Language Models and Power Laws

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Complexity Explorers Kraków, 22.02.2023

The dawn of large language models

- BERT, GPT-3, LaMDA, Palm — language models based on transformers and trained on nearly internet-sized corpora...
- ... excel at solving traditionally understood tasks in natural language processing (machine translation, etc.),
- ... can also generate quite arbitrary texts on demand:
  - rhymed poems,
  - school essays,
  - simple computer programs,
  - formalized proofs for automated theorem provers.

Question answering

Which blade was used by the hero in Jabberwocky?

In the poem "Jabberwocky" by Lewis Carroll, the hero uses a "vorpal blade" to slay the fearsome Jabberwock. The vorpal blade is described as a weapon of great power and sharpness, capable of cutting through anything in its path. It is not clear what type of blade the vorpal blade is intended to represent, as it is a fictional weapon created by the author. Some readers have speculated that it may be a sword, a knife, or some other type of bladed weapon. Ultimately, the specific type of blade used by the hero in "Jabberwocky" is a matter of interpretation.

Rhymed verses

Please write a rhymed poem about large language models.

Language models so large,
They run on a barge,
Training on data that is vast,
They generate predictions that last.

The power of these models is clear,
They can predict with no fear,
With accuracy out of sight,
They bring a helpful light.

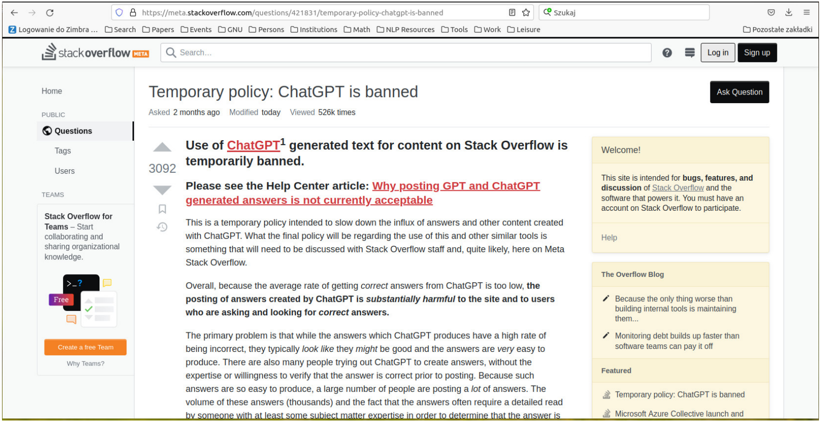
From natural language processing,
To text classification,
These models are a must,
A technology we can trust.

Programming

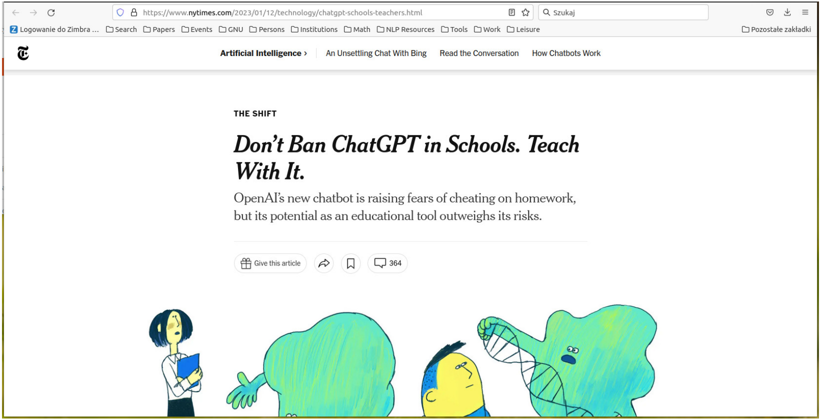
Write a Python function that checks if a number is a prime.

```
def is_prime(number):
  if number == 0 or number == 1:
    return False
  for x in range(2, number):
    if number % x == 0:
      return False
  return True
```

Ban it!


