清華大學 交叉信息研究院

Institute for Interdisciplinary Information Sciences, Tsinghua University

Understanding the Behaviors of LLMs

A perspective based on Komolgorov Complexity and Shannon Information Theory

Jian Li 李建

Institute of Interdisciplinary Information Science

Tsinghua University

交叉信息研究院 清华大学

Why DL and LLMs Work So Well?

- Tremendous success in practice
- AI models are still big black boxes
- Theory, several exciting recent results (still not so satisfying)

The Man Behind the Google Brain: Andrew Ng and the Quest for the New AI

Researcher Dreams Up Machines That Learn Without Humans

10 BREAKTHROUGH TECHNOLOGIES 2013

Meet Facebook's Head of Artificial Intelligence

FACEBOOK HEAD OF A.I.

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.

Prenatal DNA Sequencing

Reading the DNA c fetuses will be the next frontier of the genomic revolution But do you really w to know about the genetic problems o musical aptitude of your unborn child?

THE GLOBE AND MAII

Google taps U of T professor to teach context to computers 03.11.13

The New York Times

Scientists See Promise in Deep-Learning Programs

John Markoff November 23, 2012

ImageNet 1K competition, fall 2012

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, 2012

t of a waffle and a hippopotamus. This imaginative creature features the distinctive, bulky body of a hippo, but with a texture and have elements like waffle squares across its skin and a syrup-like sheen. It's set in a surreal environment that playfully combines a possibly including oversized utensils or plates in the background. The image should evoke a sense of playful absurdity and culinary fantasy.

Theory of Deep Learning & LLMs

- Theory of DL
  - Optimization (new phenomena)
  - Algorithm-dependent generalization
  - Implicit bias (towards local/global min with interesting properties)
  - Understanding useful tricks: Dropout, batchnorm, layernorm, initialization
- Theory of LLM
  - Why predicting next token yields intelligence
  - Understanding Pretraining, Fine-Tuning and In-Context Learning
  - Understanding Scaling Law
  - Hallucination and Interpretability
  - Knowledge storage
  - CoT, Reasoning

Traditional Optimization and Generalization theories do NOT work any more

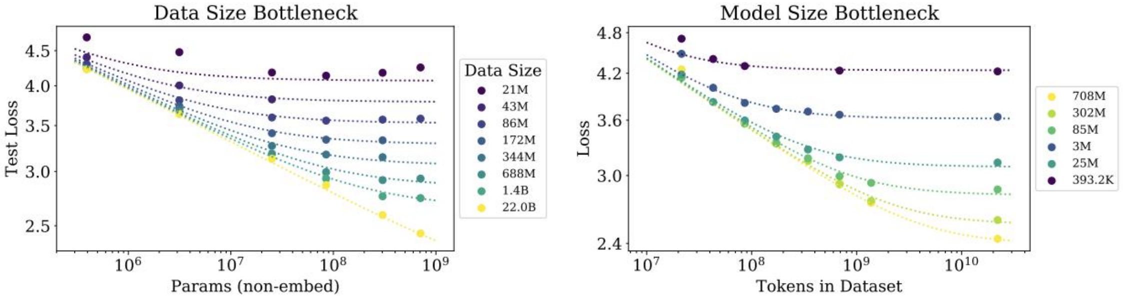
Scaling Laws

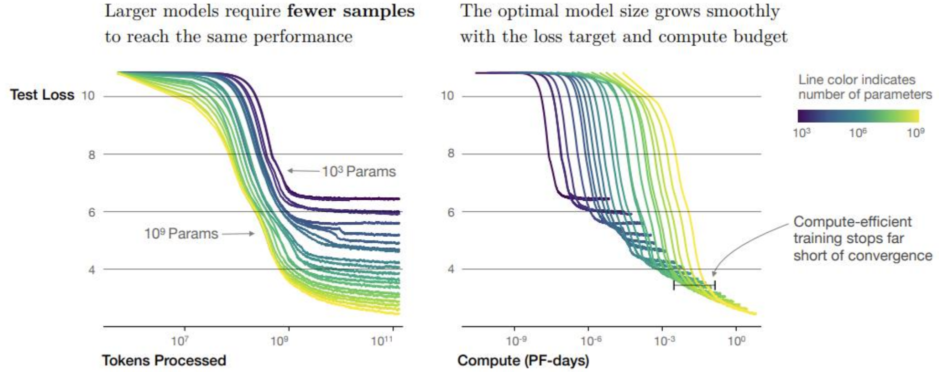
- Kaplan Scaling Law (OpenAI)

$$L(N,D)=\left[\left(\frac{N_c}{N}\right)^{\frac{\alpha_N}{\alpha_D}}+\frac{D_c}{D}\right]^{\alpha_D}$$

$\alpha_N \sim 0.076$, $N_c \sim 8.8 \times 1013$ (non-embedding parameters)

$\alpha_D \sim 0.095$, $D_c \sim 5.4 \times 1013$ (tokens)




Larger models require **fewer samples** to reach the same performance

The optimal model size grows smoothly with the loss target and compute budget

L: the loss
N: model size;
D: dataset size (the token number of training data).

- Chinchilla Scaling Laws (DeepMind)

$$L(N,D)=E+\frac{A}{N^{0.34}}+\frac{B}{D^{0.28}},$$

$E = 1.69$, $A = 406.4$, $B = 410$.

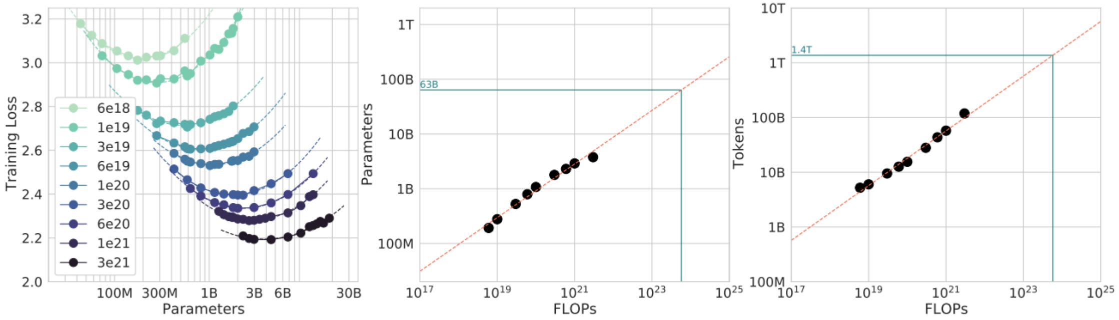
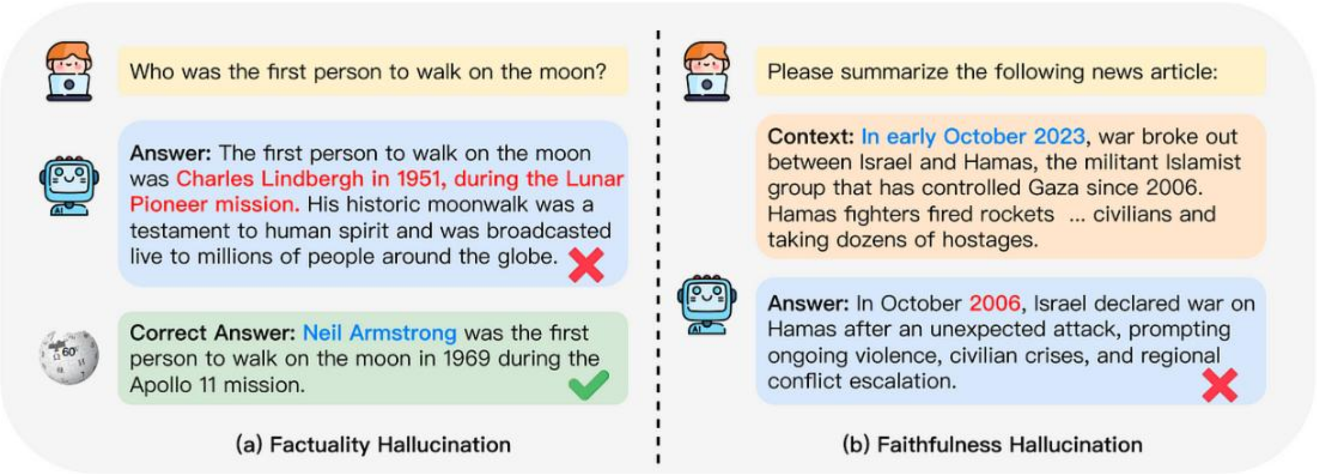


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

In Context Learning and Hallucination

- Hallucination



Who was the first person to walk on the moon?

Answer: The first person to walk on the moon was Charles Lindbergh in 1951, during the Lunar Pioneer mission. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ✗

Correct Answer: Neil Armstrong was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✓

(a) Factuality Hallucination

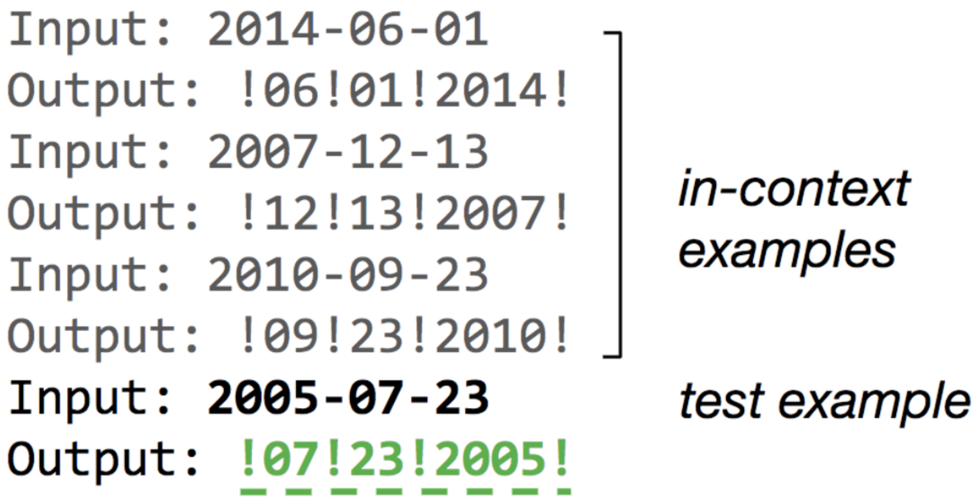
Please summarize the following news article:

Context: In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

Answer: In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ✗

(b) Faithfulness Hallucination

- In Context Learning

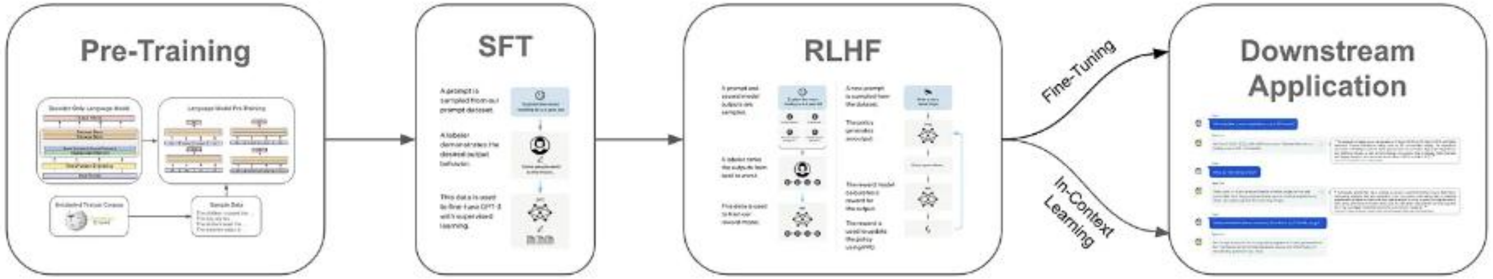


```
Input: 2014-06-01
Output: !06!01!2014!
Input: 2007-12-13
Output: !12!13!2007!
Input: 2010-09-23
Output: !09!23!2010!
Input: 2005-07-23
Output: !07!23!2005!
```

in-context examples

test example

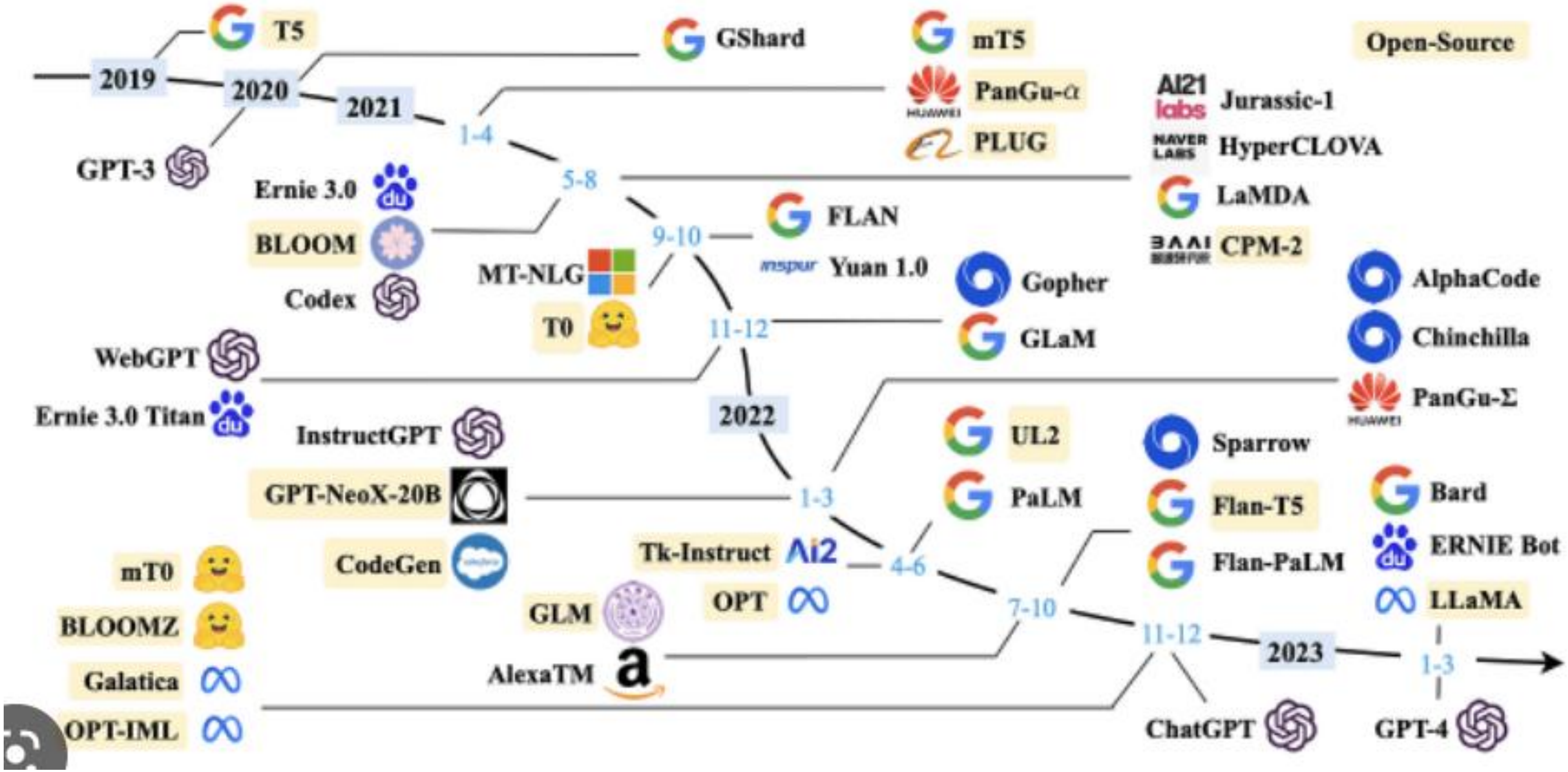
Generalization in the Pretraining-Finetuning framework



