

DRAFT

---

\*Princeton University

†Meta

‡Work done while at Google DeepMind

# A Theory for Emergence of Complex Skills in Language Models

Sanjeev Arora<sup>1,3</sup>, Anirudh Goyal<sup>2,3</sup>

This manuscript was compiled on September 13, 2024

1

## 1. Introduction

As language models scale up, via an increase in both the number of parameters and the size of the training datasets, they exhibit remarkable capabilities (1–4). This phenomenon was termed *emergence*, though today it is generally accepted that the improvement in capabilities from scaling is gradual rather than sudden (5). Some emergent properties were noticed by early model designers and have since been confirmed through experiments with substantially larger models such as GPT (1, 6), PaLM (7) and PaLM-2 (8). Two meta capabilities that emerge are *In-context Learning* (1) and *Zero-shot Learning*, whereby the model can understand task instructions given as part of its input and solve the task.

However, we lack any understanding of what capabilities will or will not emerge from scaling, and this has motivated many discussions concerning AI safety and alignment. A contrary opinion –related to the “stochastic parrots” view of LLMs (9)– downplays such concerns, holding that all model capabilities already appear in the training data. Larger models are trained with more text, and therefore we should not be surprised that they learn a broader set of behaviors.

Well-known challenges stand in the path of mathematical understanding of this issue. Rigorous analyses of deep learning exist mostly in simplistic settings, whereas emergence happens in very large models. Furthermore, a necessary prerequisite for a mathematical analysis of emergence of certain capabilities is to agree on a precise mathematical *definition* of those capabilities, which we still lack despite decades of work in linguistics and related fields such as cognitive science.

**Compositional capability:** This paper sidesteps the definitional conundrum by focusing on emergence of *compositional* ability: the model’s capability to solve new tasks by *combining* more basic skills. Letting  $S$  denote an underlying set of basic skills, we are interested in understanding how/why scaling up the model might give it the ability to solve a new task that requires a  $k$ -tuple of skills  $(s_1, s_2, \dots, s_k)$ , where each  $s_i \in S$ . This question is nontrivial because the number of skill  $k$ -tuples scales roughly as  $|S|^k$ , which even for  $k = 5$  (say) may be larger than training corpora available today. Thus for most tasks that involve combining  $k$ -tuples of skills, no corresponding training data exists in the corpus. Thus compositional ability must be learnt in a different way.

Note that the issue of “insufficient training data” is related to the old *poverty of the stimulus* argument, which motivated Chomsky’s theories about compositional structure of human text, such as context-free grammars (10).

**A. Overview of theoretical framework.** Our analysis will relate emergence of composition capability to improvements in the familiar language modeling objective, the loss on next-word

prediction. We now sketch the main ideas, including how we overcome mathematical hurdles that arise.

(i) The incomplete state of mathematical theory of deep learning will be circumvented by assuming *LLM Scaling Laws*, which are experimentally derived relationships describing how scaling up LLMs leads to reduction in language modeling loss. (Section 2).

(ii) Section 3 conceptualizes the “language distribution” as an arbitrary and unknown distribution over finite pieces of text, called *text-pieces*. These are created by “nature” via an *unknown* process that starts with a set of underlying skills, and generates text pieces by picking random  $k$ -tuples of skills for some fixed  $k$ , converting them in some arbitrary way to a text-piece (with an associated probability) whose comprehension requires the  $k$ -tuple of skills.

(iii) We assume that “comprehension” of a text-piece is testable by simple multiple-choice (“cloze”) questions (see Section 2) inserted by another (unknown) process into the text at **test** time. The model does not need to predict the content of the questions—it only needs to guess the answer.

The model’s failure rate at answering these injected multiple-choice questions is assumed to track the language-modeling loss. This assumption is intuitively justified using some notions from information theory (Section ??).

(iv) For any  $k' < k$  and any  $k'$ -tuple of skills, there is a *canonical* task associated with that  $k'$ -tuple, defined as follows: from nature’s distribution on text-pieces pick at random a text-piece whose understanding requires this particular  $k'$ -tuple of skills  $(s_1, s_2, \dots, s_{k'})$ . The task consists of answering cloze questions occurring in that text-piece. The *success rate* at this task is a measure of the model’s capability at combining the skills  $(s_1, s_2, \dots, s_{k'})$ .

This idea of defining “capabilities” on skill-tuples in terms of next-word prediction (specifically, outputting answers to

### Significance Statement

A driving phenomenon of AI today, albeit not well-understood, is that new capabilities emerge in language models when their parameter set and training corpora are scaled up. Mechanistic explanation via mathematical analysis of gradient-based training seems difficult. The current paper does a different analysis relying upon the famous (and empirical) Scaling Laws of LLMs. Contributions include: (a) A statistical framework relating LLMs’ cross-entropy loss to their competence on the basic skills that underlie language tasks. (b) Proof the model’s ability to combine a set of skills can emerge from scaling the LLM, despite lack of even a single such example in the training data. The mathematical techniques are inspired by random graph theory.

multiple-choice cloze questions) is a way to side-step the above-mentioned definitional conundrum. See Section 3 for further discussion.

(v) According to the Scaling Laws, when the model is scaled up this leads to a predictable reduction in language modeling loss. In our setup this implies similar reduction in loss on the task of answering cloze questions at test time.

(vi) Random graph theory is used to show in Section 4 that scaling up leads to improvement in the model’s ability to combine skills. This is applicable because of the assumption that text-pieces are generated using random subsets of  $k$  skills.

**B. Results and Key Takeaways.** With the problem set up as above, the following is a key takeaway from Theorem 13.

**(Informal phrasing of Corollary 12):** *If models of size  $N$  have a certain success rate at solving tasks that require  $k'$  skills, then Scaling Laws imply that scaling up models by one order of magnitude (i.e., to size  $10N$ ) will give them at least the same success rate on tasks that involve combining  $2k'$  skills.*

The proof involves an adaptation of classic methods of random graph theory to our unusual setup, when the system includes some unknown processes (see Figure 1).

As noted earlier, such results imply that LLMs can learn capabilities that had not appeared in their training data. Section 6 describes recent experiments confirming the above prediction about compositional generalization, as well as other surprising findings about LLM capabilities and skills.

**Applicability to conceptual frameworks for next-generation AI models:** In addition to our mathematical framework, we hope our conceptual framework of thinking about skills and their emergence will guide future developments in LLMs. For instance, recent paradigms for creating AI models involve a shift away from the old paradigm of simply training a model to predict the next word on a corpus of internet text. Instead, the “corpus” is a carefully weighted/curated combination of data, which could include code, math/logical reasoning, images, video etc. Training could involve new kinds of losses. Our conceptual framework can accommodate many of these. It is agnostic about what “skills” and “text” are, and how they compose. It is not wedded to a fixed language distribution (see the arbitrary measure  $\mu_2()$  in Definition 6) nor to a specific prediction loss such as cross-entropy.

**Paper organization:** Section 2 provides a brief introduction to scaling laws, emergence, and random bipartite graphs. Section B explores the connection between the reduction in excess cross-entropy and learning. Sections 4 and 5 give concrete mathematical theorems about emergence of skills as a result of scaling. Section 6 describes new experimental works that support the theory presented in the current paper and have appeared after the initial arxiv draft of the current paper.

