Explaining Emergent Capabilities of LLM for Communication Engineering Students

Cheng-Shang Chang (張正尚)

July 31, 2024

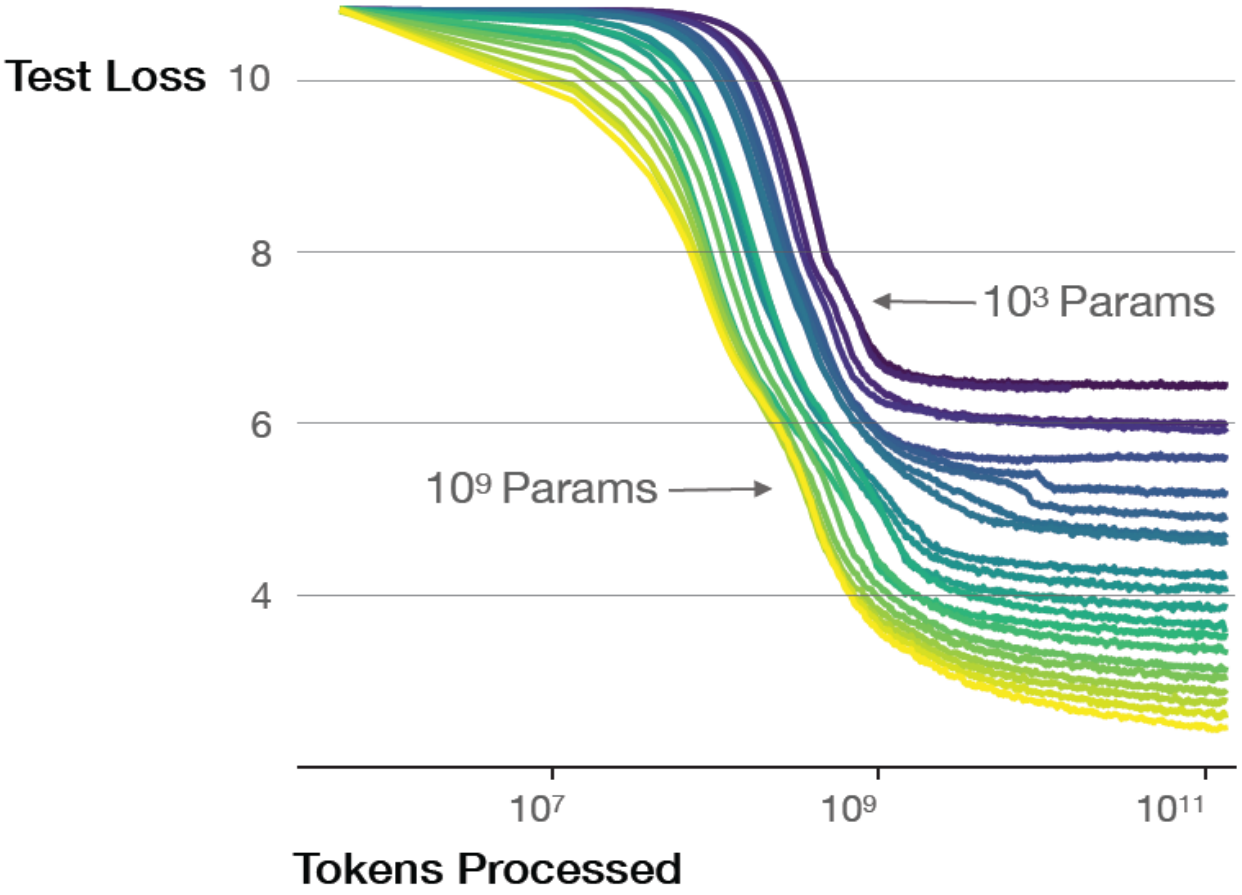
國立清華大學
智慧感知聯網研究中心
Internet of Senses Research Center

Institute of Communications Engineering
National Tsing Hua University

- Liao, Kuo-Yu, Cheng-Shang Chang, and Y-W. Peter Hong, "A Mathematical Theory for Learning Semantic Languages by Abstract Learners," *arXiv preprint arXiv:2404.07009* (2024).

Introduction to LLM for Communication Engineering Students (YouTube video)

Emergent Capabilities of LLM



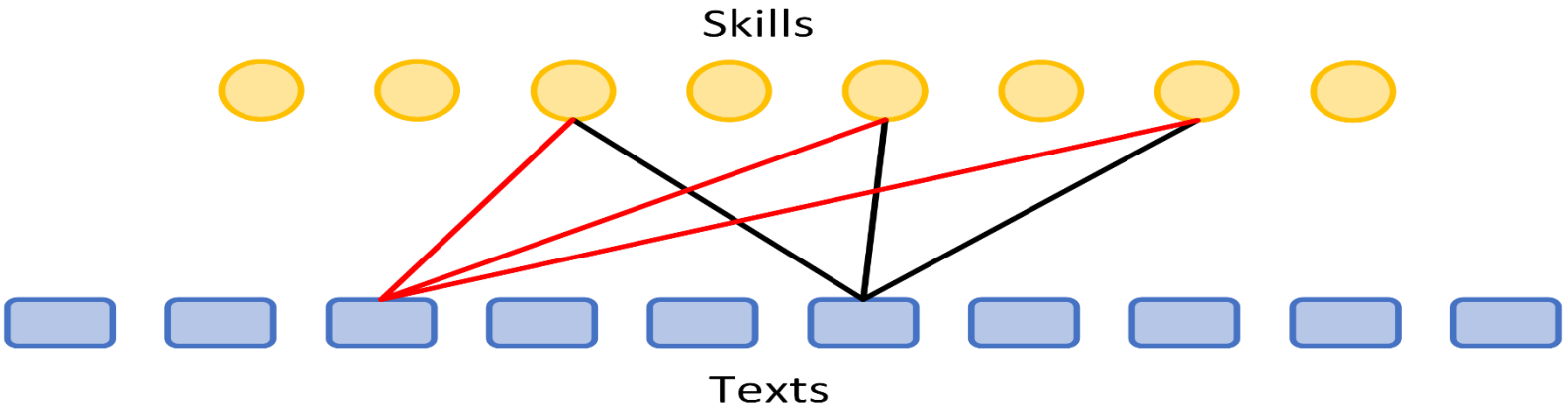
Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Insights

- Learning from Texts:
  - Skills can be acquired by reading texts.
  - A single text might contain multiple skills.
- Probability of Learning:
  - When reading a text, there is a certain probability of learning each skill mentioned.
- Repetition and Mastery:
  - By repeatedly reading a large number of texts, many skills can be learned over time.
- Questions to Consider:
  - How many readings are required to ensure the emergence of learned skills?
  - Are these learned skills interconnected for inference?

Semantic Language as a skill-text bipartite graph

A *semantic language* $\mathcal{L} = (\mathcal{A}, T, S, \phi)$ consists of (i) a set of tokens (symbols) $\mathcal{A}$, (ii) a set of texts T composed of sequences of tokens, (iii) a set of skills S, and (iv) a function ϕ that maps a text t in T to a set of *skills*.

