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

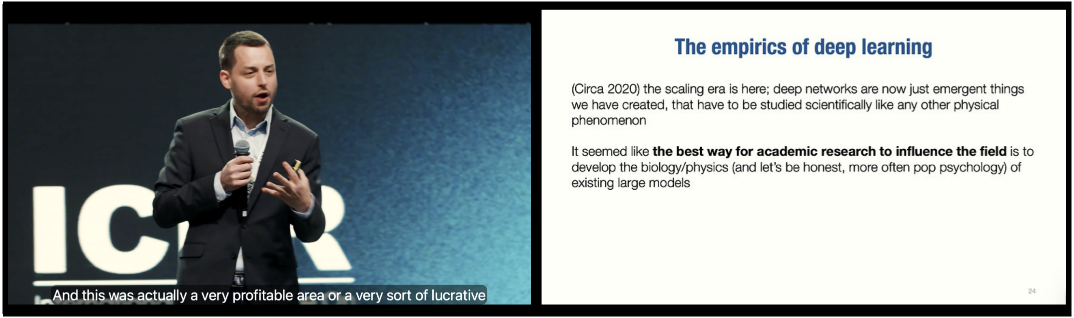
Theoretical Perspectives

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

Take Aways

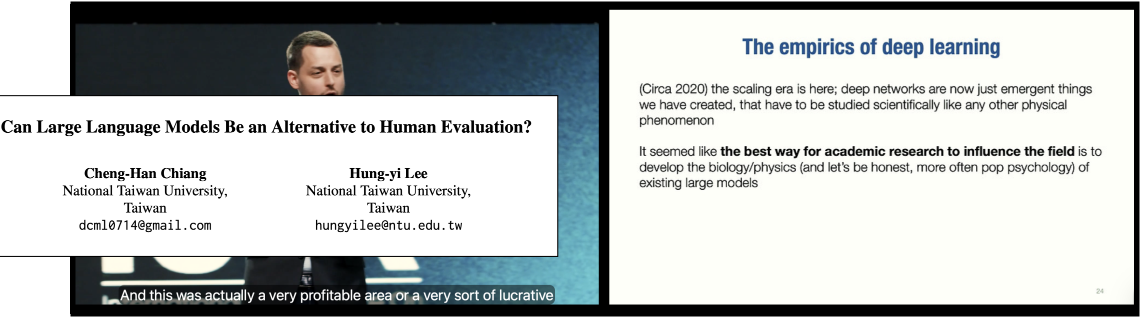
Outline

Empirical Saturnalia



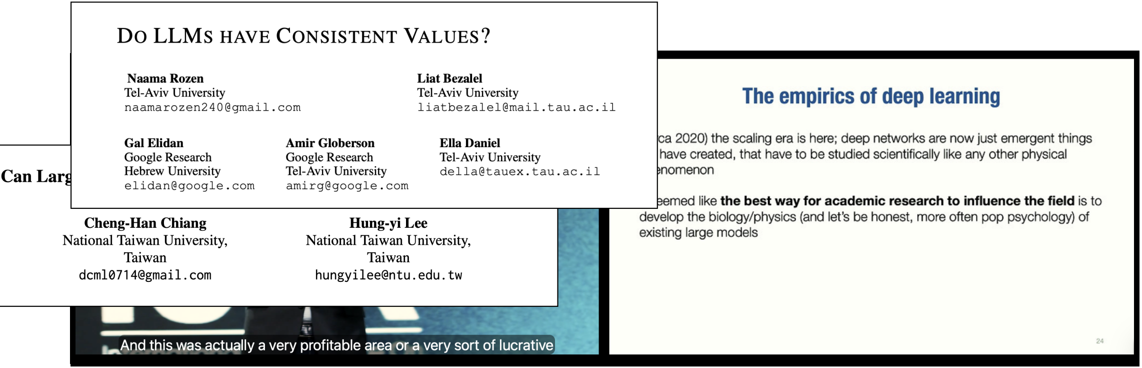
Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia

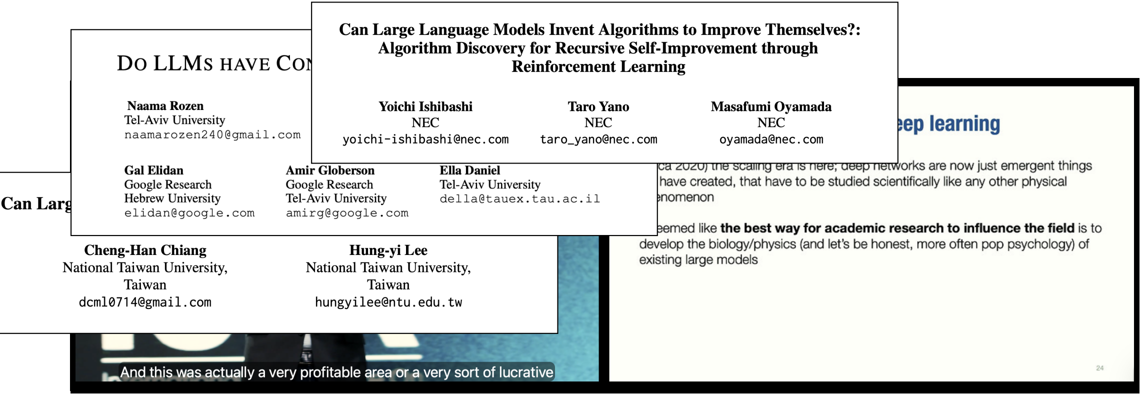


Zico Kolter, *Building Safe and Robust AI Systems*, Keynote at ICLR 2025.

Empirical Saturnalia



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



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

Yoichi Ishibashi NEC yoichi-ishibashi@nec.com
Taro Yano NEC taro_yano@nec.com
Masafumi Oyamada NEC oyamada@nec.com

DO LLMS HAVE CO

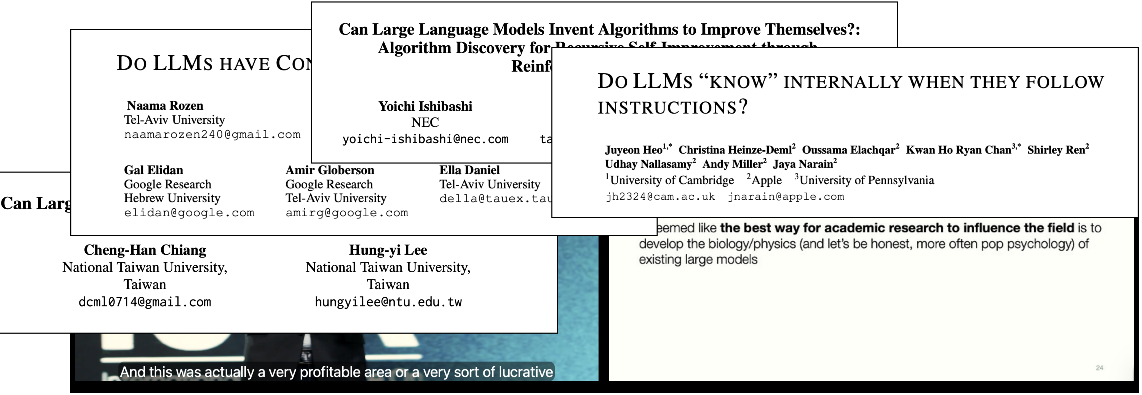
Naama Rozen Tel-Aviv University naamarozen240@gmail.com
Gal Elidan Google Research, Hebrew University elidan@google.com
Amir Globerson Google Research, Tel-Aviv University amirg@google.com
Ella Daniel Tel-Aviv University della@tauex.tau.ac.il

Can Larg

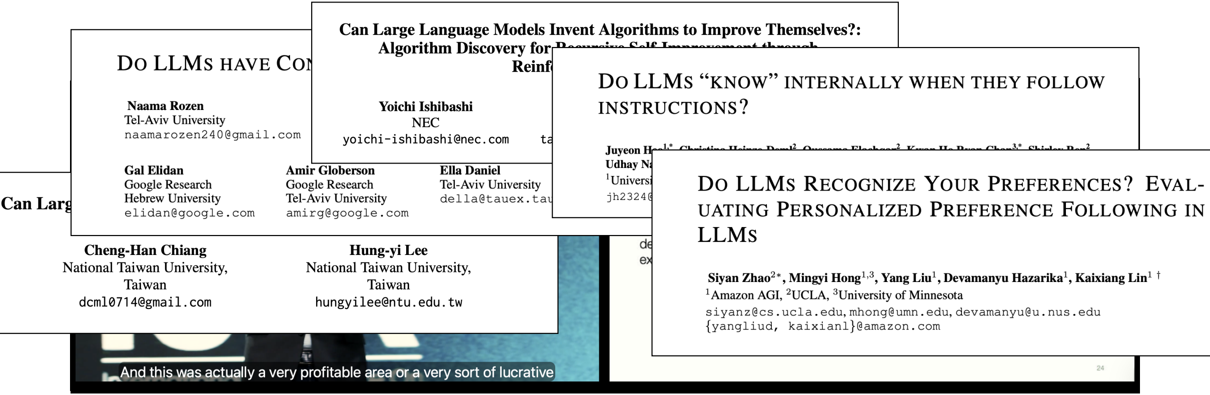
Cheng-Han Chiang National Taiwan University, Taiwan dcml0714@gmail.com
Hung-yi Lee National Taiwan University, Taiwan hungyilee@ntu.edu.tw

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

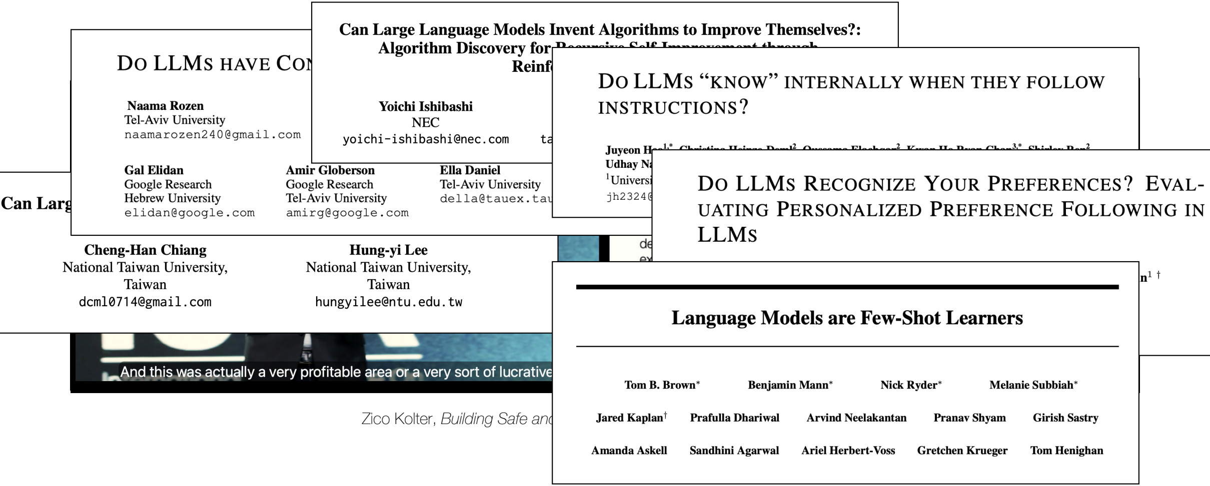
Empirical Saturnalia



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



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


Can Large Language Models Invent Algorithms to Improve Themselves?:
Algorithm Discovery for
Reinf
Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com
Do LLMs have Co
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
Do LLMs "know" internally when they follow instructions?
Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs
Can Larg
Cheng-Han Chiang
National Taiwan University, Taiwan
dcml0714@gmail.com
Hung-yi Lee
National Taiwan University, Taiwan
hungyilee@ntu.edu.tw
And this was actually a very profitable area or a very sort of lucrative
Zico Kolter, Building Safe and
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

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 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?

Can Larg

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

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

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

Language Models are Few-Shot Learners

Melanie Subbiah*
Pranav Shyam
Girish Sastry
Gretchen Krueger
Tom Henighan

And this was actually a very pr

LLMs Are Not Intelligent Thinkers: Introducing Mathematical Topic Tree Benchmark for Comprehensive Evaluation of LLMs

Arash Gholami Davoodi[1], Seyed Pouyan Mousavi Davoudi, Pouya Pezeshkpour[2]
[1]Carnegie Mellon University, [2]Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

Empirical Saturnalia

Can Larg

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

Yoichi Ishibashi
NEC
yoichi-ishibashi@nec.com ta

DO LLMS "KNOW" INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Juyeon H
Udhay Na
[1]Universi
jh2324@

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

n[1 †]

Cheng-Han Chiang
National Taiwan University,
Taiwan

Hung-yi Lee
National Taiwan University,
Taiwan

Language Models are Few-Shot Learners

k Ryder* Melanie Subbiah*

n Pranav Shyam Girish Sastry

Gretchen Krueger Tom Henighan

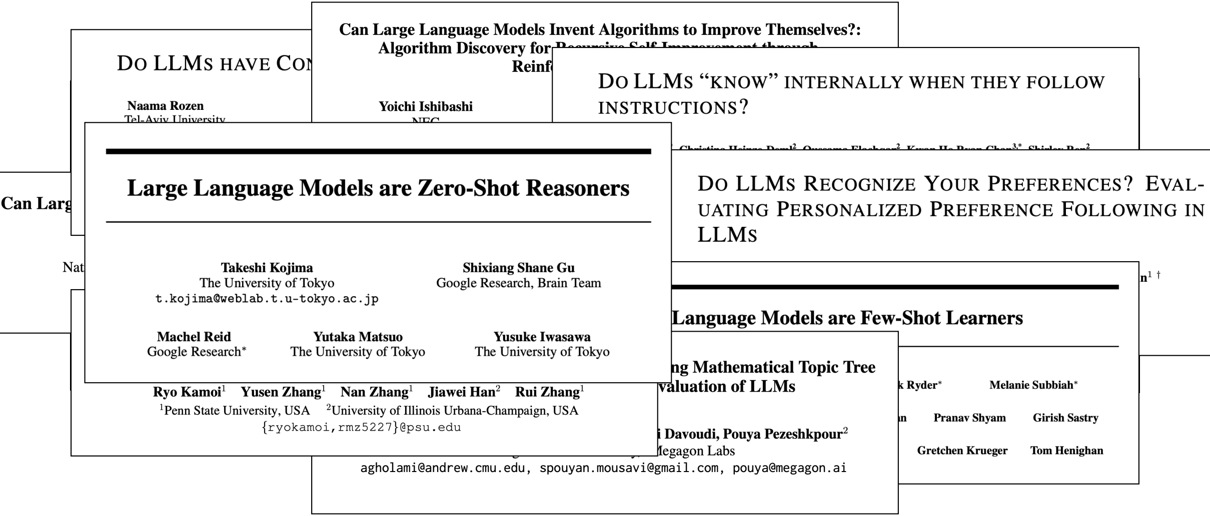
When Can LLMs *Actually* Correct Their Own Mistakes?
A Critical Survey of Self-Correction of LLMs

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

cing Mathematical Topic Tree
valuation of LLMs

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

Empirical Saturnalia


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

Yoichi Ishibashi

DO LLMS HAVE CON

Naama Rozen

DO LLMS "KNOW" INTERNALLY WHEN THEY FOLLOW INSTRUCTIONS?