## 2. Preliminaries

Language models take a statistical view of language. Pieces of text from a sufficiently large and diverse corpus are assumed to be samples from a probability distribution (11). This assumes a ground-truth (i.e., humans’) distribution for generating the next word, which assigns probability  $p_i(w \mid w_1 w_2 \dots w_i)$  to the event that the  $(i+1)$ th word is  $w$  given that the previous words were  $w_1 w_2 \dots w_i$ . Thus the probability it assigns to a

word-sequence  $w_1 w_2 \dots w_N$  is

$$p(w_1) \prod_{i=2}^N p(w_{i+1} \mid w_1 \dots w_i) \quad [1]$$

The training corpus is assumed to be sampled from this (unknown) distribution.

Training involves solving *next-word-prediction* task given the previous  $N$  words, where  $N$  can be several thousand or even more. \*. Thus the model  $M$  computes a probability  $q[w_{i+1} \mid w_1 w_2 \dots w_i]$  for all possible words  $w_{i+1}$ ’s. To improve its next-word-prediction one does gradient descent on *cross entropy loss*, which involves predicting the next word given the previous  $C$  words, where  $C$  is the context-length:

$$\ell(M) = - \sum_i \log q[w_{i+C} \mid w_i w_{i+1} \dots w_{i+C-1}] \quad (\text{Cross Entropy}) \quad [2]$$

The goodness of the trained model is computed by the *test loss*—which is same loss expression on held-out text from the same corpus.

**Scaling laws:** These empirically-derived expressions describe how test cross entropy loss on held-out data scales (in experiments) with number of model parameters ( $N$ ) and size of the dataset ( $D$ ) (12–15). For Chinchilla models (16) the law is as follows:

$$L(N, D) = A + \frac{B}{N^{0.34}} + \frac{C}{D^{0.28}} \quad (A = 1.61 \ B = 406.4, \ C = 410.7). \quad [3]$$

Here the constants  $A, B, C$  in (3) hold only for the specific architecture and training strategy—even the constant  $A$  depends upon the tokenization. This description of macro behavior using two basic parameters will help us circumvent the need for mechanistic understanding of training and generalization. Our theory will only rely upon the general form of the equations, specifically, that the dependence is inverse polynomial in  $N, D$ . So it applies to other frameworks of training (e.g., overtrained models (17)) where scaling laws have also been found.

**Emergence:** Emergence refers to an interesting empirical phenomenon that as  $D, N$  are increased together (accompanied by careful changes to other hyperparameters according to published recipes) then the model’s performance—zero shot or few-shot—on a *broad range* of language tasks improves in a correlated way. The improvement can appear as a quick transition when  $D, N$  are plotted on a log scale (which is often the case) but it is now generally accepted that for most tasks the performance improves gradually when  $D, N$  are scaled up. Thus the term *slow emergence* is more correct. Furthermore, it is known that emergence happens at different rates for different tasks, and is often quickest for tasks where the text is plausibly close to text found in training data (4). Plenty of tasks are known that stump current models, and they usually tend to be very different from what one would find in usual text corpora. See (3, 4, 18) for experimental results on emergence rates of the broad range of language tasks. One might thus posit, with some justification from the above-mentioned studies, that the emergence of skills arises from training on related tasks that were implicitly solved while solving next-word prediction in the training dataset. This is indeed our starting point.

\* Actual training involves breaking up words into smaller *tokens*, which allows a single model to handle all human languages, math formulae, computer code, etc. For simplicity, our discussion will refer to “words.”

**A. Cross-Entropy, Entropy, and Excess entropy.** Continuing in the notation of (1) we sketch the well-known conceptual framework underlying cross-entropy loss (2). We shorten  $p_i(w | w_1 w_2 \dots w_i)$  to  $p_i(w)$ . Thus the *entropy* of the  $(i + 1)$ th word

$$\sum_w p_i(w) \log \frac{1}{p_i(w)} \quad (\text{ENTROPY}) \quad [4]$$

This entropy is an inherent property of language, due to existence of many possible choices human writers can make for the next word. Given sequence  $w_1 w_2 \dots w_i$  the model has a probability distribution  $q(w | w_1 w_2 \dots w_i)$  for the next word  $w$ . Extending our compact notation, we use  $q_i(w)$  as a shorthand for this. The cross-entropy loss of the model on the  $i + 1$ th word is  $\log \frac{1}{q_i(w_{i+1})}$ , which should be seen as an empirical estimate of

$$E_{w \sim p_i(\cdot)} [\log \frac{1}{q_i(w)}] \quad (\text{EXPECTED C-E LOSS}) \quad [5]$$

*KL divergence*, also sometimes called *excess entropy*, is non-negative and defined as

$$KL(p_i || q_i) = E_{w \sim p_i(\cdot)} [\log \frac{p_i(w)}{q_i(w)}] \quad (\text{EXCESS ENTROPY}) \quad [6]$$

Thus on a per-word basis we have:

$$\text{EXPECTED C-E LOSS} = \text{ENTROPY} + \text{EXCESS ENTROPY} \quad [7]$$

Summing over the entire held out corpus, one obtains a similar estimate for the entire corpus. One can make mild assumptions to the effect that the conditional probabilities  $p_i(\cdot), q_i(\cdot)$  only depend only on (say) the previous  $10^3$  words, whereas the corpus size  $M$  is much bigger, e.g.,  $M \gg 10^7$ . So the corpus consists of a random walk of sorts, where every  $10^4$  words or so it switches to a different portion of the language distribution. Under such assumptions the above relationship, which holds in expectation at the word level, should hold fairly precisely at the corpus level.

**Data selection:** Modern training pipelines involve *data selection* which is a *re-weighting* of data depending upon its perceived utility for language modeling. This can be viewed as a reweighting of the above cross-entropy loss. Our mathematical framework will allow text-pieces to have an arbitrary measure.

**Understanding the Scaling Law in terms of excess entropy:** In (3) the  $A$  term captures the entropy of language<sup>†</sup>. No model, however good, can achieve lower cross-entropy loss than  $A$  for large corpora. The second and third terms of (3) capture *excess entropy*, and they decrease polynomially with  $N$  and  $D$ . For example when  $N, D$  are increased by a factor of 10 it reduces by roughly  $(10)^{0.28} \approx 2$ .

Section B will argue that reductions in excess entropy lead to improvements in model capabilities. Note that there is no direct way to compute/optimize excess entropy from the corpus. Relationship (7) shows that estimating excess entropy—i.e., looking at cross-entropy(2) and identifying how much of it is due to inherent entropy—requires humans in the picture.

<sup>†</sup> Here we're assuming that as the model and data set size tend to infinity in tandem, the model will perfectly learn the language distribution.

## B. (Mis)understanding, Excess entropy, and Cloze Questions.

Thinking about emergence and Scaling Laws, it is possible to get confused as follows: “When we increase  $D$  from  $10^{11}$  to  $10^{12}$  then according to (3) this changes cross-entropy by a tiny amount. Why does it lead to big changes in macroscopic behavior?” The flaw in this reasoning is that most of the loss captures merely the inherent entropy of language (the  $A$  term in (3)). We argue now that the model’s mistakes on downstream tasks (i.e., its misunderstandings) are captured by the *excess* entropy, which as noted in Section A reduces by a constant factor each time the model is scaled up by an order of magnitude<sup>‡</sup>.

We illustrate using a classic example from (20), which later inspired the *Winograd Schema Challenge (WSC)* (21):

The city councilmen refused the demonstrators a permit because they feared violence.

Here the pronoun **they** is ambiguous—grammar rules allow it to refer to either **demonstrators** or **city councilmen**. Winograd pointed out that disambiguating it (i.e., anaphora resolution) requires world knowledge that is unavailable in the text itself, namely that demonstrations can get violent, and city councilmen don’t like violence.

