NeuroMod Annual Meeting
Université Côte d'Azur
Antibes, France

What are Neural Language Models the Model of?

Epistemological and Theoretical Perspectives on LLMs

Juan Luis Gastaldi

www.giannigastaldi.com

ETH zürich

July 8, 2025

Outline

Introduction

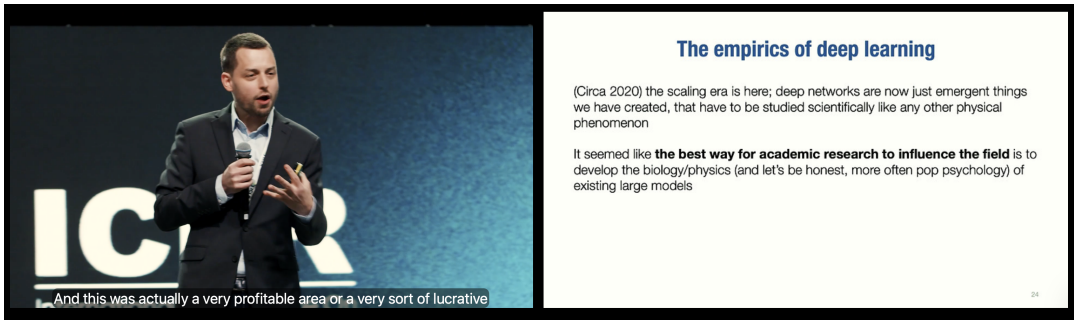
Epistemological Perspectives

Theoretical Perspectives
- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Take Aways

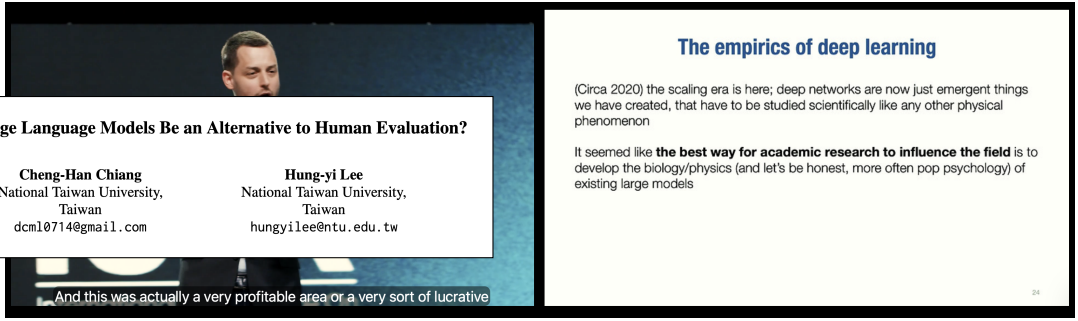
Outline

Empirical Saturnalia



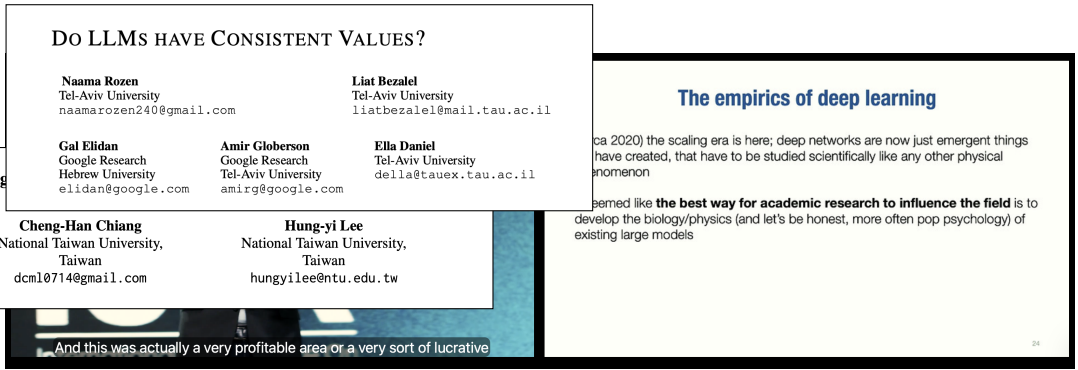
Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia



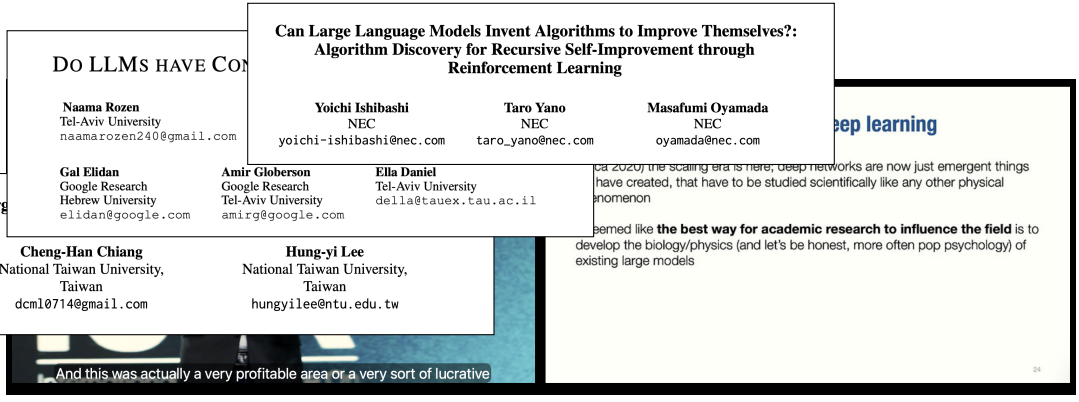
Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia



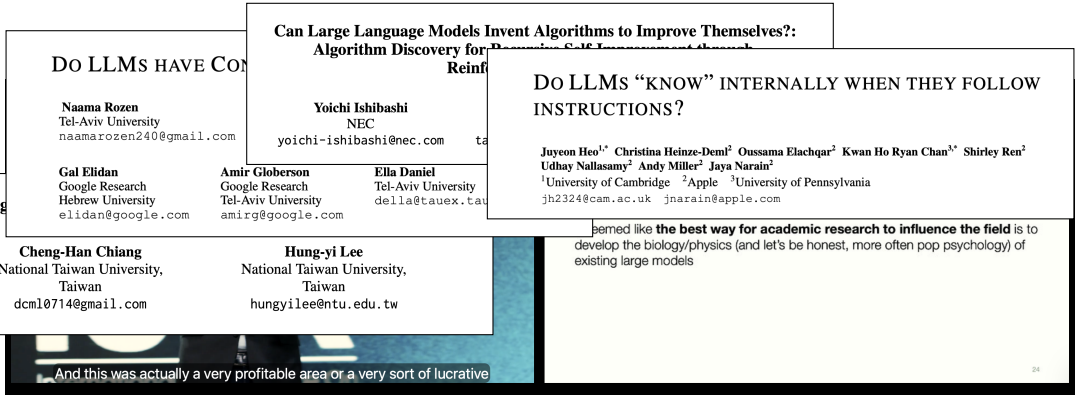
Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia



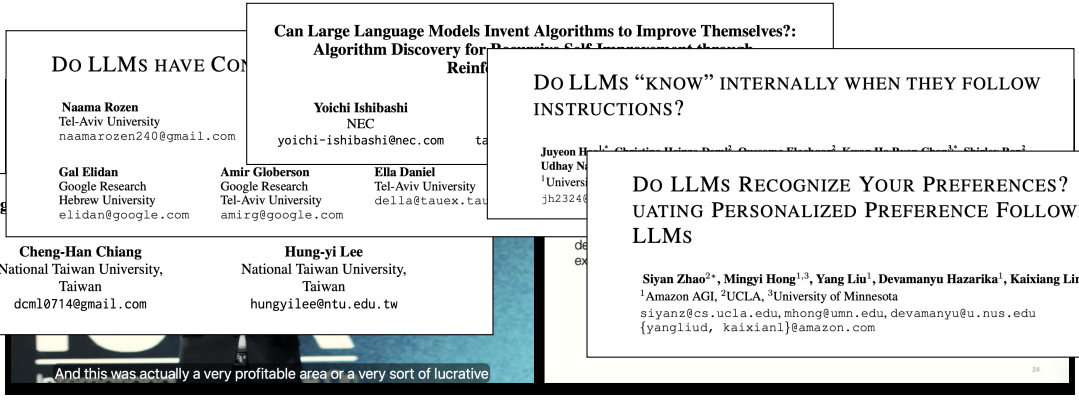
Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia



Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia



Zico Kolter, Building Safe and Robust AI Systems, Keynote at ICLR 2025.

Empirical Saturnalia

DO LLMS HAVE CON

Naama Rozen
Tel-Aviv University
naamarozen240@gmail.com

Gal Elidan
Google Research
Hebrew University
elidan@google.com

Amir Globerson
Google Research
Tel-Aviv University
amirg@google.com

Ella Daniel
Tel-Aviv University
della@tauex.tau

Can Larg

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Reinf

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com

DO LLMS "KNOW" INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Udhay Na
jh2324@

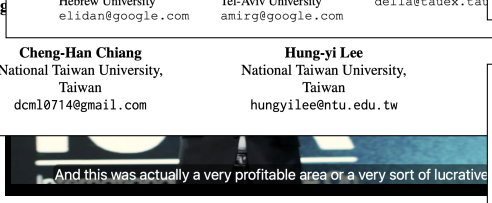
DO LLMS RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMS

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

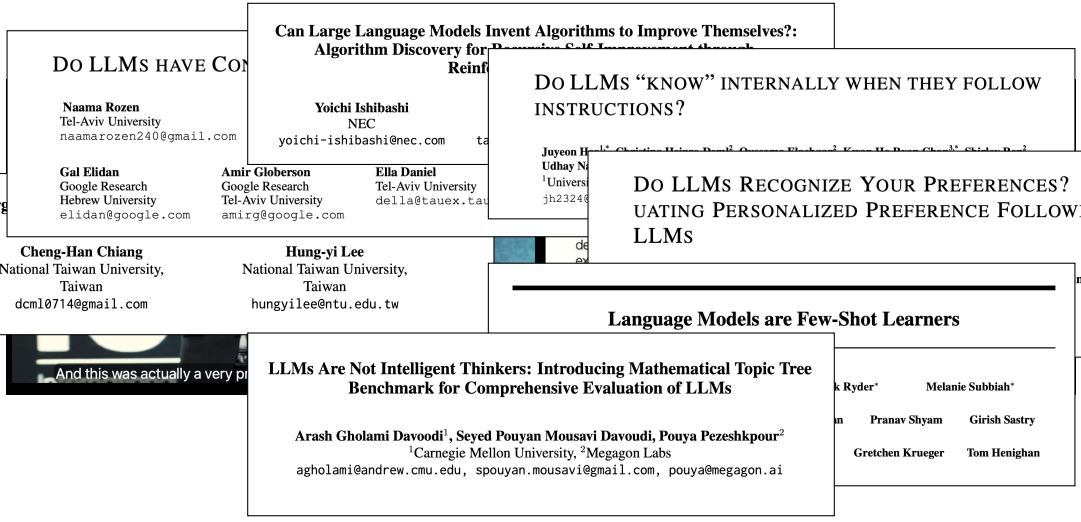
Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

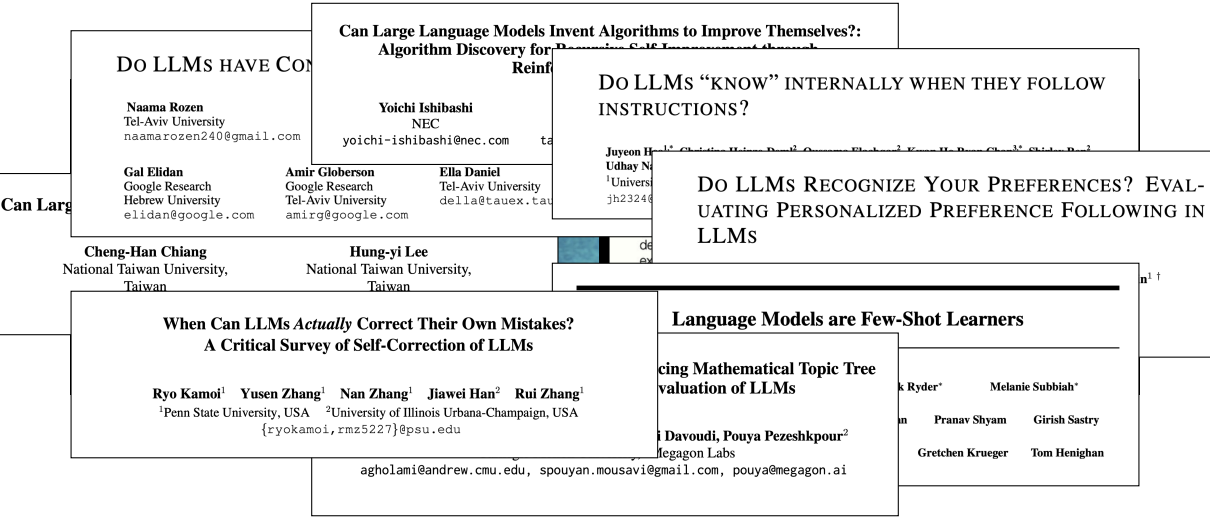


Zico Kolter, *Building Safe and*

Empirical Saturnalia



Empirical Saturnalia



Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves?: Algorithm Discovery for Recursive Self-Improvement through Reinf