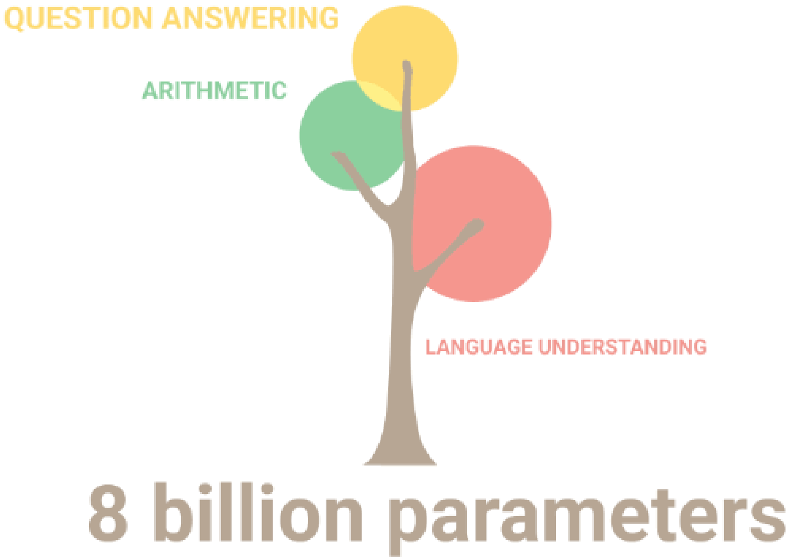
Outline

- LLM Theory
  - **Fundamental Ideas from Shannon and Kolmogorov**
  - Compression and Prediction
  - Kolmogorov’s theory
  - Data Modeling (a nonparametric model)
  - Hallucination and ICL
  - Universal Predictor
  - Research Directions

Pre-trained Foundation Models



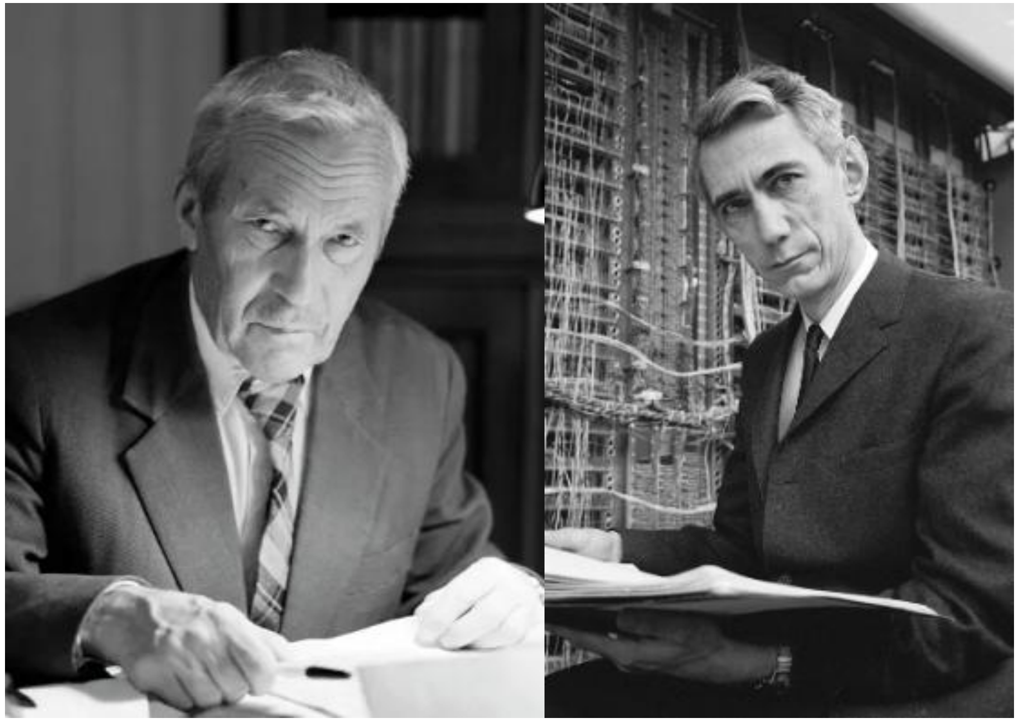
Emergence of Intelligence
智能能力的涌现



Fundamental Ideas from Shannon and Kolmogorov

Shannon, Prediction and Entropy of Printed English (1951)

- He introduced the idea of modeling language as a stochastic process.
- Experiments to estimate language entropy (perplexity).
- “Guessing games”
- Directly related to coding and compression



Kolmogorov Complexity (algorithmic information theory)

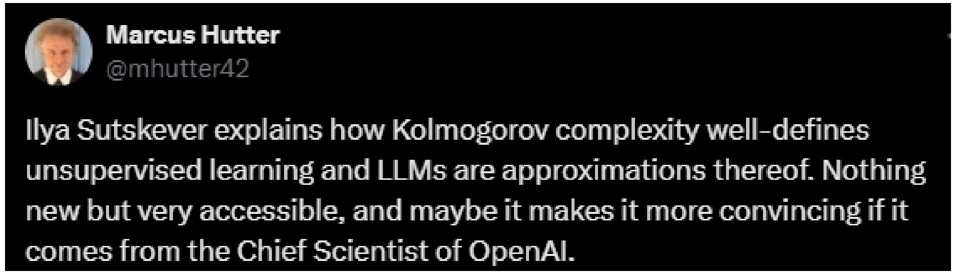
K(X):= The minimum length of TM that outputs X.

(a fine-grained structure: Kolmogorov structure function)

- Direct connection to compression
- Do not need to know the exact distribution (unlike Shannon’s information theory)
- Downside: incomputable…

Compression and Intelligence

"An Observation on Generalization" by Ilya Sutskever

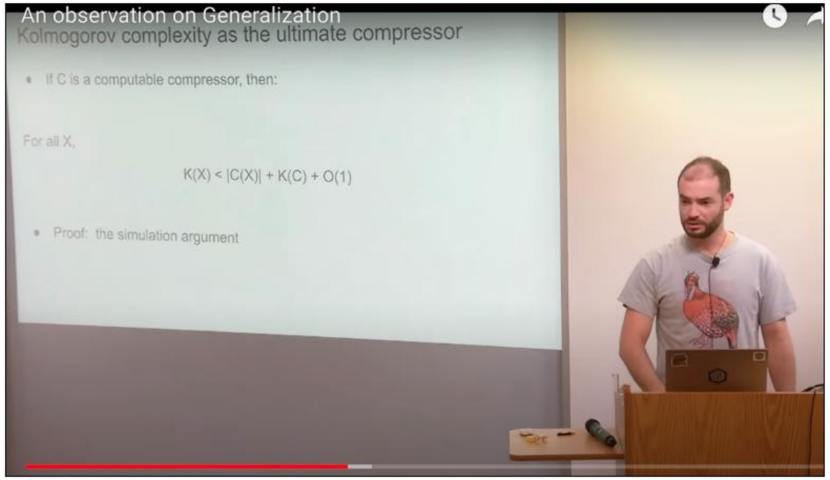


Marcus Hutter
@mhutter42

Ilya Sutskever explains how Kolmogorov complexity well-defines unsupervised learning and LLMs are approximations thereof. Nothing new but very accessible, and maybe it makes it more convincing if it comes from the Chief Scientist of OpenAI.

Very good talk (watch on youtube : https://www.youtube.com/watch?v=AKMuA_TVz3A)

View compression from Kolmogorov complexity theory



An observation on Generalization

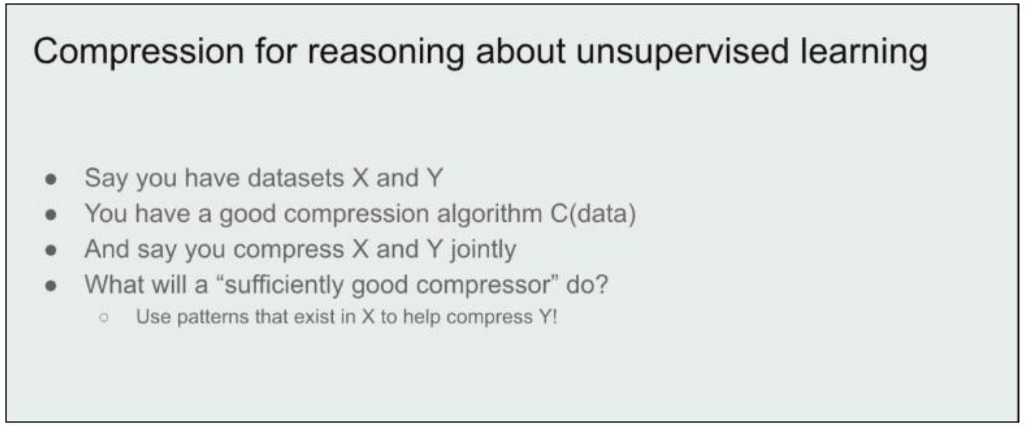
Kolmogorov complexity as the ultimate compressor

- If C is a computable compressor, then:

For all X,

K(X) < |C(X)| + K(C) + O(1)

- Proof: the simulation argument



Compression for reasoning about unsupervised learning

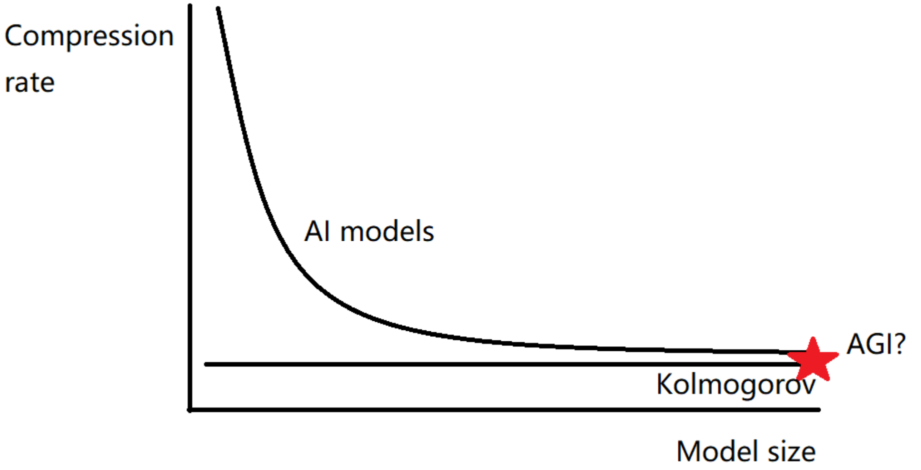
- Say you have datasets X and Y
- You have a good compression algorithm C(data)
- And say you compress X and Y jointly
- What will a "sufficiently good compressor" do?
  - Use patterns that exist in X to help compress Y!

K(X): Kolmogorov complexity of string X

The minimum length of TM that outputs X.

$$K(X) < |C(X)| + K(C) + O(1)$$

Kolmogorov compressor as the ultimate compressor



Why next-token prediction is enough for AGI - Ilya Sutskever (OpenAI Chief Scientist) https://www.youtube.com/watch?v=YEUclZdj_Sc

Outline

- LLM Theory
  - Fundamental Ideas from Shannon and Kolmogorov
  - Compression and Prediction
  - Kolmogorov’s theory
  - Data Modeling (a nonparametric model)
  - Hallucination and ICL
  - Universal Predictor
  - Research Directions

Prediction vs Compression

The training task of a pre-trained large language model (LLM) is "next token prediction," so we can naturally view a pre-trained LLM as a next token predictor.

Given an unknown source distribution P, a predictor is defined as $Q: X^* \to \Delta^N$ which given the prefix $x_{<k}$ as input and outputs the conditional probability distribution $Q(x_k \,|x_{<k})$.

Cross Entropy Loss

In practice, the objective we use to train our LLMs is cross entropy defined as

$$\begin{aligned}\mathcal{L}(M) = -\log P_M(X_{1:n}) = -\frac{1}{n}\sum_{i=1}^{n} \log\, P_M(X_i) = \mathbb{E}_{X\sim\widehat{P}_\phi}\left[-\log P_M(X)\right] \\ = H(\widehat{P}_\phi \| P_M) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{t} \log\, P_M(x_t^{(i)} \mid x_{1:t-1}^{(i)}),\end{aligned}$$

Prediction vs Compression

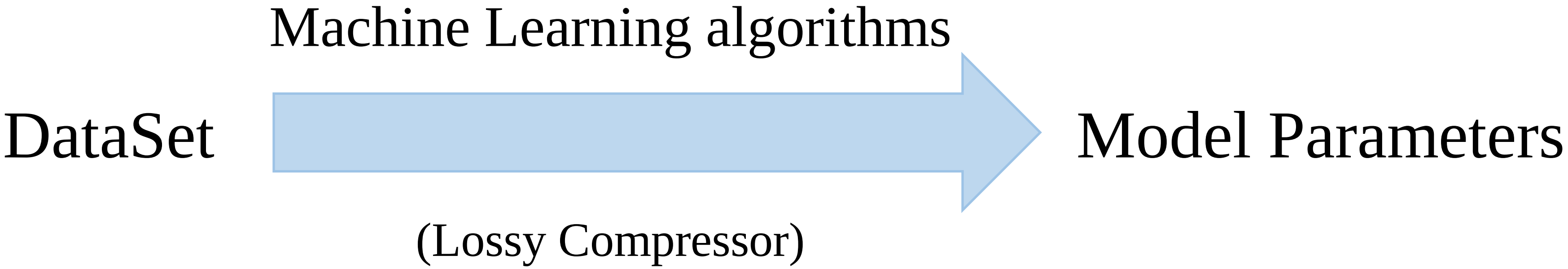
Equivalence of Prediction and Compression

The better we can predict next token, the better we can compress the sequence

We will show that if we can achieve a cross-entropy C (per token), we can essentially compress the text using C+o(1) bits (per token), and vice versa.