A key idea in designing test-beds for language understanding such as WSC is the **Cloze Procedure**<sup>§</sup>, popular also for testing language development in children (23). To test the model’s understanding of **they** in this sentence, we can append a *prompt*: Q. Who feared violence?. This is followed by either a blank, or a choice of multiple answers: A. city councilmen. B. demonstrators. For WSC examples, even though a human would be hundred percent sure of the answer, language models circa 2016 were roughly 50/50 confused between the two options.

In the above example, the human is 100% certain of the answer, which implies their entropy here is  $\log 1$ , namely 0. However if the model is split 50-50 between the two options this implies it has cross-entropy  $\log 2$ , all of which is *excess entropy*! Given the frequency of ambiguous pronouns in usual English, one concludes that a model that has not learned pronoun disambiguation will display huge excess entropy at many places in surrounding text. Thus reductions in excess entropy (which happen naturally due to scaling) will tend to squeeze out such errors. The rest of the paper tries to make this intuition mathematically precise.

Of course, text corpora do not normally contain such artificial cloze questions. But one could imagine that the model’s basic misunderstanding of the above type could, often, lead to prediction mistakes in neighboring text. Our theory in Section 3 will assume that cloze questions can closely capture the model’s misunderstanding.

## 3. A Skill-based Framework

Section 2 sketched usual statistical view of language modeling. Here, we desire a way to think in a differentiated way about the corpus as consisting of small portions, each using a small

<sup>‡</sup> A recent empirical study (19) also concludes with the finding that “language modeling perplexity correlates well with few-shot in-context learning performance along the trajectory, regardless of model sizes.” At the same time, it is known that two models with the same cross-entropy can differ somewhat in their performance on language tasks.

<sup>§</sup> Cloze questions are multiple choice, which allows testing most language skills (1). Some skills such as understanding of irony don’t lend themselves well to cloze-based testing since one of the multiple choices already explains the joke. See (22) for earlier use of Cloze prompts in developing a theory of LLMs.



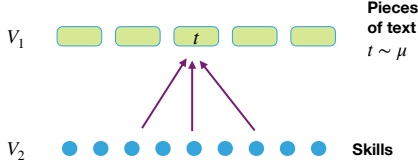


Fig. 1. Bipartite graph showing which skills are relevant to each text-piece

set of skills. So it will be more convenient think of the corpus as a long unordered list of text-pieces, each with an associated measure. This measure is a weighting of the text-piece in the original distribution and can be used to compute the average loss over the entire distribution.

**Definition 1** (Text piece). The test corpus for the model is viewed as being divided into distinct *text-pieces*, each consisting of  $C_{test}$  tokens. There is also a measure  $\mu_2()$ , with  $\mu_2(t)$  denoting the measure of text-piece  $t$ . The usual per-word cross-entropy loss for the entire corpus is computed by a weighted average of the cross-entropy loss of the text-pieces according to their measure.

**Noted:** (i) When a text-piece is small, its full comprehension may require reading the text that preceded it in the larger corpus. Such preceding text can be included with the text-piece, but the loss calculation for the text-piece is only done on words in the text-piece. (ii) The measure  $\mu_2(t)$  captures the ground-truth probability of the text-piece in the language (or reweighted probability, if training involved data selection).

Now we make some assumptions. Given lack of mathematical understanding of LLM skills and the meaning of composition, We assume that the model’s “comprehension” of a text piece is testable via suitable cloze questions analogous to the Winograd example in Section B. Specifically, we assume that an (unknown) process CLOZE has been used to add clearly-marked multiple-choice questions in simple English, which the model has to answer at test time. This is a simple form of distribution shift at test time. The prediction loss on cloze questions does not require predicting the location or contents of the cloze question—it only requires selecting the correct answer to the multiple-choice cloze question. We allow the process CLOZE to tailor the questions to the model being tested. Thus the next assumption is reasonable.

**Assumption 2.** [*Cloze Sufficiency Assumption:*] The pre-trained model’s average (multiclass) prediction loss on Cloze questions—where the average is taken over the distribution of text pieces—closely tracks (within a small multiplicative factor like 1.1) the excess cross-entropy of the model on classical next-word prediction.

**Note:** As discussed in Section B, if the cloze question is assumed to be perfectly answerable by a human then any incorrect answers by the model can be interpreted analogously as excess cross entropy. Our assumption amounts to saying that mistakes on cloze questions closely capture the excess entropy of the model as defined in (2). The next theorem, shows that there *exists* a set of cloze questions (albeit fairly artificial) where the excess cross-entropy of the model’s answer tracks the overall excess cross-entropy on next-word prediction.

SAY SOMETHING ABOUT HOW TO MEASURE FAILURE RATE IN CLOZE ANSWERS

**Theorem 3.** If a model’s excess entropy at the  $i$ th place in text is  $\epsilon$  then there is a cloze question with binary answer such that the probability that the model answers it incorrectly is at most  $\sqrt{2\epsilon}$ .

*Proof.* The proof involves Pinsker’s Inequality (wikipedia version) which relates variation distance and KL divergence. As in Section B let  $p_i()$  be the humans’ probability distribution for the  $i + 1$ th word in the text piece and  $q_i()$  be the model’s distribution. The probability that the human and the model give different answers is the variation distance between the two distributions, which is the maximum (over all subsets  $A$  of words) of  $\sum_{w \in A} (p_i(w) - q_i(w))$ . Let  $A_{i+1}$  denote the subset for which the previous expression is maximised. The cloze question consists of replacing word  $w_{i+1}$  in the text with the question: *Is the next word among the words listed in option (a) or in option (b)*, where option (a) lists words in  $A_{i+1}$  and (b) lists words in  $\overline{A_{i+1}}$ . The theorem now follows from Pinsker’s inequality.  $\square$

**A. Skills: A Statistical View.** Language is assumed to have an underlying set  $S$  of *skills*. Every text-piece  $t$  has an associated set of skills that are required for comprehending it. The theory allows this set of skills to be arbitrarily large, but somewhat smaller than the number of text-pieces in the distribution (which is an enormous number).

**Definition 4** (skill graph). A *skill graph* is a bipartite graph  $(S, T, E)$  where nodes in  $S$  correspond to skills, nodes in  $T$  correspond to text-pieces, and  $(s, t)$  is in the edge set  $E$  if “comprehending” text-piece  $t$  (i.e., answering its associated cloze questions) requires using skill  $s$ . (See Figure 1.) There is a measure  $\mu_2(\cdot)$  defined on the set of all text-pieces.

We are interested in quantifying the model’s *competence* on particular skill. This competence is not binary. For instance, say the skill is “to disambiguate pronouns from the context,” then even humans will make some mistakes at it. Thus “competence” is a fraction, quantified as follows.

**Definition 5** (Competence on skills). For a skill  $s$ , the *competence* of a model is the expectation of the following random variable: randomly sample a text-piece containing that skill (this sampling uses the measure  $\mu_2(\cdot)$  on text-pieces) and measure the model’s success-rate at answering cloze questions in that text piece. We similarly define Competence on a tuple of skills  $(s_1, s_2, \dots)$ .

Thus far our framework is a minor restatement of the standard statistical view of language models. As noted, to analyse emergence we need additional assumptions, and ours are as follows: (a) Each text-piece requires  $k$  skills<sup>¶</sup> (b) the process of generating text-pieces picks  $k$  skills independently from an underlying distribution.