Yoichi Ishibashi
NEC

DO LLMS HAVE CON

Naama Rozen
Tel-Aviv University

DO LLMS "KNOW" INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Can Larg

Nat

DO LLMS RECOGNIZE YOUR PREFERENCES? EVALUATING PERSONALIZED PREFERENCE FOLLOWING IN LLMS

n[1, †]

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Language Models are Few-Shot Learners

ng Mathematical Topic Tree
valuation of LLMs

k Ryder* Melanie Subbiah*

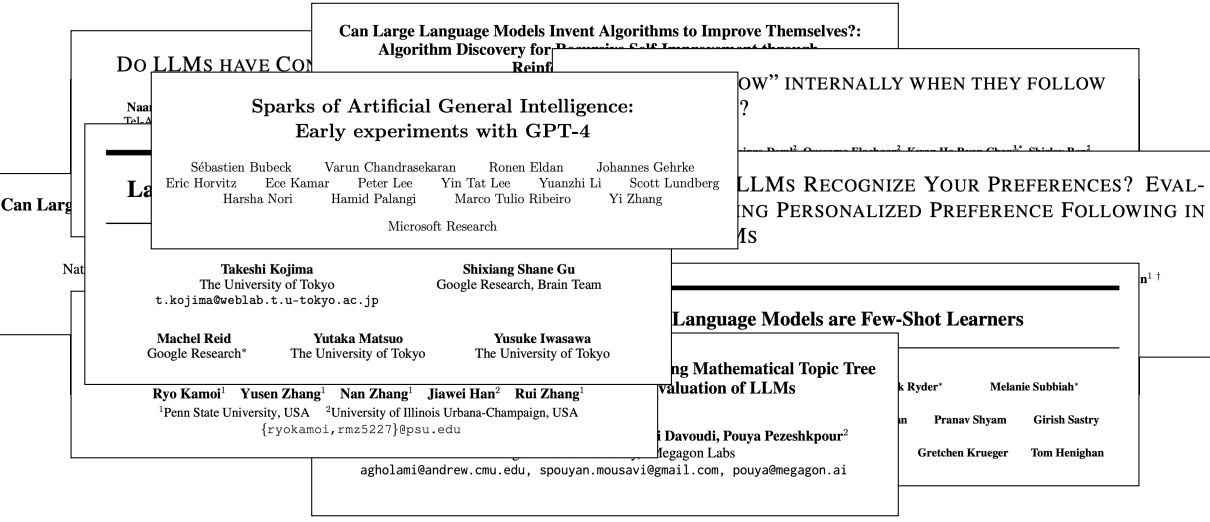
n Pranav Shyam Girish Sastry

Gretchen Krueger Tom Henighan

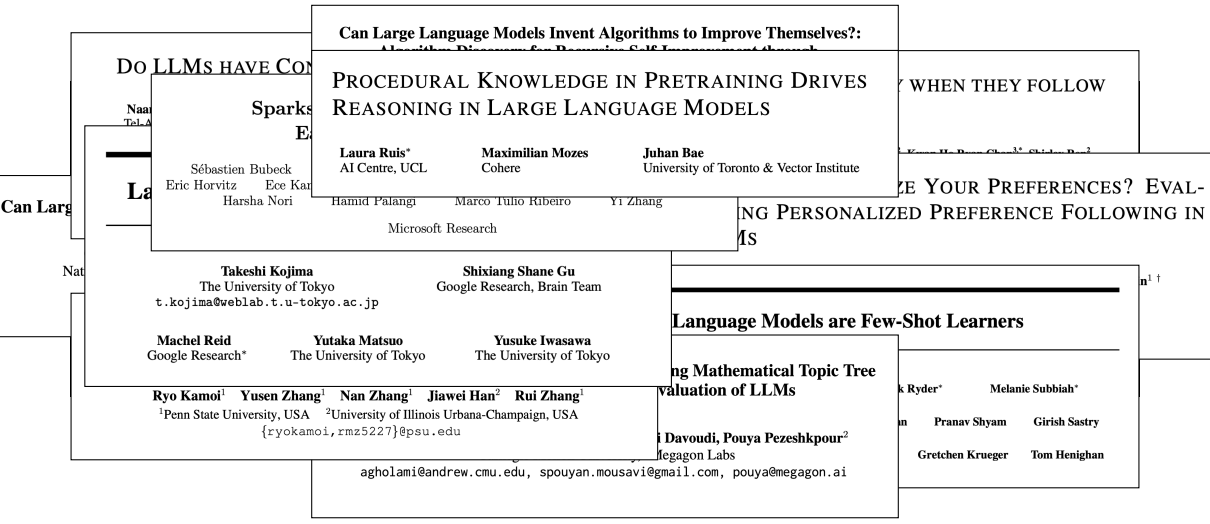
i Davoudi, Pouya Pezeshkpour[2]
Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Ryo Kamoi[1] **Yusen Zhang**[1] **Nan Zhang**[1] **Jiawei Han**[2] **Rui Zhang**[1]
[1]Penn State University, USA [2]University of Illinois Urbana-Champaign, USA
{ryokamoi,rmz5227}@psu.edu