https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned
Szukaj
Logowanie do Zimbra ... Search Papers Events GNU Persons Institutions Math NLP Resources Tools Work Leisure
Pozostałe zakładki
stackoverflow META
Search...
Log in
Sign up
Home
PUBLIC
Questions
Tags
Users
TEAMS
Stack Overflow for Teams – Start collaborating and sharing organizational knowledge.
Free
Create a free Team
Why Teams?
Temporary policy: ChatGPT is banned
Ask Question
Asked 2 months ago Modified today Viewed 526k times
3092
Use of ChatGPT[1] generated text for content on Stack Overflow is temporarily banned.
Please see the Help Center article: Why posting GPT and ChatGPT generated answers is not currently acceptable
This is a temporary policy intended to slow down the influx of answers and other content created with ChatGPT. What the final policy will be regarding the use of this and other similar tools is something that will need to be discussed with Stack Overflow staff and, quite likely, here on Meta Stack Overflow.
Overall, because the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking and looking for correct answers.
The primary problem is that while the answers which ChatGPT produces have a high rate of being incorrect, they typically look like they might be good and the answers are very easy to produce. There are also many people trying out ChatGPT to create answers, without the expertise or willingness to verify that the answer is correct prior to posting. Because such answers are so easy to produce, a large number of people are posting a lot of answers. The volume of these answers (thousands) and the fact that the answers often require a detailed read by someone with at least some subject matter expertise in order to determine that the answer is
Welcome!
This site is intended for bugs, features, and discussion of Stack Overflow and the software that powers it. You must have an account on Stack Overflow to participate.
Help
The Overflow Blog
Because the only thing worse than building internal tools is maintaining them...
Monitoring debt builds up faster than software teams can pay it off
Featured
Temporary policy: ChatGPT is banned
Microsoft Azure Collective launch and

Use it!



https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

Artificial Intelligence › An Unsettling Chat With Bing Read the Conversation How Chatbots Work

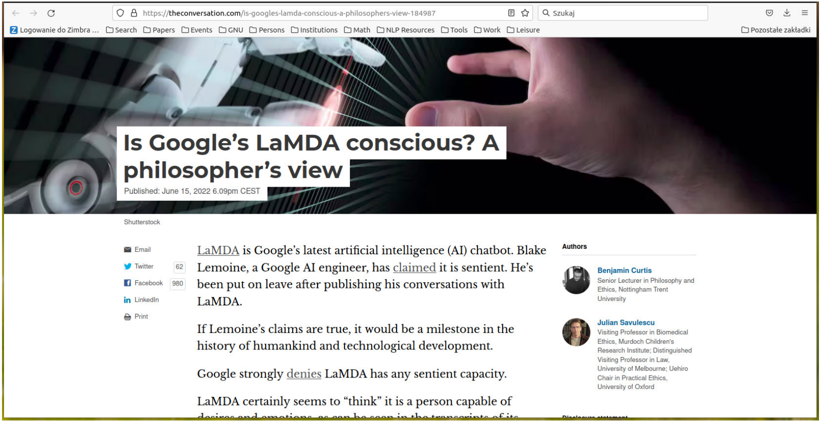
THE SHIFT

Don't Ban ChatGPT in Schools. Teach With It.

OpenAI's new chatbot is raising fears of cheating on homework, but its potential as an educational tool outweighs its risks.

Give this article 364

Believe it... or not!



https://theconversation.com/is-googles-lamda-conscious-a-philosophers-view-184987

Logowanie do Zimbra ... | Search | Papers | Events | GNU | Persons | Institutions | Math | NLP Resources | Tools | Work | Leisure | Pozostałe zakładki

Is Google's LaMDA conscious? A philosopher's view

Published: June 15, 2022 6.09pm CEST

Shutterstock

Email | Twitter 62 | Facebook 980 | LinkedIn | Print

LaMDA is Google's latest artificial intelligence (AI) chatbot. Blake Lemoine, a Google AI engineer, has claimed it is sentient. He's been put on leave after publishing his conversations with LaMDA.

If Lemoine's claims are true, it would be a milestone in the history of humankind and technological development.

Google strongly denies LaMDA has any sentient capacity.

LaMDA certainly seems to "think" it is a person capable of

Authors

Benjamin Curtis
Senior Lecturer in Philosophy and Ethics, Nottingham Trent University

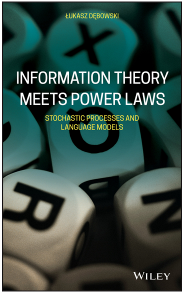
Julian Savulescu
Visiting Professor in Biomedical Ethics, Murdoch Children's Research Institute; Distinguished Visiting Professor in Law, University of Melbourne; Uehiro Chair in Practical Ethics, University of Oxford

How come and what next?