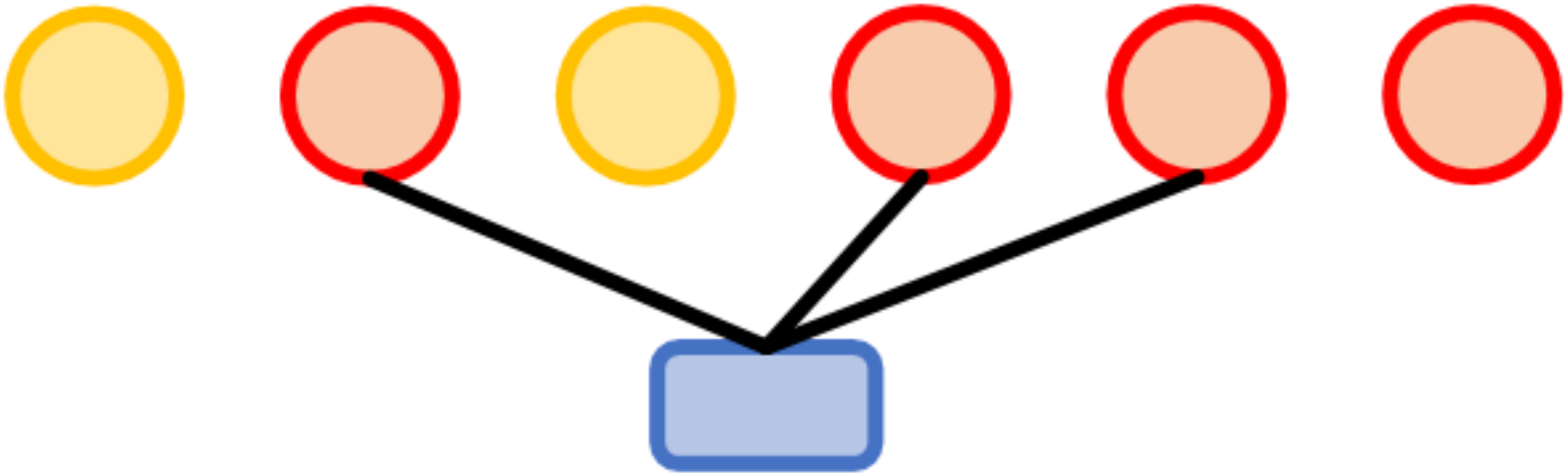


Arora, S., & Goyal, A. (2023). A theory for emergence of complex skills in language models. arXiv preprint arXiv:2307.15936.

Learning a semantic language

- We define that a skill *s* in a semantic language is *learned* if the learner can determine whether the skill *s* is present in any given text *t*.
- We consider a text *t* to be *understood* by a learner if all the skills contained within *t* are learned by the learner.
- Learning a semantic language is equivalent to learning the bipartite graph.

Semantic Language



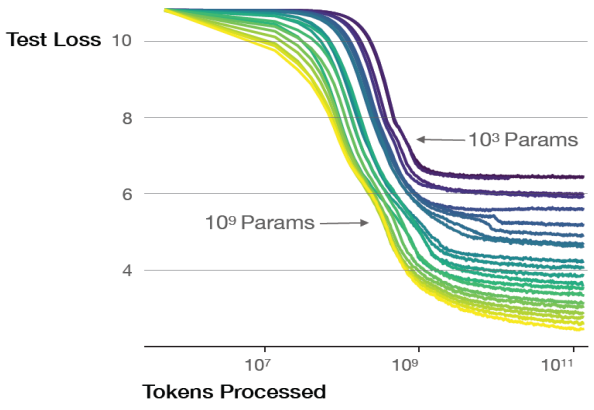
A red skill node is learned, and this text is understood by a learner since the skills contained are learned by the learner.

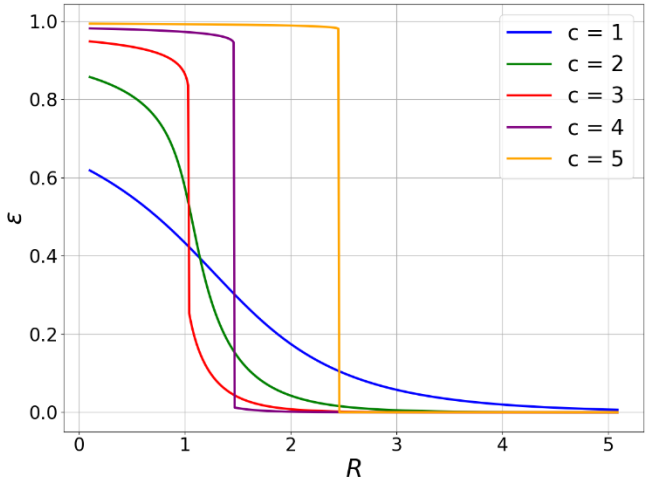
Learning a semantic language by sampling

- Collecting training texts : Sample a subset of texts D from the set of texts T.
- Let $R = |D|/|S|$ be the ratio of the number of training texts to the number of skills.
- Iterative training:
  - When a text is presented to a learner, a skill in that text might be learned with a certain probability.
  - By repeatedly presenting a large number of training texts to a learner, a fraction of skills can be learned.

Emergence of learned skills

- The emergence of learned skills when the ratio R exceeds a certain threshold.
- Once this threshold is exceeded, the testing error, defined as the probability of whether the learner can understand a randomly selected text, drops sharply.
- This also provides the scaling law of the testing error with respect to the ratio R.





Two assumptions about sampled texts:

- (A1) (Poisson degree distribution) Let $N(t) = |\phi(t)|$ be the number of skills in the text $t \in \mathcal{D}$. Then $\{N(t), t \in \mathcal{D}\}$ are independent Poisson random variables with mean c, i.e., $P(N(t) = k) = e^{-c}\frac{c^k}{k!}$ for $k = 0, 1, 2, \ldots$.

- (A2) (Uniform connection) Each edge of a text node is connected to a skill node independently and uniformly. Thus, the probability that an edge of the text node t is connected to a particular skill node s is $1/|\mathcal{S}|$.

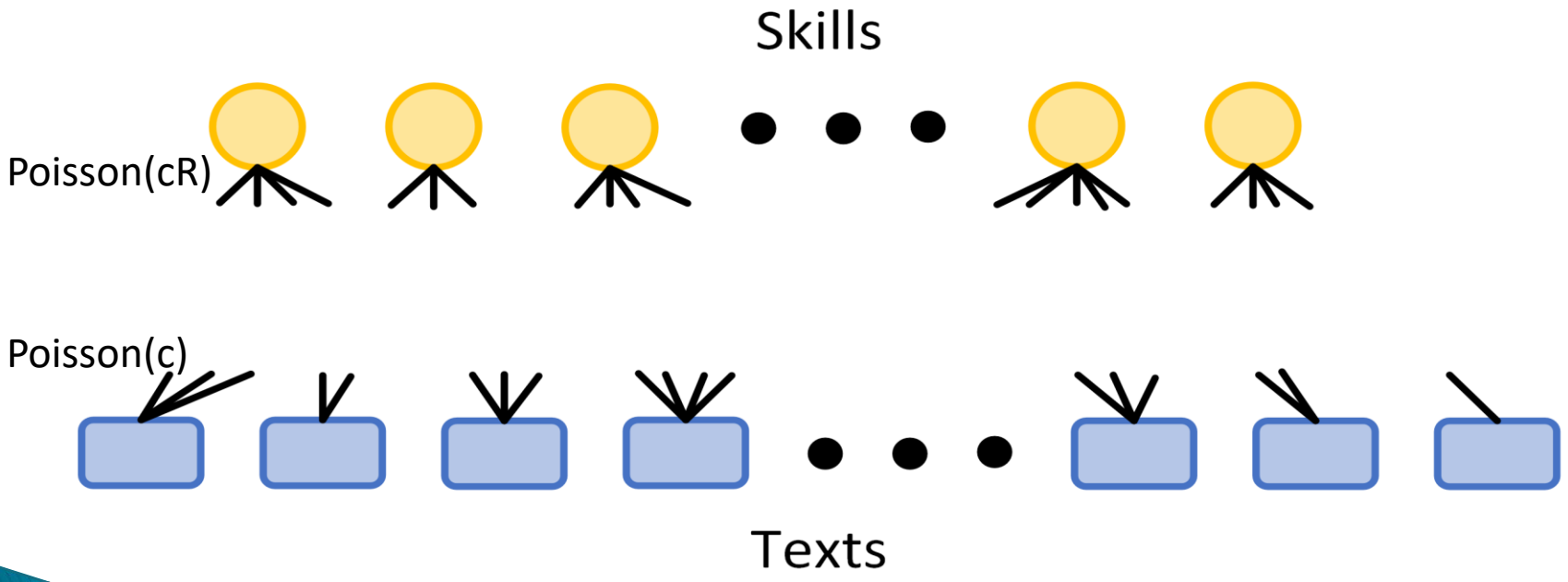
- The sampled skill-text bipartite graph is a random bipartite graph with $|S|$ skill nodes on one side and $|D|$ text nodes on the other side.