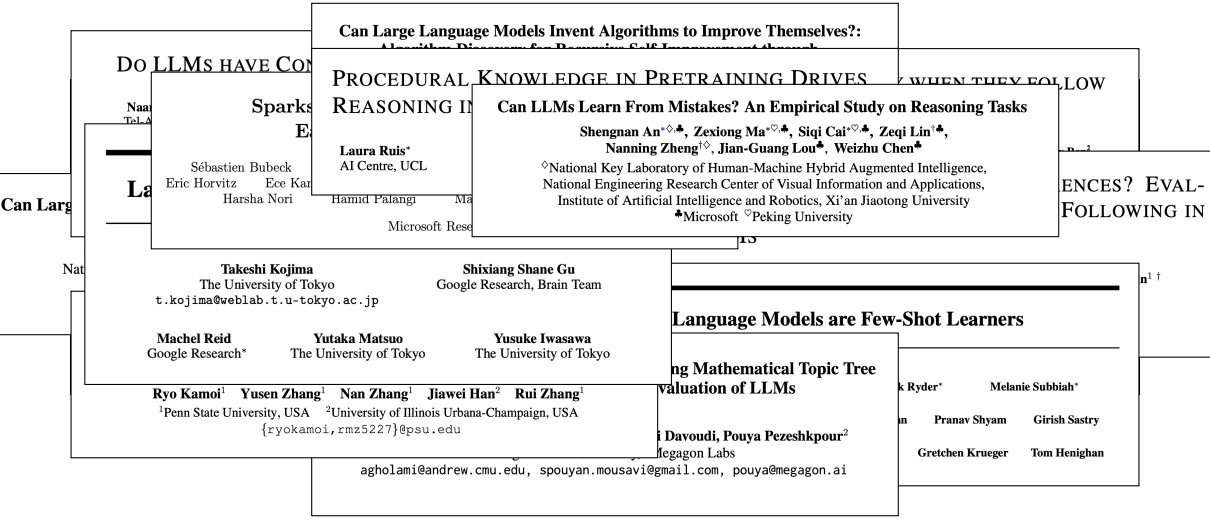
Empirical Saturnalia



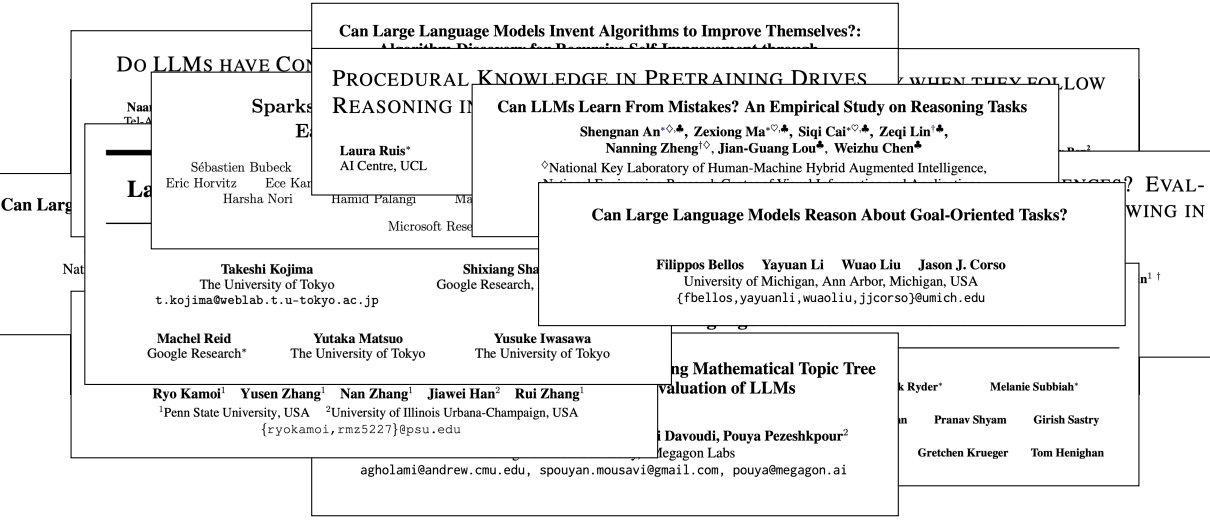
Empirical Saturnalia



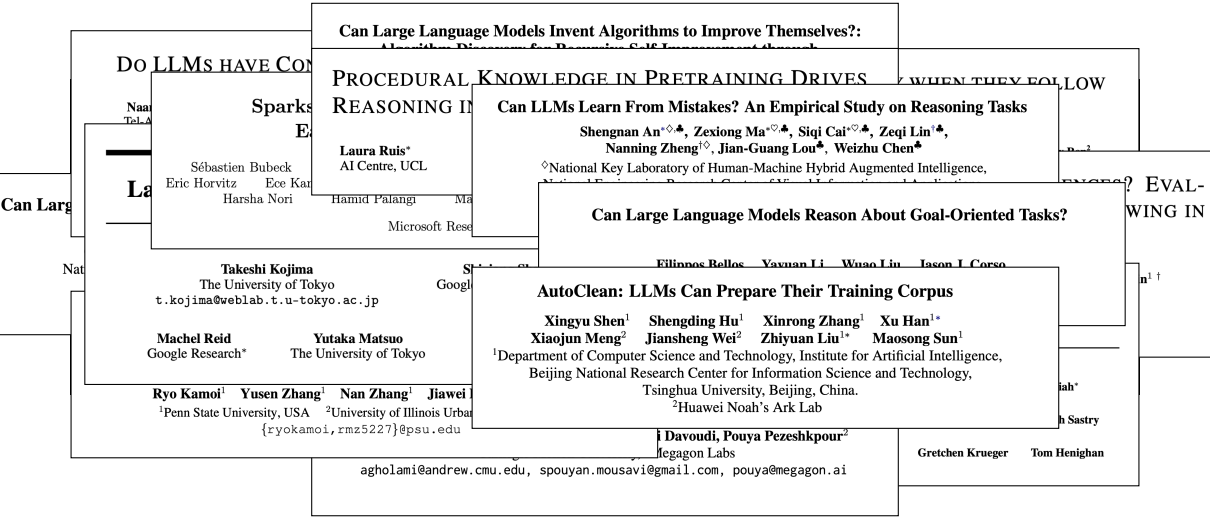
Empirical Saturnalia



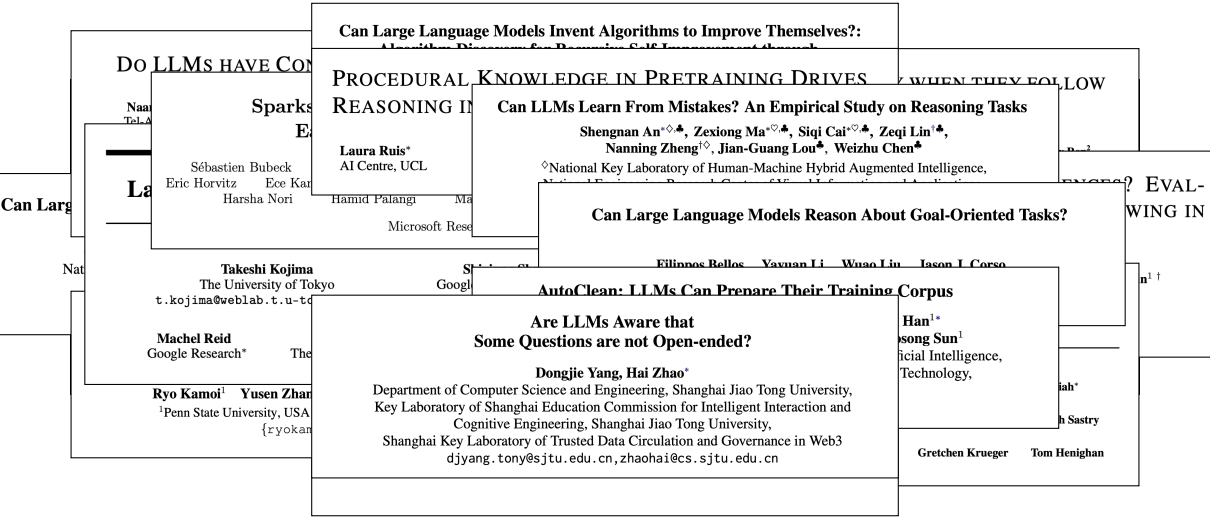
Empirical Saturnalia



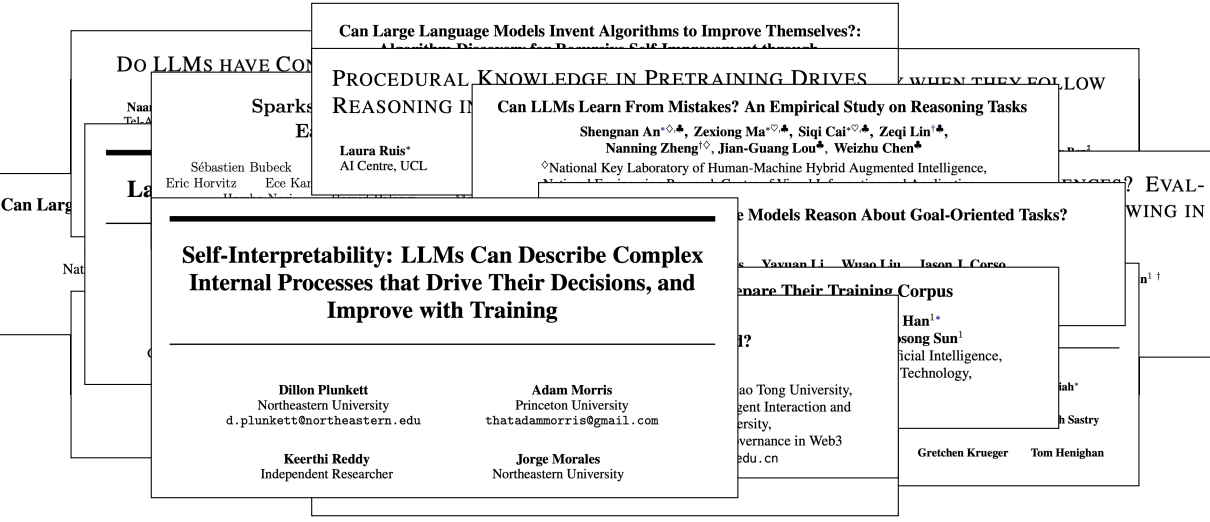
Empirical Saturnalia



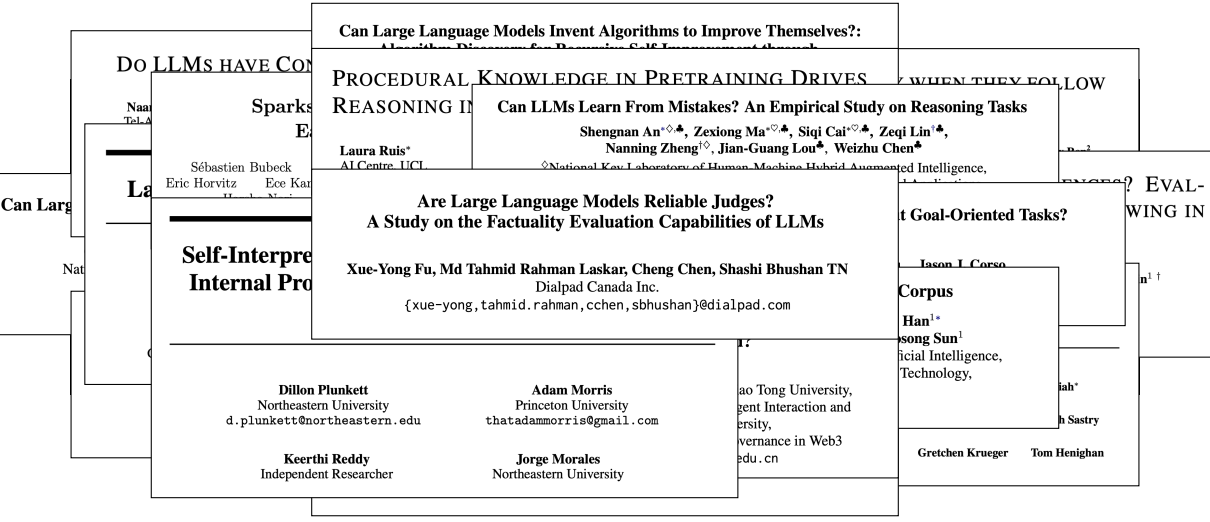
Empirical Saturnalia



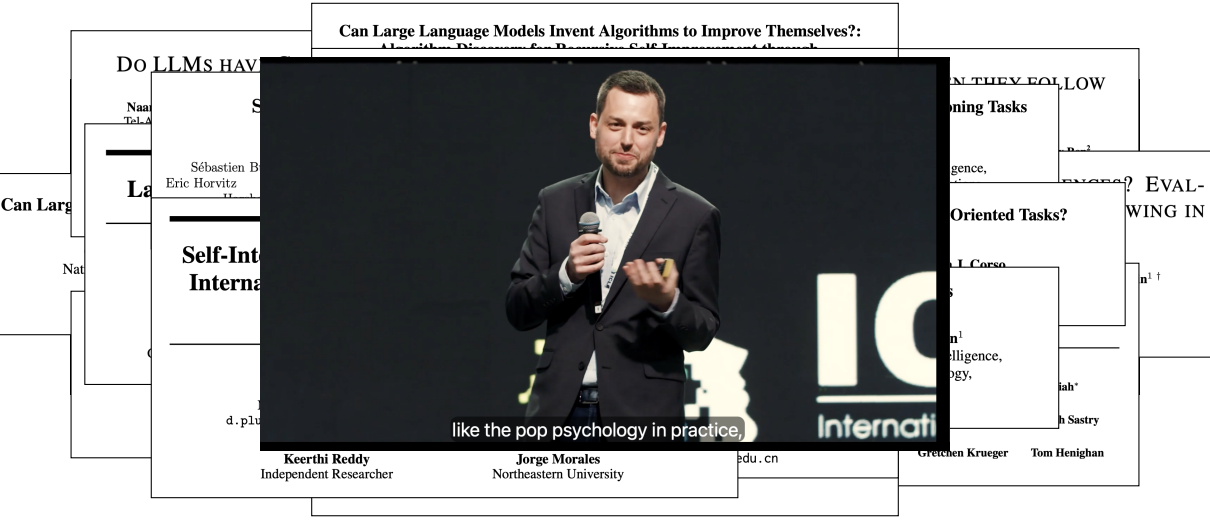
Empirical Saturnalia



Empirical Saturnalia

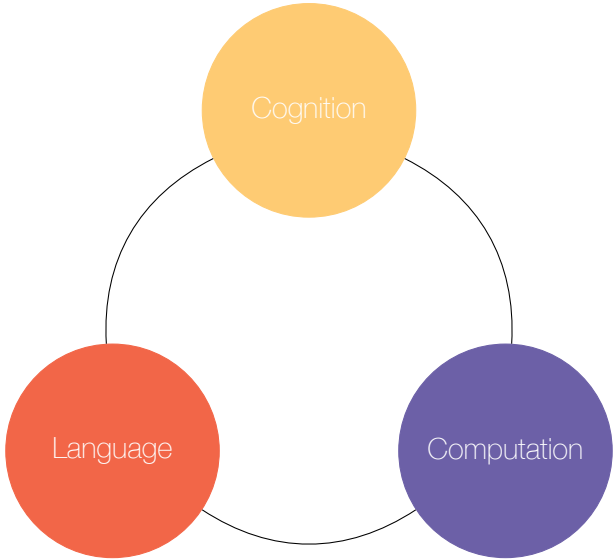


Empirical Saturnalia



Outline

Chomsky's Generativist Program and the Cognitive Revolution



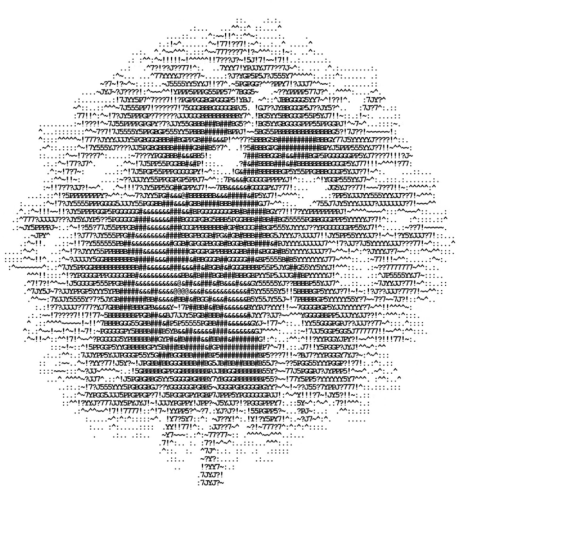
The New York Times

OPINION

GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

March 8, 2023



Chomsky against Abstraction in Principle

"Pick the properties that you like for a set of processors. Pick the criteria you like for success, whether in terms of performance or structure or whatever. Consider the class of all organisms, *abstracting in principle* from the existing world, that satisfy those things. And then you can ask whether they have some property of things in the material world. Do they breathe? Do they grow? Do they think? Do they talk? Do they walk? Do they enjoy themselves? Do they have moral rights?"

(Chomsky, 1992)

Chomsky against Abstraction in Principle

"All of these questions are stupid. And the reason they're stupid is because you've departed from naturalism. Once you've departed from naturalism, you have an algorithm for constructing stupid questions."

(Chomsky, 1992)

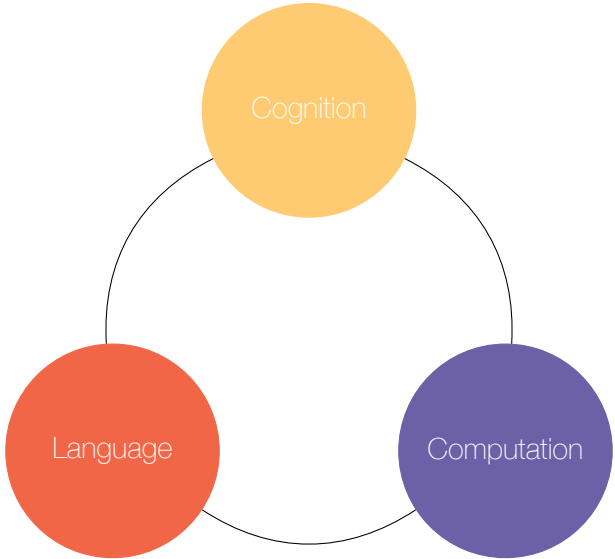
Chomsky against Abstraction in Principle

"There's nothing wrong with principled abstraction. In fact, one might think of large areas of mathematics as that. But here we have something new, principled abstraction in an empirical discipline."

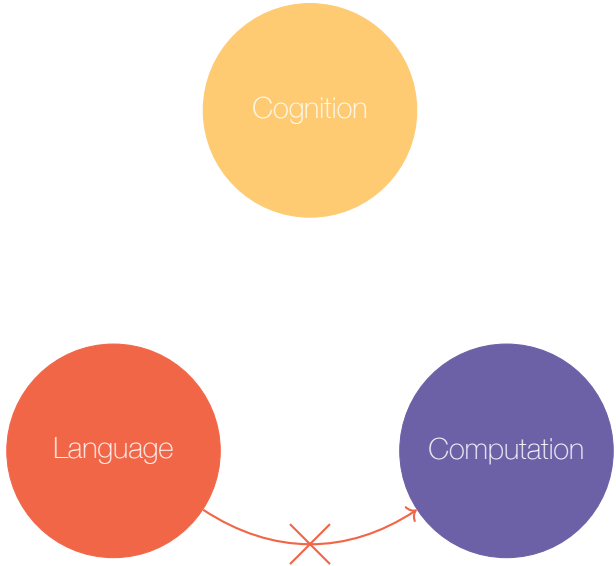
"I don't think we should cross that border, because there's no empirical claim. It is just a question of how to extend the metaphor."