Equivalence of Prediction and Compression

Lossy Compression?



Machine Learning algorithms

DataSet → Model Parameters

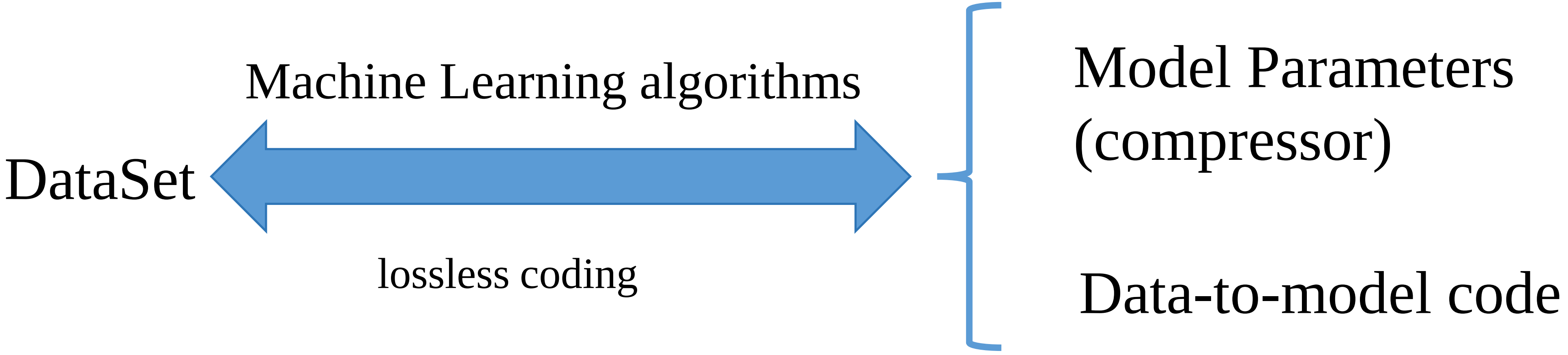
(Lossy Compressor)

Limitations: 1.Too much loss

2.No Guaranteed Generalization

Equivalence of Prediction and Compression

Lossless Compression



e.g. winzip (a compressor)

One can use LLM as a more powerful compressor

Equivalence of Prediction and Compression

Why lossless compression leads to **intelligence**

- If the model can lossless compress well, it should have learned real feature in the dataset and will generalize well.

- Minimal Description Length（MDL）

The best interpretation of a set of data is a description of that data that is accurate and as short as possible.

- Occam's Razor :

Entities should not be multiplied unnecessarily.

Solomonoff's theory of inductive inference(1964):

"If a universe is generated by an algorithm, then observation of that universe, encoded as a dataset, are best predicted by the smallest executable archive of that dataset."

Lossless Compression/coding

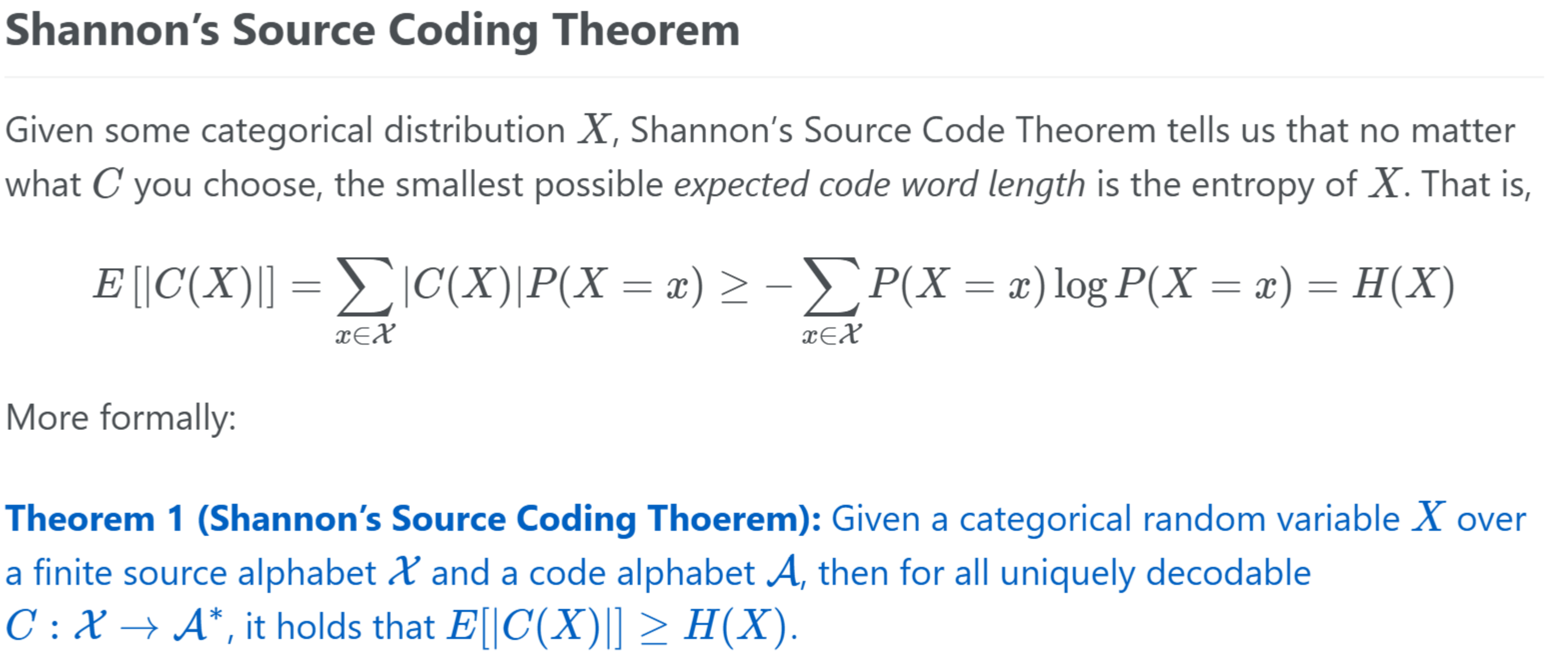
- Compressor encoder c: $X^* \to \{0,1\}^*$
- There exists a decoder d: $R(c) \to X^*$ satisfies that $d(c(x_{1:n})) = x_{1:n}$.

The goal of lossless compression is to minimize the average code length

$$L_c = E_{x\sim\rho}[l_c(x)]$$

where l_c means the bit length of $c(x)$.

Shannon Sourcing Coding Theorem



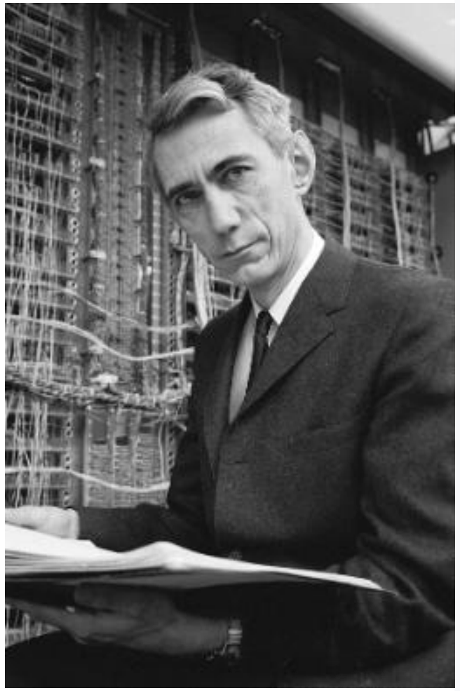
Shannon’s Source Coding Theorem

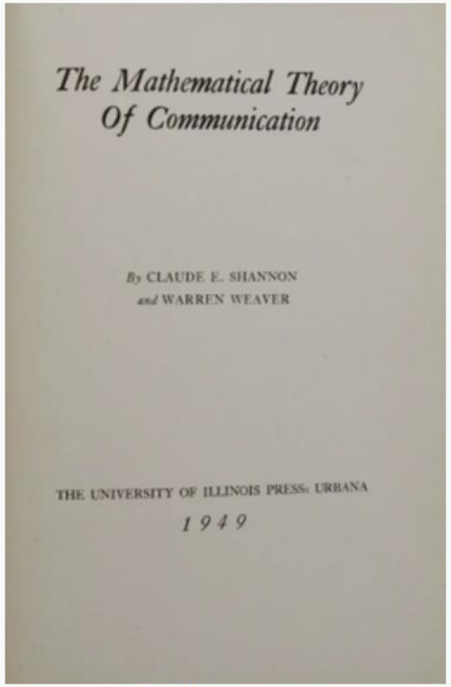
Given some categorical distribution X, Shannon’s Source Code Theorem tells us that no matter what C you choose, the smallest possible *expected code word length* is the entropy of X. That is,

$$E\left[|C(X)|\right] = \sum_{x\in\mathcal{X}} |C(X)|P(X=x) \geq -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x) = H(X)$$

More formally:

Theorem 1 (Shannon’s Source Coding Thoerem): Given a categorical random variable X over a finite source alphabet $\mathcal{X}$ and a code alphabet $\mathcal{A}$, then for all uniquely decodable $C : \mathcal{X} \to \mathcal{A}^*$, it holds that $E[|C(X)|] \geq H(X)$.





The Mathematical Theory
Of Communication

By CLAUDE E. SHANNON
and WARREN WEAVER

THE UNIVERSITY OF ILLINOIS PRESS: URBANA
1949

1949 full book edition

https://mbernste.github.io/posts/sourcecoding/

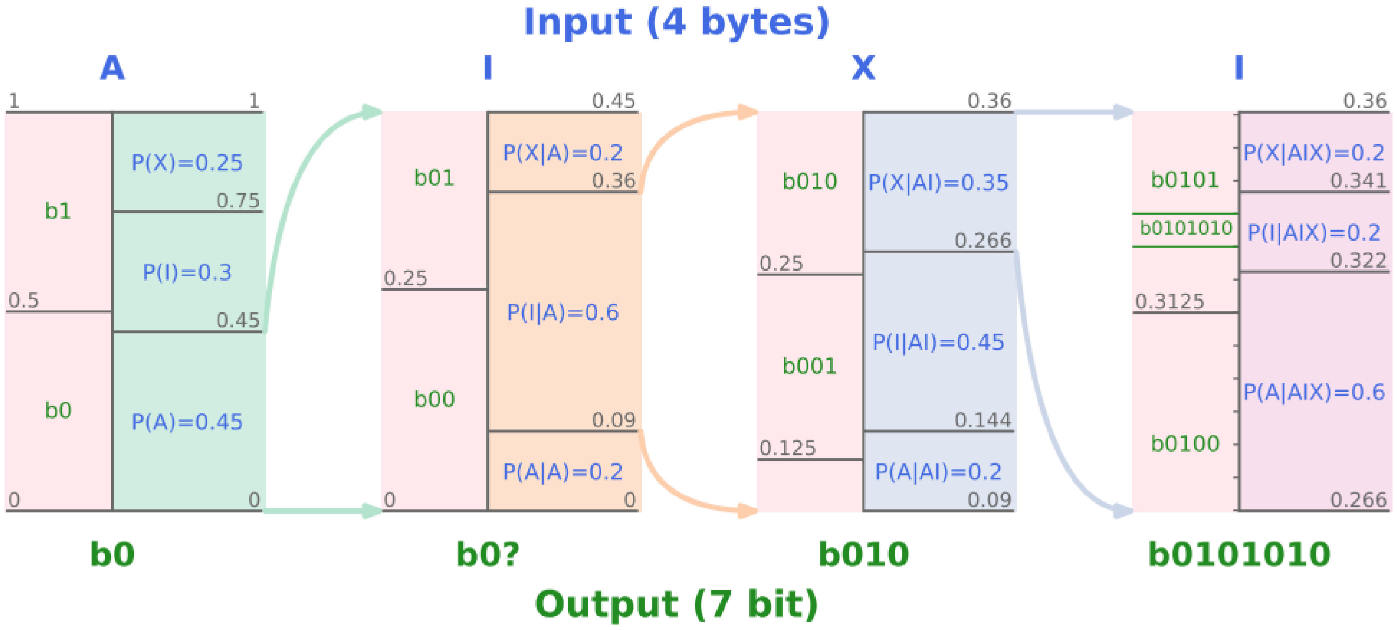
Using an Autoregressive Predictor to design a Lossless Compressor

Arithmetic Encoder:

Initially, this interval is $I_0 = [0, 1)$.