Properties of the sampled skill-text bipartite graph

- (i) (Poisson degree distribution) $\{M(s), s \in \mathcal{S}\}$ are independent Poisson random variables with mean $c|\mathcal{D}|/|\mathcal{S}| = cR$, where $M(s)$ is the degree of the skill node s.
- (ii) (Uniform connection) Each edge of a skill node is connected to a text node independently and uniformly. Thus, the probability that an edge of a skill node is connected to a particular text node is $1/|\mathcal{D}|$.
- Such a bipartite graph is a random bipartite graph, as per the configuration model, where the degree distribution of the text nodes (respectively, skill nodes) is Poisson with mean c (respectively, cR).

Configuration model

1. Use the degree distributions to generate “stubs” for skill nodes and text nodes.
2. These “stubs” are randomly connected to form edges.



Abstract learners

- To a learner, all skills are *novel* before training.
- After presenting a text *t* to the learner, a skill *s* in the text t might be *learned*.
- After repeatedly presenting the training set of $|D|$ texts, we aim to determine the fraction of skills that are learned by the learner.

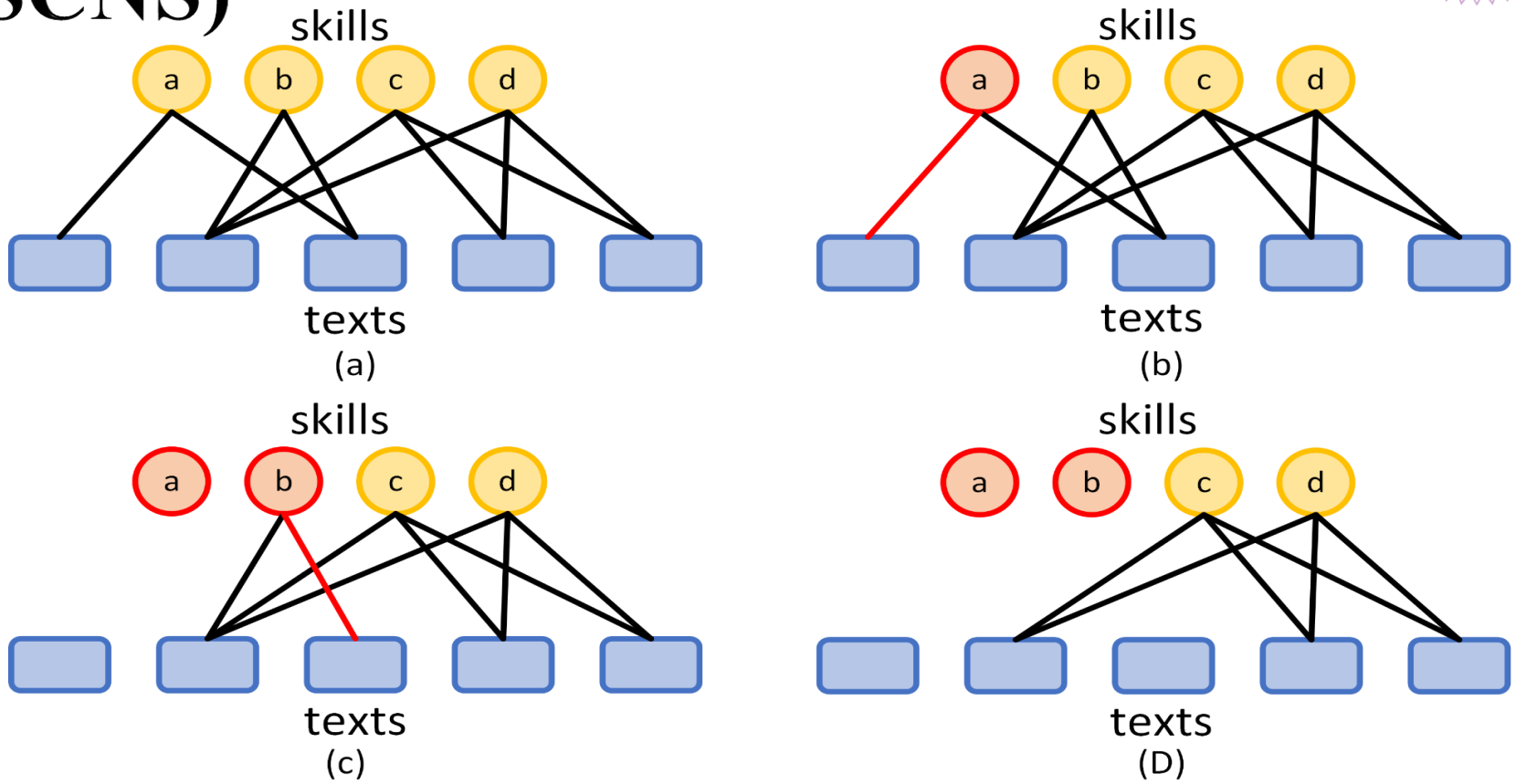
1-skill learners

- An abstract learner is called a 1-skill learner if it can learn a novel skill *s* by being presented with a text *t* where the skill *s* is the only novel skill in the text *t*.
- Once a skill *s* is learned, the learner is able to identify the skill *s* appearing in any other texts.

Successive Cancellation of Novel Skills (SCNS)

- For each iteration, we present the $|D|$ sampled texts to the 1-skill learner in parallel.
- In the first iteration, a text containing only one skill is used to learn the skill in that text. Texts containing more than one skill do not contribute to learning in the first iteration.
- Once a skill s is learned by the 1-skill learner, the number of novel skills in other texts containing s is reduced by 1.
- In other words, the edges connected to the skill node s can be removed from the skill-text bipartite graph.
- In the second iteration, texts with only one novel skill are used to learn the skills in those texts.
- As in the first iteration, skills learned in the second iteration can be used to remove the corresponding edges in the skill-text bipartite graph.
- This iteration process is repeated until no more novel skills can be learned.