(Chomsky, 1992)

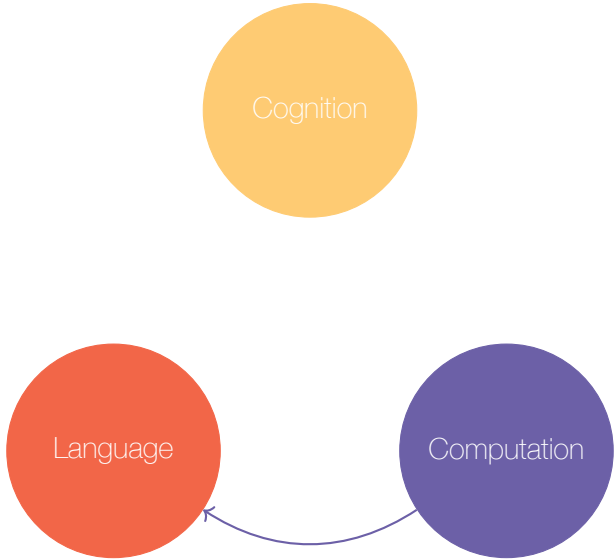
The Condition of Chomsky's Cognitive Foundations



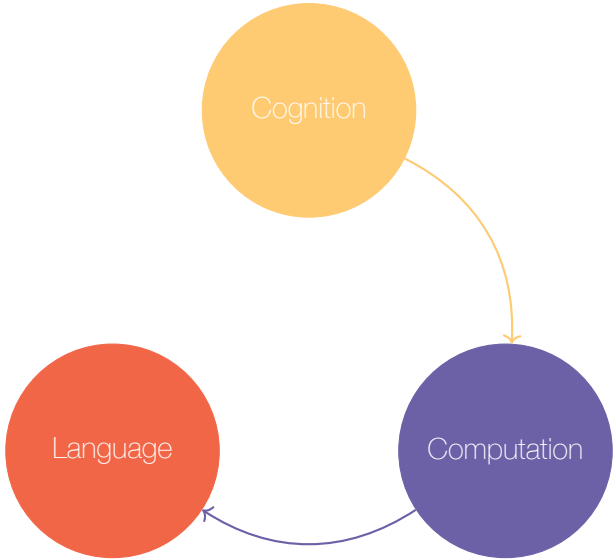
The Condition of Chomsky's Cognitive Foundations



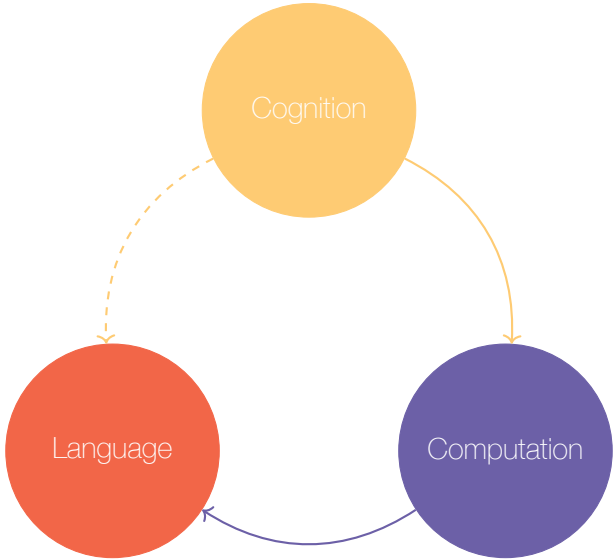
The Condition of Chomsky's Cognitive Foundations



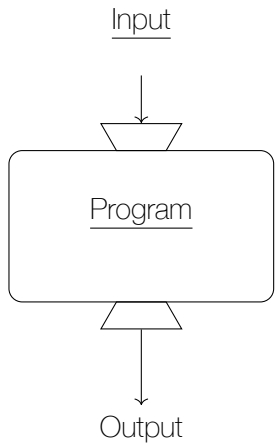
The Condition of Chomsky's Cognitive Foundations



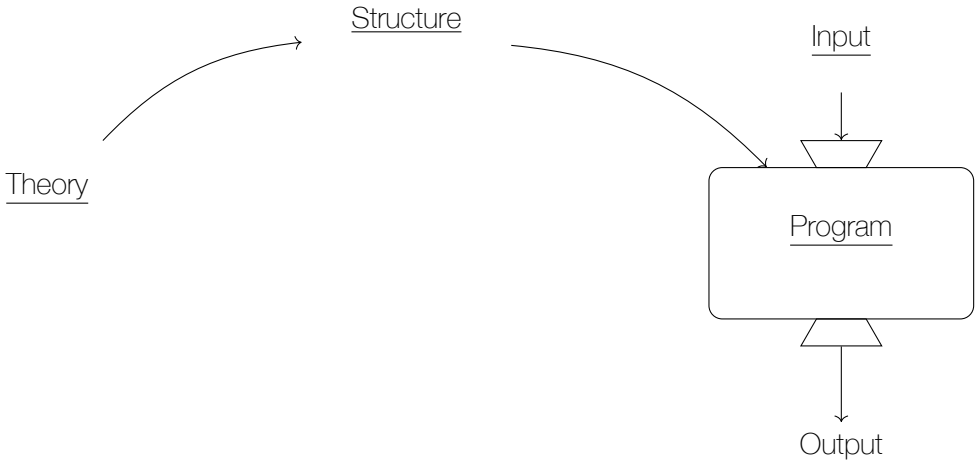
The Condition of Chomsky's Cognitive Foundations



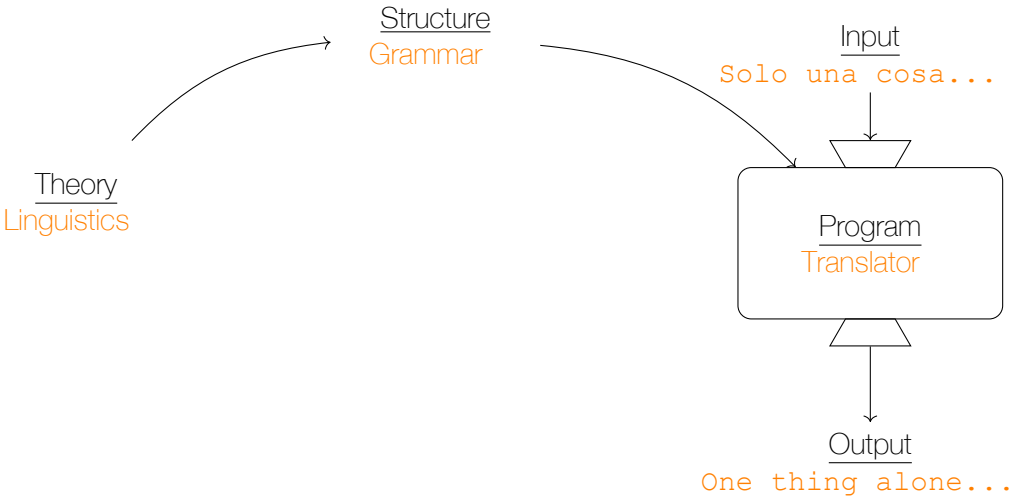
The Trap



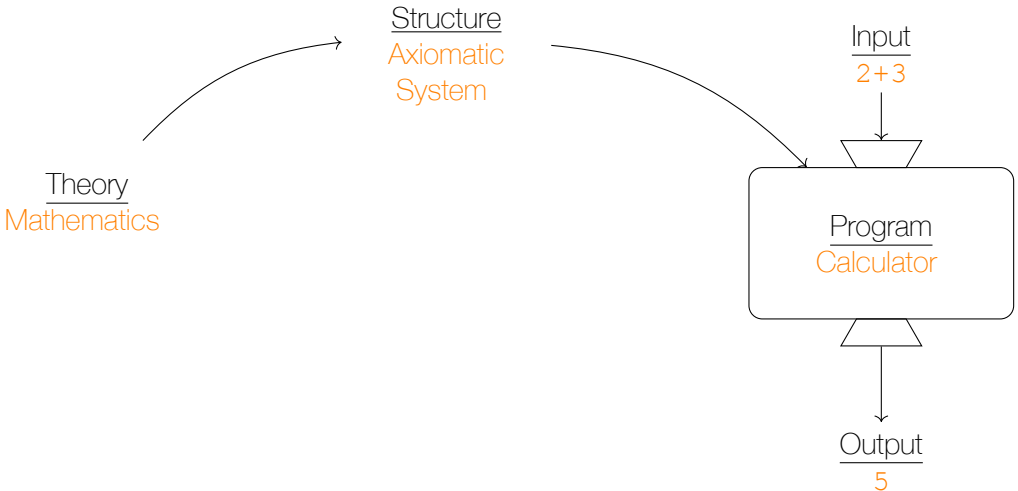
The Trap



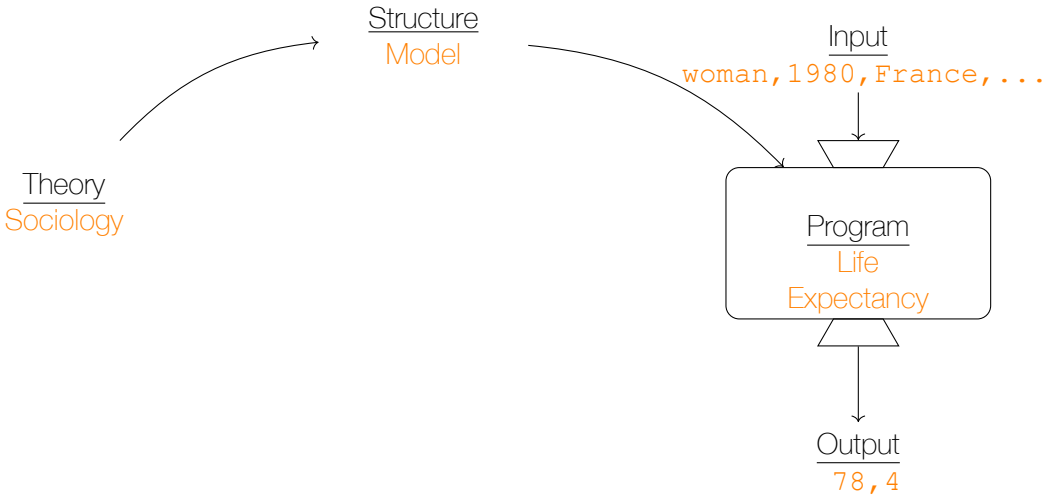
The Trap


Structure
Grammar
Input
Solo una cosa...
Theory
Linguistics
Program
Translator
Output
One thing alone...

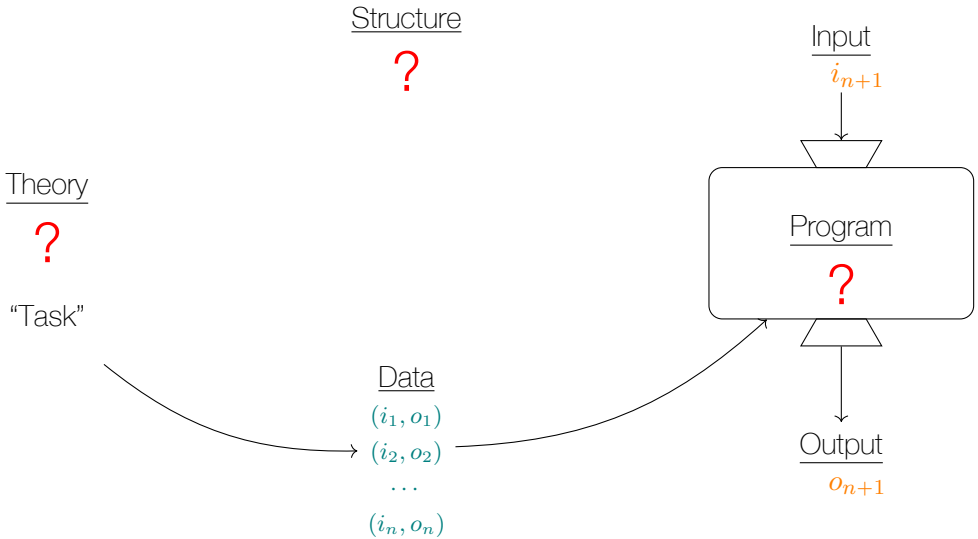
The Trap


Theory
Mathematics
Structure
Axiomatic
System
Input
2+3
Program
Calculator
Output
5

The Trap

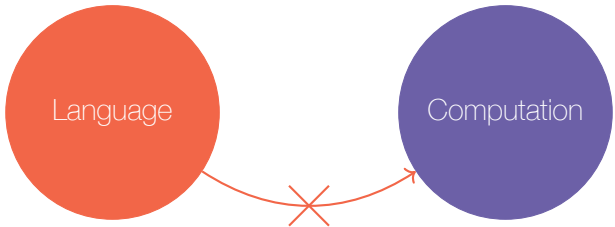

Structure
Model
Input
woman,1980,France,...
Theory
Sociology
Program
Life
Expectancy
Output
78,4

The Trap


Structure
?
Input
i_{n+1}
Theory
?
"Task"
Program
?
Data
(i_1, o_1)
(i_2, o_2)
...
(i_n, o_n)
Output
o_{n+1}

Necessary Condition?

- Inadequacy of distributional models (Chomsky, 1953)
- Limited expressive power of FSAs (Chomsky, 1956)
- The probability of a sentence is useless (Chomsky, 1957, 1959)
- Poverty of stimulus (Chomsky, 1959)



Necessary Condition?

