bias
red herring
…
spatial reasoning
modus ponens

***T* Topics**

Sewing
Dueling
…
Beekeeping
Gardening

"Skill-Mix: A Flexible and Expandable Family of Evaluations for Language Models" Dingli Yu, Simran Kaur, Arushi Gupta, J. Brown-Cohen, A. Goyal, S. Arora ICLR'24

Generate a short text about sewing that exhibits these skills:
spatial reasoning, self serving bias, metaphor

LLaMA-2 7B Chat

I'm so glad I finished that dress. It was a puzzle to piece together, but I think it turned out great.

LLaMA-2 70B Chat

I'm struggling to sew this dress because it's like trying to fit a square peg into a round hole.

GPT-4

In the labyrinth of sewing, I am the needle navigating between the intricate weaves. Any errors are due to the faulty compass of low-quality thread, not my skill.

Capability by model size

Model size	# skills
Small (10^10)	1-2
Medium (10^11)	3
Large (GPT4)	5

Grad students struggle to combine 4 skills

Compositional capability $\Longrightarrow$ LLMs are not “stochastic parrots”

N skills, T topics

Need to compose random subset of k skills

$\binom{N}{k} \times T$ possible combinations

GPT4 succeeds often for k=5 !

“Stochastic parrots” ??
[Bender et al’2

credit: DALLE-3

Simple probability calculation (based upon estimated frequencies of skills and topics in the corpus) shows that random topic + set of 5 skills are **unlikely** to have occurred in the training corpus.

Suggestion: Many emergence phenomena correspond to improved compositional capability

PRINCETON
Language + Intelligence

Mathematical understanding of emergence of new capabilities (gentle intro to the theory)

[*A theory for emergence of complex skills in LLMs from scaling, Arora and Goyal 2023*]

Theory: Some hurdles

- Mathematical analysis of deep learning is in its infancy.
- We're interested in "new capabilities" (ie tasks not seen in training)
- What are "language corpus" and "skills" (mathematically speaking)?

Theory TL;DR…

Key Assumptions: "LLM Scaling laws" + structural assumption about training data

Main prediction: Every 10x scaling of LLM size and dataset

will **double** the number of skills it is able to combine while solving tasks.

("Compositional Generalization")

(Recall: # of k-combinations of skills $\propto$ $(\#\text{skills})^k$)

This prediction was verified via SKILLMIX Evaluation on leading models (as mentioned earlier)

PRINCETON Language + Intelligence

Structural assumption about language

Mixing Assumption: If you look at a random place in text, you'll find that its comprehension requires a set of k random skills

Mathematical consequence (as shown in the paper):

Competence in individual skills arises roughly <u>in tandem</u>
Likewise, competence at applying pairs/triples of skills.

Roughly like "emergence"?

While transformers can be used to model all kinds of distribution (molecules, genes etc) it's possible that text/language is a uniquely conducive to learning.

The city councilmen denied the demonstrators a permit because they feared violence. Q: Who feared violence? A: Demonstrators B: Councilmen

Background: Cross-entropy and "understanding" (Folk-lore)

LLMs implicitly use following view of language

Consider random text-piece in the corpus, say $w_1 w_2 \ldots w_n$.

There is a ground-truth (i.e., humans') distribution for generating the next word

$p_i(w \mid w_1 w_2 \ldots w_i)$ = Probability that w is the $(i+1)$the word, given the previous i words

(will shorten this to $p_i(w)$)

$\sum_w p_i(w) \log \frac{1}{p_i(w)}$ = **Entropy** of this next-word distribution after seeing $w_1 w_2 \ldots w_i$

Cross-entropy (contd)

$p_i(w) = p_i(w \mid w_1 w_2 \ldots w_i)$ = Probability that w is the $(i+1)$the word, given $w_1 w_2 \ldots w_i$

$\sum_w p_i(w) \log \frac{1}{p_i(w)}$ = Entropy of this next-word distribution given we saw $w_1 w_2 \ldots w_i$

Let $q_i(w)$ = probability assigned by the model to w as next word given $w_1 w_2 \ldots w_i$