Successive Cancellation of Novel Skills (SCNS)



This SCNS training process is mathematically equivalent to the iterative decoding approach in LDPC codes over the binary erasure channel (BEC) and Irregular Repetition Slotted ALOHA (IRSA)

The density evolution analysis

- Let $q^{(i)}$ (respectively, $p^{(i)}$) be the probability that the skill end (respectively, the text end) of a randomly selected edge is not learned after the ith SCNS iteration of training.

$$p^{(i+1)} = 1 - e^{-q^{(i)}c}.$$

$$q^{(i+1)} = \sum_{k=0}^{\infty} e^{-cR} \frac{(cR)^k}{k!} \left(p^{(i+1)}\right)^k = e^{-cR(1-p^{(i+1)})}.$$

$$p = 1 - e^{-ce^{-cR(1-p)}}.$$

The probability of the testing error

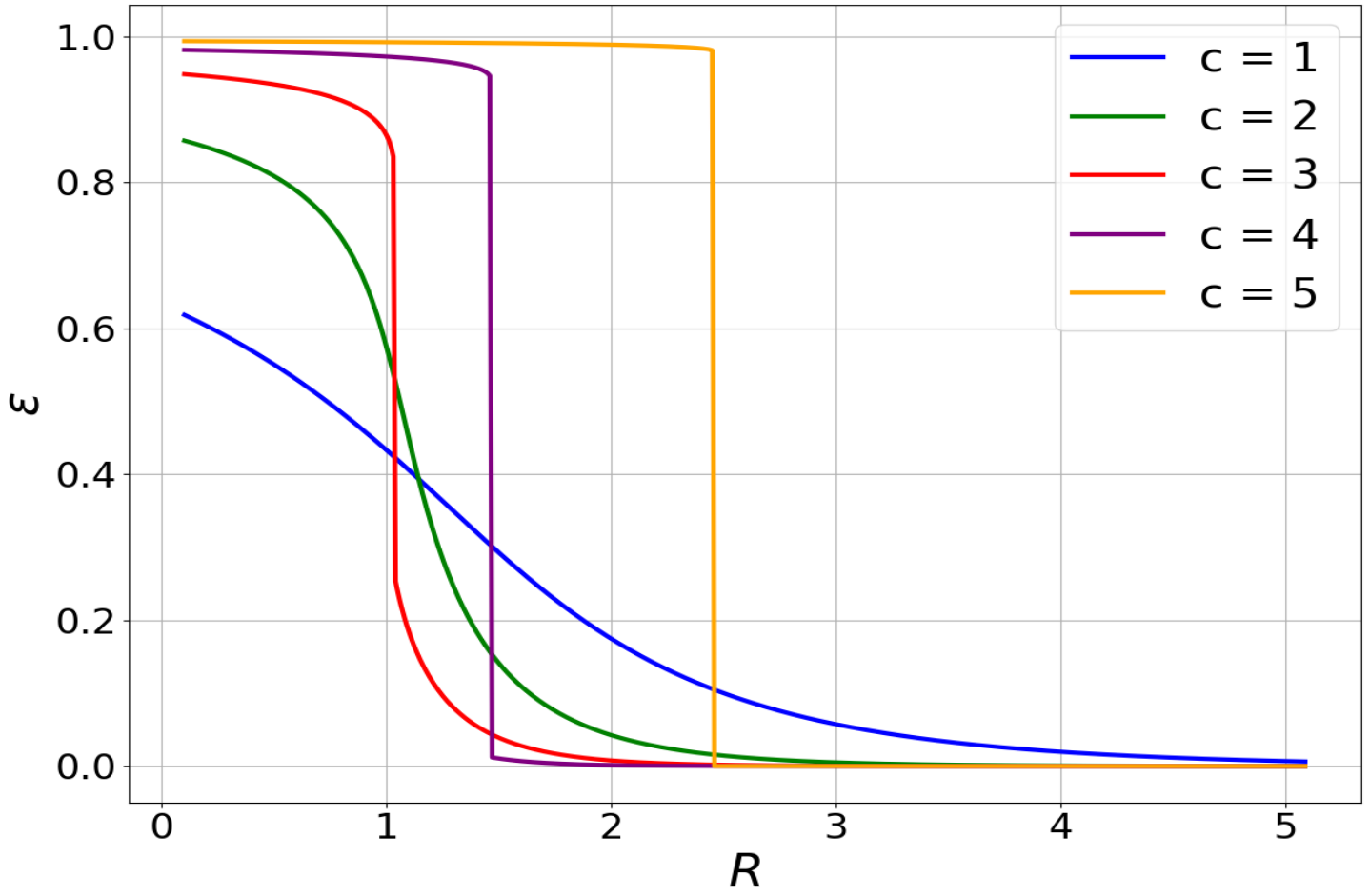
- The probability that a randomly selected skill is learned is

$$\zeta = 1 - \sum_{k=0}^{\infty} e^{-cR} \frac{(cR)^k}{k!} p^k = 1 - e^{-cR(1-p)}.$$

- For testing, we randomly select a text t from T (not included in the $|D|$ training texts).
- Assume that this randomly selected text has the same distribution as those in the training data, i.e., it also satisfies (A1) and (A2). The probability of testing error is

$$\begin{aligned} \epsilon &= 1 - \sum_{k=0}^{\infty} e^{-c} \frac{c^k}{k!} (\zeta)^k \\ &= 1 - e^{-c(1-\zeta)} = 1 - e^{-ce^{-cR(1-p)}}. \end{aligned}$$

Emergence (Threshold)



c is the average number of skills in a text

Poisson learners

A novel skill s in a text t can be learned with probability $P_{\text{suc}}(\rho)$ when the number of *novel* skills in that text is Poisson distributed with mean ρ.

- The definition of Poisson learners is analogous to that of Poisson receivers.
- The 1-skill learner is a Poisson learner with the success probability $P_{\text{suc}}(\rho) = e^{-\rho}$.
- The 2-skill learner, where skills in a text can be learned if the number of novel skills does not exceed two, is a Poisson learner with the success probability

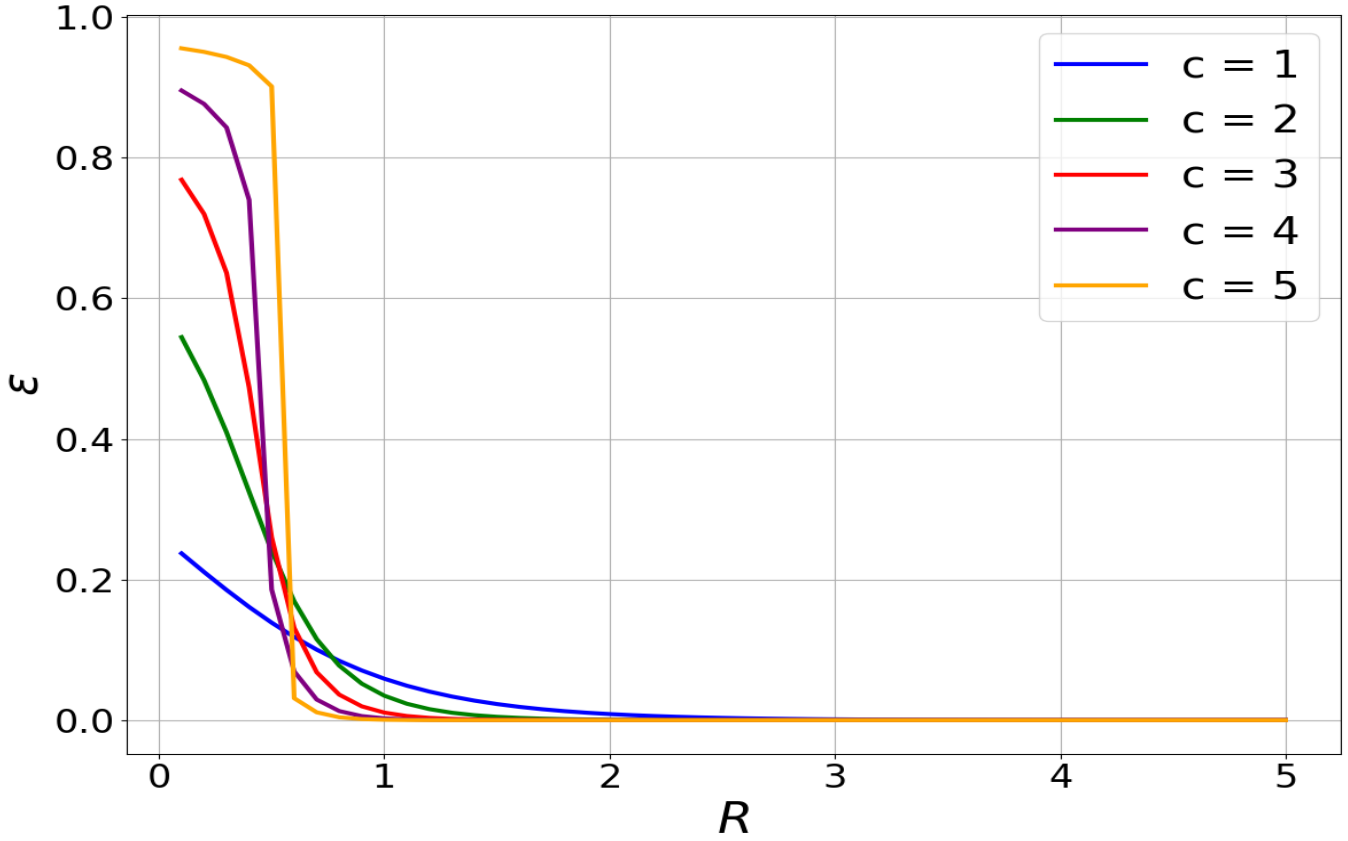
$$P_{\text{suc}}(\rho) = e^{-\rho} + \rho e^{-\rho}$$