Can Larg

Nat

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

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

Language Models are Few-Shot Learners

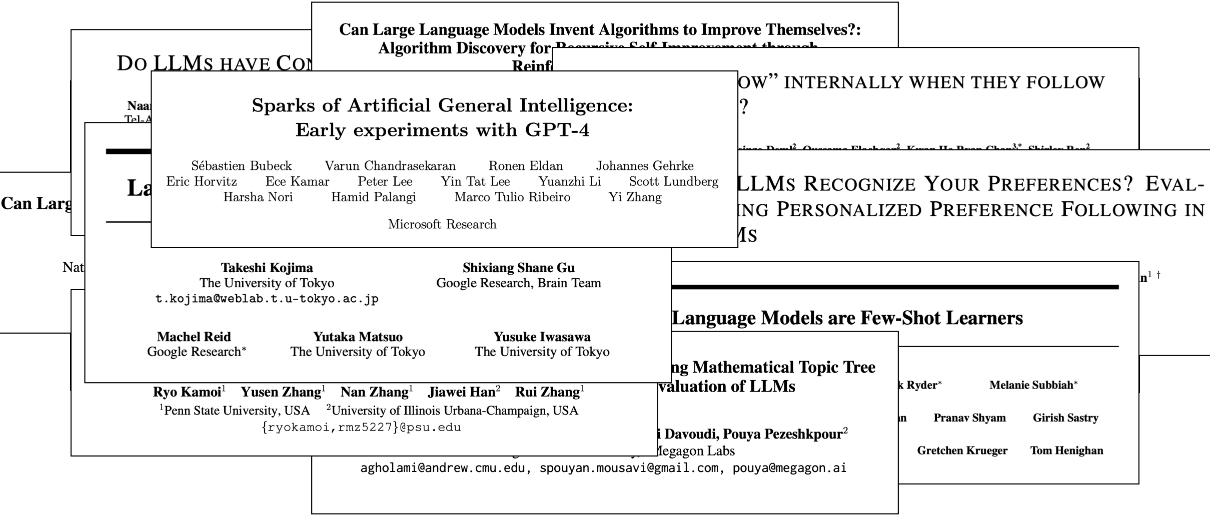
ng Mathematical Topic Tree
valuation of LLMs

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

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

k Ryder* Melanie Subbiah*
Pranav Shyam Girish Sastry
Gretchen Krueger Tom Henighan

Empirical Saturnalia



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

DO LLMS HAVE CON

Naa
Tel-A

OW" INTERNALLY WHEN THEY FOLLOW
?

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Can Larg

La

LLMS RECOGNIZE YOUR PREFERENCES? EVAL-
ING PERSONALIZED PREFERENCE FOLLOWING IN
MS

Nat

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

n[1 †]

Language Models are Few-Shot Learners

ng Mathematical Topic Tree
valuation of LLMs

k Ryder* Melanie Subbiah*

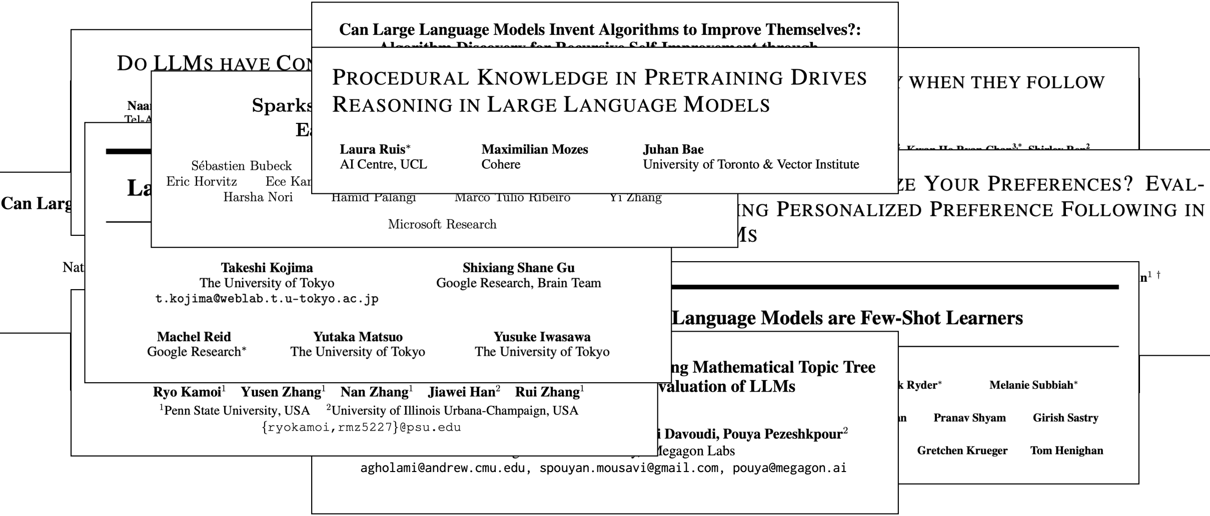
n Pranav Shyam Girish Sastry

Gretchen Krueger Tom Henighan

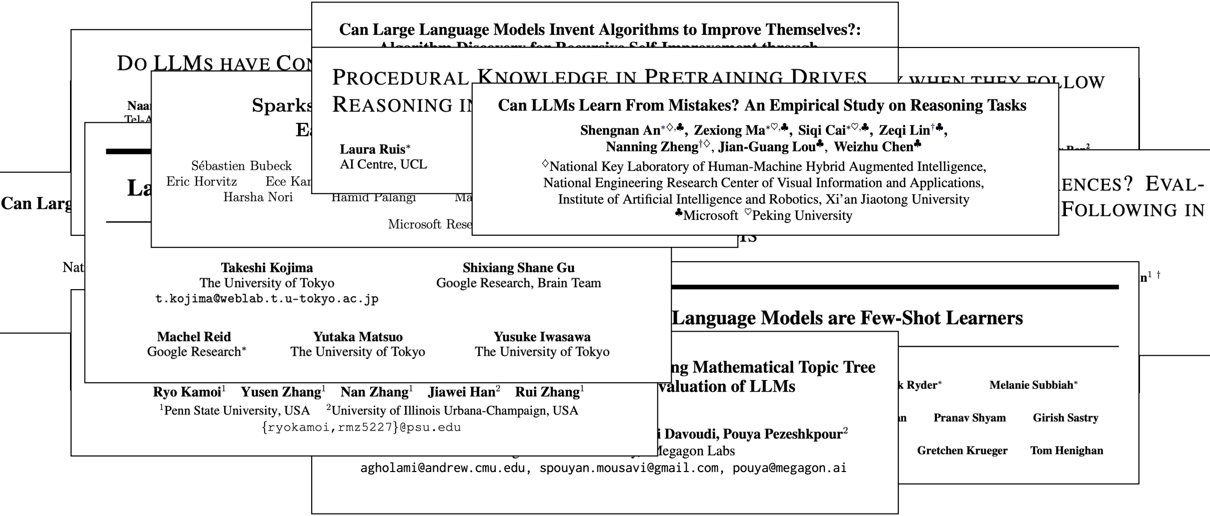
Ryo Kamoi[1] Yusen Zhang[1] Nan Zhang[1] Jiawei Han[2] Rui Zhang[1]
[1]Penn State University, USA [2]University of Illinois Urbana-Champaign, USA
{ryokamoi,rmz5227}@psu.edu

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

Empirical Saturnalia



Empirical Saturnalia



Empirical Saturnalia

Can Large Language Models Invent Algorithms to Improve Themselves?:

PROCEDURAL KNOWLEDGE IN PRETRAINING DRIVES REASONING IN

Laura Ruis*
AI Centre, UCL

Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks

Shengnan An*◇,♣, Zexiong Ma*♡,♣, Siqi Cai*♡,♣, Zeqi Lin†♣, Nanning Zheng†◇, Jian-Guang Lou♣, Weizhu Chen♣

◇National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,

Sparks

Sébastien Bubeck
Eric Horvitz Ece Kar
Harsha Nori Hamid Palangi

Microsoft Rese

Can Large Language Models Reason About Goal-Oriented Tasks?

Filippos Bellos Yayuan Li Wuao Liu Jason J. Corso
University of Michigan, Ann Arbor, Michigan, USA
{fbellos,yayuanli,wuaoliu,jjcorso}@umich.edu

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

Shixiang Sha
Google Research,

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

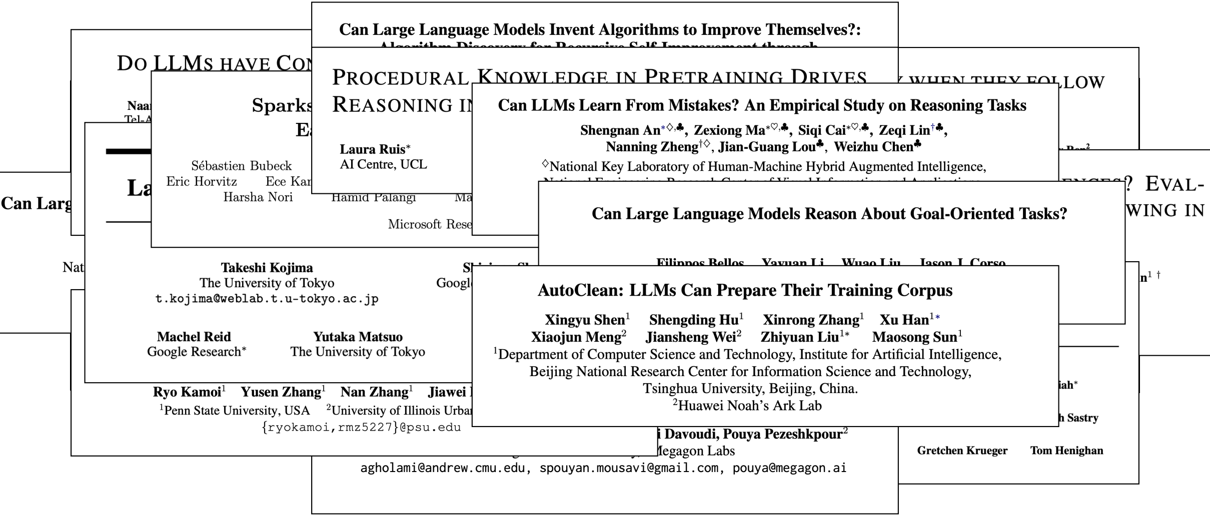
Ryo Kamoi[1] Yusen Zhang[1] Nan Zhang[1] Jiawei Han[2] Rui Zhang[1]
[1]Penn State University, USA [2]University of Illinois Urbana-Champaign, USA
{ryokamoi,rmz5227}@psu.edu

ng Mathematical Topic Tree
valuation of LLMs

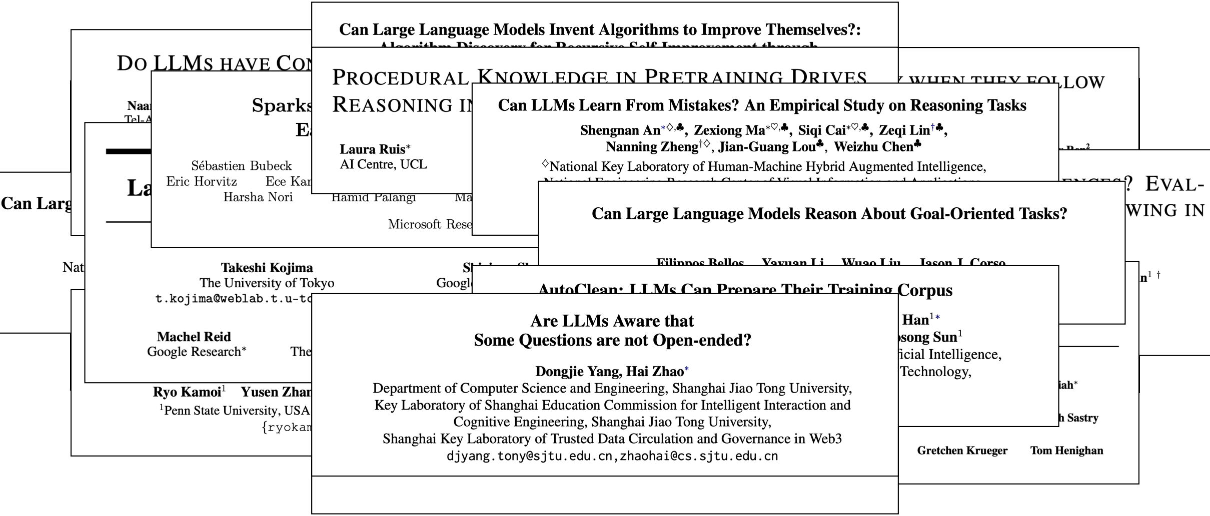
i Davoudi, Pouya Pezeshkpour[2]
Megagon Labs
agholami@andrew.cmu.edu, spouyan.mousavi@gmail.com, pouya@megagon.ai

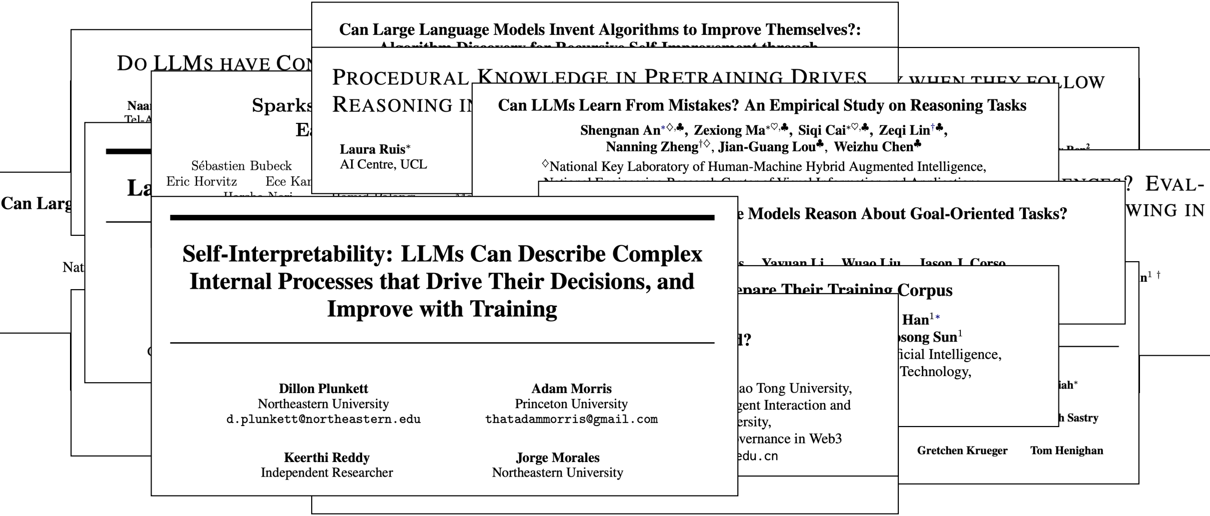
Pranav Shyam Girish Sastry
Gretchen Krueger Tom Henighan
Melanie Subbiah*