The intuition for assumption (a) can be illustrated using Winograd’s example of pronoun disambiguation in Section B. While the example focuses on pronoun disambiguation, applying the skill in that context requires other skills: understanding of causality (i.e., interpretation of “because”) as well as world

<sup>¶</sup>The definition involves the  $k$ -tuple being picked by iid sampling from measure  $\mu_1$  which, in principle, allows a skill to be picked twice. However, the probability of picking the same skill twice scales as  $O(1/|S|)$ , where  $S$  is the set of skills. Since  $S$  is assumed to be large, the distribution is almost the same as sampling  $k$  distinct skills. The small difference of  $O(1/|S|)$  between the two methods will not affect any of the random graph theory calculations.

knowledge about “city councilmen,” “permit,” “demonstrators,” etc. In other words, text-pieces in the corpus that require pronoun disambiguation will involve real-world scenarios whose comprehension requires all sorts of other skills as well. Moreover, the scenarios (and hence the relevant skills) could have different probabilities of occurring in the corpus, which is allowed in our framework.

We emphasize that the rest of the process, specifically how the generation process transforms tuples of skills into text with associated measure, does not need to be specified.

**Definition 6** (Degree- $k$  skill cluster). This is a skill graph  $(S, T, E)$  (see Figure 1) where the collection of text pieces  $T$  is generated by “nature” by applying the following process: pick a subset of  $k$  skills from  $S$  via iid sampling from an underlying measure  $\mu_1$  on skills, and then use a procedure GEN to create a text-piece  $t$  whose comprehension requires these skills, as well as a measure  $\mu_2(t)$  associated<sup>1</sup> with this text piece  $t$ . Then “nature” uses process CLOZE to add cloze prompts to test comprehension on  $t$ . The *prediction loss* on the text-piece is the cross-entropy loss on predicting the answers to the cloze questions in it. The average prediction loss over all text-pieces is computed with respect to the measure  $\mu_2()$ . We call the skill-graph thus produced a *degree- $k$  skill cluster*.

The assumption that any skill could combine with any other skill is simplistic. In real life corpora, skills used in understanding literature may not often combine more often with each other, and less often with skills used for math. Analyses of such complicated frameworks, e.g., with hierarchies of classes with richer co-occurrence structure is left for future work.

**Note:** Definition 5 should be viewed as defining a lower-bound on the competence at applying skills. To illustrate, suppose comprehending a text-piece involves 5 skills. Then that text-piece will appear in 5 statistical tasks corresponding to individual skills,  $\binom{5}{2}$  tasks corresponding to pairs of skills, and so on. However, our method of measuring the loss incurred on these statistical tasks implicitly assumes that if the model incorrectly answered this cloze question (i.e., it assigned significant probability to the wrong answer), then that loss was incurred in *all* these statistical tasks. This accounting is conservative—it gives the model no credit even if it perfectly understood skills 1 to 4, but ended up failing the cloze question because of shaky understanding of skill 5. Such conservative accounting is needed because of absence of general mathematical formulation of LLM skills and what it means to combine them.

#### 4. Analysis of Emergence (uniform cluster)

Having set up a framework for modeling skills and (via Assumption 2) connecting them to the cross-entropy loss of the model, we have arrived at a core mathematical issue around emergence: *As the model’s excess cross entropy goes down due to scaling, this improves, by Assumption 2, the model’s performance on cloze tasks inserted in the text-pieces of the full distribution (most of which was not seen during training). At what rate does this improve the model’s competence on the skills as well as on tuples of skills?*

<sup>1</sup> Note that the measure on text-pieces has to have the correct marginals e.g., the  $\mu_2$ -measure of all text-pieces containing a skill  $s$  is  $\mu_1(s)$ . There are many measures satisfying this weak condition, since the number of text pieces is way larger than the number of skills.

This section analyzes a simple setting where the distributions  $\mu_1$  and  $\mu_2$  in Definition 6 are uniform. Section 5 will extend the analysis to the general setting.

**Key Hurdle:** We point out the naive—but incorrect—way to reason about this. Since each text piece is connected to a random  $k$ -tuple of skills, say  $\vec{s}$ , one is tempted to reason about emergence via linearity of expectations, specifically, the following relation about prediction loss, where “expectation” is just average over text-pieces/skills with respect to their measure:

$$k \cdot E_t[\text{loss}(t)] = E_s[\text{failure rate of statistical task } \tau_s]. \quad [8]$$

To see that this is incorrect, let  $Y$  be the subset of such text pieces where the model makes mistakes on cloze questions. This  $Y$  depends upon the skill graph, and the unknown processes GEN and CLOZE of Definition 6, which assign measure to text pieces in an unknown way that may introduce arbitrary correlations. Since the model “saw” part of the test stream (namely, the portion corresponding to training data) it has picked some information about the skill cluster. Thus at the end of training, locations of errors in the test stream—i.e., the set  $Y$ —depend upon the skill-cluster, and since we lack understanding of  $Y$  the analysis has to treat it as arbitrary. In other words, our analysis is allowed to assume an upper bound on the test loss, but the text-pieces on which this loss occurs form an arbitrary subset that depends upon the graph structure. In particular, (8) cannot be inferred. This is the key mathematical hurdle and our proof will surmount it using random graph theory.

Let’s say the model *makes a mistake* on a text-piece if the total prediction loss on all the cloze-questions of that text-piece is at least  $1/2$  (which is the kind of error incurred if the incorrect answer is chosen with noticeable probability on even a single cloze question). Say  $Y$  contains  $\theta$  fraction of test-pieces. (Note that once the average cross-entropy loss for the text-pieces is  $\delta$  then  $Y$  contains at most  $2\delta$  fraction of text pieces.) The next theorem guarantees that for most skills  $s$ , the statistical task  $\tau_s$  corresponding to most skills do not assign significant probability to text pieces in  $Y$ . It is a simple rephrasing of Lemma 14 in the Appendix.

**Theorem 7** (Basic). *Let  $\alpha, \beta, \theta > 0, \beta > 1, \alpha\beta < 1, \theta < 1$  satisfy*

$$H(\theta) + k\theta \left( H(\beta\alpha) - \beta\alpha \log \frac{1}{\alpha} - (1 - \beta\alpha) \log \left( \frac{1}{1 - \alpha} \right) \right) < 0 \quad [9]$$

*and the distribution on skills and text pieces be uniform in the skill-cluster. Then irrespective of the details of GEN and CLOZE processes, the following holds for every  $Y$  that contains at least  $\theta$  fraction of all text-pieces: at least  $1 - \alpha$  fraction of skills each have at least  $(1 - \beta\theta)$  fraction of their edges incident to  $\bar{Y}$ .*

Note that as the model is scaled up,  $\theta$  will go down and the set  $Y$  containing erroneous answers on cloze questions will shrink. Our analysis kicks in only once  $\theta$  drops below 1. In terms of the emergence phenomenon, this corresponds to first signs of improvement on downstream tasks once the model’s loss drops below some threshold.

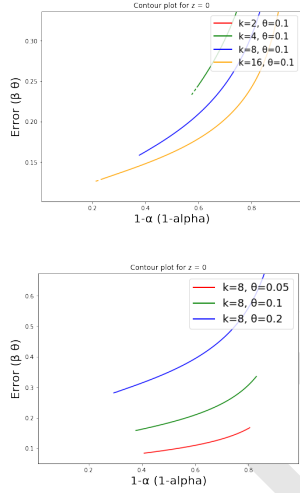
**Definition 8** (performance curve). The contour plot (i.e., the boundary) of the region of  $\alpha, \beta$  combinations satisfying

Theorem 7 is called a *performance curve* and denoted  $C_{(k,\theta)}$ . A performance curve  $C$  is *better* than another curve  $C'$  if for every  $\alpha, \beta$  on  $C$  there is a corresponding point  $(\alpha, \beta')$  on  $C'$  for  $\beta' > \beta$ .