C.-H. Yu, L. Huang, C.-S. Chang, and D.-S. Lee, "Poisson receivers: a probabilistic framework for analyzing coded random access," IEEE/ACM Transactions on Networking, vol. 29, no. 2, pp. 862–875, 2021.

Density evolution for Poisson learners

$$p = 1 - P_{\mathrm{suc}}(ce^{-cR(1-p)}).$$

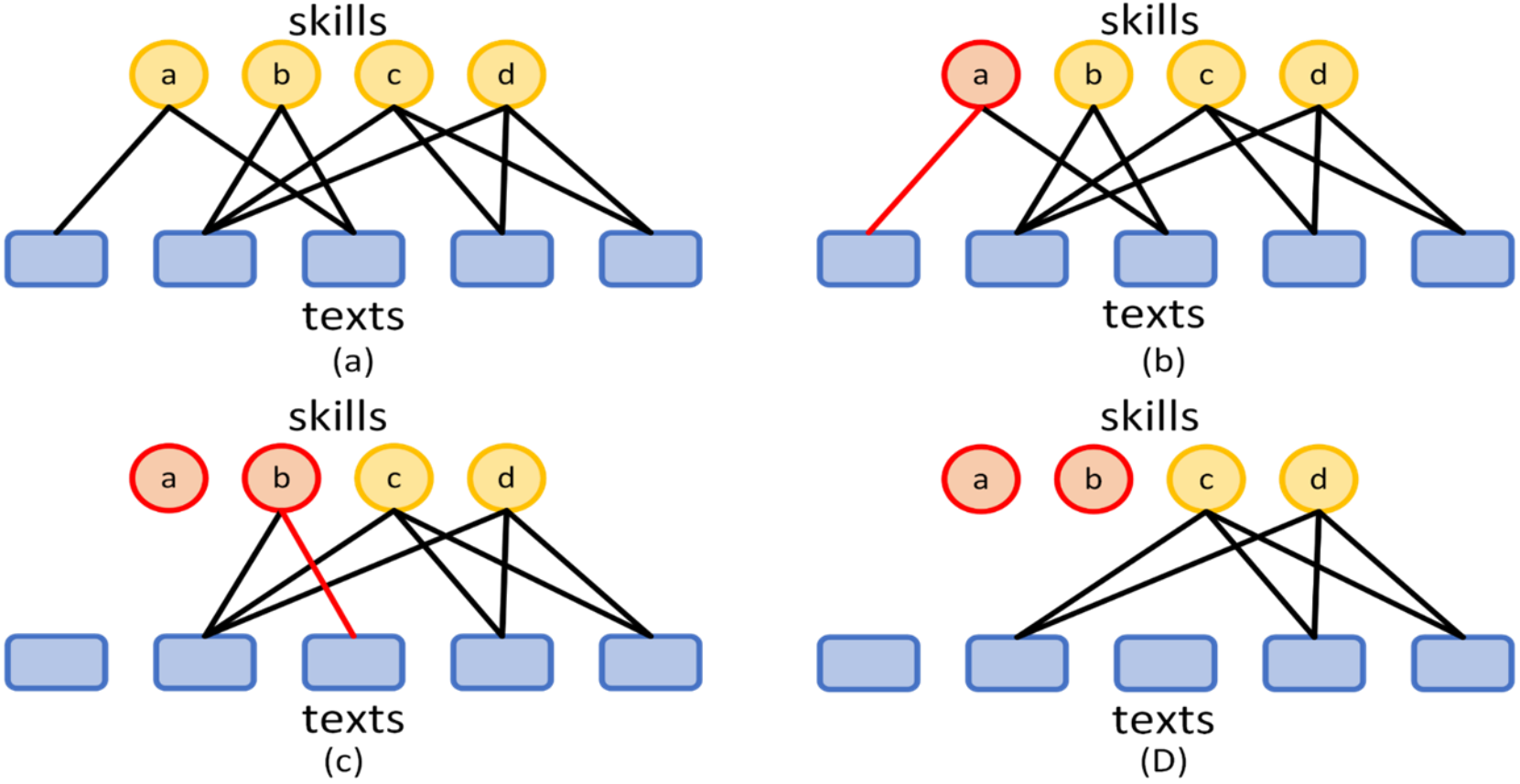
The percolation thresholds for the 2-skill learner are significantly lower than those for the 1-skill learner.



Learning the association of skills

- When the training is complete, a learner can not only learn a fraction of skills but also learn the associations between these learned skills.
- Two learned skills are associated if they appear in the same text.
- Skill association graph: adding an edge between two learned skill nodes if they appear in the same text.
- Knowing the structure of the skill association graph is crucial, as it can be utilized for inference purposes.
- Analogous to giving an LLM a prompt in the form of a text with a set of skills and asking the LLM to predict the next skill and generate a text based on the predicted skill.

Learning the association of skills



Are most of the learned skills interconnected in a way that facilitates their use for prediction?

- A giant component in a random graph is a connected subgraph whose size is proportional to the size of the graph.
- There exists at most one giant component in a random graph.
- The rest of components are called small components.
- One important property of a random graph is that small components are trees with high probability.

Learning the association of skills

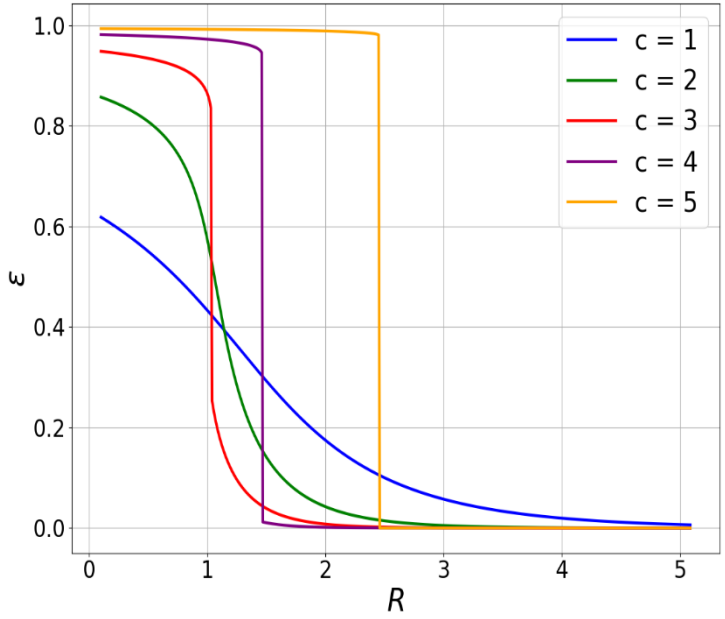
- Assume that skill nodes are learned independently with probability ζ.
- Let μ_s be the probability that a skill node is connected to a small component via one of its edges.
- Let μ_t be the probability that a text node is connected to a small component via one of its edges.