LLM loss, ie, cross-entropy (c-e) incurred = $\log \frac{1}{q_i(w_{i+1})}$

So expected c-e loss is $\sum_w p_i(w) \ \log(\frac{1}{q_i(w)})$

KL "distance"

entropy

$$= \sum_{w_{i+1}} p_i(w_{i+1}) \ \log(\frac{p_i(w_{i+1})}{q_i(w_{i+1})}) + \sum_{w_{i+1}} p_i(w_{i+1}) \log \frac{1}{p(w_{i+1})} \quad = KL(p \,||\, q) + H(p)$$

Understanding LLM scaling

KL "distance" entropy

From previous slide: c-e loss $= KL(p\,||\,q) + H(p)$

Distribution p is fixed (ie depends on humans) and so is H(p). The model controls only q

$$L(N, D) = 1.8172 + \frac{482.01}{N^{0.3478}} + \frac{2085.43}{D^{0.3658}}$$

(Scaling law from last time)

$H(p) +$ $KL(p\,||\,q)$

"10x scaling reduces KL by 2x"

Minimizing c-e loss ≡ Minimize $KL(p\,||\,q)$ ("distance" from underlying distribution)

When $D > (4000)^3, N > 10^9$ and this KL term gets fairly small < 0.1 !

Lack of understanding $\Longrightarrow$ High c-e loss

The glass fell off the table onto the marble floor and ….

Human : $\Pr[\text{"shattered"}] = 7/8 \ \Pr[\text{"bounced"}] = 1/8$

Model (w/ imperfect understanding): $\Pr[\text{"shattered"}] = 1/8 \ \Pr[\text{"bounced"}] = 7/8$

KL for Model at this place in the corpus = $\frac{7}{8}\log 7 + \frac{1}{8}\log\frac{1}{7} > 2$

This kind of KL cannot occur too often in the corpus (since avg is < 0.1)

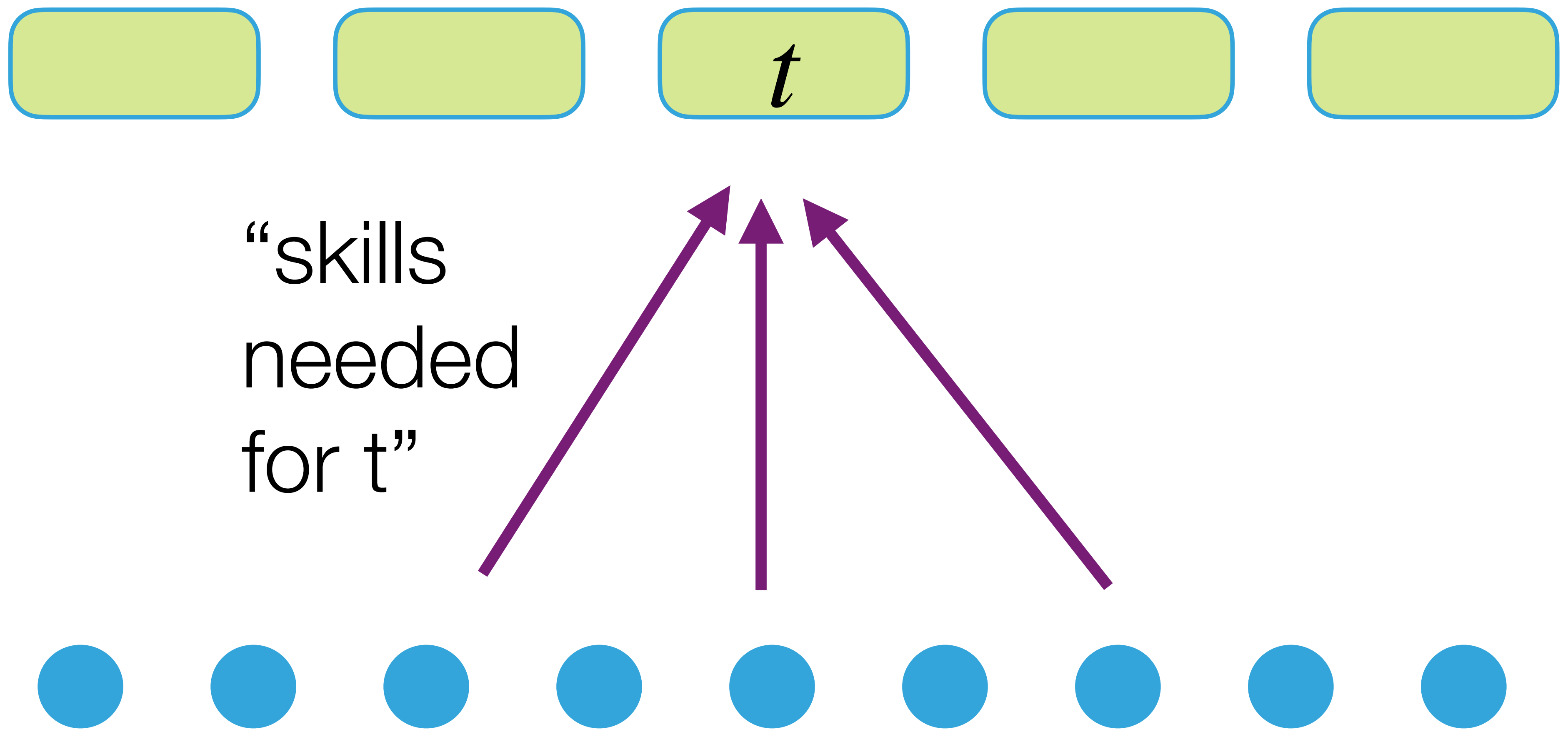
Scaling law $\Longrightarrow$ As LLMs are scaled up, they develop better understanding (e.g., that glasses more likely to shatter than to bounce)