- Inadequacy of distributional models (Chomsky, 1953)

  Inconclusive

- The probability of a sentence is useless (Chomsky, 1957, 1959)

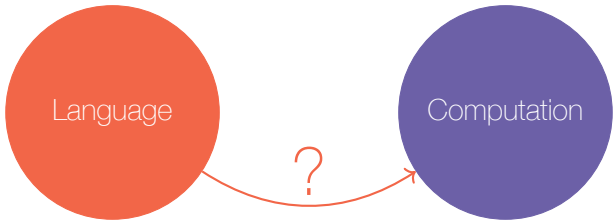
  Empirically challenged

- Limited expressive power of FSAs (Chomsky, 1956)

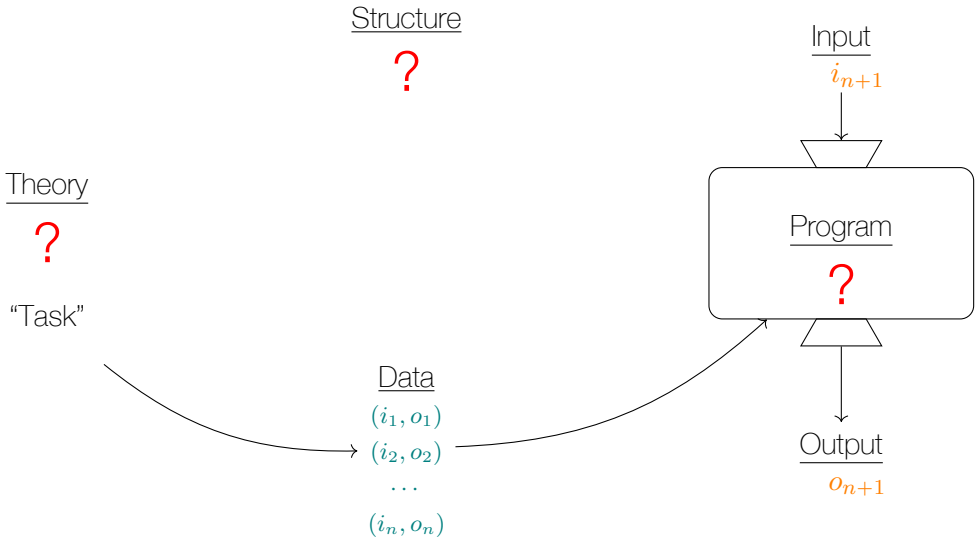
  The relevance is unclear

- Poverty of stimulus (Chomsky, 1959)

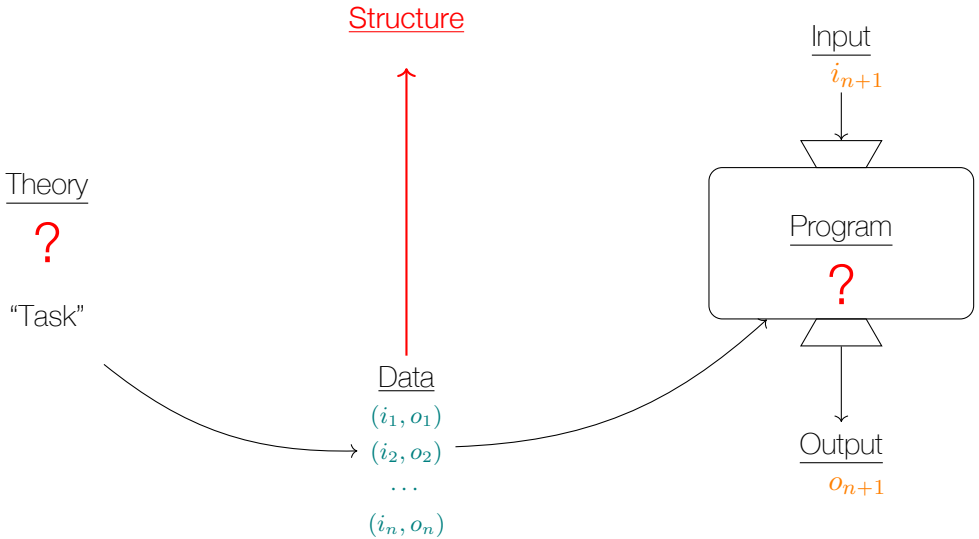
  Assumes what is to be proved



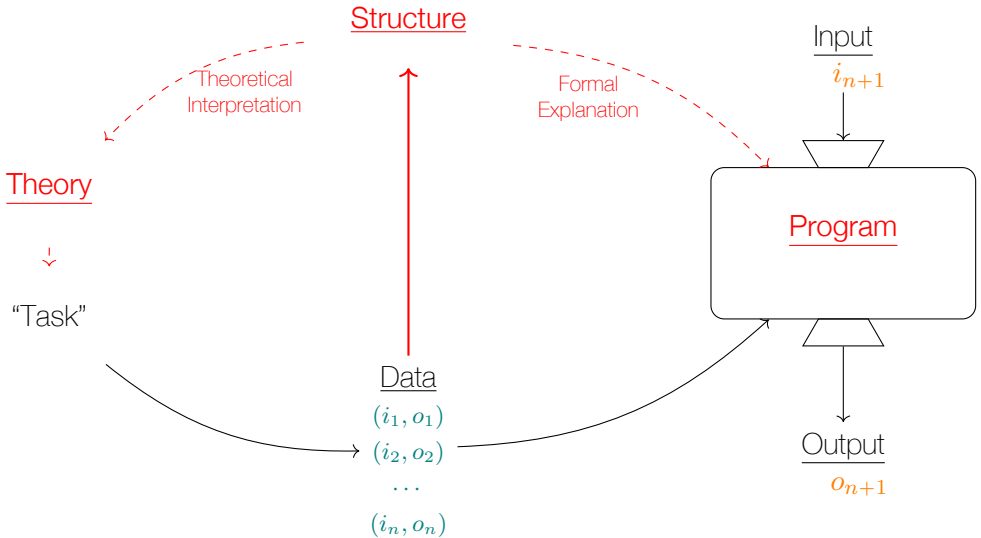
Making It Explicit



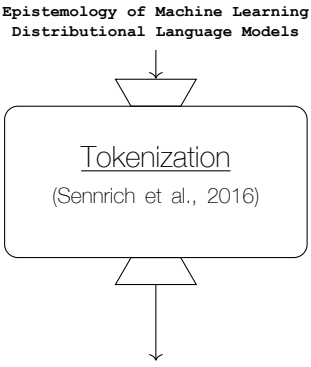
Making It Explicit



Making It Explicit

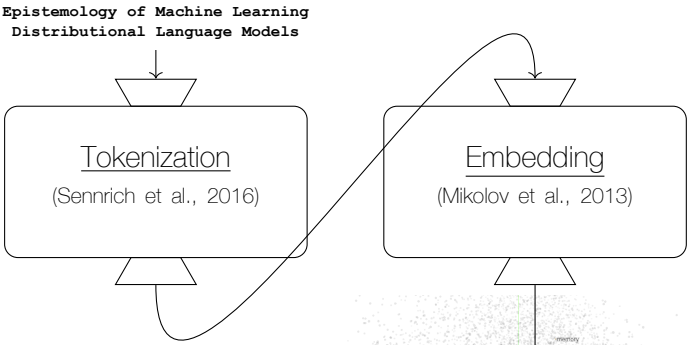


Outline



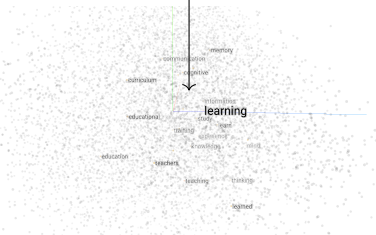
(https://tiktokenizer.vercel.app)

Formal Explainability



Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)



(https://projector.tensorflow.org)

Formal Explainability

Epistemology of Machine Learning
Distributional Language Models

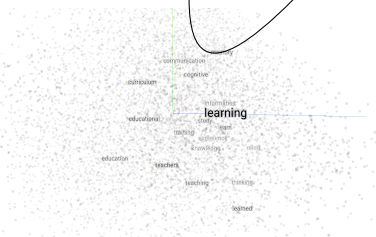
Tokenization
(Sennrich et al., 2016)

Embedding
(Mikolov et al., 2013)

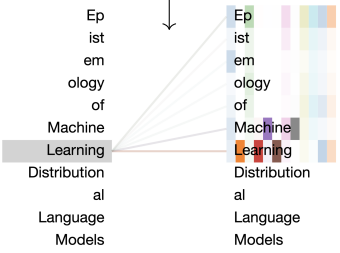
Attention
(Vaswani et al., 2017)

Epistemology of Machine Learning
Distributional Language Models

(https://tiktokenizer.vercel.app)

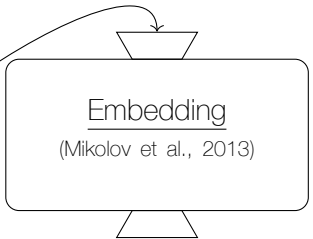


(https://projector.tensorflow.org)



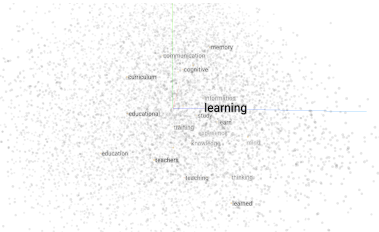
(https://github.com/jessevig/bertviz)

Formal Explainability



Epistemology of Machine Learning
Distributional Language Models

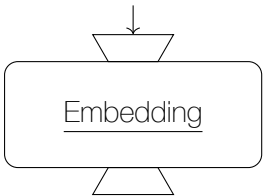
(https://tiktokenizer.vercel.app)

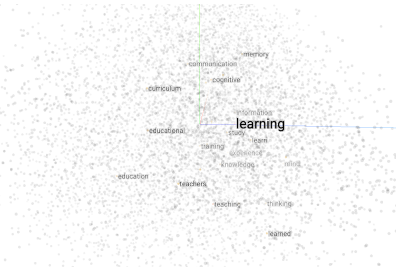


(https://projector.tensorflow.org)

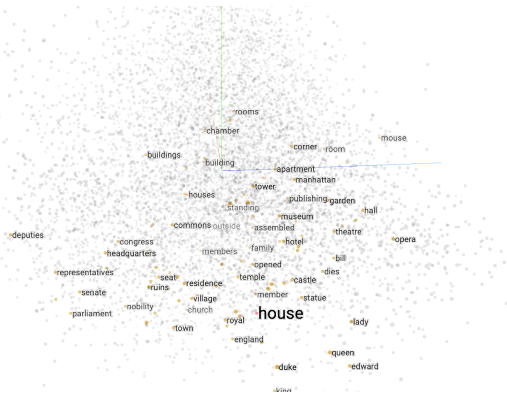
The Structure of Embeddings

Epistemology of Machine Learning
Distributional Language Models

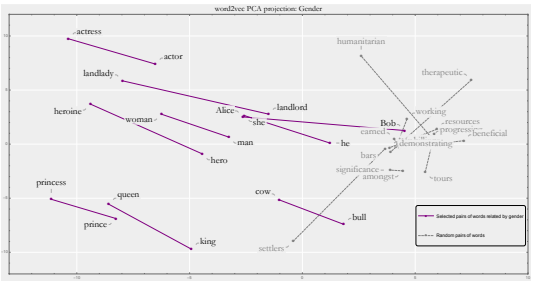




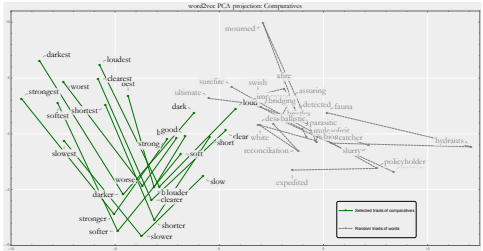
Embeddings: Similarity and Analogy

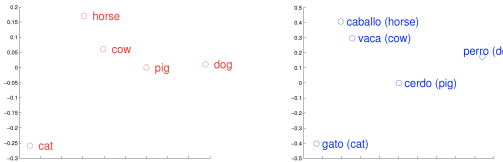


(https://projector.tensorflow.org)

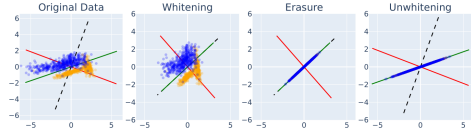


Embeddings: Other Applications

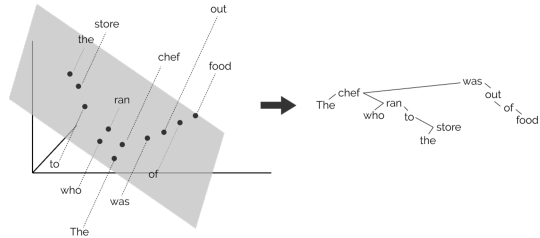




(Mikolov, Sutskever, Chen, Corrado, Dean, et al., 2013)



(Belrose et al., 2024)



(https://nlp.stanford.edu/~johnhew/structural-probe.html)

The Structure of Embeddings

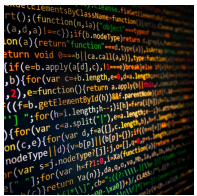
Structure

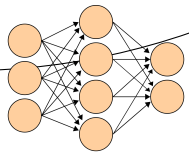
?

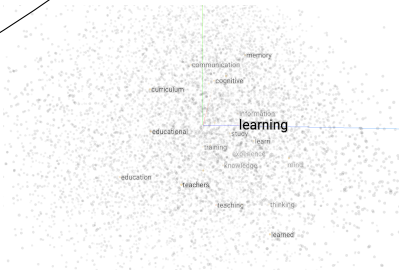
Epistemology of Machine Learning
Distributional Language Models

Embedding

Data







The Structure of Embeddings

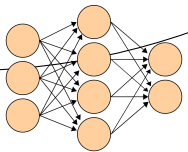
Structure

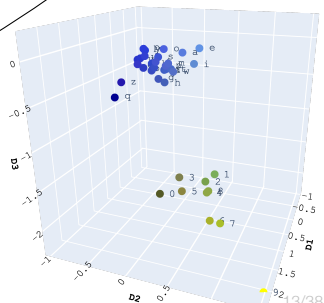
?

{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data





Outline

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.

word2vec Explained

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

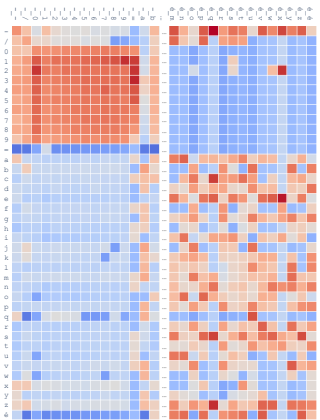
- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.
- The Singular Value Decomposition (SVD) provides an exact solution to this problem.