Figure 2 gives *performance curves*, i.e., the contour plot of the set of  $\alpha, \beta$  combinations satisfying Theorem 7 for a given  $\theta, k$ . The horizontal axis plots  $(1 - \alpha)$  and the vertical axis plots  $\beta\theta$ , so point  $(0.8, 0.16)$  on a curve means at least 0.8 fraction of skills have at most 0.16 fraction of their edges in the “error set”  $Y$  (hence 0.84 fraction of their edges are outside the error set). The emergence curves shift down noticeably (i.e., imply emergence of more skills) as we increase  $k$ . The next lemma shows this trend always holds.

**Lemma 9** (Monotonicity properties). *If  $\theta_1 < \theta_2$  then the performance curve for  $\theta_1, k$  lies below that for  $\theta_2, k$ . If  $k_2 > k_1$ , then the performance curve for  $\theta, k_2$  lies below that for  $\theta, k_1$ .*

*Proof.* Both follow by noticing that since  $H(\theta)$  is nonnegative, the only way for (11) to hold is for the expression involving  $\alpha, \beta$  inside the parenthesis to be negative.  $\square$



**Fig. 2.** Performance Curves: Left plot has  $\theta = 0.1$  and varies  $k = 2, 4, 8, 16$ . Higher values of  $k$  greatly improve performance (for  $k = 2$  valid  $\alpha, \beta$  did not exist). The right plot has  $k = 8$  and  $\theta = 0.05, 0.1, 0.2$ . Section A clarifies that it also describes the model’s performance curve for  $t$ -tuples of skills for  $\theta = 0.05$  and  $t = 1, 2, 4$  respectively (e.g., blue curve for 4-tuples).

### A. Emergence curve for $k'$ -tuple of skills via tensorization.

To derive emergence for tasks involving  $k'$ -tuples of skills we use a tensorization argument. Consider the following  *$k'$ -wise recombination* operation. First randomly partition the text-pieces into subsets of size  $k'$ , and then concatenate the  $k'$  text pieces within each subset to create a larger piece of text that we refer to as a “ $k'$ -piece,” and whose measure is the sum of the measures of the component test-pieces. All cloze questions for the old test-pieces are retained and no new cloze questions are inserted. Clearly, if the error of the model per average text-piece was  $\delta$ , then the error per average  $k'$ -piece is  $k'\delta$ . However, each  $k'$ -piece is now using a random  $k'k$ -tuple of skills. Importantly, this set of  $k'k$  skills consists of iid draws from the skill distribution. Treating  $k'$ -tuples of skills

as ‘composite-skills,’ we can reuse the previous calculation of Theorem 7 to obtain the following.

**Corollary 10** (tensorization). *In the same setting as Theorem 7, for integer  $k' \in [2, 1/\theta]$  the conclusion of that theorem holds also for  $\alpha, \beta$  pairs satisfying*

$$H(k'\theta) + k k' \theta \left( H(\beta\alpha) - \beta\alpha \log \frac{1}{\alpha} - (1 - \beta\alpha) \log \left( \frac{1}{1 - \alpha} \right) \right) < 0 \quad [10]$$

*Furthermore, if  $H(k'\theta) < k'H(\theta)$  (which holds for  $\theta > 1/2k'$ ) the emergence curve from this expression dominates that derived from Theorem 7.*

**2nd estimate (better):** Consider the following  *$k'$ -wise recombination* operation on the test stream. First randomly partition the test stream into subsets of size  $k'$ , and then concatenate the  $k'$  text pieces within each subset to create a larger piece of text that we refer to as a “ $k'$ -piece.” All cloze questions for the old test-pieces are retained and no new cloze questions are inserted. Clearly, if the error of the model per average text-piece was  $\delta$ , then the error per average  $k'$ -piece is  $k'\delta$ . However, each  $k'$ -piece is now using a random  $k'k$ -tuple of skills, which we can alternatively view as  $k$  random  $k'$ -tuples. Thus viewing  $k'$ -tuples of skills as ‘composite skills’ we can use this as the skill set in the setting of Theorem 7, which gives us an easy corollary quantifying the performance on tasks corresponding to  $k'$ -tuples of skills.

**Lemma 11** (Emergence for  $k'$ -tuples of skills). *Consider the skill-graph  $(S', T', E)$  where  $S'$  consists of all  $k'$ -tuples of skills,  $T'$  consists of  $k'$ -pieces, and  $E$  consists of  $(s', t')$  where  $s'$  is a  $k'$ -tuple of skills and  $t'$  is a  $k'$ -piece where this tuple of skills is used. Let  $Y$  consist of  $\theta$  fraction of  $k'$ -pieces. Then for any  $\alpha, \beta > 0, \beta > 1, \alpha\beta < 1$  satisfying (13) there are at least  $1 - \alpha$  fraction of  $k'$ -tuples of skills that have at most  $\alpha\beta\theta$  fraction of their edges connected to  $Y$ .*

The next corollary presents a somewhat surprising general principle that’s also hinted at in caption of Figure 2. Assume (for simplicity) a Chinchilla-like scaling law that 10x up-scaling leads to factor 2 reduction in excess entropy. If a model is considered to have reasonable performance on individual skills at current scaling, then after further up-scaling of 10x one would see similar reasonable performance on skill-pairs, and scaling up by yet another 10x after that will yield similar reasonable performance on 4-tuples of skills, etc. Note that these are *provable lower bounds* on performance gains—actual gains could be higher. Figure 2 illustrates the phenomenon.

**Corollary 12.** *When the model  $M_1$  with loss  $\delta$  is scaled up (e.g., as per equation (3)) so that the new model  $M_2$  has loss  $\delta/k'$ , then the performance curve inferred by our method for  $k'$ -tuples of skills using  $M_2$  is identical to the curve inferred for individual skills on model  $M_1$ .*

*Proof.* As noted above, a loss of  $\delta$  still allows the model to make significant mistakes on  $2\delta$  fraction of test pieces, which we denote by  $\theta$ . Thus Theorem 7 describes the performance curve for skills. Making the loss drop to  $\delta/k'$  but creating  $k'$ -pieces makes the fraction of errors  $\theta = 2\delta$  again. (Note that “error” now means an erroneous answer on *any* cloze question in the entire  $k'$ -piece—again, this is a conservative definition of error.) Applying Lemma 11 we get the same emergence curve as Theorem 7.  $\square$



## 5. Emergence analysis with general measure on text and skills

Now we turn to analysis of the general setting of Definition 6 where text piece  $t$  has measure  $\mu_2(t)$  and skill  $s$  has measure  $\mu_1(s)$ . In this setup, our lemma statements (e.g., Lemma 14 as well as the ones in Sections 4 and A) hold —the claim is the same but with cardinalities replaced by measure!

**Theorem 13** (Emergence of skills and  $k'$ -tuples of skills). *Let  $Y$  be any subset of text pieces consisting of text pieces with total measure  $\theta$ , and every text-piece has measure substantially less than  $\theta$ . Let  $\alpha, \beta > 0, \beta > 1, \alpha\beta < 1$  satisfy*

$$H(\theta) + k\theta(H(\beta\alpha) - \beta\alpha \log \frac{1}{\alpha} - (1 - \beta\alpha) \log(\frac{1}{1 - \alpha})) < 0 \quad [11]$$

*Then the measure of skills that have at most  $\beta\theta$  fraction of their edges connected to  $Y$  is at least  $1 - \alpha$ .*

*For  $k'$ -tuples of skills the statement of Lemma 11 holds with the same modification of cardinality to “measure.”*

*Proof.* The measure  $\mu_1$  on skills is trivial to reason about by just replacing each skill  $s$  by a number of copies that is proportional to  $\mu_1(s)$ . This converts the measure to a uniform measure —specifically,  $k$  iid draws from this uniform measure are equivalent to  $k$  iid draws from the  $\mu_1$ .