Sketch of Skills view, and connection to “emergence” of compositional generalization

Modeling "text corpus" and "skills"

$\mathscr{T}$ = **Text-pieces**

$t \sim \mu_2$



$\mu_2(t)$ = probability of t (unknown!)

"skills needed for t"

"Skills"could be linguistic, logic, science; common sense, theory of mind,..

Basic Skills

$\mu_1(s)$ = Prob. of skill s (unknown!).

To test understanding of t "Nature" adds cloze question(s) to it (via **unknown** process)

Statistical Task associated with skill s

" Pick random text-piece adjacent to s; answer its cloze questions"

Competence on skill s = Success rate at this statistical task

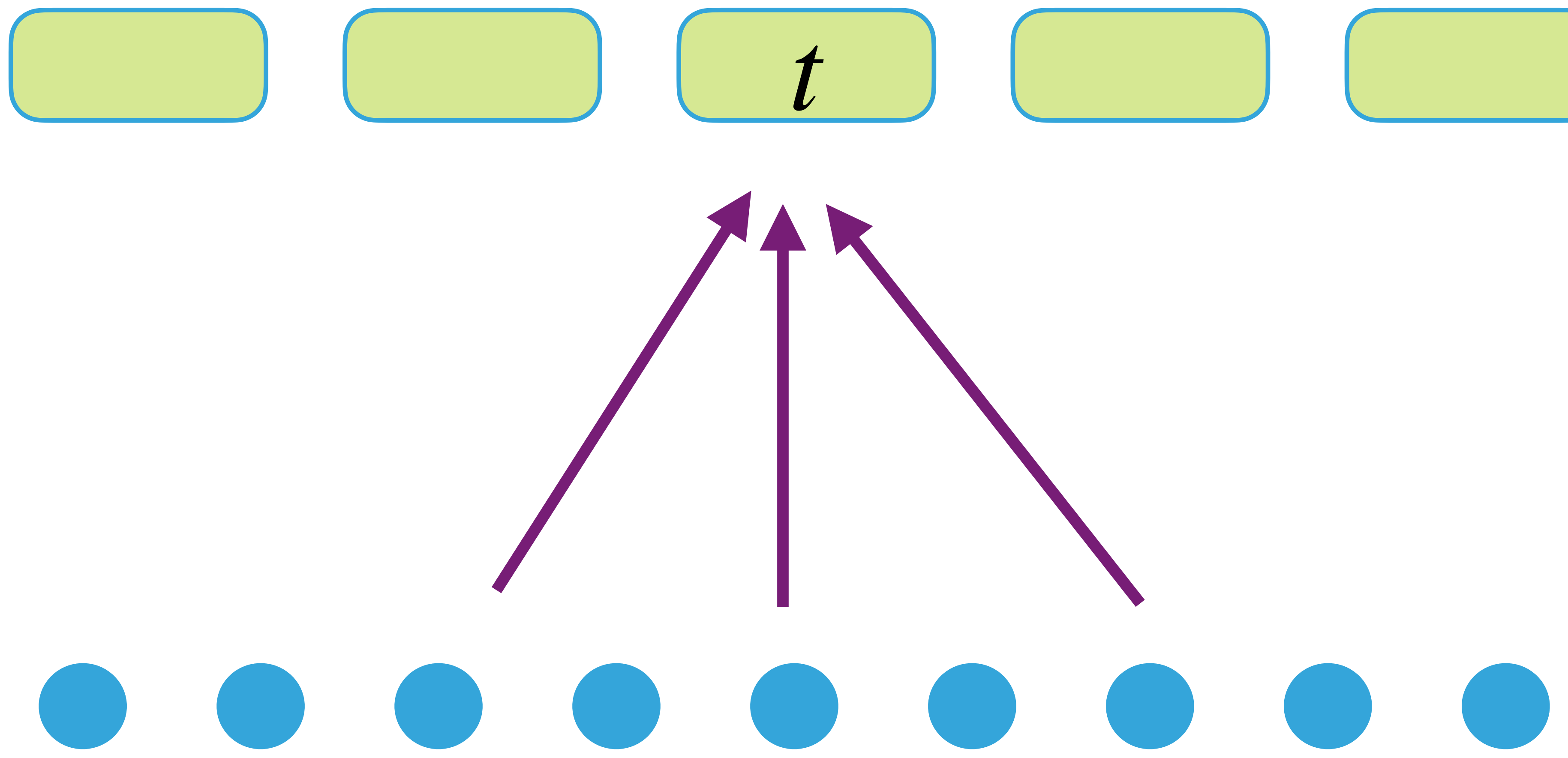
Simple Task

(uses one basic skill)

“Complex tasks” associated with skill-tuples

$\mathscr{T}$ = **Pieces of text**

$t \sim \mu_2$



$\mu_2(t)$ = probability of t

“Skills”could be linguistic, logic, science; common sense, theory of mind,..

Latent Skills

$\mu_1(s)$ = Prob. of skill s

To test understanding of t “Nature” adds cloze question(s) to it (via **unknown** process)

Statistical Task associated with ~~skill s~~ skill pair (s_1, s_2)

“ Pick random text-piece adjacent to both s_1, s_2; answer its cloze questions.”

Competence on pair (s_1, s_2) = Success rate at this statistical task

2-complex Task

(uses 2 basic skills)

Illustration

(Winograd Schema)
The city councilmen refused the demonstrators a permit because they feared violence. *Q) Who feared violence? A) councilmen (B) demonstrators*

Suppose nature produced this text using a 5-tuple of skills.

Then this piece of text appears in the distribution for:

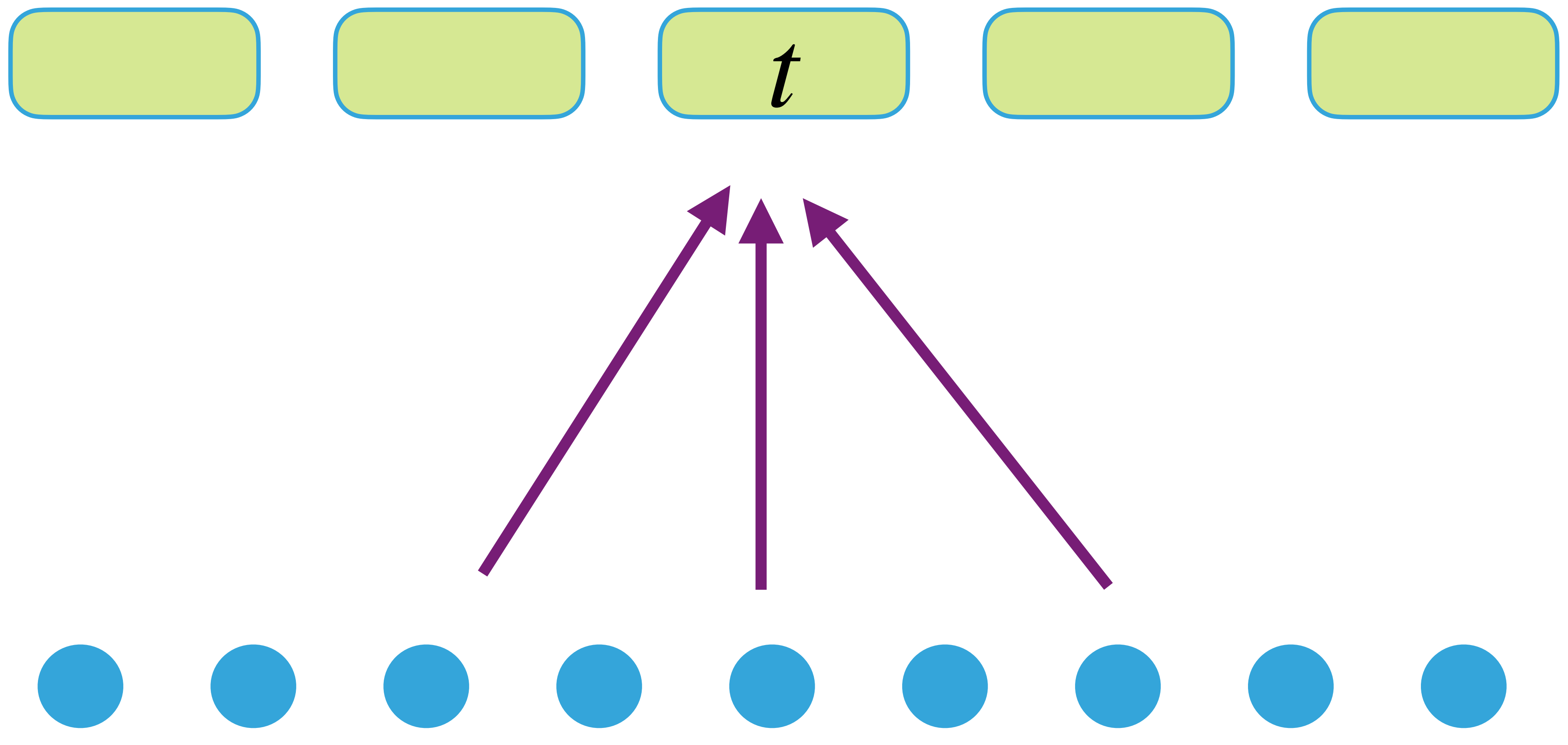
- 5 statistical tasks corresponding to those basic skills
- $\binom{5}{2}$ statistical tasks corresponding to 2-complex skills
- $\binom{5}{3}$ statistical tasks corresponding to 3-complex skills, etc.

Intuition says 3-complex skills are harder to learn than 2-complex, etc.

Key assumption: Mixing of skills

$\mathscr{T}$ = **Pieces of text**

$t \sim \mu_2$



$\mu_2(t)$ = probability of t

"Skills"could be linguistic, logic, science; common sense, theory of mind,..

Latent Skills

$\mu_1(s)$ = Prob. of skill s

To test understanding of t "Nature" adds cloze questinos(s) to it (via **unknown** process)

1. **[Mixing Assumption]**: ``Nature" picks k-tuple of skills **iid** from measure μ_1, uses **unknown** process to convert into text-piece t, with associated probability $\mu_2(t)$.