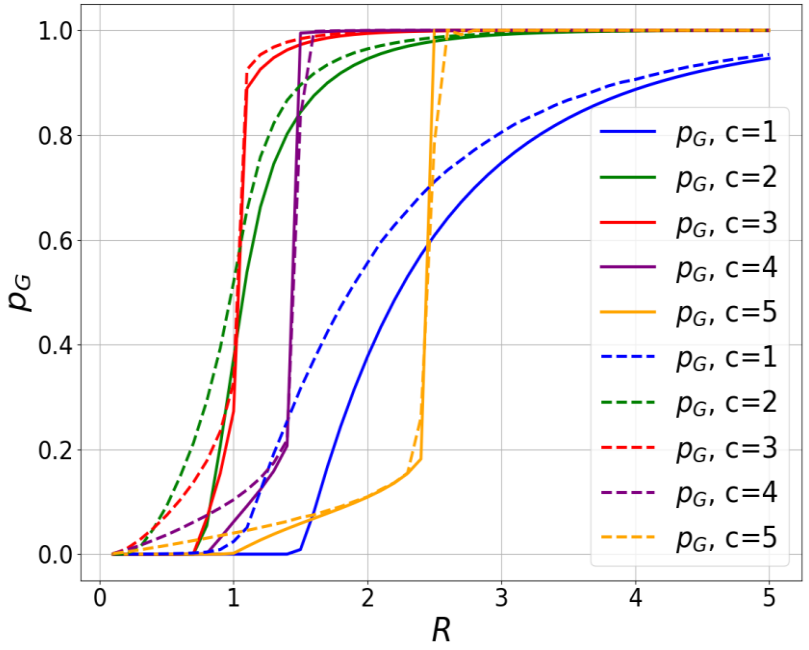
$$\mu_s = \sum_{k=0}^{\infty} e^{-c} \frac{c^k}{k!} (\mu_t)^k = e^{-c(1-\mu_t)}.$$

$$\begin{aligned} \mu_t &= (1-\zeta) + \zeta \sum_{k=0}^{\infty} e^{-cR} \frac{(cR)^k}{k!} (\mu_s)^k \\ &= (1-\zeta) + \zeta e^{-Rc(1-\mu_s)}. \end{aligned}$$

- A randomly selected skill node is in the giant component if the skill node is learned and at least one of its edges is connected to the giant component.

$$p_G = \zeta(1 - \sum_{k=0}^{\infty} e^{-cR} \frac{(cR)^k}{k!} (\mu_s)^k) = \zeta(1 - e^{-Rc(1-\mu_s)}).$$

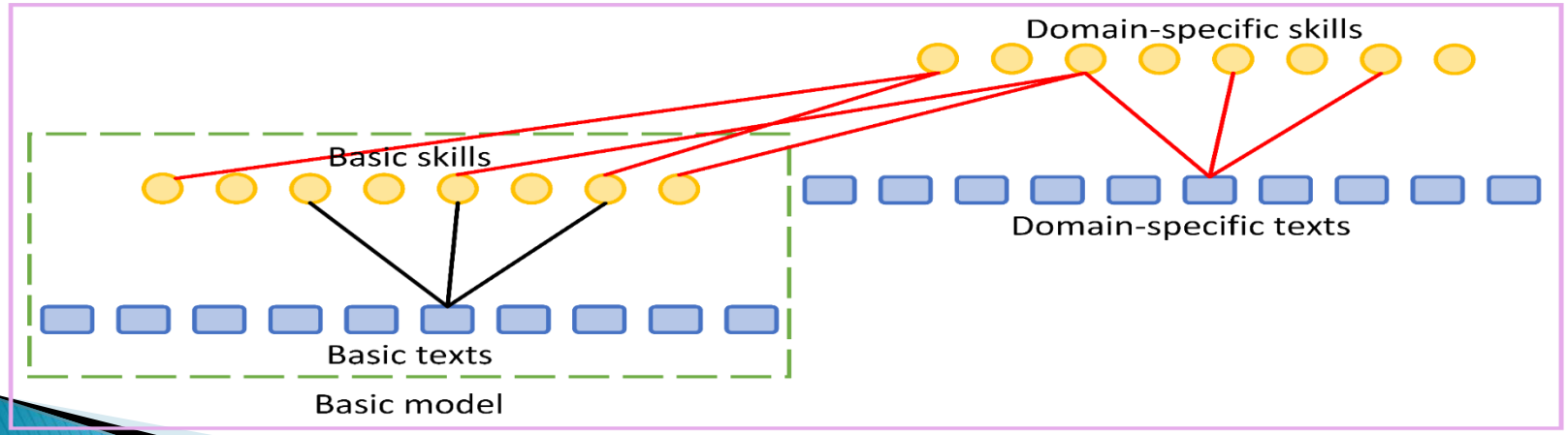




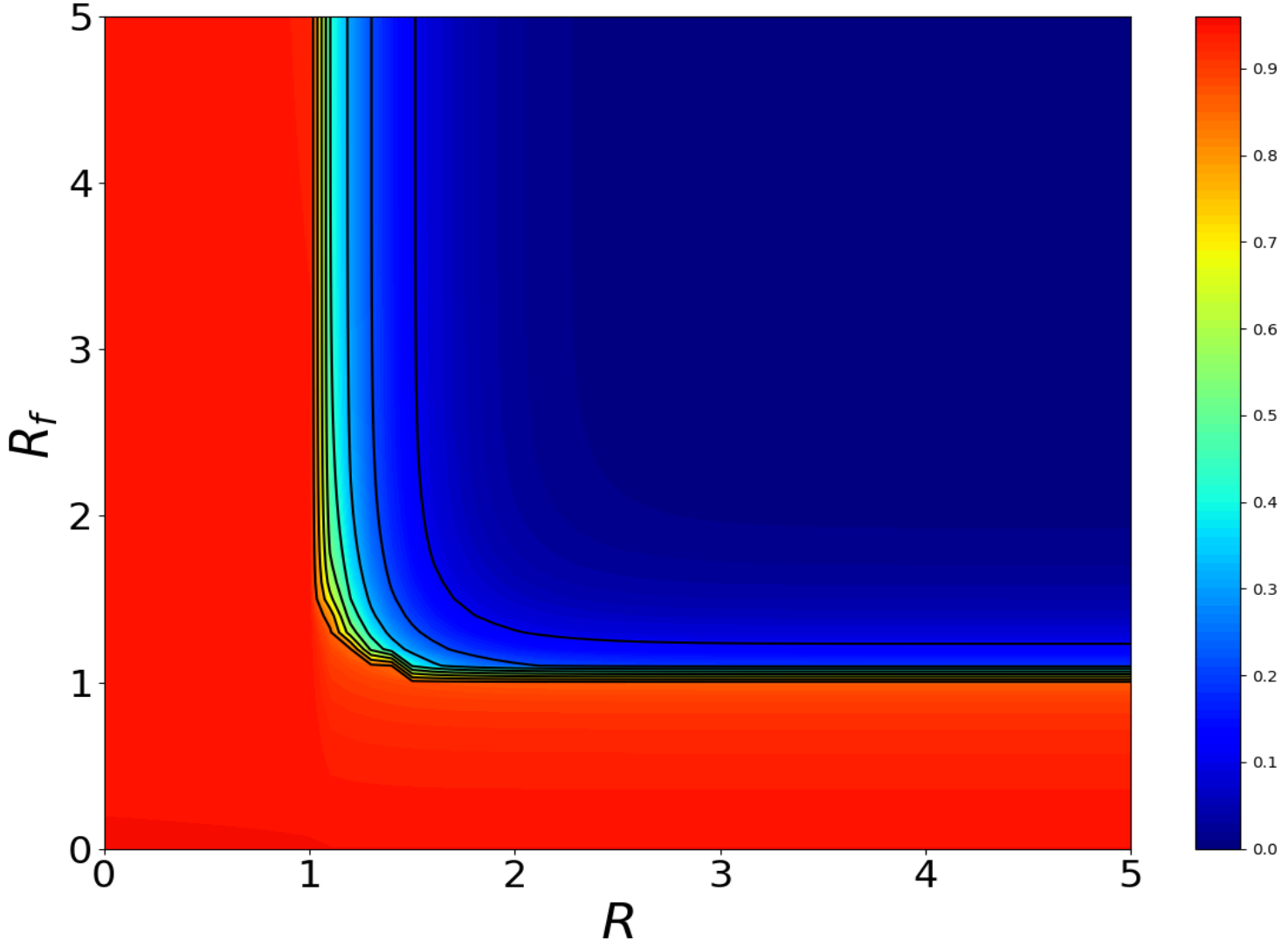
c is the average number of skills in a text