- Large language models appeared quite suddenly...
- ... made a huge progress within a few years,
- ... exhibit curious emergent behaviors.
- We are largely intellectually unprepared for their arrival.
- **Besides programming, we need theoretical insight: neuroscience, mathematics, philosophy, physics, ...**

My own interests in abstract language models

- I did my masters in theoretical physics (statistical mechanics).
- Later I worked in statistical natural language processing (part-of-speech tagging of the IPI PAN corpus of Polish).
- But my heart was taken by power laws and information theory (Zipf's law, Hilberg's hypothesis, refutation of finite-state models).
- I did my PhD in information theory and stochastic processes with long memory.
- Ever since then I have been working on mathematical foundations of statistical language modeling (measure theory, ergodic decomposition, excess entropy, Kolmogorov complexity, universal coding and universal prediction).
- **Quite a lot of pretty abstract math...**



Large language models surprised me, too!

Language Models & Power Laws

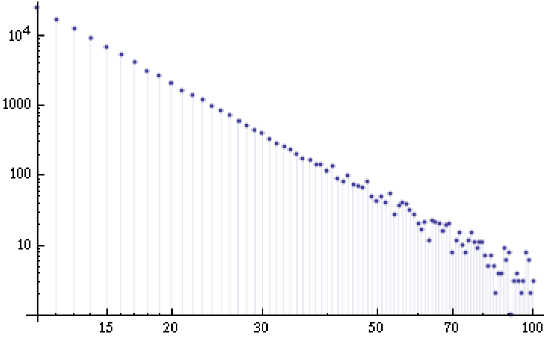
Power Laws in Language Models

What is a power law?

$\boldsymbol{N}, \boldsymbol{V}$ — some quantities of interest

$$\boldsymbol{V} \propto \boldsymbol{N}^{\gamma} \text{ for some parameter } \gamma > 0 \text{ or } \gamma < 0$$

A rough method of detection: the log-log plot



Power laws in complex systems

general math:

- Zipf's law
- fractals

physics:

- Kepler's third law
- Stefan-Boltzman's law

biology:

- Kleiber's law
- allometric laws
- Taylor's law

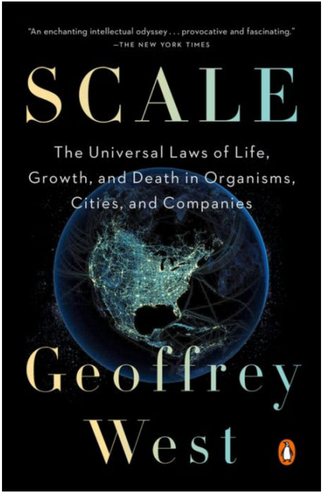
science of cities:

- Gibrat's law
- allometric laws

economics:

- distribution of income

An interesting book:



Zipf-Mandelbrot's and Herdan-Heaps' law

Shakespeare's
First Folio/35 Plays:

rank $r(w)$	freq $f(w)$	word w
1	21557	I
2	19059	and
3	16571	to
4	14921	of
5	14491	a
6	12077	my
7	10463	you
8	9789	in
9	8754	is
10	7428	that
...	...	...

Numbers of tokens and types:

$$N = \sum_w f(w), \quad V = \sum_w 1.$$

Zipf-Mandelbrot's law:

$$r(w) \approx \frac{V}{f(w)^\beta}, \quad \beta \in (0,1).$$

Herdan-Heaps' law:

$$V \propto N^\beta, \quad \beta \in (0,1).$$

Language models — Cross entropy

Let us write text $(x_1, x_2, ..., x_T)$ as x_1^T.

A language model is a (probability) measure on tokens:

$$Q(x_t|x_{t-M}^{t-1}) \geq 0, \quad \sum_{x_t} Q(x_t|x_{t-M}^{t-1}) = 1.$$

The cross entropy of the model is the mean minus log-probability:

$$-\frac{1}{T}\sum_{t=1}^{T} \log Q(x_t|x_{t-M}^{t-1}) \geq 0.$$

It is the average surprisal of model Q on text x_1^T.

We seek for Q that is a computable function of training data x_1^T and minimizes cross entropy on different data, called the test data.

Language models — Embeddings and transformers

In language models based on transformers, probabilities $Q(x_t|x_{t-M}^{t-1})$ are computed by stacking two mechanisms:

- embeddings — vectors x_t corresponding to words/concepts,
- attention — a nonlinear operation on embeddings

$$y_t = \sum_{s=t-M}^{t-1} \frac{\exp(x_t \cdot x_s)}{\sum_{r=t-M}^{t-1} \exp(x_t \cdot x_r)} x_s.$$

The GPT-3 language model:

- Number of parameters: $N = 175$ billions (800 GB RAM).
- Context length: $M = 2048$ words.
- Training data: Common Crawl (410 bln, 60%), WebText2 (19 bln, 22%), books (67 bln, 16%), Wikipedia (3 bln, 3%).