2. **[Cloze Sufficiency assumption]**:

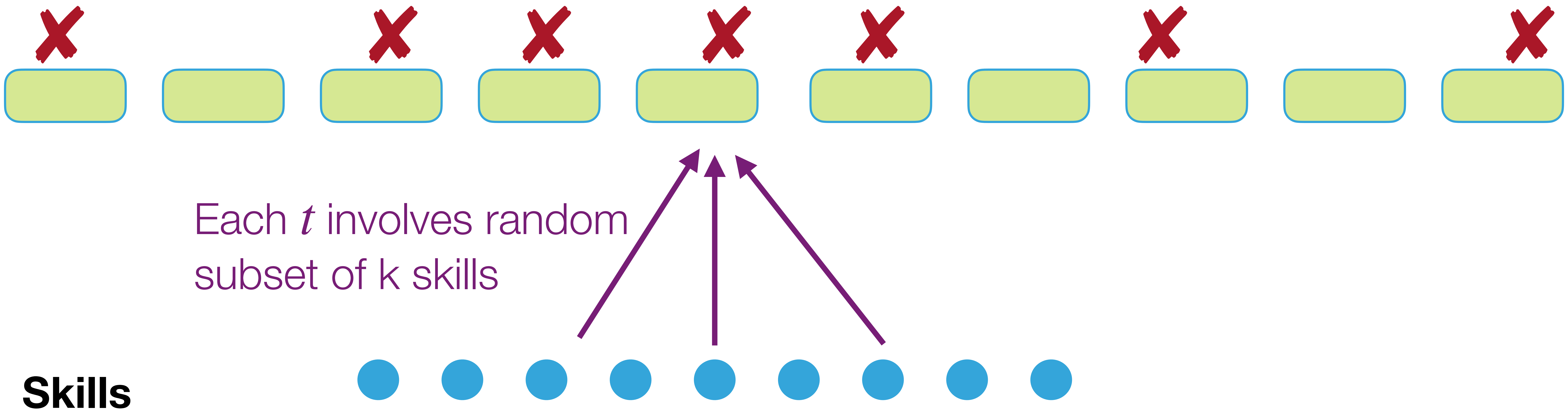
   Model's Avg error in cloze prompts ≈ KL Divergence **(hence scaling LLMs improves ability to answer cloze questions)**