When encoding x_k, we first partition the previous interval $I_{k-1} = [l_{k-1}, u_{k-1})$ into Nsub-intervals $I_k(x_1), I_k(x_2), \ldots,$ one for each letter from $\mathrm{X} = \{x_1, \ldots, x_N\}$. The size of sub-interval $I_k(y)$ that represents letter y is $(u_{k-1} - l_{k-1}) \cdot P(y \,|\, x < k)$.

e.g., encoding AIXI



The encoding length of arithmetic encoder is $l_c(x_{1:n}) = -[log_2 P(x_{1:n})] + 1$

Prediction vs Compression

Equivalence of Prediction and Compression

The better we can predict next token, the better we can compress the sequence

Shannon’s source coding theorem

$$\textit{Average Code length} \ \geq \ \textit{entropy} \quad H(P)$$

$\textit{Average Code length}$:

$$L(Q_c) \ := \ \mathbb{E}_{X\sim P_\phi}\big[\ell\big(c(X)\big)\big] \ = \ \mathbb{E}_{X\sim P_\phi}\big[-\log Q_c(X)\big] \ = \ H(P_\phi \,\|\, Q_c).$$

Redundancy of code Q_c:

KL divergence

$$\mathrm{Red}(Q_c, P) = L(Q_c) - H(P_\phi) = H(P_\phi \,\|\, Q_c) - H(P_\phi) = D_{\mathsf{KL}}(P_\phi \,\|\, Q_c).$$

Outline

- LLM Theory
    - Fundamental Ideas from Shannon and Kolmogorov
    - Compression and Prediction
    - Kolmogorov’s theory
    - Data Modeling (a nonparametric model)
    - Hallucination and ICL
    - Universal Predictor
    - Research Directions

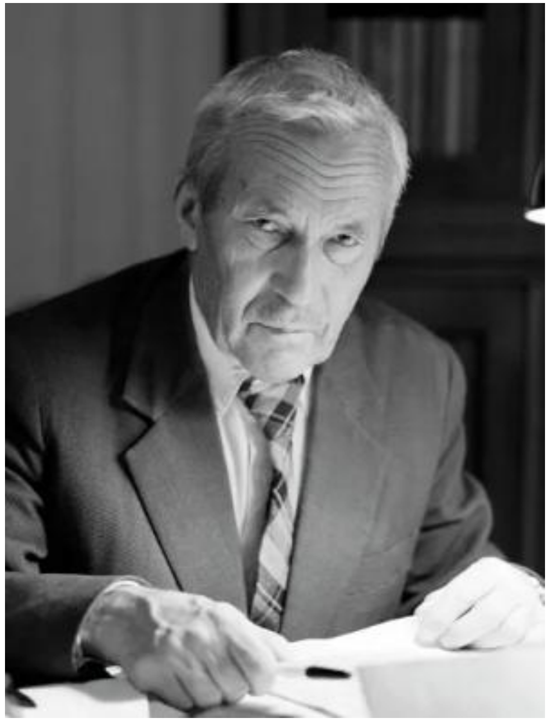
Kolmogorov Complexity

Kolmogorov Complexity (algorithmic information theory)

K(X):= The minimum length of TM that outputs X.

(a fine-grained structure: Kolmogorov structure function)

- Direct connection to compression
- In some sense, the ultimate notion of compression
- Do not need to know the exact distribution (unlike Shannon's information theory)
- Downside: incomputable…



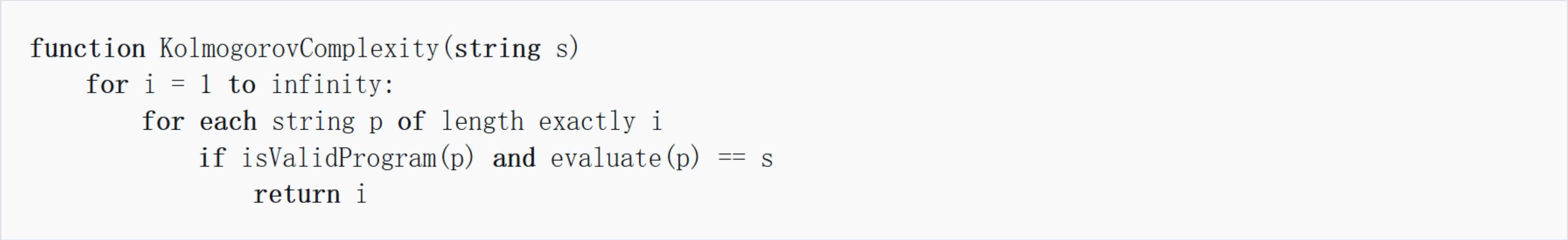
Examples:

- abababababababababababababababa….
- 4c1j5b2p0cv4w1x8rx2y39umgw5q85s7…
- 31415926535897932384626433832795…

see the classic book “elements of Information Theory”

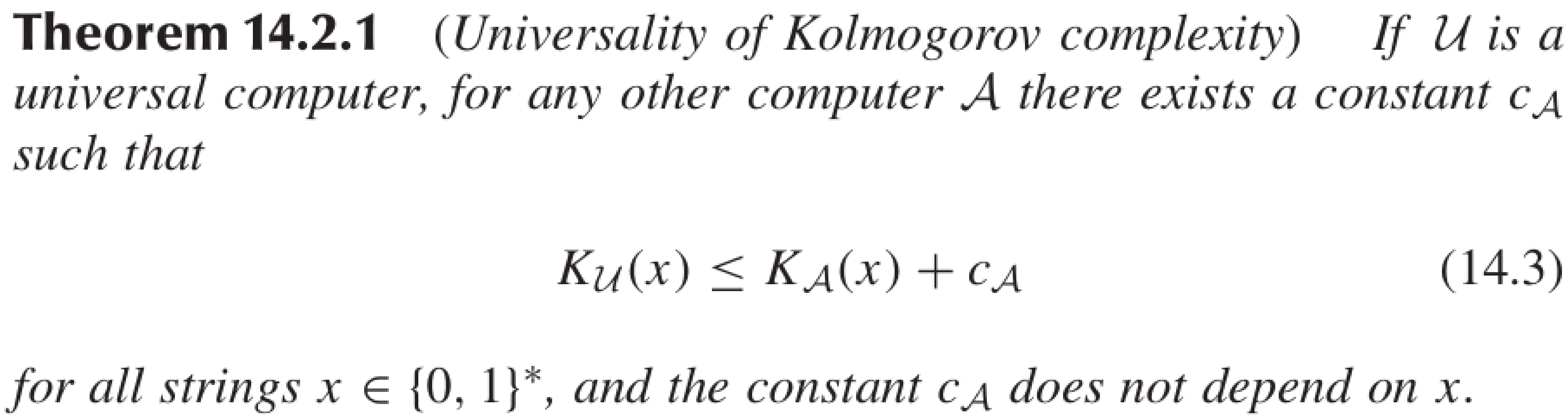
Kolmogorov Complexity

Can the following program output K(s)?



```
function KolmogorovComplexity(string s)
    for i = 1 to infinity:
        for each string p of length exactly i
            if isValidProgram(p) and evaluate(p) == s
                return i
```

Universality of Kolmogorov complexity



Theorem 14.2.1 (*Universality of Kolmogorov complexity*) *If $\mathcal{U}$ is a universal computer, for any other computer $\mathcal{A}$ there exists a constant $c_{\mathcal{A}}$ such that*

$$K_{\mathcal{U}}(x) \le K_{\mathcal{A}}(x) + c_{\mathcal{A}} \qquad (14.3)$$

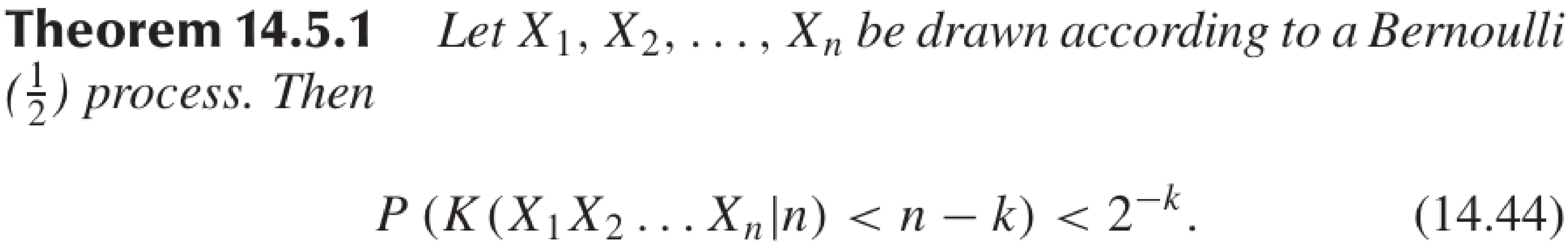
for all strings $x \in \{0, 1\}^*$, and the constant $c_{\mathcal{A}}$ does not depend on x.

Kolmogorov Complexity

- **Algorithmic randomness**
  - What is a random string? A impressible one.
  - We say a seq $x_1 \ldots x_n$ is algorithmically random if

$$K(x_1 x_2 \ldots x_n | n) \geq n.$$

  - Most sequences are random (interesting sequences are rare)



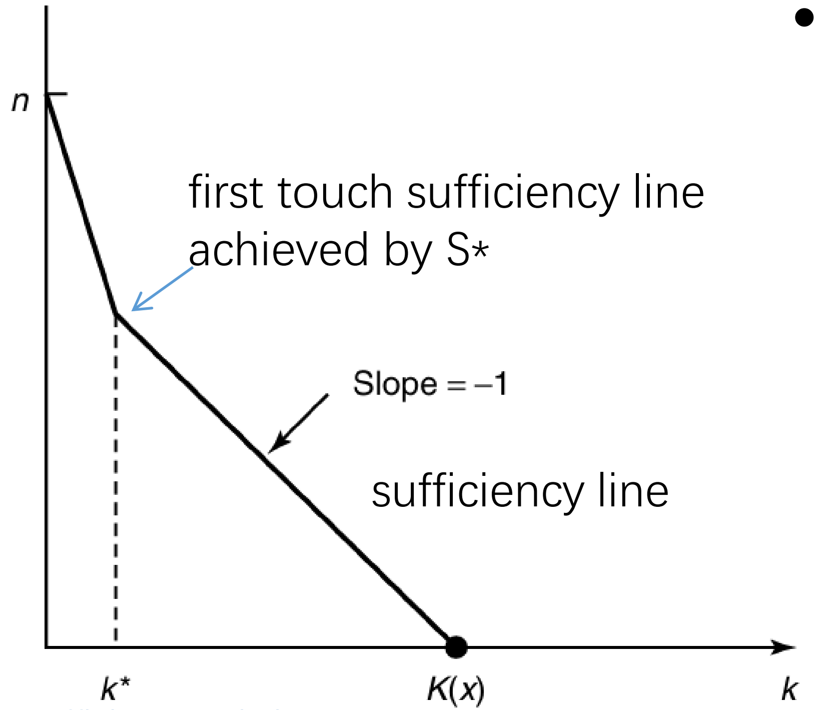
Theorem 14.5.1 *Let $X_1, X_2, \ldots, X_n$ be drawn according to a Bernoulli $(\frac{1}{2})$ process. Then*

$$P\left(K(X_1 X_2 \ldots X_n | n) < n - k\right) < 2^{-k}. \tag{14.44}$$

This can be easily seen from the fact that $|\{x \in \{0,1\}^* : K(x) < k\}| < 2^k$.

Kolmogorov structure function

$$h_X(\alpha) = \min_S\{\log|S| : S \ni X; K(S) \leq \alpha, \}$$



first touch sufficiency line achieved by S∗

sufficiency line

minimum sufficient statistics

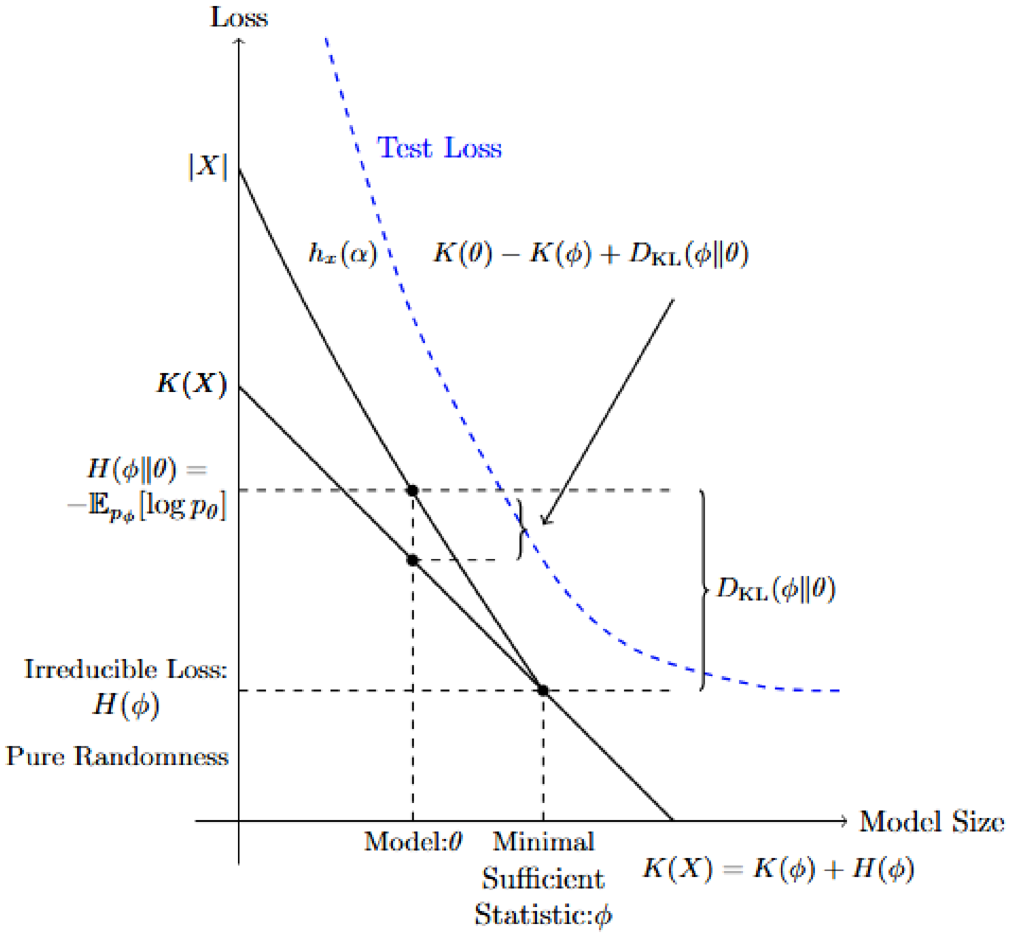
FIGURE 14.4. Kolmogorov sufficient statistic.

- Two part code (model-to-data code)

$$K(X) < |C(X)| + K(C) + O(1)$$

- Hence the set S∗ captures all the structure within x.
- The remaining description of x within S∗ is essentially the description of the randomness within the string.
- Hence S∗ is called the Kolmogorov sufficient statistic for x.