For the measure  $\mu_2(\cdot)$  on texts, the above trick doesn't work. Recall that a text-piece is connected in the skill graph to a random  $k$ -tuple of skills. If we try to replace  $\mu_2(\cdot)$  with a uniform measure by replacing the text piece with identical copies, then these copies must still all connect to the *same* subset of  $k$  skills —meaning these connections are correlated and not independent. We need a more subtle argument. The key part in the proof of Lemma 14 is where we choose random subset of text-pieces,  $Y$  whose size is  $\theta|T|$  and subset  $Z$  of skills of size  $\alpha|S|$ , and then upper bound by ( ) the expectation of the event that the latter has more than  $\alpha\beta\theta k$  fraction of its edges going to  $Y$ . In presence of measure  $\mu_2(\cdot)$  let's pick  $Y$  as follows: Independently pick text-pieces, choosing  $t$  with probability  $\theta\mu_2(t)$ . (Note:  $|Y|$  is tightly concentrated around  $\theta|T|$ .) We still pick  $Z$  randomly as before. Then we apply Jensen's Inequality on the same calculation to end up with the same upper bound as before. See Lemma 15 in the Appendix.  $\square$

**A. Extending theory to multiple clusters.** Above we assumed a single skill cluster in the language. Real-life text might contain multiple skill clusters. For example, standard corpora must contain a large skill cluster involving pieces of “everyday” text pieces and a set of basic language skills and world knowledge needed to comprehend them. Smaller clusters may correspond to specialized topics, e.g., finance, science, mathematical reasoning, etc. We assume each piece of text appears in only one cluster but skills may appear in different clusters. When each text-piece appears in a single cluster, the analysis of Section ?? continues to apply. The overall loss is the weighted sum of measure of text in the individual clusters. Thus overall reduction in loss will drive emergence within individual clusters. But lacking any mechanistic insight, our theory cannot predict the rate at which loss decrease (and hence emergence) happens within clusters. This pertains to the point made earlier in the paper about lack of detailed study of scaling laws for different kinds of corpora, as well as for training on mixes of corpora.

We leave a more fine-grained analysis, including possibly allowing hierarchical structure in clusters, for future work. As usual, simpler settings probably give the main insight.

## 6. Additional takeaways and experimental evidence

It may be useful to note the following takeaways about skill emergence as per our theory. We also mention some papers that have appeared since the first arxiv appearance of our manuscript that verify some of the predictions.

**1. Surprising emergence of compositional generalization on tuples of skills:** Theorem 13 and Corollary 12 implies that the effect of reducing  $\theta$  by a factor 2 (which as per scaling laws corresponds to roughly one order of scaling up in model parameters) has the effect of raising competence on  $2k'$ -tuples to at least the same level as what it was on  $k'$ -tuples before scaling.

Since the appearance of first version of this paper, the prediction has been experimentally verified by evaluating many models of different sizes (24)

**2. Number of skills per text pieces as a surrogate for “high quality”:** Theorem 7 and Lemma 9 shows that for a fixed prediction loss  $\theta$ , using higher  $k$  implies better emergence of skills. Since  $k$  is the number of skills being used in a single text-piece, it intuitively measures how *complex* the text is —e.g., a college text would be expected to have higher  $k$  than a primary school text. If the scaling law is same for both types of text (i.e., how  $\theta$  reduces from scaling) our theorem predicts that more complex text will be more effective at inducing skills. This prediction generally matches experts' intuition, although we are not aware of a study of scaling laws that tries to separate out texts of different quality.

A recent paper (25) shows that compositionality can be learned from examples with supervised fine-tuning. Small models can learn to combine as many (or more) skills while solving a task as a much larger model. Using data with higher number of skills per paragraph (the  $k$  in our assumption) does indeed induce compositionality much faster. Thus targeted training can be much more effective than pretraining.

**3. More frequent skills tend to reach competence level quicker than less frequent skills:** This effect is hidden in the proof of Theorem 13. Specifically, the proof reduces the case of skills appearing with different frequencies in the corpus to the uniform case by replacing a skill node with a set of nodes whose cardinality scales in proportion to the skill frequency. But note that by definition, the competence on all copies of the same skill must be the same. Thus essentially the calculation says that  $k'$ -tuples that include more frequent skills will tend to emerge faster.

**4. Compositional generalization can be learned despite paucity of stimulus.**

Competence on  $k'$ -tuple of skills (as discussed in item 1. above) for large  $k'$  leads to a paucity of stimulus situation as discussed in the main body of the paper. The SFT experiments of (25) show this phenomenon is real, in context of supervised fine-tuning. When trained on samples displaying 3 skills, i.e., with  $k = 3$ , even small models acquire capability at combining 4 skills a good fraction of time, and occasionally even 5.

**5. Use of LLMs to elicit lists of task-specific skills**

Our framework assumes existence of a large set of basic skills and that models use subsets of such skills while solving



tasks. While this is intuitive, we lack catalogs of such skills for most interesting tasks.

Cognitive scientists assemble such skill catalogs for humans but one imagines LLMs could factor their capabilities into skills differently than humans.

Recently Didolkar et al (26) show that, given an LLM dataset (e.g., MATH or GSM8K), one can use a powerful LLM to recover a list of named skills needed to solve it. Another paper (27) adapts this methodology to recover a list of named skills needed for chat capability, using just conversation with GPT4. In other words, it does not need an underlying dataset from which to extract skills.

## 7. Related Work

The stochastic parrots term comes from (9) albeit the main concern in that paper was the potential harms of LLMs. In human pedagogy there is a long history of viewing learning as a process of skill acquisition. A recent "Skill-It" framework (?) tries to operationalize this for LLMs. ANY OTHERS?

## 8. Conclusions

We have proposed a theoretical framework for understanding emergence of skills when language models are scaled up. A key insight (see Figure 2) is that reduction in excess cross entropy loss drives skill acquisition, together with the assumption that normal language —down to short paragraph level—already utilizes multiple skills, mixed up randomly. Need for mechanistic insight is sidestepped using Scaling Law, which quantifies a powerful inductive bias in pre-trained models. One concrete example of this inductive bias is that in our framework proficiency in combinations of skills arises just as naturally as proficiency in the individual skills themselves, and need not require seeing examples of all (or even most) of these combinations in the training set. This has relevance to the ongoing debate about the extent of “understanding” that current models have, and their ability to address novel settings.

We hope the simplicity of our framework will also encourage further experimental and theoretical study, including extensions to more general language skills such as generation and dialog; and modeling inductive bias at a finer level than the Scaling Laws. (For example, what are the scaling laws for interesting parts of language such as math or coding?) It is also possible that our theory underestimates the rate of emergence, due to unknown mechanisms —e.g., having to do with workings of transformers—whereas our framework treats them as blackboxes.