Example: Characters in Wikipedia

$$W = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

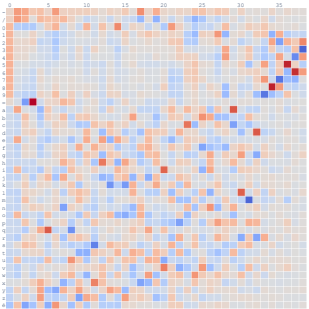
$$C = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

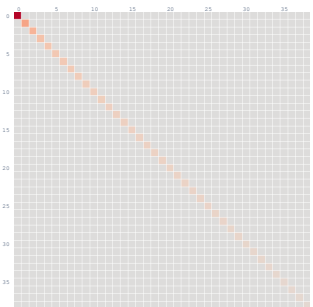
$$\begin{aligned} M_{wc} &= \text{pmi}(w, c) \\ &= \log \frac{p(w, c)}{p(w)p(c)} \end{aligned}$$

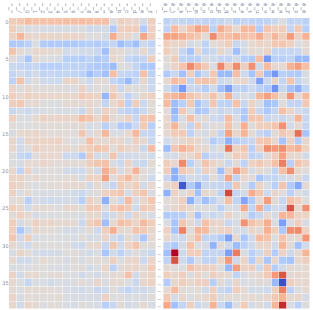


SVD of Wikipedia Character PMI Matrix

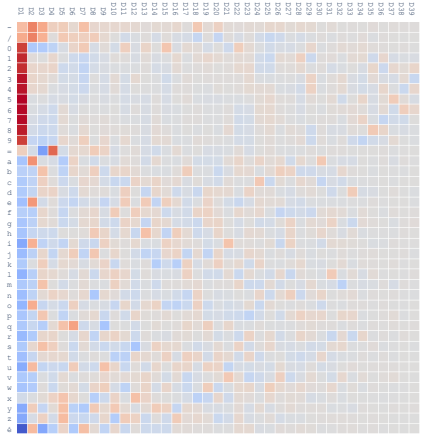
U Σ V^T







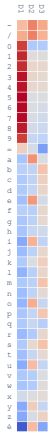
$$U \times \Sigma$$

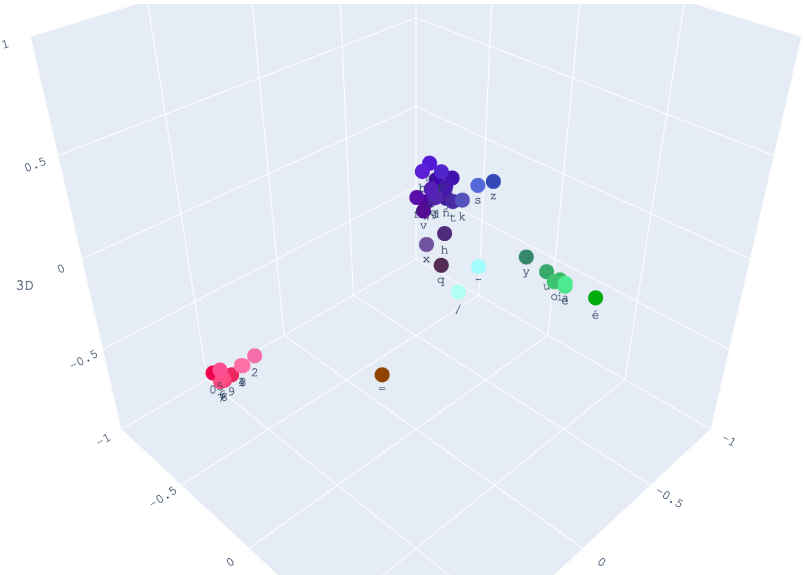


$$\hat{U} \times \hat{\Sigma}$$

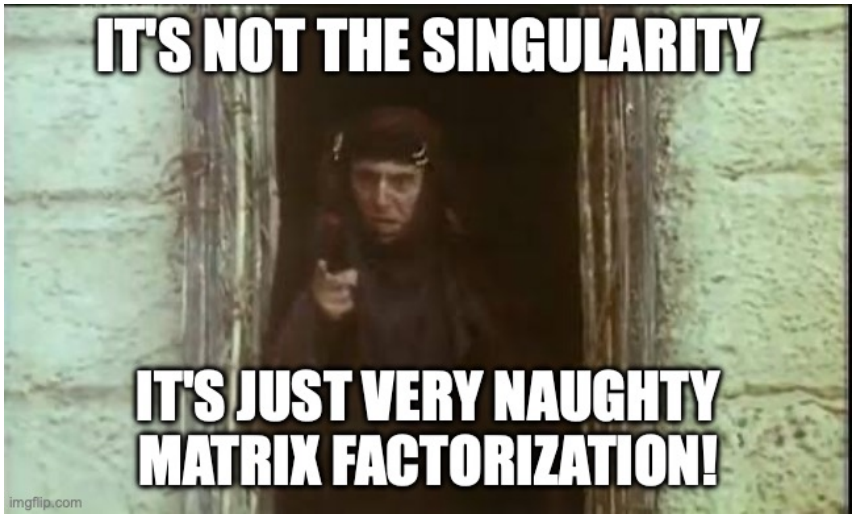
Plot

$$\hat{U} \times \hat{\Sigma}$$





What to conclude?



The Structure of Embeddings

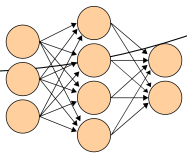
Structure

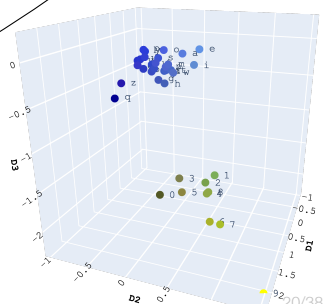
?

{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data





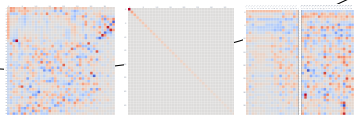
The Structure of Embeddings

Structure

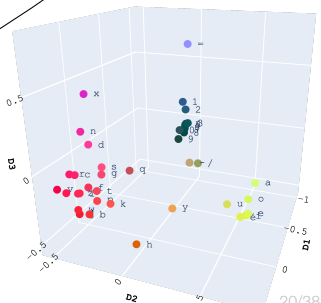
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data



SVD



4 Why does this produce good word representations?

Good question. We don't really know.

The distributional hypothesis states that words in similar contexts have similar meanings. The objective above clearly tries to increase the quantity $v_w \cdot v_c$ for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy.

Can we make this intuition more precise? We'd really like to see something more formal.

(Goldberg and Levy, 2014)

Outline

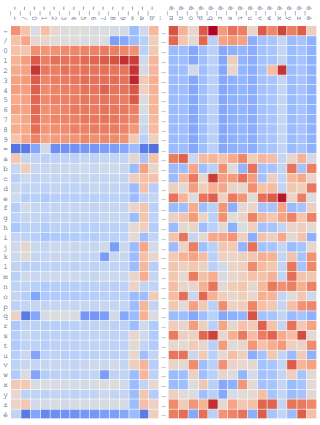
Embeddings as Functions Over Sets

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \mathrm{pmi}(x, y)$$



Embeddings as Functions Over Sets

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

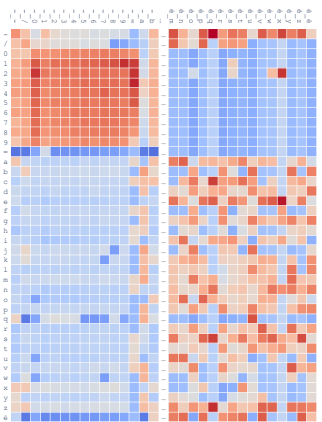
$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \operatorname{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$

$$x \mapsto M(x, -)$$



Embeddings as Functions Over Sets

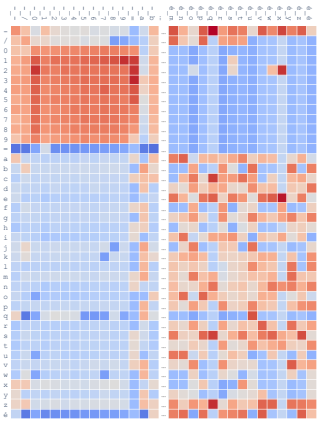
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \operatorname{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



Embeddings as Functions Over Sets

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$

$$X \xrightarrow{M_x} \mathbb{R}^Y$$

$$\mathbb{R}^X \xleftarrow[M_y]{} Y$$

Embeddings as Functions Over Sets

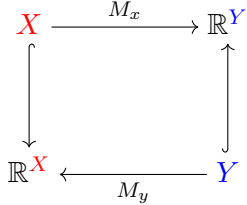
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



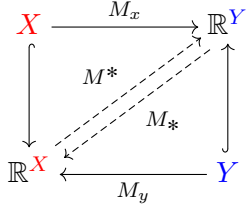
Embeddings as Functions Over Sets

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$
$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

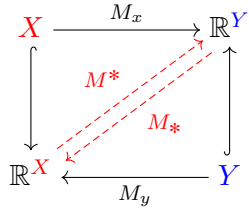
$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



$$M^* \colon \mathbb{R}^X \to \mathbb{R}^Y$$
$$M_* \colon \mathbb{R}^Y \to \mathbb{R}^X$$

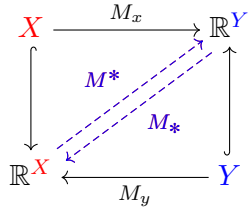
Embeddings as Functions Over Sets

$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$



Embeddings as Functions Over Sets

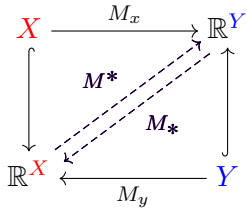
$$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$$



Embeddings as Functions Over Sets

$$M_* M^* \colon \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y$$

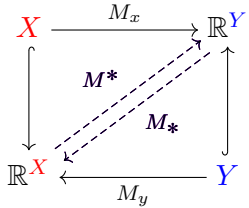
$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$



Embeddings as Functions Over Sets

$$M_* M^* \colon \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y$$

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$



$$M = U \Sigma V^T$$

$$U := [u_1, \ldots, u_m]$$
$$V := [v_1, \ldots, v_n]$$
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

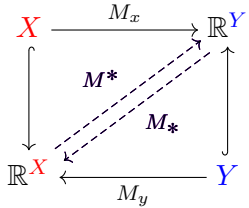
Embeddings as Functions Over Sets

$$M_* M^* \colon \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y$$

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$

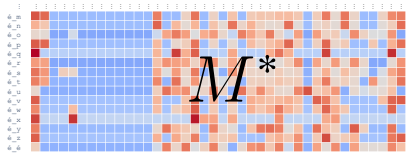
$$M_* M^* u_i = \lambda_i u_i$$
$$M^* M_* v_i = \lambda_i v_i$$

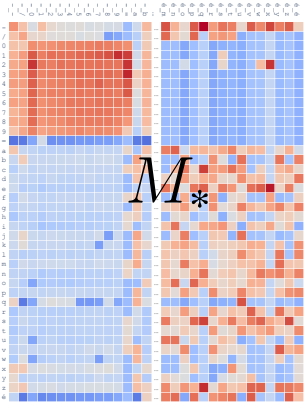
The u_i and v_i are (linear) fixed points!