Kolmogorov structure function



- Two part code (model-to-data code)

$$K(X) < |C(X)| + K(C) + O(1)$$

- Closely related to Scaling Law
- Distinguish structure & random noise
  - a fine-grained hierarchy
- Why a power law shape?
- A characterization of “what should be learnt” and “what is learnt first”

Outline

- LLM Theory
  - Fundamental Ideas from Shannon and Kolmogorov
  - Compression and Prediction
  - Kolmogorov’s theory
  - Data Modeling (a nonparametric model)
  - Hallucination and ICL
  - Universal Predictor
  - Research Directions

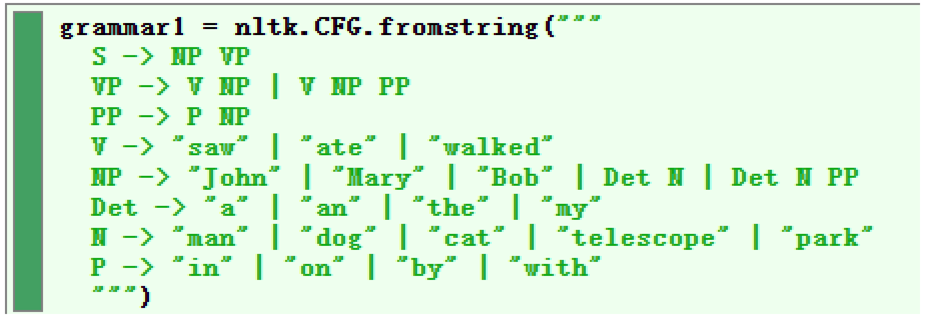
Data Modeling – A Hierarchical Nonparametric Model

A Hierarchical Nonparametric Model:

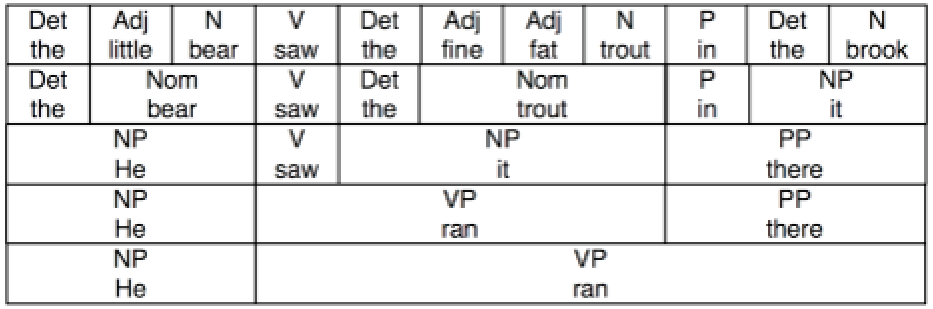
- An encoder (syntax, most common knowledge, basic logic): can be captured by a fair small-sized probabilistic TM that is fairly easy to learn (low sample complexity).

- World (factual) Knowledge: a large body of knowledge, that is constantly growing (consider the number of facts, set of proteins, species, chemical substances etc.)

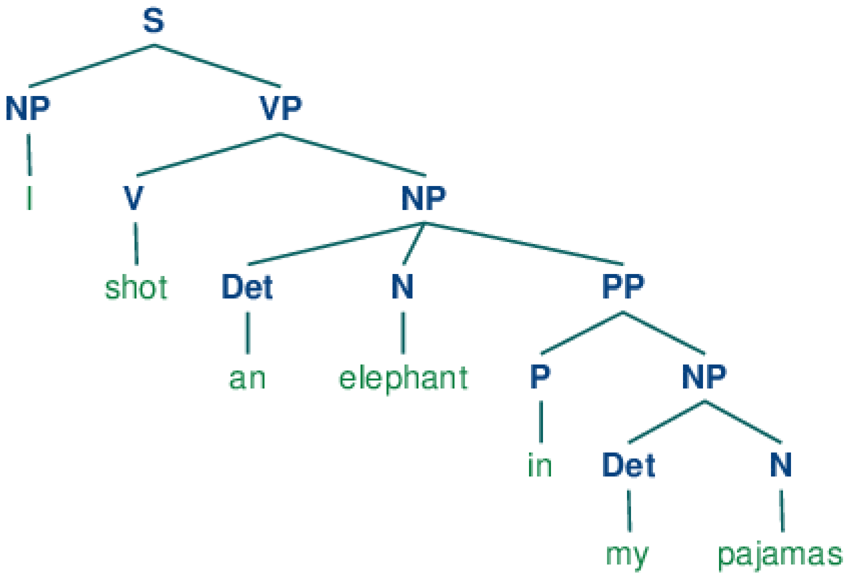
A syntax encoder: (probabilistic) grammar



```
grammar1 = nltk.CFG.fromstring("""
  S -> NP VP
  VP -> V NP | V NP PP
  PP -> P NP
  V -> "saw" | "ate" | "walked"
  NP -> "John" | "Mary" | "Bob" | Det N | Det N PP
  Det -> "a" | "an" | "the" | "my"
  N -> "man" | "dog" | "cat" | "telescope" | "park"
  P -> "in" | "on" | "by" | "with"
  """)
```



Det the	Adj little	N bear	V saw	Det the	Adj fine	Adj fat	N trout	P in	Det the	N brook
Det the	Nom bear		V saw	Det the	Nom trout			P in	NP it	
NP He			V saw	NP it				PP there		
NP He			VP ran					PP there		
NP He			VP ran							



Data Modeling – A Hierarchical Nonparametric Model

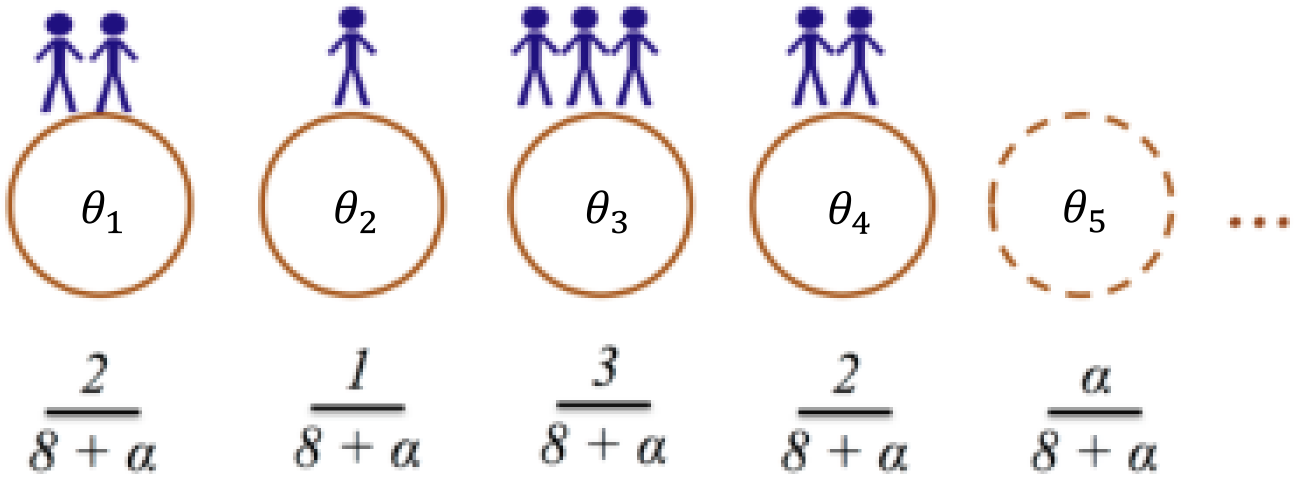
A Hierarchical Nonparametric Model:

- An encoder (syntax, most common knowledge, basic logic): can be captured by a fair small-sized probabilistic TM that is fairly easy to learn (low sample complexity).

- World (factual) Knowledge: a large body of knowledge, that is constantly growing (consider the number of facts, set of proteins, species, chemical substances etc.)

World (factual) Knowledge:

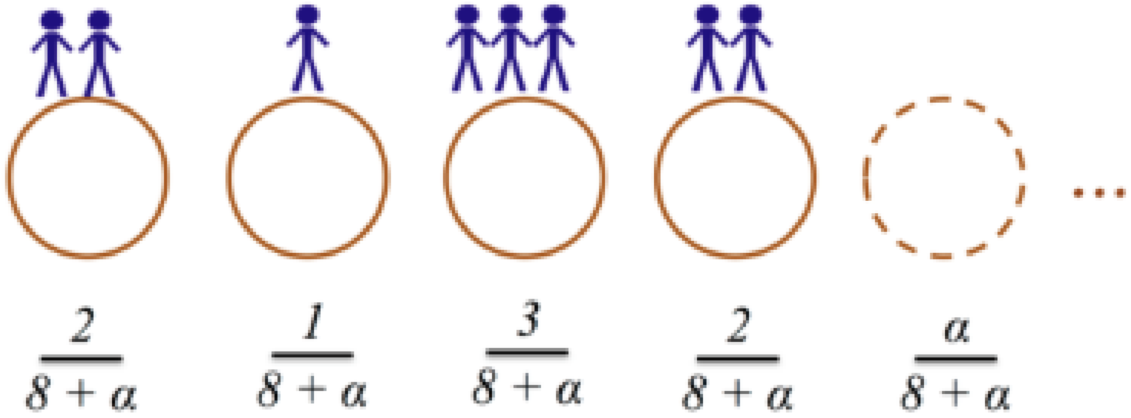
- Can't be captured by a fixed parametric model
- Modeled by **Pitman-Yor Chinese Restaurant Process (PY-CRP)**



Pitman-Yor Chinese Restaurant Process (PY-CRP)

- Preferential attachment

$$\text{For the } n\text{-th customer: } \begin{cases} \dfrac{N_k - \alpha}{n - 1 + \beta}, & \text{if joining an existing table } k, \\ \dfrac{\beta + \alpha K}{n - 1 + \beta}, & \text{if starting a new table,} \end{cases}$$



- leads to a power-law distribution

Lemma G.2 (Theorem 3.13 in Pitman (2006)). *Let* $p = (p_1, p_2, \dots) \sim \mathrm{PYCRP}(\alpha, \beta)$ *be the sequence of mixing weights drawn from a Pitman–Yor process. Then, the following limit almost surely exists:*

$$S_{\alpha,\beta} = \lim_{i \to \infty} i^{1/\alpha} p_i.$$

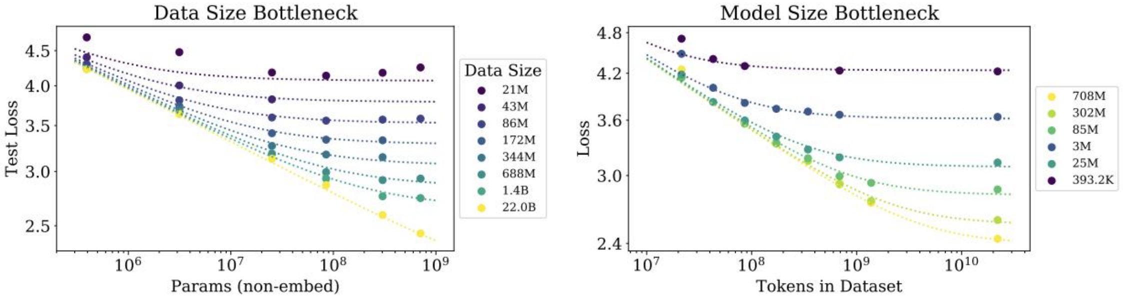
Scaling Laws

- Kaplan Scaling Law (OpenAI)

$$L(N, D) = \left[\left(\frac{N_c}{N}\right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D}\right]^{\alpha_D}$$

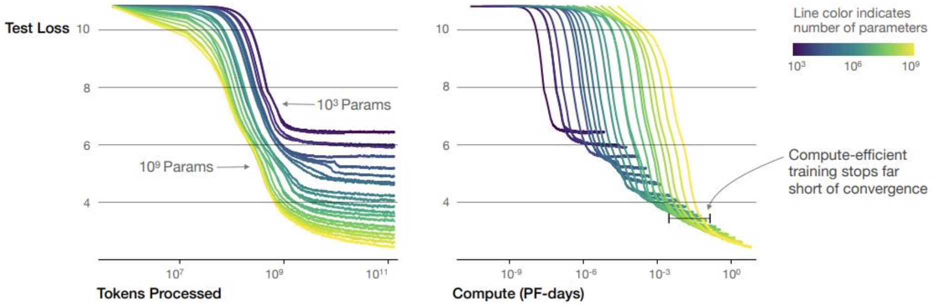
$\alpha_N \sim 0.076$, $N_c \sim 8.8 \times 1013$ (non-embedding parameters)

$\alpha_D \sim 0.095$, $D_c \sim 5.4 \times 1013$ (tokens)



Larger models require **fewer samples** to reach the same performance

The optimal model size grows smoothly with the loss target and compute budget



L: the loss
N: model size;
D: dataset size (the token number of training data).

- Chinchilla Scaling Laws (DeepMind)

$$L(N, D) = E + \frac{A}{N^{0.34}} + \frac{B}{D^{0.28}},$$

$E = 1.69$, $A = 406.4$, $B = 410$.