The simple and statistical nature of our theory should be seen as a plus — it helps identify which emergence phenomena should not be considered surprising, most notably emergence of competence on skills as well as on their combinations. But it shares limitations with other statistical frameworks. Competence is guaranteed only on text-pieces drawn from the data distribution, and governed by usual  $\epsilon$ - $\delta$  guarantees — many skills as well as combinations of skills may not get learnt, and the ones that do get learnt may incorrectly applied on some fraction of the data distribution. Nevertheless we hope this inspires more thorough experimental study (our simple experiments give a starting point) of whether or not current language models have capabilities that go beyond simple statistical explanations. Empirical properties or phenomena that are not derivable in our framework (or its natural extensions)

may be of interest for AI alignment as well as better design and understanding of language models.

**Acknowledgements:** We are very grateful to Jonah Brown-Cohen for many discussions that motivated us to improve the theory and its expositions. We thank Boaz Barak, Rong Ge, Yuxi Liu, and Nikunj Saunshi for their feedback on the manuscript.

1. Brown T, et al. (2020) Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.
2. Ganguli D, et al. (2022) Predictability and surprise in large generative models in *2022 ACM Conference on Fairness, Accountability, and Transparency*. pp. 1747–1764.
3. Srivastava A, et al. (2022) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
4. Wei J, et al. (2022) Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
5. Schaeffer R, Miranda B, Koyejo S (2023) Are emergent abilities of large language models a mirage? *Neurips*.
6. OpenAI (2023) GPT-4 technical report. *CoRR* abs/2303.08774.
7. Chowdhery A, et al. (2022) Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
8. Anil R, et al. (2023) Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
9. Bender E, Gebru T, McMillan-Major A, Shmitchel S (2021) On the dangers of stochastic parrots: Can language models be too big? *FACCT*.
10. Chomsky N (1957) *Syntactic Structures*. (Mouton Co.).
11. Bengio Y, Ducharme R, Vincent P (2000) A neural probabilistic language model. *Advances in neural information processing systems* 13.
12. Cortes C, Jackel LD, Solla S, Vapnik V, Denker J (1993) Learning curves: Asymptotic values and rate of convergence. *Advances in neural information processing systems* 6.
13. Hestness J, et al. (2017) Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.
14. Kaplan J, et al. (2020) Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
15. Bahri Y, Dyer E, Kaplan J, Lee J, Sharma U (2021) Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*.
16. Hoffmann J, et al. (2022) Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
17. Muennighoff N, et al. (2023) Scaling data-constrained language models.
18. Schaeffer R, Miranda B, Koyejo S (2023) Are emergent abilities of large language models a mirage? *ArXiv e-prints*.
19. Xia M, et al. (2023) Training trajectories of language models across scales. *ArXiv e-prints*.
20. Winograd T (1971) Procedures as a representation for data in a computer program for understanding natural language, (MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC), Technical report.
21. Levesque H, Davis E, Morgenstern L (2012) The winograd schema challenge in *Thirteenth international conference on the principles of knowledge representation and reasoning*.
22. Saunshi N, Malladi S, Arora S (2021) A mathematical exploration of why language models help solve downstream tasks in *ICLR 2021*.
23. Brown H (2018) *Language Assessment: Principles and Classroom Practice (6th Ed)*. (Pearson Education).
24. Yu D, et al. (2024) Skill-mix: a flexible and expandable family of evaluations for ai models. *ICLR*.
25. et al HZ (2024) Can models learn to compose skills from examples?
26. Didolkar A, et al. (2024) Metacognitive capabilities of llms: An exploration in mathematical problem solving.
27. Kaur S, Park S, Goyal A, Arora S (2024) Instruct-skillmix: A powerful pipeline for llm instruction tuning.
28. Alon N, Spencer J (2016) *The Probabilistic Method (4th Ed)*. (Wiley).

## A. Appendix

### A. Example of current chatbots' ability to combine skills.

**Human question:** Please give me a couple lines of text that illustrate all of the following language understanding skills: Anaphora resolution, simple logical reasoning, simple understanding of physics, and understanding of sentiment.

**Model response:** Sure! Here's a couple of lines that illustrate the mentioned language understanding skills:

"As the sun set, casting a warm glow over the horizon, John watched in awe. He couldn't help but feel a sense of tranquility wash over him, as if all his worries had melted away with the fading daylight."

Explanation:

- **Anaphora resolution:** The pronoun "he" in the second sentence refers back to the antecedent "John" in the first sentence, resolving the anaphoric reference.
- **Simple logical reasoning:** The logical reasoning here is based on the cause and effect relationship between the setting sun and John's feeling of tranquility. The implied logical connection suggests that the calming effect of the sunset is responsible for John's serene state of mind.
- **Simple understanding of physics:** The mention of the sun setting and casting a warm glow over the horizon reflects a basic understanding of physics, as it relates to the natural phenomena of the Earth's rotation and the resulting visual effect.
- **Understanding of sentiment:** The phrase "a sense of tranquility" and the description of John's worries melting away indicate a positive sentiment associated with the experience. The mention of John feeling awe and a sense of tranquility further emphasizes a positive emotional state.

**Human question:** Okay. Can you change the above example to introduce the element of "hip-hop".

**Model response:** Certainly! Here's an updated example that incorporates the element of "hip-hop": "As the sun set, casting a warm glow over the city skyline, the rhythmic beats of hip-hop music filled the air. John stood mesmerized, his body instinctively moving to the infectious grooves. He couldn't help but feel a surge of energy and joy, as if the music had unlocked a vibrant world within him."

Explanation:

- **Anaphora resolution:** The pronoun "he" in the second sentence still refers back to the antecedent "John" in the first sentence, resolving the anaphoric reference.
- **Simple logical reasoning:** The logical connection here lies in the cause and effect relationship between the hip-hop music and John's feeling of energy and joy. The music's rhythmic beats and infectious grooves are responsible for his instinctive movement and the surge of positive emotions he experiences.
- **Simple understanding of physics:** Although the example doesn't directly mention physics,

## B. Technical Theorems about Random Bipartite Graphs

The theory will need some facts about random bipartite graph  $(V_1, V_2, E)$  with  $N_i$  denoting  $|V_i|$ , and  $N_1 \gg N_2$ . When we say it has degree  $k$ , we mean that in  $E$  every vertex in  $N_1$  is connected to  $k$  vertices in  $N_2$ , where those  $k$  vertices are chosen i.i.d. with replacement. Recall that  $V_1$  corresponded to text-pieces and  $V_2$  to skills in the main body of the paper.

The next lemma uses the famous Probabilistic Method (28). In this method, one is trying to show that in a certain probability space, there are no *bad* outcomes. This is done by letting  $W$  be an integer random variable denoting the number of bad outcomes that happened, and showing that the  $E[W] \approx 0$ . Then it follows that  $W = 0$  with probability at least  $1 - E[W]$ . Concretely, in the next Lemma  $W$  will be the number of “bad” set pairs  $(Y, Z)$  of a certain size that violate the lemma.