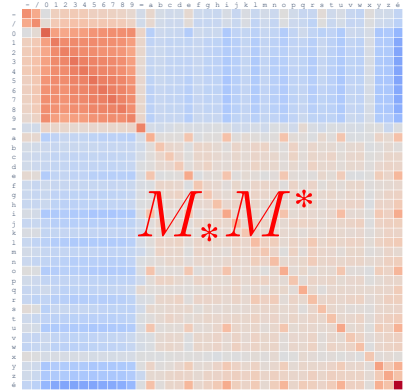


$$M = U \Sigma V^T \qquad U := [u_1, \ldots, u_m] \qquad V := [v_1, \ldots, v_n] \qquad \Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

$M_* M^*$ as a Covariance Matrix

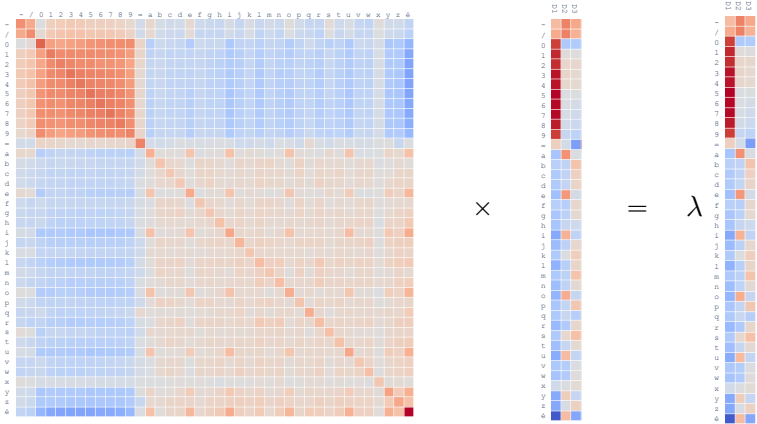






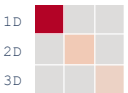
Eigenvectors as Fixed Points

$$M_* M^* u = \lambda u$$

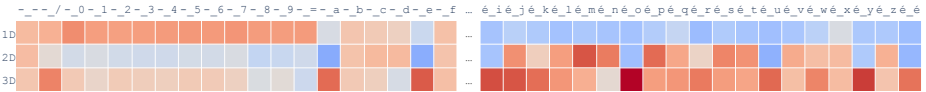


Eigenvectors of $M_* M^*$:

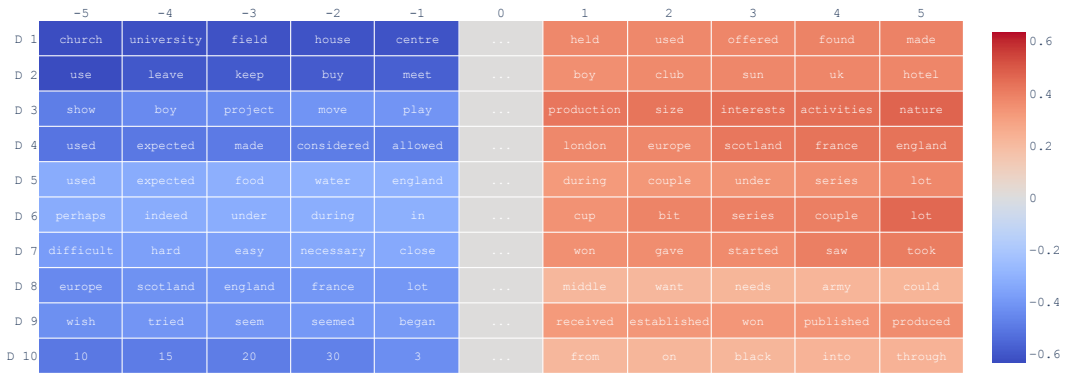
Eigenvalues of $M_* M^*$ and $M^* M_*$:



Eigenvectors of $M^* M_*$:



Words


-5 -4 -3 -2 -1 0 1 2 3 4 5
D 1 church university field house centre ... held used offered found made
D 2 use leave keep buy meet ... boy club sun uk hotel
D 3 show boy project move play ... production size interests activities nature
D 4 used expected made considered allowed ... london europe scotland france england
D 5 used expected food water england ... during couple under series lot
D 6 perhaps indeed under during in ... cup bit series couple lot
D 7 difficult hard easy necessary close ... won gave started saw took
D 8 europe scotland england france lot ... middle want needs army could
D 9 wish tried seem seemed began ... received established won published produced
D 10 10 15 20 30 3 ... from on black into through
0.6
0.4
0.2
0
-0.2
-0.4
-0.6

Outline

Introduction

Epistemological Perspectives

Theoretical Perspectives

- The Algebra Behind the Embeddings
- The Structure Behind the Algebra
- The Categories Behind the Structure

Take Aways

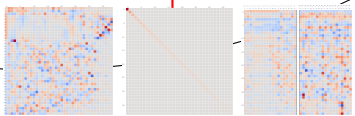
The Structure of Embeddings

Structure

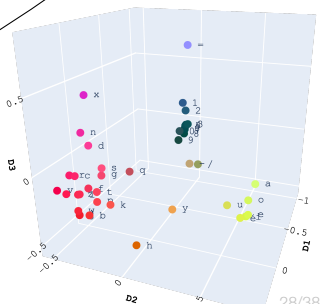
{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data



SVD



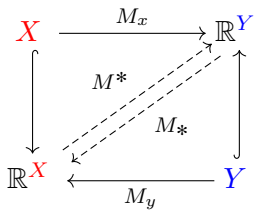
The Structure of Embeddings

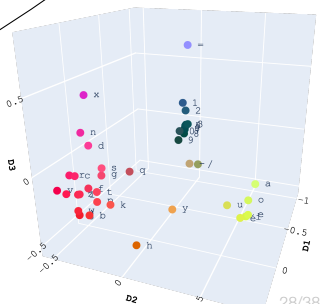
Structure

{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding

Data





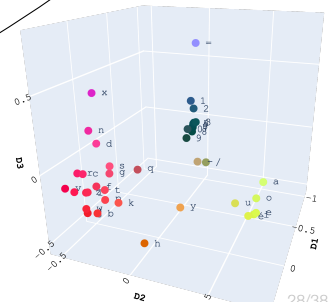
The Structure of Embeddings

Structure

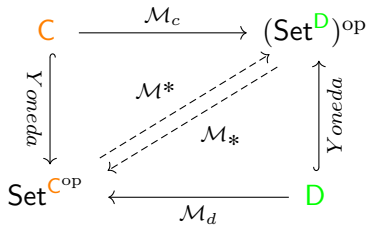
?

{-, /, 0, 1, 2, . . . , 8, 9, =,
a, b, c, . . . , w, x, y, z, é}

Embedding



Data



The Structure of Embeddings

Structure

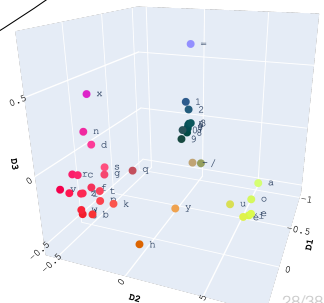
?

Data

$C^{op} \times D \rightarrow Set$

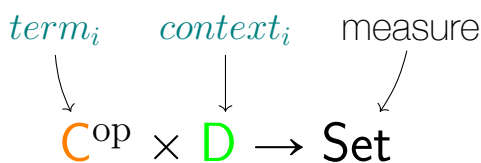
$\{-, /, 0, 1, 2, \ldots, 8, 9, =,$
$a, b, c, \ldots, w, x, y, z, é\}$

Embedding



Structure

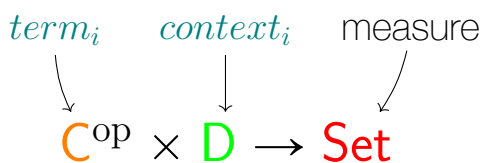
?


$term_i$
$context_i$
measure
$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \mathsf{Set}$

Nucleus and Types

Structure

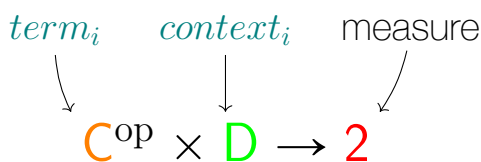
?



Nucleus and Types

Structure

?

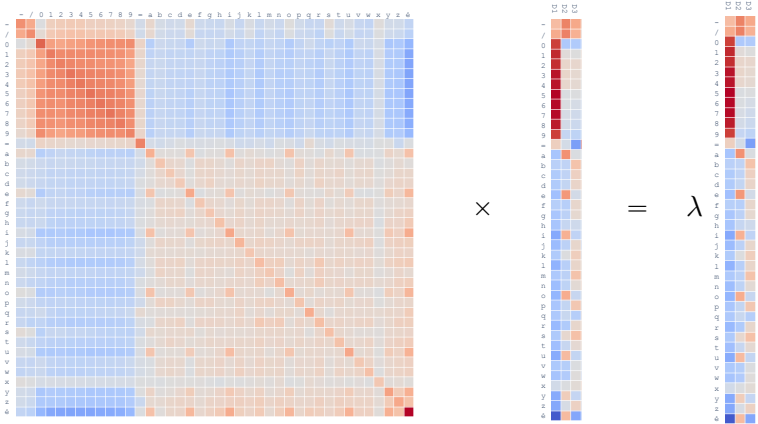


Structure

$$\begin{array}{c} \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow 2 \\ \Downarrow \\ \mathcal{M}^* : 2^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (2^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_* \end{array}$$

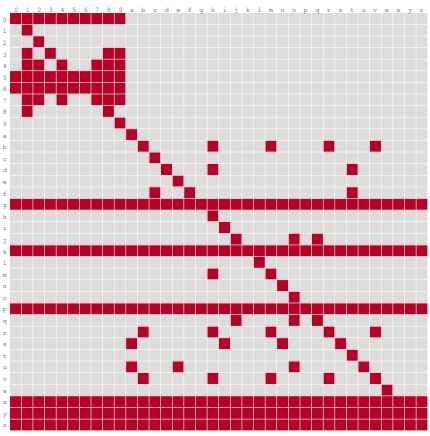
Binary Fixed Points

$$M_* M^* u = \lambda u$$

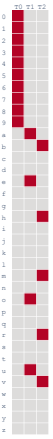


Binary Fixed Points

$$\mathcal{M}_* \mathcal{M}^* f = f$$



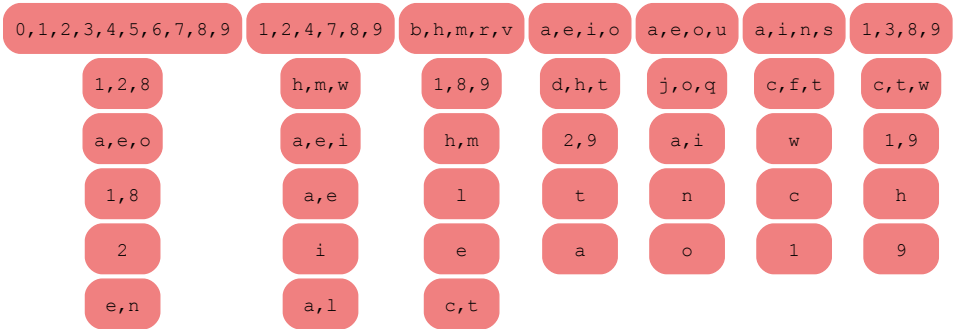
$\star$



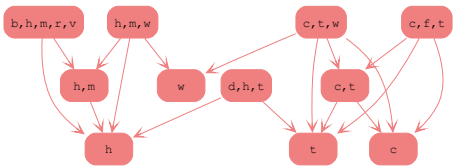
$\stackrel{?}{=}$

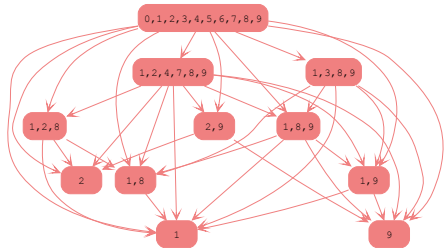
“Eigensets”

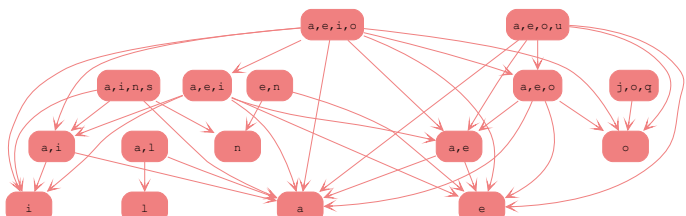
$$\mathcal{M}_* \mathcal{M}^* f = f$$



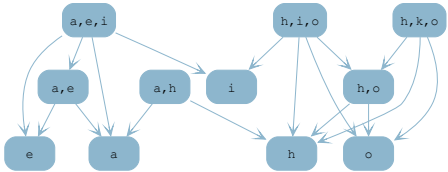
Partial Order Structure

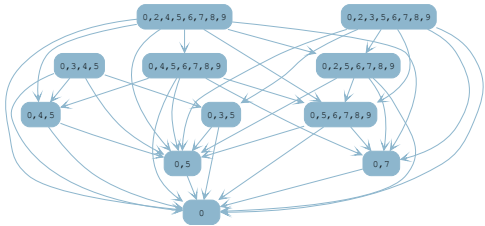


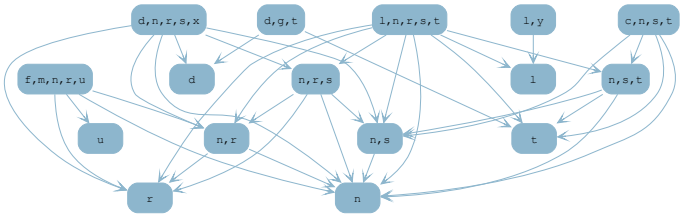




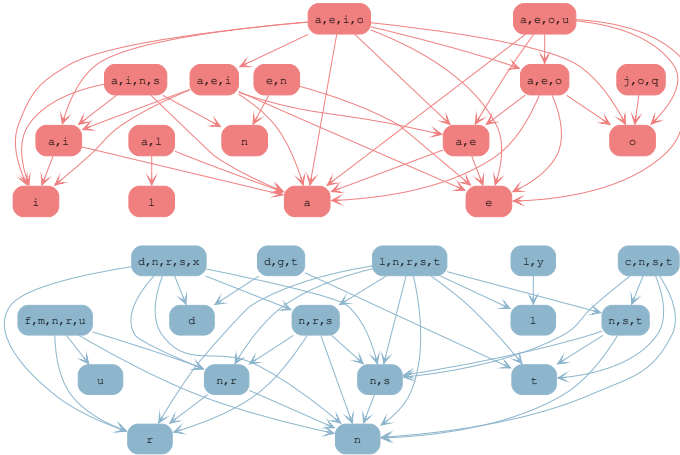
Dual Partial Order



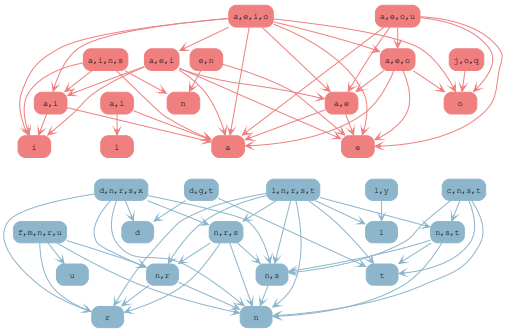




Paring of Partial Ordered Fixed Points



Structure



$$C^{\mathrm{op}} \times D \longrightarrow 2$$

$$\Downarrow$$

$$\mathcal{M}^* \colon 2^{C^{\mathrm{op}}} \leftrightarrows (2^D)^{\mathrm{op}} \colon \mathcal{M}_*$$

Structure

?