Empirical Saturnalia

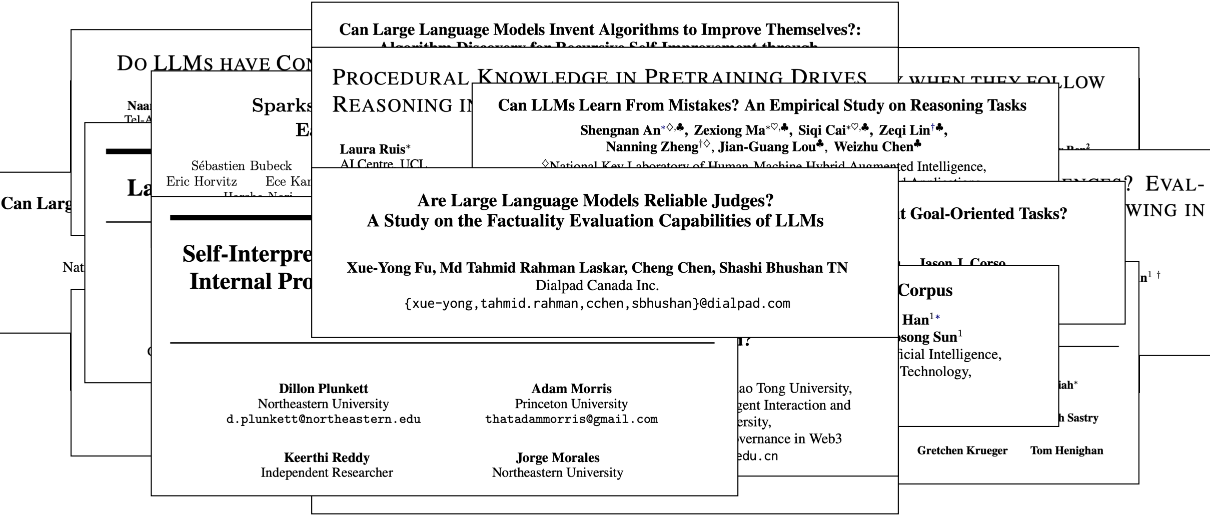


Empirical Saturnalia

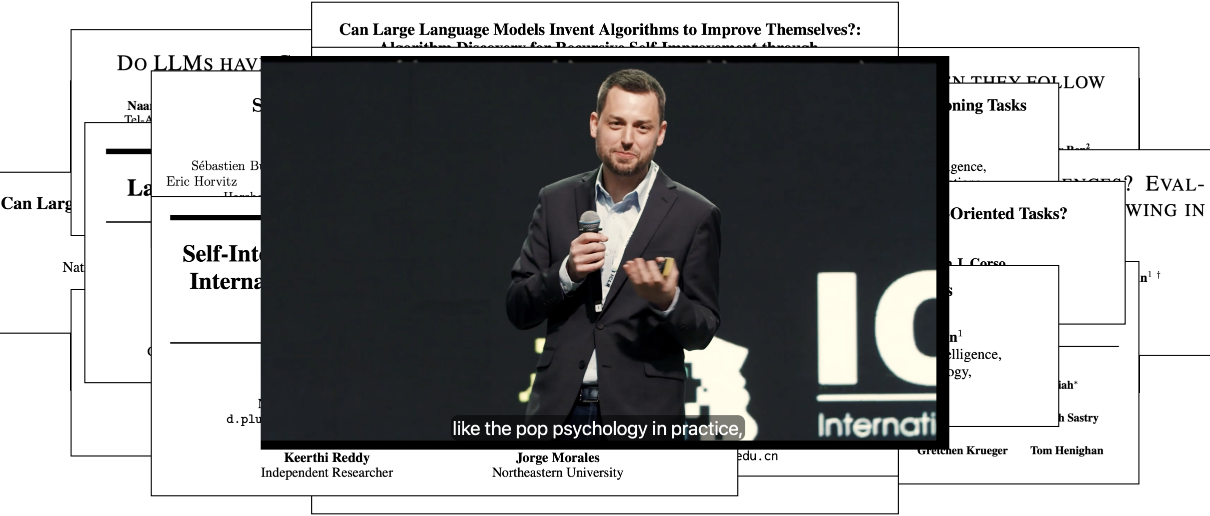



Can Large Language Models Invent Algorithms to Improve Themselves?:
Procedural Knowledge in Pretraining Drives
Reasoning in
Laura Ruis*
AI Centre, UCL
Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks
Shengnan An, Zexiong Ma, Siqi Cai, Zeqi Lin,
Nanning Zheng, Jian-Guang Lou, Weizhu Chen
National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Do LLMs have Con
Sparks
Sébastien Bubeck
Eric Horvitz
e Models Reason About Goal-Oriented Tasks?
when they follow
Self-Interpretability: LLMs Can Describe Complex Internal Processes that Drive Their Decisions, and Improve with Training
Dillon Plunkett
Northeastern University
d.plunkett@northeastern.edu
Adam Morris
Princeton University
thatadammorris@gmail.com
Keerthi Reddy
Independent Researcher
Jorge Morales
Northeastern University
Corpus
Technology,
ao Tong University,
gent Interaction and
vernance in Web3
Gretchen Krueger
Tom Henighan

Empirical Saturnalia

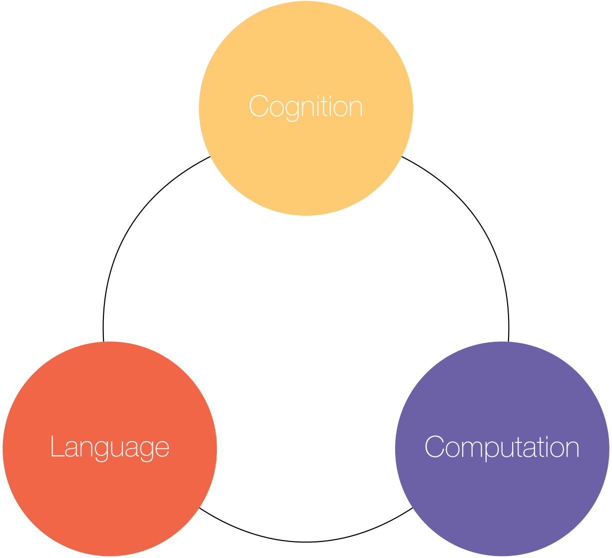


Empirical Saturnalia



Outline

Chomsky's Generativist Program and the Cognitive Revolution



Chomsky vs. LLMs

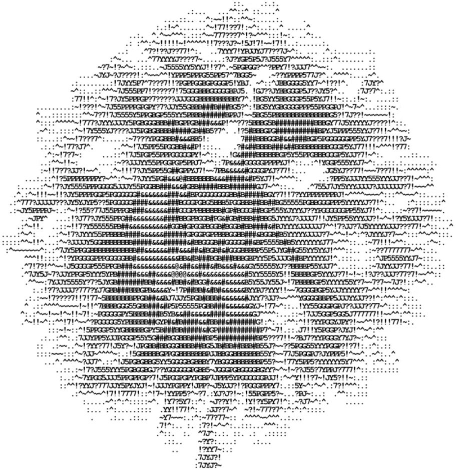
The New York Times

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)

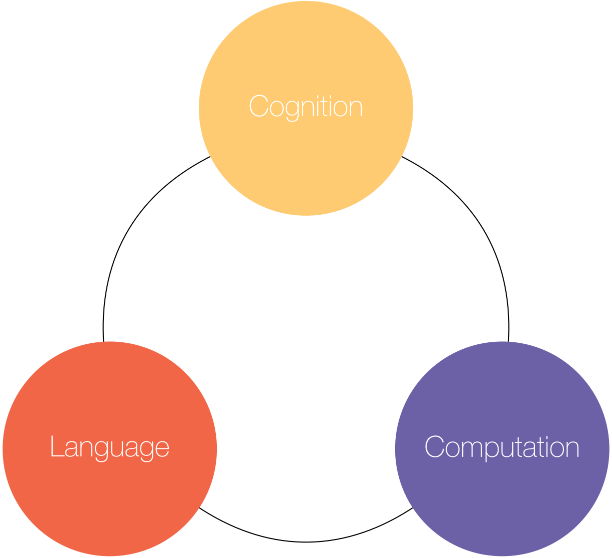
Chomsky against Abstraction in Principle

"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."

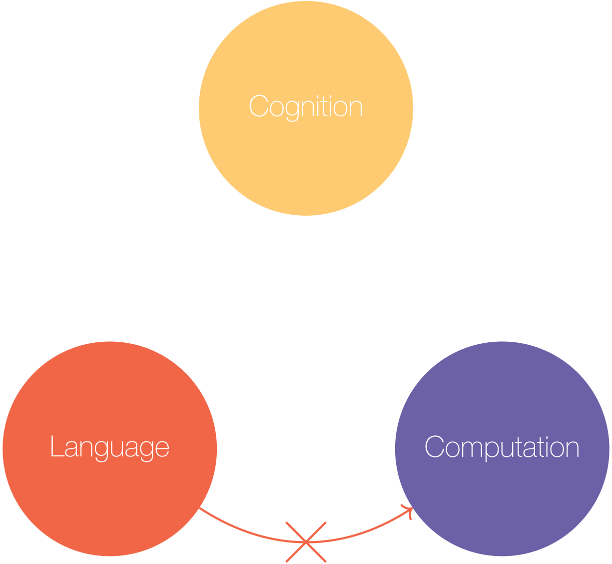
"I don't think we should cross that border, because there's no empirical claim. It is just a question of how to extend the metaphor."

(Chomsky, 1992)

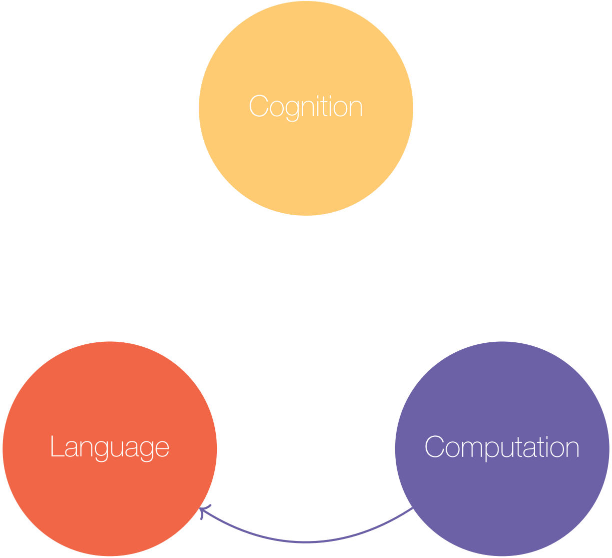
The Condition of Chomsky's Cognitive Foundations



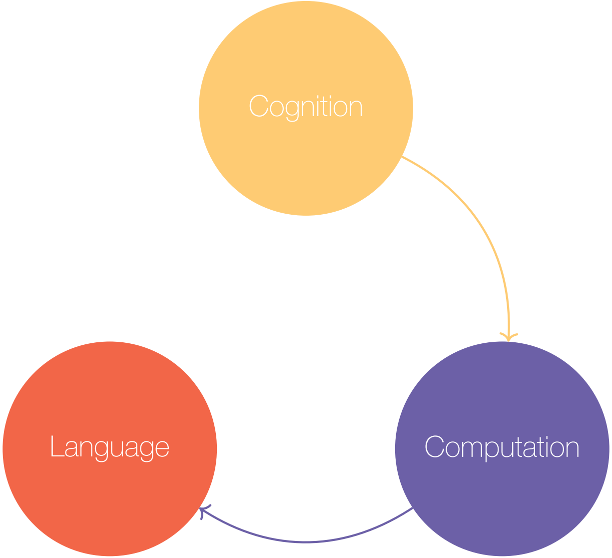
The Condition of Chomsky's Cognitive Foundations



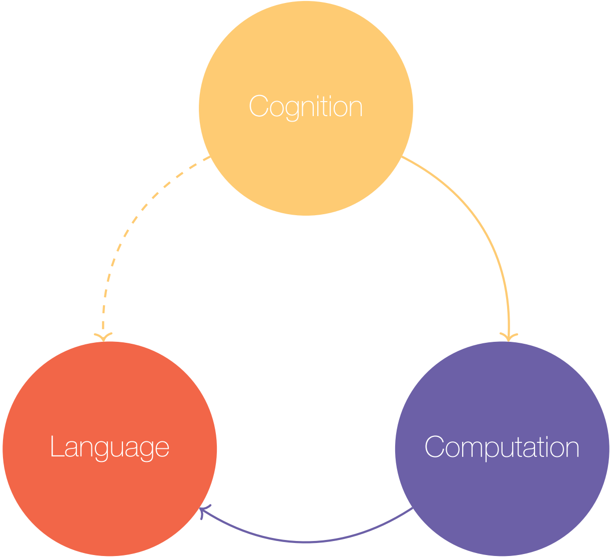
The Condition of Chomsky's Cognitive Foundations



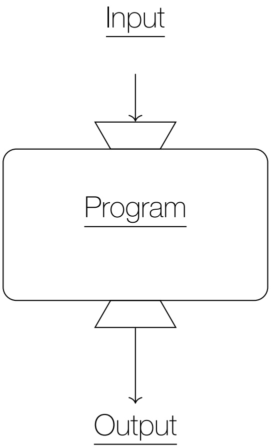
The Condition of Chomsky's Cognitive Foundations



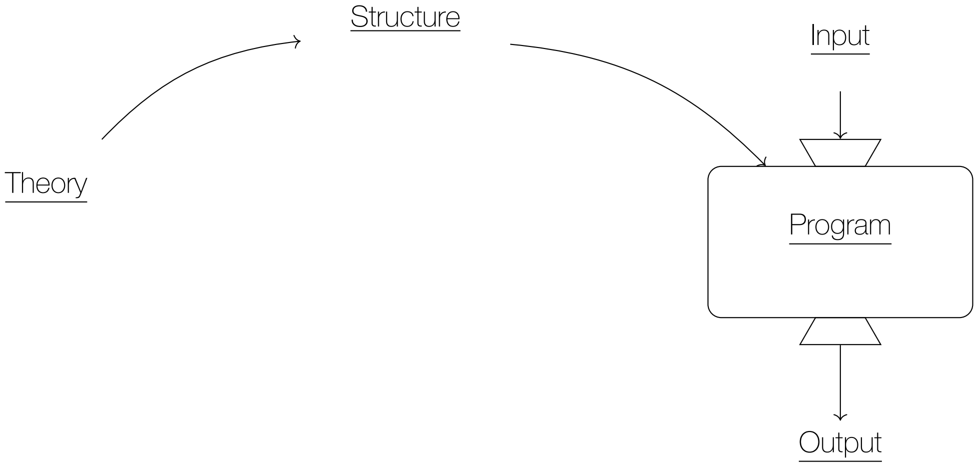
The Condition of Chomsky’s Cognitive Foundations



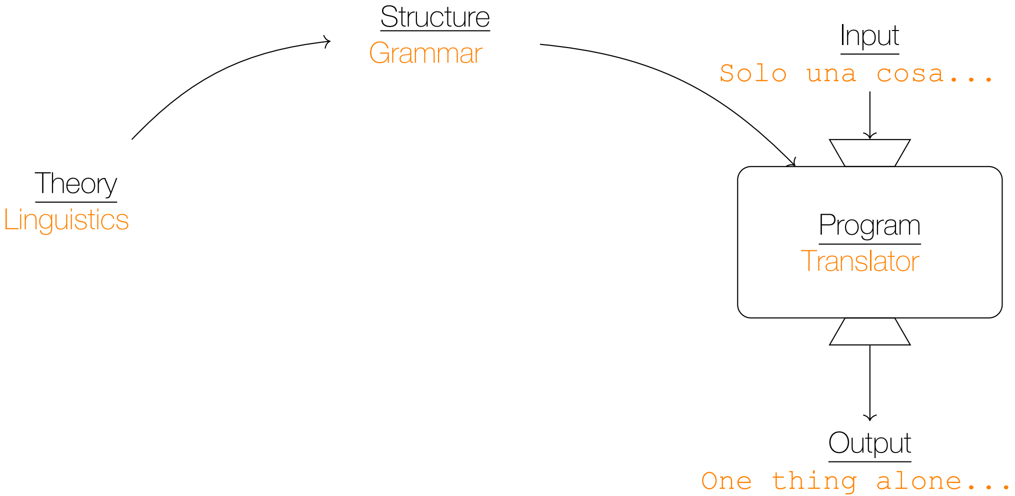
The Trap



The Trap

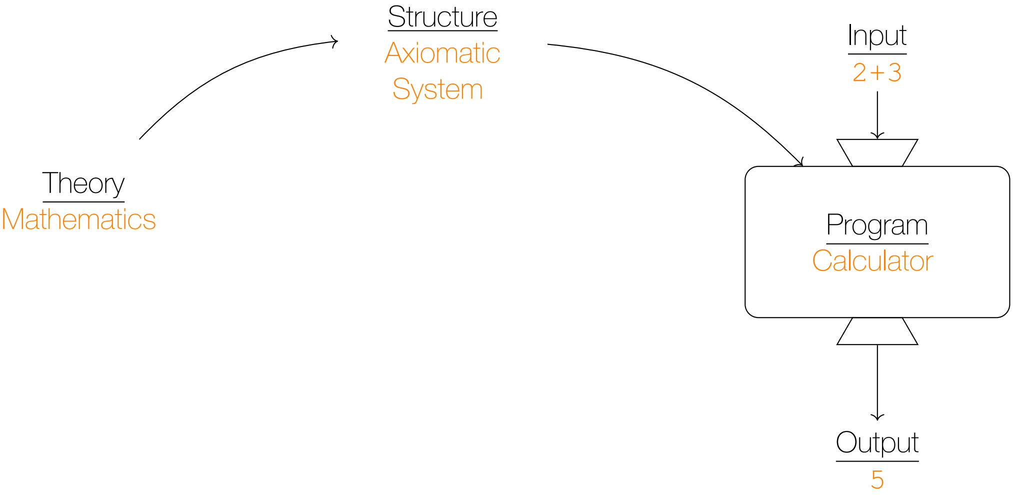


The Trap

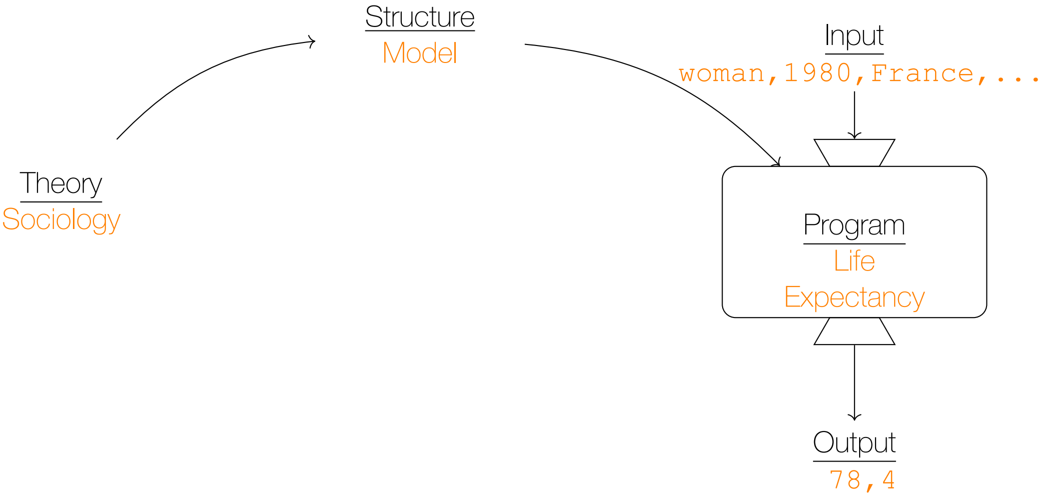


Structure
Grammar
Input
Solo una cosa...
Theory
Linguistics
Program
Translator
Output
One thing alone...