Language models — Power laws

$\boldsymbol{Q}(\boldsymbol{N},\boldsymbol{T})$ — model with $\boldsymbol{N}$ parameters trained on $\boldsymbol{T}$ tokens.

$\mathcal{L}(\boldsymbol{N},\boldsymbol{T})$ — cross entropy of $\boldsymbol{Q}(\boldsymbol{N},\boldsymbol{T})$ on the test data.

Kaplan et al. (2020) observed empirically that

$$\mathcal{L}(\boldsymbol{N},\boldsymbol{T}) \approx \left[\left(\frac{\boldsymbol{N}_0}{\boldsymbol{N}}\right)^{\frac{\gamma_{\boldsymbol{N}}}{\gamma_{\boldsymbol{T}}}} + \frac{\boldsymbol{T}_0}{\boldsymbol{T}}\right]^{\gamma_{\boldsymbol{T}}} \approx \max\left\{\left(\frac{\boldsymbol{N}_0}{\boldsymbol{N}}\right)^{\gamma_{\boldsymbol{N}}}, \left(\frac{\boldsymbol{T}_0}{\boldsymbol{T}}\right)^{\gamma_{\boldsymbol{T}}}\right\}$$

for $\boldsymbol{N}_0 = 6.4 \times 10^{13}$, $\boldsymbol{T}_0 = 1.8 \times 10^{13}$, $\gamma_{\boldsymbol{N}} = 0.076$, $\gamma_{\boldsymbol{T}} = 0.103$.

The more data and the more parameters, the better is the model:

$$\mathcal{L}(\infty,\boldsymbol{T}) \approx \left(\frac{\boldsymbol{T}_0}{\boldsymbol{T}}\right)^{\gamma_{\boldsymbol{T}}}, \quad \mathcal{L}(\boldsymbol{N},\infty) \approx \left(\frac{\boldsymbol{N}_0}{\boldsymbol{N}}\right)^{\gamma_{\boldsymbol{N}}}, \quad \mathcal{L}(\infty,\infty) \approx 0.$$

For each $\boldsymbol{T}$ there is roughly an optimal $\boldsymbol{N} = \boldsymbol{N}_0(\boldsymbol{T}/\boldsymbol{T}_0)^{\gamma_{\boldsymbol{T}}/\gamma_{\boldsymbol{N}}}$.

A Toy Language Model

A toy model of language — Santa Fe processes

Santa Fe processes are sequences $(X_t)_{t\in\mathbb{N}}$ of pairs

$$X_t = (K_t, Z_{K_t})$$

where $(K_t)_{t\in\mathbb{N}}$, called narration, is a sequence of natural numbers and $(Z_k)_{k\in\mathbb{N}}$, called knowledge, is a sequence of coin flips.

A semantic interpretation

Process $(X_t)_{t\in\mathbb{N}}$ is a sequence of propositions describing knowledge $(Z_k)_{k\in\mathbb{N}}$ at random but consistently:

- Proposition $X_t = (k, z)$ asserts that the k-th coin flip is z, in such way that one can determine both k and z.
- For $X_t = (k, z)$ and $X_s = (k', z')$ we do not know in advance which coin flips they describe but $k = k' \implies z = z'$.

Narration model — Multiperiodic sequences

A multiperiodic sequence:

$1, 2, 1, 3, 1, 4, 1, 2, 1, 5, 1, 6, 1, 2, 1, 3, 1, 7, 1, 2, 1, 8, 1, 4, 1, 2, 1, \ldots$

The rule of generation:

If we delete tokens < 1, type 1 appers every $\pi_1 = 2$ tokens.
If we delete tokens < 2, type 2 appers every $\pi_2 = 3$ tokens.
If we delete tokens < 3, type 3 appers every $\pi_3 = 4$ tokens.
...
If we delete tokens $< r$, type r appers every $\pi_r = r + 1$ tokens.