**Lemma 14.** *For every positive integer  $k$  and  $\theta \in [0, 1]$  there are  $\alpha, \beta > 0$  such that  $\alpha\beta \leq 1$  and the following holds with probability almost 1. For every  $Y \subseteq V_1$  of size  $\theta N_1$ , there are at least  $(1 - \alpha)$  fraction of vertices in  $V_2$  each of which has at most  $\beta\theta D$  edges going to  $Y$ , where  $D = kN_1/N_2$  is the expected degree of a node in  $V_2$ . The parameter values for which this occurs are specified by the condition*

$$H(\theta) + k\theta(H(\beta\alpha) - \beta\alpha \log \frac{1}{\alpha} - (1 - \beta\alpha) \log(\frac{1}{1 - \alpha})) < 0 \quad [12]$$

*Proof.* For  $Y \subseteq V_1, |Y| = \theta N_1$  and  $Z \subseteq V_2, |Z| \leq \alpha N_2$  we say that  $(Y, Z)$  are *bad* if  $Z$  has at least  $\alpha\beta\theta k N_1$  edges to  $Y$ . Let  $W$  denote the number of such  $Z$ 's. The expectation is upper bounded by

$$N_2 N_1 \binom{N_2}{\alpha N_2} \times \binom{N_1}{\theta N_1} \times \binom{k\theta N_1}{\beta\alpha k\theta N_1} \times \alpha^{\beta\alpha\theta k N_1} \times (1 - \alpha)^{(1 - \beta\alpha)\theta k N_1} \quad [13]$$

For (13) to be  $\ll 1$  it suffices for its logarithm to be negative. By Stirling's approximation  $\binom{N}{t} \leq 2^{(H(t) + \epsilon_N)N}$  where  $H(t) = -t \log t - (1 - t) \log(1 - t)$  is the binary entropy function and  $\epsilon_N$  goes to zero rapidly as  $N \rightarrow \infty$ . Applying this to (13) and taking logarithms, and assuming  $N_2 \ll N_1$ , we arrive at the condition (14) for large  $N_1$ .  $\square$

**Note:** Such arguments allow a fair bit of slop. The expectation was exponentially small, and then we took its logarithm and then divided out by a large number,  $N_1$  to reach (12). Thus additional polynomial factors in the expectation—such as  $N_1 N_2$  above—have no effect on asymptotics.

**A. Analysis of “General Measure” case in Section 5.** We give more details of the proof of Theorem 13 in Section 5. Again, we phrase it via a general lemma about bipartite graph  $(V_1, V_2, E)$  where each vertex in  $V_1$  has edges to  $k$  random vertices in  $V_2$ . We use the shorthand  $N_i = |V_i|$ . As noted in proof of Theorem 13 it suffices to consider the case when  $V_2$  has uniform measure and there is a general measure  $\mu(\cdot)$  on vertices of  $V_1$ , namely  $\mu(v_1)$  is nonnegative and  $\sum_{v_1 \in V_1} \mu(v_1) = 1$ . The measure of an edge  $(v_1, v_2)$  is defined as  $\mu(v_1)$ . We assume all  $\mu(v_1)$  are sufficiently small.

The proof will use a discretization of the measure. We conceptually divide  $V_1$  (and hence also the set of edges) into *classes*, where the  $i$ th class  $C_i$  consists of  $v_1$  such that  $\mu(v_1) \in [(1 + \epsilon)^{-i-1}, (1 + \epsilon)^{-i}]$  for  $\epsilon$  an arbitrarily small constant. We assume all  $\mu(v_1)$  are sufficiently small (meaning some large-ish

$i_0$ , class  $i$  is empty for  $i < i_0$ ) and the number of nonempty levels is much smaller than  $N_1$ . Thus each class has reasonable size—say, much larger than  $N_2$ , the number of skills—which allows the asymptotic arguments appearing below to hold within each class. The above assumptions all seem reasonable for the probability measure associated with text pieces, which should be fairly well spread out.

**Lemma 15 (Main).** *For every positive integer  $k$  and  $\theta \in [0, 1]$  and  $\alpha, \beta > 0$  satisfying  $\alpha\beta \leq 1$  and*

$$H(\theta) + k\theta(H(\beta\alpha) - \beta\alpha \log \frac{1}{\alpha} - (1 - \beta\alpha) \log(\frac{1}{1 - \alpha})) < 0 \quad [14]$$

*the following holds with probability almost 1 in the random bipartite graph  $(V_1, V_2, E)$  of degree  $k$ :*

*For every  $Y \subseteq V_1$  of total measure  $\theta$ , there is a set of least  $(1 - \alpha)$  fraction of vertices in  $V_2$  such that for each  $v_2$  in this set,*

$$\sum_{v_1: (v_1, v_2) \in E, v_2 \in Y} \mu(v_1) \leq \beta\theta. \quad [15]$$

*Proof.* Consider  $Y \subseteq V_1$  that has measure  $\theta$ , and  $Z \subseteq V_2$  has size  $\alpha N_2$ . We say  $(Y, Z)$  is *bad* if every  $v_2 \in Z$  fails condition (15). (Consequently, the measure of edges between  $Z$  and  $Y$  is at least  $\alpha\beta\theta$ .) We will argue that in the random graph, the expected number of bad  $(Y, Z)$  is  $\ll 1$ . In other words, for any fixed measure  $\mu$  with high probability the graph contains no bad  $(Y, Z)$ . (As explained in the note following Lemma 14, we can ignore the contribution to the expectation of  $Z$ 's that have size  $< \alpha N_2$ .)

For any fixed  $Y \subseteq V_1$  we denote  $Y \cap C_i$  as  $Y_i$  and let  $y_i = |Y_i|$ . If  $\mu(Y) = \theta$  then the  $y_i$ 's satisfy the following

$$\sum_i y_i (1 + \epsilon)^{-i} \in [\theta, (1 + \epsilon)\theta]. \quad [16]$$

For a fixed  $(Y, Z)$  let  $\beta_i$  be such that the number of edges between  $Y_i, Z$  is  $\alpha\beta_i y_i k$ . Then the probability (over the choice of the random graph) that  $(Y, Z)$  is bad is at most:

$$\prod_i \binom{ky_i}{\alpha\beta_i ky_i} (\alpha)^{-\alpha\beta_i ky_i} (1 - \alpha)^{ky_i(1 - \alpha\beta_i)} \quad [17]$$

Since  $\sum_i y_i = |Y|$  and  $\binom{ky_i}{\alpha\beta_i ky_i} \approx 2^{H(\alpha\beta_i)ky_i}$ , the left hand side is an expression of type

$$\prod_i 2^{ky_i(H(\alpha\beta_i) - \alpha\beta_i \log \frac{\alpha}{1 - \alpha} + \alpha \log \frac{1}{1 - \alpha})}.$$

Using first order optimality wrt  $y_i$ 's, this is maximized when all  $\beta_i$ 's are equal. So for deriving an upper bound it suffices to let all  $\beta_i = \beta$ , which simplifies (17) to

$$\prod_i \binom{ky_i}{\alpha\beta ky_i} (\alpha)^{-\alpha\beta y_i} (1 - \alpha)^{ky_i(1 - \alpha\beta)} \approx 2^{H(\alpha\beta)k|Y|} (\alpha)^{-\alpha\beta|Y|} (1 - \alpha)^{k|Y|(1 - \alpha\beta)} \quad [18]$$

Now we finish the proof using reasoning similar to that in Lemma 14. The number of choices for  $y_1, y_2, \dots$ , is  $\prod_i |C_i|$ , which is at most  $N_1^P$  where  $P$  is the number of classes.

For a fixed sequence of  $y_i$ 's the number of sets  $Y$  consistent with those intersections is

$$\prod_i \binom{|C_i|}{y_i} \approx \prod_i 2^{H(y_i/|C_i|)|C_i|}.$$

Since  $y_i$ 's satisfy Eq. (16) and  $H()$  is a concave function, this number is maximised when  $y_i/|C_i| \in [\theta, \theta(1 + \epsilon)]$ , and hence the number of possible  $Y$ 's is upper bounded by

$$\prod_i 2^{H(\theta)|C_i|} = 2^{H(\theta)N_1}. \quad [19]$$

By the union bound, the probability that there exists a  $Y$  such that  $(Y, Z)$  is bad is at most  $2^{H(\theta)N_1}$  times (18) times (19). Since  $|Y| \approx \theta N_1$  this completes the proof of Lemma 15.  $\square$

DRAFT