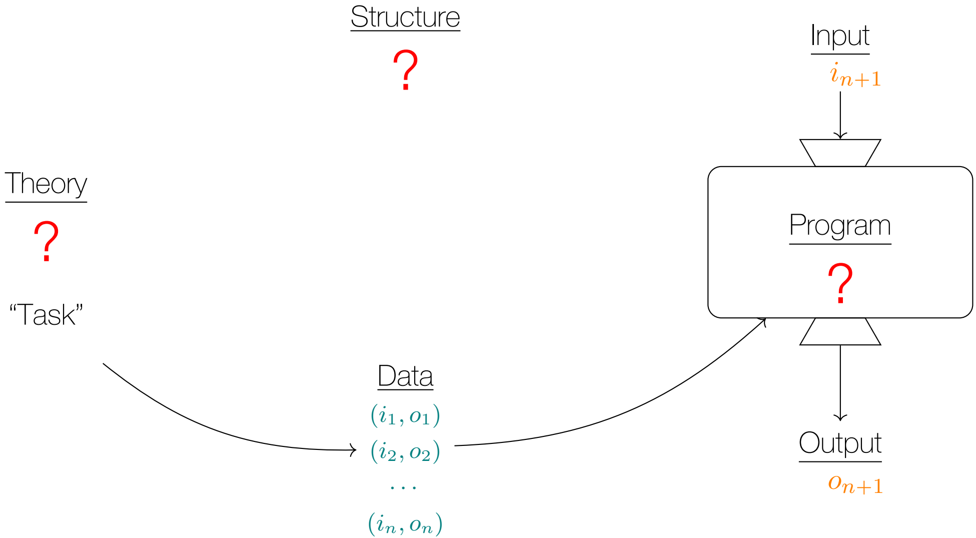
The Trap


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

The Trap


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

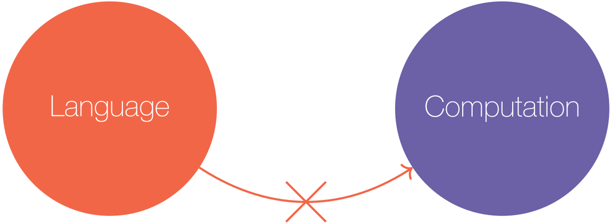
The Trap



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

Necessary Condition?

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



Necessary Condition?

- Inadequacy of distributional models (Chomsky, 1953)

  Inconclusive

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

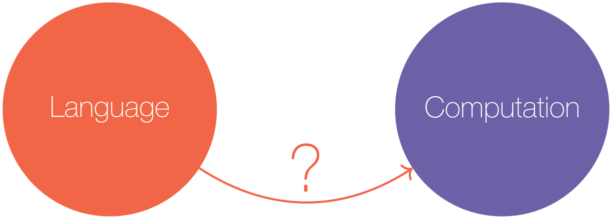
  Empirically challenged

- Limited expressive power of FSAs (Chomsky, 1956)

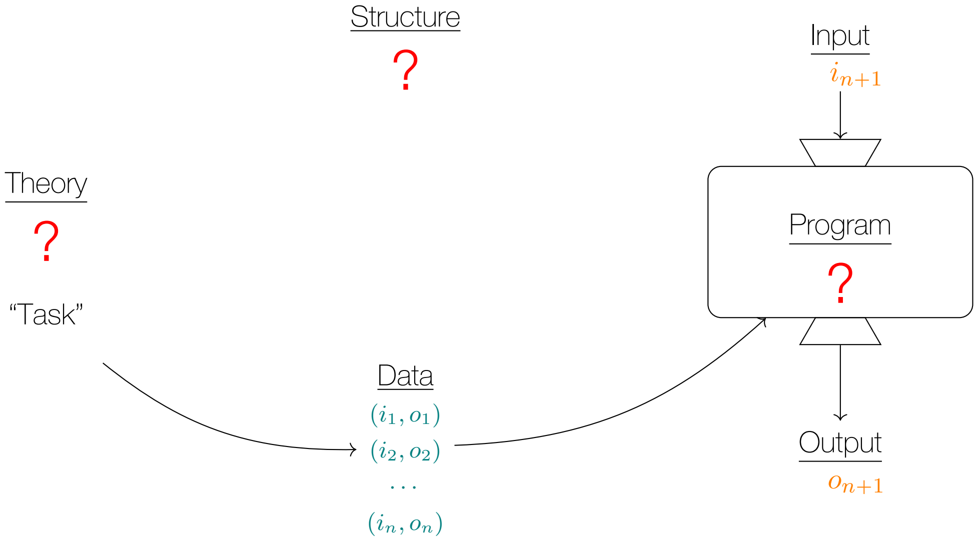
  The relevance is unclear

- Poverty of stimulus (Chomsky, 1959)

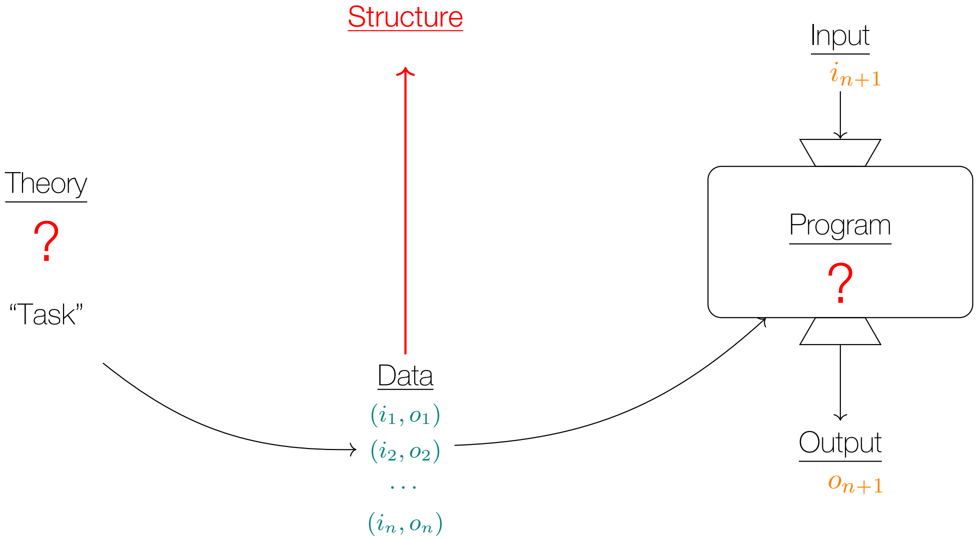
  Assumes what is to be proved



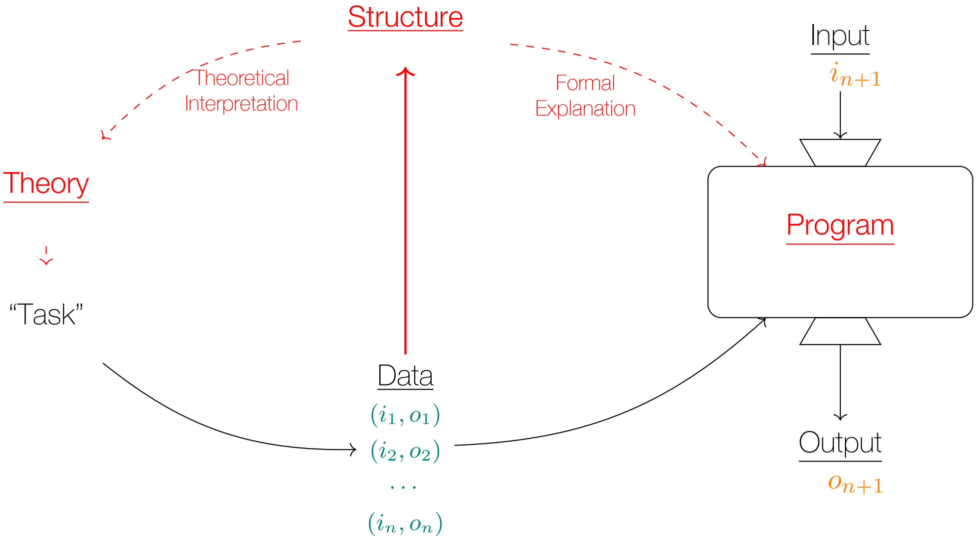
Making It Explicit



Making It Explicit



Making It Explicit



Outline

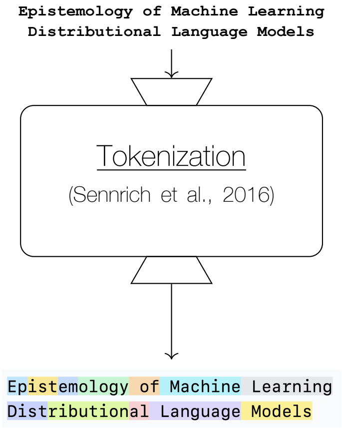
Introduction

Epistemological Perspectives

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

Take Aways

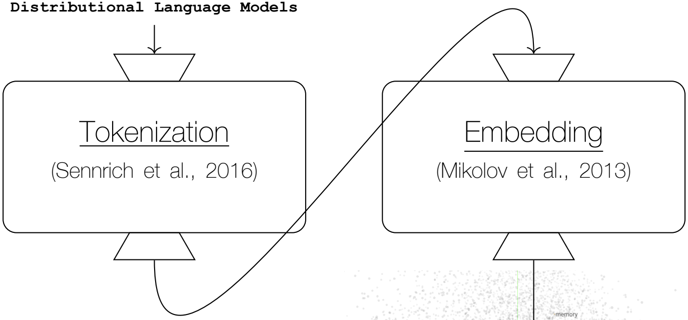
Formal Explainability



(https://tiktokenizer.vercel.app)

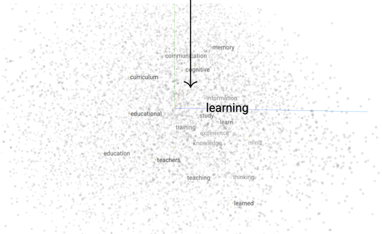
Formal Explainability

Epistemology of Machine Learning
Distributional Language Models



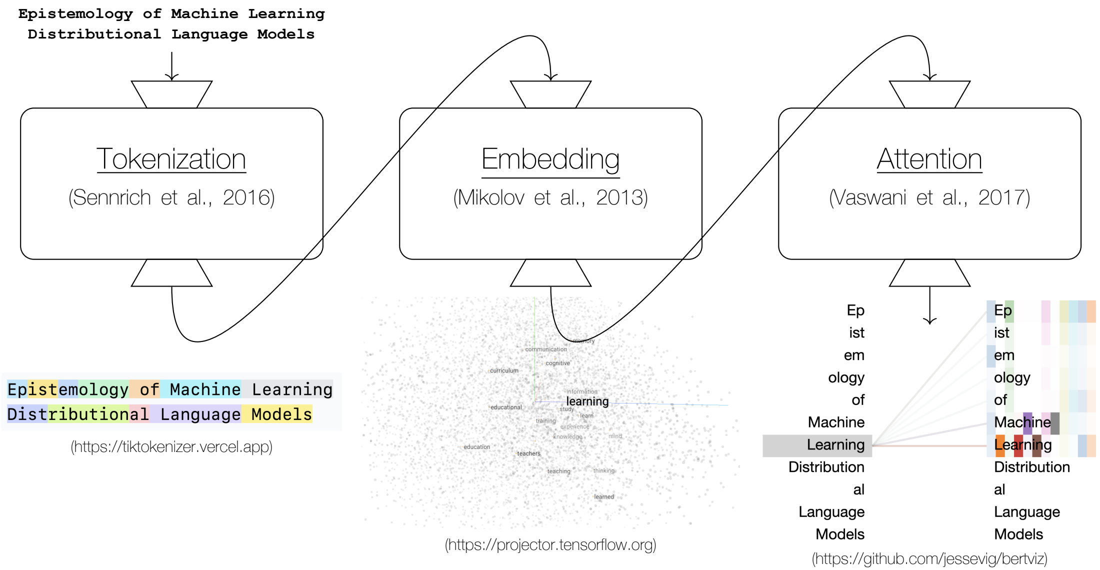
`Epistemology of Machine Learning`
`Distributional Language Models`

(https://tiktokenizer.vercel.app)

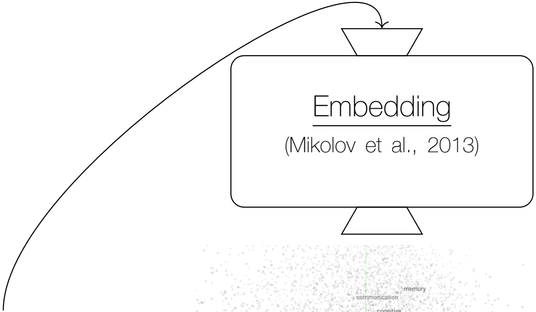


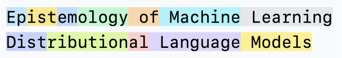
(https://projector.tensorflow.org)

Formal Explainability

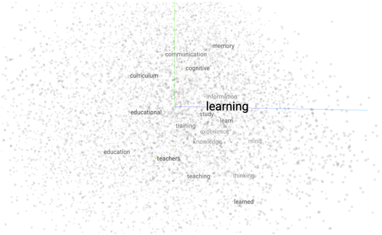


Formal Explainability



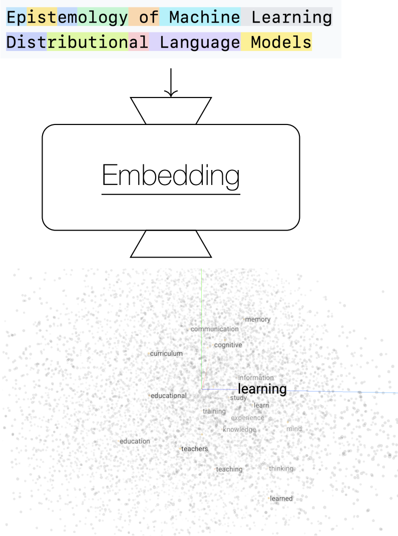


(https://tiktokenizer.vercel.app)