Multiperiodic sequences — The algorithm

Require: List $\pi[r] \in \mathbb{N}$ for $r \in \mathbb{N}$. ▷ periods
Require: List $\phi[r] = 1$ for $r \in \mathbb{N}$. ▷ clocks
Ensure: List $k[t] \in \mathbb{N}$ for $t \in \mathbb{N}$. ▷ multiperiodic sequence

for $t \in \mathbb{N}$ **do**
    $r_{\text{active}} := 0$
    $r_{\text{iter}} := 1$
    while $r_{\text{active}} = 0$ **do**
        if $\phi[r_{\text{iter}}] > 1$ **then**
            $\phi[r_{\text{iter}}] := \phi[r_{\text{iter}}] - 1$
        else
            $r_{\text{active}} := r_{\text{iter}}$
        $r_{\text{iter}} := r_{\text{iter}} + 1$
    $\phi[r_{\text{active}}] := \pi[r_{\text{active}}]$
    $k[t] := r_{\text{active}}$

Multiperiodic sequences — Relative frequency

The relative frequency of types $\geq r$:

$$
\begin{aligned}
f_r &:= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} 1\{k_t \geq r\} \\
&= \left(1 - \frac{1}{\pi_1}\right)\left(1 - \frac{1}{\pi_2}\right) \dots \left(1 - \frac{1}{\pi_{r-1}}\right)
\end{aligned}
$$

Example

Let $\pi_r \approx cr$ for some $c > 0$ and all $r \in \mathbb{N}$. We may estimate

$$
\begin{aligned}
f_r &\approx \exp \sum_{i=1}^{r-1} \log\left(1 - \frac{1}{ci}\right) \approx \exp \int_1^r \log\left(1 - \frac{1}{cx}\right) dx \\
&\approx \exp\left(-\int_1^r \frac{dx}{cx}\right) = \exp\left(-\frac{\log r}{c}\right) = r^{-1/c}.
\end{aligned}
$$

The waiting time and the number of types

The waiting time and the number of types:

$$w_r := \min \{t \in \mathbb{N} : k_t = r\} \geq r$$
$$n_t := \# \{k_1, k_2, ..., k_t\} = \max \{r \in \mathbb{N} : w_r \leq t\} \leq t$$

A sandwich bound:

$$\frac{1}{f_r} \leq w_r < \sum_{j=1}^{r} \frac{1}{f_j}$$

Example

Let $\pi_r \approx cr$ for some $c > 0$ and all $r \in \mathbb{N}$. We have

$$w_r \sim r^{(c+1)/c}, \qquad n_t \sim t^{c/(c+1)}.$$

Multiperiodic Santa Fe process — Model of learning

Environment:
A learning agent observes $(X_t)_{t\in\mathbb{N}}$ with $X_t = (k_t, Z_{k_t})$, where narration $(k_t)_{t\in\mathbb{N}}$ is a known multiperiodic sequence and knowledge $(Z_k)_{k\in\mathbb{N}}$ is a sequence of independent coin flips.

Goal:
The learning agent has to read first T data points X_1^T, then to compute N binary parameters $B_1^N = g_1(X_1^T; N)$, and finally to predict the remaining sequence as $\hat{X}_{T+i} = g_2(T+i; B_1^N)$.

Loss: We want to minimize the error rate

$$\mathcal{L}(N, T) := \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} 1\left\{X_{T+i} \neq \hat{X}_{T+i}\right\}.$$

Multiperiodic Santa Fe process — Optimal learner

Optimal parameters:

Parameters B_1^N should be chosen as the optimal estimators of coin flips Z_1^N. If token (r, Z_r) appears in data X_1^T, setting $B_r = Z_r$ can be actually carried out. If token (r, Z_r) does not appear in data X_1^T then we may put $B_r = 0$. In this way, we obtain

$$B_r = \begin{cases} Z_r, & r \le N \wedge n_T, \\ 0, & r > N \wedge n_T. \end{cases}$$

We apply notation $a \wedge b := \min \{a, b\}$ and $a \vee b := \max \{a, b\}$.