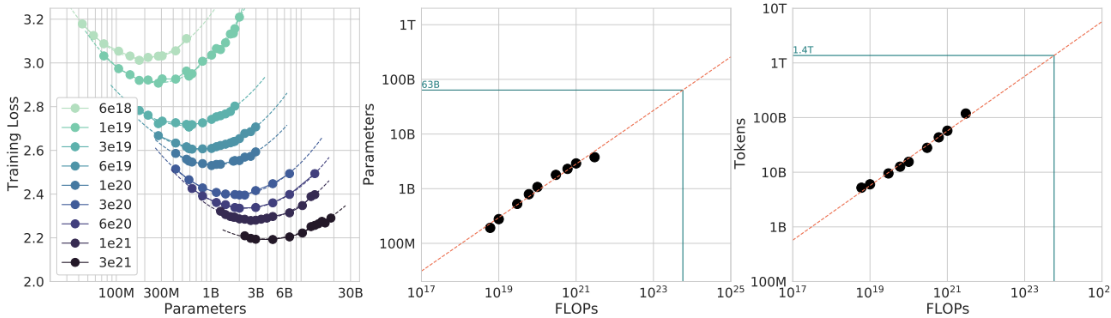


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

A Coding Game

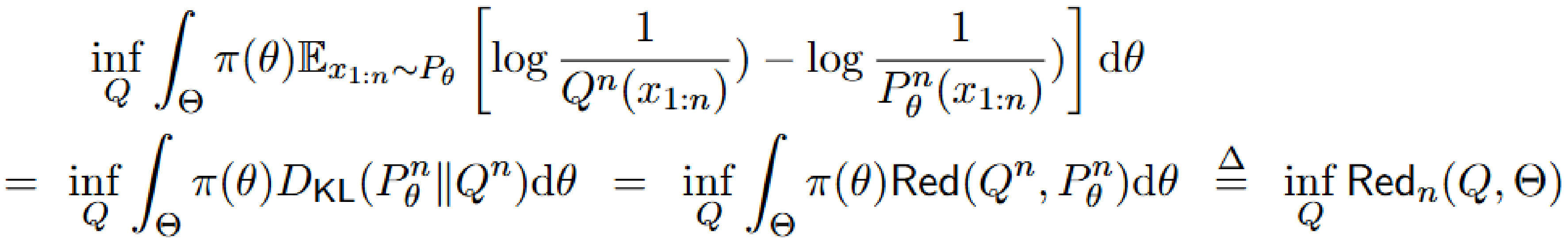
- Suppose P_{θ} is a distribution indexed by θ.
- Bayesian setting: there is a prior π over θ
- The player would like to model P_{θ} using Q
- Observe x1…xn one by one
- Minimize the following log-loss

$$\inf_{Q} \int_{\Theta} \pi(\theta) \mathbb{E}_{x_{1:n} \sim P_{\theta}} \left[\log \frac{1}{Q(x_{1:n})}) \right] \mathrm{d}\theta = \inf_{Q} \int_{\Theta} \pi(\theta) \sum_{i=1}^{n} \mathbb{E}_{P_{\theta}} \left[\log \frac{1}{Q(x_i \mid x_{1:i-1})}) \right] \mathrm{d}\theta.$$

minimize the bayesian code length
(cross-entropy)

A Coding Game

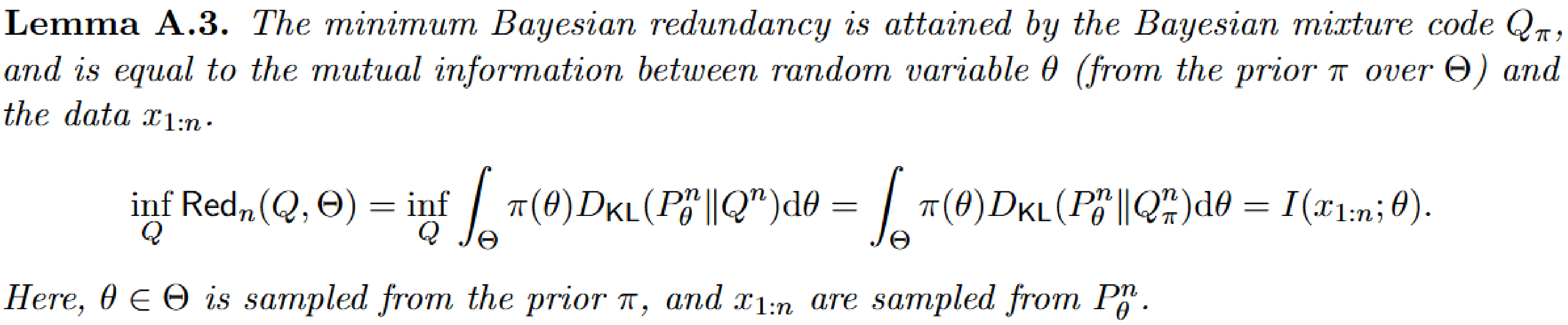
Bayesian Redundancy:



$$\inf_{Q}\int_{\Theta}\pi(\theta)\mathbb{E}_{x_{1:n}\sim P_\theta}\left[\log\frac{1}{Q^n(x_{1:n})})-\log\frac{1}{P_\theta^n(x_{1:n})})\right]\mathrm{d}\theta$$

$$=\inf_{Q}\int_{\Theta}\pi(\theta)D_{\mathsf{KL}}(P_\theta^n\|Q^n)\mathrm{d}\theta=\inf_{Q}\int_{\Theta}\pi(\theta)\mathsf{Red}(Q^n,P_\theta^n)\mathrm{d}\theta\triangleq\inf_{Q}\mathsf{Red}_n(Q,\Theta)$$

Connection of redundancy and mutual information



Lemma A.3. *The minimum Bayesian redundancy is attained by the Bayesian mixture code Q_π, and is equal to the mutual information between random variable θ (from the prior π over Θ) and the data $x_{1:n}$.*

$$\inf_{Q}\mathsf{Red}_n(Q,\Theta)=\inf_{Q}\int_{\Theta}\pi(\theta)D_{\mathsf{KL}}(P_\theta^n\|Q^n)\mathrm{d}\theta=\int_{\Theta}\pi(\theta)D_{\mathsf{KL}}(P_\theta^n\|Q_\pi^n)\mathrm{d}\theta=I(x_{1:n};\theta).$$

Here, $\theta\in\Theta$ is sampled from the prior π, and $x_{1:n}$ are sampled from P_θ^n.

Understanding Scaling Law

Under the Bayesian prediction framework, one can show that the minimum Bayesian redundancy is equal to the mutual information (between the data and prior)

$$\inf_{Q} \mathsf{Red}_n(Q,\Theta) = \inf_{Q} \int_{\Theta} \pi(\theta) D_{\mathsf{KL}}(P_\theta^n \| Q^n) \mathrm{d}\theta = \int_{\Theta} \pi(\theta) D_{\mathsf{KL}}(P_\theta^n \| Q_\pi^n) \mathrm{d}\theta = I(x_{1:n};\theta).$$

Theorem (informal): Under the above data generative model and unlimited model size, we can prove that the **Bayesian predictor** (or Maximum Likelihood predictor in large data regime) has the following loss

$$\inf_{M} \frac{1}{N} \mathbb{E}_{\phi_{data}\sim\pi, X_{1:N}\sim P_{\phi_{data}}}\left[-\log P_M(X_{1:N})\right] = \tilde{O}\left(\underbrace{\frac{d_{knw}}{N^{1-\alpha}}}_{\text{Loss incurred by learning the knowledge (power law)}} + \underbrace{\frac{n_s d_{syn}}{N}}_{\text{Loss incurred by learning the syntax encoder}}\right) + \underbrace{\frac{1}{N} H(X_{1:N}|\Phi_{data})}_{\text{Irreducible loss (pure randomness)}}.$$

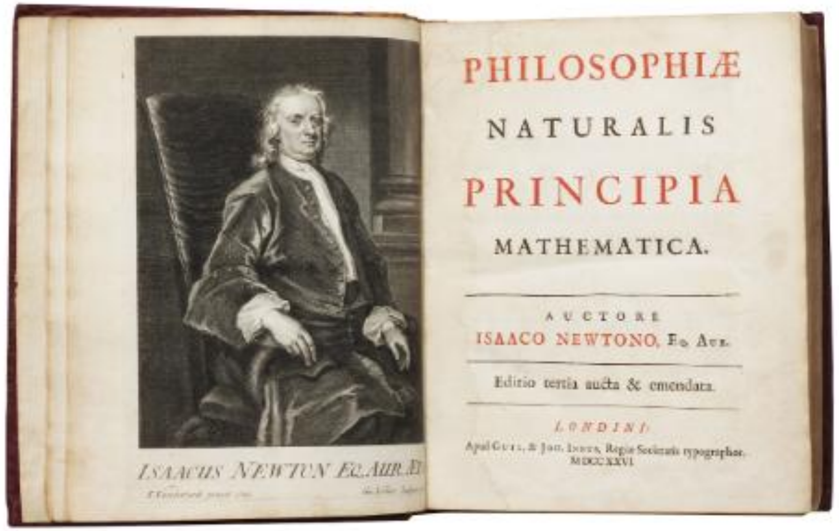
- Related to **Heap's Law (Heaps, 1978)** an empirical relationshipstating that the vocabulary size grows sublinearly with the size of a corpus N ,and **Zipf's Law (Zipf, 2016).**

Controlled experiments

Methodology popularized by Allen-Zhu and Li in a series of papers **“Physics of LLMs” -1,2,3**

Ethological approach

Controlled experiments
“LLM monkeys”

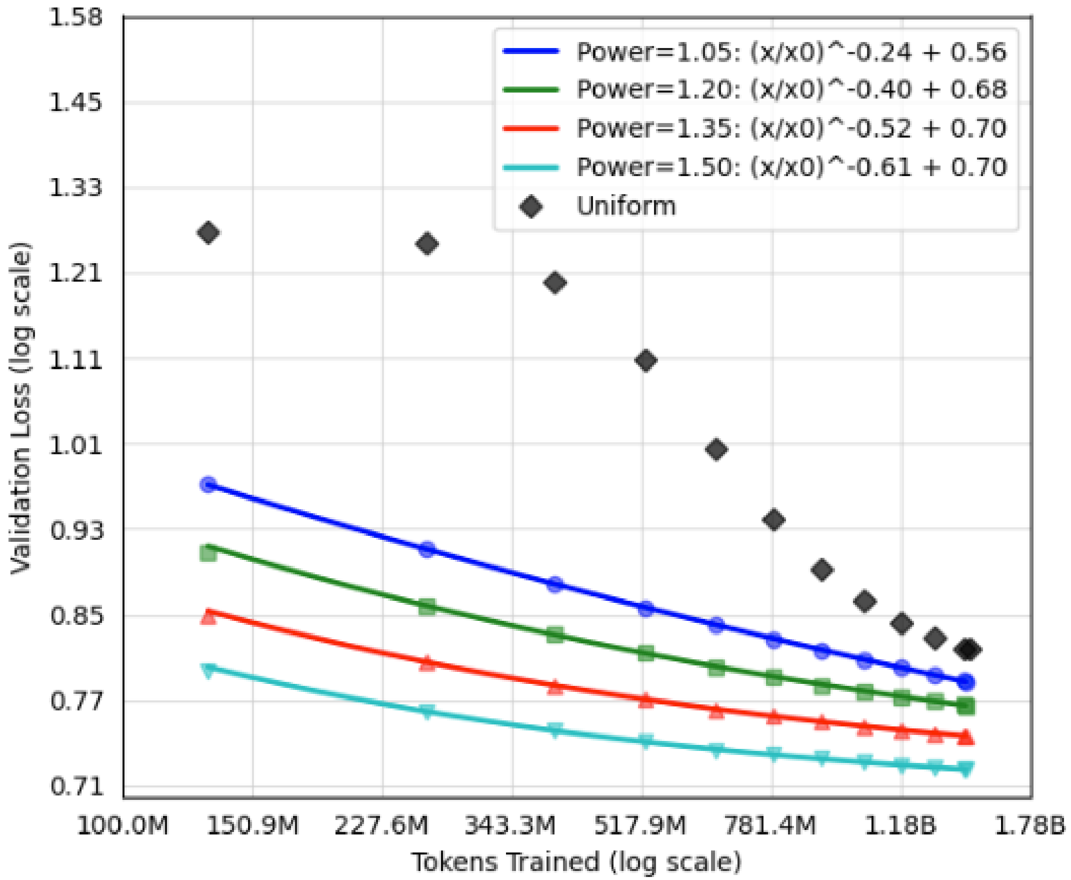


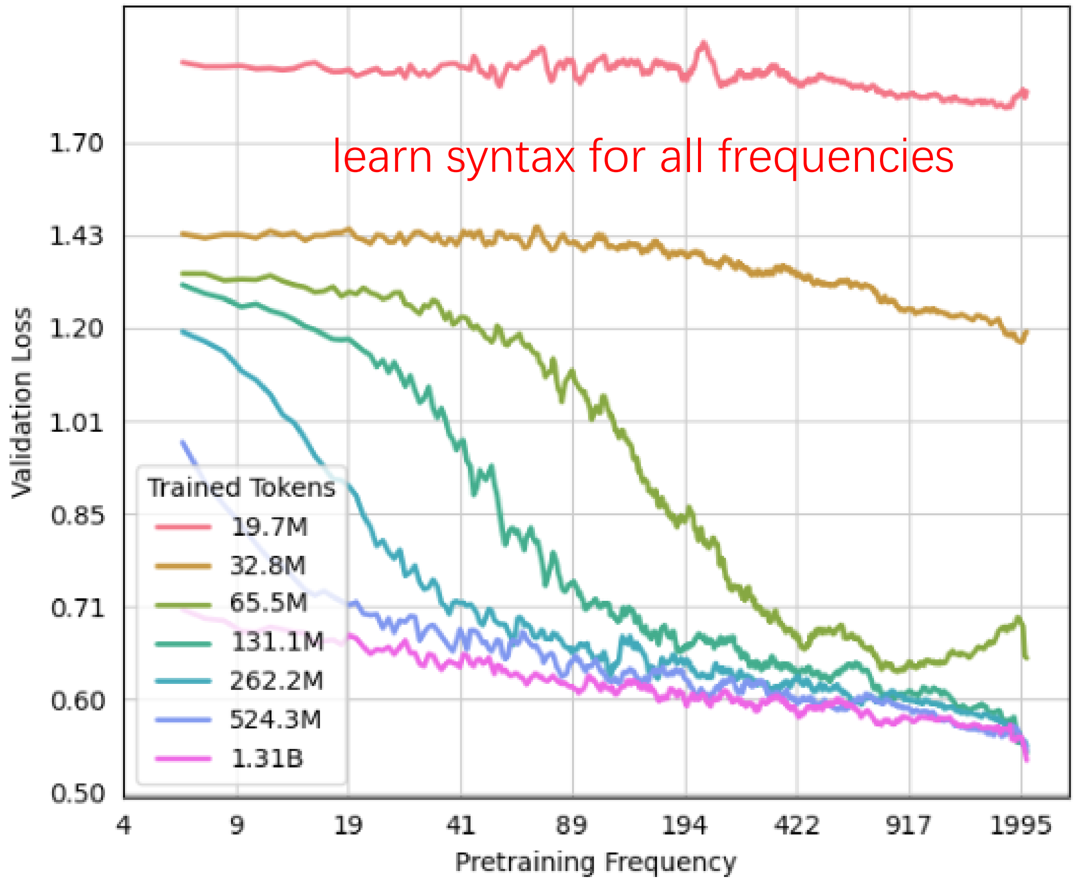
Physics

bioS dataset: we generate profiles for 400,000 individuals. Each profile contains six attributes: date of birth, birth city, university, major, employer, and employer city

“Gracie Tessa Howell wasborn in Camden, NJ. He studied Biomedical Engineering and worked at UnitedHealth Group. He entered the world on April 15, 2081, and is employed in Min-netonka. He is an alumnus/alumna of Buena Vista College.”