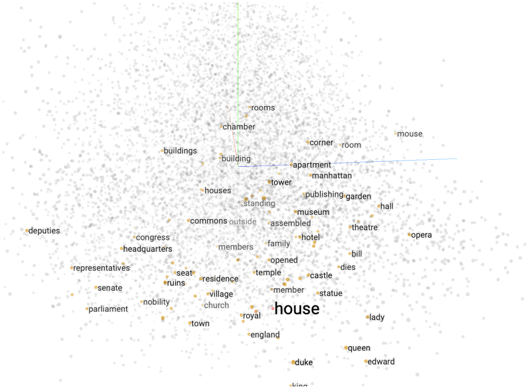


(https://projector.tensorflow.org)

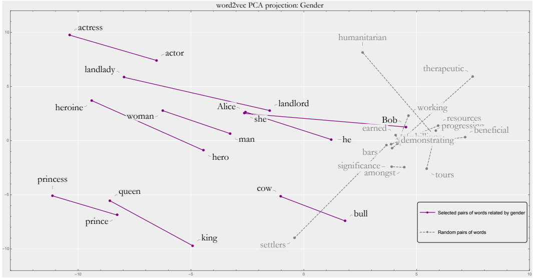
The Structure of Embeddings



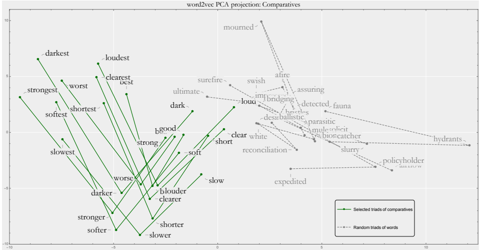
Embeddings: Similarity and Analogy

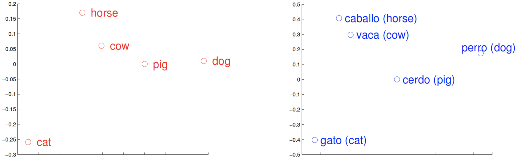


(https://projector.tensorflow.org)

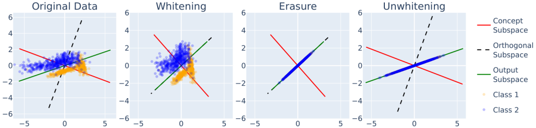


Embeddings: Other Applications

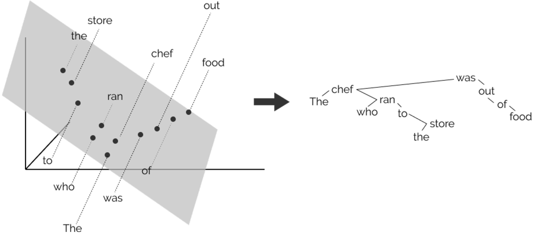




(Mikolov, Sutskever, Chen, Corrado, Dean, et al., 2013)



(Belrose et al., 2024)



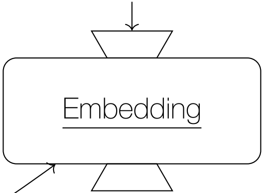
(https://nlp.stanford.edu/~johnhew/structural-probe.html)

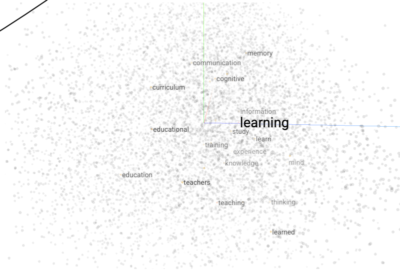
The Structure of Embeddings

Structure

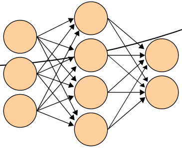
?

Epistemology of Machine Learning
Distributional Language Models

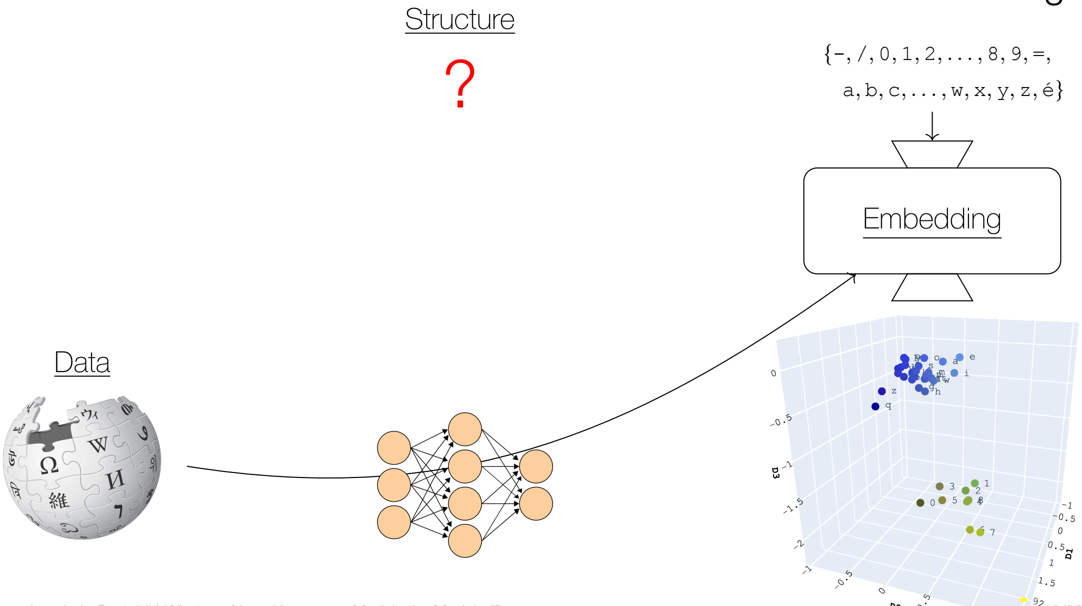




Data



The Structure of Embeddings



Outline

Introduction

Epistemological Perspectives

Theoretical Perspectives

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

Take Aways

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained

(Levy and Goldberg, 2014)

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

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

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

word2vec Explained
(Levy and Goldberg, 2014)

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

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

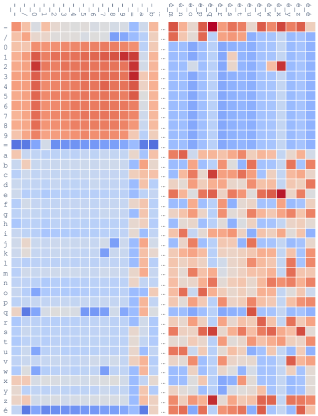
- Word2vec performs an implicit, low-dimensional factorization of a pointwise mutual information (pmi), word-context matrix.
- The Singular Value Decomposition (SVD) provides an exact solution to this problem.

Example: Characters in Wikipedia

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

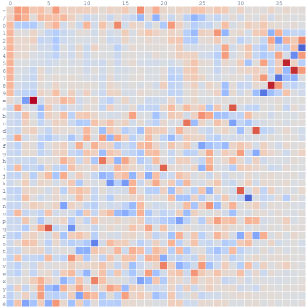
$$C = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \dots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M_{wc} = \operatorname{pmi}(w, c) = \log \frac{p(w, c)}{p(w)p(c)}$$

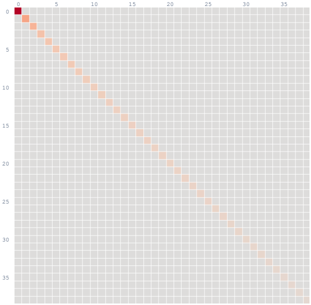


SVD of Wikipedia Character PMI Matrix

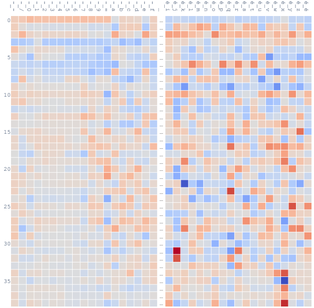
U



Σ

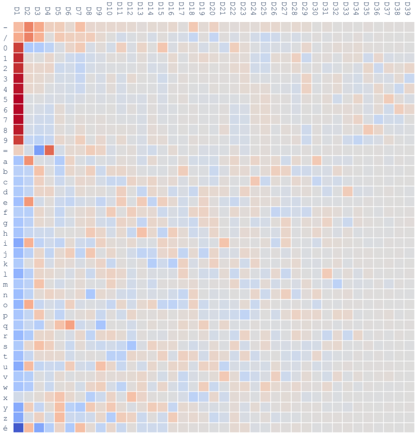


V^T



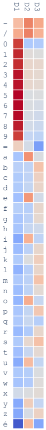
Truncate

$$U \times \Sigma$$



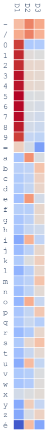
Truncate

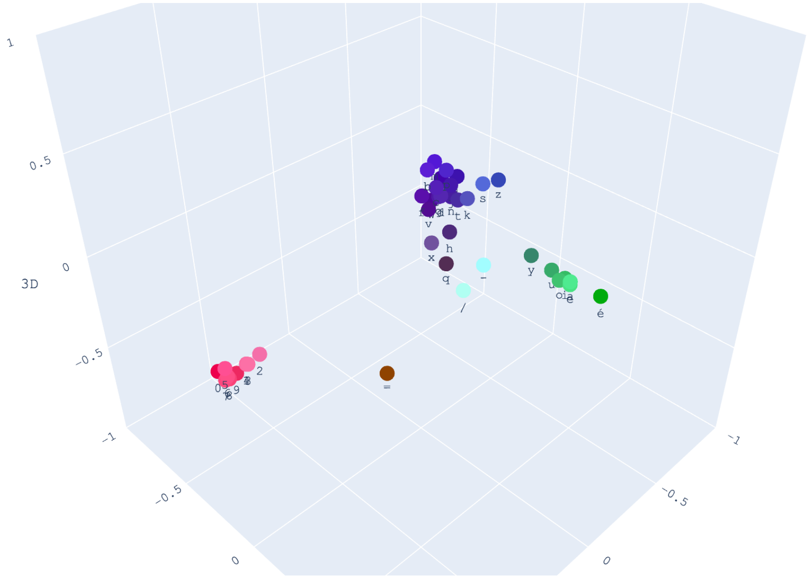
$$\hat{U} \times \hat{\Sigma}$$



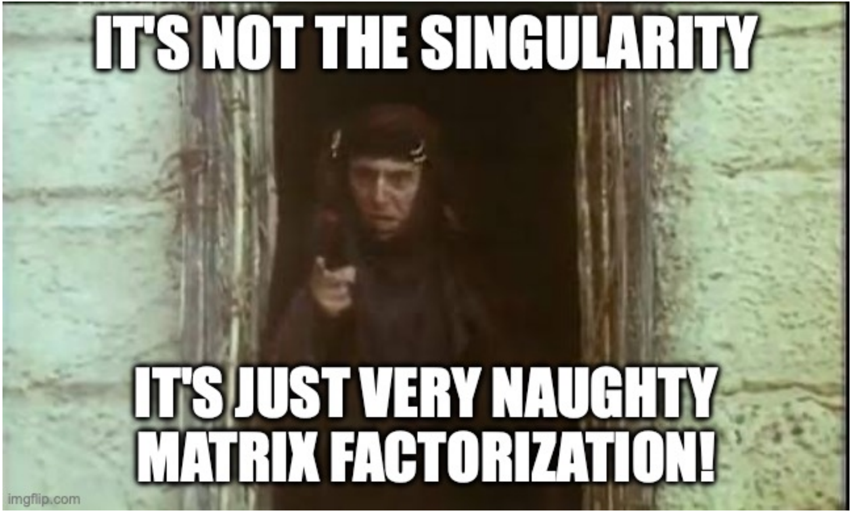
Plot

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





What to conclude?

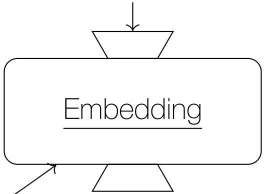


The Structure of Embeddings

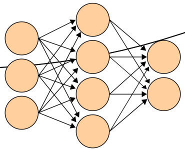
Structure

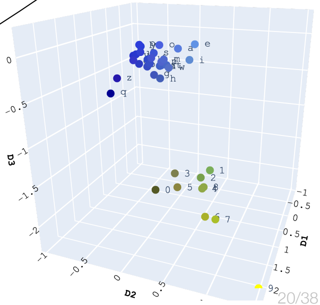
?

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



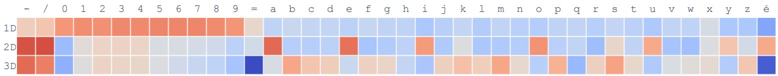
Data



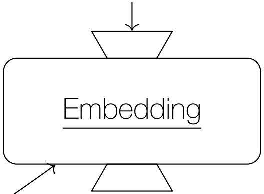


The Structure of Embeddings

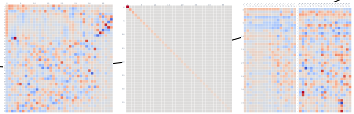
Structure



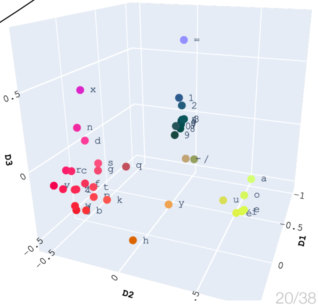
$\{-, /, 0, 1, 2, \ldots, 8, 9, =,$
$a, b, c, \ldots, w, x, y, z, é\}$



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

Introduction

Epistemological Perspectives

Theoretical Perspectives

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

Take Aways

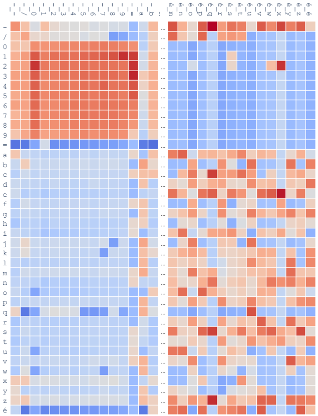
Embeddings as Functions Over Sets

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

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

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

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



Embeddings as Functions Over Sets

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

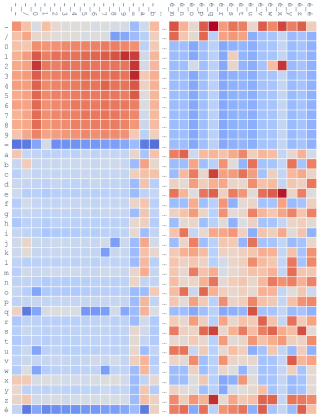
$$\textcolor{blue}{Y} = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M : \textcolor{red}{X} \times \textcolor{blue}{Y} \to \mathbb{R}$$

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

$$M_x : \textcolor{red}{X} \to \mathbb{R}^{\textcolor{blue}{Y}}$$

$$\textcolor{red}{x} \mapsto \textcolor{blue}{M(x, -)}$$



Embeddings as Functions Over Sets

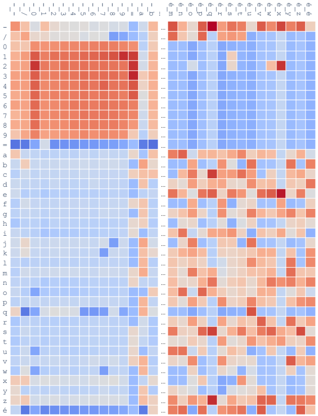
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

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

$$\begin{aligned} M : X \times Y &\to \mathbb{R} \\ (x, y) &\mapsto \operatorname{pmi}(x, y) \end{aligned}$$

$$\begin{aligned} M_x : X &\to \mathbb{R}^Y \\ x &\mapsto M(x, -) \end{aligned}$$

$$\begin{aligned} M_y : Y &\to \mathbb{R}^X \\ y &\mapsto M(-, y) \end{aligned}$$



Embeddings as Functions Over Sets

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

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

$$\begin{aligned} M : X \times Y &\to \mathbb{R} \\ (x, y) &\mapsto \mathrm{pmi}(x, y) \end{aligned}$$

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

$$\begin{aligned} M_x : X &\to \mathbb{R}^Y \\ x &\mapsto M(x, -) \end{aligned}$$

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

$$\begin{aligned} M_y : Y &\to \mathbb{R}^X \\ y &\mapsto M(-, y) \end{aligned}$$

Embeddings as Functions Over Sets

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

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

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

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

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

$$\begin{tikzcd} X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \\ \mathbb{R}^X & Y \arrow[l, "M_y"] \arrow[u, hook] \end{tikzcd}$$

Embeddings as Functions Over Sets

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

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

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

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

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

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^{Y} \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^{X} \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

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

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

$$\begin{tikzcd} X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, shift left, "M_*"] \\ \mathbb{R}^X \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook] \end{tikzcd}$$

Embeddings as Functions Over Sets

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

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

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

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

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

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$

$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

Embeddings as Functions Over Sets

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

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

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$

$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$

$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

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

$$U := [u_1, \ldots, u_m]$$

$$V := [v_1, \ldots, v_n]$$

$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

Embeddings as Functions Over Sets

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

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

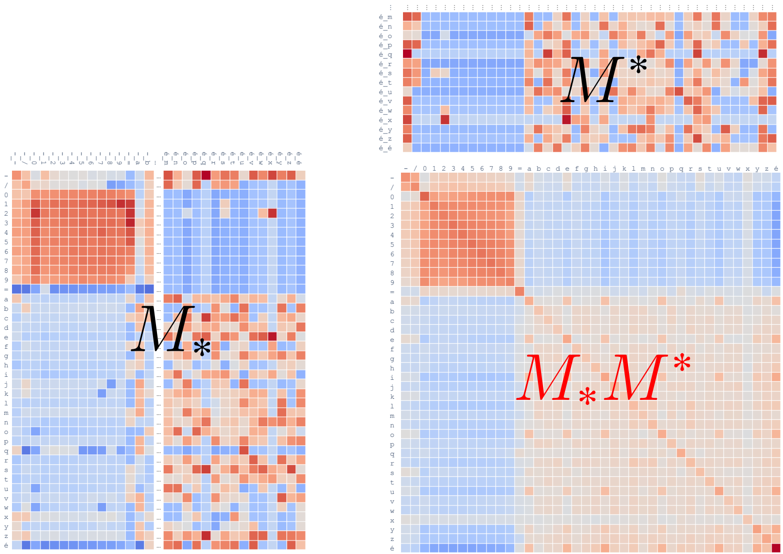
$$M_* M^* u_i = \lambda_i u_i$$
$$M^* M_* v_i = \lambda_i v_i$$

The u_i and v_i are (linear) **fixed points**!