Hierarchy of skills

- One common approach to train a domain-specific LLM is to adopt a pre-trained model, commonly referred to as a foundation model or a basic model, and fine tune it with additional domain-specific texts.
- Two classes of skills: the class of basic skills and the class of domain-specific skills.
- A prerequisite of learning a domain-specific skill requires learning a random number of basic skills first.



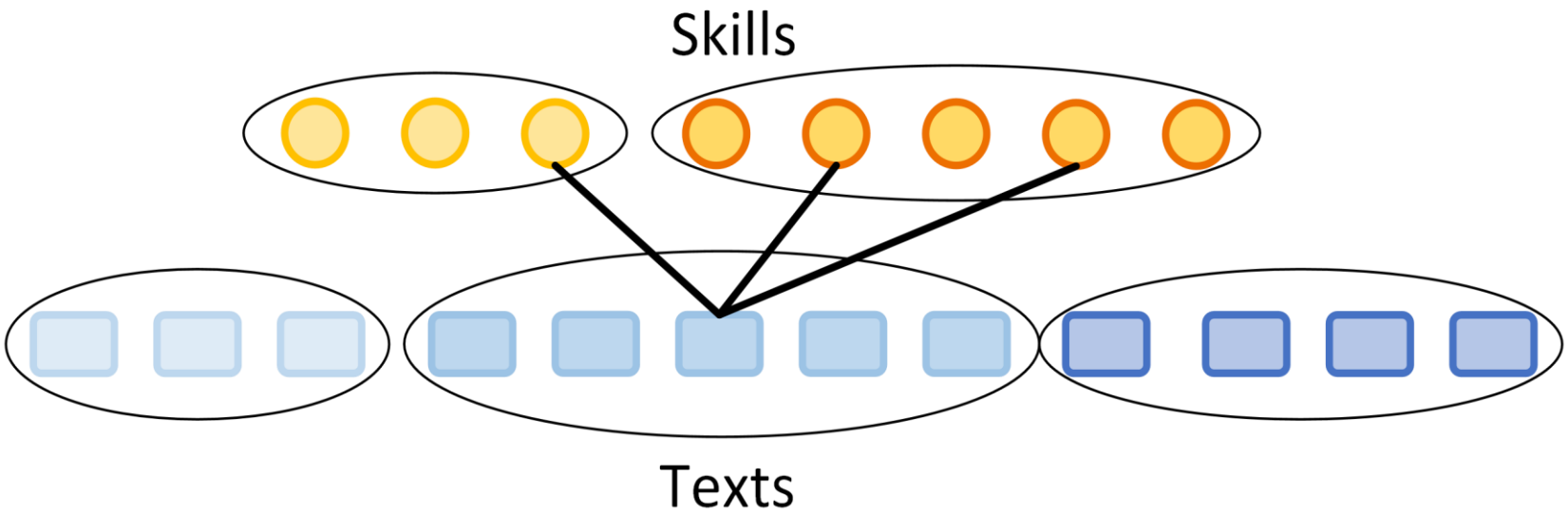
The contour plot of the probability of the testing error for a randomly selected domain-specific skill



- There are two thresholds for achieving low testing errors in a domain-specific task:
- (i) a large number of basic skills are learned when the number of basic texts exceeds the threshold in the foundation model, and
- (ii) a large number of domain-specific skills are learned when the number of domain-specific texts exceeds the threshold in the fine-tuning model.

Multiple classes of skills and texts

- The motivation for the extension to the multiple class setting is the existence of multiple subjects in texts, such as math, physics, chemistry, law, etc.