Data Scaling Law





learn syntax for all frequencies

higher frequency data learnt first

Model Scaling Law

Redundancy for the ith knowledge cluster:

$$\mathcal{D}_i(R) = \min_{\mathbb{I}(\phi_i; M_C^*) \le R} \mathbb{E}_{\phi_i \sim \pi_{\mathrm{knw}}} \left[\mathbb{E}_{q \sim P_{\phi_i}} \left[D_{\mathsf{KL}} \left(P_{\phi_i}(A \mid q) \parallel P_{M_C^*}(A \mid q) \right) \right] \right]$$

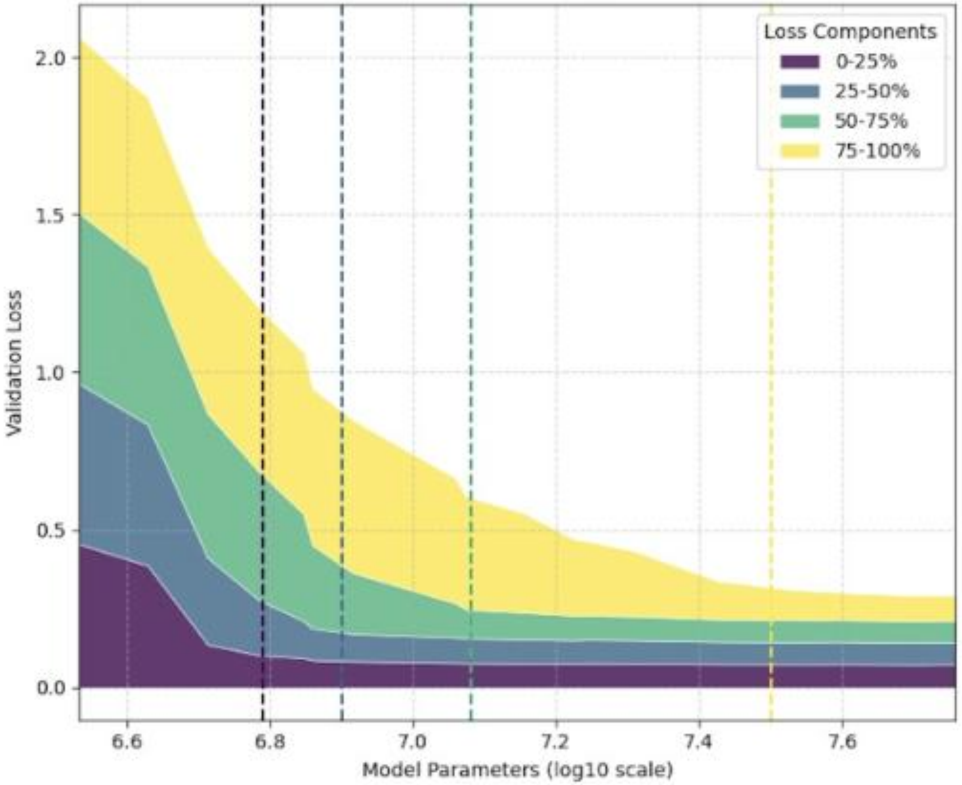
Model scaling law (total redundancy) can be captured by the following optimization problem

$$\begin{aligned} &\textbf{minimize} \quad \mathbb{E}_i[\mathcal{D}_i(m_i)] = \sum_{i=1}^{\infty} p_i \mathcal{D}_i(m_i), \\ &\textbf{subject to} \quad \mathbb{I}(\Phi_0; M_C^*) \le C, \quad m_i = \mathbb{I}(\phi_i; M_C^*) \ge 0 \text{ for all } i \in \mathbb{N}^+. \end{aligned}$$

can be solved using KKT condition introduced in the last class

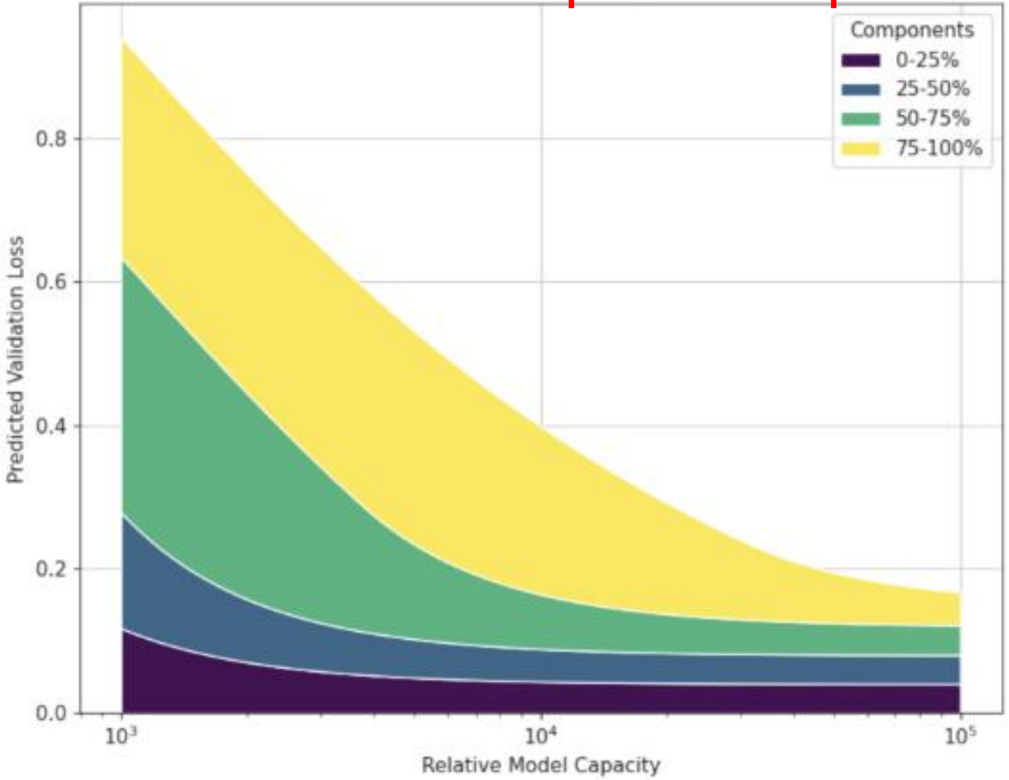
Model Scaling Law

Under the same data model, we have a theory of model scaling law (inspired by Kolmogorov structure function)



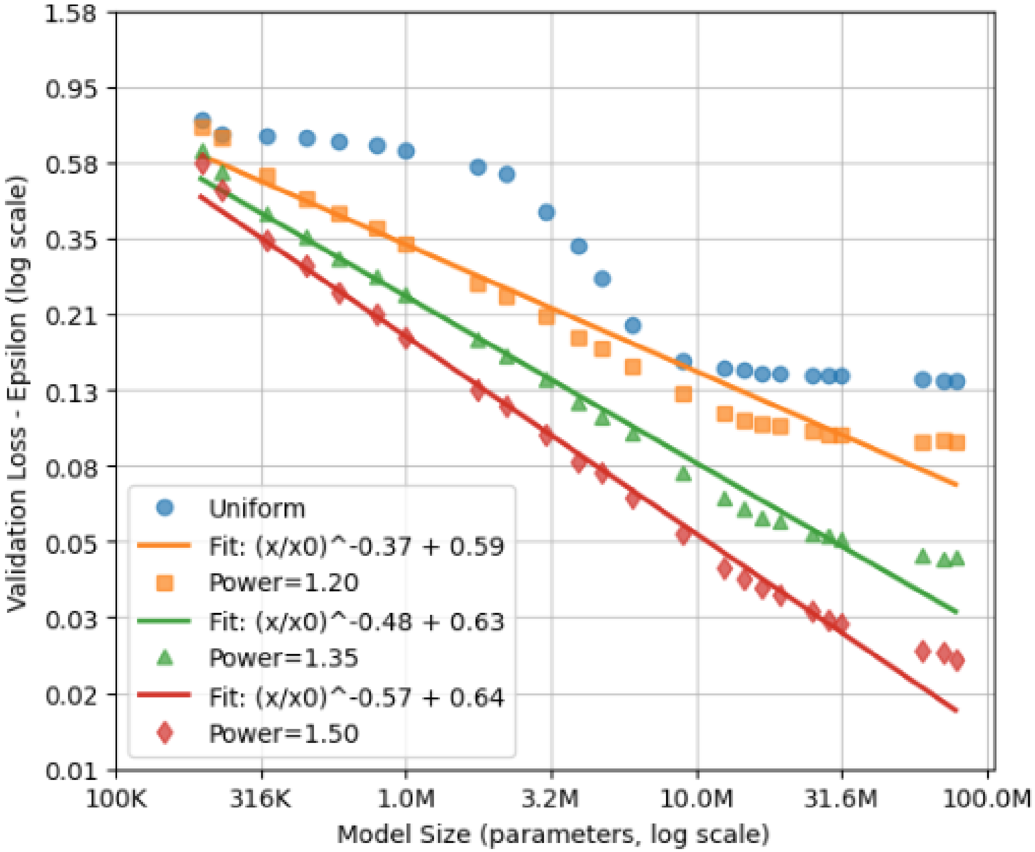
Empirical results

The solution of the optimization problem



$$
\begin{aligned}
\text{minimize} \quad & \mathbb{E}_i[\mathcal{D}_i(m_i)] = \sum_{i=1}^{\infty} p_i \mathcal{D}_i(m_i), \\
\text{subject to} \quad & \mathbb{I}(\Phi_0; M_C^*) \le C, \quad m_i = \mathbb{I}(\phi_i; M_C^*) \ge 0 \text{ for all } i \in \mathbb{N}^+
\end{aligned}
$$

Model Scaling Law



- For data generated from a uniform distribution, the loss
- curve significantly deviates from a power law

- It may be advantageous to have power-law-distributed data.

- The model can gradually learn knowledge in the order of frequency

- More effective than the uniform case, where no one element stands out and the model lacks guidance on what to prioritize.

Understanding Instruction Fine-Tuning

- Instruction fine-tuning:
  - Replace the encoder with another one (change the way the knowledge is encoded)
- The knowledge component stays the same

We can show the loss for Instruction fine-tuning can be bounded as (n' is the size of Instruction F-T data)

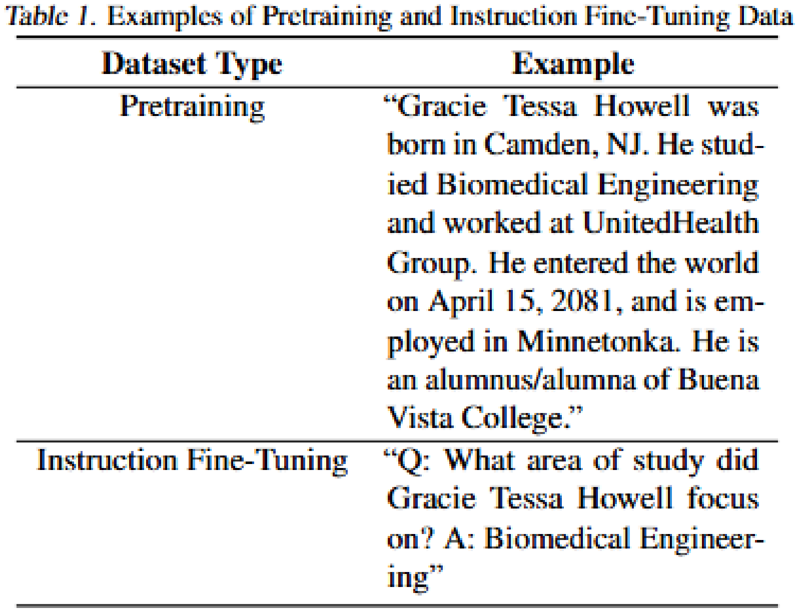


Table 1. Examples of Pretraining and Instruction Fine-Tuning Data

Dataset Type	Example
Pretraining	"Gracie Tessa Howell was born in Camden, NJ. He studied Biomedical Engineering and worked at UnitedHealth Group. He entered the world on April 15, 2081, and is employed in Minnetonka. He is an alumnus/alumna of Buena Vista College."
Instruction Fine-Tuning	"Q: What area of study did Gracie Tessa Howell focus on? A: Biomedical Engineering"

$$\frac{1}{n}\mathrm{Red}_{\mathrm{knw}} = \mathbb{I}(\boldsymbol{\kappa}_{1:N+n};\phi_{\mathrm{knw}}) - \mathbb{I}(\boldsymbol{\kappa}_{1:N};\phi_{\mathrm{knw}}) = \frac{(N+n)^{\alpha} - N^{\alpha}}{n} \overset{(1)}{=} O(N^{\alpha-1})$$

Knowledge in pretrained phase is retained

$$\frac{1}{n}\mathrm{Red}_{\mathrm{in}} = \mathbb{I}(S_{N+1:N+n};\phi_{\mathrm{ins}}) = \tilde{O}(n^{-1}).$$

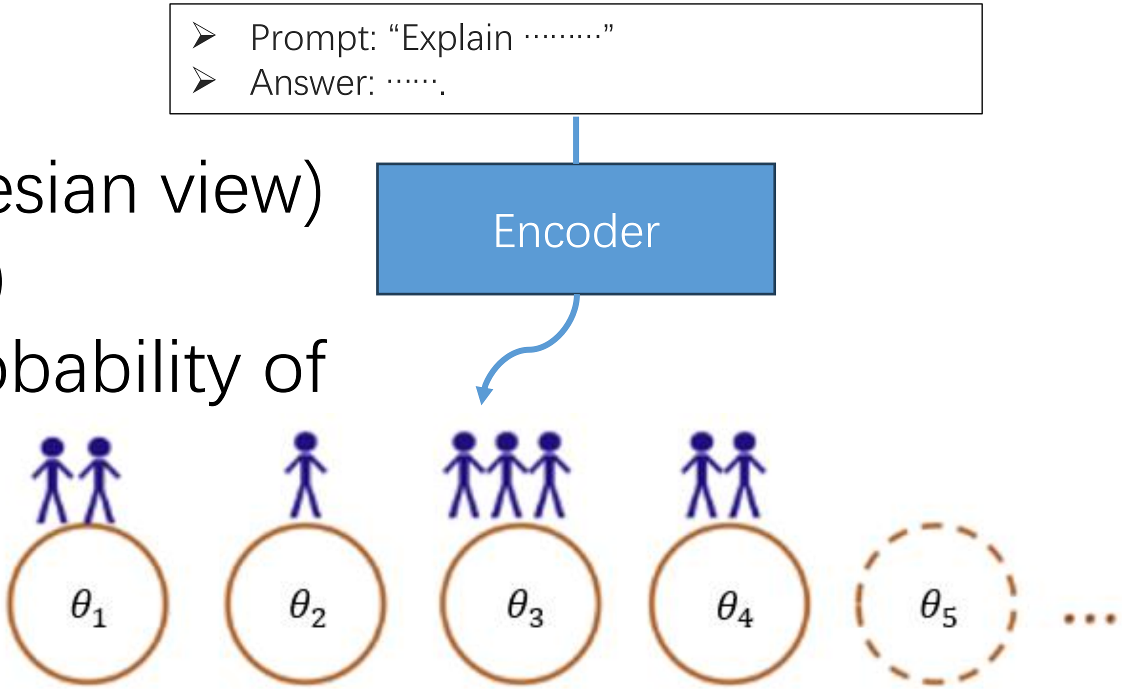
Redundancy incurred by learning the new encoder

Practical Implication: Instruction F-T is more useful for generation according to the instruction, but less effective for injecting new knowledge

Hallucination and ICL

- Causes of hallucination on (factual knowledge)
  - Insufficient samples to learn the fact
  - Insufficient model size (related to model scaling law)
  - Confusing/ambiguous prompt
  - Conflicting knowledge in training data
- Other hallucinations (not covered, future work)

- Explaining In-Context Learning (Bayesian view)
  - Consider $P(next\ token \mid prompt)$
  - Prompt increase the posterior probability of the relevant table



- ➢ Prompt: “Explain ………”
- ➢ Answer: …….