Optimal predictors:

The optimal predictors are

$$\hat{X}_{T+i} = (k_{T+i}, B_{k_{T+i}}).$$

Multiperiodic Santa Fe process — Error rate

Hence, the test loss is the relative frequency of $(Z_{k_{T+i}} \neq B_{k_{T+i}})$,

$$\mathcal{L}(N,T) = \lim_{I\to\infty} I^{-1} \sum_{i=1}^{I} 1\{Z_{k_{T+i}} \neq B_{k_{T+i}}\}.$$

Averaging over random knowledge $(Z_k)_{k\in\mathbb{N}}$, we derive

$$\begin{aligned}\mathbb{E}\,\mathcal{L}(N,T) &= \lim_{I\to\infty} I^{-1} \sum_{i=1}^{I} P(Z_{k_{T+i}} \neq B_{k_{T+i}}) \\ &= \frac{1}{2}\lim_{I\to\infty} I^{-1} \sum_{i=1}^{I} 1\{k_{T+i} > N \wedge n_T\} = \frac{f_{N\wedge n_T}}{2}.\end{aligned}$$

Example

Let $\pi_r \approx cr$. We have $f_r \sim r^{-1/c}$ and $n_t \sim t^{c/(c+1)}$. Hence

$$\mathbb{E}\,\mathcal{L}(N,T) \sim \left[N \wedge T^{c/(c+1)}\right]^{-1/c} = \frac{1}{N^{1/c}} \vee \frac{1}{T^{1/(c+1)}}.$$

Is This Relevant?

Santa Fe decomposition

- When we read a text in natural language, we may feel that it consists of contiguous propositions describing discrete facts.
- Since there are only countably many distinct propositions $\boldsymbol{x_t}$ and countably many distinct mentioned facts $\boldsymbol{b_k}$, we may enumerate them by natural numbers and arrive at a representation of individual propositions $\boldsymbol{x_t} = (\boldsymbol{k_t}, \boldsymbol{b_t})$ that resembles Santa Fe decomposition $\boldsymbol{x_t} = (\boldsymbol{k_t}, \boldsymbol{z_{k_t}})$.
- Two delicate questions are:

— Can decompositions $(\boldsymbol{k_t}, \boldsymbol{b_t})$ be effectively computed?

— Does $\boldsymbol{k_t} = \boldsymbol{k_{t'}}$ imply $\boldsymbol{b_t} = \boldsymbol{b_{t'}}$?

Only then we may define immutable facts $\boldsymbol{z_r} := \boldsymbol{b_t}$ for $\boldsymbol{k_t} = \boldsymbol{r}$.

- But even if $\boldsymbol{k_t} = \boldsymbol{k_{t'}}$ implies $\boldsymbol{b_t} = \boldsymbol{b_{t'}}$ only for time indices $\boldsymbol{t}$ and $\boldsymbol{t'}$ that are close enough then the text still exhibits some properties of the Santa Fe process.

Conditional determinism of narration

- The Santa Fe decomposition posits that text $(x_t)_{t\in\mathbb{N}}$ is a composition of knowledge $(z_k)_{k\in\mathbb{N}}$ and narration $(k_t)_{t\in\mathbb{N}}$.
- Is there a good reason to suspect that the narration is deterministic given the knowledge and resembles the multiperiodic process?
- Determinism of narration is equivalent to zero entropy rate and, as everyone knows, Shannon (1951) showed that the entropy rate of natural language is 1 bit per letter.
- There have been researchers like Hilberg (1990), looking at the same data and claiming the zero entropy rate.
- The stake is high and it is better to stay cautious.

Tampering with the multiperiodic algorithm

- The multiperiodic algorithm seems an interesting model for combining determinism and randomness in narration.
- We may tamper with clocks ϕ_r, set them at random values, reset them with certain probabilities, introduce correlations.
- All of this can make the output sequence $(k_t)_{t\in\mathbb{N}}$ more similar to the rhythm of daily chores or human utterances:
  - — there may be cycles of varying time scales,
  - — there may be repetitions,
  - — there may be hierarchical structures,
  - — there may be bursts and lulls,
  - — there may be some residual randomness.
- The open problem seems to uncover the true dynamics of clocks ϕ_r. Is it a more transparent approach to artificial intelligence than transformers?

Further reading

M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. https://arxiv.org/abs/2105.14368, 2021.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. Wiley & Sons, 2021.

Ł. Dębowski. A simplistic model of neural scaling laws: Multiperiodic Santa Fe processes. https://arxiv.org/abs/2302.09049, 2023.

M. Hutter. Learning curve theory. https://arxiv.org/abs/2102.04074, 2021.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. https://arxiv.org/abs/2001.08361, 2020.

A. Maloney, D. A. Roberts, and J. Sully. A solvable model of neural scaling laws. https://arxiv.org/abs/2210.16859, 2022.

D. A. Roberts, S. Yaida, and B. Hanin. *The Principles of Deep Learning Theory*. Cambridge University Press, 2022.

G. West. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. New York: Penguin Press, 2017.