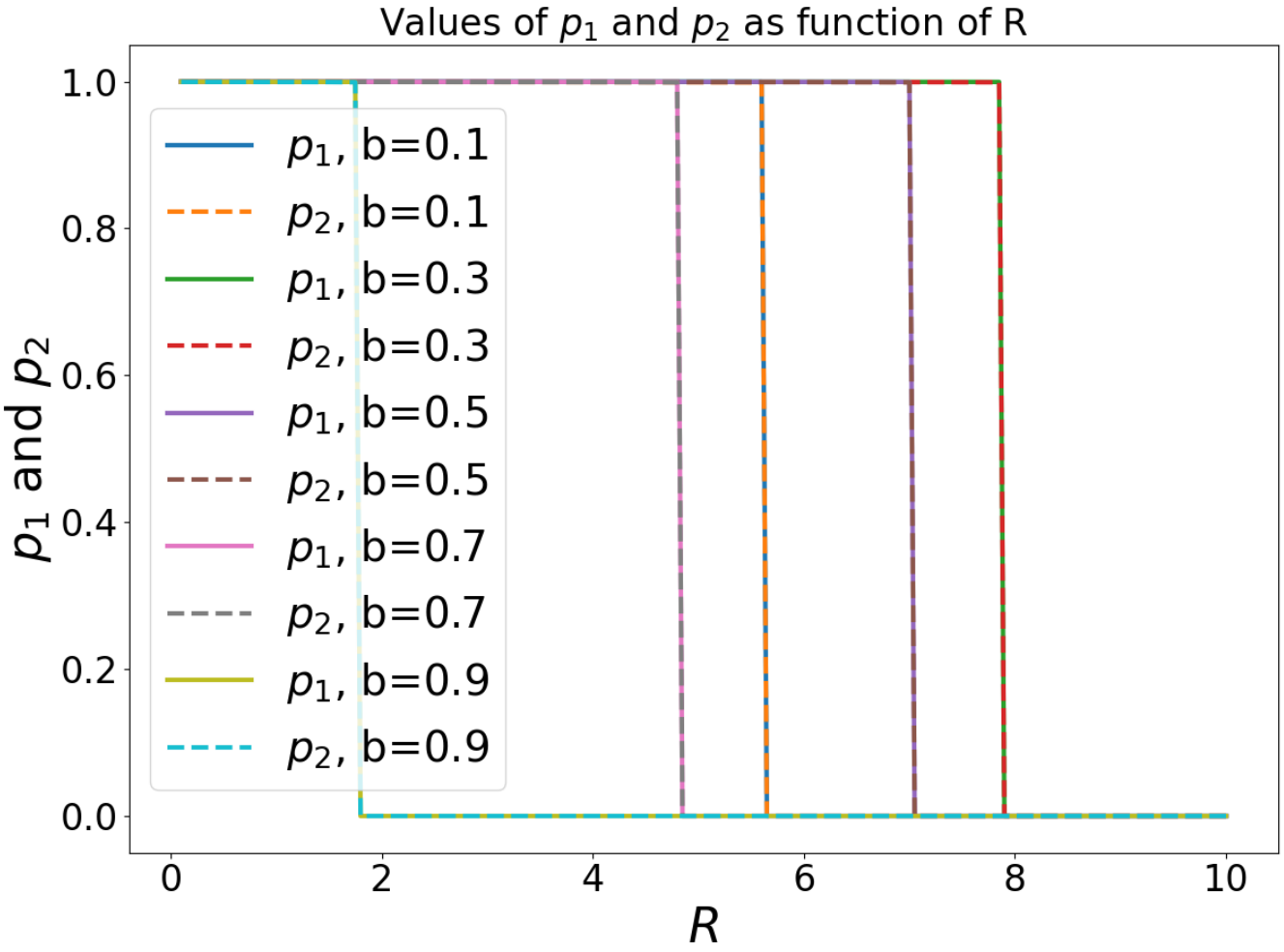


Multiple classes of skills and texts

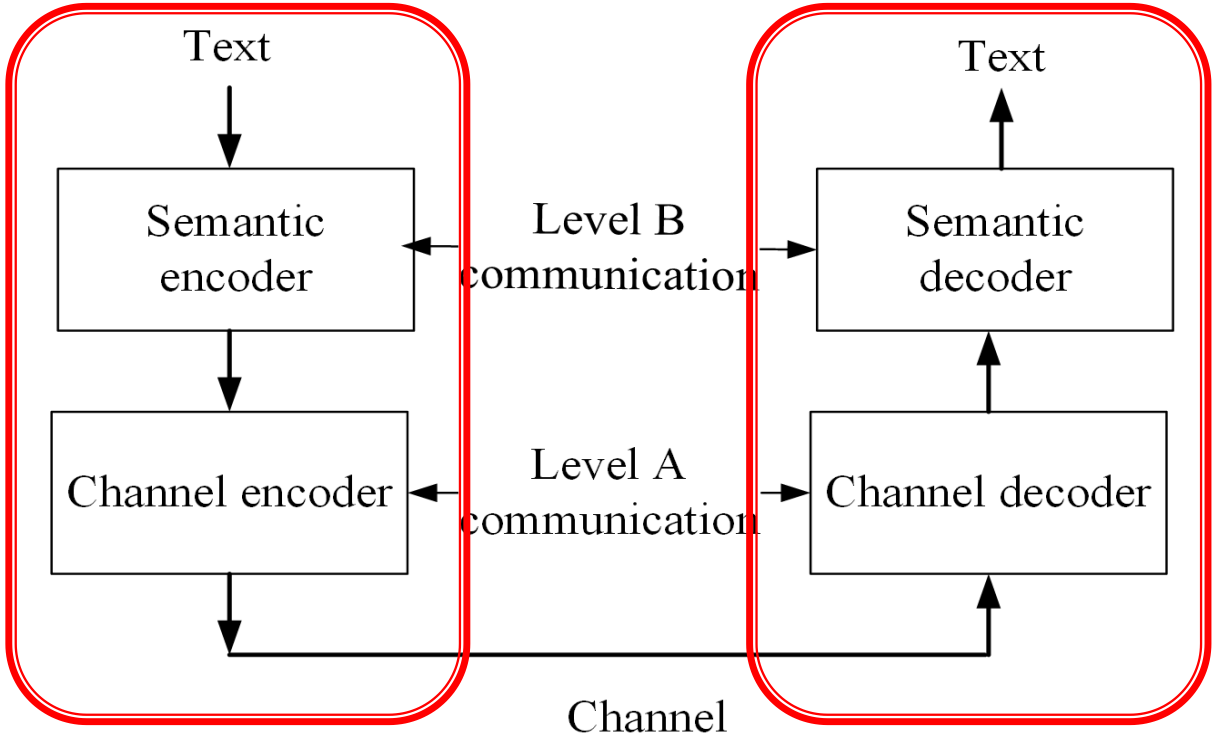
- Introduce Poisson learners with multiple classes of skills by Poisson receivers with multiple classes of users and receivers.
- Repeatedly present the multiple classes of texts to a Poisson learner with multiple classes of skills to learn the skills.
- Extend the density evolution analysis to the multiple classes of skills and texts and derive a system of coupled nonlinear equations.

C.-M. Chang, Y.-J. Lin, C.-S. Chang, and D.-S. Lee, "On the stability regions of coded Poisson receivers with multiple classes of users and receivers," IEEE/ACM Transactions on Networking, vol. 31, no. 1, pp. 234 – 247, 2022.

Simultaneous emergence of multiple classes of skills



Semantic Communication



semantic transmitter semantic receiver

Semantic compression

- (Level A) Lossless compression of a text requires exact recovery of the sequence of tokens in a text.
- (Level B) However, if our interest lies in recovering only the semantic meaning of a text, we might be able to compress it using fewer bits than required for lossless compression.

W. Weaver, “Recent contributions to the mathematical theory of communication,” ETC: a review of general semantics, pp. 261–281, 1953.

Semantic Compression

- A compression method is termed a semantic compression method if the recovered text is semantically equivalent to the original text.

> Two texts t_1 and t_2 are said to be *semantically equivalent* (or simply equivalent) if they both require the same set of skills, i.e., $\phi(t_1) = \phi(t_2)$.

- An abstract learner is called generative if it can generate a text of tokens given a set of learned skills.
- Here we assume that the abstract learners discussed in the talk are also generative.

Semantic Compression

- Once the training is complete, the expected number of skills learned is $|S|(1 - e^{-cR(1-p)})$.
- Index the learned skills.
- The number of bits required to represent a learned skill is $\log_2(|S|(1 - e^{-cR(1-p)}))$.
- On average, there are c skills in a text.
- For a text understood by the learner, it requires on average $c \log_2(|S|(1 - e^{-cR(1-p)}))$ bits to encode the text.

Semantic Compression

- Conversely, if a randomly selected text is not understood by the learner, it can be encoded using a lossless compression encoder.
- Suppose the lossless compression encoder requires, on average, z bits to compress a text.
- Then the semantic compression method described above requires, on average

$$e^{-(ce^{-cR(1-p)})}c\log_2(|S|(1-e^{-cR(1-p)}))+(1-e^{-(ce^{-cR(1-p)})})z$$

bits for a text.

Semantic Compression

- By Shannon's analysis, the entropy per word in the English language is approximately 11.82 bits.
- By assuming that an average sentence has approximately 20 words, the number of bits required to represent a text is approximately $z = 236.4$ bits.
- With R sufficiently large such that

$$c \log_2(|\mathcal{S}|(1 - e^{-cR(1-p)})) \approx c \log_2 |\mathcal{S}|.$$

- With c set as 5, a compression gain is obtained as long as

$$|\mathcal{S}| \leq 2^{47.28}.$$

C. E. Shannon, "Prediction and entropy of printed english," The Bell System Technical Journal, vol. 30, no. 1, pp. 50–64, 1951.

Semantic Compression

- Recent literature on semantic communications has proposed an end-to-end approach for jointly training the semantic and channel encoder/decoder. H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," IEEE Transactions on Signal Processing, vol. 69, pp. 2663–2675, 2021.
- This method is claimed to be superior to separate training.
- However, the end-to-end approach does not scale efficiently with data size.
- A large dataset is necessary for training to exhibit the emergence of semantic capability.
- For the transmission of texts from a general semantic language, an LLM model is required at both the semantic transmitter and receiver, which would be difficult to retrain and adapt to varying physical channels.
- In light of this, a modular design (i.e., separate source and channel coding) may be more effective for semantic communication in practice.

Conclusion

- Learning can be viewed as an iterative decoding process.
- There is a percolation threshold for an abstract learner.
- Once the number of training texts exceeds this threshold, the learner shows the emergence of capabilities (many skills are learned).
- Moreover, these learned skills are associated and form giant components in the skill association graph.
- For multiple classes of coupled texts, there is a simultaneous emergence of multiple classes of skills.
- Learned skills could be used for semantic compression and communication.