Outline

- LLM Theory
  - **Fundamental Ideas from Shannon and Kolmogorov**
  - Compression and Prediction
  - Kolmogorov's theory
  - Data Modeling (a nonparametric model)
  - Hallucination and ICL
  - Universal Predictor
  - Research Directions

Universal Predictor: Solomonoff's theory

- Dartmouth Summer Research Conference on Artificial Intelligence, where Solomonoff was one of the original 10 invitees
- A formal framework for universal inductive inference based on Kolmogorov complexity and Bayesian inference
- A **universal prior** distr $m(x)$ over all strings

$$m(x) = \sum_{p:U(p)=x*} 2^{-|p|}$$

- To predict future data, given a sequence x observed so far, and a prediction y (next token), the conditional probability is the posterior

$$m(y \mid x) = \frac{m(xy)}{m(x)}$$

- The Solomonoff predictor is **universal** $m(x) \geq c_\mu \cdot \mu(x).$

Universal Predictor: Solomonoff's theory

- Hypothesis: modern LLMs is a (rough) approximation of Solomonoff’s predictor.
  - We can explain various behaviors of LLMs using this theory
- Drawbacks
  - Unfortunately, Solomonoff’s predictor is again incomputable.
  - Solomonoff’s predictor ignores the following important aspects
    - the model size constraint
    - Architecture constraints of transformer (can not represent certain function composition)
    - Some (even simple) TMs are not efficiently learnable (XOR problem, one way functions etc.)

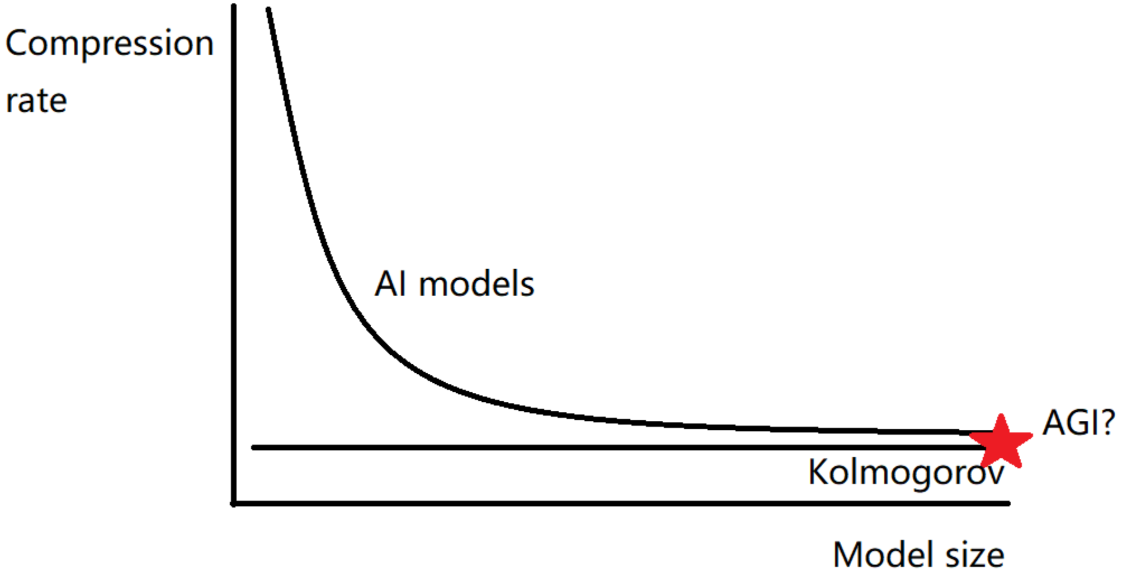
  - We should refine Solomonoff's theory to take account of the above (computational and statistical) barriers

Research Questions (LLM theory)

- Theory on more general data generative model (beyond syntax and factual knowledge models)
  - Kolmogrov's theory can be a good guideline
  - Refined Universal Predictor (Solomonoff's) Model
  - Methodology: Physics of LLMs (controlled experiments)
  - Scaling law, ICL, instruction following, emergence etc.
- Data <----> Skills
  - Data importance
  - High quality synthetic data (motivated from theory)
- Detecting Hallucination/Safety issues (from activation pattern etc.)

Discussions

Compression vs AGI ? （Sutskever’s view)



Ilya Sutskever “an observation of generalization”

Discussions
Compression vs AGI？ (Sutskever's view)

Compression ≠ AGI!!

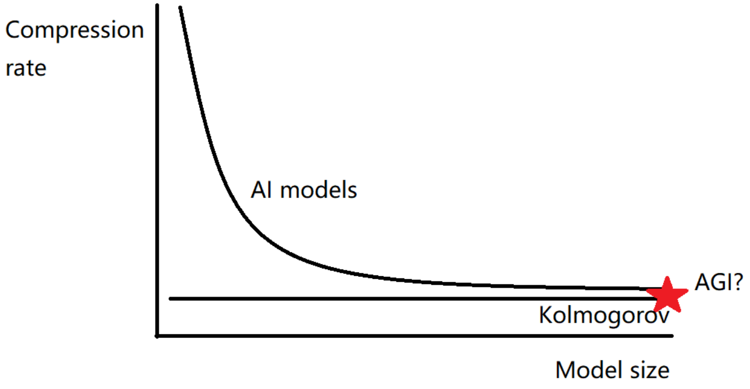
压缩即智能 （这个深刻的句子是有局限性的！）

Compression = Intelligence in inductive inference

压缩即推理智能

压缩无法产生探索智能（基于探索的创造力）

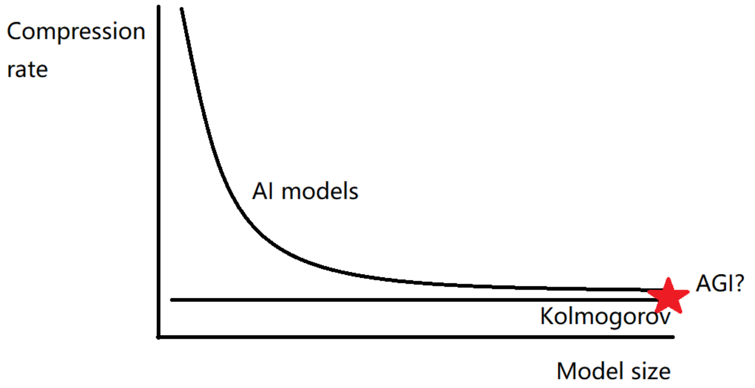
Ilya Sutskever "an observation of generalization"



Compression by itself cannot reflect exploration of physical world and mathematical/logical reasoning, in my opinion

Discussions

Compression vs AGI ?



Ilya Sutskever “an observation of generalization”

Compression = Intelligence in inductive inference

Compression by itself cannot reflect exploration of physical world and mathematical/logical reasoning

我觉得”压缩即智能“只在inductive inference范畴下是对得。没有capture智能体主动探索explore环境（和理论）和收集数据的过程。

我很难把explore环境的刺激动力用compression描述出来。一个牵强的说法：为了更好的predict next token，需要主动探索世界发现规律，然后predict next token和压缩有关系。

当然探索得到了数据，分析总结这些数据，来发现规律，需要的就是更好的压缩这些数据。

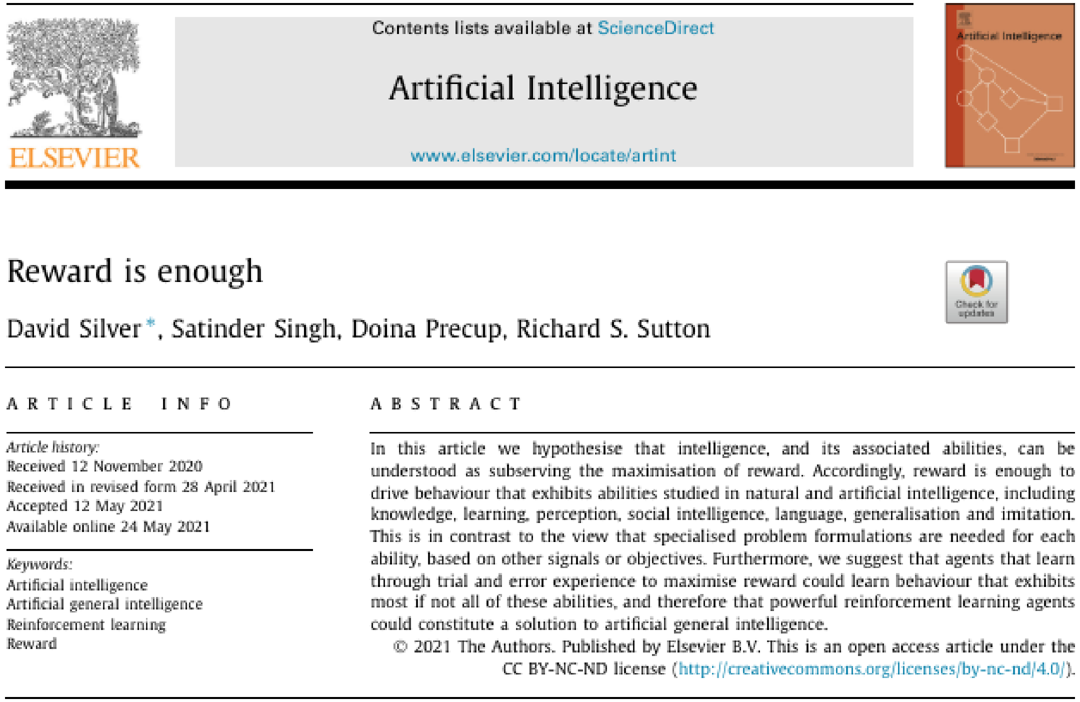
压缩即智能指的是获得数据后的智能过程，但是什么驱动探索，我觉得用压缩解释不了。也就是Ilya talk里有个图光minimize perplexity到不了AGI。

Discussion: RL and Reasoning

这个观点上我觉得reward is enough那个paper观点更根本。

智能产生的更根本的原因是无限的世界对有限资源智能体在特定情况下给与reward，有限资源智能体目标就是更有效的获得这些reward。that is all。为了获得reward，去explore，获得的data通过compress首先产生了intelligent的对世界的认识，然后又intelligent的behavior，然后再更好的收集数据。

比如寻找math proof这个事情，和人类认识世界类似，我说的那个proof构成的无限的图里有无数人类目前不关心的定理（比如一个有1000个条件的定理），目前这些个点上没有啥reward。目前的数据（有reward的）只cover其中很小的proportion。这个reward是目前人类数学家赋予的（这个reward function和数学研究里面的一些主义相关，比如直觉主义的，比如审美，还有数学是否需要和现实世界连接，还是存粹的推理 （von Neumann, Tao） 。



Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Reward is enough

David Silver*, Satinder Singh, Doina Precup, Richard S. Sutton

ARTICLE INFO

Article history:
Received 12 November 2020
Received in revised form 28 April 2021
Accepted 12 May 2021
Available online 24 May 2021

Keywords:
Artificial intelligence
Artificial general intelligence
Reinforcement learning
Reward

ABSTRACT

In this article we hypothesise that intelligence, and its associated abilities, can be understood as subserving the maximisation of reward. Accordingly, reward is enough to drive behaviour that exhibits abilities studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language, generalisation and imitation. This is in contrast to the view that specialised problem formulations are needed for each ability, based on other signals or objectives. Furthermore, we suggest that agents that learn through trial and error experience to maximise reward could learn behaviour that exhibits most if not all of these abilities, and therefore that powerful reinforcement learning agents could constitute a solution to artificial general intelligence.

But do we really need RL?? RL vs SFT on smartly collected data

Theoretical AGI:
AIXI(Artificial Intelligence exploration Institute)

- Environment Model: Assume the environment can be represented as a probability distribution $P(o_{t+1}, r_{t+1}|a_t, h_t)$ where o_{t+1} is the observation at time $t+1$, r_{t+1} is the reward at time $t+1$, a_t is the action at time t, h_t is the history up to time t.
- Prediction: AIXI uses the Solomonoff predictor to estimate future observations and rewards.
- Decision making: AIXI selects the action that maximizes the expected cumulative reward.

Marcus Hutter. “Universal Artificial Intellegence - Sequential Decisions Based on Algorithmic Probability.” Springer,2005.
http://hutter1.net/ai/suaibook.pdf

Thanks