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, shift left, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}$$

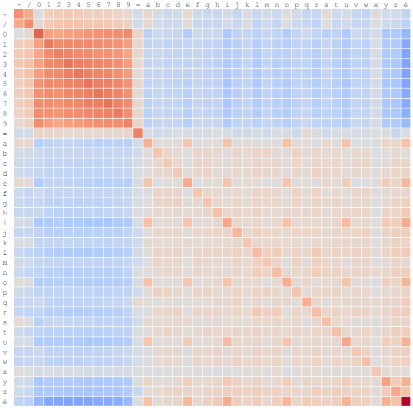
$$M = U \Sigma V^T \qquad \begin{aligned} U &:= [u_1, \ldots, u_m] \\ V &:= [v_1, \ldots, v_n] \\ \Sigma &:= \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix} \end{aligned}$$

$M_* M^*$ as a Covariance Matrix

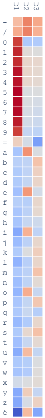


Eigenvectors as Fixed Points

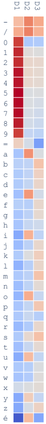
$$M_* M^* u = \lambda u$$



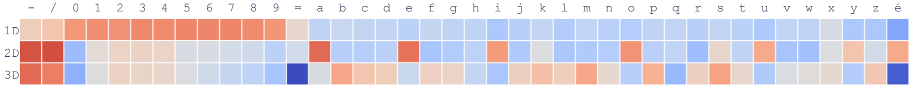
$\times$



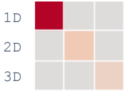
$= \lambda$



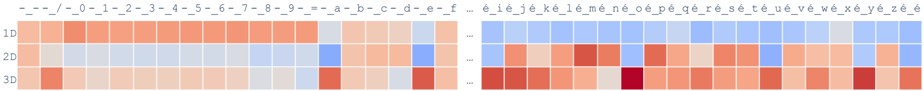
Eigenvectors of $M_* M^*$:



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



Eigenvectors of $M^* M_*$:



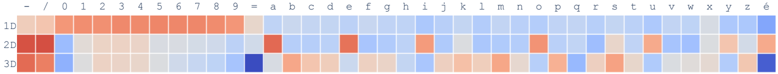
Words

	-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

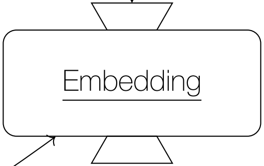
Outline

The Structure of Embeddings

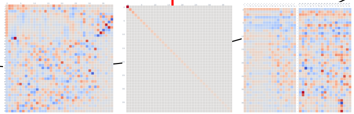
Structure



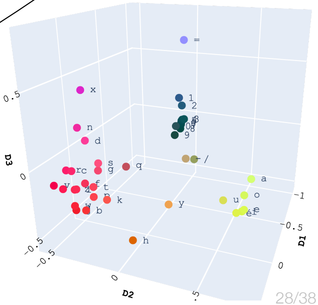
$$\{-, /, 0, 1, 2, \ldots, 8, 9, =, \\ a, b, c, \ldots, w, x, y, z, é\}$$



Data

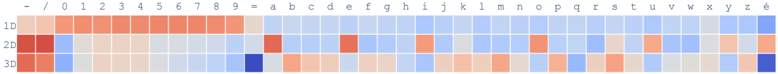


SVD

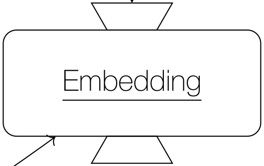


The Structure of Embeddings

Structure

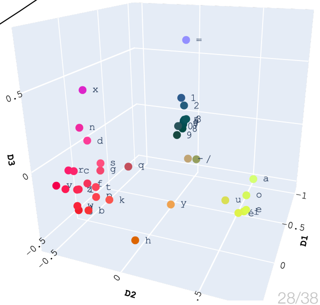


$\{-, /, 0, 1, 2, \ldots, 8, 9, =, a, b, c, \ldots, w, x, y, z, é\}$



Data

$$
\begin{tikzcd}
{\color{red}X} \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^{\color{blue}Y} \arrow[dl, dashed, shift left, "M_*"] \\
\mathbb{R}^{\color{red}X} \arrow[ur, dashed, shift left, "M^*"] & {\color{blue}Y} \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}
$$

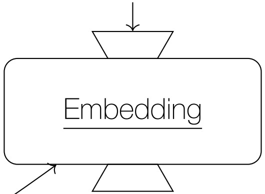


The Structure of Embeddings

Structure

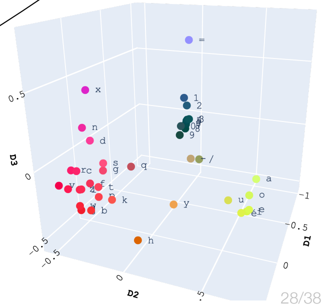
?

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



Data

$$
\begin{tikzcd}
C \arrow[r, "\mathcal{M}_c"] \arrow[d, "Yoneda"'] & (\mathsf{Set}^{D})^{op} \arrow[dl, dashed, shift left, "\mathcal{M}_*"] \\
\mathsf{Set}^{C^{op}} \arrow[ur, dashed, shift left, "\mathcal{M}^*"] & D \arrow[l, "\mathcal{M}_d"] \arrow[u, "Yoneda"']
\end{tikzcd}
$$



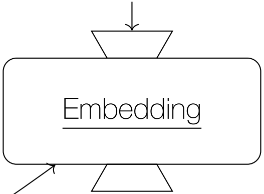
The Structure of Embeddings

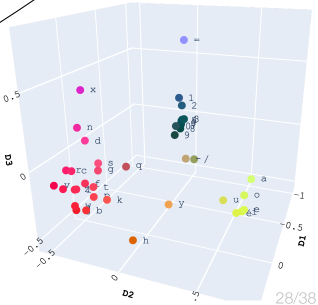
Structure

?

$\{-, /, 0, 1, 2, \ldots, 8, 9, =,$

$a, b, c, \ldots, w, x, y, z, é\}$





Data

$$C^{\mathrm{op}} \times D \longrightarrow \mathbf{Set}$$

Structure

?

$$\underbrace{\mathsf{C}^{\mathrm{op}}}_{}\times\mathsf{D}\longrightarrow\mathsf{Set}$$

$term_i$ → C^{op}

$context_i$ → D

measure → Set

Nucleus and Types

Structure

?

$term_i$ $\qquad$ $context_i$ $\qquad$ measure

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

Structure

?

$$\underset{term_i}{\mathsf{C}^{\mathrm{op}}} \times \underset{context_i}{\mathsf{D}} \longrightarrow \underset{\text{measure}}{\mathsf{2}}$$

Structure

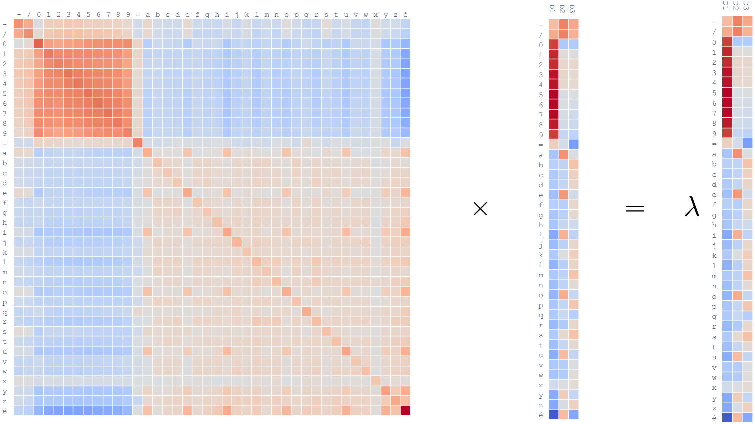
$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow 2$$

$$\Downarrow$$

$$\mathcal{M}^* : 2^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(2^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

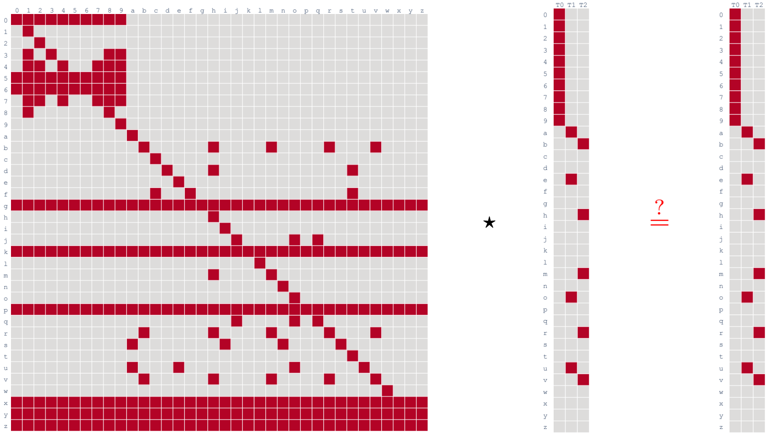
Binary Fixed Points

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



Binary Fixed Points

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

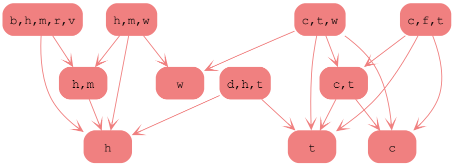


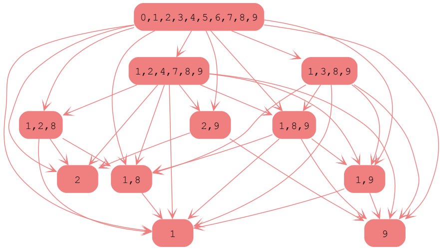
“Eigensets”

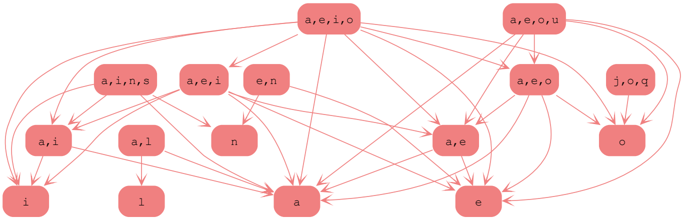
$$\mathcal{M}_*\mathcal{M}^* f = f$$

0,1,2,3,4,5,6,7,8,9	1,2,4,7,8,9	b,h,m,r,v	a,e,i,o	a,e,o,u	a,i,n,s	1,3,8,9
1,2,8	h,m,w	1,8,9	d,h,t	j,o,q	c,f,t	c,t,w
a,e,o	a,e,i	h,m	2,9	a,i	w	1,9
1,8	a,e	l	t	n	c	h
2	i	e	a	o	1	9
e,n	a,l	c,t				

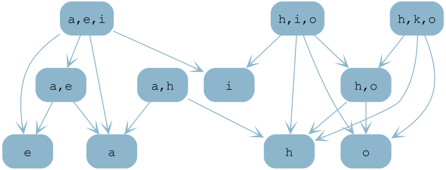
Partial Order Structure

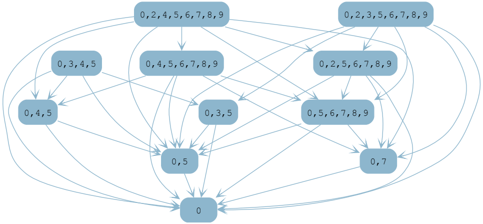


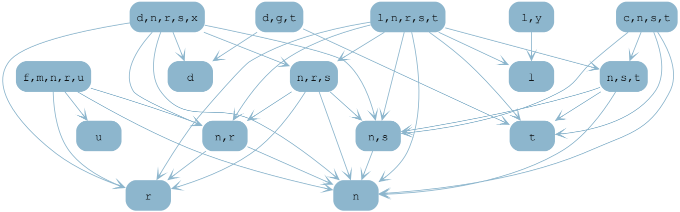




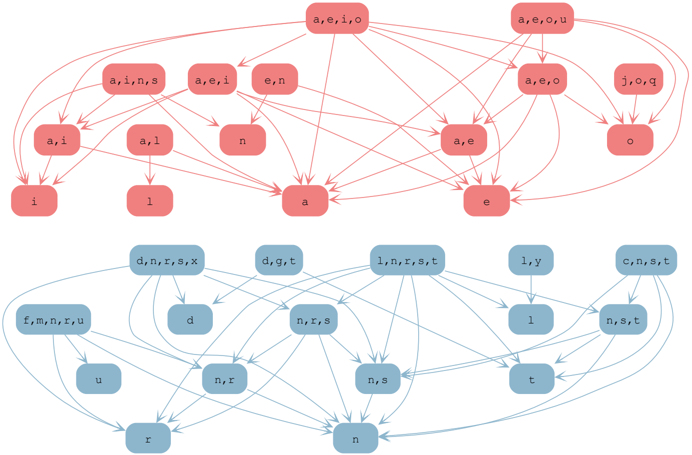
Dual Partial Order





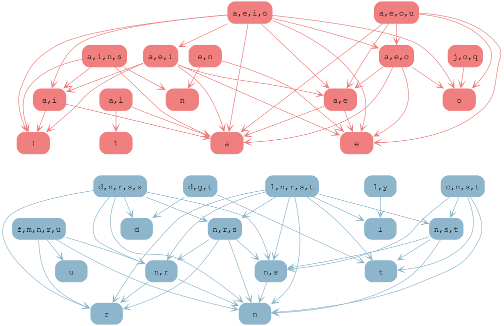


Paring of Partial Ordered Fixed Points



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

Structure



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

$$\Downarrow$$

$$\mathcal{M}^* : \mathsf{2}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows \left(\mathsf{2}^{\mathsf{D}}\right)^{\mathrm{op}} : \mathcal{M}_*$$

Structure

?