PRINCETON Language + Intelligence

Key Calculation

$\theta =$ fraction of text pieces labeled ✗

Pieces of text $t \sim \mu_2$

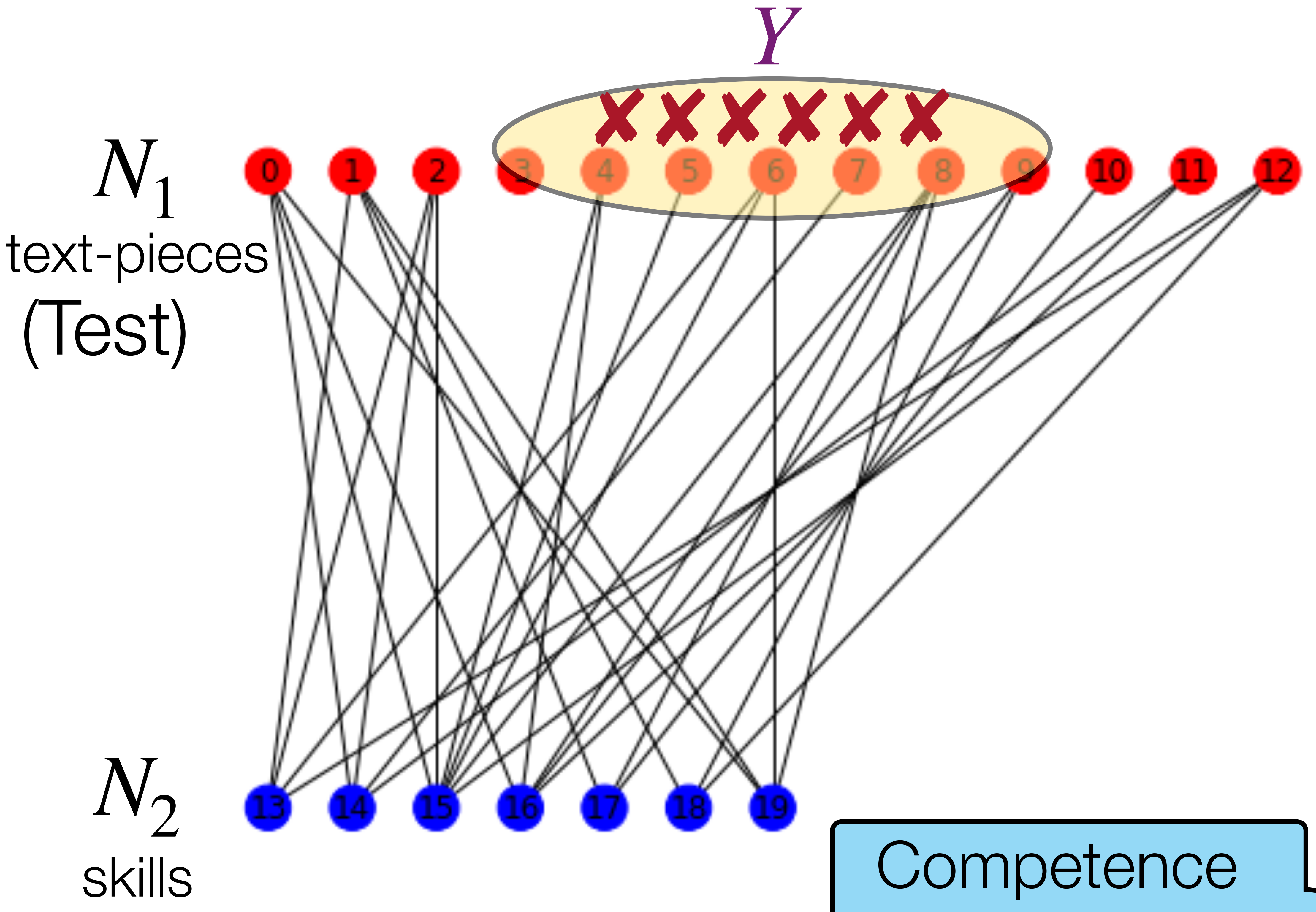


Scaling up the model reduces errors. ($\theta \rightarrow \theta/2$ when model is scaled up 10x)

How does this improve competency on tasks related to skills and skill-tuples?

✗ = cloze question incorrectly answered in this text-piece (so excess cross-entropy is $> \log_2(3/2)$ (say))

Key Calculation (via Random graph theory)



Y = Text pieces with errors. $(|Y| = \theta N_1)$

$\Longrightarrow$ Competence on a skill = fraction of its edges that do not go into Y

Competence curve (α, β, θ)

Theorem: For at least $(1-\alpha)$ fraction of skills, $\leq \beta\theta$ fraction of their edges go to Y,

$$H(\theta) + k\theta(H(\beta\alpha) - \beta\alpha \log \frac{1}{\alpha} - (1-\beta\alpha)\log(\frac{1}{1-\alpha})) = 0$$

"Entropy" $H(x) = x\log_2 1/x + (1-x)\log_2 1/(1-x)$

Proof Idea: Use **probabilistic method** to show the above holds whp for **all** Y of size θN_1

NB: minimum guaranteed competence. (could be better in practice)