$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

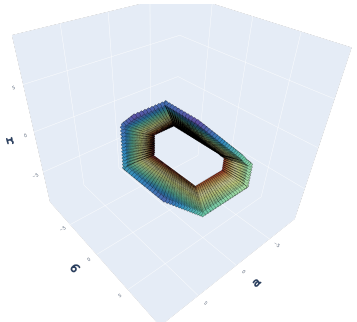
$$\Downarrow$$

$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

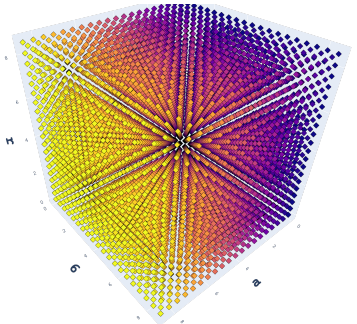
Enriching over $\bar{\mathbb{R}}$

Structure

?



$\mathcal{M}_* \mathcal{M}^*$

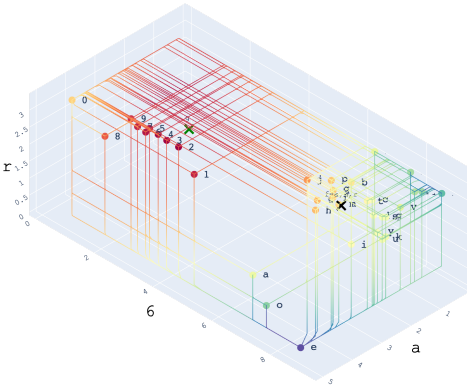


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_*$$

Structure

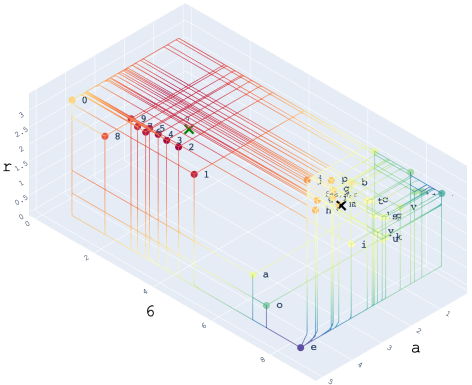


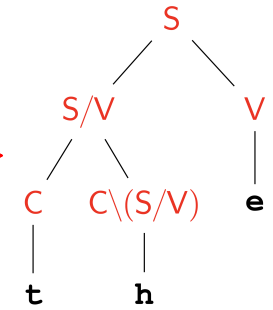
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

$$\mathcal{M}^* \colon \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{\mathsf{D}}\right)^{\mathrm{op}} \colon \mathcal{M}_*$$

Structure



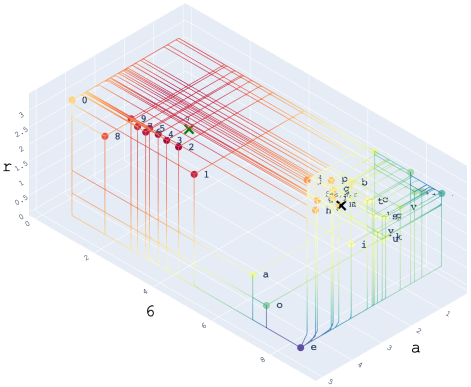


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\mathcal{M}^* : \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathsf{D}})^{\mathrm{op}} : \mathcal{M}_*$$

Formal Explainability

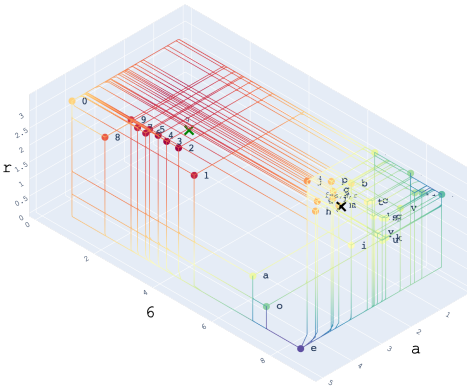
Structure



Embedding
Nucleus and Types

Formal Explainability

Structure



Tokenization
Renormalization

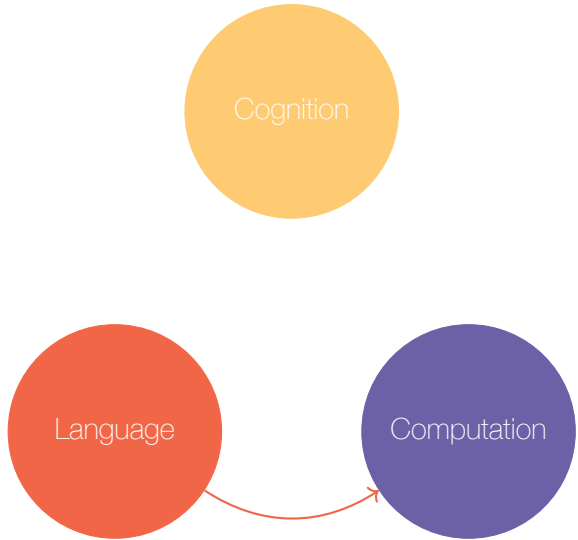
Embedding
Nucleus and Types

Attention
Tensor Product

Outline

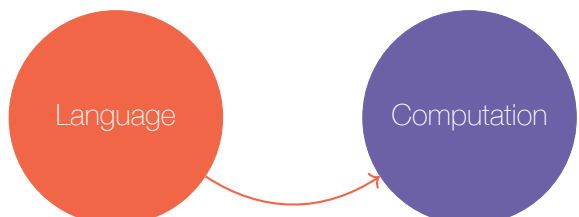
Take Aways

- A formal approach to data analysis can contribute to inferring symbolic language models from linguistic data.



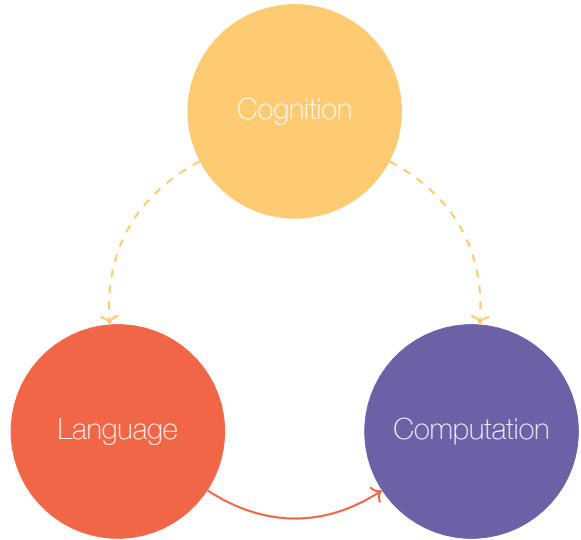
Take Aways

- A formal approach to data analysis can contribute to inferring symbolic language models from linguistic data.
- Resulting models are, a priori, models of the data.



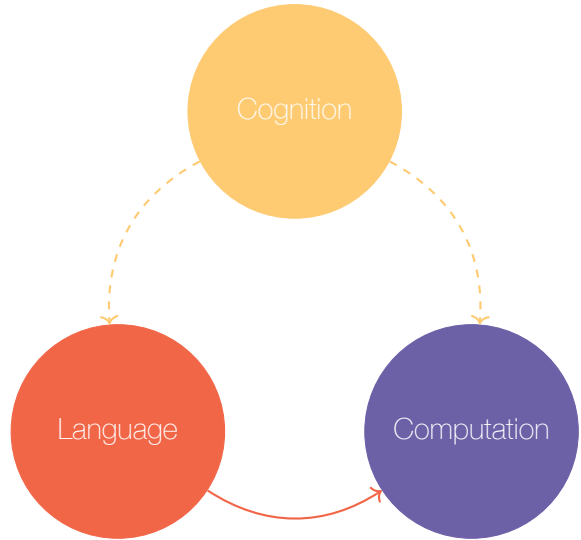
Take Aways

- A formal approach to data analysis can contribute to inferring symbolic language models from linguistic data.
- Resulting models are, a priori, models of the data.
- The cognitive content of such models is suspended, and cannot be restored without raising the problem of the data.



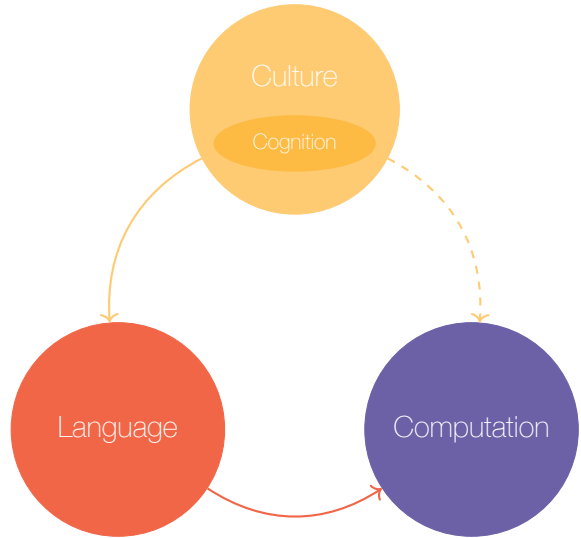
Take Aways

- A formal approach to data analysis can contribute to inferring symbolic language models from linguistic data.
- Resulting models are, a priori, models of the data.
- The cognitive content of such models is suspended, and cannot be restored without raising the problem of the data.
- The scale of the data for such models exceeds the individual scale.

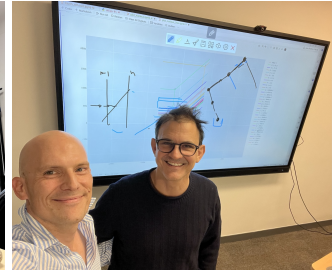


Take Aways

- A formal approach to data analysis can contribute to inferring symbolic language models from linguistic data.
- Resulting models are, a priori, models of the data.
- The cognitive content of such models is suspended, and cannot be restored without raising the problem of the data.
- The scale of the data for such models exceeds the individual scale.
- Cultural conditions of data production become constitutive in the relation between cognitive contents and language models.



Collaborations



J. Terilla (CUNY), T.-D. Bradley (SandboxAQ), L. Pellissier (Paris-Est Créteil), Th. Seiller (CNRS), S. Jarvis (CUNY)

Reference Papers

- Gastaldi, J. L. (2021).Why Can Computers Understand Natural Language? *Philosophy & Technology*, *34*(1), 149–214. https://doi.org/10.1007/s13347-020-00393-9

- Gastaldi, J. L., & Pellissier, L. (2021).The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, *46*(4), 569–590. https://doi.org/10.1080/03080188.2021.1890484

- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024).The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. https://api.semanticscholar.org/CorpusID:263613625

References I

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2024).Leace: Perfect linear concept erasure in closed form. *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024).The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. https://api.semanticscholar.org/CorpusID:263613625

Chomsky, N. (1953).Systems of syntactic analysis. *Journal of Symbolic Logic*, *18*(3), 242–256. https://doi.org/10.2307/2267409

Chomsky, N. (1956).Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124. https://doi.org/10.1109/TIT.1956.1056813

Chomsky, N. (1957). *Syntactic structures*. Mouton; Co.

Chomsky, N. (1959). *Language*, *35*(1), 26–58. Retrieved July 7, 2025, from http://www.jstor.org/stable/411334

Chomsky, N. (1992, November). Language and the "cognitive revolutions" [Delivered November 23–27, 1992].

Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L. K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., & Ortega, P. A. (2023). Neural networks and the chomsky hierarchy. https://arxiv.org/abs/2207.02098

Gastaldi, J. L. (2021).Why Can Computers Understand Natural Language? *Philosophy & Technology*, *34*(1), 149–214. https://doi.org/10.1007/s13347-020-00393-9

Gastaldi, J. L., & Pellissier, L. (2021).The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, *46*(4), 569–590. https://doi.org/10.1080/03080188.2021.1890484

Girard, J.-Y. (2011, September). *The blind spot*. European Mathematical Society.

References II

Goldberg, Y., & Levy, O. (2014).Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.

Levy, O., & Goldberg, Y. (2014).Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., & Strohmann, T. (2013). *Learning representations of text using neural networks. NIPS deep learning workshop 2013 slides*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013).Distributed representations of words and phrases and their compositionality. *CoRR*, *abs/1310.4546*.

Sennrich, R., Haddow, B., & Birch, A. (2016).Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the ACL*, 1715–1725.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

NeuroMod Annual Meeting
Université Côte d'Azur
Antibes, France

What are Neural Language Models the Model of?

Epistemological and Theoretical Perspectives on LLMs

Juan Luis Gastaldi

www.giannigastaldi.com

ETH zürich

July 8, 2025