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

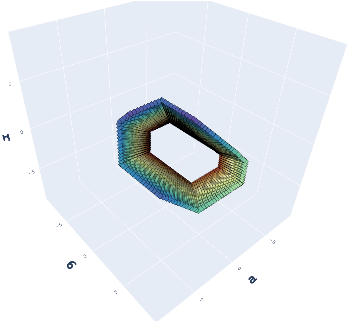
$$\Downarrow$$

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

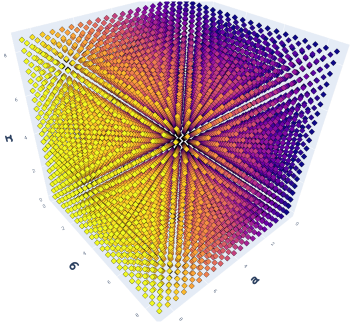
Enriching over $\bar{\mathbb{R}}$

Structure

?



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

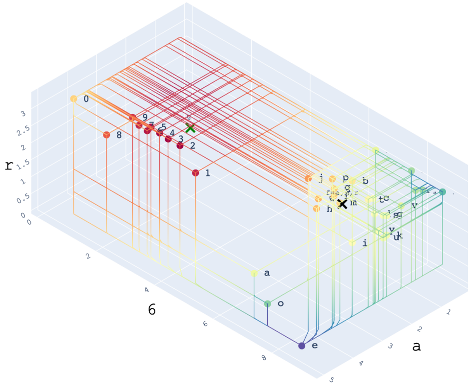


$$\mathsf{C}^{\mathrm{op}} \times \mathsf{D} \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure

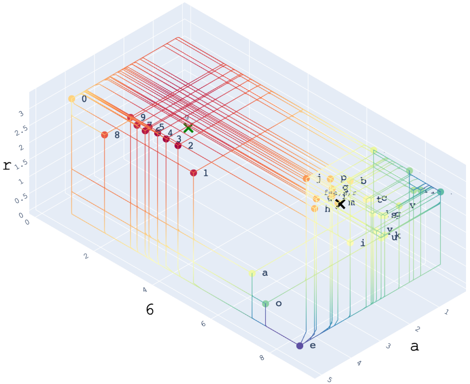


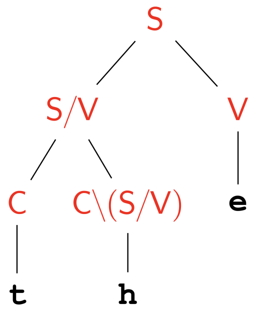
$$C^{\mathrm{op}} \times D \longrightarrow \bar{\mathbb{R}}$$

$$\Downarrow$$

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

Structure





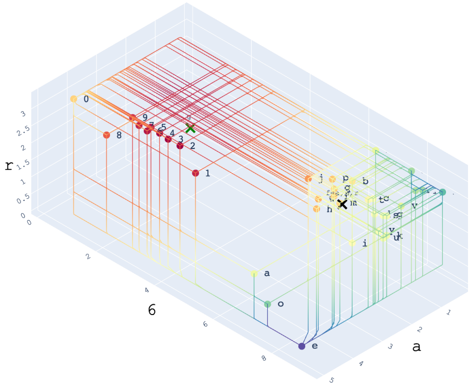
$$\color{orange}{C}^{\mathrm{op}} \times \color{lime}{D} \longrightarrow \color{red}{\bar{\mathbb{R}}}$$

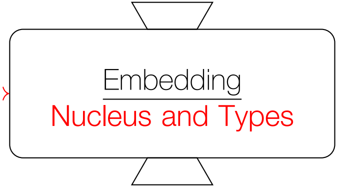
$$\Downarrow$$

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

Formal Explainability

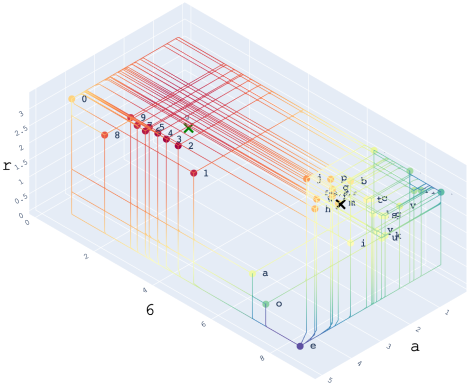
Structure





Formal Explainability

Structure



Tokenization
Renormalization

Embedding
Nucleus and Types

Attention
Tensor Product

Outline

Introduction

Epistemological Perspectives

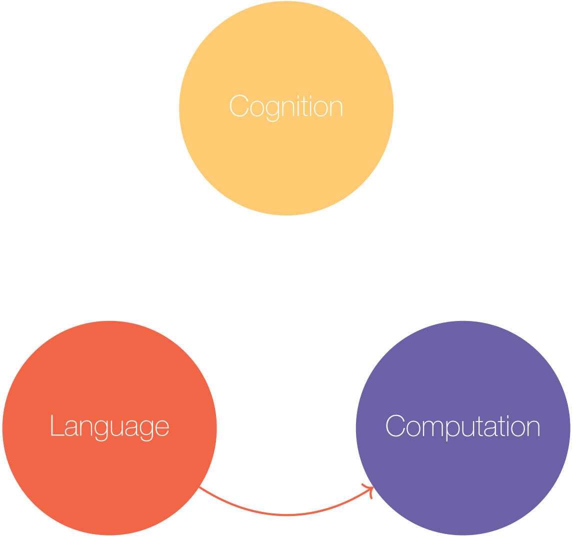
Theoretical Perspectives

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

Take Aways

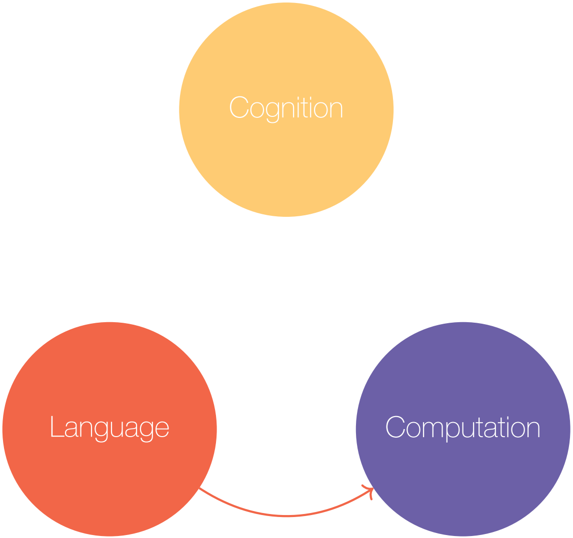
Take Aways

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



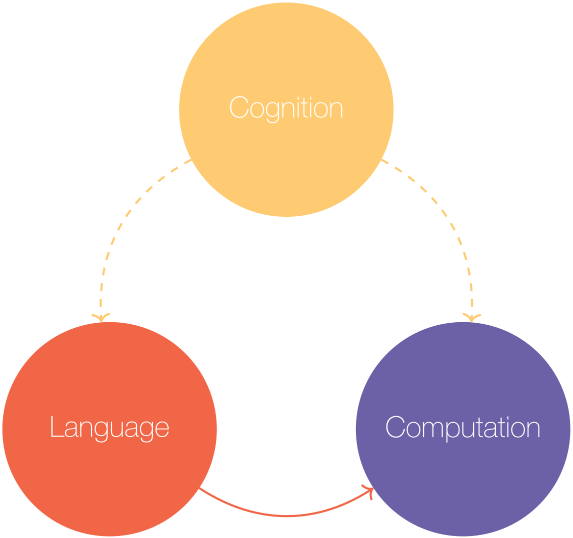
Take Aways

- 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.



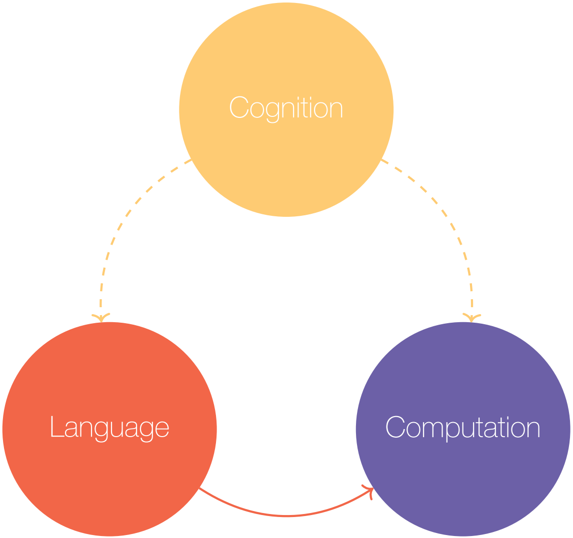
Take Aways

- 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.



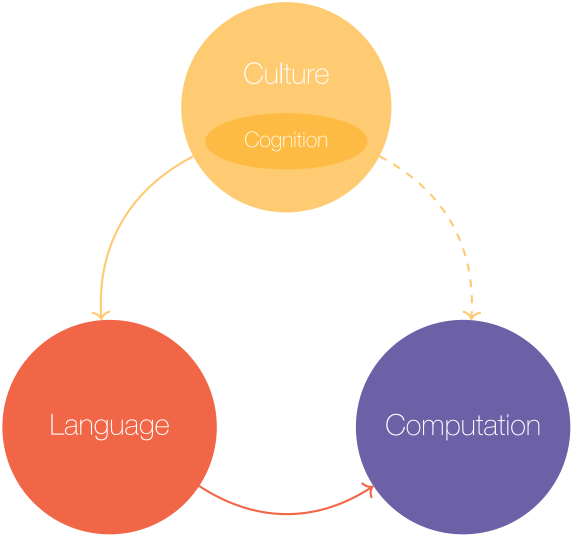
Take Aways

- 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.

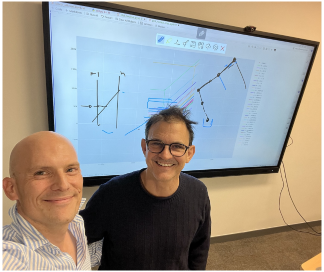


Take Aways

- 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

Empirical Research Pattern

Are Large Language Models Reliable Judges?
A Study on the Factuality Evaluation Capabilities of LLMs

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN
Dialpad Canada Inc.
{xue-yong,tahmid.rahman,cchen,sbhushan}@dialpad.com

Empirical Research Pattern

Are Large Language Models Reliable Judges?
A Study on the Factuality Evaluation Capabilities of LLMs

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN
Dialpad Canada Inc.
{xue-yong,tahmid.rahman,cchen,sbhushan}@dialpad.com

Present Work. We study whether LLMs can detect content that would be persuasive to individuals with specific demographics or beliefs. We center our investigation around three research questions. Namely, can LLMs...

- **RQ1:** judge the quality of arguments and identify convincing arguments and humans?
- **RQ2:** judge how demographics and beliefs influence people's stances on specific topics?
- **RQ3:** determine how arguments appeal to individuals depending on their demographics?

Empirical Research Pattern

Are Large Language Models Reliable Judges?
A Study on the Factuality Evaluation Capabilities of LLMs

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN
Dialpad Canada Inc.
{xue-yong,tahmid.rahman,cchen,sbhushan}@dialpad.com

Pres
tect c
with
our i
Nam

2 Related Work

We review related work in three broad directions closely related to the tasks proposed.

Demographics, beliefs, and persuasion. Demographics have long been known to impact people's political beliefs and attitudes. Group-level demographic factors such as race, religion, and education shape individuals' perspectives on various political issues and voting behavior in the U.S. context (Campbell et al., 1960; Erikson and Tedin, 2019). For example, 78% of Black, 72% of Asian, and 65% of Hispanic workers see efforts on increasing diversity, equity, and inclusion at work positively, compared to 47% of White workers (Minkin, 2023). Similarly, previous work indicates that persuasion depends on the message recipients' existing values and that individual differences can influence persuasion (O'Keefe, 2015).

Empirical Research Pattern

Are Large Language Models Reliable Judges?
A Study on the Factuality Evaluation Capabilities of LLMs

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN
Dialpad Canada Inc.
{xue-yong,tahmid.rahman,cchen,sbhushan}@dialpad.com

Pres
tect c
with
our i
Nam
•
•
•

2 Related Work

We r
close

Dem
grapl
ple's
demc
educ
ous p
conte
din,
Asiaı
on ir
work
ers (
dicat
recip
ferences can influence persuasion (O'Keefe, 2015).

System Role	*You are an expert debate judge with experience in determining who in a debate made more convincing arguments. Today's date is **\$debate-date\$**. You have no information on any events that happened after this date. You have no access to information released after this date.*
Context	*Consider the following proposition:*
	\$proposition\$
	I will give you a debate on the above proposition containing rounds of a 'Pro' and 'Con' argument in JSON format. Here is the debate:
	\$debate\$
Question	*Respond with 'Pro', 'Con', or 'Tie' based on which of the debaters you believe made more convincing arguments.*
Constraint	*Even if you are uncertain, you must answer with either 'Pro', 'Con' or 'Tie' without using any other words or punctuation.*

Figure 2: Prompt structure used in **RQ1**.

Empirical Research Pattern

Are Large Language Models Reliable Judges?
A Study on the Factuality Evaluation Capabilities of LLMs

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN
Dialpad Canada Inc.
{xue-yong,tahmid.rahman,cchen,sbhushan}@dialpad.com

Pres
tect c
with
our i
Nam

-
-
-

2 Related Work

We r
close

Dem
grapl
ple's
demo
educ
ous p
conte
din, 2
Asiar
on ir
work
ers (
dicat
recip
ferences can influence persuasion (O'Keefe, 2015).

#	Question	Model	Dataset	Accuracy (%)	95% Confidence Interval
1	**RQ1**	Llama 2	**PP**	24.91	(20.65, 26.53)
2		Mistral 7B	**PP**	37.69	(32.89, 39.26)
3		GPT-3.5	**PP**	42.74	(39.38, 46.1)
4		GPT-4	**PP**	60.50	(57.26, 63.87)
5		Stacked	**PP**	61.94	(58.54, 65.34)
6		Majority	**PP**	60.69	(59.56, 61.79)
7		Random	**PP**	33.33	—
8	**RQ2**	Llama 2	**IS**	41.56	(38.16, 45.1)
9		Mistral 7B	**IS**	41.39	(38.4, 44.86)
10		GPT-3.5	**IS**	41.73	(38.52, 44.98)
11		GPT-4	**IS**	42.82	(39.59, 46.53)
12		MTurk	**IS**	39.32	(35.89, 42.88)
13		Stacked	**IS**	45.91	(40.02, 51.81)
14		XGBoost	**IS**	55.34	(50.45, 60.22)
15		Random	**IS**	33.33	—
16	**RQ3**	Llama 2	**IS**	41.24	(37.2, 44.02)
17		Mistral 7B	**IS**	42.28	(32.06, 38.28)
18		GPT-3.5	**IS**	38.97	(35.41, 42.11)
19		GPT-4	**IS**	44.38	(40.91, 47.73)
20		MTurk	**IS**	39.86	(36.44, 43.42)
21		Stacked	**IS**	46.86	(41.17, 52.55)
22		Random	**IS**	33.33	—

Empirical Research Pattern

Are Large Language Models Reliable Judges?
A Study on the Factuality Evaluation Capabilities of LLMs

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, Shashi Bhushan TN
Dialpad Canada Inc.
`{xue-yong,tahmid.rahman,cchen,sbhushan}@dialpad.com`

2 Related Work

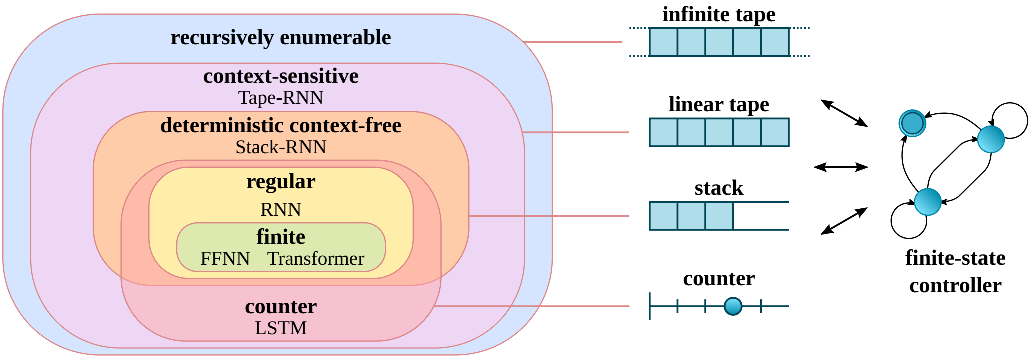
ferences can influence persuasion (O'Keefe, 2015).