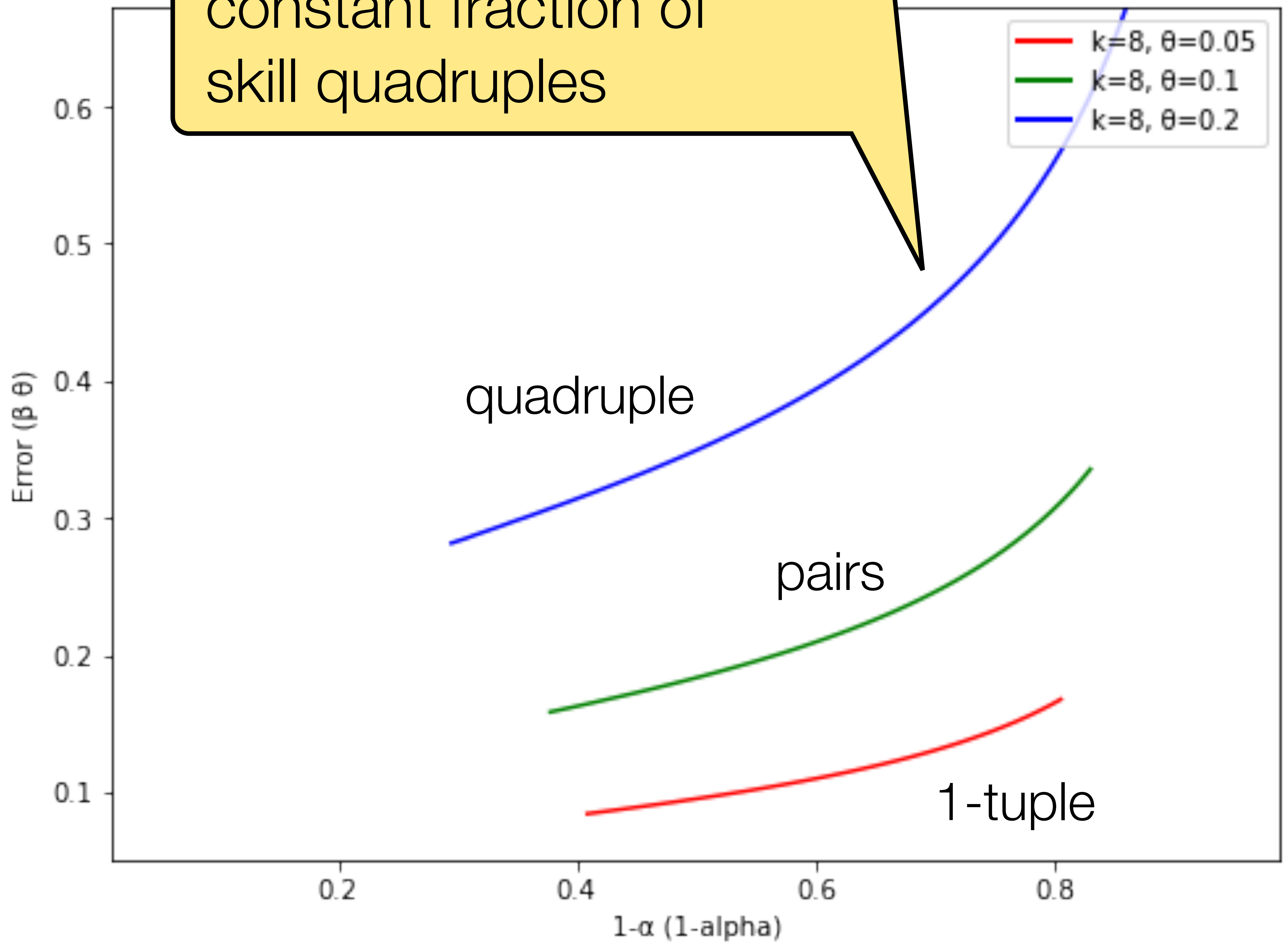
Emergence Law for k'-complex skills

(uses tensorization argument)

"If competence on k'-tuples of skills is currently described by some curve, then after 10x scaling of the model the same curve holds for competence on $2k'$-tuples"

$(\#\text{skills})^{k'}$ could be $\gg$ training corpus size

nontrivial performance on constant fraction of skill quadruples



("lower curve is better")

.. *poverty of stimulus!*

N. Chomsky

"Emergence" phenomenon is fascinating. The full range of LLM capabilities (and how training affects them) is still being mapped

Happy to chat more about skill-based view

Later in the term: LLM Metacognition