6 Discussion and Conclusion

Here we studied LLM's persuasive capabilities by considering its ability to identify convincing arguments in general and for people with specific arguments. We argue that if LLMs can detect content that is highly persuasive to specific demographics, they may be used to detect and amplify tailor-made misinformation and propaganda. Our findings indicate that LLMs demonstrate human-level performance in (1) judging argument quality, (2) predicting users' stances on specific topics given users' demographics and basic beliefs, and (3) detecting arguments that would be persuasive to individuals with specific demographics or beliefs.

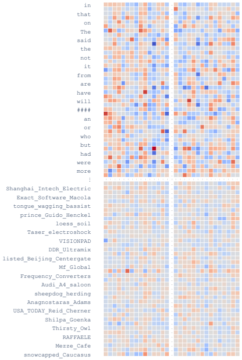
NNs and the Chomsky Hierarchy

(Delétang et al., 2023)



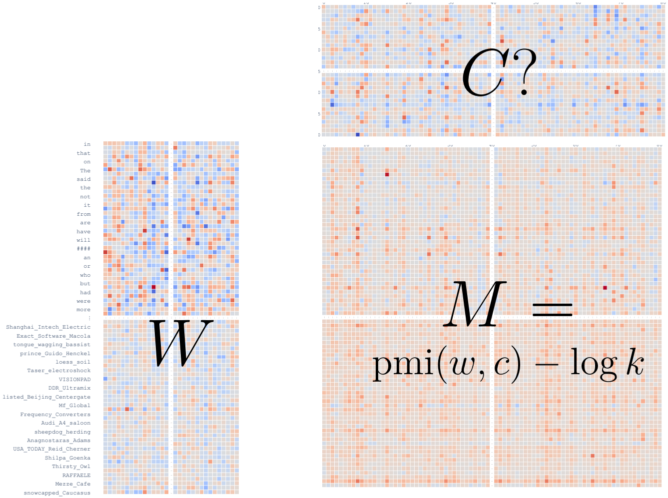
word2vec as Implicit Matrix Factorization

(Levy and Goldberg, 2014)



word2vec as Implicit Matrix Factorization

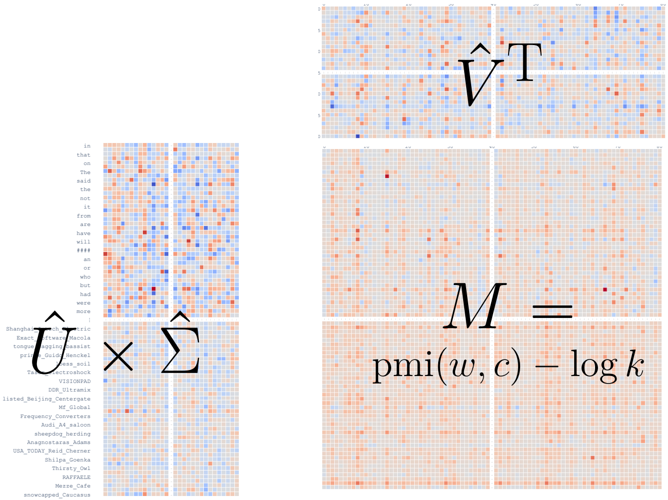
(Levy and Goldberg, 2014)



$$W \times C \approx M$$

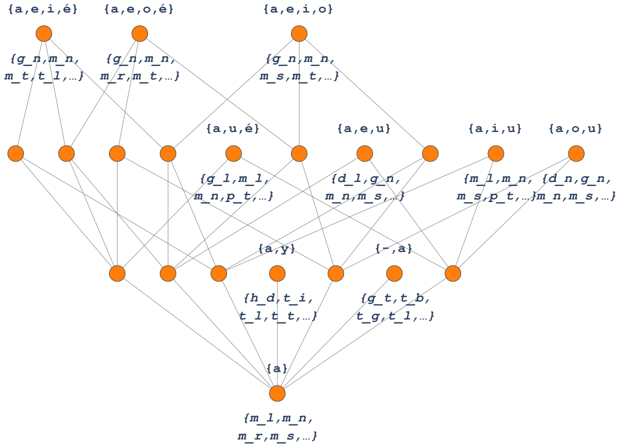
Word Embeddings as Truncated SVD

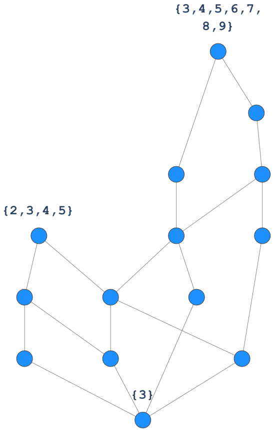
(Levy and Goldberg, 2014)



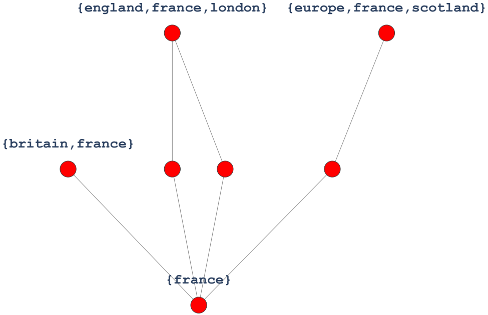
$$M \approx \hat{U} \times \hat{\Sigma} \times \hat{V}^{\mathrm{T}}$$

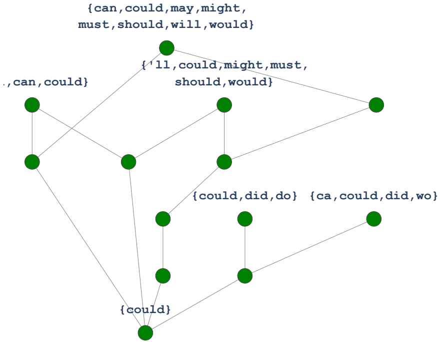
Formal Concepts



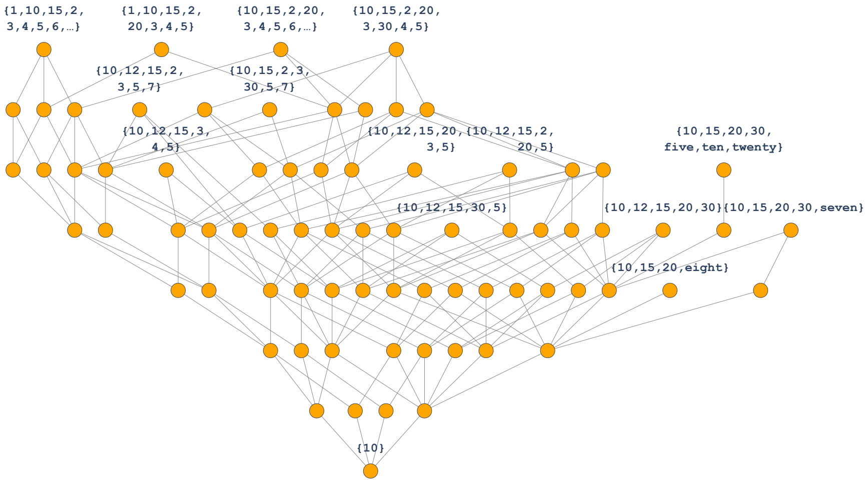


Formal Concepts (words)



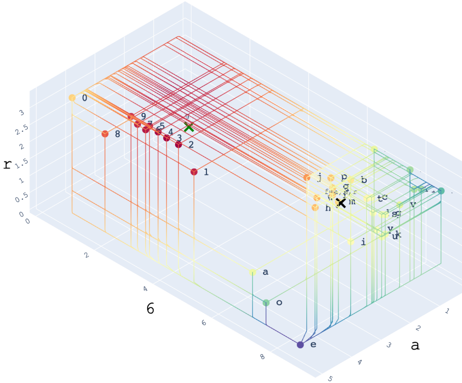


Formal Concepts (words)



Theoretical Interpretability

Structure



Theory

“Task”

?

Theoretical Interpretability

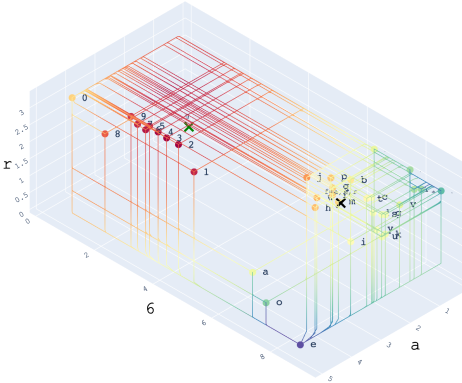
$$\mathcal{C}^{\mathrm{op}} \times \mathcal{D} \longrightarrow \bar{\mathbb{R}}$$

Distributional Hypothesis

The content of linguistic units is determined by their *distribution* in a corpus.

Theory
“Task”

Structure



Theoretical Interpretability

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

Distributional Hypothesis

The content of linguistic units is determined by their *distribution* in a corpus.

$$\Downarrow$$

Theory

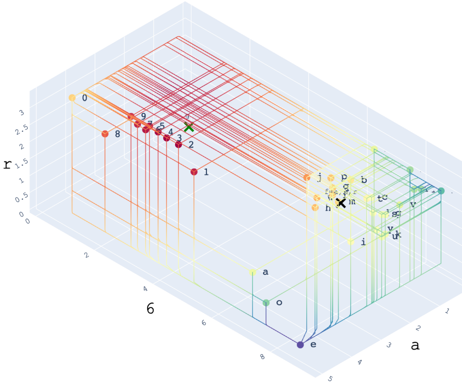
“Task”

Structuralist Hypothesis

Linguistic content is the effect of a virtual *structure* underlying linguistic practices within a community

$$\bar{\mathbb{R}}^{C^{\mathrm{op}}} \leftrightarrows \left(\bar{\mathbb{R}}^{D}\right)^{\mathrm{op}}$$

Structure



From Linear Algebra to Categories

$$\begin{tikzcd}
X \arrow[r, "M_x"] \arrow[d, hook] & \mathbb{R}^Y \arrow[dl, dashed, bend left, "M_*"] \\
\mathbb{R}^X \arrow[ur, dashed, "M^*"] & Y \arrow[l, "M_y"] \arrow[u, hook]
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
\mathsf{C} \arrow[r, "\mathcal{M}_c"] \arrow[d, hook, "Yoneda"'] & (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \arrow[dl, dashed, bend left, "\mathcal{M}_*"] \\
\mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \arrow[ur, dashed, "\mathcal{M}^*"] & \mathsf{D} \arrow[l, "\mathcal{M}_d"] \arrow[u, hook, "Yoneda"']
\end{tikzcd}$$

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

$$\mathcal{M}_*\mathcal{M}^* : \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$$
$$\mathcal{M}^*\mathcal{M}_* : (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \to (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}}$$

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

$$\mathrm{Fix}(\mathcal{M}_*\mathcal{M}^*) := \{f \in \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} | \mathcal{M}_*\mathcal{M}^*(f) \cong f\}$$
$$\mathrm{Fix}(\mathcal{M}^*\mathcal{M}_*) := \{g \in (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} | \mathcal{M}^*\mathcal{M}_*(g) \cong g\}$$

Theory of Computational Types

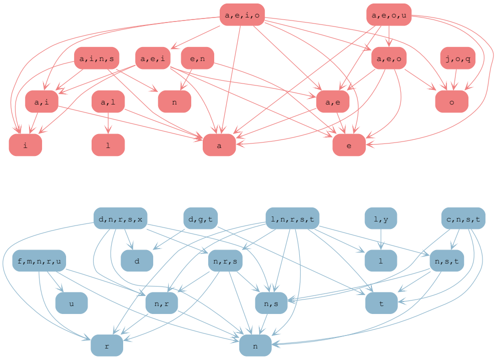
Definition (Polar/Orthogonal - Girard, 2011)

[G]iven a binary operation, noted $a, b \rightsquigarrow \langle a|b\rangle \colon A \times B \to C$ and a subset $P \subset C$ (the 'pole') one can define the *polar* $X^{\perp} \subset B$ of a subset $X \subset A$ (resp. $Y^{\perp} \subset A$ of a subset $Y \subset B$) by :

$$X^{\perp} := \{y \in B \colon \forall x \in X, \langle a|b\rangle \in P\}$$

$$Y^{\perp} := \{x \in A \colon \forall y \in Y, \langle a|b\rangle \in P\}$$

- The map 'polar' is decreasing: $X \subset X' \Rightarrow X'^{\perp} \subset X^{\perp}$.
- The set $\mathrm{Pol}(A) \subset \mathcal{P}(A)$ of *polar* sets, i.e., of the form $Y^{\perp}$, is closed under arbitrary intersections. In particular, A is polar and $X^{\perp\perp}$ is the smallest polar set containing X.
- As a consequence, $X^{\perp\perp\perp} = X^{\perp}$.



Theory of Computational Types

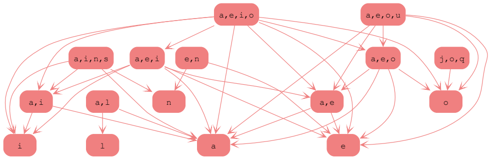
Definition (Polar/Orthogonal - Girard, 2011)

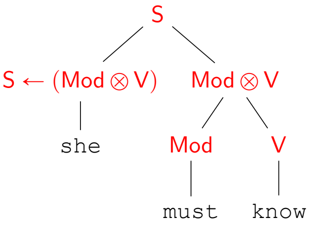
[G]iven a binary operation, noted $a, b \rightsquigarrow \langle a|b\rangle \colon A \times B \to C$ and a subset $P \subset C$ (the 'pole') one can define the *polar* $X^\perp \subset B$ of a subset $X \subset A$ (resp. $Y^\perp \subset A$ of a subset $Y \subset B$) by :

$$X^\perp := \{y \in B \colon \forall x \in X, \langle a|b\rangle \in P\}$$

$$Y^\perp := \{x \in A \colon \forall y \in Y, \langle a|b\rangle \in P\}$$

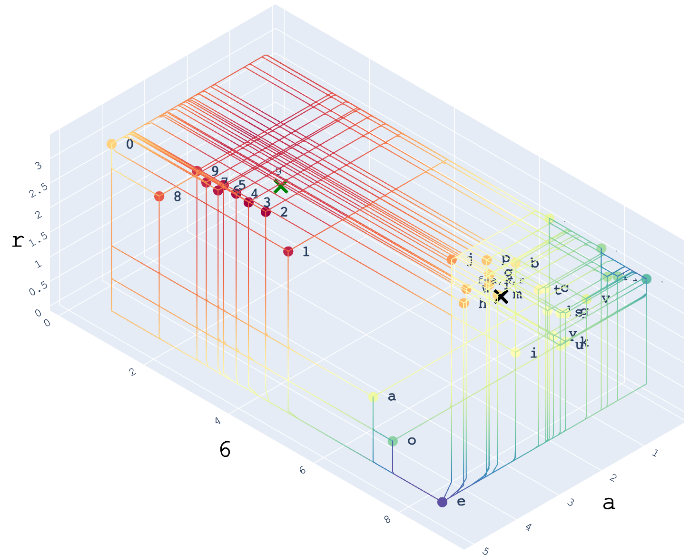
- The map 'polar' is decreasing: $X \subset X' \Rightarrow X'^\perp \subset X^\perp$.
- The set $\mathrm{Pol}(A) \subset \mathcal{P}(A)$ of *polar* sets, i.e., of the form $Y^\perp$, is closed under arbitrary intersections. In particular, A is polar and $X^{\perp\perp}$ is the smallest polar set containing X.
- As a consequence, $X^{\perp\perp\perp} = X^\perp$.





(Gastaldi2021)

Internal Structure of the Nucleus



Matrix and Analogy

a = your c = my	w = apartment x = house y = chair z = stool

your : house
my : apartment

	...	w	x	y	z	...
...	...	0	0	0	0	...
a	...	0	1	1	0	...
b	...	0	0	1	1	...
c	...	1	0	0	1	...
...	...	0	0	0	0	...